HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences

Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology (NAIST), Japan
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

Recently, we have often observed hallucinated citations or references that do not correspond to any existing work in papers under review, preprints, or published papers. Such hallucinated citations pose a serious concern to scientific reliability. When they appear in accepted papers, they may also negatively affect the credibility of conferences. In this study, we refer to hallucinated citations as ***"HalluCitation"*** and systematically investigate their prevalence and impact. We analyze all papers published at ACL, NAACL, and EMNLP in 2024 and 2025, including main conference, Findings, and workshop papers. Our analysis reveals that over 300 papers contain at least one HalluCitation, most of which were published in 2025. Notably, half of these papers were identified at EMNLP 2025, the most recent conference, indicating that this issue is rapidly increasing. Moreover, more than 100 such papers were accepted as main conference and Findings papers at EMNLP 2025, affecting the credibility.

✗ **ACL 2025 Main:** Chang et al. (2025) —

> Y. Zhang and Others. 2024. Subsampling for skill improvement in large language models. arXiv preprint arXiv:2402.12345.
> ✓ Hohloch (2024)

✗ **EMNLP 2025 Findings:** Jalori et al. (2025) —

> Wendi Zhou, Xiao Li, Lin Geng Foo, Yitan Wang, Harold Soh, Caiming Xiong, and Yoonkey Kim. 2024. TEMPO: Temporal representation prompting for large language models in time-series forecasting. arXiv preprint arXiv:2405.18384. Anticipated for NeurIPS 2024. Preprint, arXiv:2405.18384.
> ✓ Shandi et al. (2024)

✗ **EMNLP 2025 Main:** Srivastava (2025) —

> Wei Xu, Yulia Tsvetkov, and Alan Black. 2022. AI for language learning: Conversational agents and personalized feedback. Transactions of the Association for Computational Linguistics (TACL), 10:1–15. ✗ **(Non-existent)**
> ✗ Title Link: Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation (Clark et al., 2022)

Figure 1: Examples of incorrect reference information. Some references include incorrect arXiv IDs or contain uncertain phrases such as "Anticipated for". Even references that appear plausible may contain HalluCitations, e.g., incorrect link information or references to pages where the cited paper does not exist, as in the last case.

1 Introduction

Thanks to the advances in AI for Science research (Sun et al., 2024; Chen et al., 2025d; Shi et al., 2025) and AI Scientist technologies (Wei et al., 2025b; Huang et al., 2025; Yamada et al., 2025; Tang et al., 2025a; Shao et al., 2025; Xu et al., 2025b), academic research and writing have shifted from traditionally human-centered processes to ones supported by and actively leveraging AI systems such as large language models (LLMs). In particular, LLMs are enabling a wide range of tools for writing assistance (Hwang et al., 2025), literature search (Asai et al., 2024), citation recommendation (Şahinuç et al., 2024a; Çelik and Tekir, 2025), idea discussion (Zheng et al., 2025d), peer-review support and analysis (Jin et al., 2024; Starace et al., 2025; Hossain et al., 2025), and other scientific automation (Şahinuç et al., 2024b). These benefits have been especially significant for non-native English speakers, as AI tools can improve translation and grammatical quality and help create fairer opportunities for participation in the international academic community (Lepp and Smith, 2025).

However, while these advances have boosted research productivity, the research community has begun to face new challenges arising from the exponential growth in paper submissions, particularly in securing qualified reviewers. As a result, the qualification bar has been lowered[1], and reviewers are often required to evaluate many papers on unfamiliar topics within a limited time. Moreover, recent review processes involve potential penalties

[1] https://aclrollingreview.org/incentives2025

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47295–47376
July 2-7, 2026

for delayed reviews, such as desk rejection of a reviewer's own submissions[1]. Under such pressure and concern, the focus of peer review has shifted from conducting thorough and rigorous evaluations to completing reviews under strict time constraints. Consequently, we face growing difficulty in carefully inspecting the rigor and factual correctness of submitted manuscripts. In particular, as shown in Figure 1, we have recently observed non-existent citations, i.e., hallucinated references that do not correspond to any real prior work. Such false citations pose a serious threat to scientific reliability and undermine respect for existing work. When they appear in published papers, they may also negatively affect the overall quality and credibility of academic conferences and journals.

In this study, we refer to such hallucinated references as ***"Hallu<u>Citation</u>"*** (Hallucination + Citation, /həˌluː.saɪˈteɪ.ʃən/). Papers that include at least one HalluCitation are referred as ***"HalluCited"*** papers (Hallucination + Cited, /ˈhæl.uːˌsaɪ.tɪd/). We systematically investigate their prevalence and impact at top-tier NLP conferences. We analyze over 17,000 papers published at ACL, NAACL, and EMNLP in 2024 and 2025, including main conference, Findings, and workshop papers. Our analysis and discussion reveal the following findings:

- Over 300 papers contain at least one HalluCitation, mostly published in 2025. Notably, half of these papers were identified at EMNLP 2025, the most recent conference, indicating that this issue is rapidly increasing. Moreover, over 100 such papers are accepted as main and Findings papers in EMNLP 2025, affecting the credibility.
- We propose a HalluCitation detection method based on OCR and database matching, and show that it effectively supports our analysis in identifying actual HalluCitations. In particular, papers with four or more detected candidates exhibit a high incidence of HalluCitations, suggesting a practical guideline for integration into automated detection toolkits.
- Even when HalluCitations are present, they cannot be immediately attributed to AI-generated content, as they may also originate from secondary sources, making their causes more complex. Accordingly, the presence of HalluCitations alone should not lead to immediate penalties for authors. Instead, it is important to establish prior consensus on preventive measures, such as author toolkits for pre-submission checks.

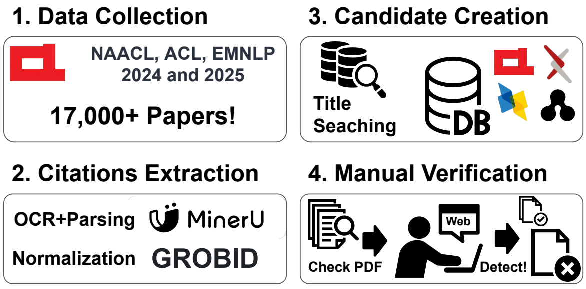


Figure 2: Our analysis method. First, we extract citations from the collected PDFs. Next, we attempt to identify the corresponding papers in a reference paper database using fuzzy matching on citation titles, and flag citations for which no matching paper can be found as HalluCitation candidates. Finally, we manually verify the existence of each candidate by referring to the original PDFs and conducting web searches. If it cannot be objectively identified, we mark it as a HalluCitation.

Year	Venue	Main	Findings	S.I.D.	WS	Total
2024	NAACL	562	296	94	609	1,561
2024	ACL	941	975	84	721	2,721
2024	EMNLP	1,268	1,003	173	510	2,954
2025	NAACL	718	474	177	560	1,929
2025	ACL	1,699	1,387	259	1,165	4,510
2025	EMNLP	1,809	1,405	269	684	4,167
Overall		6,997	5,540	1,056	4,249	17,842

Table 1: Number of PDFs. S.I.D. denotes the total number of papers from the Student Research Workshop, Industry Track, and Demonstration Track. Note that the Industry Track was not held at ACL 2024, and the Student Research Workshop was not held at EMNLP. Conferences are ordered chronologically by their dates.

2 Analysis Methods

This investigation is similar to finding a needle in a haystack, as it requires checking an enormous number of citations to detect a small number of HalluCitation. Therefore, manually verifying all citations is impractical. To address this challenge, as shown in Figure 2, we exploit the assumption that *"the majority of citations are correct"* and systematically list up candidate HalluCitation. Then, we manually verify each candidate to identify HalluCitation. Note that, since the actual number of HalluCitation is unknown, our reported results should be interpreted as a lower bound, and additional HalluCitation may exist in the wild.

2.1 Data Collection

In this study, we target all papers presented at a total of six conferences, NAACL, ACL, and EMNLP, held in 2024 and 2025, including main confer-

Year	Venue	Mean	Std.	Q1	Q2	Q3	Total Citations
2024	NAACL	36.33	20.48	21	35	48	56,717
2024	ACL	41.26	24.04	27	39	52	112,262
2024	EMNLP	42.91	23.90	29	40	53	126,744
2025	NAACL	39.33	22.42	25	37	50	75,870
2025	ACL	42.63	25.94	27	40	54	192,245
2025	EMNLP	42.67	27.68	28	39	52	177,818
Overall		41.57	25.03	27	39	52	741,656

Table 2: Number of extracted citations. We report the mean, standard deviation, quartiles (Q1, Q2=median, Q3), and the total number of citations for each venue.

ence papers, Findings papers, and all workshop papers. We collected all PDF files and metadata of archival papers registered as events in the ACL Anthology[2], excluding tutorials and proceedings. Table 1 presents the statistics of the collected PDFs. In total, our dataset covers over 17,000 papers.

2.2 Citation Extraction

Next, we extract citation information from the PDFs. Through pilot studies, we found that a pipeline combining OCR-based extraction of raw bibliographical reference strings with normalization of the corresponding bibliographical references yields the most efficient citation extraction. Accordingly, first, we use MinerU (Wang et al., 2024a) to extract reference entries from the references section at the text-block level. Then, we apply GROBID (GRO, 2008–2025) to parse and normalize each reference entry, obtaining structured bibliographic information such as titles[3]. Table 2 shows the statistics of the extracted citation counts. From these results, we observe that the median number of citations per paper is 39 and that the quartiles remain stable across venues. In total, we extracted over 740,000 citations.

2.3 Candidate Creation

As shown in Table 2, the total number of extracted citations is extremely large, making it impractical to manually inspect all citations. Therefore, we focus on citations that explicitly indicate references to papers from the ACL Anthology or arXiv, and extract a subset of candidate citations by matching them against reference databases.

First, we heuristically narrow down the target citations by checking whether the raw citation strings contain keywords related to ACL conferences or arXiv, such as ACL, EMNLP, or arXiv. Next, we search for corresponding papers primarily using metadata from the ACL Anthology[4] and arXiv[5]. In addition, since this heuristic filtering serves as a coarse screening, citations to papers outside ACL venues or arXiv may still remain. To further filter such cases, we additionally leverage DBLP[6] and OpenAlex (Priem et al., 2022)[7].

For database matching, we employ character-level fuzzy matching on citation titles using a similarity score based on the normalized Levenshtein distance with the `RapidFuzz` library (Bachmann, 2025). Specifically, we compute title similarity as $1 - d_{\text{lev}} / \max(|s_1|, |s_2|)$, where s_1 and s_2 are the character strings being compared, d_{lev} denotes the Levenshtein edit distance (Levenshtein, 1965), and scale it to the range $[0, 1]$. If no title in the reference databases achieves a similarity score of 0.9 or higher, we regard the citation as potentially non-existent and include it in a list of candidates for manual verification.

Year	Venue	#Papers	#Citations	Ave.	Max
2024	NAACL	211 (13.52%)	268 (0.47%)	1.27	6
2024	ACL	381 (14.00%)	476 (0.42%)	1.25	4
2024	EMNLP	443 (15.00%)	556 (0.44%)	1.26	7
2025	NAACL	336 (17.42%)	465 (0.61%)	1.38	10
2025	ACL	796 (17.65%)	1,117 (0.58%)	1.40	17
2025	EMNLP	783 (18.79%)	1,222 (0.69%)	1.56	31
2024		1,035 (14.30%)	1,300 (0.44%)	1.26	7
2025		1,915 (18.06%)	2,804 (0.63%)	1.46	31
Overall		2,950 (16.53%)	4,104 (0.55%)	1.39	31

Table 3: Number of candidate papers and citations. We also report the proportion of candidate papers and citations relative to the total. Ave. denotes the average number of candidate citations per paper, while Max denotes the maximum number of candidate citations observed in a single paper.

Observation. Table 3 shows the statistics of candidate citations and papers. From these results, we

[2] https://aclanthology.org/

[3] Although GROBID can directly extract bibliographic information from PDFs, our pilot studies indicate that its citation extraction is not sufficiently robust for ACL-style reference formats. Specifically, when reference lists span multiple pages, GROBID frequently fails to extract references appearing on subsequent pages, leading to under-coverage. For instance, in some cases, only around 10 citations are extracted from papers that actually contain 50 references. To mitigate this, we explicitly provide reference blocks as input to GROBID.

[4] https://github.com/acl-org/acl-anthology/tree/c9d3481/data

[5] Clement et al. (2019), https://www.kaggle.com/datasets/Cornell-University/arxiv/versions/263

[6] https://drops.dagstuhl.de/entities/artifact/10.4230/dblp.xml.2025-12-01

[7] We use the REST APIs via `PyAlex` (De Bruin, 2023).

observe that both the average number of candidate citations per paper and the maximum number increase over time. In addition to the raw counts of citations and papers, the corresponding percentages also show a consistent upward trend. In particular, when comparing 2024 and 2025, we observe different trends in the proportions of targeted papers and citations. These observations suggest that, although our coarse screening and simple database matching may introduce some noise, the increase in candidate citations exceeds what can be explained by such methodological noise alone. This indicates that factors beyond the limitations of our method are likely contributing to the observed increase, most notably the effect of ***HalluCitation***.

Year	Venue	Main	Findings	S.I.D.	WS	Total
2024	NAACL	$1^{0.18}_{50.0}$	$0^{0.00}_{0.00}$	$0^{0.00}_{0.00}$	$1^{0.16}_{50.0}$	$2^{0.12}_{-}$
2024	ACL	$3^{0.31}_{42.9}$	$1^{0.10}_{14.3}$	$0^{0.00}_{0.00}$	$3^{0.42}_{42.9}$	$7^{0.26}_{-}$
2024	EMNLP	$3^{0.24}_{27.3}$	$4^{0.40}_{36.4}$	$2^{1.16}_{18.2}$	$2^{0.39}_{18.2}$	$11^{0.37}_{-}$
2025	NAACL	$6^{0.84}_{17.1}$	$10^{2.11}_{28.6}$	$2^{1.13}_{5.71}$	$17^{3.04}_{48.6}$	$35^{1.81}_{-}$
2025	ACL	$13^{0.77}_{15.1}$	$20^{1.44}_{23.3}$	$9^{3.74}_{10.5}$	$44^{3.78}_{51.2}$	$86^{1.91}_{-}$
2025	EMNLP	$62^{3.43}_{38.8}$	$49^{3.49}_{30.6}$	$13^{4.83}_{8.13}$	$36^{5.26}_{22.5}$	$160^{3.84}_{-}$
2024		$7^{0.25}_{35.0}$	$5^{0.22}_{25.0}$	$2^{0.57}_{10.0}$	$6^{0.33}_{30.0}$	$20^{0.28}_{-}$
2025		$81^{1.92}_{28.8}$	$79^{2.42}_{28.1}$	$24^{3.40}_{8.54}$	$97^{4.03}_{35.3}$	$281^{2.65}_{-}$
Overall		$88^{1.26}_{29.2}$	$84^{1.52}_{27.9}$	$26^{2.46}_{8.63}$	$103^{2.42}_{34.2}$	$301^{1.69}_{-}$

Table 4: Number of papers including ***"HalluCitation"***. We also report the proportion of papers with HalluCitation relative to the total number of papers, as well as track-specific proportions within each venue. We denote these values as $\mathrm{NUM}^{\text{Venue\%}}_{\text{Track\%}}$. Note that in some venues and tracks with only a small number of papers, the percentages may be sensitive to individual papers. Nevertheless, they remain useful for tracking the trends.

2.4 Manual Verification

Finally, we manually verify all candidate citations identified in Section 2.3 to determine whether each citation refers to an actual existing work. Since our goal is to detect papers that contain HalluCitation, we mark a paper as *"HalluCited"* immediately when at least one HalluCitation is identified and terminate further checks for that paper. We apply this procedure to all 2,950 target papers.

If identifiable information, such as a clickable link, DOI, arXiv ID, page numbers, or venue information, is provided in the citation, we use this information to search for the referenced paper. If such information is missing or insufficient to locate the paper, we conduct a web search using the citation title to find potentially corresponding works. We classify a citation as "HalluCitation" if no corresponding paper can be found, e.g., when a search returns only the target paper itself, or if a similar paper is found but at least two key attributes, such as the title, authors, venue, or page numbers, do not match any reliable source. This verification procedure adopts a conservative criterion that basically trusts the papers, prioritizing precision over recall. Appendix B shows the list of HalluCited papers.

3 Results

Finding 1: The number of HalluCited papers increased sharply in 2025, with EMNLP accounting for over half of these papers. Table 4 reports the number of papers that contain HalluCitation. The results show that the number of hallucinated papers increased from 20 in 2024 to 281 in 2025. In particular, EMNLP 2025 alone accounts for 160 papers, nearly doubling the number observed at ACL 2025. Comparing 2024 and 2025, the number of HalluCited papers increased by more than an order of magnitude. The proportion of HalluCited papers also increased from around 0.28% in 2024 to 2.65% in 2025, reaching 3.84% at EMNLP 2025. Examining venue-specific trends in 2025, we find that while workshop papers account for roughly half of the HalluCited papers at NAACL and ACL, nearly 70% of HalluCited papers at EMNLP 2025 appear in the Main and Findings. A similar pattern is observed in the proportions of HalluCited papers across tracks in 2025. At NAACL, only workshop papers exceed 3%, whereas at ACL, the S.I.D. tracks reach 3%. In contrast, at EMNLP, all tracks exceed 3%. These results suggest that the impact of HalluCited papers is no longer confined to workshops or side tracks, but is increasingly affecting main tracks. This finding indicates that the rapid increase in HalluCited papers is less likely to be attributable to deficiencies in specific tracks and instead highlights emerging limitations in the ability of the current review system to detect HalluCitation, revealing its social impact on sustainability and trustworthiness.

Finding 2: Papers containing multiple HalluCitation candidates are highly likely to contain actual HalluCitation. Table 5 shows the detailed counts of candidate papers and the actually detected HalluCited papers. From these results, we find that when a paper contains around four HalluCi-

	Candidates		HalluCited		Hit Rate (%)	
Freq.	Num.	Cum.	Num.	Cum.	Num.	Cum.
≥9	10	10	10	10	100.0%	100.0%
8	6	16	5	15	83.3%	93.8%
7	14	30	13	28	92.9%	93.3%
6	10	40	7	35	70.0%	87.5%
5	9	49	7	42	77.7%	85.7%
4	28	77	17	59	60.7%	76.6%
3	91	168	38	97	41.8%	57.7%
2	526	694	78	175	14.8%	25.2%
1	2,256	2,950	126	301	5.59%	10.2%

Table 5: Number of candidate papers and detected HalluCited papers. We report the number of papers for each frequency of HalluCitation candidates. Hit Rate denotes the proportion of HalluCited papers within each frequency bin. "Num." indicates the raw count for each bin, and "Cum." represents the cumulative count aggregated from higher frequencies. For papers containing nine or more candidates, all HalluCited papers were successfully detected, and thus, we group them together.

tation candidates, nearly three out of four papers indeed include HalluCitation. This finding indicates that even simple title matching against reference databases can sufficiently detect HalluCitation. In contrast, cases with only a few candidates often include noise introduced by OCR or parsing errors, as well as limitations of simple character-level fuzzy matching. Although these citations may be erroneously flagged due to parsing failures or high thresholds in fuzzy matching, such mismatches rarely occur consecutively. Therefore, when three or four candidates are detected within a single paper, the paper is considered doubtful and recommended for further verification.

Finding 3: Manual detection is challenging because most HalluCited papers contain only a small number of HalluCitations. As shown in Table 5, when a paper contains only one or two HalluCitation candidates, the number of detected papers reaches nearly 200, accounting for around two-thirds of all HalluCited papers. This indicates that, rather than papers containing many HalluCitations, the majority of HalluCited papers contain only a small number of HalluCitations. This also suggests that such HalluCitations are often disguised among otherwise proper citations. At present, only reviewers or readers with strong expertise closely aligned with the cited work may notice these HalluCitations. However, when citations fall outside an area of expertise, manually detecting such hallucinations becomes nearly infeasible within the limited time available, e.g., during the review period. Moreover, we point out that some HalluCitations may not always be introduced intentionally. For instance, authors may obtain citation information from secondary sources such as Google Scholar[8], reference management tools, or LLM-based recommendation systems, where the entries themselves may already contain hallucinations. We discuss this possibility in detail in Section 4.3. These HalluCitations highlight that authors often implicitly trust such citations and do not always verify their factuality. Moreover, because such HalluCitations are often embedded among otherwise proper citations, it is preferable to introduce automated detection systems into the validation process.

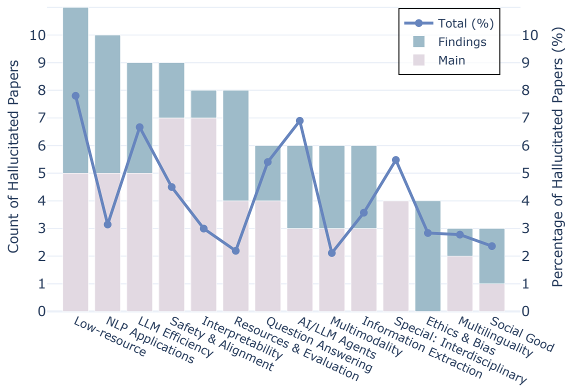


Figure 3: Number and proportion of HalluCited papers by area. We report areas with three or more papers. Area names are abbreviated using the first few words.

4 Further Analysis and Discussions

4.1 Trend Analysis

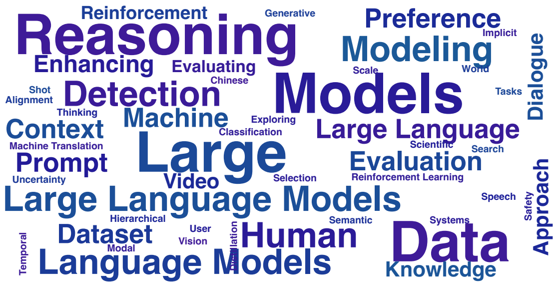
Which areas more frequently include HalluCited papers? To analyze trends in the occurrence of HalluCited papers, we examine the research areas in which such papers are accepted. We focus on EMNLP 2025[9] and analyze the Main and Findings papers by identifying the assigned research areas for each paper. Figure 3 shows the counts and proportions of HalluCited papers by area. For the official area names and the complete list of areas, please refer to the EMNLP 2025 Call for Papers[10]. We found that areas such as

[8] https://scholar.google.com/

[9] [Accessed on Dec. 31st, 2025] Area information is available at https://rrplanning2022-my.sharepoint.com/:x:/g/personal/jrachford_randrplanning_com/IQCJIl5zmNpUSZUv8j8H0z1TAVaIf1rOpMYhjCllLm703RY via https://2025.emnlp.org/program.

[10] https://2025.emnlp.org/calls/main_conference_papers/#submission-topics

(a) High TF-IDF differences in **HalluCited papers**.



(b) High TF-IDF differences in **general papers**.

Figure 4: Word clouds based on TF-IDF differences between HalluCited papers and general papers.

Low-Resource NLP, LLM Efficiency, and AI/LLM Agents exhibit relatively high counts and proportions of HalluCited papers. In contrast, areas including NLP Applications, Interpretability, and Resources and Evaluation show relatively large absolute counts but lower proportions. Notably, several areas with higher proportions of HalluCited papers, including Safety and Alignment in LLMs, AI/LLM Agents, and LLM Efficiency, were newly introduced at EMNLP 2025[11]. This suggests that emerging topics may be difficult to secure qualified reviewers for, which may make rigorous review, such as detecting hallucinations, more challenging.

Which keywords are included in HalluCited papers? To investigate characteristic terms and topics of HalluCited papers, we compute TF-IDF scores (Manning et al., 2008) over paper titles for both general papers and HalluCited papers, focusing on the Main and Findings papers at EMNLP 2025. Next, we compare the two groups by examining terms with large differences in their TF-IDF values. Figure 4 shows word clouds (Mueller, 2023) visualizing terms with high TF-IDF differences for each group. From these results, we observe that HalluCited papers more frequently contain terms related to topics such as "Multimodal", "Decoding", and "Quantization". In contrast, general papers tend to include terms associated with reinforcement learning-related topics, such as "Human", "Reasoning", and "Preference". In addition, HalluCited papers are more likely to use abbreviations such as "LLM" in their titles, whereas general papers more often use the full form, such as "Large Language Model". This suggests that HalluCited papers tend to adopt concise or abbreviated terminology and focus on efficiency-related topics, which is consistent with the results of Figure 3.

	2023		2024*				2025		
Freq.	Oct.	Dec.	Feb.	Apr.	June	Aug.	Feb.	May	July
≧9	0	0	1	0	0	0	2	6	4
8	0	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	2	3	0
6	0	0	0	0	0	0	2	3	0
5	0	0	2	0	3	0	1	5	1
4	0	0	0	0	0	0	5	6	2
3	2	0	4	1	3	0	13	14	6
2	4	18	37	3	39	3	62	60	17
1	51	87	186	30	226	17	322	237	72
cand.	57	105	230	34	271	20	409	335	102
total	270	659	967	227	922	127	1,349	1,377	371
%	**21.1**	15.9	**23.8**	15.0	**29.4**	15.7	**30.3**	**24.3**	**27.5**
AVERAGE REVIEW LOAD									
meta	3.9	5.0	6.6	2.3	5.3	2.6	6.5	5.4	3.1
review	2.2	2.4	3.7	1.9	3.1	1.8	3.9	3.2	2.5
STATISTICS OF AVAILABLE PREPRINTS									
submit	1,275	2,604	4,835	881	5,813	488	8,350	7,916	1,451
%	21.2	25.3	20.0	25.8	15.9	26.0	16.2	17.4	25.6

Table 6: Number of candidate papers among opted-in preprints in ARR by frequency of HalluCitation candidates. "cand." denotes the number of papers containing at least one HalluCitation candidate, and "total" denotes the number of opted-in anonymous preprints. **Bold** values exceed the average percentage of accepted papers (16.53%) reported in Table 3. Note that Oct. and Dec. 2024 are excluded because they are not publicly available. Dec. 2023 corresponds to NAACL 2024, Feb. 2024 to ACL 2024, June 2024 to EMNLP 2024, Feb. 2025 to ACL 2025, and May 2025 to EMNLP 2025. We also report the average number of assignments per meta-reviewer and reviewer. Underlined values indicate cases exceeding the expected value. We report the total number of submissions and the preprint disclosure rate.

4.2 How about the Peer-Review Process?

Table 6 shows the number of papers containing HalluCitation candidates among opted-in preprints in the ACL Rolling Review (ARR)[12], together with reviewer assignment statistics and the disclosure rate among all submissions. Main and Findings papers are submitted to ARR, and submissions that select the opt-in option are publicly released as preprints.

[11] https://2025.emnlp.org/track-changes/

[12] https://aclrollingreview.org/

In addition, the ARR Statistics Dashboard[13] provides aggregate statistics related to ARR. We leveraged these publicly available resources for our analysis. From Table 6, we observe that for submission iterations linked to conferences since 2024, the proportion of papers containing HalluCitation candidates is high. Nevertheless, when compared with the proportion of accepted papers reported in Table 3, such papers are filtered out to some extent during the peer-review process. We also observe that papers containing HalluCitation candidates are frequently assigned to meta-reviewers, and in some iterations, such as February 2024, they may be assigned at a rate of at least one paper per reviewer. This highlights the social impact of HalluCited papers as a practical and immediate issue for us.

However, since the preprint disclosure rate is around 20%, it remains unclear whether such papers are more likely to opt in or opt out. Therefore, we assumed that these papers are uniformly distributed with respect to disclosure. To facilitate retrospective studies and improve transparency in the peer-review process, greater disclosure of review-related information would be important.

Finally, we note that this analysis does not assess the quality of HalluCited papers. It is possible that some low-quality submissions include HalluCitations, and that the other factors already contribute to screening. Nevertheless, given the observed increase in HalluCited papers among accepted papers at EMNLP 2025, filtering such cases is becoming increasingly challenging even with reviewer effort alone. This also suggests the need to implement automated flagging systems to assist reviewer workloads and maintain rigor and quality.

4.3 Unintentional HalluCitation: Databases Are Often Contaminated

HalluCitations may be introduced unintentionally by authors. One common source of such errors is secondary databases, e.g., Google Scholar and Semantic Scholar[14], which may themselves contain incorrect or incomplete entries. Figure 5 shows several cases illustrating this issue. In Case 1, the paper title is truncated, preventing correct matching with the arXiv database. In Case 2, we confirm that the referenced paper does not exist. There was no corresponding publication by the listed authors at ACL 2024, nor could we find any paper with a similar title, despite the entry being listed in Google Scholar. In Case 3, the citation is associated with an incorrect arXiv ID or is missing an author.

These cases demonstrate that retrieving citation information from secondary sources such as Google Scholar does not guarantee correctness. In particular, as shown in Case 2, entirely hallucinated papers may be registered in such databases, and this citation is in fact referenced by other papers (Zhou, 2025). In Case 3, the incorrect arXiv ID has propagated to hundreds of subsequent citations. These findings suggest that the presence of HalluCitation does not necessarily imply that an AI system generated the citation or that AI tools were used, as erroneous citation entries are frequently observed in these databases. It is plausible that authors man-

Case 1: Cited by Li et al. (2025d)

✗ **Semantic Scholar (Ni and Li, 2024a):**
> Xuanfan Ni and Piji Li. 2024a. A systematic evaluation of large language models for natural. ArXiv, abs/2405.10251. **(Corpus ID: 261341578)**

✓ **arXiv (Formal) (Ni and Li, 2024b):**
> Xuanfan Ni and Piji Li. 2024b. A systematic evaluation of large language models for natural language generation tasks. Preprint, arXiv:2405.10251.

Case 2: Cited by Lee et al. (2025)

✗ **Google Scholar (non-existent reference):**
> Wei Chen, Arjun Kumar, and Lin Yang. 2024. Distraction-based attack prompts: An effective jailbreaking method for llms. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). **(ID: 10238337808766132122)**

Case 3: Cited by Ji et al. (2025)

✗ **Google Scholar (ID: 3546576102853818527):**
> Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. arXiv preprint arXiv:1910.01708 , 7(1):2.
> ✓ Fujimoto et al. (2019)

✗ **Semantic Scholar (Corpus ID: 208283920):**
> Josh Achiam and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. URL: https://d4mucfpksywv.cloudfront.net/safexp-short.pdf. **(Expired Link)**
> ✗ Missing First Author, Alex Ray.

✓ **OpenAI (Official) (Ray et al., 2019):**
> Alex Ray, Josh Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. URL: https://cdn.openai.com/safexp-short.pdf

Figure 5: Examples of incorrect database entries from Google Scholar and Semantic Scholar. Case 1 shows an incorrect title, Case 2 refers to a non-existent paper, and Case 3 contains inaccurate citation information.

[13] https://stats.aclrollingreview.org/
[14] https://www.semanticscholar.org/

Venue	Main	Findings	S.I.D.	WS	Total
NUMBER OF PAPERS					
AACL 2025	235	147	38	203	623
EACL 2026	446	354	181	441	1,422
NUMBER OF HALLUCITED PAPERS					
AACL 2025	17	14	1	24	56
EACL 2026	19	18	12	35	84
PERCENTAGE OF HALLUCITED PAPERS					
AACL 2025	7.23%	9.52%	2.63%	11.82%	8.99%
EACL 2026	4.26%	5.08%	6.63%	7.94%	5.91%
Overall	5.29%	6.39%	5.94%	4.50%	6.85%

Table 7: Statistics for AACL 2025 and EACL 2026. The notation is the same as in Table 1. Note that the Industry Track was not held at AACL 2025.

ually retrieved citation information from search engines or automatically imported it via reference management tools, such as Paperpile[15], thereby inadvertently incorporating incorrect entries. Furthermore, it is important to address such hallucinations as early as possible. Recent AI-Scientist agents can retrieve citations via APIs from these databases, potentially amplifying the issue and increasing both its scale and complexity over time.

Nevertheless, these cases indicate that authors did not consult the original sources directly and instead relied on secondary sources (Church, 2017). To mitigate this issue, it is necessary to acknowledge that databases are not error-free and to encourage authors to obtain citation information directly from primary sources such as the ACL Anthology or arXiv. Alternatively, the use of citation normalization tools, such as Ribiber[16], may help reduce such errors and should be more widely adopted.

4.4 Afterward: AACL 2025 and EACL 2026

The main analysis of this paper focuses on NAACL, ACL, and EMNLP, as these conferences were held consecutively over a two-year period. Fortunately, within this time window, no other major ACL conferences were held, which simplifies the analysis and helps maintain comparability. After submission to the ARR January cycle, the proceedings of AACL 2025 and EACL 2026 became publicly available. Therefore, we additionally examined these conferences. Table 7 shows the results, and Appendix C provides the list of HalluCited papers.

Unfortunately, the observed trend of HalluCitations continues to increase. Overall, more than 5% of papers contain at least one HalluCitation, suggesting that approximately one in twenty accepted papers includes such issues. In AACL, the proportion is even higher, corresponding to roughly one in eleven papers. This trend aligns with our main discussion and with the results in Table 4, as well as the trends observed in under-review papers shown in Table 6. These results may suggest that the current review process is approaching its practical capacity limits, making it increasingly difficult.

5 Suggestions and Recommendations

Introducing automatic verification systems for author toolkits and review checks. As discussed in Section 4.3, HalluCitations may originate from errors in secondary sources. Therefore, the presence of HalluCitations cannot be directly attributed to the use of LLMs or AI-generated papers alone. However, regardless of their origin, such HalluCitations indicate that authors often fail to carefully verify reference information and do not always consult the original sources. Therefore, we argue that it is necessary to introduce dedicated HalluCitation detection tools, such as ours[17] or existing LLM-based agents[18], and to integrate them into existing toolkits such as ACL pubcheck[19]. By using these automatic verification systems on both the author and organizer sides, authors can receive early warnings before submission, and organizers can automatically scan submissions at scale. This would enable fairer verification, reduce reviewer workload, and improve overall citation reliability.

Definition of HalluCitation and ensuring traceability in the peer-review process. It is important to clearly define what should be considered a HalluCitation, as not all citation errors are equally severe. In this study, we define HalluCitations based on the existence of the cited title and the consistency of key identifiers such as IDs and page numbers. However, minor citation errors have long existed in academic writing (Wang and Barabási, 2021), and inaccuracies may also originate from secondary sources (Besançon et al., 2024), making their causes complex. When only a few HalluCitations are identified, it is often sufficient to request correction. Unlike journals, where editors may correct citation information, conference papers place this responsibility largely on authors. Therefore,

[15] https://paperpile.com/
[16] https://github.com/yuchenlin/rebiber
[17] Our detection tool, HALLUCITECHECKER: https://github.com/yusuke1997/HalluCiteChecker/
[18] https://gptzero.me/news/iclr-2026/
[19] https://github.com/acl-org/aclpubcheck

minor citation errors should not be overly penalized, and providing author toolkits in advance can help prevent many formal mistakes. A more serious issue arises when it is unclear whether such corrections are reflected in the camera-ready version, highlighting the need for traceability in the revision and peer-review process. Improving it would enable verification of corrections and support rapid analyses of emerging issues such as HalluCitations.

Rethinking the purpose of conferences and peer review. What does "peer" review truly mean? We consider reviewers to be hidden co-authors of a paper, whose role is not merely to complete a formal evaluation but to engage in constructive discussion that improves the work. From this perspective, review systems should prioritize sufficient time and manageable workloads that enable meaningful feedback, rather than mechanisms that incentivize rapid review completion under strict time constraints of only a few weeks. Currently, early release through preprints, such as arXiv, and blog posts has become common. Despite the lack of formal verification through peer review, such outputs are already widely cited (Church et al., 2025). At the same time, conference reviewing is often less rigorous than journal peer review. If conferences offer neither the immediacy of preprints nor the rigor of journals, their core purpose should be reconsidered. The growing review burden risks undermining conference credibility and raises concerns about long-term sustainability.

One possible direction is a community-wide transition toward a mega-journal-style model with asynchronous, rigorous peer review, combined with conference venues that focus on presenting accepted work. Such an evolution could provide a more sustainable review process with sufficient time for verification, while still offering meaningful opportunities for authors to present their research. Originally, the ARR system and the Findings track had the potential to fulfill this role. However, ARR cycles became tightly coupled with conferences, and Findings were eventually positioned as a "companion" (Cohn et al., 2020). Now, more than five years after the introduction of ARR and Findings, it may be time to reconsider their roles. One possible redefinition is to treat them as journal-equivalent venues, with conference presentations serving as a form of certification or presentation right for accepted work, similar to TMLR. While this would not fundamentally resolve the growth in submission volume, it would at least help preserve credibility.

No paper is flawless. We argue that the presence of HalluCitations should not be treated as grounds for immediate penalties. As discussed, such errors can arise unintentionally rather than deliberate misconduct by authors. Instead of punitive measures, we emphasize developing author toolkits and pre-submission checks to proactively prevent such issues. It is also worth noting that the HalluCited papers were accepted due to their substantive contributions. Accordingly, authors of HalluCited papers should not be penalized post hoc. Since the ACL Anthology allows erratum revisions, it is important to better encourage voluntary corrections[20]. Acceptance should not be treated as a binary reward or punishment, but rather as part of an ongoing process of improving and communicating research.

Moreover, it is important to reaffirm the collaborative nature of peer review. Conferences exist to share work, not to reward perfection under batch review and strict time constraints (Church, 2005). *No paper is flawless*, and a healthy research community should remain tolerant of minor errors while actively promoting responsible correction. The ACL community has long been sustained by self-improvement mechanisms such as Findings papers and ARR. The NLP community can now take the initiative toward greater sustainability, including new approaches such as AI-assisted and asynchronous review processes.

6 Conclusion

We investigated the presence of HalluCitations in accepted papers at ACL conferences by examining all papers published at ACL, NAACL, and EMNLP in 2024 and 2025. As a result, we identified over 300 HalluCited papers, more than half of which appeared at EMNLP 2025. Moreover, Section 4.4 reveals that over one in twenty accepted papers at AACL 2025 and EACL 2026 contain at least one HalluCitation, highlighting the seriousness of the issue and the need for community discussion.

Will the trend continue to increase at ACL 2026, or will its impact be reduced? How about 2027? Our goal is to reduce the occurrence of HalluCitations through increased awareness. We hope this paper helps bring attention to challenges currently faced by the community, encourages greater care in publication practices, and stimulates active and nuanced discussion within the broader community.

[20] https://aclanthology.org/info/corrections/

7 Limitations

Scope. This study focuses on six recent top-tier NLP conferences. Although the scope could be expanded, we limit our analysis for several reasons. First, the number of HalluCited papers is negligible in earlier years, e.g., only two cases at NAACL 2024, and the trend is monotonically increasing, making more retrospective analysis less informative. Second, our primary focus is on the sharp surge observed in 2025, which motivates us to concentrate on the most recent conferences. Other venues, such as ACL 2026 or the ARR March 2026 cycle, were not publicly available at the time of writing and therefore could not be analyzed, though our findings suggest that a non-trivial number of HalluCitations may also be present there. While extending the analysis to other AI/ML conferences is possible, our goal is to assess the impact within the NLP community specifically. Overall, despite this limited scope, we believe our analysis is sufficiently comprehensive for the intended purpose, and broader investigations, e.g., other conferences and journals, are left for future work.

Analysis. We primarily analyzed accepted papers. First, accepted papers are readily accessible, which makes large-scale analysis feasible. Second, we showed that even accepted papers contain HalluCitations, suggesting that the peer-review process may not always sufficiently verify citation correctness under time constraints and heavy reviewer workloads. From this perspective, analyzing accepted papers is both reasonable and informative. Nevertheless, we also examined publicly available ARR data and conducted a comprehensive investigation within the limits of available disclosures. However, as discussed in Section 4.2, the transparency of the peer-review process remains limited, and a more thorough analysis is currently infeasible due to restricted access to review materials. Moreover, rejected papers are generally not publicly available. Although there have been efforts toward data release[21] (Dycke et al., 2023, 2022), these resources are limited to cases where explicit agreement has been obtained. In addition, the traceability of whether reviewer-requested corrections are reflected in the camera-ready versions remains unclear. We believe academic papers are the outcome of collaborative work between authors and reviewers. Thus, papers also reflect reviewer feedback. From this perspective, review materials, including rejected papers, should be made publicly available (Church et al., 2025; Yang, 2025; Kim et al., 2025d). Such transparency would facilitate analyses of review quality and outcomes.

Methods. We focused on arXiv and the ACL Anthology as solid and well-established data sources. Although extending our analysis to broader domains such as AI, ML, CV, robotics, speech, and journals is an important direction for future work, these domains currently lack a unified and consistently maintained bibliographic infrastructure comparable to that of the ACL community. Since this work focuses on ACL conferences. Despite this restricted scope, we identify nearly 300 HalluCited papers, indicating that our approach is effective. For implementation, we employ MinerU for OCR, although alternative OCR tools (Wei et al., 2025a; Cui et al., 2025a) and additional engineering improvements are possible. These implementation choices are not our primary contribution. Instead, the key contribution of this work is to quantitatively and comprehensively demonstrate the prevalence and impact of HalluCitations. Finally, our methodology prioritizes precision, and the reported results should be interpreted as a lower bound. Additional HalluCitations may exist beyond those detected here. We hope that this study will serve as a baseline and motivate follow-up work that improves detection accuracy and coverage.

Human verification. When hyperlinks were unavailable, we retrieved references using information such as proceedings identifiers, publication year, and title information. We also investigated title change histories and performed web searches for exact matches whenever possible. Most searches were conducted mainly during the first and second weeks of December 2025. Some works may have been released on arXiv or other platforms after the search period (Mohammad, 2025), or may not have been indexed by web search engines (Zhu et al., 2025a). In such cases, traceability was limited at the time of analysis, and they may have been temporarily included in the list. We thank the authors for their clarification. These cases are described as a limitation and have been removed from the list. We also received several inquiries regarding individual cases. In most cases, after providing other HalluCitations, the classification was confirmed and did not require removal. Although we have carefully verified, we welcome further feedback.

[21] https://arr-data.aclweb.org/

Ethical Considerations

Information. In this manuscript, all information reflects the state of publicly available data as of January 4, 2026. Unless explicitly stated otherwise, links, metadata, and dataset contents correspond to what was available at that time. Updates, corrections, or removals after this date may change the information and are not reflected in this study.

Licenses. Our primary data sources are the ACL Anthology and OpenReview, both of which are distributed under the Creative Commons Attribution 4.0 International License. We confirm that all additional datasets, dumps, and metadata used in this study are publicly available and permissively licensed. The code and tools developed in this work are authored by us. Human verification was conducted solely by the authors of this paper. Therefore, no licensing or consent issues arise from human annotation in this study.

Discussion on the protection of personally identifiable information and potential harm. This study analyzes real published papers. Therefore, the author information of HalluCited papers is inherently observable. However, concealing the identities of HalluCited papers would significantly hinder reproducibility. Moreover, the HalluCitations discussed here can already be identified through public search engines and thus constitute existing public information. Moreover, our analysis is comprehensive and systematic and does not target or single out specific individuals. Accordingly, this study is conducted strictly as academic research. Importantly, as explicitly stated in Section 6, we do not attribute HalluCitations to author misconduct, nor do we argue that authors should be penalized. Rather, we highlight limitations of the current review system, particularly the lack of appropriate author toolkits. Accordingly, this analysis and its claims and purpose are not intended to be offensive or accusatory. For transparency and reproducibility, we provide a list of HalluCited papers in Section B. We do not introduce new annotations or subjective judgments. We only verify and list HalluCitations that already exist in the public record. As an additional safeguard, Table 8 in Section B does not directly display author names or paper titles of HalluCited papers. Instead, each HalluCited paper is referenced only via links to the corresponding reference entries, and we show one example HalluCitation per paper. This design balances transparency and reproducibility with the protection of authors. Therefore, our discussion throughout the paper remains neutral and forward-looking, focusing on constructive solutions rather than blame. We have carefully reviewed the ACL Ethics Policy[22] and confirm this research complies with its guidelines.

Tool usage. We used DeepL, ChatGPT, and Grammarly for translation and grammatical improvement. All original content was written by the authors, and the authors take full responsibility for the final manuscript. For experiments, we used an NVIDIA A6000 GPU for OCR processing. Other analyses were conducted using a MacBook Pro, spreadsheets, and Google Chrome for literature searches. Notably, all records related to HalluCitations were managed locally and were not transmitted to any external APIs or services. Accordingly, tool usage and data management were conducted in compliance with ethical considerations.

Acknowledgements

We thank the anonymous reviewers, area chair, and senior area chair for their valuable comments and suggestions. Their feedback has strengthened this work, encouraged its publication, inspired new research directions, and motivated us. During this investigation, we also encountered many excellent papers that provided new research inspiration. We would like to express our sincere appreciation to the authors of these works.

We are also deeply grateful to the authors who contacted us after reading the early preprint version of this paper, as well as to the readers who provided encouragement and thoughtful feedback. The fact that researchers around the world engaged with our work has reaffirmed our motivation and commitment to further research. Following the release of the preprint, we exchanged nearly one hundred emails, including constructive criticism, supportive messages, and media inquiries. This level of engagement has been an important milestone in our research journey and has further motivated us to continue contributing to the community.

References

2008–2025. Grobid. https://github.com/kermitt2/grobid.

Aastik, Meghana Topu, Chinmay Kulkarni, and Pragya Paramita Sahu. 2025. NormAL LoRA: What

[22] https://ethics.aclweb.org/

is the perfect size? In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19716–19731, Suzhou, China. Association for Computational Linguistics.

Kurt Abela, Marc Tanti, and Claudia Borg. 2026. Balancing fluency and adherence: Hybrid fallback term injection in low-resource terminology translation. In Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 78–86, Rabat, Morocco. Association for Computational Linguistics.

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In Findings of the Association for Computational Linguistics: ACL 2025, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Tewodros Achamaleh, Tolulope Olalekan Abiola, Lemlem Eyob Kawo, Mikiyas Mebraihtu, and Grigori Sidorov. 2025. CIC-NLP@DravidianLangTech 2025: Detecting AI-generated product reviews in Dravidian languages. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 502–507, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Priyobroto Acharya, Haranath Mondal, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2025. JU-NLP: Improving low-resource Indic translation system with efficient LoRA-based adaptation. In Proceedings of the Tenth Conference on Machine Translation, pages 1201–1209, Suzhou, China. Association for Computational Linguistics.

Muhammad Adilazuarda. 2024. Beyond Turing: A comparative analysis of approaches for detecting machine-generated text. In Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024), pages 1–12, Mexico City, Mexico. Association for Computational Linguistics.

Sabik Aftahee, Tofayel Ahmmed Babu, MD Musa Kalimullah Ratul, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET_NetworkSociety@DravidianLangTech 2025: A transformer-based approach for detecting AI-generated product reviews in low-resource Dravidian languages. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 522–528, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Mahmoud Ahmad and Habeebah Kakudi. 2025. Stance detection on Nigerian 2023 election tweets using BERT: A low-resource transformer-based approach. In Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 54–63, Suzhou, China. Association for Computational Linguistics.

Mohammad Sadegh Akhondzadeh, Aleksandar Bojchevski, Evangelos Eleftheriou, and Martino Dazzi. 2025. KurTail : Kurtosis-based LLM quantization. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17404–17419, Suzhou, China. Association for Computational Linguistics.

Maro Akpobi. 2025. Yankari: Monolingual Yoruba dataset. In Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 1–6, Vienna, Austria. Association for Computational Linguistics.

Diyam Akra, Mohammed Khalilia, and Mustafa Jarrar. 2025. Active learning for multidialectal Arabic POS tagging. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24960–24973, Suzhou, China. Association for Computational Linguistics.

Burak Aktaş, Mehmet Can Baytekin, Süha Kağan Köse, Ömer İlbilgi, Elif Özge Yılmaz, Cagri Toraman, and Bilge Kaan Görür. 2026. BIRDTurk: Adaptation of the BIRD text-to-SQL dataset to Turkish. In Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 155–171, Rabat, Morocco. Association for Computational Linguistics.

Abdullah Al Monsur, Nitesh Vamshi Bommisetty, and Gene Louis Kim. 2026. Event detection with a context-aware encoder and LoRA for improved performance on long-tailed classes. In Findings of the Association for Computational Linguistics: EACL 2026, pages 5985–6003, Rabat, Morocco. Association for Computational Linguistics.

Fawzia Zaal Alanazi, Asma Mohammed Alamri, Arwa Bin Saleh, and Abdullah I. Alharbi. 2026. A hybrid confidence-aware framework for Arabic toxicity detection in social media. In Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, pages 364–370, Rabat, Morocco. Association for Computational Linguistics.

Flor Alberts, Ivo Bruinier, Nathalie Palm, Justin Paetzelt, and Erik Varecha. 2025. FENJI at SemEval-2025 task 3: Retrieval-augmented generation and hallucination span detection. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1143–1151, Vienna, Austria. Association for Computational Linguistics.

Cristea Alexandru-Marian. 2026. alexandru412 at MWE-2026 AdMIRe 2.0: Advancing multimodal idiomaticity representation. In Proceedings of the 22nd Workshop on Multiword Expressions (MWE

2026), pages 139–143, Rabat, Marocco. Association for Computational Linguistics.

Ghada Alfattni. 2026. A corpus-based investigation of contemporary Arabic dialects using the SADA corpus. In Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, pages 276–286, Rabat, Morocco. Association for Computational Linguistics.

Najwa Abdullah AlGhamdi, Sadam Al-Azani, Kwabena Nuamah, and Alan Bundy. 2026. A knowledge graph based diagnostic framework for analyzing hallucinations in Arabic machine reading comprehension. In Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, pages 413–421, Rabat, Morocco. Association for Computational Linguistics.

Falwah Alhamed, Rebecca Bendayan, Julia Ive, and Lucia Specia. 2024. Monitoring depression severity and symptoms in user-generated content: An annotation scheme and guidelines. In Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 227–233, Bangkok, Thailand. Association for Computational Linguistics.

Morteza Alikhani, Mohammadtaha Bagherifard, Erfan Zinvandi, and Mehran Sarmadi. 2025. PerCoR: Evaluating commonsense reasoning in Persian via multiple-choice sentence completion. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2205–2224, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Tural Alizada and Haim Dubossarsky. 2026. Benchmarking hate speech detection in Azerbaijani with Turkish cross-lingual transfer and transformer models. In Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 103–112, Rabat, Morocco. Association for Computational Linguistics.

Afnan Mohammed Aloraini, Riza Batista-Navarro, Goran Nenadic, and Viktor Schlegel. 2026. The SlangTrack dataset: Supporting the detection of words used in slang senses. In The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26), pages 1–19, Rabat, Morocco. Association for Computational Linguistics.

Dana R Alsagheer, Abdulrahman Kamal, Mohammad Kamal, Cosmo Yang Wu, and Weidong Shi. 2025. The lawyer that never thinks: Consistency and fairness as keys to reliable AI. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9943–9954, Vienna, Austria. Association for Computational Linguistics.

Mahdi Alshaikh Saleh and Irfan Ahmad. 2026. Seeing words differently: Visual embeddings for robust English-Arabic machine translation. In Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, pages 66–74, Rabat, Morocco. Association for Computational Linguistics.

Jesus Alvarez C, Daua Karajeanes, Ashley Prado, John Ruttan, Ivory Yang, Sean O'brien, Vasu Sharma, and Kevin Zhu. 2025. Advancing uto-aztecan language technologies: A case study on the endangered Comanche language. In Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 27–37, Albuquerque, New Mexico. Association for Computational Linguistics.

Rucha Ambaliya, Mahika Dugar, and Pruthwik Mishra. 2025. Niyamika at BHASHA task 1: Word-level transliteration for English-Hindi mixed text in grammar correction using MT5. In Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 135–141, Mumbai, India. Association for Computational Linguistics.

Anson Antony and Annika Schoene. 2025. Retrieval-enhanced mental health assessment: Capturing self-state dynamics from social media using in-context learning. In Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 268–278, Albuquerque, New Mexico. Association for Computational Linguistics.

Nora Aranberri. 2026. Crowd-based evaluation of emotion intensity preservation in Spanish–Basque tweet machine translation. In The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026), pages 123–133, Rabat, Morocco. Association for Computational Linguistics.

Saurav K. Aryal and Kritika Pant. 2025. Howard University-AI4PC at SemEval-2025 task 9: Using open-weight BART-MNLI for zero shot classification of food recall documents. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1919–1923, Vienna, Austria. Association for Computational Linguistics.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. Preprint, arXiv:2411.14199.

Yash Kumar Atri, Ahmed Alaa, and Thomas Hartvigsen. 2025. Lifelong model editing with graph-based external memory. In Findings of the Association for

Computational Linguistics: ACL 2025, pages 13336–13352, Vienna, Austria. Association for Computational Linguistics.

Mohammad Anas Azeez, Rafiq Ali, Ebad Shabbir, Zohaib Hasan Siddiqui, Gautam Siddharth Kashyap, Jiechao Gao, and Usman Naseem. 2025. Truth, trust, and trouble: Medical AI on the edge. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1017–1025, Suzhou (China). Association for Computational Linguistics.

Massa Baali, Xiang Li, Hao Chen, Syed Abdul Hannan, Rita Singh, and Bhiksha Raj. 2025. CAARMA: Class augmentation with adversarial mixup regularization. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9732–9742, Suzhou, China. Association for Computational Linguistics.

Max Bachmann. 2025. rapidfuzz/rapidfuzz: Release 3.13.0.

Thomas Bailleux, Tanmoy Mukherjee, Pierre Marquis, and Zied Bouraoui. 2025. Connecting concept layers and rationales to enhance language model interpretability. In Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025), pages 409–429, Suzhou, China. Association for Computational Linguistics.

Sathiyakugan Balakrishnan and Uthayasanker Thayasivam. 2026. Quality-aware adversarial ensemble for singer identification in 1960s Tamil film music. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 150–159, Rabat, Morocco. Association for Computational Linguistics.

Sagarika Banerjee, Tangatar Madi, Advait Swaminathan, Jolie Nguyen, Shivank Garg, Kevin Zhu, and Vasu Sharma. 2026. MiSCHiEF: A benchmark in minimal-pairs of safety and culture for holistic evaluation of fine-grained image-caption alignment. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 384–406, Rabat, Morocco. Association for Computational Linguistics.

Mohammad Jahid Ibna Basher, Md Kowsher, Md Saiful Islam, Rabindra Nath Nandi, Nusrat Jahan Prottasha, Mehadi Hasan Menon, Tareq Al Muntasir, Shammur Absar Chowdhury, Firoj Alam, Niloofar Yousefi, and Ozlem Garibay. 2025. BnTTS: Few-shot speaker adaptation in low-resource setting. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 4971–4983, Albuquerque, New Mexico. Association for Computational Linguistics.

Debarchan Basu, Shashwat Bhardwaj, Vaibhav Sharma, Pooja Singh, and Sandeep Kumar. 2025. GARuD: Guided alignment of representations using distillation for ultra-low-resource languages. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1867–1880, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Samuel Bell, Eduardo Sánchez, David Dale, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-Jussà. 2025. Translate, then detect: Leveraging machine translation for cross-lingual toxicity classification. In Proceedings of the Tenth Conference on Machine Translation, pages 253–268, Suzhou, China. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2025. HEDS 3.0: The human evaluation data sheet version 3.0. In Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2), pages 60–81, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Lonni Besançon, Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2024. Sneaked references: Fabricated reference metadata distort citation counts. Journal of the Association for Information Science and Technology, 75(12):1368–1379.

Mabrouka Bessghaier, Md. Rafiul Biswas, Shimaa Ibrahim, and Wajdi Zaghouani. 2026. A multi-task learning framework for modeling engagement and topic-sensitive responses in Arabic women's discourse. In Findings of the Association for Computational Linguistics: EACL 2026, pages 4846–4854, Rabat, Morocco. Association for Computational Linguistics.

Niloufar Beyranvand, Hamidreza Dastmalchi, Aijun An, Heidar Davoudi, Winston Chan, and Ron DiCarlantonio. 2025. GEAR: A scalable and interpretable evaluation framework for RAG-based car assistant systems. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2792–2810, Suzhou (China). Association for Computational Linguistics.

Shubhang Bhatnagar, Renxiong Wang, Kapil Krishnakumar, Adel Ahmadyan, Zhaojiang Lin, Lambert Mathias, Xin Luna Dong, Babak Damavandi, Narendra Ahuja, and Seungwhan Moon. 2026. Video-Mind: Thinking in steps for long video understanding. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 406–416, Rabat, Morocco. Association for Computational Linguistics.

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. 2025. FESTA: Functionally equivalent sampling for trust assessment of multimodal LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12277–12295, Suzhou, China. Association for Computational Linguistics.

Nikhil Bhendawade, Irina Belousova, Qichen Fu, Henry Mason, Antonie Lin, Mohammad Rastegari, and Mahyar Najibi. 2025. Speculative streaming: Efficient and scalable speculative decoding with multi-stream attention. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19536–19559, Suzhou, China. Association for Computational Linguistics.

Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiyao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, and Minlie Huang. 2025. MAGI: Multi-agent guided interview for psychiatric assessment. In Findings of the Association for Computational Linguistics: ACL 2025, pages 24898–24921, Vienna, Austria. Association for Computational Linguistics.

Pedro Alejandro Dal Bianco, Jean Paul Nunes Reinhold, Facundo Manuel Quiroga, and Franco Ronchetti. 2025. Augmenting sign language translation datasets with large language models. In Proceedings of the Workshop on Sign Language Processing (WSLP), pages 20–26, IIT Bombay, Mumbai, India (Co-located with IJCNLP–AACL 2025). Association for Computational Linguistics.

Suhas Bn, Dominik O. Mattioli, Andrew M. Sherrill, Rosa I. Arriaga, Christopher Wiese, and Saeed Abdullah. 2025. How real are synthetic therapy conversations? evaluating fidelity in prolonged exposure dialogues. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20986–20995, Suzhou, China. Association for Computational Linguistics.

Anna Bogdanova and Ileana Bucur. 2026. PMI MWE scorer at PARSEME 2.0 subtask 1: identifying multiword expressions using pointwise mutual information and universal dependencies. In Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 165–169, Rabat, Marocco. Association for Computational Linguistics.

Ondrej Bohdal, Mete Ozay, Jijoong Moon, Kyenghun Lee, Hyeonmok Ko, and Umberto Michieli. 2025. Efficient compositional multi-tasking for on-device large language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28141–28165, Suzhou, China. Association for Computational Linguistics.

Abhilekh Borah, Shubhra Ghosh, Kedar Joshi, Aditya Kumar Guru, and Kripabandhu Ghosh. 2026. Don't judge a book by its cover: Testing LLMs' robustness under logical obfuscation. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2162–2180, Rabat, Morocco. Association for Computational Linguistics.

Abhilekh Borah, Chhavi Sharma, Danush Khanna, Utkarsh Bhatt, Gurpreet Singh, Hasnat Md Abdullah, Raghav Kaushik Ravi, Vinija Jain, Jyoti Patel, Shubham Singh, Vasu Sharma, Arpita Vats, Rahul Raja, Aman Chadha, and Amitava Das. 2025. Alignment quality index (AQI) : Beyond refusals: AQI as an intrinsic alignment diagnostic via latent geometry, cluster divergence, and layer wise pooled representations. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2888–2947, Suzhou, China. Association for Computational Linguistics.

Abdessalam Bouchekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al Khatib, and Mohammed Ghaly. 2025. Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. In Proceedings of The Third Arabic Natural Language Processing Conference, pages 246–257, Suzhou, China. Association for Computational Linguistics.

Hossam Boudraa, Benoit Favre, and Raquel Urena. 2025. Implicit hate target span detection in zero- and few-shot settings with selective sub-billion parameter models. In Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 228–240, Vienna, Austria. Association for Computational Linguistics.

Nicolas Bougie and Narimawa Watanabe. 2025. Generative reviewer agents: Scalable simulacra of peer review. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 98–116, Suzhou (China). Association for Computational Linguistics.

Hongyi Cai, Jie Li, Mohammad Mahdinur Rahman, and Wenzhen Dong. 2025. Low-confidence gold: Refining low-confidence samples for efficient instruction tuning. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8233–8240, Suzhou, China. Association for Computational Linguistics.

Crivoi Carla and Ana Sabina Uban. 2025. SciBERT meets contrastive learning: A solution for scientific hallucination detection. In Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 336–343, Vienna, Austria. Association for Computational Linguistics.

Ege Yiğit Çelik and Selma Tekir. 2025. CiteBART: Learning to generate citations for local citation recommendation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 1703–1719, Suzhou, China. Association for Computational Linguistics.

Abhijit Chakraborty, Chahana Dahal, and Vivek Gupta. 2025. Federated retrieval-augmented generation: A systematic mapping study. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7362–7374, Suzhou, China. Association for Computational Linguistics.

Ernie Chang, Yang Li, Patrick Huber, Vish Vogeti, David Kant, Yangyang Shi, and Vikas Chandra. 2025. AutoMixer: Checkpoint artifacts as automatic data mixers. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19942–19953, Vienna, Austria. Association for Computational Linguistics.

Yu Chao, Siyu Lin, Xiaorong Wang, Zhu Zhang, Zihan Zhou, Haoyu Wang, Shuo Wang, Jie Zhou, Zhiyuan Liu, and Maosong Sun. 2025. LLM×MapReduce-v3: Enabling interactive in-depth survey generation through a MCP-driven hierarchically modular agent system. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 688–695, Suzhou, China. Association for Computational Linguistics.

Stergios Chatzikyriakidis, Erofili Psaltaki, Dimitrios Papadakis, Erik Henriksson, and Veronika Laippala. 2026. Perplexity as a metric for dialectal distance: A computational study of Greek varieties. In Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects, pages 101–112, Rabat, Morocco. Association for Computational Linguistics.

Shraddha Chauhan and Abhinav Kumar. 2025. MNLP@DravidianLangTech 2025: Transformer-based multimodal framework for misogyny meme detection. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 248–253, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Divij Chawla, Ashita Bhutada, Duc Anh Do, Abhinav Raghunathan, Vinod Sp, Cathy Guo, Dar Win Liew, Prannaya Gupta, Rishabh Bhardwaj, Rajat Bhardwaj, and Soujanya Poria. 2025. Evaluating AI for finance: Is AI credible at assessing investment risk appetite? In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2828–2839, Suzhou (China). Association for Computational Linguistics.

Boqi Chen, Xudong Liu, and Jianing Qiu. 2026a. Mask what matters: Mitigating object hallucinations in multimodal large language models with object-aligned visual contrastive decoding. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 9–16, Rabat, Morocco. Association for Computational Linguistics.

Bowen Chen, Zhao Wang, and Shingo Takamatsu. 2025a. OMS: On-the-fly, multi-objective, self-reflective ad keyword generation via LLM agent. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 18583–18601, Suzhou, China. Association for Computational Linguistics.

Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Kirkpatrick. 2024. LogogramNLP: Comparing visual and textual representations of ancient logographic writing systems for NLP. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14238–14254, Bangkok, Thailand. Association for Computational Linguistics.

Diyang Chen. 2025. pingan-team at SemEval-2025 task 2: LoRA-augmented qwen2.5 with Wikidata-driven entity translation. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2065–2070, Vienna, Austria. Association for Computational Linguistics.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025b. Benchmarking large language models on answering and explaining challenging medical questions. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2026b. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5848–5877, Rabat, Morocco. Association for Computational Linguistics.

Meng Chen, Philip Arthur, Qianyu Feng, Cong Duy Vu Hoang, Yu-Heng Hong, Mahdi Kazemi Moghaddam, Omid Nezami, Duc Thien Nguyen, Gioacchino Tangari, Duy Vu, Thanh Vu, Mark Johnson, Krishnaram Kenthapadi, Don Dharmasiri, Long Duong, and Yuan-Fang Li. 2025c. Mastering the craft of data synthesis for CodeLLMs. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 12484–12500, Albuquerque, New Mexico. Association for Computational Linguistics.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. 2025d. Ai4research: A survey of artificial intelligence for scientific research. Preprint, arXiv:2507.01903.

Yilong Chen, Junyuan Shang, Zhenyu Zhang, Yanxi Xie, Jiawei Sheng, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. 2025e. Inner thinking transformer: Leveraging dynamic depth scaling to foster adaptive internal thinking. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28241–28259, Vienna, Austria. Association for Computational Linguistics.

Yuhang Chen, Zhen Tan, Ajay Kumar Jaiswal, Huaizhi Qu, Xinyu Zhao, Qi Lin, Yu Cheng, Andrew Kwong, Zhichao Cao, and Tianlong Chen. 2025f. Bit-flip error resilience in LLMs: A comprehensive analysis and defense framework. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10414–10424, Suzhou, China. Association for Computational Linguistics.

Fedor Chernogorskii, Sergei Averkiev, Liliya Kudraleeva, Zaven Martirosian, Maria Tikhonova, Valentin Malykh, and Alena Fenogenova. 2026. DRAGOn: Designing RAG on periodically updated corpus. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 622–638, Rabat, Morocco. Association for Computational Linguistics.

Seyyed Saeid Cheshmi, Hahnemann Ortiz, James Mooney, and Dongyeop Kang. 2026. Reasoning beyond literal: Cross-style multimodal reasoning for figurative language understanding. In Findings of the Association for Computational Linguistics: EACL 2026, pages 5942–5956, Rabat, Morocco. Association for Computational Linguistics.

Veer Chheda, Aaditya Ghaisas, Avantika Sankhe, and Narendra Shekokar. 2025a. Extract-explain-abstract: A rhetorical role-driven domain-specific summarisation framework for Indian legal documents. In Proceedings of the Natural Legal Language Processing Workshop 2025, pages 439–455, Suzhou, China. Association for Computational Linguistics.

Veer Chheda, Avantika Sankhe, and Aaditya Uday Ghaisas. 2025b. Iterative critique-driven simplification: Targeted enhancement of complex definitions with small language models. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1081–1096, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Moulik Choraria, Xinbo Wu, Akhil Bhimaraju, Nitesh Sekhar, Yue Wu, Xu Zhang, Prateek Singhal, and Lav R. Varshney. 2026. DeepInsert: Early layer bypass for efficient and performant multimodal understanding. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7059–7079, Rabat, Morocco. Association for Computational Linguistics.

Manan Roy Choudhury, Adithya Chandramouli, Mannan Anand, and Vivek Gupta. 2026. Better call CLAUSE: A discrepancy benchmark for auditing LLMs legal reasoning capabilities. In Findings of the Association for Computational Linguistics: EACL 2026, pages 5776–5818, Rabat, Morocco. Association for Computational Linguistics.

Rafi Hassan Chowdhury and Rahanuma Ryaan Ferdous. 2025. BOIGENRE: A large-scale Bangla dataset for genre classification from book summaries. In Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 249–258, Mumbai, India. Association for Computational Linguistics.

Kenneth Church. 2005. Last words: Reviewing the reviewers. Computational Linguistics, 31(4):575–578.

Kenneth Ward Church. 2017. Word2vec. Natural Language Engineering, 23(1):155–162.

Kenneth Ward Church, Raman Chandrasekar, John E. Ortega, and Ibrahim Said Ahmad. 2025. Is peer-reviewing worth the effort? In Proceedings of the 31st International Conference on Computational Linguistics, pages 3589–3599, Abu Dhabi, UAE. Association for Computational Linguistics.

Christian Clark, Byung-Doh Oh, and William Schuler. 2025. How well does first-token entropy approximate word entropy as a psycholinguistic predictor? In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 47–57, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. Transactions of the Association for Computational Linguistics, 10:73–91.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset. Preprint, arXiv:1905.00075.

Trevor Cohn, Yulan He, and Yang Liu, editors. 2020. Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online.

Gaye Colakoglu, Gürkan Solmaz, and Jonathan Fürst. 2025. Problem solved? information extraction design space for layout-rich documents using LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17908–17927, Suzhou, China. Association for Computational Linguistics.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025a. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. Preprint, arXiv:2510.14528.

Wendi Cui, Jiaxin Zhang, Zhuohang Li, Hao Sun, Damien Lopez, Kamalika Das, Bradley A. Malin, and

Sricharan Kumar. 2025b. Heuristic-based search algorithm in automatic instruction-focused prompt optimization: A survey. In Findings of the Association for Computational Linguistics: ACL 2025, pages 22093–22111, Vienna, Austria. Association for Computational Linguistics.

Xia Cui. 2025. xiacui at SemEval-2025 task 11: Addressing data imbalance in transformer-based multilabel emotion detection with weighted loss. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 247–256, Vienna, Austria. Association for Computational Linguistics.

Duy Minh Dao Sy, Trung Kiet Huynh, Nguyen Chi Tran, Nguyen Lam Phu Quy, Pham Phu Hoa, and Nguyen Dinh Ha Duong. 2026a. HCMUS_The fangs at AbjadStyleTransfer shared task: Learning to query style, contrastive representations for zero-shot Arabic authorship style transfer. In Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, pages 438–442, Rabat, Morocco. Association for Computational Linguistics.

Duy Minh Dao Sy, Nguyen Chi Tran, Trung Kiet Huynh, Nguyen Lam Phu Quy, Pham Phu Hoa, and Nguyen Dinh Ha Duong. 2026b. HCMUS_TheFangs at AbjadGenEval shared task: Weighted layer pooling with attention fusion for Arabic AI-generated text detection. In Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, pages 433–437, Rabat, Morocco. Association for Computational Linguistics.

Amitava Das, Yaswanth Narsupalli, Gurpreet Singh, Vinija Jain, Vasu Sharma, Suranjana Trivedy, Aman Chadha, and Amit Sheth. 2025a. YinYang-align: A new benchmark for competing objectives and introducing multi-objective preference based text-to-image alignment. In Findings of the Association for Computational Linguistics: ACL 2025, pages 23518–23598, Vienna, Austria. Association for Computational Linguistics.

Amitava Das, Suranjana Trivedy, Danush Khanna, Yaswanth Narsupalli, Basab Ghosh, Rajarshi Roy, Gurpreet Singh, Vinija Jain, Vasu Sharma, Aishwarya Naresh Reganti, and Aman Chadha. 2025b. DPO kernels: A semantically-aware, kernel-enhanced, and divergence-rich paradigm for direct preference optimization. In Findings of the Association for Computational Linguistics: ACL 2025, pages 22174–22270, Vienna, Austria. Association for Computational Linguistics.

Kunal Kingkar Das, Manoj Balaji Jagadeeshan, Nallani Chakravartula Sahith, Jivnesh Sandhan, and Pawan Goyal. 2025c. Still not there: Can LLMs outperform smaller task-specific Seq2Seq models on the poetry-to-prose conversion task? In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 343–357, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Sourya Dipta Das, Shubham Kumar, and Kuldeep Yadav. 2025d. Zero-shot grammar competency estimation using large language model generated pseudo labels. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3167–3179, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Jonathan De Bruin. 2023. PyAlex.

Rupesh Dhakad, Naveen Kumar, and Shrikant Malviya. 2025. Team-SVNIT at JUST-NLP 2025: Domain-adaptive fine-tuning of multilingual models for English–Hindi legal machine translation. In Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 148–154, Mumbai, India. Association for Computational Linguistics.

Manav Dhamecha, Gaurav Damor, Sunil Jaat, and Pruthwik Mishra. 2025a. Team horizon at BHASHA task 2: Fine-tuning multilingual transformers for Indic word grouping. In Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 175–179, Mumbai, India. Association for Computational Linguistics.

Manav Dhamecha, Sunil Jaat, Gaurav Damor, and Pruthwik Mishra. 2025b. Team horizon at BHASHA task 1: Multilingual IndicGEC with transformer-based grammatical error correction models. In Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 142–146, Mumbai, India. Association for Computational Linguistics.

Varun Dhanraj and Chris Eliasmith. 2025. Improving rule-based reasoning in LLMs using neurosymbolic representations. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30589–30608, Suzhou, China. Association for Computational Linguistics.

Yiwen Ding, Jiarui Liu, Zhiheng Lyu, Kun Zhang, Bernhard Schölkopf, Zhijing Jin, and Rada Mihalcea. 2025. Voices of her: Analyzing gender differences in the AI publication world. In Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 196–214, Vienna, Austria. Association for Computational Linguistics.

Vrund Dobariya, Jatayu Baxi, Bhavika Gambhava, and Brijesh Bhatt. 2025. Smruti: Grammatical error correction for Gujarati using LLMs with non-parametric memory. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 473–485, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Bonaventure F. P. Dossou and Henri Aïdasso. 2025. Towards open-ended discovery for low-resource NLP. In Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 287–297, Suzhou, China. Association for Computational Linguistics.

Razvan-Gabriel Dumitru, Minglai Yang, Vikas Yadav, and Mihai Surdeanu. 2025. CopySpec: Accelerating LLMs with speculative copy-and-paste. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26301–26332, Suzhou, China. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Yes-yes-yes: Proactive data collection for ACL rolling review and beyond. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 300–318, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Nellia Dzhubaeva, Katharina Trinley, and Laura Pissani. 2025. Unstructured minds, predictable machines: A comparative study of narrative cohesion in human and LLM stream-of-consciousness writing. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1079–1096, Vienna, Austria. Association for Computational Linguistics.

Sana Ebrahimi, Mohsen Dehghankar, and Abolfazl Asudeh. 2025. An adversary-resistant multi-agent LLM system via credibility scoring. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1676–1693, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Roxanne El Baff, Dominik Opitz, and Diaoulé Diallo. 2025. CriticalBrew at CQs-gen 2025: Collaborative multi-agent generation and evaluation of critical questions for arguments. In Proceedings of the 12th Argument mining Workshop, pages 314–321, Vienna, Austria. Association for Computational Linguistics.

Mohamed Elgaar and Hadi Amiri. 2025. Linguistically-controlled paraphrase generation. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20842–20864, Suzhou, China. Association for Computational Linguistics.

Mohamed Elgaar and Hadi Amiri. 2026. LingGen: Scalable multi-attribute linguistic control via power-law masking. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1925–1942, Rabat, Morocco. Association for Computational Linguistics.

Auda Elshokry and Mohammed Alhanjouri. 2026. Optimizer choice and calibration for QARiB on Arabic-script social media offensive language detection. In Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, pages 26–36, Rabat, Morocco. Association for Computational Linguistics.

Vignesh Ethiraj, Ashwath D, Sidhanth Menon, Divya Vijay, and Vidhyakshaya Kannan. 2025. T-VEC: A telecom-specific vectorization model with enhanced semantic understanding via deep triplet loss fine-tuning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2449–2460, Suzhou (China). Association for Computational Linguistics.

Egecan Evgin, Ilknur Karadeniz, and Olcay Taner Yıldız. 2025. MetninOzU at BioLaySumm2025: Text summarization with reverse data augmentation and injecting salient sentences. In Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks), pages 179–184, Vienna, Austria. Association for Computational Linguistics.

Neo Eyal, Nachum Dershowitz, and Kfir Bar. 2025. Layer duplication in LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17797–17807, Suzhou, China. Association for Computational Linguistics.

Daniel Fadlon and Kfir Bar. 2026. How much pretraining does structured data need? In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3352–3365, Rabat, Morocco. Association for Computational Linguistics.

Daniel Fadlon, David Dov, Aviya Bennett, Daphna Heller-Miron, Gad Levy, Kfir Bar, and Ahuva Weiss-Meilik. 2025. Decode like a clinician: Enhancing LLM fine-tuning with temporal structured data representation. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1906–1922, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Ahmad Fahmid, Fahim Foysal, Wasif Haider, Shafin Rahman, and Md Adnan Arefeen. 2025.

NSU_PiedPiper at BLP-2025 task 2: A chain-of-thought with iterative debugging approach for code generation with Bangla instruction. In Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 595–600, Mumbai, India. Association for Computational Linguistics.

Yijia Fan, Jusheng Zhang, Kaitong Cai, Jing Yang, and Keze Wang. 2025. CCG: Rare-label prediction via neural SEM–driven causal game. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 6243–6256, Suzhou, China. Association for Computational Linguistics.

Maryam Fatima. 2025. FIRMA: Bidirectional formal-informal mathematical language alignment with proof-theoretic grounding. In Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025), pages 62–76, Suzhou, China. Association for Computational Linguistics.

Arvid Frydenlund. 2025. Language models, graph searching, and supervision adulteration: When more supervision is less and how to make more more. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29011–29059, Vienna, Austria. Association for Computational Linguistics.

Zichuan Fu, Xian Wu, Guojing Li, Yingying Zhang, Yefeng Zheng, Tianshi Ming, Yejing Wang, Wanyu Wang, and Xiangyu Zhao. 2025. Model merging for knowledge editing. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 433–443, Vienna, Austria. Association for Computational Linguistics.

Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. 2019. Benchmarking batch deep reinforcement learning algorithms. Preprint, arXiv:1910.01708.

Achille Fusco, Maria Letizia Piccini Bianchessi, Tommaso Sgrizzi, Asya Zanollo, and Cristiano Chesi. 2025. Linguistic units as tokens: Intrinsic and extrinsic evaluation with BabyLM. In Proceedings of the First BabyLM Workshop, pages 496–507, Suzhou, China. Association for Computational Linguistics.

Jack Gallifant, Shan Chen, Kuleen Sasse, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. 2025. Sparse autoencoder features for classifications and transferability. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29939–29963, Suzhou, China. Association for Computational Linguistics.

Fan Gao, Cheng Huang, Yutong Liu, Nyima Tashi, Xiangxiang Wang, Thupten Tsering, Ban Ma-bao, Renzeng Duojie, Gadeng Luosang, Rinchen Dongrub, Dorje Tashi, Xiao Feng Cd, Yongbin Yu, and Hao Wang. 2025. TLUE: A Tibetan language understanding evaluation benchmark. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35071–35097, Suzhou, China. Association for Computational Linguistics.

Rahul Garg, Trilok Padhi, Hemang Jain, Ugur Kursuncu, and Ponnurangam Kumaraguru. 2025. Just KIDDIN' : Knowledge infusion and distillation for detection of INdecent memes. In Findings of the Association for Computational Linguistics: ACL 2025, pages 23067–23086, Vienna, Austria. Association for Computational Linguistics.

Carol-Luca Gasan and Vasile Pais. 2025. SAG: Enhancing domain-specific information retrieval with semantic-augmented graphs. In Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025), pages 362–371, Suzhou, China. Association for Computational Linguistics.

Arpit Singh Gautam, Kailash Talreja, and Saurabh Jha. 2026. The energy of falsehood: Detecting hallucinations via diffusion model likelihoods. In Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER), pages 47–58, Rabat, Morocco. Association for Computational Linguistics.

Haonan Ge, Yiwei Wang, Ming-Hsuan Yang, and Yujun Cai. 2025. MRFD: Multi-region fusion decoding with self-consistency for mitigating hallucinations in LVLMs. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15860–15879, Suzhou, China. Association for Computational Linguistics.

Ryan George, Akshay Govind Srinivasan, Jayden Koshy Joe, Harshith M R, Vijayavallabh J, Hrushikesh Kant, Rahul Vimalkanth, Sachin S, and Sudharshan Suresh. 2025. Enhancing financial RAG with agentic AI and multi-HyDE: A novel approach to knowledge retrieval and hallucination reduction. In Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing, pages 19–32, Suzhou, China. Association for Computational Linguistics.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8920–8936, Suzhou, China. Association for Computational Linguistics.

Lynn Greschner and Roman Klinger. 2025. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and LLMs. In Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 628–646, Albuquerque, USA. Association for Computational Linguistics.

Kaiwen Guo and Yimeng Wu. 2025. MAJI: A multi-agent workflow for augmenting journalistic interviews. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the

Association for Computational Linguistics, pages 1061–1083, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Minghao Guo, Qingcheng Zeng, Xujiang Zhao, Yanchi Liu, Wenchao Yu, Mengnan Du, Haifeng Chen, and Wei Cheng. 2026. DeepSieve: Information sieving via LLM-as-a-knowledge-router. In Findings of the Association for Computational Linguistics: EACL 2026, pages 3054–3077, Rabat, Morocco. Association for Computational Linguistics.

Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Dinesh Khandelwal, Dinesh Raghu, and Sachindra Joshi. 2025a. Selective self-to-supervised fine-tuning for generalization in large language models. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 6255–6264, Albuquerque, New Mexico. Association for Computational Linguistics.

Taneesh Gupta, Shivam Shandilya, Xuchao Zhang, Rahul Madhavan, Supriyo Ghosh, Chetan Bansal, Huaxiu Yao, and Saravan Rajmohan. 2025b. CARMO: Dynamic criteria generation for context aware reward modelling. In Findings of the Association for Computational Linguistics: ACL 2025, pages 2202–2261, Vienna, Austria. Association for Computational Linguistics.

Namrata Bhalchandra Patil Gurav, Akashdeep Ranu, Archchana Sindhujan, and Diptesh Kanojia. 2026. Domain-specific quality estimation for machine translation in low-resource scenarios. In Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 630–650, Rabat, Morocco. Association for Computational Linguistics.

Akio Hayakawa, Nouran Khallaf, Horacio Saggion, and Serge Sharoff. 2025. UoL-UPF at TSAR 2025 shared task a generate-and-select approach for readability-controlled text simplification. In Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 193–210, Suzhou, China. Association for Computational Linguistics.

Chenyuan He, Yuxiang Jia, Fei Gao, Senbin Zhu, Hongde Liu, Hongying Zan, and Min Peng. 2025a. Task-aware contrastive mixture of experts for quadruple extraction in conversations with code-like replies and non-opinion detection. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4144–4159, Suzhou, China. Association for Computational Linguistics.

Zoe Wanying He, Sean Trott, and Meenakshi Khosla. 2025b. Seeing through words, speaking through pixels: Deep representational alignment between vision and language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35657–35672, Suzhou, China. Association for Computational Linguistics.

Arshia Hemmat, Matin Moqadas, Ali Mamanpoosh, Amirmasoud Rismanchian, and Afsaneh Fatemi. 2025. VAGUE-Gate: Plug-and-Play Local-Privacy shield for Retrieval-Augmented generation. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3715–3730, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Rushikesh Hiray and Venelin Kovatchev. 2025. Storybranch - generating multimedia content from novels. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 485–493, Albuquerque, New Mexico. Association for Computational Linguistics.

Sonja Hohloch. 2024. Homoclinic floer homology via direct limits. Preprint, arXiv:2402.12345.

Faye Holt, William Held, and Diyi Yang. 2024. Perceptions of language technology failures from South Asian English speakers. In Findings of the Association for Computational Linguistics: ACL 2024, pages 4067–4081, Bangkok, Thailand. Association for Computational Linguistics.

Eftekhar Hossain, Sanjeev Kumar Sinha, Naman Bansal, R. Alexander Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Ram Pavan Kumar Guttikonda, Mousumi Akter, Md. Mahadi Hassan, Matthew Freestone, Matthew C. Williams Jr., Dongji Feng, and Santu Karmaker. 2025. LLMs as meta-reviewers' assistants: A case study. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7763–7803, Albuquerque, New Mexico. Association for Computational Linguistics.

Faezeh Hosseini, Mohammadali Yousefzadeh, and Yadollah Yaghoobzadeh. 2026. FFE-hallu: Hallucinations in fixed figurative expressions: A benchmark of idioms and proverbs in the Persian language. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5222–5235, Rabat, Morocco. Association for Computational Linguistics.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Large language models are cross-lingual knowledge-free reasoners. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1525–1542, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, Kun Shao, and Jun Wang. 2025. Deep research agents: A systematic examination and roadmap. Preprint, arXiv:2506.18096.

Ziyi Huang and Xia Cui. 2025. PromotionGo at SemEval-2025 task 11: A feature-centric framework for cross-lingual multi-emotion detection in short texts. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2140–2148, Vienna, Austria. Association for Computational Linguistics.

Angel Hsing-Chi Hwang, Q. Vera Liao, Su Lin Blodgett, Alexandra Olteanu, and Adam Trischler. 2025. 'it was 80% me, 20% ai': Seeking authenticity in co-writing with large language models. Proc. ACM Hum.-Comput. Interact., 9(2).

Jahidul Islam, Md Ataullha, and Saiful Azad. 2025. PyBangla at BLP-2025 task 2: Enhancing Bangla-to-python code generation with iterative self-correction and multilingual agents. In Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 566–570, Mumbai, India. Association for Computational Linguistics.

Ahmad Ibrahim Ismail, Bashirudeen Opeyemi Ibrahim, Olubayo Adekanmbi, and Ife Adebara. 2025. Retrieval-augmented generation meets local languages for improved drug information access and comprehension. In Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 108–114, Vienna, Austria. Association for Computational Linguistics.

Akriti Jain and Aparna Garimella. 2026. Knowing what's missing: Assessing information sufficiency in question answering. In Findings of the Association for Computational Linguistics: EACL 2026, pages 4163–4174, Rabat, Morocco. Association for Computational Linguistics.

Raunak Jain and Srinivasan Rengarajan. 2025. Emergent wisdom at BEA 2025 shared task: From lexical understanding to reflective reasoning for pedagogical ability assessment. In Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), pages 1108–1120, Vienna, Austria. Association for Computational Linguistics.

Charlott Jakob, David Harbecke, Patrick Parschan, Pia Wenzel Neves, and Vera Schmitt. 2025. PolBiX: Detecting LLMs' political bias in fact-checking through X-phemisms. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17192–17210, Suzhou, China. Association for Computational Linguistics.

Gunjan Jalori, Preetika Verma, and Sercan O Arik. 2025. FLAIRR-TS - forecasting LLM-agents with iterative refinement and retrieval for time series. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15427–15437, Suzhou, China. Association for Computational Linguistics.

Surgan Jandial, Yinheng Li, Justin Wagle, and Kazuhito Koishida. 2026. Do GUI grounders truly understand UI elements? In Findings of the Association for Computational Linguistics: EACL 2026, pages 2772–2785, Rabat, Morocco. Association for Computational Linguistics.

Dongsuk Jang, Ziyao Shangguan, Kyle Tegtmeyer, Anurag Gupta, Jan T Czerminski, Sophie Chheang, and Arman Cohan. 2025. MedTutor: A retrieval-augmented LLM system for case-based medical education. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 319–353, Suzhou, China. Association for Computational Linguistics.

Ala Jararweh, Oladimeji Macaulay, David Arredondo, Yue Hu, Luis E Tafoya, Kushal Virupakshappa, and Avinash Sahu. 2025. Protein2Text: Resampling mechanism to translate protein sequences into human-interpretable text. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track), pages 918–937, Albuquerque, New Mexico. Association for Computational Linguistics.

Sungho Jeon and Michael Strube. 2025. Entity tracking in small language models: An attention-based study of parameter-efficient fine-tuning. In Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 42–53, Suzhou, China. Association for Computational Linguistics.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. 2025. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31983–32016, Vienna, Austria. Association for Computational Linguistics.

Jiazhou Ji and Xinru Lu. 2025. ReFLAIR: Enhancing multimodal reasoning via structured reflection and reward-guided learning. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25401–25413, Suzhou, China. Association for Computational Linguistics.

Feiran Jia, Tong Wu, Xin Qin, and Anna Squicciarini. 2025a. The task shield: Enforcing task alignment to defend against indirect prompt injection in LLM agents. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29680–29697, Vienna, Austria. Association for Computational Linguistics.

Runsong Jia, Mengjia Wu, Ying Ding, Jie Lu, and Yi Zhang. 2025b. HetGCoT: Heterogeneous graph-enhanced chain-of-thought LLM reasoning for academic question answering. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15950–15963, Suzhou, China. Association for Computational Linguistics.

Jin Jiang, Jianing Wang, Yuchen Yan, Yang Liu, Jianhua Zhu, Mengdi Zhang, and Liangcai Gao. 2025. Do large language models excel in complex logical reasoning with formal language? In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16878–16903, Suzhou, China. Association for Computational Linguistics.

Lesheng Jin, Zhenyuan Ruan, Haohui Mai, and Jingbo Shang. 2025. VeriLocc: End-to-end cross-architecture register allocation via LLM. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30252–30262, Suzhou, China. Association for Computational Linguistics.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring peer review dynamics with LLM agents. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.

Aniket Joshi, Cyrus Andre DSouza, Sejal Jain, Jitenkumar Babubhai Rana, and Promod Yenigalla. 2025a. I-SEE: An instruction-tuned, SOP-enhanced quality evaluator for product content. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1379–1388, Suzhou (China). Association for Computational Linguistics.

Nisheeth Joshi, Palak Arora, Anju Krishnia, Riya Lonchenpa, and Mhasilenuo Vizo. 2025b. BVSLP: Machine translation using linguistic embellishments for IndicMT shared task 2025. In Proceedings of the Tenth Conference on Machine Translation, pages 1265–1270, Suzhou, China. Association for Computational Linguistics.

Srihari V K, Vijay Karthick Vaidyanathan, and Thenmozhi Durairaj. 2025. NLP_goats@DravidianLangTech 2025: Detecting fake news in Dravidian languages: A text classification approach. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 345–349, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Mukund K. Roy, Karunesh Arora, Praveen Kumar Chandaliya, Rohit Kumar, and Pruthwik Mishra. 2025. CDAC-SVNIT submission for IWSLT 2025 Indic track shared task. In Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 180–185, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Maithili Sanjay Kadam and Francis Ferraro. 2025. TAG–EQA: Text–and–graph for event question answering via structured prompting strategies. In Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025), pages 304–315, Suzhou, China. Association for Computational Linguistics.

Xiaoqiang Kang, Shengen Wu, Zimu Wang, Yilin Liu, Xiaobo Jin, Kaizhu Huang, Wei Wang, Yutao Yue, Xiaowei Huang, and Qiufeng Wang. 2025. Can GRPO boost complex multimodal table understanding? In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12631–12644, Suzhou, China. Association for Computational Linguistics.

Yunus Karatepe, Mert Sülük, Zeynep Tuğçe Kırımlı, and Begüm Özbay. 2026. Sahara tokenizers at PARSEME 2.0 subtask 1: Combining contextual embeddings with structural decoding for multi-word expression detection. In Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 154–159, Rabat, Marocco. Association for Computational Linguistics.

Aly M. Kassem, Zhuan Shi, Negar Rostamzadeh, and Golnoosh Farnadi. 2025. REVIVING YOUR MNEME: Predicting the side effects of LLM unlearning and fine-tuning via sparse model diffing. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32250–32263, Suzhou, China. Association for Computational Linguistics.

Md Siddikul Imam Kawser, Jidan Al Abrar, Mehebub Bin Kabir, Md. Rayhan Chowdhury, and Md Ataullah Bahari. 2025. A hybrid transformer–sequential model for depression detection in Bangla–English code-mixed text. In Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 107–112, Mumbai, India. Association for Computational Linguistics.

Samreen Kazi and Shakeel Ahmed Khoja. 2026. QARI: Neural architecture for Urdu extractive machine reading comprehension. In Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 168–177, Rabat, Morocco. Association for Computational Linguistics.

Mamadou K. Keita, Adwoa Bremang, Huy Le, Dennis Owusu, Marcos Zampieri, and Christopher Homan. 2026. Grammatical error correction for low-resource languages: The case of Zarma. In Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 98–109, Rabat, Morocco. Association for Computational Linguistics.

Olga Kellert, Nemika Tyagi, Muhammad Imran, Nelvin Licona-Guevara, and Carlos Gómez-Rodríguez.

2025. Parsing the switch: LLM-based UD annotation for complex code-switched and low-resource languages. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15934–15949, Suzhou, China. Association for Computational Linguistics.

Ashinee Kesanam, Gummuluri Venkata Ravi Ram, Chaithanya Swaroop Banoth, and G Rama Mohana Reddy. 2025. NITK-VITAL at SemEval-2025 task 11: Focal-RoBERTa: Addressing class imbalance in multi-label emotion classification. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1077–1081, Vienna, Austria. Association for Computational Linguistics.

Elham Khabiri, Jeffrey O. Kephart, Fenno F. Heath Iii, Srideepika Jayaraman, Yingjie Li, Fateh A. Tipu, Dhruv Shah, Achille Fokoue, and Anu Bhamidipaty. 2025. Declarative techniques for NL queries over heterogeneous data. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1744–1761, Suzhou (China). Association for Computational Linguistics.

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. 2025. CoCoA: Confidence- and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6835–6855, Suzhou, China. Association for Computational Linguistics.

Maira Khatoon, Arooj Kiyani, Tehmina Farid, and Sadaf Abdul Rauf. 2025. FJWU_Squad at SemEval-2025 task 1: An idiom visual understanding dataset for idiom learning. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1759–1765, Vienna, Austria. Association for Computational Linguistics.

Adity Khisa, Nusrat Jahan Lia, Tasnim Mahfuz Nafis, Zarif Masud, Tanzir Pial, Shebuti Rayana, and Ahmedul Kabir. 2025. Exploring cross-lingual knowledge transfer via transliteration-based MLM fine-tuning for critically low-resource Chakma language. In Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 280–291, Mumbai, India. Association for Computational Linguistics.

Md Tawkat Islam Khondaker, Numaan Naeem, Fatimah Khan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Benchmarking LLaMA-3 on Arabic language generation tasks. In Proceedings of the Second Arabic Natural Language Processing Conference, pages 283–297, Bangkok, Thailand. Association for Computational Linguistics.

Ruslan Khrulev. 2025. CHECK-MAT: Probing the mathematical reasoning and rubric-alignment of vision-language models on handwritten solutions. In Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025), pages 77–94, Suzhou, China. Association for Computational Linguistics.

Ahrii Kim. 2025. Context is ubiquitous, but rarely changes judgments: Revisiting document-level MT evaluation. In Proceedings of the Tenth Conference on Machine Translation, pages 81–97, Suzhou, China. Association for Computational Linguistics.

Dongjun Kim, Gyuho Shim, Yongchan Chun, Minhyuk Kim, Chanjun Park, and Heuiseok Lim. 2025a. Benchmark profiling: Mechanistic diagnosis of LLM benchmarks. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15635–15650, Suzhou, China. Association for Computational Linguistics.

Heehyeon Kim, Kyeongryul Lee, and Joyce Jiyoung Whang. 2025b. Beneath the facade: Probing safety vulnerabilities in LLMs via auto-generated jailbreak prompts. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17668–17700, Suzhou, China. Association for Computational Linguistics.

Hyeongsik Kim, Yanheng Xu, Chaoqun Dong, and Fei Du. 2025c. Over-generation and compaction: A prompting strategy for procedural text adaptation with large language models. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19306–19337, Suzhou, China. Association for Computational Linguistics.

Jaeho Kim, Yunseok Lee, and Seulki Lee. 2025d. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pages 81634–81651. PMLR.

Jongwoo Kim, SeongYeub Chu, Bryan Wong, and Mun Yong Yi. 2025e. Not all options are created equal: Textual option weighting for token-efficient LLM-based knowledge tracing. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16114–16128, Suzhou, China. Association for Computational Linguistics.

Juhyeong Kim, Sangyeon Yu, Gyunyeop Kim, and Sangwoo Kang. 2025f. FractalLLM: Lossless self-speculative decoding with layer embedded self-compression. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23666–23673, Suzhou, China. Association for Computational Linguistics.

Hiroaki Kingetsu, Kaoru Yokoo, Kenji Fukumizu, and Manohar Kaul. 2026. Look before you leap: A lookahead reasoning quality gate for speculative decoding. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7831–7847, Rabat, Morocco. Association for Computational Linguistics.

Đorđe Klisura, Astrid R Bernaga Torres, Anna Karen Gárate-Escamilla, Rajesh Roshan Biswal, Ke Yang, Hilal Pataci, and Anthony Rios. 2025. A multi-agent framework for mitigating dialect biases in privacy policy question-answering systems. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32318–32337, Vienna, Austria. Association for Computational Linguistics.

Shaghayegh Kolli, Richard Rosenbaum, Timo Cavelius, Lasse Strothe, Andrii Lata, and Jana Diesner. 2025. Hybrid fact-checking that integrates knowledge graphs, large language models, and search-based retrieval agents improves interpretable claim verification. In Proceedings of the 9th Widening NLP Workshop, pages 106–115, Suzhou, China. Association for Computational Linguistics.

Kubilay Kağan Kömürcü and Tugce Temel. 2026. 3K2T at MWE-2026 AdMIRe 2: CARIM– category-aware reasoning for idiomatic multimodality. In Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 160–164, Rabat, Marocco. Association for Computational Linguistics.

Olga Kononykhina, Anna-Carolina Haensch, and Frauke Kreuter. 2025. Mind the gap: Gender-based differences in occupational embeddings. In Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 83–91, Vienna, Austria. Association for Computational Linguistics.

Yeganeh Kordi, Nihal V. Nayak, Max Zuo, Ilana Nguyen, and Stephen Bach. 2026. Revisiting generalization across difficulty levels: It's not so easy. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7014–7042, Rabat, Morocco. Association for Computational Linguistics.

Maria Korobeynikova, Alessia Battisti, Lukas Fischer, and Yingqiang Gao. 2026. DETECT: Determining ease and textual clarity of German text simplifications. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2852–2882, Rabat, Morocco. Association for Computational Linguistics.

Vignesh Kothapalli, Hamed Firooz, and Maziar Sanjabi. 2025. CoT-ICL lab: A synthetic framework for studying chain-of-thought learning from in-context demonstrations. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14620–14642, Vienna, Austria. Association for Computational Linguistics.

Yevheniia Kryklyvets, Mohammed Irfan Kurpath, Sahal Shaji Mullappilly, Jinxing Zhou, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, and Hisham Cholakkal. 2025. MAviS: A multimodal conversational assistant for avian species. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28613–28639, Suzhou, China. Association for Computational Linguistics.

Atharv Kulkarni, Kushagra Dixit, Vivek Srikumar, Dan Roth, and Vivek Gupta. 2025. LLM-symbolic integration for robust temporal tabular reasoning. In Findings of the Association for Computational Linguistics: ACL 2025, pages 19914–19940, Vienna, Austria. Association for Computational Linguistics.

Gaurav Kumar, Qiangjian Xi, Tanmaya Shekhar Dabral, Hooshang Ghasemi, Abishek Krishnamoorthy, Danqing Fu, Rui Min, Emilio Antunez, Zhongli Ding, and Pradyumna Narayana. 2026. Enhancing user safety: Context-aware detection of offensive query-ad pairs in multimodal search advertising. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 475–482, Rabat, Morocco. Association for Computational Linguistics.

Rohit Kumar and Chandan Nolbaria. 2025. Bridging the data gap in financial sentiment: LLM-driven augmentation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1246–1254, Vienna, Austria. Association for Computational Linguistics.

Sandeep Kumar, Samarth Garg, Sagnik Sengupta, Tirthankar Ghosal, and Asif Ekbal. 2025. MixRevDetect: Towards detecting AI-generated content in hybrid peer reviews. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 944–953, Albuquerque, New Mexico. Association for Computational Linguistics.

Sumit Kumar and Tanmay Basu. 2025. AdaBioBERT: Adaptive token sequence learning for biomedical named entity recognition. In Proceedings of the 24th Workshop on Biomedical Language Processing, pages 56–62, Viena, Austria. Association for Computational Linguistics.

Gitanjali Kumari, Jitendra Solanki, and Asif Ekbal. 2025. MemeDetoxNet: Balancing toxicity reduction and context preservation. In Findings of the Association for Computational Linguistics: ACL 2025, pages 25076–25098, Vienna, Austria. Association for Computational Linguistics.

Beong-woo Kwak, Minju Kim, Dongha Lim, Hyungjoo Chae, Dongjin Kang, Sunghwan Kim, Dongil Yang, and Jinyoung Yeo. 2025. ToolHaystack: Stress-testing tool-augmented language models in realistic long-term interactions. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24696–24727, Suzhou, China. Association for Computational Linguistics.

Roman Kyslyi, Yuliia Maksymiuk, and Ihor Pysmennyi. 2025a. Vuyko mistral: Adapting LLMs for low-resource dialectal translation. In Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 86–95, Vienna, Austria (online). Association for Computational Linguistics.

Roman Kyslyi, Artem Orlovskyi, Pavlo Khomenko, Bohdan Onyshchenko, and Zakhar Guzii. 2026. Building ASR resources for the hutsul dialect of Ukrainian. In Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects, pages 186–195, Rabat, Morocco. Association for Computational Linguistics.

Roman Kyslyi, Nataliia Romanyshyn, and Volodymyr Sydorskyi. 2025b. The UNLP 2025 shared task on detecting social media manipulation. In Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 105–111, Vienna, Austria (online). Association for Computational Linguistics.

Ilya Lasy, Peter Knees, and Stefan Woltran. 2025. Understanding verbatim memorization in LLMs through circuit discovery. In Proceedings of the First Workshop on Large Language Model Memorization (L2M2), pages 83–94, Vienna, Austria. Association for Computational Linguistics.

Sunbowen Lee, Junting Zhou, Chang Ao, Kaige Li, Xeron Du, Sirui He, Haihong Wu, Tianci Liu, Jiaheng Liu, Hamid Alinejad-Rokny, Min Yang, Yitao Liang, Zhoufutu Wen, and Shiwen Ni. 2025. Quantification of large language model distillation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4985–5004, Vienna, Austria. Association for Computational Linguistics.

Kuangdai Leng, Jia Bi, Samuel Pinilla, and Jaehoon Cha. 2026. Token-wise kernels (TWiKers) for vicinity-aware attention in transformers. In Findings of the Association for Computational Linguistics: EACL 2026, pages 5819–5835, Rabat, Morocco. Association for Computational Linguistics.

Haley Lepp and Daniel Scott Smith. 2025. “you cannot sound like gpt": Signs of language discrimination and resistance in computer science publishing. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’25, page 3162–3181. ACM.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics. Doklady, 10:707–710.

Amit LeVi, Rom Himelstein, Yaniv Nemcovsky, Avi Mendelson, and Chaim Baskin. 2025. Jailbreak attack initializations as extractors of compliance directions. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 6672–6705, Suzhou, China. Association for Computational Linguistics.

Chenhui Li and Weihai Lu. 2026. Decoding the market’s pulse: Context-enriched agentic retrieval augmented generation for predicting post-earnings price shocks. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3055–3073, Rabat, Morocco. Association for Computational Linguistics.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025a. Language models can easily learn to reason from demonstrations. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15979–15997, Suzhou, China. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025b. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Jiazheng Li, Hanqi Yan, and Yulan He. 2025c. Drift: Enhancing LLM faithfulness in rationale generation via dual-reward probabilistic inference. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6850–6866, Vienna, Austria. Association for Computational Linguistics.

Matthew Li, Santiago Torres-Garcia, Shayan Halder, Phani Kuppa, Sean O’Brien, Vasu Sharma, Kevin Zhu, and Sunishchal Dev. 2025d. FrontierScience bench: Evaluating AI research capabilities in LLMs. In Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025), pages 428–453, Vienna, Austria. Association for Computational Linguistics.

Muyao Li, Zihao Wang, Kaichen He, Xiaojian Ma, and Yitao Liang. 2025e. JARVIS-VLA: Post-training large-scale vision language models to play visual games with keyboards and mouse. In Findings of the Association for Computational Linguistics: ACL 2025, pages 17878–17899, Vienna, Austria. Association for Computational Linguistics.

Pingzhi Li, Prateek Yadav, Jaehong Yoon, Jie Peng, Yi-Lin Sung, Mohit Bansal, and Tianlong Chen. 2025f. Glider: Global and local instruction-driven expert router. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6240–6301, Suzhou, China. Association for Computational Linguistics.

Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, and Zhifang Sui. 2025g. How far are LLMs from being our digital twins? a benchmark for persona-based behavior chain simulation.

In Findings of the Association for Computational Linguistics: ACL 2025, pages 15738–15763, Vienna, Austria. Association for Computational Linguistics.

Runchao Li, Yao Fu, Mu Sheng, Xianxuan Long, Haotian Yu, and Pan Li. 2025h. FAEDKV: Infinite-window Fourier transform for unbiased KV cache compression. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16856–16866, Suzhou, China. Association for Computational Linguistics.

Shenglan Li, Jia Xu, and Mengjiao Zhang. 2025i. EmByte: Decomposition and compression learning for small yet private NLP. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7182–7201, Suzhou, China. Association for Computational Linguistics.

Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2024. DrAttack: Prompt decomposition and reconstruction makes powerful LLMs jailbreakers. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 13891–13913, Miami, Florida, USA. Association for Computational Linguistics.

Zehan Li, Ruhua Pan, and Xinyu Pi. 2025j. Beyond LLMs a linguistic approach to causal graph generation from narrative texts. In Proceedings of the The 7th Workshop on Narrative Understanding, pages 36–51, Albuquerque, New Mexico. Association for Computational Linguistics.

Zichao Li. 2025a. Formula-text cross-retrieval: A benchmarking study of dense embedding methods for mathematical information retrieval. In Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025), pages 124–133, Suzhou, China. Association for Computational Linguistics.

Zichao Li. 2025b. Knowledge-grounded detection of cryptocurrency scams with retrieval-augmented LMs. In Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM), pages 40–48, Vienna, Austria. Association for Computational Linguistics.

Zichao Li. 2025c. Retrieval-augmented forecasting with tabular time series data. In Proceedings of the 4th Table Representation Learning Workshop, pages 192–199, Vienna, Austria. Association for Computational Linguistics.

Zichao Li and Zong Ke. 2025a. Cross-modal augmentation for low-resource language understanding and generation. In Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025), pages 90–99, Vienna, Austria. Association for Computational Linguistics.

Zichao Li and Zong Ke. 2025b. Domain meets typology: Predicting verb-final order from Universal Dependencies for financial and blockchain NLP. In Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 156–164, Vienna, Austria. Association for Computational Linguistics.

Zichao Li, Zong Ke, and Puning Zhao. 2025k. Injecting structured knowledge into LLMs via graph neural networks. In Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025), pages 16–25, Vienna, Austria. Association for Computational Linguistics.

Chuang Liang, Wei Li, and Yanqiu Shao. 2025. Ctrl-Shift: Steering language models for dense quotation retrieval with dynamic prompts. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2192–2204, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Ren-Wei Liang, Chin Ting Hsu, Chan-Hung Yu, Saransh Agrawal, Shih-Cheng Huang, Chieh-Yen Lin, Shang-Tse Chen, Kuan-Hao Huang, and Shao-Hua Sun. 2026. Adaptive helpfulness–harmlessness alignment with preference vectors. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1646–1668, Rabat, Morocco. Association for Computational Linguistics.

Yuan Ling, Chunqing Yuan, Shujing Dong, Yongjian Yang, Nataraj Mocherla, and Ayush Goyal. 2026. ShopperBench: A benchmark for personalized shopping with persona-guided simulation. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 837–846, Rabat, Morocco. Association for Computational Linguistics.

Emmy Liu, Amanda Bertsch, Lintang Sutawika, Lindia Tjuatja, Patrick Fernandes, Lara Marinov, Michael Chen, Shreya Singhal, Carolin Lawrence, Aditi Raghunathan, Kiril Gashteovski, and Graham Neubig. 2025a. Not-just-scaling laws: Towards a better understanding of the downstream impact of language model design decisions. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16396–16427, Suzhou, China. Association for Computational Linguistics.

Jiarui Liu, Yueqi Song, Yunze Xiao, Mingqian Zheng, Lindia Tjuatja, Jana Schaich Borg, Mona T. Diab, and Maarten Sap. 2025b. Synthetic socratic debates: Examining persona effects on moral decision and persuasion dynamics. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16428–16458, Suzhou, China. Association for Computational Linguistics.

Quancai Liu, Haihui Fan, Jinchao Zhang, Xiangfang Li, Chuanrong Li, and Bo Li. 2025c. DisComp: A two-stage prompt optimization framework combining task-agnostic and task-aware compression.

In Findings of the Association for Computational Linguistics: NAACL 2025, pages 1033–1044, Albuquerque, New Mexico. Association for Computational Linguistics.

Siyi Liu and Dan Roth. 2025. Conflicts in texts: Data, implications and challenges. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10073–10091, Suzhou, China. Association for Computational Linguistics.

Wanlong Liu, Junying Chen, Ke Ji, Li Zhou, Wenyu Chen, and Benyou Wang. 2025d. RAG-instruct: Boosting LLMs with diverse retrieval-augmented instructions. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 3865–3888, Suzhou, China. Association for Computational Linguistics.

Yifan Liu, Qianfeng Wen, Mark Zhao, Jiazhou Liang, and Scott Sanner. 2025e. MA-DPR: Manifold-aware distance metrics for dense passage retrieval. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31085–31103, Suzhou, China. Association for Computational Linguistics.

Zhiwei Liu, Jielin Qiu, Shiyu Wang, Jianguo Zhang, Zuxin Liu, Roshan Ram, Haolin Chen, Weiran Yao, Shelby Heinecke, Silvio Savarese, Huan Wang, and Caiming Xiong. 2025f. MCPEval: Automatic MCP-based deep evaluation for AI agent models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 373–402, Suzhou, China. Association for Computational Linguistics.

Zhixiang Lu, Mian Zhou, Angelos Stefanidis, and Jionglong Su. 2025. UniMath-CoT: A unified framework for multimodal mathematical reasoning with re-inference affirmation. In Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025), pages 176–185, Suzhou, China. Association for Computational Linguistics.

Hanjun Luo, Yingbin Jin, Yiran Wang, Xinfeng Li, Tong Shang, Xuecheng Liu, Ruizhe Chen, Kun Wang, Hanan Salam, Qingsong Wen, and Zuozhu Liu. 2025a. DynamicNER: A dynamic, multilingual, and fine-grained dataset for LLM-based named entity recognition. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16511–16535, Suzhou, China. Association for Computational Linguistics.

Xiangfeng Luo, Ruoxin Zheng, Jianqiang Huang, and Hang Yu. 2025b. HDiff: Confidence-guided denoising diffusion for robust hyper-relational link prediction. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7417–7434, Suzhou, China. Association for Computational Linguistics.

Lijia Lv, Yuanshu Zhao, Guan Wang, Xuehai Tang, Wen Jie, Jizhong Han, and Songlin Hu. 2025. Gamma-guard: Lightweight residual adapters for robust guardrails in large language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12219–12231, Suzhou, China. Association for Computational Linguistics.

Jiawen Lyn and Yvette Graham. 2025. TransLaTeX: Exposing the last-mile execution gap in LLM-agent for scientific formatting. In Proceedings of The First Workshop on Human–LLM Collaboration for Ethical and Responsible Science Production (SciProdLLM), pages 19–24, Mumbai, India (Hybrid). Association for Computational Linguistics.

Adrian Denzel Macayan, Luis Andrew Sunga Madridijo, Ellexandrei Esponilla, and Zachary Mitchell Francisco. 2026. Tao–Filipino neural machine translation: Strategies for ultra–low-resource settings. In Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 27–36, Rabat, Morocco. Association for Computational Linguistics.

Jabez Magomere, Elena Kochkina, Samuel Mensah, Simerjot Kaur, and Charese Smiley. 2025. FinNLI: Novel dataset for multi-genre financial natural language inference benchmarking. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 4545–4568, Albuquerque, New Mexico. Association for Computational Linguistics.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, and Sriparna Saha. 2025. SANSKRITI: A comprehensive benchmark for evaluating language models' knowledge of Indian culture. In Findings of the Association for Computational Linguistics: ACL 2025, pages 4434–4451, Vienna, Austria. Association for Computational Linguistics.

Vijit Malik, Akshay Jagatap, Vinayak S Puranik, and Anirban Majumder. 2024. PEARL: Preference extraction with exemplar augmentation and retrieval with LLM agents. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1536–1547, Miami, Florida, US. Association for Computational Linguistics.

Stasa Mandic, Georg Niess, and Roman Kern. 2025. From in-distribution to out-of-distribution: Joint loss for improving generalization in software mention and relation extraction. In Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 146–153, Vienna, Austria. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK.

Utsav Maskey, Chencheng Zhu, and Usman Naseem. 2025. Benchmarking large language models for cryptanalysis and side-channel vulnerabilities. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19849–19865, Suzhou, China. Association for Computational Linguistics.

Blake Matheny, Phuong Minh Nguyen, and Minh Le Nguyen. 2025. JNLP at SemEval-2025 task 1: Multimodal idiomaticity representation with large language models. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1479–1484, Vienna, Austria. Association for Computational Linguistics.

Abhijit Mishra, Mingda Li, Hsiang Fu, Richard Noh, and Minji Kim. 2025. ReVision: A dataset and baseline VLM for privacy-preserving task-oriented visual instruction rewriting. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2877–2889, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Kshitij Mishra, Nils Lukas, and Salem Lahlou. 2026. SD-e2: Semantic exploration for reasoning under token budgets. In Findings of the Association for Computational Linguistics: EACL 2026, pages 6144–6157, Rabat, Morocco. Association for Computational Linguistics.

Saif M. Mohammad. 2025. Words of warmth: Trust and sociability norms for over 26k English words. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18830–18850, Vienna, Austria. Association for Computational Linguistics.

Seyedali Mohammadi, Bhaskara Hanuma Vedula, Hemank Lamba, Edward Raff, Ponnurangam Kumaraguru, Francis Ferraro, and Manas Gaur. 2025. Do LLMs adhere to label definitions? examining their receptivity to external label definitions. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32380–32393, Suzhou, China. Association for Computational Linguistics.

Joydeb Mondal and Pramir Sarkar. 2025. Modgenix at SemEval-2025 task 1: Context aware vision language ranking (CAViLR) for multimodal idiomaticity understanding. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 780–784, Vienna, Austria. Association for Computational Linguistics.

Hoyeon Moon, Byeolhee Kim, and Nikhil Verma. 2025. Quality-aware translation tagging in multilingual RAG system. In Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 161–177, Suzhuo, China. Association for Computational Linguistics.

Guy Mor-Lan, Naama Rivlin-Angert, Yael R. Kaplan, Tamir Sheafer, and Shaul R. Shenhav. 2025. HebID: Detecting social identities in Hebrew-language political text. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9850–9870, Suzhou, China. Association for Computational Linguistics.

Yasmin Moslem, Aman Kassahun Wassie, and Amanuel Gizachew Abebe. 2026. AfriNLLB: Efficient translation models for African languages. In Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), pages 281–292, Rabat, Morocco. Association for Computational Linguistics.

Hamdy Mubarak, Majd Hawasly, and Abubakr Mohamed. 2026. Nahw: A comprehensive benchmark of Arabic grammar understanding, error detection, correction, and explanation. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6310–6328, Rabat, Morocco. Association for Computational Linguistics.

Andreas C Mueller. 2023. Wordcloud.

Gagan Mundada, Yash Vishe, Amit Namburi, Xin Xu, Zachary Novack, Julian McAuley, and Junda Wu. 2025. WildScore: Benchmarking MLLMs in-the-wild symbolic music reasoning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16847–16863, Suzhou, China. Association for Computational Linguistics.

Nandini Mundra, Aditya Nanda Kishore Khandavally, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M Khapra. 2024. An empirical comparison of vocabulary expansion and initialization approaches for language models. In Proceedings of the 28th Conference on Computational Natural Language Learning, pages 84–104, Miami, FL, USA. Association for Computational Linguistics.

Tafazzul Nadeem, Riyansha Singh, Suyamoon Pathak, and Ashutosh Modi. 2025. Exploration lab IITK at SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1859–1865, Vienna, Austria. Association for Computational Linguistics.

Mahmud Wasif Nafee, Maiqi Jiang, Haipeng Chen, and Yanfu Zhang. 2025. Dynamic retriever for in-context knowledge editing via policy optimization. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16744–16757, Suzhou, China. Association for Computational Linguistics.

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj Alam. 2025.

TituLLMs: A family of Bangla LLMs with comprehensive benchmarking. In Findings of the Association for Computational Linguistics: ACL 2025, pages 24922–24940, Vienna, Austria. Association for Computational Linguistics.

Kamel Nebhi, Amrita Panesar, and Hans Bantilan. 2025. End-to-end automated item generation and scoring for adaptive English writing assessment with large language models. In Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), pages 968–977, Vienna, Austria. Association for Computational Linguistics.

Grace Kevine Ngoufo, Shamsuddeen Hassan Muhammad, and Kevin Jeff Fogang Fokoa. 2026. Building a conversational AI assistant for African travel services with LLMs and RAG. In Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), pages 171–178, Rabat, Morocco. Association for Computational Linguistics.

Bao Nguyen, Binh Nguyen, Duy Nguyen, and Viet Anh Nguyen. 2025a. Distributional surgery for language model activations. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8192–8212, Suzhou, China. Association for Computational Linguistics.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2025b. Do large language models align with core mental health counseling competencies? In Findings of the Association for Computational Linguistics: NAACL 2025, pages 7503–7526, Albuquerque, New Mexico. Association for Computational Linguistics.

Xuanfan Ni and Piji Li. 2024a. A systematic evaluation of large language models for natural. ArXiv, abs/2405.10251. Note: Cited from Semantic Scholar.

Xuanfan Ni and Piji Li. 2024b. A systematic evaluation of large language models for natural language generation tasks. Preprint, arXiv:2405.10251. Note: Cited from ArXiv.

Natapong Nitarach, Warit Sirichotedumrong, Panop Pitchayarthorn, Pittawat Taveekitworachai, Potsawee Manakul, and Kunat Pipatanakul. 2025. FinCoT: Grounding chain-of-thought in expert financial reasoning. In Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing, pages 93–123, Suzhou, China. Association for Computational Linguistics.

Syeda Alisha Noor, Sadia Anjum, Syed Ahmad Reza, and Md Rashadur Rahman. 2025. Celestia@DravidianLangTech 2025: Malayalam-BERT and m-BERT based transformer models for fake news detection in Dravidian languages. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 688–693, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Farhad Nooralahzadeh, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. 2025. Multi-modal data exploration via language agents. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 795–813, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Badal Nyalang. 2026. NE-BERT: A multilingual language model for nine Northeast Indian languages. In Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 1–12, Rabat, Morocco. Association for Computational Linguistics.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal LLMs with cultural knowledge. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24187–24231, Suzhou, China. Association for Computational Linguistics.

Akhil Rajeev P. 2025. akhilrajeevp at BHASHA task 1: Minimal-edit instruction tuning for low-resource Indic GEC. In Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 164–174, Mumbai, India. Association for Computational Linguistics.

Akhil Rajeev P and Annarao Kulkarni. 2025. Accent placement models for rigvedic Sanskrit text. In Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 122–126, Mumbai, India. Association for Computational Linguistics.

Samarth P and Sanjay Balaji Mahalingam. 2025. The gemma sutras: Fine-tuning gemma 3 for Sanskrit sandhi splitting. In Proceedings of the 9th Widening NLP Workshop, pages 235–241, Suzhou, China. Association for Computational Linguistics.

Marc Pàmies, Javier Aula-Blasco, Aitor Gonzalez-Agirre, and Marta Villegas. 2026. Vinclat: Evaluating reasoning, cognition and culture in one game. In Proceedings of the First Workshop on Multilingual Multicultural Evaluation, pages 49–66, Rabat, Morocco. Association for Computational Linguistics.

Solmaz Panahi, John Kelleher, and Vasudevan Nedumpozhimana. 2026. When LLMs annotate: Reliability challenges in low-resource NLI. In Proceedings of the Second Workshop on Language

Models for Low-Resource Languages (LoResLM 2026), pages 178–188, Rabat, Morocco. Association for Computational Linguistics.

Mihir Panchal, Deeksha Varshney, Mamta ., and Asif Ekbal. 2026. Indic-TunedLens: Interpreting multilingual models in Indian languages. In Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects, pages 172–185, Rabat, Morocco. Association for Computational Linguistics.

Manas Pandya, Avinash Kumar Sharma, and Arpit Shukla. 2025. Swahili news classification: Performance, challenges, and explainability across ML, DL, and transformers. In Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 203–209, Vienna, Austria. Association for Computational Linguistics.

Lei Pang, Hanyi Mao, Quanjia Xiao, Chen Ruihan, Jingjun Zhang, Haixiao Liu, and Xiangyi Li. 2025. In2X at WMT25 translation task. In Proceedings of the Tenth Conference on Machine Translation, pages 671–679, Suzhou, China. Association for Computational Linguistics.

Angelina Parfenova and Jürgen Pfeffer. 2025. Measuring what matters: Evaluating ensemble LLMs with label refinement in inductive coding. In Findings of the Association for Computational Linguistics: ACL 2025, pages 10803–10816, Vienna, Austria. Association for Computational Linguistics.

Seongwan Park, Taeklim Kim, and Youngjoong Ko. 2025. Decoding dense embeddings: Sparse autoencoders for interpreting and discretizing dense retrieval. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26468–26485, Suzhou, China. Association for Computational Linguistics.

Nirvan Patil, Malhar Abhay Inamdar, Agnivo Gosai, Guruprasad Pathak, Anish Joshi, Anish Joshirao, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. 2025. Regional-TinyStories: A small language model framework for evaluating language learning, tokenizers, and datasets. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2318–2367, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Priyaranjan Pattnayak, Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, and Srikant Panda. 2025a. Hybrid AI for responsive multi-turn online conversations with novel dynamic routing and feedback adaptation. In Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing, pages 215–229, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Priyaranjan Pattnayak and Sanchari Chowdhuri. 2026. IndicJR: A judge-free benchmark of jailbreak robustness in South Asian languages. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 649–668, Rabat, Morocco. Association for Computational Linguistics.

Priyaranjan Pattnayak, Sanchari Chowdhuri, Amit Agarwal, and Hitesh Laxmichand Patel. 2025b. LLM-guided lifecycle-aware clustering of multi-turn customer support conversations. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3180–3206, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Soma Paul, Sukhada Sukhada, Bidisha Bhattacharjee, Kumari Riya, Sashank Tatavolu, Kamesh R, Isma Anwar, and Pratibha Rani. 2025. Indian grammatical tradition-inspired universal semantic representation bank (USR bank 1.0). In Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 11–22, Mumbai, India. Association for Computational Linguistics.

Tzu-Cheng Peng, Chien Chin Chen, and Yung-Chun Chang. 2026. Hierarchical user intent inference with knowledge graph grounding. In Findings of the Association for Computational Linguistics: EACL 2026, pages 5371–5377, Rabat, Morocco. Association for Computational Linguistics.

Nguyen Pham Hoang Le, An Dinh Thien, Son T. Luu, and Kiet Van Nguyen. 2025. DocIE@XLLM25: ZeroSemble - robust and efficient zero-shot document information extraction with heterogeneous large language model ensembles. In Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025), pages 288–297, Vienna, Austria. Association for Computational Linguistics.

Pakawat Phasook, Rapepong Pitijaroonpong, Jiramet Kinchagawat, Amrest Chinkamol, Tossaporn Saengja, Kiartnarin Udomlapsakul, Jitkapat Sawatphol, and Piyalitt Ittichaiwong. 2025. VeReaFine: Iterative verification reasoning refinement RAG for hallucination-resistant on open-ended clinical QA. In Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks), pages 281–288, Vienna, Austria. Association for Computational Linguistics.

Arpan Phukan, Anupam Pandey, Deepjyoti Bodo, and Asif Ekbal. 2025. VideoChain: A transformer-based framework for multi-hop video question generation. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the

Association for Computational Linguistics, pages 2246–2266, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Renjie Pi, Haoping Bai, Qibin Chen, Xiaoming Simon Wang, Jiulong Shan, Xiaojiang Liu, and Meng Cao. 2025. MR. judge: Multimodal reasoner as a judge. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20181–20205, Suzhou, China. Association for Computational Linguistics.

Varivashya Poladi and Sandipan Dandapat. 2025. Comparing language models of different scales for security-focused tabular query generation and reasoning. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1002–1016, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. Preprint, arXiv:2205.01833.

Emiliana Pulido, Robert Pugh, and Zoey Liu. 2025. I speak for the árboles: Developing a dependency treebank for Spanish L2 and heritage speakers. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 814–822, Vienna, Austria. Association for Computational Linguistics.

Saif Punjwani and Larry Heck. 2025. Weight-of-thought reasoning: Exploring neural network weights for enhanced LLM reasoning. In Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025), pages 471–485, Vienna, Austria. Association for Computational Linguistics.

Junlang Qian, Zixiao Zhu, Hanzhang Zhou, Zijian Feng, Zepeng Zhai, and Kezhi Mao. 2025. Beyond the next token: Towards prompt-robust zero-shot classification via efficient multi-token prediction. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7093–7115, Albuquerque, New Mexico. Association for Computational Linguistics.

Peijun Qing, Xingjian Diao, Chiyu Ma, Saeed Hassanpour, and Soroush Vosoughi. 2026. Tailoring memory granularity for multi-hop reasoning over long contexts. In Findings of the Association for Computational Linguistics: EACL 2026, pages 3648–3666, Rabat, Morocco. Association for Computational Linguistics.

Neel Prabhanjan Rachamalla, Aravind Konakalla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. Pragyaan: Designing and curating high-quality cultural post-training datasets for Indian languages. In Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 285–321, Suzhuo, China. Association for Computational Linguistics.

Javad Rafiei Asl, Sidhant Narula, Mohammad Ghasemigol, Eduardo Blanco, and Daniel Takabi. 2025. NEXUS: Network exploration for eXploiting unsafe sequences in multi-turn LLM jailbreaks. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24267–24295, Suzhou, China. Association for Computational Linguistics.

S M Rafiuddin and Muntaha Nujat Khan. 2025a. A formal analysis of chain-of-thought prompting via Turing reductions. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 87–98, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

S M Rafiuddin and Muntaha Nujat Khan. 2025b. Learning what to remember: Adaptive probabilistic memory retention for memory-efficient language models. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3969–3981, Suzhou, China. Association for Computational Linguistics.

Shahid Iqbal Rai, Danilo Croce, and Roberto Basili. 2025. Injecting frame semantics into large language models via prompt-based fine-tuning. In Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025), pages 31–47, Suzhou, China. Association for Computational Linguistics.

Nishat Raihan, Mohammad Anas Jawad, Md Mezbaur Rahman, Noshin Ulfat, Pranav Gupta, Mehrab Mustafy Rahman, Santu Karmaker, and Marcos Zampieri. 2025a. Overview of BLP-2025 task 2: Code generation in Bangla. In Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 373–387, Mumbai, India. Association for Computational Linguistics.

Nishat Raihan, Joanna C. S. Santos, and Marcos Zampieri. 2025b. MojoBench: Language modeling and benchmarks for mojo. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 4109–4128, Albuquerque, New Mexico. Association for Computational Linguistics.

Rahul Raja and Arpita Vats. 2025. Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review. In Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT

2025), pages 129–143, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Gokul Srinath Seetha Ram. 2025. Citation drift: Measuring reference stability in multi-turn LLM conversations. In Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 186–191, Mumbai, India and virtual. Association for Computational Linguistics.

Manikandan Ravikiran, Rohit Saluja, and Arnav Bhavsar. 2025. INDRA: Iterative difficulty refinement attention for MCQ difficulty estimation for Indic languages. In Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 37–51, Mumbai, India. Association for Computational Linguistics.

Alex Ray, Josh Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning.

Saed Rezayi, Le An Ha, Yiyun Zhou, Andrew Houriet, Angelo D'Addario, Peter Baldwin, Polina Harik, Ann King, and Victoria Yaneva. 2025. Automated scoring of communication skills in physician-patient interaction: Balancing performance and scalability. In Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), pages 891–897, Vienna, Austria. Association for Computational Linguistics.

Naama Rivlin-Angert and Guy Mor-Lan. 2025. The enemy from within: A study of political delegitimization discourse in israeli political speech. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16634–16647, Suzhou, China. Association for Computational Linguistics.

Keonwoo Roh, Yeong-Joon Ju, and Seong-Whan Lee. 2025. XLQA: A benchmark for locale-aware multilingual open-domain question answering. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28809–28821, Suzhou, China. Association for Computational Linguistics.

Pablo Romero, Zihao Li, and Matthew Shardlow. 2025. Efficient on-device text simplification for firefox with synthetic data fine-tuning. In Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 105–115, Suzhou, China. Association for Computational Linguistics.

Diego Rossini and Lonneke Van Der Plas. 2026. Binary token-level classification with DeBERTa for all-type MWE identification: A lightweight approach with linguistic enhancement. In Findings of the Association for Computational Linguistics: EACL 2026, pages 2600–2610, Rabat, Morocco. Association for Computational Linguistics.

Moutushi Roy and Dipankar Das. 2025. NLP4Health: Multilingual clinical dialogue summarization and QA with mT5 and LoRA. In NLP-AI4Health, pages 93–97, Mumbai, India. Association for Computational Linguistics.

Kalaivani K S, Sanjay R, Thissyakkanna S M, and Nirenjhanram S K. 2025a. KSK@DravidianLangTech 2025: Political multiclass sentiment analysis of Tamil X (Twitter) comments using incremental learning. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 221–225, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Vishal S, Rajalakshmi Sivanaiah, and Angel Deborah S. 2025b. TechSSN3 at SemEval-2025 task 11: Multi-label emotion detection using ensemble transformer models and lexical rules. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 186–192, Vienna, Austria. Association for Computational Linguistics.

Jon Saad-Falcon, Rajan Pathe Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2025. LMUNIT: Fine-grained evaluation with natural language unit tests. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3303–3324, Suzhou, China. Association for Computational Linguistics.

Syeda Jannatus Saba and Steven Skiena. 2025. Evaluating language translation models by playing telephone. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10321–10336, Suzhou, China. Association for Computational Linguistics.

Mushtari Sadia, Zhenning Yang, Yunming Xiao, Ang Chen, and Amrita Roy Chowdhury. 2025. SQUiD: Synthesizing relational databases from unstructured text. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31987–32012, Suzhou, China. Association for Computational Linguistics.

Taha Sadikot and Sarika Jain. 2025. Legal document summarization: A zero-shot modular agentic workflow approach. In Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 29–40, Mumbai, India. Association for Computational Linguistics.

Muhammed Saeed, Shaina Raza, Ashmal Vayani, Muhammad Abdul-Mageed, Ali Emami, and Shady Shehata. 2025. Beyond content: How grammatical gender shapes visual representation in text-to-image models. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24673–24695, Suzhou, China. Association for Computational Linguistics.

Rishav Sahay, Lavanya Sita Tekumalla, and Anoop Saladi. 2026. MIRAGE: Metadata-guided image retrieval and answer generation for E-commerce troubleshooting. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 764–776, Rabat, Morocco. Association for Computational Linguistics.

Furkan Şahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2024a. Systematic task exploration with LLMs: A study in citation text generation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4832–4855, Bangkok, Thailand. Association for Computational Linguistics.

Furkan Şahinuç, Thy Thy Tran, Yulia Grishina, Yufang Hou, Bei Chen, and Iryna Gurevych. 2024b. Efficient performance tracking: Leveraging large language models for automated construction of scientific leaderboards. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7963–7977, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Sakunkoo and Annabella Sakunkoo. 2025. Lost and found: Computational quality assurance of crowdsourced knowledge on morphological defectivity in Wiktionary. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 998–1003, Vienna, Austria. Association for Computational Linguistics.

Hadia Mohmmedosman Ahmed Samil and David Ifeoluwa Adelani. 2026. Sudanese-Flores: Extending FLORES+ to Sudanese Arabic dialect. In Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), pages 243–247, Rabat, Morocco. Association for Computational Linguistics.

Sirat Samyoun, Yingtai Xiao, and Jian Du. 2026. Attribution-guided multi-object hallucination and bias detection in vision-language models. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4529–4548, Rabat, Morocco. Association for Computational Linguistics.

Msvpj Sathvik, Raj Sonani, and Ravi Teja Potla. 2025a. Detection of religious hate speech during elections in Karnataka. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 562–566, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Sai Sathvik, Muralidhar Palli, Keerthana NNL, Balasubramanian Palani, Jobin Jose, and Siranjeevi Rajamanickam. 2025b. Team-risers@DravidianLangTech 2025: AI-generated product review detection in Dravidian languages using transformer-based embeddings. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 33–37, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 720–736, Mexico City, Mexico. Association for Computational Linguistics.

Miriam Schirmer, Angelina Voggenreiter, Juergen Pfeffer, and Agnes Horvat. 2025. Detecting child objectification on social media: Challenges in language modeling. In Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 396–412, Vienna, Austria. Association for Computational Linguistics.

Aishwarya Selvamurugan. 2025. DravLingua@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages using late fusion of muril and Wav2Vec models. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 694–699, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Uzair Shah, Md. Rafiul Biswas, Marco Agus, Mowafa Househ, and Wajdi Zaghouani. 2024. MemeMind at ArAIEval shared task: Generative augmentation and feature fusion for multimodal propaganda detection in Arabic memes through advanced language and vision models. In Proceedings of the Second Arabic Natural Language Processing Conference, pages 467–472, Bangkok, Thailand. Association for Computational Linguistics.

Md Kaf Shahrier, Suhana Binta Rashid, Hasan Mesbaul Ali Taher, and Mohammed Moshiul Hoque. 2025. CUET_Expelliarmus at BLP2025 task 2: Leveraging instruction translation and refinement for Bangla-to-python code generation with open-source LLMs. In Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 608–614, Mumbai, India. Association for Computational Linguistics.

Krithi Shailya, Akhilesh Kumar Mishra, Gokul S Krishnan, and Balaraman Ravindran. 2025. Where should I study? biased language models decide! evaluating fairness in LMs for academic recommendations. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2291–2317, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Naim Shandi, Jason M. Merlo, and Jeffrey A. Nanzer. 2024. Decentralized picosecond synchronization for distributed wireless systems. Preprint, arXiv:2405.18384.

Chenyang Shao, Dehao Huang, Yu Li, Keyu Zhao, Weiquan Lin, Yining Zhang, Qingbin Zeng, Zhiyu Chen, Tianxing Li, Yifei Huang, Taozhong Wu, Xinyang Liu, Ruotong Zhao, Mengsheng Zhao, Jiaoyang Li, Xuhua Zhang, Yue Wang, Yuanyi Zhen, Fengli Xu, and 2 others. 2025. Omniscientist: Toward a co-evolving ecosystem of human and ai scientists. Preprint, arXiv:2511.16931.

Jianshu She, Wenhao Zheng, Zhengzhong Liu, Hongyi Wang, Eric P. Xing, Huaxiu Yao, and Qirong Ho. 2025. Token level routing inference system for edge devices. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 159–166, Vienna, Austria. Association for Computational Linguistics.

Hanwen Shen, Jiajie Lu, Yupeng Cao, and Xiaonan Yang. 2025. Enhancing scene transition awareness in video generation via post-training. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 706–721, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Haonan Sheng, Dou Hu, Lingwei Wei, Wei Zhou, and Songlin Hu. 2025. Regularized contrastive decoding with hard negative samples for LLM hallucination mitigation. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 6061–6073, Suzhou, China. Association for Computational Linguistics.

Braeden Sherritt, Isar Nejadgholi, Efstratios Aivaliotis, Khaled Mslmani, and Marzieh Amini. 2025. WildFireCan-MMD: A multimodal dataset for classification of user-generated content during wildfires in Canada. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 884–902, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Rajvee Sheth, Shubh Nisar, Heenaben Prajapati, Himanshu Beniwal, and Mayank Singh. 2024. Commentator: A code-mixed multilingual text annotation framework. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 101–109, Miami, Florida, USA. Association for Computational Linguistics.

Kaiwen Shi, Zheyuan Zhang, Zhengqing Yuan, Keerthiram Murugesan, Vincent Galassi, Chuxu Zhang, and Yanfang Ye. 2026. NG-router: Graph-supervised multi-agent collaboration for nutrition question answering. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7508–7527, Rabat, Morocco. Association for Computational Linguistics.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, Qiujie Xie, Xinyu Guo, Qu Yang, Jiayi Wu, Jujia Zhao, Xiaqiang Tang, Xinbei Ma, Cunxiang Wang, Jiaxin Mao, and 7 others. 2025. Deep research: A systematic survey. Preprint, arXiv:2512.02038.

Fariha Anjum Shifa, Muhtasim Ibteda Shochcho, Abdullah Ibne Hanif Arean, Mohammad Ashfaq Ur Rahman, Akm Moshiur Rahman Mazumder, Ahaj Mahhin Faiak, Md Fahim, M Ashraful Amin, Amin Ahsan Ali, and Akmmahbubur Rahman. 2025. SOMAJGYAAN: A dataset for evaluating LLMs on Bangla culture, social knowledge, and low-resource language adaptation. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2157–2177, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. 2025. Efficient unstructured pruning of mamba state-space models for resource-constrained environments. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11098–11126, Suzhou, China. Association for Computational Linguistics.

Jungwoo Shim, Dae Won Kim, Sunwook Kim, Sooyoung Kim, Myungcheol Lee, Jaegeun Cha, and Hyunhwa Choi. 2026. POaaS: Minimal-edit prompt optimization as a service to lift accuracy and cut hallucinations on on-device sLLMs. In Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER), pages 13–27, Rabat, Morocco. Association for Computational Linguistics.

Siddharth Shukla, Tanuj Tyagi, Abhay Singh Bisht, Ashish Sharma, and Basant Agarwal. 2025. NyayGraph: A knowledge graph enhanced approach for legal statute identification in Indian law using large language models. In Proceedings of the Natural Legal Language Processing Workshop 2025, pages 147–156, Suzhou, China. Association for Computational Linguistics.

Shikhhar Siingh, Abhinav Rawat, Chitta Baral, and Vivek Gupta. 2025. GETReason: Enhancing image context extraction through hierarchical multi-agent reasoning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29779–29800, Vienna, Austria. Association for Computational Linguistics.

Pritam Sil, Durgaprasad Karnam, Vinay Reddy Venumuddala, and Pushpak Bhattacharyya. 2026. How effective are VLMs in assisting humans in inferring the quality of mental models from multimodal short answers? In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1130–1140, Rabat, Morocco. Association for Computational Linguistics.

Divyansh Singh, Brodie Mather, Demi Zhang, Patrick Lehman, Justin Ho, and Bonnie J Dorr. 2025. FIDELITY: Fine-grained interpretable distillation for effective language insights and topic yielding. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 2460–2472, Albuquerque, New Mexico. Association for Computational Linguistics.

Pranaydeep Singh, Orphee De Clercq, and Els Lefever. 2026. Lost in activations: A neuron-level analysis of encoders for cross-lingual emotion detection. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 154–159, Rabat, Morocco. Association for Computational Linguistics.

Clemencia Siro, Pourya Aliannejadi, and Mohammad Aliannejadi. 2026. Learning to judge: LLMs designing and applying evaluation rubrics. In Findings of the Association for Computational Linguistics: EACL 2026, pages 6371–6389, Rabat, Morocco. Association for Computational Linguistics.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. 2024. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10406–10421, Bangkok, Thailand. Association for Computational Linguistics.

Konstantinos Skianis, A. Seza Doğruöz, and John Pavlopoulos. 2024. Leveraging LLMs for translating and classifying mental health data. In Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024), pages 236–241, Miami, Florida, USA. Association for Computational Linguistics.

Yejin Son, Saejin Kim, Dongjun Min, and Youngjae Yu. 2025. Multimodal UNcommonsense: From odd to ordinary and ordinary to odd. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17586–17609, Suzhou, China. Association for Computational Linguistics.

Woomin Song, Saket Dingliwal, Sai Muralidhar Jayanthi, Bhavana Ganesh, Jinwoo Shin, Aram Galstyan, and Sravan Babu Bodapati. 2025a. Accelerated test-time scaling with model-free speculative sampling. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30623–30636, Suzhou, China. Association for Computational Linguistics.

Xiaoying Song, Sharon Lisseth Perez, Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2025b. Echoes of discord: Forecasting hater reactions to counterspeech. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 4907–4920, Albuquerque, New Mexico. Association for Computational Linguistics.

Shashank Sonkar, Naiming Liu, and Richard Baraniuk. 2024. Student data paradox and curious case of single student-tutor model: Regressive side effects of training LLMs for personalized learning. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15543–15553, Miami, Florida, USA. Association for Computational Linguistics.

Himadri Sonowal and Saisab Sadhu. 2025. Structure-aware chunking for abstractive summarization of long legal documents. In Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 171–178, Mumbai, India. Association for Computational Linguistics.

Mohsen Sorkhpour, Abbas Yazdinejad, and Ali Dehghantanha. 2025. RedHit: Adaptive red-teaming of large language models via search, reasoning, and preference optimization. In Proceedings of the The First Workshop on LLM Security (LLMSEC), pages 7–16, Vienna, Austria. Association for Computational Linguistics.

Alexander Spangher, Tenghao Huang, Yiqin Huang, Lucas Spangher, Sewon Min, and Mark Dredze. 2025a. A novel multi-document retrieval benchmark: Journalist source-selection in newswriting. In Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing, pages 180–204, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Alexander Spangher, Michael Lu, Sriya Kalyan, Hyundong Justin Cho, Tenghao Huang, Weiyan Shi, and Jonathan May. 2025b. NewsInterview: a dataset and a playground to evaluate LLMs' grounding gap via informational interviews. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32895–32925, Vienna, Austria. Association for Computational Linguistics.

Maya Srikanth, Run Chen, and Julia Hirschberg. 2025. Mixed signals: Understanding model disagreement in multimodal empathy detection. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1978–1991, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Shashank Srivastava. 2025. Large language models threaten language's epistemic and communicative foundations. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28662–28676, Suzhou, China. Association for Computational Linguistics.

Walker Stanislas Rocksane COMPAORE, Maimouna Ouattara, Rodrique Kafando, Tegawendé F. Bissyandé, Abdoul Kader Kabore, and Aminata Sabane. 2026. Neural machine translation for French–mooré: Adapting large language models to low-resource languages. In Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 615–622, Rabat, Morocco. Association for Computational Linguistics.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. PaperBench: Evaluating AI' s ability to replicate AI research. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pages 56843–56873. PMLR.

Dmitrii Stoianov, Danil Taranets, Olga Tsymboi, Ramil Latypov, Almaz Dautov, Vladislav Kruglikov, Surkov Nikita, German Abramov, Pavel Gein, Dmitry Abulkhanov, Mikhail Gashkov, Viktor Zelenkovskiy, Artem Batalov, Aleksandr Medvedev, and Anatolii Potapov. 2026. T-pro 2.0: An efficient Russian hybrid-reasoning model and playground. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 297–319, Rabat, Marocco. Association for Computational Linguistics.

Ruiran Su, Markus Leippold, and Janet B. Pierrehumbert. 2026. Actors, frames and arguments: A multi-decade computational analysis of climate discourse in financial news using large language models. In Findings of the Association for Computational Linguistics: EACL 2026, pages 1994–2014, Rabat, Morocco. Association for Computational Linguistics.

Jie Sun, Junkang Wu, Jiancan Wu, Zhibo Zhu, Xingyu Lu, Jun Zhou, Lintao Ma, and Xiang Wang. 2025. Robust preference optimization via dynamic target margins. In Findings of the Association for Computational Linguistics: ACL 2025, pages 5399–5416, Vienna, Austria. Association for Computational Linguistics.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17):19053–19061.

Liwen Sun, Xiang Yu, Ming Tan, Zhuohao Chen, Anqi Cheng, Ashutosh Joshi, and Chenyan Xiong. 2026. Linking knowledge to care: Knowledge graph-augmented medical follow-up question generation. In Findings of the Association for Computational Linguistics: EACL 2026, pages 846–853, Rabat, Morocco. Association for Computational Linguistics.

Sanjay Suryanarayanan, Haiyue Song, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, and Raj Dabre. 2025. PRALEKHA: Cross-lingual document alignment for Indic languages. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 662–676, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Gokul Swamy, Aman Gulati, Srinivas Virinchi, and Anoop Saladi. 2025. An address intelligence framework for E-commerce deliveries. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1026–1034, Suzhou (China). Association for Computational Linguistics.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025a. Ai-researcher: Autonomous scientific innovation. Preprint, arXiv:2505.18705.

Zuojin Tang, Bin Hu, Chenyang Zhao, De Ma, Gang Pan, and Bin Liu. 2025b. VLASCD: A visual language action model for simultaneous chatting and decision making. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9212–9232, Suzhou, China. Association for Computational Linguistics.

Mehrab Tanjim, Yeonjun In, Xiang Chen, Victor Bursztyn, Ryan A. Rossi, Sungchul Kim, Guang-Jie Ren, Vaishnavi Muppala, Shun Jiang, Yongsung Kim, and Chanyoung Park. 2025. Disambiguation in conversational question answering in the era of LLMs and agents: A survey. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9537–9550, Suzhou, China. Association for Computational Linguistics.

Chong Tian, Qirong Ho, and Xiuying Chen. 2025. A symbolic adversarial learning framework for evolving fake news generation and detection. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12296–12310, Suzhou, China. Association for Computational Linguistics.

Pavel Tikhonov, Ivan Oseledets, and Elena Tutubalina. 2026. Confidence leaps in LLM reasoning: Early stopping and cross-model transfer. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 602–616, Rabat, Morocco. Association for Computational Linguistics.

Manyana Tiwari. 2025. Extended abstract: Probing-guided parameter-efficient fine-tuning for balancing linguistic adaptation and safety in LLM-based social influence systems. In Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025), pages 145–147, Vienna, Austria. Association for Computational Linguistics.

Tasnimul Hossain Tomal, Anam Borhan Uddin, Intesar Tahmid, Mir Sazzat Hossain, Md Fahim, and Md Farhad Alam Bhuiyan. 2025. LP-FT-LoRA: A three-stage PEFT framework for efficient domain adaptation in Bangla NLP tasks. In Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 212–222, Mumbai, India. Association for Computational Linguistics.

Raj Vardhan Tomar, Preslav Nakov, and Yuxia Wang. 2025. UnsafeChain: Enhancing reasoning model safety via hard cases. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1233–1247, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

MeiHan Tong and Shuai Wang. 2025. NovelCR: A large-scale bilingual dataset tailored for long-span coreference resolution. In Findings of the Association for Computational Linguistics: ACL 2025, pages 5161–5173, Vienna, Austria. Association for Computational Linguistics.

Octavian Alexandru Trifan, Jason Lee Weber, Marc Titus Trifan, Alexandru Nicolau, and Alexander Veidenbaum. 2025. Grammar pruning: Enabling low-latency zero-shot task-oriented language models for edge AI. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16944–16957, Suzhou, China. Association for Computational Linguistics.

Aakash Trivedi, Aniket Upadhyay, Pratik Narang, Dhruv Kumar, and Praveen Kumar. 2026. A hybrid supervised-LLM pipeline for actionable suggestion mining in unstructured customer reviews. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 823–836, Rabat, Morocco. Association for Computational Linguistics.

Maria Tseytlin, Paul Roit, Omri Abend, Ido Dagan, and Ayal Klein. 2025. QA-noun: Representing nominal semantics via natural language question-answer pairs. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2727–2741, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Umar Baba Umar, Sulaimon Adebayo Bashir, and Abdulmalik Danlami Mohammed. 2026. Anchoring the judge: Curriculum-based adaptation and reference-anchored MQM for LLM-based machine translation of an unseen low-resource language - a case of nupe. In Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 200–211, Rabat, Morocco. Association for Computational Linguistics.

Ankit Kumar Upadhyay, John S. Erickson, and Deborah L. McGuinness. 2026. Evidence grounding vs. memorization: Why neural semantics matter for knowledge graph fact verification. In Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER), pages 28–46, Rabat, Morocco. Association for Computational Linguistics.

Vijay Karthick Vaidyanathan, Srihari V K, and Thenmozhi Durairaj. 2025a. NLP_goats@DravidianLangTech 2025: Towards safer social media: Detecting abusive language directed at women in Dravidian languages. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 350–354, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Vijay Karthick Vaidyanathan, Srihari V K, Mugilkrishna D U, and Saritha Madhavan. 2025b. NLP_goats at SemEval-2025 task 11: Multi-label emotion classification using fine-tuned roberta-large tranformer. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1023–1027, Vienna, Austria. Association for Computational Linguistics.

Bhoomit Vasani, Jack FitzGerald, Anjie Fang, and Sushmit Vaish. 2025. PHLoRA: data-free post-hoc low-rank adapter extraction from full-rank checkpoint. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1992–1999, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Henrik Vatndal and Vinay Setty. 2025. Short-Check: Checkworthiness detection of multilingual Short-Form videos. In Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations, pages 77–85, Mumbai, India. Association for Computational Linguistics.

Arpita Vats, Rahul Raja, Ashish Kattamuri, and Abhinav Bohra. 2025. A systematic survey of quantum natural language processing: Models, encoding paradigms, and evaluation methods. In Proceedings of the QuantumNLP{:} Integrating Quantum Computing with Natural Language Processing, pages 53–64, Mumbai, India (Hybrid). Association for Computational Linguistics.

Harshil Vejendla. 2025a. RewriteNets: End-to-end trainable string-rewriting for generative sequence modeling. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 446–451, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Harshil Vejendla. 2025b. SliceMoE: Routing embedding slices instead of tokens for fine-grained and balanced transformer scaling. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15971–15978, Suzhou, China. Association for Computational Linguistics.

Harshil Vejendla. 2025c. Teaching by failure: Counter-example–driven curricula for transformer self-improvement. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 423–431, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Akshay Verma, Swapnil Gupta, Deepak Gupta, Prateek Sircar, and Siddharth Pillai. 2026. SELENE: Selective and evidence-weighted LLM debating for efficient and reliable reasoning. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 95–104, Rabat, Morocco. Association for Computational Linguistics.

Abhiram Vinjamuri and Weiwei Sun. 2025. Transfer learning for dependency parsing of Vedic Sanskrit. In Proceedings of the 9th Widening NLP Workshop, pages 50–55, Suzhou, China. Association for Computational Linguistics.

Sahil Wadhwa, Himanshu Kumar, Guanqun Yang, Abbaas Alif Mohamed Nishar, Pranab Mohanty, Swapnil Shinde, and Yue Wu. 2026. ART: Adaptive reasoning trees for explainable claim verification. In Findings of the Association for Computational Linguistics: EACL 2026, pages 571–586, Rabat, Morocco. Association for Computational Linguistics.

Jan Philip Wahle, Terry Ruas, Yang Xu, and Bela Gipp. 2024. Paraphrase types elicit prompt engineering capabilities. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11004–11033, Miami, Florida, USA. Association for Computational Linguistics.

Neng Wan, Sicong Huang, Esha Ubale, and Ian Lane. 2025. UCSC at SemEval-2025 task 8: Question answering over tabular data. In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2050–2058, Vienna, Austria. Association for Computational Linguistics.

Benlu Wang, Iris Xia, Yifan Zhang, Junda Wang, Feiyun Ouyang, Shuo Han, Arman Cohan, Hong Yu, and Zonghai Yao. 2025a. From scores to steps: Diagnosing and improving LLM performance in evidence-based medical calculations. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10809–10833, Suzhou, China. Association for Computational Linguistics.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. Mineru: An open-source solution for precise document content extraction. Preprint, arXiv:2409.18839.

Dashun Wang and Albert-László Barabási. 2021. The Science of Science. Cambridge University Press.

Jiayou Wang, Rundong Liu, Yue Hu, Huijia Wu, and Zhaofeng He. 2025b. SecDecoding: Steerable decoding for safer LLM generation. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20504–20521, Suzhou, China. Association for Computational Linguistics.

Qidong Wang, Junjie Hu, and Ming Jiang. 2025c. V-SEAM: Visual semantic editing and attention modulating for causal interpretability of vision-language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17396–17420, Suzhou, China. Association for Computational Linguistics.

Song Wang, Zhen Tan, Zihan Chen, Shuang Zhou, Tianlong Chen, and Jundong Li. 2025d. AnyMAC: Cascading flexible multi-agent collaboration via next-agent prediction. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11555–11567, Suzhou, China. Association for Computational Linguistics.

WenHao Wang, Mengying Yuan, Zijie Yu, Guangyi Liu, Rui Ye, Tian Jin, Siheng Chen, and Yanfeng Wang. 2025e. MobileA3gent: Training mobile GUI agents using decentralized self-sourced data from diverse users. In Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 79–112, Suzhou, China. Association for Computational Linguistics.

Xiaochen Wang, Jiaqi Wang, Houping Xiao, Jinghui Chen, and Fenglong Ma. 2024b. FEDKIM: Adaptive federated knowledge injection into medical foundation models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8141–8154, Miami, Florida, USA. Association for Computational Linguistics.

Xiaonan Wang, Bo Shao, and Hansaem Kim. 2025f. AI knows where you are: Exposure, bias, and inference in multimodal geolocation with KoreaGEO. In

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9886–9903, Suzhou, China. Association for Computational Linguistics.

Yao Wang, Di Liang, and Minlong Peng. 2025g. Not all parameters are created equal: Smart isolation boosts fine-tuning performance. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9871–9885, Suzhou, China. Association for Computational Linguistics.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025a. Deepseek-ocr: Contexts optical compression. Preprint, arXiv:2510.18234.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Ming Hu, Chenglong Ma, Shixiang Tang, Junjun He, Chunfeng Song, Xuming He, Qiang Zhang, Chenyu You, and 8 others. 2025b. From ai for science to agentic science: A survey on autonomous scientific discovery. Preprint, arXiv:2508.14111.

Qiming Wu, Zichen Chen, Will Corcoran, Misha Sra, and Ambuj Singh. 2025a. GraphEval36K: Benchmarking coding and reasoning capabilities of large language models on graph datasets. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 8110–8132, Albuquerque, New Mexico. Association for Computational Linguistics.

Zekun Wu, Seonglae Cho, Umar Mohammed, Cristian Enrique Munoz Villalobos, Kleyton Da Costa, Xin Guan, Theo King, Ze Wang, Emre Kazim, and Adriano Koshiyama. 2025b. LibVulnWatch: A deep assessment agent system and leaderboard for uncovering hidden vulnerabilities in open-source AI libraries. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 608–656, Vienna, Austria. Association for Computational Linguistics.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and Andre Freitas. 2025. SylloBio-NLI: Evaluating large language models on biomedical syllogistic reasoning. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7235–7258, Albuquerque, New Mexico. Association for Computational Linguistics.

Eric Xia, Karthik Srikumar, Keshav Karthik, Advaith Renjith, and Ashwinee Panda. 2025. Beyond the haystack: Sensitivity to context in legal reference recall. In Proceedings of the Natural Legal Language Processing Workshop 2025, pages 48–53, Suzhou, China. Association for Computational Linguistics.

Chengrui Xiang, Tengfei Ma, Xiangzheng Fu, Yiping Liu, Bosheng Song, and Xiangxiang Zeng. 2025. From knowledge to treatment: Large language model assisted biomedical concept representation for drug repurposing. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13967–13982, Suzhou, China. Association for Computational Linguistics.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6012–6025, Miami, Florida, USA. Association for Computational Linguistics.

Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T. Diab. 2025. Humanizing machines: Rethinking LLM anthropomorphism through a multi-level framework of design. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 3331–3350, Suzhou, China. Association for Computational Linguistics.

Pengcheng Xu, Sicheng Shen, Jieli Zhou, and Hongyi Xin. 2025a. Team XSZ at BioLaySumm2025: Section-wise summarization, retrieval-augmented LLM, and reinforcement learning fine-tuning for lay summaries. In Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks), pages 275–280, Vienna, Austria. Association for Computational Linguistics.

Ruiyao Xu, Noelle I. Samia, and Han Liu. 2026. DS2-instruct: Domain-specific data synthesis for large language models instruction tuning. In Findings of the Association for Computational Linguistics: EACL 2026, pages 3368–3384, Rabat, Morocco. Association for Computational Linguistics.

Wanghan Xu, Yuhao Zhou, Yifan Zhou, Qinglong Cao, Shuo Li, Jia Bu, Bo Liu, Yixin Chen, Xuming He, Xiangyu Zhao, Xiang Zhuang, Fengxiang Wang, Zhiwang Zhou, Qiantai Feng, Wenxuan Huang, Jiaqi Wei, Hao Wu, Yuejin Yang, Guangshuai Wang, and 88 others. 2025b. Probing scientific general intelligence of llms with scientist-aligned workflows. Preprint, arXiv:2512.16969.

Yuebin Xu, Zhiyi Chen, and Zeyi Wen. 2025c. EcoTune: Token-efficient multi-fidelity hyperparameter optimization for large language model inference. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7735–7745, Suzhou, China. Association for Computational Linguistics.

Zicheng Xu, Guanchu Wang, Guangyao Zheng, Yu-Neng Chuang, Alex Szalay, Xia Hu, and Vladimir Braverman. 2025d. Self-ensemble: Mitigating confidence distortion for large language models. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16603–16615, Suzhou, China. Association for Computational Linguistics.

Tanushree Ravindra Pratap Yadav. 2026. Competence collapse in code-mixed generation: Spectral evidence and mechanistic recovery via cross-lingual activation steering. In Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 29–40, Rabat, Morocco. Association for Computational Linguistics.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. Preprint, arXiv:2504.08066.

Bohao Yang, Dong Liu, Chenghao Xiao, Kun Zhao, Chen Tang, Chao Li, Lin Yuan, Yang Guang, and Chenghua Lin. 2025a. Crafting customisable characters with LLMs: A persona-driven role-playing agent framework. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20216–20240, Suzhou, China. Association for Computational Linguistics.

Dayu Yang, Antoine Simoulin, Xin Qian, Xiaoyi Liu, Yuwei Cao, Zhaopu Teng, and Grey Yang. 2025b. DocAgent: A multi-agent system for automated code documentation generation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 460–471, Vienna, Austria. Association for Computational Linguistics.

Jing Yang. 2025. Position: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pages 82399–82408. PMLR.

Zihe Ye, Jingyuan Huang, Weixin Chen, and Yongfeng Zhang. 2026. H-mem: Hybrid multi-dimensional memory management for long-context conversational agents. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7756–7775, Rabat, Morocco. Association for Computational Linguistics.

Seungyoun Yi, Minsoo Khang, and Sungrae Park. 2025. ZERA: Zero-init instruction evolving refinement agent – from zero instructions to structured prompts via principle-based optimization. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23323–23337, Suzhou, China. Association for Computational Linguistics.

Kai Yin, Xiangjue Dong, Chengkai Liu, Lipai Huang, Yiming Xiao, Zhewei Liu, Ali Mostafavi, and James Caverlee. 2025. DisastIR: A comprehensive information retrieval benchmark for disaster management. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1836–1867, Suzhou, China. Association for Computational Linguistics.

Ko Yoshida, Daiki Shiono, Kai Sato, Toko Miura, Momoka Furuhashi, and Jun Suzuki. 2025. Batch-wise convergent pre-training: Step-by-step learning inspired by child language development. In Proceedings of the First BabyLM Workshop, pages 508–524, Suzhou, China. Association for Computational Linguistics.

Yao-Ching Yu, Tsun-Han Chiang, Cheng-Wei Tsai, Chien-Ming Huang, and Wen-Kwang Tsao. 2025a. Primus: A pioneering collection of open-source datasets for cybersecurity LLM training. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10391–10413, Suzhou, China. Association for Computational Linguistics.

Zhaojian Yu, Yinghao Wu, Yilun Zhao, Arman Cohan, and Xiao-Ping Zhang. 2025b. Z1: Efficient test-time scaling with code. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2688–2712, Suzhou (China). Association for Computational Linguistics.

Chenchen Yuan, Zheyu Zhang, Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. 2025a. Probabilistic aggregation and targeted embedding optimization for collective moral reasoning in large language models. In Findings of the Association for Computational Linguistics: ACL 2025, pages 11151–11168, Vienna, Austria. Association for Computational Linguistics.

Mengying Yuan, WenHao Wang, Zixuan Wang, Yujie Huang, Kangli Wei, Fei Li, Chong Teng, and Donghong Ji. 2025b. Cross-document cross-lingual NLI via RST-enhanced graph fusion and interpretability prediction. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31607–31629, Suzhou, China. Association for Computational Linguistics.

Mengying Yuan, WenHao Wang, Zixuan Wang, Yujie Huang, Kangli Wei, Fei Li, Chong Teng, and Donghong Ji. 2025c. Cross-document cross-lingual NLI via RST-enhanced graph fusion and interpretability prediction. In Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 11–33, Suzhuo, China. Association for Computational Linguistics.

Aizan Zafar, Kshitij Mishra, and Asif Ekbal. 2024. MedLogic-AQA: Enhancing medicare question answering with abstractive models focusing on logical structures. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16844–16867, Miami, Florida, USA. Association for Computational Linguistics.

Ofir Zafrir, Igor Margulis, Dorin Shteyman, Shira Guskin, and Guy Boudoukh. 2025. FastDraft: How to train your draft. In Findings of the Association for Computational Linguistics: ACL 2025, pages 22488–22505, Vienna, Austria. Association for Computational Linguistics.

Arun Zechariah, Balu Krishna, Hannah Mary Thomas, Joy Mammen, Dipti Misra Sharma, Parameswari Krishnamurthy, Vandan Mujadia, Priyanka Dasari, and Vishnuraj Arjunaswamy. 2025. Patient-centric question answering- overview of the shared task at the second workshop on NLP and AI for multilingual and healthcare communication. In NLP-AI4Health, pages 55–68, Mumbai, India. Association for Computational Linguistics.

Tom Zehle and Matthias Aßenmacher. 2026. Can calibration of positional encodings enhance long context utilization? In Findings of the Association for Computational Linguistics: EACL 2026, pages 2268–2280, Rabat, Morocco. Association for Computational Linguistics.

Kamyar Zeinalipour, Neda Jamshidi, Seyedehbahareh Hejazi, Marco Maggini, Monica Bianchini, Simone Paoletti, and Marco Gori. 2025. FarSense: A comprehensive commonsense benchmark and evaluation framework for the Farsi language. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3529–3599, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Kamyar Zeinalipour, Andrea Zugarini, Asya Zanollo, and Leonardo Rigutini. 2026. PharmaQA.IT: an Italian dataset for Q&A in the pharmaceutical domain. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 937–947, Rabat, Morocco. Association for Computational Linguistics.

Yiming Zeng, Wanhao Yu, Zexin Li, Tao Ren, Yu Ma, Jinghan Cao, Xiyan Chen, and Tingting Yu. 2025. Bridging the editing gap in LLMs: FineEdit for precise and targeted text modifications. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2193–2206, Suzhou, China. Association for Computational Linguistics.

Hao Zhang, Bo Huang, Zhenjia Li, Xi Xiao, Hui Yi Leong, Zumeng Zhang, Xinwei Long, Tianyang Wang, and Hao Xu. 2025a. Sensitivity-LoRA : Low-load sensitivity-based fine-tuning for large language models. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13185–13199, Suzhou, China. Association for Computational Linguistics.

Hengrui Zhang, Pin-Siang Huang, Zhen Zhang, Peican Lin, Yao-Ching Yu, Bo Hu, and Yulu Du. 2025b. PathwiseRAG: Multi-dimensional exploration and integration framework. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22904–22925, Suzhou, China. Association for Computational Linguistics.

Jiaxin Zhang. 2025. Confidence-aware reasoning: Optimizing self-guided thinking trajectories in large reasoning models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2081–2095, Suzhou (China). Association for Computational Linguistics.

Shiwen Zhang, Lingxiang Wang, Hainan Zhang, Ziwei Wang, Sijia Wen, and Zhiming Zheng. 2025c. Beyond the surface: A solution-aware retrieval model for competition-level code generation. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 5237–5246, Suzhou, China. Association for Computational Linguistics.

Tianyu Zhang, Xinyu Wang, Lu Li, Zhenghan Tai, Jijun Chi, Jingrui Tian, Hailin He, and Suyuchen Wang. 2025d. STRICT: Stress-test of rendering image containing text. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21137–21150, Suzhou, China. Association for Computational Linguistics.

Wei Zhang, Hongcheng Guo, Jian Yang, Zhoujin Tian, Yi Zhang, Yan Chaoran, Zhoujun Li, Tongliang Li, Xu Shi, Liangfan Zheng, and Bo Zhang. 2024. mABC: Multi-agent blockchain-inspired collaboration for root cause analysis in micro-services architecture. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4017–4033, Miami, Florida, USA. Association for Computational Linguistics.

Yiming Zhang, Siyue Zhang, Junbo Zhao, and Chen Zhao. 2025e. RPDR: A round-trip prediction-based data augmentation framework for long-tail question answering. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21998–22012, Suzhou, China. Association for Computational Linguistics.

Yizhe Zhang and Navdeep Jaitly. 2025. SAGE: Steering dialog generation with future-aware state-action augmentation. In Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP, pages 123–132, Suzhou, China. Association for Computational Linguistics.

Yukun Zhang and Xueqing Zhou. 2025. Continuous-time attention: PDE-guided mechanisms for long-sequence transformers. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21643–21663, Suzhou, China. Association for Computational Linguistics.

Zhuoxuan Zhang, Jinhao Duan, Edward Kim, and Kaidi Xu. 2025f. Sparse neurons carry strong signals of question ambiguity in LLMs. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16081–16099, Suzhou, China. Association for Computational Linguistics.

Chuanyang Zheng, Yihang Gao, Guoxuan Chen, Han Shi, Jing Xiong, Xiaozhe Ren, Chao Huang, Zhenguo Li, and Yu Li. 2025a. Self-adjust softmax. In

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7816–7836, Suzhou, China. Association for Computational Linguistics.

Shunfeng Zheng, Meng Fang, and Ling Chen. 2025b. SpatialWebAgent: Leveraging large language models for automated spatial information extraction and map grounding. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 252–266, Vienna, Austria. Association for Computational Linguistics.

Shunfeng Zheng, Yudi Zhang, Meng Fang, Zihan Zhang, Zhitan Wu, Mykola Pechenizkiy, and Ling Chen. 2025c. Benchmarking foundation models with retrieval-augmented generation in olympic-level physics problem solving. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21927–21956, Suzhou, China. Association for Computational Linguistics.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025d. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 414–431, Suzhou, China. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.

Jingzhuo Zhou. 2025. Can mllms detect phishing? a comprehensive security benchmark suite focusing on dynamic threats and multimodal evaluation in academic environments. Preprint, arXiv:2511.15165v1. Note: The issue was fixed in version 2.

Shu Zhou, Yunyang Xuan, Yuxuan Ao, Xin Wang, Tao Fan, and Hao Wang. 2025a. MERIT: Multi-agent collaboration for unsupervised time series representation learning. In Findings of the Association for Computational Linguistics: ACL 2025, pages 24011–24028, Vienna, Austria. Association for Computational Linguistics.

WenJie Zhou, Bohan Wang, Wei Chen, and Xueqi Cheng. 2025b. BSFA: Leveraging the subspace dichotomy to accelerate neural network training. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 18834–18849, Suzhou, China. Association for Computational Linguistics.

Yue Zhou and Barbara Di Eugenio. 2025. Veracity bias and beyond: Uncovering LLMs' hidden beliefs in problem-solving reasoning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21298–21310, Vienna, Austria. Association for Computational Linguistics.

Zebiao Zhou, Hui Li, Xiangxun Zhu, and Kangzhen Liu. 2025c. TranssionMT's submission to the Indic MT shared task in WMT 2025. In Proceedings of the Tenth Conference on Machine Translation, pages 1271–1275, Suzhou, China. Association for Computational Linguistics.

Zihan Zhou, Simon Kurz, and Zhixue Zhao. 2025d. Revisiting pruning vs quantization for small language models. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12055–12070, Suzhou, China. Association for Computational Linguistics.

He Zhu, Guanhua Chen, and Wenjia Zhang. 2025a. PlanGPT: Enhancing urban planning with a tailored agent framework. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 764–783, Vienna, Austria. Association for Computational Linguistics.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025b. MultiAgentBench : Evaluating the collaboration and competition of LLM agents. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8580–8622, Vienna, Austria. Association for Computational Linguistics.

Lichao Zhu, Maria Zimina-Poirot, Cristian Valdez, and Stephane Patin. 2025c. Fine-tuning NMT models and LLMs for specialised EN-ES translation using aligned corpora, glossaries, and synthetic data: MULTITAN at WMT25 terminology shared task. In Proceedings of the Tenth Conference on Machine Translation, pages 1284–1291, Suzhou, China. Association for Computational Linguistics.

Runchuan Zhu, Bowen Jiang, Lingrui Mei, Fangkai Yang, Lu Wang, Haoxiang Gao, Fengshuo Bai, Pu Zhao, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2025d. AdaptFlow: Adaptive workflow optimization via meta-learning. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3287–3302, Suzhou, China. Association for Computational Linguistics.

Runchuan Zhu, Zinco Jiang, Jiang Wu, Zhipeng Ma, Jiahe Song, Fengshuo Bai, Dahua Lin, Lijun Wu, and Conghui He. 2025e. GRAIT: Gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 4006–4021, Albuquerque, New Mexico. Association for Computational Linguistics.

A More Suggestions

Toward easy verification formats and proper citation practices. HalluCitations are often disguised within the reference section, making perceptual detection by human reviewers costly. Although this study relies on OCR-based parsing and manual verification to identify such cases, parsing is not always perfect. Therefore, we believe that reference sections in ACL style formats could be improved to be more machine-friendly to facilitate automatic verification, e.g., by introducing clearer structural delimiters between citations, more explicit formatting of key fields such as titles, and printing URLs rather than embedding them only as hyperlinks. Moreover, we argue that verifiable identifiers, e.g., clickable links and DOIs, should be mandatory. This would facilitate verification, reduce manual effort, and enable automatic verification.

Eliminating area selection or improving area transparency. As discussed in Section 4.1, the proportion of HalluCited papers varies across research areas, with higher rates in emerging areas. We argue that this discrepancy is partly an artifact of area partitioning. Accordingly, we suggest either moving to a single area or improving transparency by publicly releasing area information for accepted papers. As cross-area research becomes more popular, fixed area selection can lead to reviewer mismatches and reduced review rigor. Given that modern review systems such as OpenReview support automated reviewer matching, explicit area selection may no longer be necessary. Alternatively, since area information is currently disclosed only partially[9], archiving it in resources such as the ACL Anthology would help reduce mismatches and support more effective reviews, including the detection of HalluCitations.

B List of HalluCited Papers

Table 8 shows all identified HalluCited papers and the corresponding evidence, i.e., HalluCitations. For privacy reasons, we present the information in reference format rather than directly displaying the HalluCited paper details. To identify the corresponding paper, please follow the provided link or refer to the reference section.

Note that the reported numbers correspond to the cases identified in our analysis. The actual number may be higher or lower depending on the definition adopted. Nevertheless, it is reasonable to conclude that the number of HalluCited papers exceeds 300. The list has also been updated since the initial preprint version and may be subject to further changes. Due to space limitations, we present only one example per paper, although some papers contain multiple HalluCitations. Therefore, if revisions are planned for the ACL Anthology and other platforms, we recommend carefully rechecking all references.

C List of HalluCited Papers Identified at AACL 2025 and EACL 2026

Table 9 shows all identified HalluCited papers and the corresponding evidence, i.e., HalluCitations, for AACL 2025 and EACL 2026. The remaining details are the same as in Section B.

Table 8: List of hallucinated papers, i.e., HalluCited Papers, and their corresponding HalluCitations.

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
1	naacl-2024	2024.naacl-short.63	Scaria et al. (2024)	Himanshu Gupta, Kevin Scaria, Swaroop Mishra, and Chitta Baral. 2024b. Beyond the data bottleneck: Optimizing instruction tuning with difficulty-based exemplar selection. ArXiv preprint.
2	naacl-2024	2024.trustnlp-1.1	Adilazuarda (2024)	Daphne Ippolito, Daniel Duckworth, Chris CallisonBurch, and Douglas Eck. 2020. Discriminating between human-produced and machine-generated text: A survey. arXiv preprint arXiv:2012.03358.
3	acl-2024	2024.acl-long.560	Siska et al. (2024)	Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2023. An adversarial winograd schema challenge at scale. arXiv preprint arXiv:2305.06300.
4	acl-2024	2024.acl-long.611	Zheng et al. (2024)	Chun-Hung Yeh, Anuradha Welivita, and Pearl Pu Faltings. 2015. A dialogue dataset containing emotional support for people in distress. arXiv preprint arXiv:1503.08895.
5	acl-2024	2024.acl-long.768	Chen et al. (2024)	Mengjie Fang, Linlin Xu, and Pascale Fung. 2020. Unsupervised cross-lingual transfer learning for contextualized word embeddings. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2616–2629.
6	acl-2024	2024.arabicnlp-1.24	Khondaker et al. (2024)	Isabel Moreno and Wei Zhang. 2024. Evaluating multilingual models on nlp tasks in arabic. Computational Linguistics, 50(3):425–445.
7	acl-2024	2024.arabicnlp-1.45	Shah et al. (2024)	Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2021. Interplay between preferences and machine learning in language model fine-tuning. arXiv preprint arXiv:2110.08413.
8	acl-2024	2024.findings-acl.241	Holt et al. (2024)	Junjie Li, Jieyu Wu, and Richard Socher. 2021. Understanding code-switching in language models. arXiv preprint arXiv:2109.04278.
9	acl-2024	2024.wassa-1.18	Alhamed et al. (2024)	David Benetka, Alicia Moreno-Moral, Lorena RomeroFombuena, and Juan Lopez-Gazpio. 2020. An annotation scheme for mental health discussions in social media. In International Conference on Computational Linguistics (Proceedings of the Conference: Long Papers, 2020), pages 2617–2627. Association for Computational Linguistics.
10	emnlp-2024	2024.conll-1.8	Mundra et al. (2024)	Adam Alabi, Saleha Nawaz, and Vincent Ng. 2022. Alabi: A light-weight approach for multilingual biomedical language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9717–9727.
11	emnlp-2024	2024.emnlp-demo.11	Sheth et al. (2024)	Stephan Druskat, Ulrike Gut, Nils Reiter, Stefan Schweter, and Manfred Stede. 2014. Atomic: An open-source tool for working with anaphora in multiple languages. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 71–76.

Continued on next page

	EVENT ID	PAPER ID	HALLUCITED PAPER	EXAMPLE OF HALLUCITATION
12	emnlp-2024	2024.emnlp-industry.112	Malik et al. (2024)	Jiajia Liu, Mengyuan Yang, Yankai Yu, Haixia Xu, Kang Li, and Xiaobo Zhou. 2023a. Enhancing customer experience with ai-driven conversational agents. arXiv preprint arXiv:2306.04325.
13	emnlp-2024	2024.emnlp-main.345	Xiao et al. (2024)	Nelson F Liu, Tony Wu, Duane S Boning, and Tanmoy Choudhury. 2020b. AI bug detector: Adversarial input detection for natural language processing models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, page 187–196.
14	emnlp-2024	2024.emnlp-main.464	Wang et al. (2024b)	Yu Gu, Robert Tinn, Hao Cheng, Youzheng Ben, Zhuozhao Liu, Jingqi Zhou, Michael Wang, Shizhuo Wang, Hongfang Zhou, and Yanshan Shen. 2021. Biomedgpt: A large-scale biomedical generative pretrained transformer for biomedical text mining. arXiv preprint arXiv:2104.07497.
15	emnlp-2024	2024.emnlp-main.617	Wahle et al. (2024)	Michael A. Covington. 2007. Mattr: A new approach to vocabulary range in natural language processing. Computational Linguistics, 33(1):70–73.
16	emnlp-2024	2024.findings-emnlp.232	Zhang et al. (2024)	Yi Chen et al. 2023. Empowering cloud rca with augmented large language models. arXiv preprint arXiv:2311.00000.
17	emnlp-2024	2024.findings-emnlp.813	Li et al. (2024)	Ruochen Wang, Ting Liu, Cho-jui Hsieh, and Boqing Gong. 2023. DPO-DIFF:on Discrete Prompt Optimization for text-to-image DIFFusion modelsgenerating Natural Language Adversarial Examples. Preprint, arXiv:2311.07998.
18	emnlp-2024	2024.findings-emnlp.912	Sonkar et al. (2024)	Yoav Goldberg. 2022. Assessing claims about large language models. arXiv preprint arXiv:2212.09273.
19	emnlp-2024	2024.findings-emnlp.981	Zafar et al. (2024)	Xinyu Wang, Tao Sun, Deqing Zou, Wei Wu, and Jiawei Han. 2021. Logic-guided data augmentation and regularization for consistency learning. arXiv preprint arXiv:2104.04379.
20	emnlp-2024	2024.mrl-1.20	Skianis et al. (2024)	Shaoxiong Ji, Yanzhao Zhang, Leilei Sun, and Jia Wang. 2022. Mentalbert: A pretrained language model for mental healthcare. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).
21	naacl-2025	2025.americasnlp-1.4	Alvarez C et al. (2025)	Manuel Mager, Arturo Oncevay, Annette Rios, Jamshidbek Mirzakhalov, and Katharina Kann. 2023. The role of computational linguistics in indigenous language revitalization: Challenges and opportunities. In Proceedings of the 1st Workshop on Computation for Indigenous Languages (C3NLP), pages 23–31.
22	naacl-2025	2025.clpsych-1.23	Antony and Schoene (2025)	Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2022. An empirical study on neural methods for legal judgment prediction. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4783–4798. Association for Computational Linguistics.

CONTINUED ON NEXT PAGE

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
23	naacl-2025	2025.dravidianlangtech-1.117	Noor et al. (2025)	G. Shimi et al. 2024. Language identification for dravidian languages: A crucial step for fake news detection in multilingual settings. TBD.
24	naacl-2025	2025.dravidianlangtech-1.118	Selvamurugan (2025)	Anil Kumar, Ajay Kumar Ojha, Shervin Malmasi, and Marcos Zampieri. 2021. Benchmarking aggression identification in social media for low-resource languages. In Proceedings of the Workshop on Online Abuse and Harms (ACL).
25	naacl-2025	2025.dravidianlangtech-1.38	S et al. (2025a)	Simran Khanuja, Kaustav Dey, El Moatez Billah Karim Nagoudi, et al. 2020. A new dataset and strong baselines for the detection of code-mixed offensive language. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1719–1726.
26	naacl-2025	2025.dravidianlangtech-1.43	Chauhan and Kumar (2025)	Douwe Kiela, Mohamed Elhoseiny, Lin Zhang, Marco Baroni, and Geoffrey Hinton. 2019. Supervised multimodal hashing for scalable cross-modal retrieval. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1740–1751.
27	naacl-2025	2025.dravidianlangtech-1.61	K et al. (2025)	F. Wu and M. Dredze. 2019. Social media as a sensor of public opinion. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), 2019:1001–1010.
28	naacl-2025	2025.dravidianlangtech-1.62	Vaidyanathan et al. (2025a)	Jiho Park, Jihyung Shin, Sangyoun Lee, and Changhyun Seo. 2018. A survey of hate speech detection: Data, methods, and challenges. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), pages 385–395. International Committee on Computational Linguistics.
29	naacl-2025	2025.dravidianlangtech-1.7	Sathvik et al. (2025b)	M. H. Al-Adhaileh and F. W. Alsaade. 2022. Bidirectional long-short term memory (bilstm) networks for fake review detection: A comparative study. Springer.
30	naacl-2025	2025.dravidianlangtech-1.88	Achamaleh et al. (2025)	Anna Kolesnikova and Sergey Ivanov. 2023. Exploring multilingual text representations with transformer models. Transactions of the ACL, 11:212–230.
31	naacl-2025	2025.dravidianlangtech-1.91	Aftahee et al. (2025)	R. Raja, A. Kumar, and S. Joseph. 2023. Fake news detection in low-resource languages: Challenges and advancements. Computational Linguistics Review, 15(2):123–137.
32	naacl-2025	2025.dravidianlangtech-1.97	Sathvik et al. (2025a)	Simran Khanuja, Sandipan Dandapat, Ritesh Kumar, Sunayana Sitaram, K. P. Soman, and Anup Kumar. 2021. Mahanlp: Towards indic language understanding using bert models for hindi, marathi, and kannada. arXiv preprint arXiv:2106.07469.
33	naacl-2025	2025.findings-naacl.132	Singh et al. (2025)	W. Liu, X. Li, W. Yang, Y. Lin, X. Liu, S. Wang, C. Xie, L. Xu, W. Zhuang, X. Zhao, and L. Li. 2020. Minilm: Deep self-attention distillation for tiny transformers. arXiv preprint arXiv:2002.10957.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
34	naacl-2025	2025.findings-naacl.223	Zhu et al. (2025e)	Krishnateja Killamsetty and Rishabh Iyer. 2021. Subset selection with gradient matching. arXiv preprint arXiv:2103.00123.
35	naacl-2025	2025.findings-naacl.230	Raihan et al. (2025b)	OpenAI. 2024. Gpt-4 omni: A comprehensive multimodal model for language, vision, and beyond. arXiv preprint arXiv:2408.01234.
36	naacl-2025	2025.findings-naacl.257	Magomere et al. (2025)	Chaitanya Shivade et al. 2019. Mednli-a natural language inference dataset for the clinical domain. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics, pages 1586–1596.
37	naacl-2025	2025.findings-naacl.275	Song et al. (2025b)	M Honnibal and I Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. neural machine translation. In Proceedings of the Association for Computational Linguistics (ACL), pages 688–697.
38	naacl-2025	2025.findings-naacl.279	Basher et al. (2025)	Y. Xu et al. 2023. Cross-lingual transfer for lowresource text-to-speech. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 510–520. Association for Computational Linguistics.
39	naacl-2025	2025.findings-naacl.349	Gupta et al. (2025a)	Zijun Zhang, Yue Wu, Hao Guan, Xinlei Chen, and Yue Zhang. 2023. Continual learning with transformers: Challenges and solutions. arXiv preprint arXiv:2302.13713.
40	naacl-2025	2025.findings-naacl.418	Nguyen et al. (2025b)	Danny Hernandez, Jared Kaplan, Tom Henighan, Tom Brown, Jack Clark, Preetum Nakkiran, Catherine Olsson, and Afshin Rahimi. 2023. Measuring data contamination in large language models. arXiv preprint arXiv:2303.13375.
41	naacl-2025	2025.findings-naacl.452	Wu et al. (2025a)	Abhijit Srivastava, Nathan Major, Jasmine D. Hernandez, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating llm capabilities with big-bench. arXiv preprint arXiv:2206.04615.
42	naacl-2025	2025.findings-naacl.58	Liu et al. (2025c)	Xiang Jiang et al. 2023c. Mistral-7b-instruct-v0.2: An efficient large language model for instruction following. arXiv preprint arXiv:2309.00001.
43	naacl-2025	2025.knowledgenlp-1.18	Spangher et al. (2025a)	Tushar Khot, Ashish Sabharwal, and et al. 2023. Decomposition-driven reasoning in language models. arXiv preprint arXiv:2304.xxxxx.
44	naacl-2025	2025.knowledgenlp-1.20	Pattnayak et al. (2025a)	et al. Zhao, W. 2020. A retrieval-augmented encoder-decoder for knowledge-intensive nlp tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).
45	naacl-2025	2025.loresmt-1.12	Raja and Vats (2025)	Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. Code-mixing: A challenge for language identification in the indian context. In Proceedings of the First Workshop on Computational Approaches to Code Switching (EMNLP), pages 13–23.
46	naacl-2025	2025.naacl-demo.39	Hiray and Kovatchev (2025)	T. Lee et al. 2023. Bivdiff: A training-free framework for general-purpose video synthesis. arXiv preprint arXiv:2302.02918.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
47	naacl-2025	2025.naacl-industry.68	Jararweh et al. (2025)	Haotian Liu, Chunyuan Lin, Fangyun Zeng, and et al. 2023a. Llava: Large language and vision assistant. arXiv preprint arXiv:2304.08485.
48	naacl-2025	2025.naacl-long.182	Chen et al. (2025b)	Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023. Falcon-40b: an open large language model with stateof-the-art performance. Findings of the Association for Computational Linguistics: ACL, 2023:10755–10773.
49	naacl-2025	2025.naacl-long.363	Qian et al. (2025)	Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. Can prompt-based learning be practical for low-resource tasks? arXiv preprint arXiv:2001.07676.
50	naacl-2025	2025.naacl-long.371	Wysocka et al. (2025)	Jakob Prange, Khalil Mrini, and Noah A. Smith. 2023. Challenges and opportunities in nlp for systematic generalization: Beyond compositionality. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.
51	naacl-2025	2025.naacl-long.620	Chen et al. (2025c)	Jiawei Liu, Yifeng Ding, Naman Jain, Harm de Vries, Leandro von Werra, Arjun Guha, Lingming Zhang, and Yuxiang Wei. 2024a. Starcoder2-instruct: Fully transparent and permissive self-alignment for code generation. Preprint, arXiv:2307.08701.
52	naacl-2025	2025.naacl-long.72	Hu et al. (2025)	Mor Geva, Yoav Goldberg, and Jonathan Berant. 2021. Strategyqa: A question answering benchmark for reasoning about strategies. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.
53	naacl-2025	2025.naacl-short.79	Kumar et al. (2025)	Sankha Bhattacharjee et al. 2023. Adversarial learning for robust ai-generated text detection. arXiv preprint arXiv:2304.07812.
54	naacl-2025	2025.nlp4dh-1.52	Greschner and Klinger (2025)	Chiara Leoni, Mauro Coccoli, Ilaria Torre, and Gianni Vercelli. 2018. Your paper title here. In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy. Accademia University Press.
55	naacl-2025	2025.wnu-1.10	Li et al. (2025j)	Xingxing Lu, Yuning Mao, and Jason Wei. 2023. Autoeval: Llm-based automatic evaluation framework for text summarization. In Findings of ACL.
56	acl-2025	2025.acl-demo.16	She et al. (2025)	Wenhao Zheng and 1 others. 2025a. Citer: Confidencebased token routing for collaborative inference with large language models. arXiv preprint arXiv:2503.01013.
57	acl-2025	2025.acl-demo.25	Zheng et al. (2025b)	Wei Zhang, Pengfei Liu, and Qiang Shen. 2020. Handling ambiguities in geographic texts using nlp techniques. Transactions in GIS, 24(3):519–538.
58	acl-2025	2025.acl-demo.44	Yang et al. (2025b)	Ziniu Wu, Cheng Liu, Jindong Zhang, Xinyun Li, Yewen Wang, Jimmy Xin, Lianmin Zhang, Eric Xing, Yuxin Lu, and Percy Liang. 2023. Autogen: Enabling next-generation multi-agent communication with language models. arXiv preprint arXiv:2309.07864.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
59	acl-2025	2025.acl-industry.30	Fu et al. (2025)	Yihan Chen, Dongkuan Zhang, Xiang Wang, Yifan Yang, and Heng Wang. 2023. Dare: ldirect parameter editing for adaptive mode reconfiguration. arXiv preprint arXiv:2310.09570.
60	acl-2025	2025.acl-long.1034	Zhou and Di Eugenio (2025)	Jane Doe and Wei Chen. 2023. Debiasing reasoning in language models: A quantitative study. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
61	acl-2025	2025.acl-long.1369	Chen et al. (2025e)	Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, and Wolfgang Macherey. 2024a. Scaling laws across model architectures: A comparative analysis of dense and mixture of experts models. arXiv preprint arXiv:2410.05661.
62	acl-2025	2025.acl-long.1409	Frydenlund (2025)	Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregresinterpreting gpt: the logit lenssive neural sequence modeling by iterative refinement. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.
63	acl-2025	2025.acl-long.1435	Jia et al. (2025a)	Mei Li et al. 2024. Securing tool use in llm agents: Challenges and strategies. arXiv preprint arXiv:2402.03014.
64	acl-2025	2025.acl-long.1439	Siingh et al. (2025)	Wei Xu, Shujie Liu, Minlie Huang, Si Wei, and Hang Li. 2018. A survey on hallucination in neural machine translation. arXiv preprint arXiv:1803.02577. Accessed: 2025-02-14.
65	acl-2025	2025.acl-long.1554	Klisura et al. (2025)	Dorottya Demeszky, Weijie Gong, and Diyi Yang. 2019. Analyzing bias and framing in news articles on artificial intelligence: The case of sentiment analysis. In Proceedings of the 2019 Annual Conference on Empirical Methods in Natural Language Processing, pages 1655–1661.
66	acl-2025	2025.acl-long.1580	Spangher et al. (2025b)	Minjun Chang et al. 2023. Self-reflection with generative agents. arXiv preprint arXiv:2311.09214.
67	acl-2025	2025.acl-long.248	Lee et al. (2025)	Wei Chen, Arjun Kumar, and Lin Yang. 2024. Distraction-based attack prompts: An effective jailbreaking method for llms. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).
68	acl-2025	2025.acl-long.340	Li et al. (2025c)	openai. 2024. Learning to reason with llm. Preprint, arXiv:2406.19949.
69	acl-2025	2025.acl-long.421	Zhu et al. (2025b)	Xu Sun, Xiaoya Zhang, Yicheng Feng, Shiyang Wang, Shuming Ma, Jiuding He, Zhixu Zhang, Yuxian Gu, Yi Xu, Hao Zhou, and Zhiyuan Liu. 2023. A systematic capability evaluation of large vision-language models. arXiv preprint arXiv:2305.16372.
70	acl-2025	2025.acl-long.491	Alsagheer et al. (2025)	Hao Zhong, Jieyu Tang, Tianyang Xu, et al. 2020. Does nlp benefit legal system? a case study of document representation. arXiv preprint arXiv:2005.01647.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
71	acl-2025	2025.acl-long.712	Kothapalli et al. (2025)	Max Nye, Anders Johan Andreassen, David A Abolafia, Fran¸cois Chollet, and Geoffrey Irving. 2021. Show Your Work: Scratchpads for intermediate computation. arXiv preprint arXiv:2112.00114.
72	acl-2025	2025.acl-long.979	Chang et al. (2025)	Xinyu Li, Rahul Gupta, and Emily Brown. 2024. Datacomplm: Benchmarking data quality in large language models. arXiv preprint arXiv:2402.23456.
73	acl-2025	2025.acl-srw.40	Wu et al. (2025b)	[first names omitted for brevity] Zhang and colleagues. 2024. Metagpt: Meta programming sota autonomous multi-agent cooperative llm workflows. arXiv preprint arXiv:2404.14496.
74	acl-2025	2025.acl-srw.56	Pulido et al. (2025)	Sabrina J. Mielke, Tal Linzen, and Jason Eisner. 2021. What kind of knowledge is captured by contextualized word representations? In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, pages 1250–1265.
75	acl-2025	2025.acl-srw.73	Sakunkoo and Sakunkoo (2025)	Christo Kirov, John Sylak-Glassman, Ryan Que, David Yarowsky, Jason Eisner, and Ryan Cotterell. 2018. Universal morphological inflection generation using a multilingual dataset. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pages 52–62.
76	acl-2025	2025.acl-srw.85	Dzhubaeva et al. (2025)	Y. Wang, H. Li, and X. Zhang. 2024. Consistency of personality traits in quantized role-playing dialogue agents. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 123–130.
77	acl-2025	2025.acl-srw.98	Kumar and Nolbaria (2025)	Raghuraj Sailesh Shah, Ankur Anand, Srini Chodisetti, Ankit Gupta, Raj Sanjay Patel, Sameer Patel, and Manish Gupta. 2022. FinservNLP: A library of financial shared tasks and benchmarks. In Proceedings of the Third Workshop on Economics and Natural Language Processing, pages 144–150. Association for Computational Linguistics.
78	acl-2025	2025.africanlp-1.1	Akpobi (2025)	Adewale Oladipo et al. 2023. Wura: A multilingual dataset of african languages. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.
79	acl-2025	2025.africanlp-1.15	Ismail et al. (2025)	J. Sohn and 1 others. 2024. Rationale-guided rag for medical question answering. arXiv preprint arXiv:2411.00300.
80	acl-2025	2025.africanlp-1.30	Pandya et al. (2025)	M. Martin et al. 2022. Swahbert: Enhancing swahili nlp with pretrained transformers. In Proc. EACL.
81	acl-2025	2025.argmining-1.30	El Baff et al. (2025)	Yiqiu Xu, Alessio Frosini, Mattia Vanni, Anisa Rula, and Roberto Navigli. 2024. Frase: Frame-based semantic enrichment for sparql query generation. arXiv preprint arXiv:2503.22144.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
82	acl-2025	2025.bea-1.66	Rezayi et al. (2025)	Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2024. Qwen2.5: The next generation of qwen large language models. arXiv preprint arXiv:2407.10671.
83	acl-2025	2025.bea-1.73	Nebhi et al. (2025)	Elizabeth Mayfield and Alan W. Black. 2020. Automated scoring of written essays with transformer models. In Proc. BEA at ACL 2020.
84	acl-2025	2025.bea-1.86	Jain and Rengarajan (2025)	Tianyu Zhang, Sayak Basu, Douwe Kiela, and Oriol Vinyals. 2024. Cascade-aware training and inference for more resource-efficient language models. arXiv preprint arXiv:2405.12345.
85	acl-2025	2025.bionlp-1.6	Kumar and Basu (2025)	Ilias Chalkidis, Michail Fergadiotis, Roger Stradling, Nikolaos Pappas, and Prodromos Malakasiotis. 2020. Transformer-based models for legal and biomedical document classification: A comparative study. In Proceedings of ACL.
86	acl-2025	2025.bionlp-share.22	Evgin et al. (2025)	Qingxiu Jia, Qipeng Xu, Weiting Yu, Yitong Duan, JianYun Nie, and Zhiyuan Liu. 2022. Alignscore: Evaluating factual consistency with contextual alignment. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).
87	acl-2025	2025.bionlp-share.33	Xu et al. (2025a)	Yang Liu, Ani Nenkova, and Kathleen McKeown. 2022. Incorporating controllability in neural abstractive summarization. In Proceedings of the 60th Annual Meeting of the ACL, pages 876–890. ACL.
88	acl-2025	2025.bionlp-share.34	Phasook et al. (2025)	Sahil Dhuliawala, Raghav Gupta, and Shashi Narayan. 2023. Chain-of-verification: Enhancing llm factuality with self-verification. In NAACL.
89	acl-2025	2025.findings-acl.1022	Kulkarni et al. (2025)	Chen et al. 2023a. Nl2fol: Natural language to firstorder logic. arXiv preprint arXiv:2304.09102.
90	acl-2025	2025.findings-acl.114	Gupta et al. (2025b)	Yao Du, Kaitao Qian, Sanmi Koyejo, Zheng Zheng, and Chao Zhang. 2023. Alpaca: A strong, replicable instruction-following model. arXiv preprint arXiv:2303.16199.
91	acl-2025	2025.findings-acl.1140	Cui et al. (2025b)	Xiaoyu Li, Wei Zhang, Jiahui Chen, and Hongwei Liu. 2024c. Iris: Benchmarking large language models for information retrieval tasks. arXiv preprint arXiv:2401.12345.
92	acl-2025	2025.findings-acl.1143	Das et al. (2025b)	Raphael Rafailov, Orion Redwood, et al. 2023. Direct preference optimization: You don’ t need rewards to finish rlhf. arXiv preprint arXiv:2305.11517. Preprint, arXiv:2305.11517.
93	acl-2025	2025.findings-acl.1156	Zafrir et al. (2025)	Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Xsum: A new dataset for abstractive summarization of news articles. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 701–711.

Continued on next page

	EVENT ID	PAPER ID	HALLUCITED PAPER	EXAMPLE OF HALLUCITATION
94	acl-2025	2025.findings-acl.1184	Garg et al. (2025)	William Berrios, Yonglong Tian, Chiraag Lala, Alec Parrish, and Trevor Darrell. 2023. Lens: A learnable embedding name space for multi-modal, multi-task learning. arXiv preprint arXiv:2306.11527.
95	acl-2025	2025.findings-acl.1208	Das et al. (2025a)	Rakesh Gupta, Susan Johnson, and Wei Li. 2023. Prompt abstraction: Leveraging language-image embeddings to scale creativity. arXiv preprint arXiv:2302.12345.
96	acl-2025	2025.findings-acl.1231	Zhou et al. (2025a)	Oron Anschel, Nir Baram, and Nahum Shimkin. 2018. Learning representations in a multi-agent environment. In International Conference on Learning Representations.
97	acl-2025	2025.findings-acl.1278	Bi et al. (2025)	DeepSeek-AI. 2024. Deepseek llm: Scaling opensource language models with longevity value. arXiv preprint arXiv:2401.02954.
98	acl-2025	2025.findings-acl.1279	Nahin et al. (2025)	W Lian, B Goodson, E Pentland, et al. 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. arXiv preprint arXiv:2305.11206.
99	acl-2025	2025.findings-acl.1286	Kumari et al. (2025)	Douwe Kiela et al. 2018. Efficient large-scale multimodal classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
100	acl-2025	2025.findings-acl.228	Maji et al. (2025)	Ramesh Patel et al. 2023. Evaluating cultural understanding in multilingual models. arXiv preprint arXiv:2302.12345.
101	acl-2025	2025.findings-acl.268	Tong and Wang (2025)	David Bamman. 2020. Booknlp: Natural language processing for long-form text. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 1–7. Association for Computational Linguistics.
102	acl-2025	2025.findings-acl.282	Sun et al. (2025)	Tianle Li, Wei-Lin Chiang, Lisa Dunlap Evan Frick, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From live data to high-quality benchmarks: The arena-hard pipeline. arXiv preprint, abs/2406.11939.
103	acl-2025	2025.findings-acl.563	Parfenova and Pfeffer (2025)	Alexander Fabbri et al. 2021a. Summac: Re-visiting nli-based models for consistency evaluation. Proceedings of EMNLP 2021.
104	acl-2025	2025.findings-acl.581	Yuan et al. (2025a)	Kevin Meng, Alex Andonian, David Bau, Yonatan Belinkov, and Fei-Fei Li. 2023. MEMIT: Massediting memory in a transformer. arXiv preprint arXiv:2302.04761.
105	acl-2025	2025.findings-acl.690	Atri et al. (2025)	Peter Hase and Mohit Bansal. 2023. Compositional edits in language models. EMNLP.
106	acl-2025	2025.findings-acl.813	Li et al. (2025g)	OpenAI. 2024. Gpt-4o: Scaling and performance improvements. arXiv preprint arXiv:2401.XXXXX.
107	acl-2025	2025.findings-acl.861	Abootorabi et al. (2025)	Gunnar A Sigurdsson, Gul Varol, Giovanni Maria Farinella, et al. 2018. Charadesego: A dataset for egocentric video understanding. arXiv preprint arXiv:1804.09626.

CONTINUED ON NEXT PAGE

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
108	acl-2025	2025.findings-acl.920	Li et al. (2025e)	Zihao Wang, Haowei Lin, Ruilin Yan, Xiangyu Wang, Jiaqi Li, Weiye Shi, Xiaojian Ma, Anji Liu, Yitao Liang, et al. 2024e. Optimizing inference-time reasoning in llms via retrieval-augmented reflection. arXiv preprint arXiv:2403.05313.
109	acl-2025	2025.gebnlp-1.7	Kononykhina et al. (2025)	Mingyuan Zhang, Jiayi Wang, and et al. 2024a. Retrieval-augmented icd coding: A two-stage system beats vanilla llms by 94 arXiv preprint arXiv:2401.12345.
110	acl-2025	2025.gem-1.6	Belz and Thomson (2025)	Anya Belz, Simon Mille, and Craig Thomson. 2025b. A taxonomy of quality criterion names and definitions for evaluating nlp systems in terms of standard comparable aspects of quality.
111	acl-2025	2025.iwslt-1.16	K. Roy et al. (2025)	Elizabeth Salesky, Ramon Sanabria, Alan W. Black, and Florian Metze. 2021. Exploring low-resource speechto-text translation across modalities. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, pages 1134–1145.
112	acl-2025	2025.knowllm-1.4	Li (2025b)	Wei Zhang, Li Chen, and Nicolas Christin. 2022. Dynamic graph learning for cryptocurrency fraud detection. In IEEE International Conference on Blockchain and Cryptocurrency, pages 1–9.
113	acl-2025	2025.l2m2-1.7	Lasy et al. (2025)	Nikhil Stoehr, Kha Pham Doan, and Louis-Philippe Morency. 2023. Localizing memorization in transformer language models. arXiv preprint arXiv:2310.01331.
114	acl-2025	2025.llmsec-1.2	Sorkhpour et al. (2025)	Robert Shelby, Carl Vondrick, and 1 others. 2023. Can llms be safely released? evaluating the impact of red teaming on language model behavior. arXiv preprint arXiv:2304.10685.
115	acl-2025	2025.magmar-1.9	Li and Ke (2025a)	Rohit Gupta, Ankit Kumar, and Rahul Singh. 2022. Crisismm: A framework for multimodal crisis response. arXiv preprint arXiv:2203.01234.
116	acl-2025	2025.nlp4pi-1.17	Ding et al. (2025)	Zhijing Jin, Zhiheng Lyu, Yiwen Ding, Mrinmaya Sachan, Kun Zhang, Rada Mihalcea, and Bernhard Schoelkopf. 2022. AI Scholars: A dataset for NLPinvolved causal inference.
117	acl-2025	2025.realm-1.33	Punjwani and Heck (2025)	Harsha Nori, Andrew He, Chenguang Wang, Po-Sen Huang, Yejin Yang, Hao-Tong Tung, Shashank Singh, Hossein Hosseini, Mark Hughes, Matei Zaharia, and 1 others. 2023. Can language models teach weaker agents? training reasoning teachers that can explain their actions. arXiv preprint arXiv:2311.10731.
118	acl-2025	2025.sdp-1.14	Mandic et al. (2025)	Kai Sun, Rui Zhang, and Xiang Ren. 2022. FS-NER: A few-shot joint entity and relation extraction framework. In Findings of the Association for Computational Linguistics: ACL 2022, pages 4292–4303, Dublin, Ireland. Association for Computational Linguistics.
119	acl-2025	2025.sdp-1.32	Carla and Uban (2025)	Xiang Deng, Han Zhang, Wenhao Yu, Clare Lee, and Mohit Bansal. 2024. Factscore 2.0: Dualencoder contrastive pre-training for factual consistency. Transactions of the Association for Computational Linguistics, 12:1–19.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
120	acl-2025	2025.semeval-1.106	Mondal and Sarkar (2025)	Xiao Tan and Yuan Jiang. 2021. Exploring idioms in question answering systems. In Proceedings of the 2021 Conference of the North American Chapter of the ACL.
121	acl-2025	2025.semeval-1.135	Vaidyanathan et al. (2025b)	Pranav Goel, Shrey Agarwal, and Amir Hussain. 2021. Emotion intensity regression using transformers and loss function optimization. In ACL 2021 Findings, pages 1432–1440.
122	acl-2025	2025.semeval-1.142	Kesanam et al. (2025)	Shailza Choudhary and Tanmoy Chakraborty. 2020. A hybrid feature extraction approach for multi-label emotion classification. IEEE Transactions on Affective Computing.
123	acl-2025	2025.semeval-1.151	Alberts et al. (2025)	Jian Wu, Junwei Xie, Bing Tang, Hongyan Yan, and Lijun Wang. 2022. Dense passage retrieval: A study of the representations of multiple sentences in a passage. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022).
124	acl-2025	2025.semeval-1.195	Matheny et al. (2025)	Silvio Ricardo Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. A computational model of nominal compound compositionality. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pages 2762–2773. Association for Computational Linguistics.
125	acl-2025	2025.semeval-1.231	Khatoon et al. (2025)	Vered Shwartz and Ido Dagan. 2018. A sequential neural model for multiword expression identification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT).
126	acl-2025	2025.semeval-1.243	Nadeem et al. (2025)	Frank A Acheampong, Henry Nunoo-Mensah, and Wei Chen. 2020. Transformer models for text-based emotion detection: A comparative analysis. arXiv preprint arXiv:2004.13704.
127	acl-2025	2025.semeval-1.250	Aryal and Pant (2025)	Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. arXiv preprint arXiv:1812.01813.
128	acl-2025	2025.semeval-1.266	Wan et al. (2025)	Danqi Chen, Pengcheng Xie, Shiyu Wang, et al. 2020. Tabfact: A large-scale dataset for table-based fact verification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).
129	acl-2025	2025.semeval-1.268	Chen (2025)	Min Li, Liangyou Li, and Philipp Koehn. 2022. Lowfrequency entity translation in neural machine translation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 3021–3033.
130	acl-2025	2025.semeval-1.27	S et al. (2025b)	S. Kumar et al. 2023. Sentilex: Enhancing emotion classification with lexicon-based methods. In ACL Workshop on Sentiment Analysis.
131	acl-2025	2025.semeval-1.278	Huang and Cui (2025)	Linhao Zhang, Shuai Wang, and Bing Liu. 2018. Learning to detect multi-label emotions in tweets with label co-occurrence and dependency. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3299–3309.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
132	acl-2025	2025.semeval-1.37	Cui (2025)	Yuqing Zhang, Yansong Feng, Yinwen Zhang, and Shujie Liu. 2021. Attention-based multi-label emotion classification with label-wise attention mechanism. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021), pages 5943–5952.
133	acl-2025	2025.sicon-1.12	Tiwari (2025)	Yixuan Zhang, Leo Cheng, Lichang Zhang, Lu Liu, Jingcheng Li, Haoning Liu, and Z G Xu. 2024. Alert: A comprehensive safety benchmark suite for large language models. Preprint, arXiv:2402.12441.
134	acl-2025	2025.sigtyp-1.15	Li and Ke (2025b)	Kevin Zhang and 1 others. 2022. Code meets prose: Syntactic patterns in cryptocurrency whitepapers. In LAW@ACL. Manual analysis of 200 whitepapers showing SVO/SOV mixing.
135	acl-2025	2025.trl-1.16	Li (2025c)	Haoyang Ding, Zijian Liu, Xiao Chen, Lidong Bing, and Wai Lam. 2021. Financial table processing with pretrained language models. EMNLP.
136	acl-2025	2025.unlp-1.10	Kyslyi et al. (2025a)	Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, and Stephan Vogel. 2017. Arabic dialect identification for the dsl 2017 shared task. In Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial).
137	acl-2025	2025.unlp-1.12	Kyslyi et al. (2025b)	Firoj Alam, Shaden Shaar, Alex Nikolov, and 1 others. 2021. A survey on nlp for fake news detection. Computational Linguistics, 47(4):905–960.
138	acl-2025	2025.woah-1.21	Boudraa et al. (2025)	Hande Kartal, Dilek Hakkani-Tür, and Gokhan Tur. 2022. Span-based detection of biased statements in news articles. In Proceedings of the 2022 Conference on Computational Linguistics. COLING.
139	acl-2025	2025.woah-1.36	Schirmer et al. (2025)	Xing Jiang, Hugo Touvron, Ludovic Denoyer, Hervé Jégou, and Guillaume Lample. 2023. Mistral 7b: A small model that thinks big. arXiv preprint arXiv:2310.06825.
140	acl-2025	2025.xllm-1.25	Pham Hoang Le et al. (2025)	Micha Livne, Roberto Dessì, and Douwe Kiela. 2023. How to train a large language model to be a reliable information extraction system. Computing Research Repository, arXiv:2309.16396. Version 3.
141	acl-2025	2025.xllm-1.3	Li et al. (2025k)	Yujia Zhang, Xiaobo Liu, and Zhenhua Li. 2020. Graphbased hierarchical relationships for text representation. arXiv preprint arXiv:2007.12345.
142	emnlp-2025	2025.arabicnlp-main.20	Bouchekif et al. (2025)	Faiza Qamar, Seemab Latif, and Rabia Latif. 2024. A benchmark dataset with larger context for non-factoid question answering over islamic text. arXiv preprint arXiv:2409.09844.
143	emnlp-2025	2025.babylm-main.35	Fusco et al. (2025)	Leonie Weissweiler, Valentin Hofmann, David Mortensen, and Janet Pierrehumbert. 2023. Evaluating morphological generalization with wug tests. arXiv preprint. ArXiv:230x.xxxxx.
144	emnlp-2025	2025.babylm-main.36	Yoshida et al. (2025)	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Preprint, arXiv:1901.04513.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
145	emnlp-2025	2025.codi-1.4	Jeon and Strube (2025)	Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. BERT: A study of attention in BERT. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5717–5726, Hong Kong, China. Association for Computational Linguistics.
146	emnlp-2025	2025.codi-1.5	Ahmad and Kakudi (2025)	Yuan Lan, Chao Huang, Wayne Xin Zhao, and Jun Li. 2024. Cola: Collaborative role infused multiagent debate framework for stance detection. arXiv preprint arXiv:2403.01234.
147	emnlp-2025	2025.emnlp-demos.24	Jang et al. (2025)	Rong Chen, Siyun Zhang, Yiyi Zheng, Qiuhua Yu, and Chu-huai Wang. 2025. Enhancing treatment decision-making for low back pain: A novel framework integrating large language models with retrieval-augmented generation technology. arXiv preprint arXiv:2501.34567.
148	emnlp-2025	2025.emnlp-demos.27	Liu et al. (2025f)	Jihan Zhu, Zhizheng Lin, Jun Gao, Zhaoxuan Zhou, Yaodong Zhang, Zhaofan Liu, Ceyao Zheng, Cheng Li, Zhaokai Wang, Zili Wang, and 1 others. 2024. A survey of ai agent evaluation: A hundred unsolved problems and a one-stop open-source library. arXiv preprint arXiv:2406.09844.
149	emnlp-2025	2025.emnlp-demos.51	Chao et al. (2025)	Jinhao Liang, Yihang Chen, Haozheng Zhang, and 1 others. 2025. Surveyx: An end-to-end solution for automated survey generation. arXiv preprint arXiv:2501.12345.
150	emnlp-2025	2025.emnlp-industry.123	Khabiri et al. (2025)	Takeshi Kojima, Shinn Yao, Jinyi Zhao, Xiang Ren, et al. 2023. ToolLLM: Facilitating LLMs to master 16000+ real-world APIs. arXiv preprint arXiv:2312.11568.
151	emnlp-2025	2025.emnlp-industry.146	Zhang (2025)	Emmanouil Panousis and Tom Rainforth. 2022. A principled and unified framework for bayesian optimal experimental design. arXiv preprint arXiv:2202.04647.
152	emnlp-2025	2025.emnlp-industry.168	Ethiraj et al. (2025)	Zhen Li and 1 others. 2023. Towards general text embeddings with mixture of experts. Preprint, arXiv:2311.05723. https://arxiv.org/abs/2311.05723.
153	emnlp-2025	2025.emnlp-industry.182	Yu et al. (2025b)	Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. Adaptthink: Llm can learn when to think. https://arxiv.org/abs/2501.12570.
154	emnlp-2025	2025.emnlp-industry.187	Beyranvand et al. (2025)	Jun Xu, Fan Wang, and Shuo Chen. 2022. Automated question-answering for vehicle manuals using nlp and knowledge graphs. IEEE Transactions on Intelligent Transportation Systems.
155	emnlp-2025	2025.emnlp-industry.189	Chawla et al. (2025)	Yifan Liu, Jiaqi Wang, and Zhi Zhang. 2024. Bias in ai investment advice: An analysis of home bias in large language models. arXiv preprint arXiv:2209.04538.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
156	emnlp-2025	2025.emnlp-industry.69	Azeez et al. (2025)	Juho Kim et al. 2023. Does gpt-4 pass the bar exam? a case study on the performance of large language models on legal multiple choice questions. In Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics.
157	emnlp-2025	2025.emnlp-industry.70	Swamy et al. (2025)	Florian Schuster et al. 2023. Getml: Pareto-based feature learning for machine learning applications. arXiv preprint arXiv:XXXX.XXXXX.
158	emnlp-2025	2025.emnlp-industry.8	Bougie and Watanabe (2025)	Anonymous and 1 others. 2023b. Is llm a reliable reviewer? a comprehensive evaluation of llm on reviewrevision multiple-choice questions. ACL Anthology 2024.lrec-main.816.
159	emnlp-2025	2025.emnlp-industry.96	Joshi et al. (2025a)	Siva Reddy, Daniel Khashabi, Benjamin Lo, Chandra Bhagavatula, Maarten Sap, and Yejin Choi. 2023. Do large language models understand instructions? arXiv preprint arXiv:2301.01193.
160	emnlp-2025	2025.emnlp-main.1021	Pi et al. (2025)	Ziyang Xiong et al. 2024. Llava-critic: Visual instruction tuning with feedback. arXiv preprint arXiv:2402.12345.
161	emnlp-2025	2025.emnlp-main.1070	Zhang et al. (2025d)	Yiming Zhang and Zhouhui Lian. 2024a. Brush your text: Enhancing text rendering in diffusion models with attention map interventions. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).
162	emnlp-2025	2025.emnlp-main.1097	Zhang and Zhou (2025)	X. Li, P. Wang, and Y. Chen. 2020. Iterative transformer: Recurrent attention updates for sequence modeling. arXiv preprint arXiv:2010.02536.
163	emnlp-2025	2025.emnlp-main.1119	Zhang et al. (2025e)	Chenyan Yu, Lee Xiong, and Jamie Callan. 2022. A survey on conversational dense retrieval: Methods and challenges. arXiv preprint arXiv:2201.08452.
164	emnlp-2025	2025.emnlp-main.1167	Zhang et al. (2025b)	Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2024. Flare: Active retrieval augmentation for hallucination mitigation. arXiv preprint arXiv:2401.07019.
165	emnlp-2025	2025.emnlp-main.1190	Yi et al. (2025)	Bogdan Gliwa, Iwona Mochol, Michał Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue summary dataset. arXiv preprint arXiv:1911.12237.
166	emnlp-2025	2025.emnlp-main.1232	Nyandwi et al. (2025)	Wenxuan Zhang, Da Yin, Zhong Ding, and Radu Soricut. 2023. M3exam: A multilingual, multimodal exam-style qa benchmark. arXiv preprint arXiv:2306.05179.
167	emnlp-2025	2025.emnlp-main.1235	Rafiei Asl et al. (2025)	Mark Russinovich, Jingwen Cai, Yifan Hou, Yuxiao Jiang, Yiping Jiang, Jieyu Kang, Prithviraj Kang, Urvashi Khandelwal, Nitish Khurana, Pang Wei Koh, et al. 2024. Crescendo: Towards realistic redteaming of llms with language agents. arXiv preprint arXiv:2402.13249.
168	emnlp-2025	2025.emnlp-main.1337	Dumitru et al. (2025)	Jia Chen and Hao Xu. 2023. Parallel decoding with speculative sampling for large language models. arXiv preprint arXiv:2306.15478.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
169	emnlp-2025	2025.emnlp-main.1345	Park et al. (2025)	Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docttttquery. arXiv preprint arXiv:1904.08375.
170	emnlp-2025	2025.emnlp-main.138	Li et al. (2025b)	Weixi Tong and Tianyi Zhang. 2024. Codejudge: Evaluating code generation with large language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20032–20051.
171	emnlp-2025	2025.emnlp-main.1429	Bohdal et al. (2025)	Aliaksei Severyn, Eric Emil Malmi, Jonathan Stephen Mallinson, Sascha Rothe, and Sebastian Krause. 2021. Grammatical error correction using large multilingual language models. In ACL.
172	emnlp-2025	2025.emnlp-main.145	Borah et al. (2025)	Junnan Li, Ziqing Wang, Yilun Xu, et al. 2024. Alignment Degradation in LLMs: Measuring, Visualizing, and Mitigating Latent Drift. arXiv preprint arXiv:2401.04200.
173	emnlp-2025	2025.emnlp-main.1455	Kryklyvets et al. (2025)	S. Thoppilan and et al. 2023. Gemini: A unified multimodal large language model. arXiv preprint arXiv:2303.06601.
174	emnlp-2025	2025.emnlp-main.1457	Srivastava (2025)	Wei Xu, Yulia Tsvetkov, and Alan Black. 2022. AI for language learning: Conversational agents and personalized feedback. Transactions of the Association for Computational Linguistics (TACL), 10:1–15.
175	emnlp-2025	2025.emnlp-main.1466	Roh et al. (2025)	Amanpreet Singh, Yujia Wang, Yulia Tsvetkov, and Percy Liang. 2024. Global-mmlu: Evaluating cultural and linguistic biases in multilingual language understanding. arXiv preprint arXiv:2412.03304.
176	emnlp-2025	2025.emnlp-main.1521	Gallifant et al. (2025)	Trenton Bricken, Jonathan Marcus, Siddharth MishraSharma, Meg Tong, Ethan Perez, Mrinank Sharma, Kelley Rivoire, and Thomas Henighan. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. ArXiv, abs/2309.08600.
177	emnlp-2025	2025.emnlp-main.1538	Jin et al. (2025)	Liang Liu, Marc Brockschmidt, Vivek Murali, and et al. 2024. The meta llm compiler: A suite of opensource models for code optimization. arXiv preprint arXiv:2407.02524.
178	emnlp-2025	2025.emnlp-main.1556	Dhanraj and Eliasmith (2025)	Jianfeng Wang, Fei Huang, and 1 others. 2023. Efficient fine-tuning of large language models with lora. arXiv preprint arXiv:2303.01234.
179	emnlp-2025	2025.emnlp-main.1558	Song et al. (2025a)	Mor Geva, Tal Schuster, Jonathan Berant, and Omer Levy. 2023. Token recycling: Making llms faster and more data-efficient. arXiv preprint arXiv:2310.02548.
180	emnlp-2025	2025.emnlp-main.1582	Liu et al. (2025e)	Sabine Ploux and Hong Ji. 2003. Semantic spaces: Modeling the distribution of word senses with unsupervised learning techniques. Computational Linguistics, 29(2):259–275.
181	emnlp-2025	2025.emnlp-main.1611	Yuan et al. (2025b)	Laura A Banarescu, Claire Bonial, Sheila Condon, Emily Faries, Jon Niekrasz, and Tim O' Connor. 2018. Amr for multi-document summarization. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 577–583.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
182	emnlp-2025	2025.emnlp-main.1629	Sadia et al. (2025)	Srivatsan Sundar, Dhruv Jain, Yuwei Zhang, and H. V. Jagadish. 2024. gtbls: Generating tables from text by learning table structures. In Proceedings of the 2024 Conference of the Association for Computational Linguistics.
183	emnlp-2025	2025.emnlp-main.164	Xiao et al. (2025)	Kai Jiang, Hao Liu, and Rodney Brooks. 2022. Trustperformance paradox in industrial hri: Empirical evidence from warehouse robotics. arXiv preprint arXiv:2208.14637.
184	emnlp-2025	2025.emnlp-main.1641	Kassem et al. (2025)	Baolin Peng and 1 others. 2024. Efficient model editing at scale. arXiv preprint arXiv:2401.05911.
185	emnlp-2025	2025.emnlp-main.1648	Mohammadi et al. (2025)	Jerry Lin, Roy Fox, Isa Dalai, Aditya Kusupati, Sewon Min, Luke Zettlemoyer, Hannaneh Hajishirzi, Jon Shlens, Sham Kakade, and Hossein Adeli. 2024. In-context learning creates task vectors. arXiv preprint arXiv:2402.16509.
186	emnlp-2025	2025.emnlp-main.1777	Gao et al. (2025)	Jack FitzGerald, Ryan Cotterell, Bahadir Avci, Sebastian Ruder, Graham Neubig, Melvin Johnson, and Orhan Firat. 2022. Massive: A 51-language multilingual dataset for task oriented semantic parsing. arXiv preprint arXiv:2204.08582.
187	emnlp-2025	2025.emnlp-main.1806	He et al. (2025b)	A. Kirstain and 1 others. 2023. Pick-a-pic: A dataset for evaluating the robustness of vision-language models. Preprint, arXiv:2305.01569.
188	emnlp-2025	2025.emnlp-main.192	Liu et al. (2025d)	Rishi Taori et al. 2023. Alpaca: A 175b-parameter model for instruction following. ArXiv preprint arXiv:2303.11366.
189	emnlp-2025	2025.emnlp-main.207	He et al. (2025a)	Shuang Chen, Yang Liu, and Erik Cambria. 2023. Knowledge-enhanced aspect-based sentiment analysis with external commonsense graphs. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), pages 13567–13579. Association for Computational Linguistics.
190	emnlp-2025	2025.emnlp-main.319	Li et al. (2025f)	Kevin Lin, Ben Tan, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2019b. Ropes: Reading comprehension over paragraphs. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).
191	emnlp-2025	2025.emnlp-main.348	Khandelwal et al. (2025)	Alex Mallen, Zexuan Zhang, Xinyu Liu, Yichong Wu, Yiming Zhang, and William Yang Wang. 2023b. When not to trust language models: Investigating effectiveness of parametric and retrieval-based knowledge. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1234–1245.
192	emnlp-2025	2025.emnlp-main.394	Xu et al. (2025c)	Todor Mihaylov and 1 others. 2018. Openbookqa: Factual knowledge assessment in question answering. EMNLP.
193	emnlp-2025	2025.emnlp-main.397	Zheng et al. (2025a)	Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021b. Efficient routing transformers: Dynamic token interaction models for natural language processing. arXiv preprint arXiv:2003.05997.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
194	emnlp-2025	2025.emnlp-main.468	Tang et al. (2025b)	Justin Fu, Kelvin Zhang, Utkarsh Sanyal, Lantao Yu, Collin Moses, Fan Yang, Stefano Ermon, and Zhibin Zhao. 2023. Driving with reasoning: Reinforcement learning with generalist language models for interpretable policies. arXiv preprint arXiv:2303.00745.
195	emnlp-2025	2025.emnlp-main.482	Tanjim et al. (2025)	Meta AI. 2024. Llama 3.2: Advancing open-weight language models. ArXiv preprint arXiv:2408.00001.
196	emnlp-2025	2025.emnlp-main.500	Wang et al. (2025g)	Michael Chen, Jan Tworek, Heewoo Jun, Qingqing Yuan, Tom Jacobsen, Tijana Radulovic, Jungo Kasai, Guy Ephrath, and Nando De Freitas. 2020. Just ask for general knowledge: Training universal retrieval models for open-domain question answering. arXiv preprint arXiv:2008.03886.
197	emnlp-2025	2025.emnlp-main.501	Wang et al. (2025f)	Xinyu Fan, Li Zheng, Bing Yu, Yifan Liu, Yichi Zhang, Yu Zhang, Hao Tang, Yikai Wang, Weinan Chen, Yizhou Wang, and 1 others. 2023. Urban identityaware geo-reasoning in street views. arXiv preprint arXiv:2311.16456.
198	emnlp-2025	2025.emnlp-main.524	Saba and Skiena (2025)	Markus Freitag, Yaser Al-Onaizan, Shuo Sun Ma, and 1 others. 2022a. High-quality low-resource machine translation: A new benchmark. In Findings of EMNLP.
199	emnlp-2025	2025.emnlp-main.527	Yu et al. (2025a)	Wei Zhang, Ming Li, Hao Wang, and Yang Liu. 2024b. Deepseekmath: Scalable math pre-training and group relative policy optimization for mathematical reasoning. arXiv preprint arXiv:2402.03300.
200	emnlp-2025	2025.emnlp-main.528	Chen et al. (2025f)	Salah Lhoussaine, Georgios Tziantzioulis, Michael B. Sullivan, Vilas Sridharan, Nathan DeBardeleben, Christian Engelmann, and Simranjit Singh. 2024. A first look at bfloat16 soft-error resilience in large language models. Preprint, arXiv:2412.07192.
201	emnlp-2025	2025.emnlp-main.548	Wang et al. (2025a)	Harsha Nori, Allison Webster, Matthew McInnis, et al. 2023b. Medprompt: Large language models are reasoning engines with structured prompts. Preprint, arXiv:2307.13880.
202	emnlp-2025	2025.emnlp-main.562	Shihab et al. (2025)	Szymon Tworkowski, Konrad Przybysz, Tomasz Korbak, Pedro Rodriguez, Wojciech Rozemłyn, Kamil Rakowski, Maciej Grzelak, Albert Webson, Maciej Szafraniec, Robin Sorsch, and 1 others. 2024. Mambaformer: Efficient language modeling with selective state spaces and attention. arXiv preprint arXiv:2312.00752.
203	emnlp-2025	2025.emnlp-main.584	Wang et al. (2025d)	DeepSeek. 2025. Deepseek r1: Towards deep reinforcement learning for language models. arXiv preprint arXiv:2501.12948.
204	emnlp-2025	2025.emnlp-main.614	Lv et al. (2025)	Yifan Xu, Xiaoyu Liu, Baolin Peng, and et al. 2024. Evaluating the instruction-following robustness of large language models. In Proceedings of EMNLP 2024.
205	emnlp-2025	2025.emnlp-main.619	Tian et al. (2025)	Jing Qian and et al. 2018. Neural network-based fake news detection: Learning to identify deceptive content. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
206	emnlp-2025	2025.emnlp-main.637	Kang et al. (2025)	Hang Yu, Tianle Li, and et al. 2025. Dapo: Data-aware policy optimization for open-source llm alignment. arXiv preprint arXiv:2503.04111.
207	emnlp-2025	2025.emnlp-main.789	Kim et al. (2025a)	Mor Geva, Daniel Khashabi, Elad Segal, Jonathan Berant, and Ido Dagan. 2021. Aristotle: A dataset for focused logical reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1–14, Punta Cana, Dominican Republic. Association for Computational Linguistics.
208	emnlp-2025	2025.emnlp-main.807	Vejendla (2025b)	Zexuan Chen, Zihui Li, Xuan Zhao, A-Long Zhou, Lichao Yu, Wei Zhang, and Ming Zhou. 2023. PRMoE: Post-routing mixture of experts. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
209	emnlp-2025	2025.emnlp-main.813	Zhang et al. (2025f)	Mark Finlayson and 1 others. 2023. Causal analysis of language model behavior with induced interventions. arXiv preprint arXiv:2301.12928.
210	emnlp-2025	2025.emnlp-main.830	Liu et al. (2025a)	Yixin Nie, Jing Wang, Mohit Bansal, and Kai-Wei Chang. 2020. Adversarial natural language inference. In Proceedings of ACL.
211	emnlp-2025	2025.emnlp-main.831	Liu et al. (2025b)	Paola Espinosa, Sarah Kadi, Stefanie Kamps, and et al. 2017. Gender differences in empathy-related processes: Exploring behavioral and neurophysiological correlates. Psychoneuroendocrinology, 85:34–43.
212	emnlp-2025	2025.emnlp-main.835	Luo et al. (2025a)	Xuezhe Li and Xiaodan Sun. 2020. Dice loss for dataimbalanced nlp tasks: Application to named entity recognition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4653–4661.
213	emnlp-2025	2025.emnlp-main.841	Rivlin-Angert and Mor-Lan (2025)	David Hebron, Avi Shmidman, and Moshe Koppel. 2023. Hero: A hebrew roberta model for hebrew nlp tasks. Preprint, arXiv:2309.12345.
214	emnlp-2025	2025.emnlp-main.848	Nafee et al. (2025)	Linyong Lu, Qingxiu Zhang, Xiang Li, and Jingjing Liu. 2022. Promptpg: Prompting for policy gradient training. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).
215	emnlp-2025	2025.emnlp-main.853	Mundada et al. (2025)	Lukasz Czajka, Malihe Alikhani, Patrick Verga, and Yonatan Belinkov. 2024b. Musictheory-bench. ArXiv preprint arXiv:2410.02084.
216	emnlp-2025	2025.emnlp-main.855	Jiang et al. (2025)	Victor Zhong, Chandra Bhagavatula, Ronan Le Bras, Yejin Choi, and Noah A Smith. 2022. Analytical reasoning of text: Unifying machine reading and logical reasoning. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 2307–2323.
217	emnlp-2025	2025.emnlp-main.858	Trifan et al. (2025)	Liang Pan, Lei Zhang, and Zhenzhong Yang. 2023. Can chatgpt do zero-shot dialogue state tracking? In arXiv preprint arXiv:2302.01180.
218	emnlp-2025	2025.emnlp-main.880	Wang et al. (2025c)	Weijie Chen, Yizhe Zhang, Qian Wu, and 1 others. 2024. Internvl: Scaling up vision-language pretraining with multimodal reinforcement learning. arXiv preprint arXiv:2402.00028.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
219	emnlp-2025	2025.emnlp-main.938	Chen et al. (2025a)	Xiaoxi Zhong et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploration with llms. arXiv preprint arXiv:2309.07864.
220	emnlp-2025	2025.emnlp-main.952	Zhou et al. (2025b)	Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with warmup-stable-decay learning rate scheduler. arXiv preprint arXiv:2404.06395.
221	emnlp-2025	2025.emnlp-main.986	Bhendawade et al. (2025)	Yi Tay, Dara Bahri, Donald Metzler, et al. 2022. Scale efficiently: Insights from training and scaling large language models. arXiv preprint arXiv:2210.03863.
222	emnlp-2025	2025.findings-emnlp.1052	Kim et al. (2025c)	H. Yang and K. Li. 2023. Boostaug: Hybrid instance filtering framework for boosting text augmentation. In Findings of the Association for Computational Linguistics: ACL 2023.
223	emnlp-2025	2025.findings-emnlp.1074	Aastik et al. (2025)	Xianzhi Li, Zihang Dai, and 1 others. 2022. Branchtuning: Efficient fine-tuning of pre-trained vision models via branching. arXiv preprint arXiv:2202.06924.
224	emnlp-2025	2025.findings-emnlp.1082	Maskey et al. (2025)	Rodrigo Gomez et al. 2018. Ciphergan: Unsupervised cipher cracking using gans. arXiv preprint arXiv:1801.04883.
225	emnlp-2025	2025.findings-emnlp.1100	Yang et al. (2025a)	Mo Yu, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Jing Li, and Jie Zhou. 2022. Character understanding in movies: A benchmark for movie character analysis. arXiv preprint arXiv:2211.04684.
226	emnlp-2025	2025.findings-emnlp.1118	Wang et al. (2025b)	Zhen Chiang, Noah Shinn, Siyuan Zhuang, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. Preprint, arXiv:2304.05335.
227	emnlp-2025	2025.findings-emnlp.1137	Elgaar and Amiri (2025)	Xiaofei Lu. 2020. Automatic analysis of syntactic complexity in second language writing. ArXiv preprint, abs/2005.02013.
228	emnlp-2025	2025.findings-emnlp.1144	Bn et al. (2025)	Suyeon Lee, Sunghwan Kim, et al. 2024b. Counselingeval: Towards evaluating the quality of llmbased psychological counseling. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).
229	emnlp-2025	2025.findings-emnlp.118	Zeng et al. (2025)	X. Du et al. 2022a. Grit1: A grammar error correction dataset for llm evaluation. In Proceedings of EMNLP 2022. X. Du et al. 2022b. Grit2: Extending grammar error correction for multilingual llms. In Findings of EMNLP 2022.
230	emnlp-2025	2025.findings-emnlp.1196	Zheng et al. (2025c)	OpenAI. 2023a. Gpt-4: Openai’ s generative pre-trained transformer 4 model. Preprint, arXiv:arXiv:2301.00000.
231	emnlp-2025	2025.findings-emnlp.1286	Kim et al. (2025f)	Baptiste Roziere, Gautier Izacard, Jan Leike Botha, and Others. 2023. Code LLaMA: Large language models for code. Preprint, arXiv:2308.08545.

Continued on next page

	EVENT ID	PAPER ID	HALLUCITED PAPER	EXAMPLE OF HALLUCITATION
232	emnlp-2025	2025.findings-emnlp.1343	Saeed et al. (2025)	Damai Bi, Kunhao Lv, Xiang Ji, Yining Wang, Zecheng Lyu, Zihan Du, et al. 2024. Deepseek: Advancing ai through enhanced language models. arXiv preprint arXiv:2401.14196.
233	emnlp-2025	2025.findings-emnlp.1344	Kwak et al. (2025)	Aditya Maharana and 1 others. 2024. Locomo: Long context modeling benchmark for conversational agents. arXiv preprint arXiv:2402.13968.
234	emnlp-2025	2025.findings-emnlp.1359	Akra et al. (2025)	Ling Chen, Ming Zhao, and 1 others. 2024. Efficient active learning for biomedical image segmentation with minimal annotations. arXiv preprint arXiv:2405.01701.
235	emnlp-2025	2025.findings-emnlp.1384	Ji and Lu (2025)	Yiming Li, Xia Chen, Liang Zhou, and Zheng Wang. 2024b. Hindsight: Selective reflection for robust reasoning. arXiv preprint arXiv:2406.12050.
236	emnlp-2025	2025.findings-emnlp.175	Zhu et al. (2025d)	Yujia Hong, Junyang Wu, Fan Zhang, and Shuo Zhang. 2024b. Adaptive agents with code and memory for solving math word problems. arXiv preprint arXiv:2403.01290.
237	emnlp-2025	2025.findings-emnlp.176	Saad-Falcon et al. (2025)	Dahyun Koo, Yejin Choi, and Eunsol Choi. 2023. Cognitive Biases in Large Language Models as Evaluators. ArXiv Preprint arXiv:2312.05441.
238	emnlp-2025	2025.findings-emnlp.212	Rafiuddin and Khan (2025b)	Tim Dettmers, Yannic Kilcher, Henry Minsky, Anna McDowell, Neha Nangia, Andreas Vlachos, and the Microsoft Phi Team. 2025. Phi-4-mini: Compact yet powerful multimodal models. arXiv preprint arXiv:2503.01743.
239	emnlp-2025	2025.findings-emnlp.281	Zhang et al. (2025c)	Yinlin Liu, Pengcheng Yin, and Graham Neubig. 2019. Conala: The code/natural language challenge. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).
240	emnlp-2025	2025.findings-emnlp.322	Sheng et al. (2025)	Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, et al. 2023. Moss: Training conversational language models from synthetic data. arXiv preprint arXiv:2307.15020, 7:3.
241	emnlp-2025	2025.findings-emnlp.331	Fan et al. (2025)	Yuyang Huang and Clark Glymour. 2016. A survey of causal discovery and causal inference. arXiv preprint.
242	emnlp-2025	2025.findings-emnlp.354	LeVi et al. (2025)	Q. Zhang and 1 others. 2025. Gradient-free adversarial attacks on llms: Transferability and optimization. arXiv:2502.01567v1 [cs.CL]. This paper primarily focuses on gradient-free attacks on LLMs but includes gradient-based attacks, with transferability in black-box settings using surrogate models. It does not emphasize jailbreaking or Greedy Coordinate Gradient (GCG), making it less aligned with the text' s focus. Wang et al. (2025, arXiv:2503.02219v1 [cs.CL]) on gradient-based jailbreaking with GCG is a more precise alternative.

CONTINUED ON NEXT PAGE

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
243	emnlp-2025	2025.findings-emnlp.379	Li et al. (2025i)	Yinhan Tang, Xavier Garcia, Alessandro Raganato, Vishrav Chaudhary, and Jiatao Gu. 2022. An efficient multilingual byte-to-byte model for sequenceto-sequence tasks. In Findings of the Association for Computational Linguistics: EMNLP 2022.
244	emnlp-2025	2025.findings-emnlp.388	Chakraborty et al. (2025)	Pengfei Liu, Weizhe Yuan, Jinlan Fu, et al. 2021. Pretrain prompt tune: Towards generalizing to unseen tasks. In Proceedings of ACL-IJCNLP.
245	emnlp-2025	2025.findings-emnlp.391	Luo et al. (2025b)	Wenbin Xiong, Zhan Liu, Yihan Yang, Ge Yu, Ge Zhang, and Tian Song. 2023. A survey on hyperrelational knowledge graphs: From data models to applications. arXiv preprint arXiv:2301.03869.
246	emnlp-2025	2025.findings-emnlp.435	Nguyen et al. (2025a)	Yi Cheng, Yilun Wu, Adrian Weller, Matt J. Kusner, and Pushmeet Kohli. 2024. Integrative decoding: Improving factuality in large language model generation via implicit self-consistency. In International Conference on Learning Representations.
247	emnlp-2025	2025.findings-emnlp.437	Cai et al. (2025)	Wenhui Wang, Li Dong, Hao Cheng, Furu Wei, and Ming Zhou. 2022. Minilmv2: Multi-task pre-training for multi-task all-purpose text representations. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2907–2918.
248	emnlp-2025	2025.findings-emnlp.474	Ghosh et al. (2025)	Yankai Lin, Jiapeng Zhou, Yiming Shen, Wenxuan Zhou, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Xcsqa: A benchmark for cross-lingual conversational question answering. In EMNLP.
249	emnlp-2025	2025.findings-emnlp.517	Baali et al. (2025)	Jue Gao, Tri M Dang, Michael L Seltzer, and Rif A Saurous. 2022. Conversasynth: Exploring the landscape of synthetic conversations for audio understanding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6543–6558.
250	emnlp-2025	2025.findings-emnlp.523	Mor-Lan et al. (2025)	David Hebron, Avi Shmidman, and Moshe Koppel. 2023. Hero: A hebrew roberta model for hebrew nlp tasks. Preprint, arXiv:2309.12345.
251	emnlp-2025	2025.findings-emnlp.534	Liu and Roth (2025)	Ellie Pavlick and Joel Tetreault. 2016. Semantically motivated future directions in linguistic ambiguity detection. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).
252	emnlp-2025	2025.findings-emnlp.645	Zhou et al. (2025d)	Sho Kurita, Jonas Pfeiffer, Ivan Vulic, Edoardo Ponti, ´ and Anna Korhonen. 2020. A weighted approach to unsupervised multilingual transformer fine-tuning. In Proc. of AACL-IJCNLP.
253	emnlp-2025	2025.findings-emnlp.657	Bhattacharya et al. (2025)	Yifei Li, Zhou Zhao, Xiaohan Yu, and Deng Cai. 2023. Multimodal uncertainty estimation for deep learning models. arXiv preprint arXiv:2304.02637.
254	emnlp-2025	2025.findings-emnlp.709	Zhang et al. (2025a)	E. Zaken, Y. Goldberg, and S. Ravfogel. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformers. Transactions of the Association for Computational Linguistics (TACL), 10:1–16.

Continued on next page

	EVENT ID	PAPER ID	HALLUCITED PAPER	EXAMPLE OF HALLUCITATION
255	emnlp-2025	2025.findings-emnlp.751	Xiang et al. (2025)	VN Ioannidis, X Song, S Manchanda, M Li, X Pan, D Zheng, X Ning, X Zeng, and G Karypis. 2022. Drkg-drug repurposing knowledge graph for covid19. 2020. arXiv preprint arXiv:2010.09600, pages 1–13.
256	emnlp-2025	2025.findings-emnlp.834	Jalori et al. (2025)	Wendi Zhou, Xiao Li, Lin Geng Foo, Yitan Wang, Harold Soh, Caiming Xiong, and Yoonkey Kim. 2024. TEMPO: Temporal representation prompting for large language models in time-series forecasting. arXiv preprint arXiv:2405.18384. Anticipated for NeurIPS 2024. Preprint, arXiv:2405.18384.
257	emnlp-2025	2025.findings-emnlp.858	Ge et al. (2025)	Tianrui Han, Ruisheng Zhang, and Siyu Li. 2024. Analyzing and mitigating object hallucination in large vision-language models from perspective of training data bias. arXiv preprint arXiv:2402.16011.
258	emnlp-2025	2025.findings-emnlp.863	Kellert et al. (2025)	Shruti Rijhwani, Lawrence Wolf-Sonkin, Victor Kuperman, Timothy Baldwin, and Thamar Solorio. 2017. Analyzing code-switched social media text. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.
259	emnlp-2025	2025.findings-emnlp.864	Jia et al. (2025b)	Jiawei Chen, Yuanhang He, Yada Zhang, Jiachen Ji, and Jie Tang. 2024. Generate-on-graph: Treat llm as both agent and knowledge graph for incomplete kgqa. arXiv preprint arXiv:2404.14741.
260	emnlp-2025	2025.findings-emnlp.866	Li et al. (2025a)	Eric Mitchell, Roberta Raileanu, Colin Raffel, John Levine, Yulia Tsvetkov, and Christopher D Manning. 2023. Ieval: An instruction following benchmark. arXiv preprint arXiv:2310.07724.
261	emnlp-2025	2025.findings-emnlp.874	Kim et al. (2025e)	Zhiwei Chen, Yichao Lu, and Mingjun Zheng. 2020. Dkt+: Enhanced deep knowledge tracing with regularization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1234–1244.
262	emnlp-2025	2025.findings-emnlp.902	Xu et al. (2025d)	Zixuan Wang, Xiaocheng Li, and Yang Liu. 2023. Posterior prompt tuning: Toward faithful and calibrated llms. In Empirical Methods in Natural Language Processing (EMNLP).
263	emnlp-2025	2025.findings-emnlp.914	Li et al. (2025h)	RangRang Ge, ShiYe Song, Zhaorui Liu, Wei Liu, Yuesheng Wang, Dongling Wang, Bofang Zhou, Zhicheng Dou, and Ji-Rong Wen. 2023. A survey on KV cache compression for large language models. arXiv preprint arXiv:2312.10546.
264	emnlp-2025	2025.findings-emnlp.932	Jakob et al. (2025)	Shaina Ashraf, Isabel Bezzaoui, Ionut Andone, Alexander Markowetz, Jonas Fegert, and Lucie Flek. 2024. Defakts: A fine-grained dataset for analyzing disinformation in german media. In Proceedings of The 2024 Joint International Conference, on Computational Linguistics, Language Resources and Evaluation, Torino, Italia. European Language Resources Association.

CONTINUED ON NEXT PAGE

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
265	emnlp-2025	2025.findings-emnlp.943	Akhondzadeh et al. (2025)	Michael Boratko, Harsh Padigela, Deepak Mikkilineni, Pavan Yuvraj, Rajarshi Das, Andrew McCallum, Mihai Chang, Achille Fokoue, Pavan Kapanipathi, Nicholas Mattei, et al. 2018. Arc: A machine reading comprehension dataset for reasoning over science text. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1414–1423.
266	emnlp-2025	2025.findings-emnlp.954	Son et al. (2025)	Jack Hessel, Jingkang Zhao, Ranjay Krishna, Angel Chang, and Yonatan Bisk. 2022. Abductionrules: Leveraging commonsense knowledge and probabilistic reasoning for visual abduction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1451–1463.
267	emnlp-2025	2025.findings-emnlp.960	Kim et al. (2025b)	Xudong Liu, Paul Röttger, Johannes Welbl, Yonatan Belinkov, Hila Gonen, Eric Wallace, Samuel R. Bowman, Ryan Cotterell, and Noah A. Smith. 2023a. Harmbench: Measuring the propensity of language models to produce harmful content. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 11033–11051.
268	emnlp-2025	2025.findings-emnlp.967	Eyal et al. (2025)	Jackson Petty, Sjoerd van Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2024. The impact of depth on compositional generalization in transformer-based neural networks. arXiv preprint arXiv:2310.19956.
269	emnlp-2025	2025.findings-emnlp.97	Yin et al. (2025)	Alice Oh, Kalpesh Krishna, Eric Wallace, Yichong Zhao, Patrick Lewis, and Antoine Bosselut. 2023. Instructir: Making dense retrievers follow instructions. Preprint, arXiv:2305.14252.
270	emnlp-2025	2025.findings-emnlp.973	Colakoglu et al. (2025)	Jian Zhong et al. 2020. Docextractor: An end-to-end system for information extraction from forms and receipts. arXiv preprint arXiv:2012.04573.
271	emnlp-2025	2025.finnlp-2.3	George et al. (2025)	Peter Henderson, Koustuv Sinha, Nicolas AngelardGontier, Nan Rosemary Ke, Geneviève Fried, Ryan Lowe, and Joelle Pineau. 2023. Foundation models for legal reasoning. arXiv preprint arXiv:2307.03557.
272	emnlp-2025	2025.finnlp-2.8	Nitarach et al. (2025)	Xuezhi Ma, Yan Zhou, Yining Wang, and 1 others. 2023. Financialqa: A reasoning benchmark for financial question answering. arXiv preprint arXiv:2302.07304.
273	emnlp-2025	2025.hcinlp-1.8	Wang et al. (2025e)	OpenAI. 2023. Gpt-4: A large-scale multimodal model. arXiv preprint arXiv:2303.08774.
274	emnlp-2025	2025.mathnlp-main.13	Lu et al. (2025)	Zhe Chen, Weiyun Zhang, Wen Wang, Yiliang Liu, Zhaoyang Zhang, Jian Wang, Jie Luo, Yu Qiao, and Wenhai Wang. 2024. Internvl 1.5: A general visionlanguage model. arXiv preprint arXiv:2404.16821.
275	emnlp-2025	2025.mathnlp-main.5	Fatima (2025)	Jingwen Xin, Zhengying Liu, Yifan Luo, and 1 others. 2025. Deepseek-prover-v2: Scaling natural-language graph-based test time compute for automated theorem proving. arXiv preprint arXiv:2503.11657.

Continued on next page

	EVENT ID	PAPER ID	HALLUCITED PAPER	EXAMPLE OF HALLUCITATION
276	emnlp-2025	2025.mathnlp-main.6	Khrulev (2025)	Zhen Yuan, Yifan Zhang, Jing Liu, Yuxiang Wang, Jie Zhang, Hanwang Liu, and Tat-Seng Chua. 2024. Fermat: A benchmark for evaluating vlm’ s ability in factual error correction of handwritten math solutions. arXiv preprint arXiv:2405.10100.
277	emnlp-2025	2025.mathnlp-main.9	Li (2025a)	Qwen Team. 2024. Qwen2.5-math: Scaling reasoning in mathematical domains. arXiv preprint arXiv:2409.XXXXX. Model card and technical report.
278	emnlp-2025	2025.mrl-main.12	Moon et al. (2025)	Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. CrossRAG: Cross-lingual retrieval-augmented generation for knowledge-intensive tasks. arXiv preprint arXiv:2504.03616.
279	emnlp-2025	2025.mrl-main.2	Yuan et al. (2025c)	Laura A Banarescu, Claire Bonial, Sheila Condon, Emily Faries, Jon Niekrasz, and Tim O’ Connor. 2018. Amr for multi-document summarization. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 577–583.
280	emnlp-2025	2025.mrl-main.20	Rachamalla et al. (2025)	NVIDIA. 2025. Llama-nemotron-post-training dataset: A comprehensive collection of instruction tuning and alignment data. arXiv preprint arXiv:2505.00949.
281	emnlp-2025	2025.nllp-1.11	Shukla et al. (2025)	Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2022b. Inlegalbert: Pre trained language models for indian legal texts. arXiv preprint arXiv:2209.06049.
282	emnlp-2025	2025.nllp-1.32	Chheda et al. (2025a)	Shivansh Nigam, Sarvesh Dubey, Ayush Agarwal, Dhananjay Kumar, and Saket Maheshwary. 2025. Legalseg: Unlocking the structure of indian legal documents. arXiv preprint arXiv:2502.05836.
283	emnlp-2025	2025.nllp-1.5	Xia et al. (2025)	Ghita Houir Alami and 1 others. 2024. Legalbench-rag: A benchmark for retrieval-augmented systems in the legal domain. Preprint, arXiv:2408.10343.
284	emnlp-2025	2025.nlperspectives-1.11	Zhang and Jaitly (2025)	Barbara Plank. 2022. Human label variation: Challenges and opportunities. Computational Linguistics, 48(4):999–1015.
285	emnlp-2025	2025.starsem-1.24	Kadam and Ferraro (2025)	esse Dunietz, Sam Thomson, Chris Dyer, and Noah A. Smith. 2020. An interpretable, lexicalized model for implicit event causality. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1703–1713.
286	emnlp-2025	2025.starsem-1.29	Gasan and Pais (2025)	Jinhyuk Kim, Wonjin Kim, Jinhyuk Lee, Joongbo Lee, Kyunghyun Lee, Sunghwan Yoon, and Jaewoo Kang. 2024. Meerkat: A medical reasoning benchmark for large language models. arXiv preprint arXiv:2402.00000.
287	emnlp-2025	2025.starsem-1.3	Rai et al. (2025)	Harish Madabushi and 1 others. 2024. Frame-based embeddings for coherent question answering. In Proceedings of EACL.

CONTINUED ON NEXT PAGE

	EVENT ID	PAPER ID	HALLUCITED PAPER	EXAMPLE OF HALLUCITATION
288	emnlp-2025	2025.starsem-1.33	Bailleux et al. (2025)	Sumedha Bhan, Aaditya Prabhu, Huaxiu Ma, and Zachary C. Lipton. 2025. Complete textual concept bottleneck models: Addressing concept completeness and classification leakage. arXiv preprint arXiv:2502.12345.
289	emnlp-2025	2025.tsar-1.16	Hayakawa et al. (2025)	Ben Dodd, Betty van Aken, Paul Röttger, and Isabelle Augenstein. 2021. AUTORANK: A systematic approach to benchmark and compare machine learning models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 170–185, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics
290	emnlp-2025	2025.tsar-1.7	Romero et al. (2025)	Mozilla NLP Team. 2023. Mozilla translations: Opensource neural translation in the browser. In Proceedings of Machine Translation Summit.
291	emnlp-2025	2025.uncertainlp-main.24	Dossou and Aïdasso (2025)	Yi Zhang, Jialin Li, and Danqi Chen. 2023. Coannotating: Human-ai collaborative annotation via uncertainty-guided task allocation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
292	emnlp-2025	2025.winlp-main.12	Vinjamuri and Sun (2025)	Junxian Li, Xuezhe Ma, and Eduard Hovy. 2019. Dependency parsing with partial bi-affine attention. In Proceedings of NAACL-HLT.
293	emnlp-2025	2025.winlp-main.19	Kolli et al. (2025)	Yujia Tan, Wenpeng Zhang, Xiang Ren, and Qiji Chen. 2023b. Oe-fact: Open-domain explanation-enhanced fact-checking with large language models. In Findings of the Association for Computational Linguistics: EMNLP 2023.
294	emnlp-2025	2025.winlp-main.35	P and Mahalingam (2025)	Rahul Aralikatte, Neelamadhav Gantayat, Naveen Panwar, Anush Sankaran, and Senthil Mani. 2018. Sanskrit sandhi splitting using a double decoder rnn. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4909–4914. Association for Computational Linguistics.
295	emnlp-2025	2025.wmt-1.105	Joshi et al. (2025b)	Ahmed, T., Hasan, M. K., Hoque, M. T., & Sultana, N. (2023). A comparative study on subword segmentation strategies for low-resource neural machine translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)* (pp. 912–920). Association for Computational Linguistics. https://aclanthology.org/2023.wmt1.87
296	emnlp-2025	2025.wmt-1.106	Zhou et al. (2025c)	Anoop Kunchukuttan and 1 others. 2023. Indictrans2: Towards high-quality and low-resource machine translation for indic languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
297	emnlp-2025	2025.wmt-1.108	Zhu et al. (2025c)	Martin Vechev Luca Beurer-Kellner, Marc Fischer. 2024. Guiding llms the right way: Fast, noninvasive constrained decoding. arXiv.
298	emnlp-2025	2025.wmt-1.15	Bell et al. (2025)	Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting lexical alignment for cross language textual entailment. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 33–40, Trento, Italy. Association for Computational Linguistics.

CONTINUED ON NEXT PAGE

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
299	emnlp-2025	2025.wmt-1.43	Pang et al. (2025)	Aaron Chiang and et al. 2023. Autoreviewer: Enabling model self-review for dataset quality control. arXiv preprint arXiv:2303.14112.
300	emnlp-2025	2025.wmt-1.5	Kim (2025)	Zheng Jiang, Yang Yu, Yang Feng, Bing Qin, and Ting Liu. 2022. Blonde: An automatic evaluation metric for document-level natural language generation. In Proceedings of NAACL, pages 1679–1698.
301	emnlp-2025	2025.wmt-1.95	Acharya et al. (2025)	Gowtham Ramesh, Vishrav Chaudhary, Divyanshu Kakwani, Sai Praneeth Golla, Abhishek Philip, et al. 2023. Indictrans2: Towards high-quality and efficient multilingual translation for indic languages. arXiv preprint arXiv:2304.09105.
				Concluded

Table 9: List of hallucinated papers, i.e., HalluCited Papers, and their corresponding HalluCitations from AACL 2025 and EACL 2026.

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
1	aacl-2025	2025.banglalp-1.17	Tomal et al. (2025)	Wei Zhao, Li Chen, and Jin-seo Park. 2024. The new era of foundation models: A survey on pre-training, finetuning, and adaptation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pages 1–18.
2	aacl-2025	2025.banglalp-1.20	Chowdhury and Ferdous (2025)	R. Zhang and 1 others. 2024. Cross-lingual crosstemporal summarization: Dataset creation, modeling, and evaluation. Computational Linguistics, 50(3):1001–1023.
3	aacl-2025	2025.banglalp-1.23	Khisa et al. (2025)	Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: A reproduction of the llama training dataset. arXiv preprint arXiv:2411.12372.
4	aacl-2025	2025.banglalp-1.31	Raihan et al. (2025a)	Tianyi Wang, Yang Ye, Panupong Pasupat, Aohan Wan, Grant Friedman, Jiacheng Tu, Maya Schaar, Jason Wei, Suriya Gunasekar, Matthew Richardson, et al. 2023. Babelcode: Llm as a polyglot programmer. arXiv preprint arXiv:2303.03845.
5	aacl-2025	2025.banglalp-1.55	Islam et al. (2025)	Abhishek Bhattacharjee, Shammur Aktar, Md. Saidul Islam Zishan, Sudipta Ghosh, Md. Mahadi Hasan, M. Sohel Rahman, Mohammad Ruhul Alam, Tetsuji Nakagawa, Teruhisa Misu, and Farig Shah. 2023. BanglaBERT: A Denoising Autoencoding based Pre-trained Language Model for Bengali. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2577–2591.
				Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
6	aacl-2025	2025.banglalp-1.60	Fahmid et al. (2025)	Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Arjun Guha, and 1 others. 2023. Multi-lingual code generation and evaluation with mbxp. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pages 1064–1075.
7	aacl-2025	2025.banglalp-1.62	Shahrier et al. (2025)	Frank F Xu, Suraj Nair, Shuyan Lin, Pengcheng Yin, and Graham Neubig. 2022. Polycoder: A model for programming by example in multiple languages. arXiv preprint arXiv:2202.13169.
8	aacl-2025	2025.banglalp-1.8	Kawser et al. (2025)	Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. Borrowing and code-mixing in social media texts: A case study of hindi-english. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 116–126.
9	aacl-2025	2025.bhasha-1.11	P and Kulkarni (2025)	Yoshitomo Tsukagoshi, Nikhil Sharma, and Pawan Goyal. 2025. Towards accent-aware vedic sanskrit optical character recognition. In Computational Sanskrit and Digital Humanities, World Sanskrit Conference, pages 71–80, Kathmandu, Nepal. ACL.
10	aacl-2025	2025.bhasha-1.13	Ambaliya et al. (2025)	Md. Akmal Hassan, Md. Saiful Islam, and Mumit Khan. 2016. Sentiment Analysis on Bangla and Romanized Bangla Text Using Long Short-Term Memory Recurrent Neural Network. arXiv preprint arXiv:1610.00369.
11	aacl-2025	2025.bhasha-1.14	Dhamecha et al. (2025b)	Abhik Bhattacharjee, Tahmid Hasan, Anindya Sahu, Kumar Shikhar Deep Anand, Md. Rokibul Hasan, Rakesh Sreenivas, Roshni Ghosal, Asif Salekin, and Mohammad Mahbubur Rahman. 2023. Banglat5: A suite of pre-trained language models for Bangla. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 12345–12356.
12	aacl-2025	2025.bhasha-1.17	P (2025)	Eleri Luhtaru and Mark Fishel. 2024. Multilingual grammatical error correction using pre-trained translation models. In Proceedings of the 18th Conference of the European Chapter of the ACL (EACL 2024). Long paper.
13	aacl-2025	2025.bhasha-1.18	Dhamecha et al. (2025a)	Sumanth Doddapaneni, Divyanshu Kakwani, Simran Khanuja, Pushpak Bhattacharyya, and Sunayana Sitaram. 2023. Indicbert v2 and the revival of classical indic nlp benchmarks. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL).
14	aacl-2025	2025.bhasha-1.2	Paul et al. (2025)	Omri Abend and Ari Rappaport. 2017. The universal anaphora annotation framework. Transactions of the Association for Computational Linguistics, 5:561– 577.
15	aacl-2025	2025.bhasha-1.4	Ravikiran et al. (2025)	Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. 2025. Psn-irt: Psychometrics + neural for llm benchmarks. arXiv preprint arXiv:2505.15055.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
16	aacl-2025	2025.findings-ijcnlp.117	Basu et al. (2025)	D. I. Adelani and 1 others. 2022. Afriberta: Large-scale self-supervised pretraining for african languages. arXiv preprint arXiv:2204.06487.
17	aacl-2025	2025.findings-ijcnlp.124	Srikanth et al. (2025)	Zhou' an Zhu, Xin Li, Jicai Pan, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. Medic: A multimodal dataset for empathic dialogue in counseling. arXiv preprint arXiv:2305.14221.
18	aacl-2025	2025.findings-ijcnlp.125	Vasani et al. (2025)	Xiaohui Zhang, Xiang Lisa Li, and 1 others. 2023. Adaptive low-rank adapter (adalora): Compressing lora further via rank allocation. arXiv preprint arXiv:2303.10512.
19	aacl-2025	2025.findings-ijcnlp.134	Shifa et al. (2025)	Alberto Poli, Ilya Sergey, and Greg Morrisett. 2024. Scalable verification of mitigations for memory safety. arXiv preprint arXiv:2404.12464.
20	aacl-2025	2025.findings-ijcnlp.141	Shailya et al. (2025)	Gemma Team, Thomas Mesnard, Cassidy Hardin, Sebastian Borgeaud, Aidan Clark, Arthur Mensch, Michael Ring, Laurent Hoffmann, Eliza Buchatskaya, Andy Cassirer, and 1 others. 2024. Gemma: Open models for responsible ai. Technical report, arXiv preprint arXiv:2403.08295.
21	aacl-2025	2025.findings-ijcnlp.142	Patil et al. (2025)	Shiyue Zheng, Tuhin Ghosal, Ethan Chi, and et al. 2023b. Judging llm-as-a-judge: Comparison and calibration of llms for evaluating open-ended text generation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL).
22	aacl-2025	2025.findings-ijcnlp.24	Vejendla (2025c)	Ren Wang, Bo Dai, and Li Liu. 2021. Self-supervised curriculum learning for deep reinforcement learning. In Advances in Neural Information Processing Systems, volume 34.
23	aacl-2025	2025.findings-ijcnlp.26	Vejendla (2025a)	Albert Gu, Tri Dao, and 1 others. 2023. Mamba: Lineartime sequence modeling with selective state spaces. In NeurIPS.
24	aacl-2025	2025.findings-ijcnlp.28	Dobariya et al. (2025)	John Lee. 2004. A classifier-based approach to preposition and determiner errors. In Proceedings of the 20th International Conference on Computational Linguistics.
25	aacl-2025	2025.findings-ijcnlp.41	Shen et al. (2025)	Yufei Gu, Wenhao Huang, Xinyang Zhang, and 1 others. 2023. Celebv-text: A large-scale video-text dataset for realistic human generation. arXiv preprint arXiv:2312.00734.
26	aacl-2025	2025.findings-ijcnlp.47	Nooralahzadeh et al. (2025)	Jane Doe, John Smith, and Wei Zhang. 2025. Magma: Multi-agent generalist for multimodal alignment. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). To appear.
27	aacl-2025	2025.findings-ijcnlp.53	Sherritt et al. (2025)	Juan Ramirez, Chinasa Okolo, Mahima Hegde, Soumya Banerjee, Aayush Wadhwa, Rashmi Gangadharaiah, and Niranjan Balasubramanian. 2025. Using multimodal models for informative classification of disaster tweets. In Proceedings of the 3rd Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 221–229, St. Julians, Malta. Association for Computational Linguistics.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
28	aacl-2025	2025.findings-ijcnlp.64	Chheda et al. (2025b)	Abdulaziz Barayan et al. 2025. Zero-shot readabilitycontrolled sentence simplification with instructiontuned large language models. In Proceedings of the 30th International Conference on Computational Linguistics (COLING 2025).
29	aacl-2025	2025.findings-ijcnlp.75	Tomar et al. (2025)	Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. Strongreject: Can LLMs strongly say no to jailbreaks? CoRR, abs/2402.10260.
30	aacl-2025	2025.ijcnlp-demo.9	Vatndal and Setty (2025)	Haotian Liu, Pengcheng Lin, Chaoyi Wu, Xiangning Li, Kevin Lin, Ying Zhang, Jingren Du, and Jian Fang. 2023b. Llava: Visual instruction tuning. arXiv preprint arXiv:2304.08485.
31	aacl-2025	2025.ijcnlp-long.103	Fadlon et al. (2025)	Ofir Ben Shoham. and Nadav Rappoport. 2025. CPLLM: Leveraging LLMs for data representation in clinical prediction. ArXiv.
32	aacl-2025	2025.ijcnlp-long.119	Liang et al. (2025)	Xiaoyang Liu, Yi Zhang, et al. 2021. Metaphor-aware poem generation with conceptual mappings. In ACL.
33	aacl-2025	2025.ijcnlp-long.120	Alikhani et al. (2025)	Jiacheng Chen, Dinghan Shen, William Wang, and Diyi Zhang. 2023. Retrieval-augmented pretraining for multiple-choice question answering with knowledge triplets. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1914–1929.
34	aacl-2025	2025.ijcnlp-long.122	Phukan et al. (2025)	Y. Xiao, X. Li, and 1 others. 2021. Next-qa: A dataset for causal and temporal reasoning in video. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4093–4103.
35	aacl-2025	2025.ijcnlp-long.147	Tseytlin et al. (2025)	Shayne Longpre Wanner, Jason Weston, Mikel Artetxe, Arthur Szlam, and Angela Fan. 2024. DecompScore: Reference-based evaluation of sentence decompositions into atomic facts. In arXiv preprint, arXiv. arXiv. ArXiv:2403.11903.
36	aacl-2025	2025.ijcnlp-long.154	Mishra et al. (2025)	Sheng Zhang et al. 2024b. Analyzing the utility of llms for improved query rewriting. arXiv preprint arXiv:2404.12872.
37	aacl-2025	2025.ijcnlp-long.169	Das et al. (2025d)	Xiaodong Wang, Xiaole Zhou, Matthew Fulton, and Sowmya Vajjala. 2023. Prompting for trait-wise supervision: Using rubrics and llms for automated essay scoring. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 122–134.
38	aacl-2025	2025.ijcnlp-long.170	Pattnayak et al. (2025b)	Wenxin Zhu, Ling Han, and Jialin Wang. 2024. Multiturn dialogue understanding for customer support. arXiv preprint.
39	aacl-2025	2025.ijcnlp-long.187	Zeinalipour et al. (2025)	Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024b. Multilingual commonsenseqa (mcsqa) with unified creation strategy by language models and humans. arXiv preprint arXiv:2406.04215.
40	aacl-2025	2025.ijcnlp-long.194	Hemmat et al. (2025)	Martin Azar, Yulia Tsvetkov, and Noah A. Smith. 2024. Hierarchical retrieval for large language models. ArXiv:2401.01234.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
41	aacl-2025	2025.ijcnlp-long.37	Suryanarayanan et al. (2025)	Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135 Million Parallel Sentences from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1625–1639. Association for Computational Linguistics.
42	aacl-2025	2025.ijcnlp-long.55	Poladi and Dandapat (2025)	arthur mensch chris bamford devendra singh chaplot diego de las casas florian bressand gianna lengyel guillaume lample lucile saulnier lélio renard lavaud marie-anne lachaux pierre stock teven le scao thibaut lavril thomas wang timothée lacroix william el sayed Albert q. jiang, alexandre sablayrolles. 2023. arXiv preprint arXiv:2310.06825, 3.
43	aacl-2025	2025.ijcnlp-long.58	Guo and Wu (2025)	Yifu Li, Jialu Li, Yufang Lai, and Nanyun Peng. 2023b. Accord: A multi-document approach to generating diverse and coherent summaries. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8341–8356, Toronto, Canada. Association for Computational Linguistics.
44	aacl-2025	2025.ijcnlp-long.90	Ebrahimi et al. (2025)	Xian Yeow Lee, Shunichi Akatsuka, Lasitha Vidyaratne, Aman Kumar, Ahmed Farahat, and Chetan Gupta. 2025. Reliable decision-making for multi-agent llm systems. arXiv preprint arXiv:2406.04092.
45	aacl-2025	2025.ijcnlp-short.29	Das et al. (2025c)	Anoop Kunchukuttan Krishna, Abhijit Mishra, Monojit Choudhury, and Pushpak Bhattacharyya. 2020b. Graph-based structured models for joint morphological tagging and syntactic parsing of sanskrit. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7431–7442.
46	aacl-2025	2025.ijcnlp-short.4	Clark et al. (2025)	Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. ArXiv.
47	aacl-2025	2025.ijcnlp-short.8	Rafiuddin and Khan (2025a)	Tim Dettmers, Yannic Kilcher, Henry Minsky, Anna McDowell, Neha Nangia, Andreas Vlachos, and the Microsoft Phi Team. 2025. Phi-4-mini: Compact yet powerful multimodal models. arXiv preprint arXiv:2503.01743.
48	aacl-2025	2025.justnlp-main.16	Dhakad et al. (2025)	Ramakrishna Appicharla, Baban Gain, Santanu Pal, and Asif Ekbal. 2025. Beyond the sentence: Contextaware MT survey. arXiv:2506.07583.
49	aacl-2025	2025.justnlp-main.19	Sonowal and Sadhu (2025)	S. Kumar and 1 others. 2024. A survey on chunking strategies for large language models. arXiv preprint.
50	aacl-2025	2025.justnlp-main.5	Sadikot and Jain (2025)	Anand Kumar and Saurabh Sharma. 2022. Indian legal corpus (ilc): A dataset for abstractive summarization. arXiv preprint arXiv:2210.10398.
51	aacl-2025	2025.nlpai4health-main.10	Roy and Das (2025)	Parameswari Krishnamurthy, Dipti Misra Sharma, R. Singh, and 1 others. 2023. Clinical nlp resources and benchmarks for indian languages. In Proceedings of the Workshop on Healthcare NLP for Indian Languages. Workshop paper; proceedings not indexed in ACL Anthology.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
52	aacl-2025	2025.nlpai4health-main.5	Zechariah et al. (2025)	Di Jin, Eileen Pan, Nassim Oufattole, Gerald Wicks, Hua Luo, and Frank Rudzicz. Disease knowledge transfer across languages and modalities. In Findings of the Association for Computational Linguistics: EMNLP 2021, page 2021.
53	aacl-2025	2025.quantumnlp-1.8	Vats et al. (2025)	Xiaoming Liu, Yifan Zhao, and Ming Chen. 2025b. Qcse: Pretrained quantum context-sensitive word embeddings. arXiv preprint arXiv:2509.05729.
54	aacl-2025	2025.sciprodllm-1.3	Lyn and Graham (2025)	Yicheng Chen, Chujie Zhao, Yankai Lin, and Zhiyuan Liu. 2024a. Humans are still better judges: On the evaluation of large language models in text generation. In Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand. Association for Computational Linguistics.
55	aacl-2025	2025.wasp-main.20	Ram (2025)	Yong Lin Rui Pan-Xiang Liu Tong Zhang Shizhe Diao, Pengcheng Wang. 2024. Active prompting with chain-of-thought. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).
56	aacl-2025	2025.wslp-main.4	Bianco et al. (2025)	Ehsan Davoodi and 1 others. 2022. Improving lowresource classification via large language models for data augmentation. In Proceedings of the 60th Annual Meeting of the ACL (Short Papers).
57	eacl-2026	2026.abjadnlp-1.35	Alfattni (2026)	Omar Obeid, Go Inoue, and Muhammad Abdul-Mageed. 2023. A benchmark suite for arabic nlp: Dialectal, classical, and modern standard evaluation. Natural Language Engineering, 29(5):765–789.
58	eacl-2026	2026.abjadnlp-1.4	Elshokry and Alhanjouri (2026)	Ahmed Abdelali, Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2021. QARiB: QCRI arabic and dialectal BERT. arXiv preprint.
59	eacl-2026	2026.abjadnlp-1.42	Alanazi et al. (2026)	Bert Kamphuis and Sjoerd van der Meer. 2024. Tinytoxic-detector: A compact transformer for toxicity detection. Transactions of the ACL, 12(3):215–229.
60	eacl-2026	2026.abjadnlp-1.49	AlGhamdi et al. (2026)	Wenxuan Zhang, Kun Song, Sebastian Riedel, and Antoine Bosselut. 2023c. Benchmarking graph-based hallucination detection for factual consistency in summarization. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4403–4418.
61	eacl-2026	2026.abjadnlp-1.51	Dao Sy et al. (2026b)	Jiucai Li and Yong Jiang. 2023. Deepfake text detection: A study of detection methods and model robustness. arXiv preprint arXiv:2303.16876.
62	eacl-2026	2026.abjadnlp-1.52	Dao Sy et al. (2026a)	Yujin Shao, Yue Zhang, and Ruifeng He. 2024. Authorship style transfer with inverse transfer learning. In ACL.
63	eacl-2026	2026.abjadnlp-1.9	Alshaikh Saleh and Ahmad (2026)	Asli Eyecioglu and Bill Keller. 2016. Character-based neural models for semantic similarity. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 248–255. Association for Computational Linguistics.
64	eacl-2026	2026.africanlp-main.16	Ngoufo et al. (2026)	Y Li, J Chen, and X Wang. 2023. Prompt engineering for large language models: A survey. arXiv preprint.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
65	eacl-2026	2026.africanlp-main.25	Samil and Adelani (2026)	Maja Popovic. 2017. chrF++: Towards character and word n-gram F-score for automatic mt evaluation. In Proceedings of the Second Conference on Machine Translation, pages 568–575. Association for Computational Linguistics.
66	eacl-2026	2026.africanlp-main.30	Moslem et al. (2026)	Laban Kumbuga, Joyce Nakatumba-Nabende, Jonathan Mukiibi, and Andrew Katumba. 2024. SALT: Sunbird African language translation corpus. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 5462– 5472. ELRA and ICCL.
67	eacl-2026	2026.eacl-demo.22	Stoianov et al. (2026)	David Rein and 1 others. 2023. Gpqa: A google-proof q&a benchmark for large language models. arXiv preprint arXiv:2311.16452.
68	eacl-2026	2026.eacl-industry.30	Bhatnagar et al. (2026)	Hongxiao Ma, Ziyang Chen, Ya-chu Wang, Ao Fan, and Jing Yang. 2024. Adaretake: A novel approach for long-video understanding. arXiv preprint arXiv:2406.11029.
69	eacl-2026	2026.eacl-industry.36	Kumar et al. (2026)	Wei-Lin Yuan, Zhaode Wang, Hong-Ying Zan, Zhaohui Lin, Siyuan Bao, Yixin Wang, Jiacheng Liu, Yichi Zhang, Zhen Liu, Lisha Wang, and 1 others. 2024. A survey on leveraging large language models for natural language generation evaluation. arXiv preprint arXiv:2401.07103.
70	eacl-2026	2026.eacl-industry.50	Pattnayak and Chowdhuri (2026)	Phillip Rust, Jonas Pfeiffer, Ivan Vulic, and Sebastian Ruder. 2021. Good-enough compositional data augmentation and sampling for word-piece tokenization. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), pages 3500–3512.
71	eacl-2026	2026.eacl-industry.56	Sahay et al. (2026)	Yan Liu and 1 others. 2024. Imageeval: Benchmarking the factuality of image generation and editing. arXiv preprint arXiv:2403.XXXX.
72	eacl-2026	2026.eacl-industry.61	Trivedi et al. (2026)	Budi Wicaksono and Sung-Hyon Myaeng. 2013. Mining product improvement suggestions from customer reviews. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL).
73	eacl-2026	2026.eacl-industry.62	Ling et al. (2026)	Nguyen Hoang Tien, Nguyen Duc Tai, Huynh Ngoc Thang, and Le Hoang Tan. 2024. Generating Customer Personas for E-commerce Applications. arXiv preprint arXiv:2405.12745.
74	eacl-2026	2026.eacl-industry.7	Verma et al. (2026)	Jie Li, Jian Zhou, Tao Lin, Han Wang, and Fang Liu. 2024. Calibrating large language models with logprobability guidance for factual reliability. arXiv preprint arXiv:2402.08032.
75	eacl-2026	2026.eacl-industry.70	Zeinalipour et al. (2026)	Telmo Menezes and Benjamin Roth. 2019. Distant supervision for ner: A systematic study. arXiv.
76	eacl-2026	2026.eacl-long.131	Korobeynikova et al. (2026)	Yiheng Li, Yulong Qian, Yuxuan Zhang, Yihan Liu, Junyi Yin, and Jiaguang Sun. 2024. A Survey on Knowledge Distillation and its Application in Large Language Models. arXiv preprint arXiv:2402.13116.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
77	eacl-2026	2026.eacl-long.141	Li and Lu (2026)	Jingyuan Cao, Hongzhan Chen, Zirui CHEN, Yujie Ding, Guanqun Hou, Zhiyuan Liu, Zhixu Li, Jiaxing Liu, Xiao-Yang Liu, Yang Liu, Maosong Sun, Hengrui Zhang, Yifei Zhang, and Peng Zhou. 2023. A survey on financial large language models. arXiv preprint arXiv:2311.13788.
78	eacl-2026	2026.eacl-long.154	Fadlon and Bar (2026)	Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2023. Dropping layers in pretrained transformers: How much do you need? arXiv preprint arXiv:2305.00017.
79	eacl-2026	2026.eacl-long.210	Samyoun et al. (2026)	Anthony Brohan, Yevgen Chebotar, and Chelsea Finn. 2023. Robotic manipulation through spatial hallucination mitigation. arXiv preprint arXiv:2303.03534.
80	eacl-2026	2026.eacl-long.241	Hosseini et al. (2026)	Yejin Bang, Haotian Zhang, Yoonki Kim, Frank F Xu, and Wenpeng Yu. 2023. Hallulens: Identifying and categorizing hallucinations in open-ended text generation. arXiv preprint arXiv:2305.16477.
81	eacl-2026	2026.eacl-long.274	Chen et al. (2026b)	Amir H. Zadeh, Ming Chen, Soujanya Poria, and Louis-Philippe Morency. 2018. Cmu-mosei: A multimodal sequence-to-sequence dataset for emotion recognition. In Proceedings of the 27th International Conference on Computational Linguistics.
82	eacl-2026	2026.eacl-long.296	Mubarak et al. (2026)	El Moatez Billah Nagoudi et al. 2022. Arat5: A sequence-to-sequence model for arabic text correction. In Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP).
83	eacl-2026	2026.eacl-long.330	Kordi et al. (2026)	Lucia Zheng, Joel Niklaus, Daniel E. Ho, Christopher D. Manning, and Percy Liang. 2023. Ifeval: A benchmark for evaluating instruction following abilities of large language models. arXiv preprint arXiv:2311.07911.
84	eacl-2026	2026.eacl-long.332	Choraria et al. (2026)	Z Lin, S Shen, H Liu, Y Li, and YJ Lee. 2024. VTW: Visual token withdrawal for large vision-language models. Preprint, arXiv:2405.05803.
85	eacl-2026	2026.eacl-long.352	Shi et al. (2026)	D. Stripelis et al. 2024. A multi-model router for efficient llm inference. In EMNLP.
86	eacl-2026	2026.eacl-long.363	Ye et al. (2026)	Xia Jiang, Wei Xu, Ming Zhao, Li Chen, and Zhicheng Dou. 2024. Graphrag: Unlocking llms for graph-structured reasoning. arXiv preprint arXiv:2403.09236.
87	eacl-2026	2026.eacl-long.367	Kingetsu et al. (2026)	Hanyu Xia, Ming Zhou, and Yunxin Wang. 2024. A comprehensive survey of speculative decoding. In Findings of ACL.
88	eacl-2026	2026.eacl-long.50	Sil et al. (2026)	Google DeepMind. 2025a. Gemini 2.5 flash: Price-performance balanced multimodal thinking model. arXiv preprint arXiv:2507.06261. Also refers to Gemini-2.5 model family including Flash, Flash-Lite, Pro, model card available online.
89	eacl-2026	2026.eacl-long.77	Liang et al. (2026)	Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. arXiv preprint arXiv:1910.01708.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
90	eacl-2026	2026.eacl-long.85	Elgaar and Amiri (2026)	Xiaofei Lu. 2020. Automatic analysis of syntactic complexity in second language writing. ArXiv preprint, abs/2005.02013.
91	eacl-2026	2026.eacl-long.95	Borah et al. (2026)	OpenAI. 2025a. Gpt-5: Scaling laws and emergent reasoning abilities. arXiv preprint arXiv:2505.12345.
92	eacl-2026	2026.eacl-short.29	Banerjee et al. (2026)	Mario Kovac, Michael Moosmüller, Stjepan Marjanovic, and Miloš Stanojevic. 2023. Llms as cultural personas: Benchmarking persona-steered value judgments across cultures. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 13815–13828. Association for Computational Linguistics.
93	eacl-2026	2026.eacl-short.43	Tikhonov et al. (2026)	Yao Fu, Xiaochuang Tang, Hao Peng, and 1 others. 2023. Chain-of-thought early exit: When to stop thinking in large language models. arXiv preprint arXiv:2310.04369.
94	eacl-2026	2026.eacl-short.9	Singh et al. (2026)	Logan Foote, Neel Nanda, Ján Kramár, Ioannis Konstas, Jonathan Cohen, and Fazl Barez. 2023. Neuron-to-graph: Interpretable representation learning for neurosymbolic reasoning. arXiv preprint arXiv:2305.19911.
95	eacl-2026	2026.eacl-srw.10	Balakrishnan and Thayasivam (2026)	J. Shi, J. Xu, Y. Fujita, S. Watanabe, and B. Xu. 2024. Self-supervised singing voice pre-training towards speech-to-singing conversion. arXiv preprint arXiv:2401.05064.
96	eacl-2026	2026.eacl-srw.2	Chen et al. (2026a)	X Wan and 1 others. 2024. Contrastive response generation for mitigating object hallucination in lvlms. arXiv preprint arXiv:2405.xxxxx.
97	eacl-2026	2026.eacl-srw.48	Chernogorskii et al. (2026)	Patrick Es, Menno van Zaanen, Rob Koeling, and Mark Stevenson. 2024. Ragas: An evaluation framework for retrieval-augmented generation. In Proceedings of the 2024 Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations, pages 157–166.
98	eacl-2026	2026.fever-1.2	Shim et al. (2026)	Jun Xie, Zhaopeng Wang, Zhen Li, and 1 others. 2017a. Data noising as smoothing in neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.
99	eacl-2026	2026.fever-1.3	Upadhyay et al. (2026)	Zhaowei Yuan and Andreas Vlachos. 2024. Zero-shot fact verification with semantic triples and knowledge graphs. In Findings of the Association for Computational Linguistics: ACL 2024, pages 12844–12865.
100	eacl-2026	2026.fever-1.4	Gautam et al. (2026)	Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross model consistency. arXiv preprint arXiv:2305.13281.
101	eacl-2026	2026.findings-eacl.104	Su et al. (2026)	Yang Liu, Sheng Zhang, Jingbo Shang, and Jiawei Han. 2020. A top-down neural architecture towards text-level parsing of discourse relations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 6386–6395. Association for Computational Linguistics.

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
102	eacl-2026	2026.findings-eacl.120	Zehle and Aßenmacher (2026)	Albert Qiaochu Jiang, Xiaohan Ren, Yixiao Chen, Michael Zhang, Tianle Wang, and 1 others. 2024. Longllmlingua: Compressing prompts for accelerating long-context llms. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).
103	eacl-2026	2026.findings-eacl.135	Rossini and Plas (2026)	Kevin Walsh, Carlos Ramisch, Agata Savary, and Silvio Cordeiro. 2018. A dataset of verbal multiword expressions annotated with syntactic variants. In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 101–110. Association for Computational Linguistics.
104	eacl-2026	2026.findings-eacl.144	Jandial et al. (2026)	OpenAI. 2025b. Operator: A computer-using ai agent. arXiv preprint arXiv:2501.10114. Research preview of the "Operator" AI agent; released January 23, 2025.
105	eacl-2026	2026.findings-eacl.160	Guo et al. (2026)	Daniel Gutiérrez, Xueguang Dai, Xi Victoria Lin, Chunting Liu, and 1 others. 2025. RAGMemory v2: Nonparametric continual memory for language models. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).
106	eacl-2026	2026.findings-eacl.176	Xu et al. (2026)	Yixin Xu, Siyuan Chen, Yutong Wu, Chenguang Li, and Cheng Li. 2023b. Instruction-tuned language models are better listeners. arXiv preprint arXiv:2308.12038.
107	eacl-2026	2026.findings-eacl.189	Qing et al. (2026)	Priya Jain et al. 2024. Knowledge structure generation and utilization by llms: Emerging methods. arXiv preprint arXiv:2402.03546.
108	eacl-2026	2026.findings-eacl.217	Jain and Garimella (2026)	Kalpesh Krishna and et al. 2024. Hallucination in retrieval-augmented generation: Benchmarking and mitigation. arXiv preprint arXiv:2403.00000.
109	eacl-2026	2026.findings-eacl.253	Bessghaier et al. (2026)	Noura AlShenaifi, Nora Al-Twairesh, and Hend AlKhalifa. 2024b. Llm-in-the-loop active learning for arabic sentiment analysis. arXiv preprint arXiv:2509.23515.
110	eacl-2026	2026.findings-eacl.28	Wadhwa et al. (2026)	Xinya Chen and 1 others. 2023. Benchmarking large language models in rag settings. arXiv preprint arXiv:2311.09383.
111	eacl-2026	2026.findings-eacl.285	Peng et al. (2026)	Xinyang Xu, Zeyu Wang, Xinyi Li, Hang Dong, Qingqing Zheng, Yijia Liu, Wayne Xin Zhao, and Ji-Rong Wen. 2024a. Text2cypher: Training llms to generate executable cypher queries over knowledge graphs. arXiv preprint arXiv:2412.10064.
112	eacl-2026	2026.findings-eacl.305	Choudhury et al. (2026)	Zihan Ji, Daniel Lee, Yixin Pan, Yichong Li, and et al. 2023. Halueval: A large-scale hallucination evaluation benchmark for open-domain dialogue models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL).
113	eacl-2026	2026.findings-eacl.306	Leng et al. (2026)	Yanyan Cui. 2022. Ctap: A linguistic complexity feature extraction tool. In Proceedings of the Language Resources and Evaluation Conference (LREC).

Continued on next page

	Event ID	Paper ID	HalluCited Paper	Example of HalluCitation
114	eacl-2026	2026.findings-eacl.311	Cheshmi et al. (2026)	Zheng Yang and 1 others. 2022. Multimodal sarcasm detection with image-text fusion. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL).
115	eacl-2026	2026.findings-eacl.314	Al Monsur et al. (2026)	Hugo Touvron et al. 2024. Llama 3: Open foundation and instruction models. arXiv preprint arXiv:2404.12345.
116	eacl-2026	2026.findings-eacl.323	Mishra et al. (2026)	Qingxiu Huang, Jun Wang, Lei Wu, and Jingbo Zhou. 2024. Frost: Fine-grained reward optimization for step-wise thinking. arXiv preprint arXiv:2403.00604.
117	eacl-2026	2026.findings-eacl.335	Siro et al. (2026)	Yvan Dubois, Adam Jacob, Camille Dubois, Yonatan Belinkov, and Aviv Efrat. 2023. Alpacaeval: A strong set of reference rankings for instruction following llms. arXiv preprint arXiv:2310.01377.
118	eacl-2026	2026.findings-eacl.43	Sun et al. (2026)	John W Ely, Jerome A Osheroff, Melisa L Chambliss, Mark H Ebell, and Mark E Rosenbaum. 2011. The value of clinical questions: identifying research priorities in primary care. BMJ Quality & Safety, 20(9):787–792.
119	eacl-2026	2026.lchange-1.1	Aloraini et al. (2026)	Li Lucy and David Bamman. 2021. Style variation and social meaning in online communities. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), pages 4745–4760. Association for Computational Linguistics.
120	eacl-2026	2026.loreslm-1.1	Nyalang (2026)	Agniva Maiti, Manya Pandey, and Murari Mandal. 2025. NagaNLP: Bootstrapping NLP for low-resource Nagamese with closed-loop synthetic data. arXiv preprint arXiv:2512.12537.
121	eacl-2026	2026.loreslm-1.16	Kazi and Khoja (2026)	Zihan Wang and 1 others. 2023. Character-word entanglement. In ACL.
122	eacl-2026	2026.loreslm-1.17	Panahi et al. (2026)	Mikayla Sclar, Yejin Choi, Yulia Tsvetkov, and Maarten Sap. 2024. Quantifying language models' sensitivity to spurious features in prompt design. arXiv preprint arXiv:2402.16892.
123	eacl-2026	2026.loreslm-1.19	Umar et al. (2026)	David Ifeoluwa Adelani, Jesujoba O. Alabi, Angela Fan, Julia Kreutzer, Samuel Olowobo, Ituen Okayo, Bernard Opoku, Chiamaka Chukwuneke, and 1 others. 2022. A few-shot learning approach for lowresource african languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 904–917.
124	eacl-2026	2026.loreslm-1.3	Yadav (2026)	Shuyuan Xie, Suraj Maharjan, and Thamar Solorio. 2023. An empirical study of in-context learning in large language models with code-switching. arXiv preprint arXiv:2305.09066.
125	eacl-2026	2026.loreslm-1.53	Stanislas Rocksane COMPAORE et al. (2026)	Laura Martinus and Jade Z. Abbott. 2020. Vukuzenzele and IsiZulu: Introducing two new castilian-South African machine translation datasets. In Proceedings of the Fourth Widening NLP Workshop, pages 146–148, Seattle, USA. Association for Computational Linguistics.
126	eacl-2026	2026.loreslm-1.55	Gurav et al. (2026)	Amir Kargaran and 1 others. 2025. Layer-wise crosslingual alignment in low-resource languages. arXiv preprint.

Continued on next page

	EVENT ID	PAPER ID	HALLUCITED PAPER	EXAMPLE OF HALLUCITATION
127	eacl-2026	2026.loreslm-1.9	Keita et al. (2026)	Kevin Scannell. 2007. The challenges of developing computational resources for minority languages. In Proceedings of the ACL-2007 Workshop on Computational Approaches to Semitic Languages, pages 65–72.
128	eacl-2026	2026.loresmt-1.2	Macayan et al. (2026)	Filipe Cardoso, Rui Silva, and Luis Gomes. 2022. Improving neural MT of indigenous languages with multilingual transfer learning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 299– 307, Dublin, Ireland. Association for Computational Linguistics.
129	eacl-2026	2026.loresmt-1.6	Abela et al. (2026)	Toms Bergmanis and Marcis Pinnis. 2021. Context-independent terminology translation with neural machine translation models. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 142–153.
130	eacl-2026	2026.mme-main.4	Pàmies et al. (2026)	Nikolai Lörer, Norman Lee, and Douwe Kiela. 2023. Conceptarc: Evaluating abstract reasoning in language models with the abstraction and reasoning corpus. arXiv preprint arXiv:2306.01774.
131	eacl-2026	2026.mwe-1.17	Alexandru-Marian (2026)	Yue Fan and 1 others. 2025. An empirical study of positional bias in large language models. arXiv preprint arXiv:2501.00000.
132	eacl-2026	2026.mwe-1.20	Karatepe et al. (2026)	Joakim Nivre. 2008. Deterministic dependency parsing algorithms. In ACL, pages 513–520. Association for Computational Linguistics.
133	eacl-2026	2026.mwe-1.21	Kömürcü and Temel (2026)	Yukun Ma, Xiang Chen, and Zhiyuan Liu. 2023. Figurative language grounding with vision–language models. In Proceedings of EMNLP.
134	eacl-2026	2026.mwe-1.22	Bogdanova and Bucur (2026)	Shiva Taslimipoor and Omid Rohanian. 2019. Investigating BERT for multilingual NLP tasks: The case of multiword expressions. In Proceedings of the 15th Workshop on Multiword Expressions (MWE 2019), pages 157–163.
135	eacl-2026	2026.sigturk-1.10	Alizada and Dubossarsky (2026)	Junjie Hu and 1 others. 2020. Back-Translation for LowResource Toxic Language Detection. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1–6.
136	eacl-2026	2026.sigturk-1.13	Aktaş et al. (2026)	H. Ozan Hacıfazlıoglu, Vahid Partovi Nia, and Ercan Kuruoglu. 2024. Fine-tuning large language models for turkish. arXiv preprint arXiv:2407.15185.
137	eacl-2026	2026.vardial-1.14	Panchal et al. (2026)	Divyansh Subramani and David Bau. 2022. Directional stimulus optimization: Open-ended text generation as activation steering. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).
138	eacl-2026	2026.vardial-1.15	Kyslyi et al. (2026)	Davide Michelsanti and et al. 2019. Asr for unwritten languages: A survey. Preprint, arXiv:1907.13511.
139	eacl-2026	2026.vardial-1.8	Chatzikyriakidis et al. (2026)	Eve Fleisig, Suchin Gururangan, and Noah A. Smith. 2024. Standard language ideology in NLP. arXiv preprint arXiv:2404.11968.

CONTINUED ON NEXT PAGE

	EVENT ID	PAPER ID	HALLUCITED PAPER	EXAMPLE OF HALLUCITATION
140	eacl-2026	2026.wassa-1.11	Aranberri (2026)	Eleftheria Briakou and Marine Carpuat. 2021. A case for evaluating emotion in machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP), pages 2855–2869.
				CONCLUDED