

What Counts as Hallucination? A Scoping Review of Tasks, Symptoms, and Convergence Expectations in LLM Research

Toshiki Kuramoto^{1,2}, Ryohei Kamei¹, Jun Suzuki^{1,3,4}

¹Tohoku University, Miyagi, Japan,

²Bridgestone Corporation, Tokyo, Japan,

³Riken, Tokyo, Japan,

⁴NII LLMC, Tokyo, Japan,

Correspondence: kuramoto.toshiki.q1@dc.tohoku.ac.jp

Abstract

Hallucination in large language models (LLMs) and related generative models has become a central issue for reliability and practical deployment. Yet prior work does not consistently define what is treated as hallucination or the level of expectation under which model outputs are evaluated. This study conducts a scoping review of 244 papers published between 2018 and 2025, analyzing definitions of hallucination, output-side symptoms, and Convergence Expectation (CE). The results show that approximately 90% of the papers provide some form of operational definition, but the symptoms targeted by these operationalizations vary substantially by task. *Addition* appears across many tasks as a common symptom, whereas *Transformation* and *Deviation* strongly depend on task structure and input constraints. Analyses by CE band further show that, even within the same task, the symptoms problematized as hallucination vary according to the strictness of expected output convergence. These findings indicate that hallucination is not a fixed failure inherent in model outputs, but a phenomenon shaped by task, input constraints, evaluation criteria, and the expectations of humans or evaluators. This study reframes hallucination as expectation mismatch and argues that definitions, symptom scope, and expectation levels must be made explicit when comparing research findings and designing evaluations.

1 Introduction

Large language models (LLMs) and related generative models are widely used across a variety of natural language generation tasks. At the same time, these models may produce outputs that are inconsistent with input, context, or external knowledge. This phenomenon is commonly referred to as hallucination. Hallucination is a critical issue for the reliability of generative AI systems, decision support, practical deployment, and safety.

However, the term “hallucination” is not used consistently across prior work. Previous studies have noted that hallucination is defined in a task-dependent manner and that its manifestations, evaluation methods, and mitigation strategies differ across NLG tasks such as summarization, dialogue, question answering, data-to-text generation, and machine translation (Yang et al., 2023; Ji et al., 2023a; Xu et al., 2023b). Furthermore, it has been pointed out that there is a lack of a clear and unified definition in this field (Filippova, 2020; Raulnak et al., 2021; Rawte et al., 2023). In practice, what counts as hallucination differs between tasks with explicit input sources, such as machine translation and summarization, and tasks with weaker input constraints, such as dialogue and open-ended question answering. Moreover, the same output-side departure may be treated as problematic in one task context, while being acceptable as a creative or exploratory output in another. Under these conditions, hallucination rates and detection performance reported across studies cannot be compared straightforwardly. Recent benchmarking work has likewise identified inconsistent definitions and category coverage across hallucination benchmarks, complicating direct comparison of evaluation results (Bang et al., 2025).

This study begins from a form of uncertainty that precedes model-level uncertainty estimation. Hallucination research often examines predictive or semantic uncertainty as a signal for detecting unreliable generations (Farquhar et al., 2024; Xiao and Wang, 2021). In contrast, we focus on conceptual and operational uncertainty: uncertainty about which output phenomena are labeled as hallucination, which criteria are used to identify them, and how strictly outputs are expected to conform to a task-specific ideal. This uncertainty is located not in an individual model’s probability distribution, but in the definitions, evaluation practices, and task assumptions adopted across studies. We introduce

Convergence Expectation (CE) to make one dimension of this variation explicit, namely the strictness with which each paper expects outputs to converge toward an ideal response.

To this end, this study conducts a scoping review of how hallucination is defined and operationalized in research on LLMs and related generative models. The review consists of two phases. The first phase explores how target papers define hallucination and in which tasks and evaluation settings the term is used. The second phase uses the output-side symptom taxonomy and Convergence Expectation (CE) derived from this exploratory phase as a fixed codebook and re-codes the target papers.

The central claim of this study is that hallucination is not a single property inherent in model outputs, but a phenomenon defined through the interaction between task structure and expectations. In tasks where strong faithfulness to an input source is expected, unsupported content is more likely to be treated as hallucination. In contrast, in open-ended dialogue or creative generation, some degree of divergence may be acceptable, and the same type of deviation may not necessarily be treated as hallucination. Thus, evaluating hallucination requires explicit specification of task-specific input constraints, the range of desirable outputs, and the expected level of accuracy or groundedness.

This study makes three contributions. First, it systematically examines the clarity of hallucination definitions in related work. Second, it classifies output-side symptoms treated as hallucination and quantitatively shows how these symptoms differ across tasks. Third, it introduces CE as an analytic construct and shows that judgements of hallucination are shaped by task structure and expectation levels. The contribution is not merely the general observation that hallucination is task-dependent. Rather, this study provides systematic evidence of how that dependence is operationalized across 244 papers and offers a fixed codebook for separating definitional clarity, output-side symptom scope, and evaluative strictness. Through these steps, this study reframes hallucination not merely as a model failure, but as an evaluative concept in research, evaluation design, and practical deployment.

2 Related Work

Hallucination in natural language generation has been treated through various conceptual lenses (Rawte et al., 2023; Ji et al., 2023a). Early

work often understood hallucination as a lack of faithfulness to the input source. For example, studies of machine translation have emphasized the need to distinguish hallucination from ordinary translation errors (Dale et al., 2023a), while studies of summarization have proposed the distinction between intrinsic and extrinsic hallucination (Maynez et al., 2020). The former refers to outputs that contradict the source, whereas the latter refers to outputs containing information that cannot be determined from the source. This distinction is useful, but it primarily assumes tasks with explicit input sources, such as translation and summarization. More broadly, faithfulness research has examined source consistency across multiple NLG tasks and has shown that evaluation and optimization methods remain highly task-specific (Li et al., 2022).

With the spread of LLMs, hallucination has come to refer not only to inconsistency with input, but also to factual errors with respect to the external world, deviations from context, generation of unverifiable information, and misinterpretation of user intent (Huang et al., 2025c). This expansion has broadened the scope of research, but has also increased conceptual ambiguity. Although many classifications have been proposed, they are used at different granularities and for different purposes, and they do not function as a unified coding system across studies (Bang et al., 2025). This inconsistency extends to human evaluation. A recent survey found substantial variation in definitions, terminology, annotation guidelines, agreement reporting, and evaluator information across hallucination studies (Schmidtova et al., 2025). This study complements prior work by evaluating differences in definitional clarity and operationalization across the literature.

The operationalization of hallucination is also strongly task-dependent. In some tasks, adding information not present in the input is a serious mismatch with task expectations, while in other tasks, inference beyond the input may be expected. Existing benchmarks operationalize hallucination differently across task-specific settings, including question answering, knowledge-grounded dialogue, summarization, and general instruction following (Li et al., 2023a). Therefore, hallucination cannot be adequately understood by simply listing surface-level characteristics. It must be considered in relation to task-specific input constraints and the expected range of outputs.

To capture these differences in expectations, this

study introduces *Convergence Expectation (CE)*. CE is a paper-level proxy for the degree to which each paper expects model outputs to converge toward an ideal output. It does not redefine hallucination itself. Rather, CE provides an analytic framework for comparing how strictly prior studies expect output convergence. This perspective helps explain why one study treats a particular departure as hallucination, while another treats a similar departure as outside the evaluation target or as an acceptable output. Prior surveys organize hallucination types, causes, detection methods, or human-evaluation practices. In contrast, our review focuses on how papers operationalize the hallucination label across tasks, separating definitional clarity, output-side symptoms, and the strictness of expected output convergence.

3 Method

To clarify how hallucination is defined and operationalized in research on LLMs and related generative models, this study conducted paper-level coding based on a scoping review. The method consists of corpus construction, derivation of analytic constructs through exploratory review, re-coding using a fixed codebook, and LLM-assisted evidence extraction followed by targeted human review of sampled cases. The LLM was used to extract verbatim evidence and generate initial paper-level codes under the fixed codebook. Automated validation was applied to all outputs, while targeted human review was conducted on sampled cases, primarily those marked by uncertainty flags.

3.1 Corpus Construction

This section outlines the creation of the corpus. Refer to Appendix A for the detailed extraction protocol. The corpus was constructed primarily from papers whose titles contain “Hallucination” or its morphological variants and that were published between 2016 and 2025 in major AI and NLP venues. After applying the eligibility criteria, the final included corpus covered papers published from 2018 to 2025. Some relevant studies address hallucination without explicitly using the term in their titles. However, we restricted the corpus to papers that explicitly positioned hallucination as a primary topic.

The target venues included AAAI, ACL, COLING, CoNLL, EACL, EMNLP, ICLR, ICML, IJCAI, INLG, NAACL, and NeurIPS. Each paper

was assigned a unique ID, and its publication year, venue, title, PDF URL, and metadata were recorded. For textual interpretation, verbatim quotations from the PDF were used as evidence for definitions and scoring. A total of 244 papers were obtained through this process.

3.2 Phase 1: Exploratory Review

In the first phase, we conducted an exploratory review using an initial sample of the target literature. Rather than fixing the final categories in advance, we descriptively recorded how each paper used the term hallucination. Specifically, we extracted definitions, judgement criteria, target tasks, input sources, output examples, evaluation metrics, annotation procedures, and descriptions related to risk and trust.

When comparing analytic memos, we focused not on merging the terminology of individual papers, but on recurrent differences in judgement across papers. The same expression, such as “unsupported,” was connected to explicit labeling rules or metrics in some papers, but remained a general description in others. In this context, the expressions frequently used in the previous studies include “unrelated to the source text,” “contradicted to the source,” “Hard to be verified with the source,” “not include some of the source text”. In addition, machine translation and summarization tended to center on the relation between output and input source, whereas open-ended question answering and dialogue tended to emphasize world knowledge, context, and user intent. Based on these observations, we derived definitional clarity, output-side symptoms, and CE as analytic constructs for Phase 2. Each construct was adopted only if it appeared across multiple task domains and could be applied at the paper level.

3.3 Coding Constructs

The codebook consisted of four major components: (1) definitional clarity, (2) output-side symptoms, (3) CE, and (4) attitudinal or evaluative framing. Please refer to the Appendix for details regarding the codebook provided to the LLM.

For cross-paper analysis, the original task labels were normalized into nine higher-level categories: Data-to-Text, Dialogue, General / Conditional NLG, Image Captioning, Machine Translation (MT), Other / Low-frequency, Question Answering (QA), Retrieval-Augmented Generation (RAG), and Summarization. Conceptually similar

tasks were grouped together, whereas infrequent tasks that could not be reliably analyzed as separate categories were assigned to Other / Low-frequency. For papers addressing multiple tasks, the primary task emphasized in the paper was used for the main analysis. The original task labels were retained separately for traceability.

Definitional clarity was coded on three levels from Lv0 to Lv2. Lv2 indicates that the paper operationally specifies how hallucination is judged, for example through annotation rules, metric definitions, label criteria, or verification procedures. Lv1 indicates that the paper provides a general description or cites prior work without a clear judgement procedure in the paper itself. Lv0 indicates that the term is used but no usable definition or judgement criterion is available.

Output-side symptoms were coded using five categories. *Transformation* refers to distortion of or contradiction with the meaning of an input source. *Addition* refers to the introduction of content not supported by the input source. *Omission* refers to missing required information only when the paper treats such missing information as hallucination. *Deviation* refers to divergence from context, user intent, or world knowledge in tasks with weak input sources. *Other* refers to task-specific output phenomena not covered by the preceding categories. Details regarding these symptoms are provided in Appendix B.2. Internal indicators such as predictive entropy, attention, hidden states, and token contributions were not used as symptom flags and were recorded separately.

Convergence Expectation (CE) is an analytic construct for comparing how strictly each paper expects output convergence. CE comprises seven paper-level items: source faithfulness (CE1), external factuality (CE2), failure cost (CE3), evaluation rigor (CE4), source constraint (CE5), allowance for creative divergence (CE6), and accountability or safeguards (CE7). CE1–CE5 and CE7 were scored from 0 to 2, whereas CE6 was scored as 0, −1, or −2 and served as a negative correction when creative or exploratory divergence was permitted or valued. The item scores were summed and clipped to the 0–12 range. CE complements task labels by capturing variation in evaluative strictness among papers assigned to the same task category. The complete rubric is provided in Appendix B.3.

3.4 LLM-assisted Evidence-first Coding and Validation

To process 244 papers consistently, we used a two-stage LLM-assisted coding procedure. We used GPT-5.6 Sol (OpenAI, 2026) via the API for evidence extraction and initial paper-level coding. Each paper was processed using the fixed prompt provided in the supplementary material.

In the first stage, the LLM extracted verbatim evidence for definitions, tasks, symptoms, CE items, and attitudinal framing, together with page and section information. Missing evidence was recorded as NOT FOUND.

In the second stage, paper-level codes were generated using only the extracted evidence. Ambiguous cases were assigned the lower adjacent score, and explicit gates were applied to Lv2 judgements and the highest CE item scores. Outputs were stored as structured JSON and automatically checked for required fields, valid score ranges, and internal consistency. Human review was conducted on a targeted sample rather than on the full corpus. Targeted human review was conducted for 90 of the 244 papers (36.9%). The reviewed sample focused on cases marked by uncertainty flags, including ambiguous definition levels, CE items near the two-point gates, unclear symptom mappings, multi-task papers, and cases in which model-internal indicators could be confused with output-side symptoms. Human review was used to identify and correct coding issues in the sampled cases. It was not treated as an independent validation of every paper in the corpus. The full codebook, scoring gates, and evidence-selection rules are provided in Appendix B.

4 Results

This section reports the coding results for definitions of hallucination, output-side symptoms, CE, and their relationships with task categories. We first examine definitional clarity, then analyze task-dependent symptom distributions, and finally examine how task, symptom, and CE jointly shape what is treated as hallucination.

4.1 Definition Clarity Is Often Operationalized, but Not Uniform

As shown in Figure 1, 220 of the 244 papers (90.2%) were coded as Lv2, 24 (9.8%) as Lv1, and none as Lv0. However, the dominance of Lv2 does not imply conceptual uniformity. A qualitative

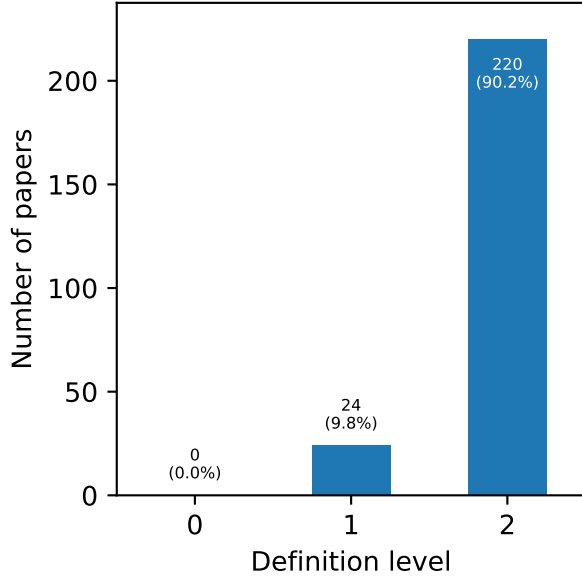


Figure 1: Distribution of definition clarity levels in the review corpus. Lv0 indicates papers without a usable definition, Lv1 indicates papers with a generic or cited definition, and Lv2 indicates papers with operational criteria.

overview showed that the ways in which hallucination was operationalized differed substantially by task and evaluation setting. The results therefore suggest that uncertainty in hallucination research does not simply arise from the absence of definitions, but from the coexistence of different definitions and evaluation procedures across task contexts.

4.2 Hallucination Symptoms Are Task-dependent

Figure 2 shows the prevalence of each output-side symptom by task category.

Analysis of output-side symptoms by task category showed that *Addition* was the most widely observed symptom. *Addition* was coded in more than 50% of papers in every task category, indicating that the introduction of content unsupported by an input source, image, knowledge base, or context is a common symptom in hallucination research. In contrast, *Transformation* varied substantially by task and appeared more frequently in tasks where distortion of input, context, visual information, or knowledge sources was problematized. *Deviation* was especially salient in QA and General / Conditional NLG, indicating that weakly constrained tasks tend to treat divergence from world knowledge, user intent, dialogue context, or expected answer space as hallucination. *Omission* was gen-

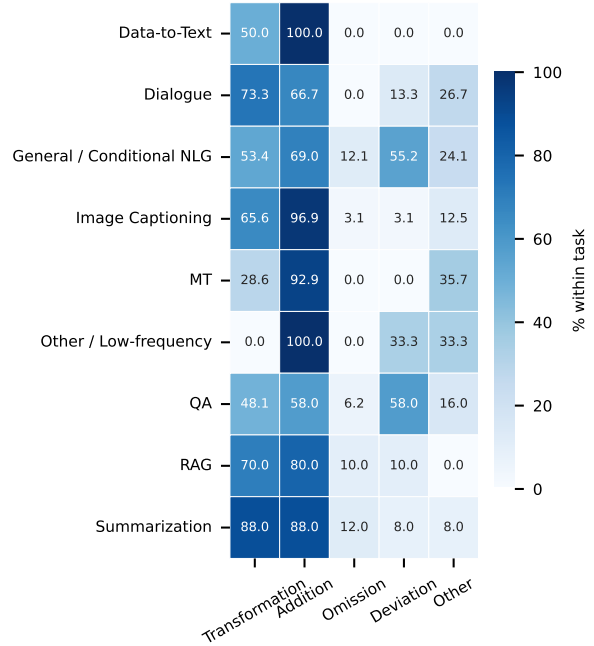


Figure 2: Task-specific prevalence of hallucination symptoms. Rows indicate task categories and columns indicate output-side symptom categories. Each cell reports the percentage of papers within that task category that treated the corresponding symptom as hallucination.

Table 1: Exploratory association tests between task category and hallucination symptom labels.

Symptom	χ^2	df	p	p_{FDR}	V
Transformation	24.617	8	<.01	<.01	.318
Addition	28.943	8	<.01	<.01	.344
Omission	7.101	8	.526	.526	.171
Deviation	67.942	8	<.01	<.01	.528
Other	11.941	8	.154	.192	.221

erally infrequent, suggesting that missing information is often treated as a different error type, such as coverage error or under-generation. However, we are cautious about interpreting these associations as independent causal effects stemming from task categories. This is because the codebook itself may be inherently dependent on the source materials referenced. Therefore, this analysis quantifies how hallucinations are operationally defined and classified based on these distinctions within the literature under review.

To support these descriptive patterns, we conducted chi-square tests of association between symptom presence and task category, reporting Cramér’s V as an effect size (Table 1). *Transformation*, *Addition*, and *Deviation* remained associated with task category after FDR correction. *Devia-*

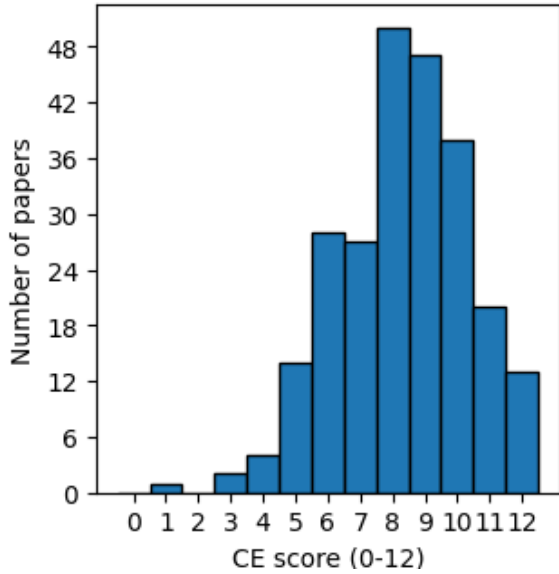


Figure 3: Distribution of CE scores in the review corpus. CE scores are clipped to the 0–12 range.

tion showed the largest effect size, consistent with its higher prevalence in QA and General / Conditional NLG. *Omission* showed no significant association, and Other was not significant after FDR correction. Because some cells had small expected counts, these tests are interpreted in conjunction with effect sizes and visualization results.

4.3 CE Scores Vary Across Tasks but Are Not Determined by Task Alone

Figure 3 presents a histogram showing the number of papers based on CE scores. CE scores were concentrated around 8 to 9, indicating that most target papers assume moderate to high levels of output convergence.

Next, Figure 4 plots the distribution of CE scores for each task. QA, Dialogue, RAG, Summarization, and Image Captioning tended to show higher CE scores, reflecting stronger expectations that outputs align with input sources, external knowledge, visual information, or answer labels. In contrast, General / Conditional NLG showed a wider range of CE scores, reflecting heterogeneity in evaluation expectations across free generation, long-form generation, general-purpose NLG, and multi-task settings. A Kruskal-Wallis test did not show a statistically significant difference in CE score distributions across task categories at the 5% level ($H = 12.501$, $p = .130$), and the effect size was small ($\epsilon^2 = .019$). This result suggests that CE is not strongly explained by task category alone. CE should therefore be interpreted as a distinct axis

that includes input constraints, evaluation procedures, methods for verifying factuality or faithfulness, and framing around risk or accountability.

4.4 Task, Symptom, and CE Jointly Shape What Counts as Hallucination

Finally, we analyzed the relationship among task, symptom, and CE. Because the observed CE scores were concentrated around 8–9, we used auxiliary observed CE bands: Lower observed CE (1–6), Middle observed CE (7–9), and Higher observed CE (10–12). This is not a redefinition of CE, but an aid for interpreting the observed distribution.

Figure 5 shows that symptom prevalence varies across observed CE bands. We further break this down and analyze it by task. Figure 6 shows how these patterns differ by task. Because output-side symptom coding was multi-label, a paper could contribute to more than one symptom column. Because several task-by-band cells contained few papers, we interpret Figure 6 primarily at the aggregate pattern level and avoid drawing substantive conclusions from individual low-count cells.

In Lower observed CE, symptom distributions varied by task: *Deviation* was high in QA, *Transformation* was high in Summarization. In Middle observed CE, *Addition* was high across many tasks. At the same time, *Deviation* remained high in QA and General / Conditional NLG, indicating that divergence from world knowledge or context continues to matter in tasks with weak sources. In Higher observed CE, both *Transformation* and *Addition* were prominent. In QA, *Transformation* reached 76.7%, *Addition* 86.7%, and *Deviation* 43.3%, suggesting that high-CE QA studies treat a broad range of output inconsistencies as hallucination.

These findings show that the hallucination label is not determined by an output-side symptom alone, but is shaped by the interaction among task structure, output-side symptoms, and CE.

5 Discussion

5.1 Hallucination as Expectation Mismatch

The central implication of this study is that hallucination is not determined by model outputs alone, but is defined through the relation between task structure and evaluation expectations. *Addition* appeared across many tasks, whereas *Transformation* and *Deviation* varied substantially by task.

This indicates that whether an output is treated as an inappropriate departure depends on the clar-

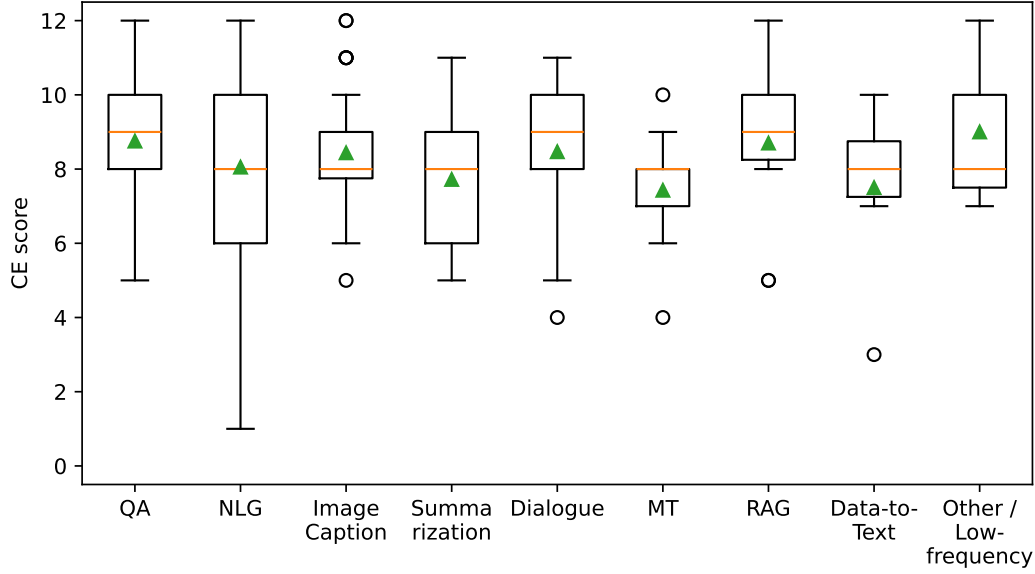


Figure 4: Distribution of CE scores by task category. Boxes indicate interquartile ranges, and triangles indicate task-level means.

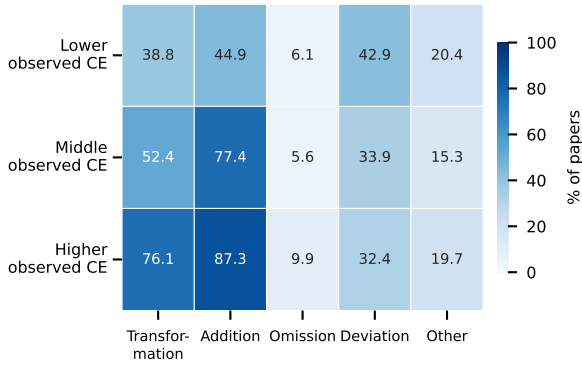


Figure 5: Symptom prevalence by observed CE band.

ity of the input source, the nature of the reference object, and the task objective. From this perspective, hallucination can be understood as expectation mismatch. An output is called hallucination when it fails to satisfy the level of faithfulness, factuality, groundedness, contextual consistency, or accountability expected in a given task. Thus, the definition of hallucination depends not only on the formal properties of output, but also on the task in which the output is generated, the input constraints under which it is evaluated, and the expectations against which it is judged. This interpretation is descriptive rather than causal. The results show how the prior works organize hallucination judgements, not that task or CE independently determines the symptoms produced by a model.

5.2 From Undefined Hallucination to Diverse Operationalizations

Most papers in this study were coded as Lv2. This does not mean that the definitional problem in hallucination research has been resolved. The existence of operationalization is different from the existence of a unified operationalization. Many papers operationalized hallucination through annotation criteria, benchmarks, evaluation metrics, or detection procedures, but the target phenomena differed by task and evaluation setting. The current issue is not simply that hallucination is undefined, but that different operational definitions coexist while being compared under the same term.

5.3 Implications for Evaluation Design

The findings have three implications for evaluation design. First, researchers should specify not only their definition of hallucination, but also the scope of symptoms being evaluated. *Addition*, *Transformation*, *Deviation*, *Omission*, and *Other* do not appear with equal frequency and are distributed differently across tasks. Second, evaluation metrics and benchmarks should be designed in relation to the CE of the task. High-CE tasks require mechanisms such as citation, external verification, expert review, human-in-the-loop procedures, and auditability, whereas low-CE or creative tasks require evaluations that do not over-suppress diversity or exploration. Third, comparative studies of hallucination should align task categories and CE, or at least report them explicitly. Directly comparing

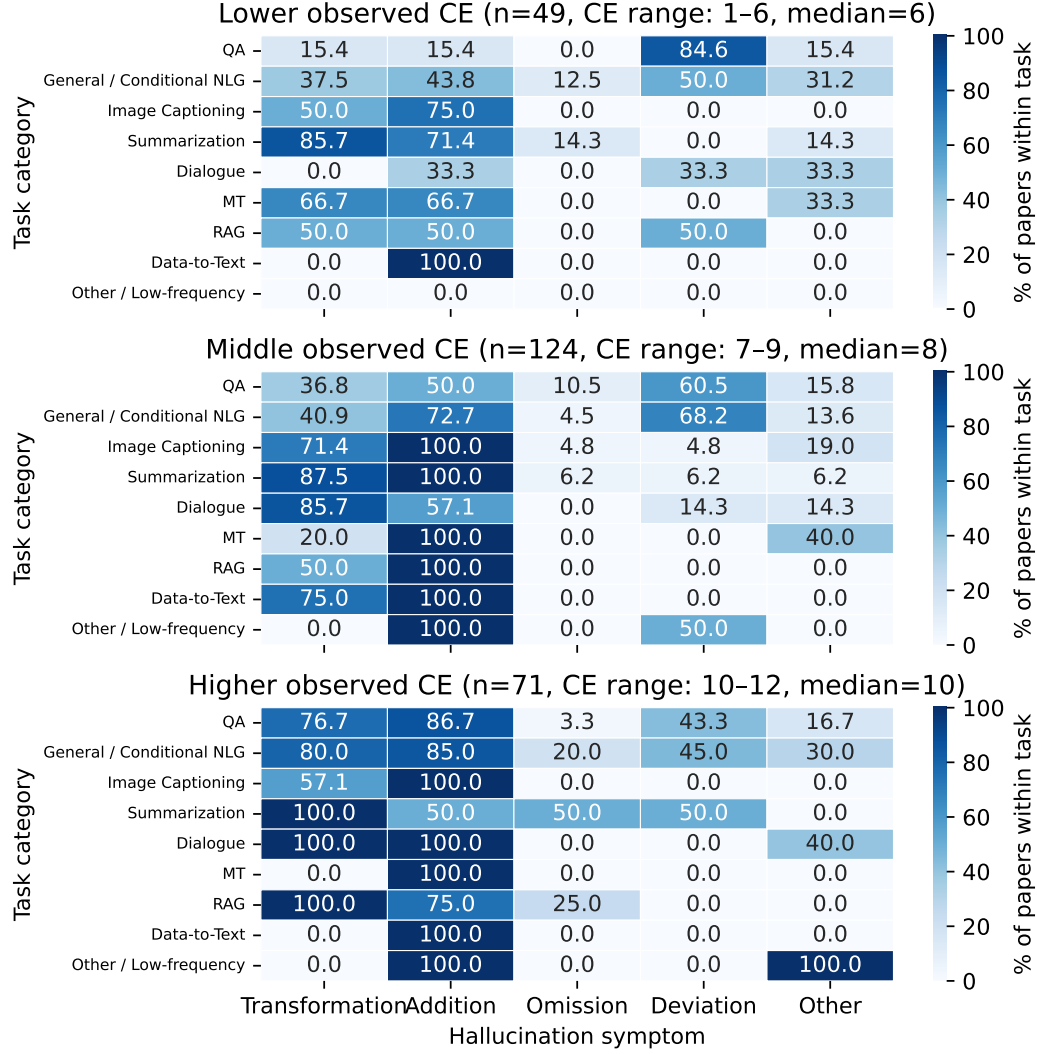


Figure 6: Task-specific hallucination symptoms by observed CE band. Papers are grouped into Lower observed CE (1–6), Middle observed CE (7–9), and Higher observed CE (10–12) based on the observed distribution of CE scores. Each cell reports the percentage of papers within the corresponding task and CE band that received the symptom label. Because symptom coding was multi-label, percentages across a row may sum to more than 100%. Underlying paper and symptom counts are reported in Appendix Table 8.

hallucination rates or detection performance across different tasks, CE levels, and symptom scopes may lead to misleading interpretations.

This study extends the discussion of uncertainty around hallucination beyond model-internal predictive uncertainty. Prior work has estimated the likelihood of hallucination using entropy, self-consistency, hidden states, token probabilities, epistemic uncertainty, and aleatoric uncertainty. The present findings show that there is uncertainty at an earlier stage: uncertainty about what is to be treated as hallucination. In this sense, uncertainty exists not only in model probability distributions, but also in evaluative concepts, symptom scope, and expectation levels. Therefore, future studies require a multidimensional evaluation design that

encompasses tasks, symptoms, and expected standards, rather than treating hallucination as a single global metric.

6 Conclusion

This study analyzed how hallucination is defined and operationalized in research on LLMs and related generative models through a scoping review of 244 papers. The results show that many papers operationalize hallucination in some form, but that what is treated as hallucination differs substantially by task category. *Addition* appeared as a common core symptom across many tasks, whereas *Transformation* and *Deviation* strongly depended on task structure and input constraints. The Convergence Expectation (CE) analysis further showed that hal-

lucination judgements are conditioned by the strictness of output convergence required by the task.

These findings demonstrate the limits of treating hallucination as a single failure inherent in model outputs. Rather, hallucination is a phenomenon defined through the interaction among task, input constraints, evaluation criteria, and the expectations of humans or evaluators. By reframing hallucination as expectation mismatch, this study highlights the need to explicitly report definitions, symptom scope, and CE in comparative research and evaluation design.

The contribution of this study is the systematic quantification of how task dependence is operationalized across the literature. By separating definitional clarity, output-side symptom scope, and convergence expectations, the proposed codebook provides a basis for more transparent comparisons across hallucination studies.

Limitations

This study has several limitations. First, the corpus was constructed from major NLP and AI venues, and does not exhaustively cover all hallucination research. Applied research in HCI or information systems may not be sufficiently represented. Therefore, the results should be interpreted as trends in the target corpus, not as comprehensive population estimates for all generative AI research. Second, Convergence Expectation (CE) is a proxy inferred from textual descriptions in papers, not a direct measurement of user or evaluator expectations. Third, LLM-assisted coding may involve extraction errors or interpretive variation. Although this study combined verbatim quotations, conservative scoring, JSON validation, uncertainty flags, and targeted human review to reduce overreliance on LLM outputs, human review was conducted on a sample focused primarily on flagged cases rather than on the full corpus. Consequently, extraction or coding errors may remain among cases that were not manually reviewed. Furthermore, the targeted audit did not involve independent coding by multiple human annotators. Therefore, inter-annotator agreement could not be estimated. Fourth, some observed task-symptom associations are partly induced by the source-sensitive coding rules. Transformation requires an identifiable source, whereas Deviation is restricted to weak-source or source-free settings. Similarly, CE1 and CE5 encode source faithfulness and source constraint. The corresponding analyses

should therefore be interpreted as descriptions of how operational choices co-occur in the literature, rather than as estimates of independent causal relationships. Fifth, some task-by-CE cells contain few papers, making individual percentages unstable. We therefore emphasize aggregate patterns.

References

- Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024a. [Do language models know when they're hallucinating references?](#) In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 912–928, St. Julian's, Malta. Association for Computational Linguistics.
- Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024b. [Can knowledge graphs reduce hallucinations in LLMs? : A survey.](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico. Association for Computational Linguistics.
- Sumukh K Aithal, Pratyush Maini, Zachary Lipton, and J. Zico Kolter. 2024. [Understanding hallucinations in diffusion models through mode interpolation.](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 134614–134644. Curran Associates, Inc.
- Eunice Akani, Benoit Favre, Frederic Bechet, and Romain Gemignani. 2023. [Reducing named entity hallucination risk to ensure faithful summary generation.](#) In *Proceedings of the 16th International Natural Language Generation Conference*, pages 437–442, Prague, Czechia. Association for Computational Linguistics.
- Spandan Anaokar, Shrey Ganatra, Harshvivek Kashid, Swapnil Bhattacharyya, Shruthi Nair, Reshma Sekhar, Siddharth Manohar, Rahul Hemrajani, and Pushpak Bhattacharyya. 2025. [HalluDetect: Detecting, mitigating, and benchmarking hallucinations in conversational systems in the legal domain.](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1822–1847, Suzhou (China). Association for Computational Linguistics.
- Jiaqi Bai, Hongcheng Guo, Zhongyuan Peng, Jian Yang, Zhoujun Li, Mohan Li, and Zhihong Tian. 2025a. [Mitigating hallucinations in large vision-language models by adaptively constraining information flow.](#) In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 23442–23450. AAAI Press.

- Tianyi Bai, Yuxuan Fan, Qiu Jiantao, Fupeng Sun, Jiayi Song, Junlin Han, Zichen Liu, Conghui He, Wentao Zhang, and Binhang Yuan. 2025b. [Hallucination at a glance: Controlled visual edits and fine-grained multimodal learning](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 135360–135393. Curran Associates, Inc.
- Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. [HalluLens: LLM hallucination benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.
- Guy Bar-Shalom, Fabrizio Frasca, Yaniv Galron, Yiftah Ziser, and Haggai Maron. 2025. [Beyond token probes: Hallucination detection via activation tensors with act-vit](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 85531–85561. Curran Associates, Inc.
- Patrice Béchard and Orlando Marquez Ayala. 2024. [Reducing hallucination in structured outputs via retrieval-augmented generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238, Mexico City, Mexico. Association for Computational Linguistics.
- Catarina G Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. [From single to multi: How LLMs hallucinate in multi-document summarization](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5291–5324, Albuquerque, New Mexico. Association for Computational Linguistics.
- Masha Belyi, Robert Friel, Shuai Shao, and Atindriyo Sanyal. 2025. [Luna: A lightweight evaluation model to catch language model hallucinations with high accuracy and low cost](#). In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 398–409, Abu Dhabi, UAE. Association for Computational Linguistics.
- Nihed Bendahman, Karen Pinel-Sauvagnat, Gilles Hubert, and Mokhtar Boumedyen Billami. 2025. [Not all hallucinations are good to throw away when it comes to legal abstractive summarization](#). 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 5331–5344, Albuquerque, New Mexico. Association for Computational Linguistics.
- Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. 2024. [Machine translation hallucination detection for low and high resource languages using large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9647–9665, Miami, Florida, USA. Association for Computational Linguistics.
- Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. 2025. [Hallucination detection in LLMs using spectral features of attention maps](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24354–24385, Suzhou, China. Association for Computational Linguistics.
- Adam Bouyamourn. 2023. [Why LLMs hallucinate, and how to get \(evidential\) closure: Perceptual, intensional, and extensional learning for faithful natural language generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3181–3193, Singapore. Association for Computational Linguistics.
- Meng Cao, Yue Dong, and Jackie Cheung. 2022. [Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.
- Qingxing Cao, Junhao Cheng, Xiaodan Liang, and Liang Lin. 2024. [VisDiaHalBench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12161–12176, Bangkok, Thailand. Association for Computational Linguistics.
- Kyubyung Chae, Jaepill Choi, Yohan Jo, and Taesup Kim. 2024. [Mitigating hallucination in abstractive summarization with domain-conditional mutual information](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1809–1820, Mexico City, Mexico. Association for Computational Linguistics.
- Trishna Chakraborty, Udit Ghosh, Xiaopan Zhang, Fahim Faisal Niloy, Yue Dong, Jiachen Li, Amit Roy-Chowdhury, and Chengyu Song. 2025. [HEAL: An empirical study on hallucinations in embodied agents driven by large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21226–21243, Suzhou, China. Association for Computational Linguistics.
- Tyler A. Chang, Katrin Tomanek, Jessica Hoffmann, Nithum Thain, Erin MacMurray van Liemt, Kathleen Meier-Hellstern, and Lucas Dixon. 2024. [Detecting hallucination and coverage errors in retrieval augmented generation for controversial topics](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4729–4743, Torino, Italia. ELRA and ICCL.
- Yurui Chang, Bochuan Cao, and Lu Lin. 2025. [Monitoring decoding: Mitigating hallucination via evaluating](#)

- the factuality of partial response during generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14574–14587, Vienna, Austria. Association for Computational Linguistics.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. [Inside: LLMs' internal states retain the power of hallucination detection](#). In *International Conference on Learning Representations*, volume 2024, pages 3056–3076.
- Guoqing Chen, Fu Zhang, Jinghao Lin, Chenglong Lu, and Jingwei Cheng. 2025a. [RRHF-V: Ranking responses to mitigate hallucinations in multimodal large language models with human feedback](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6798–6815, Abu Dhabi, UAE. Association for Computational Linguistics.
- Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, and Feng Zheng. 2025b. [Enhancing uncertainty modeling with semantic graph for hallucination detection](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 23586–23594. AAAI Press.
- Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024b. [In-context sharpness as alerts: An inner representation perspective for hallucination mitigation](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7553–7567. PMLR.
- Wei Chen, Xin Yan, Bin Wen, Fan Yang, Tingting Gao, Di ZHANG, and Long Chen. 2025c. [Decoupling contrastive decoding: Robust hallucination mitigation in multimodal large language models](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 166734–166756. Curran Associates, Inc.
- Wei-Fan Chen, Zhixue Zhao, Akbar Karimi, and Lucie Flek. 2025d. [Explainable hallucination through natural language inference mapping](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1888–1896, Vienna, Austria. Association for Computational Linguistics.
- Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lyu, Dan Zhang, and Huajun Chen. 2024c. [Factchd: Benchmarking fact-conflicting hallucination detection](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6216–6224. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024d. [Unified hallucination detection for multimodal large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252, Bangkok, Thailand. Association for Computational Linguistics.
- Xuwei Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Jianing Yang, David Fouhey, Joyce Chai, and Shengyi Qian. 2024e. [Multi-object hallucination in vision language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 44393–44418. Curran Associates, Inc.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024f. [HALC: Object hallucination reduction via adaptive focal-contrast decoding](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7824–7846. PMLR.
- Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. 2025a. [Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1272–1305, Suzhou, China. Association for Computational Linguistics.
- Xiaoxia Cheng, Zeqi Tan, Zhe Zheng, and Weiming Lu. 2025b. [G2ldetect: A global-to-local approach for hallucination detection](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 102–109. AAAI Press.
- Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025c. [Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7979–7990, Vienna, Austria. Association for Computational Linguistics.
- Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. [KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14053, Singapore. Association for Computational Linguistics.
- Prafulla Kumar Choubey, Alex Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Rajani. 2023. [CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10755–10773, Toronto, Canada. Association for Computational Linguistics.

- Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2022. [HOLM: Hallucinating objects with language models for referring expression recognition in partially-observed scenes](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5440–5453, Dublin, Ireland. Association for Computational Linguistics.
- Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. [Plausible may not be faithful: Probing object hallucination in vision-language pre-training](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2136–2148, Dubrovnik, Croatia. Association for Computational Linguistics.
- David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. [Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.
- David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023b. [HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653, Singapore. Association for Computational Linguistics.
- Jisheng Dang, Shengjun Deng, Haochen Chang, Teng Wang, Bimei Wang, Shude Wang, Nannan Zhu, Guo Niu, Jingwen Zhao, and Jizhao Liu. 2025. [Hallucination reduction in video-language models via hierarchical multimodal consistency](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 9167–9175. International Joint Conferences on Artificial Intelligence Organization. AI4Tech: AI Enabling Technologies.
- Souvik Das, Sougata Saha, and Rohini Srihari. 2022. [Diving deep into modes of fact hallucinations in dialogue systems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 684–699, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. [Chain-of-verification reduces hallucination in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.
- Yue Dong, John Wieting, and Pat Verga. 2022. [Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chengfeng Dou, Ying Zhang, Yanyuan Chen, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, Zhenwei Tao, and Yun Huang. 2024. [Detection, diagnosis, and explanation: A benchmark for chinese medical hallucination evaluation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4784–4794, Torino, Italia. ELRA and ICCL.
- Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. [Halo-scope: Harnessing unlabeled llm generations for hallucination detection](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 102948–102972. Curran Associates, Inc.
- Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. [Neural path hunter: Reducing hallucination in dialogue systems via path grounding](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. [On the origin of hallucinations in conversational models: Is it the datasets or the models?](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.
- Xinyue Fang, Zhen Huang, Zhiliang Tian, Minghui Fang, Ziyi Pan, Quntian Fang, Zhihua Wen, Hengyue Pan, and Dongsheng Li. 2025a. [Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 23868–23877. AAAI Press.
- Yi Fang, Moxin Li, Wenjie Wang, Lin Hui, and Fuli Feng. 2025b. [Counterfactual debating with preset stances for hallucination elimination of LLMs](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10554–10568, Abu Dhabi, UAE. Association for Computational Linguistics.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nature*, 630:625 – 630.
- Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. [Scene graph as pivoting](#).

- Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, Toronto, Canada. Association for Computational Linguistics.
- Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don’t hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.
- Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.
- Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. 2025. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16563–16577, Vienna, Austria. Association for Computational Linguistics.
- Yujian Gan, Yuan Liang, Jinxia Xie, Yanni Lin, Juntao Yu, and Massimo Poesio. 2025. Low-hallucination and efficient coreference resolution with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17243–17256, Suzhou, China. Association for Computational Linguistics.
- Zorik Gekhman, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.
- Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Alexander Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2025. DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10003–10039, Suzhou, China. Association for Computational Linguistics.
- Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. 2022. Visual correspondence hallucination. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024a. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111, Mexico City, Mexico. Association for Computational Linguistics.
- Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024b. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 18126–18134. AAAI Press.
- Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023a. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics.
- Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 18135–18143. AAAI Press.
- Zhengyi Ho, Siyuan Liang, Sen Zhang, Yibing Zhan, and Dacheng Tao. 2025. Novo: Norm voting off hallucinations with attention heads in large language models. In *International Conference on Learning Representations*, volume 2025, pages 88043–88066.
- Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiao-Yong Wei, and Li Qing. 2025. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853, Vienna, Austria. Association for Computational Linguistics.
- Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024a. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975, Miami, Florida, USA. Association for Computational Linguistics.
- Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. 2024b.

- Embedding and gradient say wrong: A white-box method for hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1959, Miami, Florida, USA. Association for Computational Linguistics.
- Baixiang Huang, Canyu Chen, Xiong Xiao Xu, Ali Payani, and Kai Shu. 2025a. [Can knowledge editing really correct hallucinations?](#) In *International Conference on Learning Representations*, volume 2025, pages 88116–88149.
- Chenyang Huang, Abbas Ghaddar, Ivan Kobyzev, Mehdi Rezagholizadeh, Osmar Zaiane, and Boxing Chen. 2024a. [OTTAWA: Optimal Transport adaptive word aligner for hallucination and omission translation errors detection](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6322–6334, Bangkok, Thailand. Association for Computational Linguistics.
- Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. 2025b. [Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24564–24579, Vienna, Austria. Association for Computational Linguistics.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025c. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Trans. Inf. Syst.*, 43(2).
- Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024b. [Visual hallucinations of multi-modal large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631, Bangkok, Thailand. Association for Computational Linguistics.
- Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. [Self-introspective decoding: Alleviating hallucinations for large vision-language models](#). In *International Conference on Learning Representations*, volume 2025, pages 24272–24295.
- Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. 2025. [The illusion of progress: Re-evaluating hallucination detection in LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34728–34745, Suzhou, China. Association for Computational Linguistics.
- Andrew Jesson, Nicolas Beltran-Velez, Quentin Chu, Sweta Karlekar, Jannik Kossen, Yarin Gal, John P. Cunningham, and David Blei. 2024. [Estimating the hallucination rate of generative ai](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 31154–31201. Curran Associates, Inc.
- Lanlan Ji, Dominic Seyler, Gunkirat Kaur, Manjunath Hegde, Koustuv Dasgupta, and Bing Xiang. 2025. [Phantom: A benchmark for hallucination detection in financial long-context qa](#). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. [ANAH: Analytical annotation of hallucinations in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023b. [RHO: Reducing hallucination in open-domain dialogues with knowledge grounding](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, Toronto, Canada. Association for Computational Linguistics.
- Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023c. [Towards mitigating LLM hallucination via self reflection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.
- Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024. [On large language models’ hallucination with regard to known facts](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053, Mexico City, Mexico. Association for Computational Linguistics.
- Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. 2025. [Interpreting and editing vision-language representations to mitigate hallucinations](#). In *International Conference on Learning Representations*, volume 2025, pages 63582–63605.
- Erik Jones, Hamid Palangi, Clarisse Ribeiro, Varun Chandrasekaran, Subhabrata Mukherjee, Arindam Mitra, Ahmed H Awadallah, and Ece Kamar. 2024. [Teaching language models to hallucinate less with synthetic tasks](#). In *International Conference on Learning Representations*, volume 2024, pages 52036–52055.

- Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2025. [Unfamiliar finetuning examples control how language models hallucinate](#). 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 3600–3612, Albuquerque, New Mexico. Association for Computational Linguistics.
- Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. 2025. [Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32310–32322, Suzhou, China. Association for Computational Linguistics.
- Mayank Kothiyari, Dhruva Dhingra, Sunita Sarawagi, and Soumen Chakrabarti. 2023. [CRUSH4SQL: Collective retrieval using schema hallucination for Text2SQL](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14054–14066, Singapore. Association for Computational Linguistics.
- Atharva Kulkarni, Yuan Zhang, Joel Ruben Antony Moniz, Xiou Ge, Bo-Hsiang Tseng, Dhivya Piraviprumal, Swabha Swayamdipta, and Hong Yu. 2025. [Evaluating evaluation metrics – the mirage of hallucination detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19013–19032, Suzhou, China. Association for Computational Linguistics.
- Mateusz Lango and Ondrej Dusek. 2023. [Critic-driven decoding for mitigating hallucinations in data-to-text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2853–2862, Singapore. Association for Computational Linguistics.
- Dongyub Lee, Eunhwan Park, Hodong Lee, and Heuiseok Lim. 2024a. [Ask, assess, and refine: Rectifying factual consistency and hallucination in LLMs with metric-guided feedback learning](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2422–2433, St. Julian’s, Malta. Association for Computational Linguistics.
- Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024b. [Volcano: Mitigating multimodal hallucination through self-feedback guided revision](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 391–404, Mexico City, Mexico. Association for Computational Linguistics.
- Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2025. [The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio](#). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2024a. [Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination?](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7675–7688, Mexico City, Mexico. Association for Computational Linguistics.
- Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024b. [The dawn after the dark: An empirical study on factuality hallucination in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.
- Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. [HaluEval: A large-scale hallucination evaluation benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.
- Junyi Li and Hwee Tou Ng. 2025. [Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 151064–151085. Curran Associates, Inc.
- Li Li, Jiashu Qu, Linxin Song, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. 2025a. [Treble counterfactual VLMs: A causal approach to hallucination](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18423–18434, Suzhou, China. Association for Computational Linguistics.
- Miaoran Li, Jiangning Chen, Minghua Xu, and Xiaolong Wang. 2025b. [Hallucination detection in structured query generation via LLM self-debating](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16102–16113, Suzhou, China. Association for Computational Linguistics.
- Qiang Li, Mingkun Tan, Xun Zhao, Dan Zhang, Daoan Zhang, Shengzhao Lei, Anderson S. Chu, Lujun Li, and Porawit Kamnoedboon. 2025c. [How LLMs react to industrial spatio-temporal data? assessing hallucination with a novel traffic incident benchmark dataset](#). 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 36–53, Albuquerque, New Mexico. Association for Computational Linguistics.

- Shuo Li, Jiajun Sun, Guodong Zheng, Xiaoran Fan, Yujiong Shen, Yi Lu, Zhiheng Xi, Yuming Yang, Wenming Tan, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025d. [Mitigating object hallucinations in MLLMs via multi-frequency perturbations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1230–1247, Suzhou, China. Association for Computational Linguistics.
- Tianyi Li, Erenay Dayanik, Shubhi Tyagi, and Andrea Pierleoni. 2025e. [HALLUCANA: Fixing LLM hallucination with a canary lookahead](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 213–230, Albuquerque, New Mexico. Association for Computational Linguistics.
- Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. [Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods](#). *Preprint*, arXiv:2203.05227.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenying Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. 2025f. [The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 35799–35819. PMLR.
- Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Cheng Peng, Zhonghao Wang, and Haiying Deng. 2024. [UHGEval: Benchmarking the hallucination of Chinese large language models via unconstrained generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5266–5293, Bangkok, Thailand. Association for Computational Linguistics.
- Youngsun Lim, Hoon Choi, and Hyun Jung Shim. 2025. [Evaluating image hallucination in text-to-image generation with question-answering](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 26290–26298. AAAI Press.
- Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. 2025a. [Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28568–28584, Suzhou, China. Association for Computational Linguistics.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. [Mitigating hallucination in large multi-modal models via robust instruction tuning](#). In *International Conference on Learning Representations*, volume 2024, pages 57689–57733.
- Hui Liu and Xiaojun Wan. 2023. [Models see hallucinations: Evaluating the factuality in video captioning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11807–11823, Singapore. Association for Computational Linguistics.
- Lingyuan Liu and Mengxiang Zhang. 2025. [GOLFer: Smaller LMs-generated documents hallucination filter & combiner for query expansion in information retrieval](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 153–162, Vienna, Austria. Association for Computational Linguistics.
- Sheng Liu, Haotian Ye, and James Y Zou. 2025b. [Reducing hallucinations in large vision-language models via latent space steering](#). In *International Conference on Learning Representations*, volume 2025, pages 72402–72419.
- Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yasmine Benajiba, and Dan Roth. 2025c. [Towards long context hallucination detection](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7842–7850, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. [A token-level reference-free hallucination detection benchmark for free-form text generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.
- Xu Liu, Guanyi Chen, Kees van Deemter, and Tingting He. 2025d. [Annotating hallucinations in question-answering using rewriting](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 823–832, Hanoi, Vietnam. Association for Computational Linguistics.
- Zihang Liu, Jiawei Guo, Hao Zhang, Hongyang Chen, Jiajun Bu, and Haishuai Wang. 2025e. [Long-form hallucination detection with self-elicitation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4082–4100, Vienna, Austria. Association for Computational Linguistics.
- Rui Lu, Runzhe Wang, Kaifeng Lyu, Xitai Jiang, Gao Huang, and Mengdi Wang. 2025a. [Towards understanding text hallucination of diffusion models via local generation bias](#). In *International Conference on Learning Representations*, volume 2025, pages 24192–24219.

- Yifan Lu, Ziqi Zhang, Chunfeng Yuan, Jun Gao, Congxuan Zhang, Xiaojuan Qi, Bing Li, and Weiming Hu. 2025b. [Mitigating hallucinations in large vision-language models by self-injecting hallucinations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13861–13877, Suzhou, China. Association for Computational Linguistics.
- Qitan Lv, Jie Wang, Hanzhu Chen, Bin Li, Yongdong Zhang, and Feng Wu. 2024. [Coarse-to-fine highlighting: Reducing knowledge hallucination in large language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33594–33623. PMLR.
- Xinyu Lyu, Beita Chen, Lianli Gao, Jingquan Song, and Heng Tao Shen. 2024. [Alleviating hallucinations in large vision-language models through hallucination-induced optimization](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 122811–122832. Curran Associates, Inc.
- Kishan Maharaj, Ashita Saxena, Raja Kumar, Abhijit Mishra, and Pushpak Bhattacharyya. 2023. [Eyes show the way: Modelling gaze behaviour for hallucination detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11424–11438, Singapore. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Laura Mascarell, Ribin Chalumattu, and Annette Rios. 2024. [German also hallucinates! inconsistency detection in news summaries with the absinth dataset](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7696–7706, Torino, Italia. ELRA and ICCL.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. [Sources of hallucination by large language models on inference tasks](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.
- Shahrad Mohammadzadeh, Juan David Guerra, Marco Bonizzato, Reihaneh Rabbany, and Golnoosh Farnadi. 2025. [Hallucination detox: Sensitivity dropout \(SenD\) for large language model training](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5538–5554, Vienna, Austria. Association for Computational Linguistics.
- Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. [Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015, Torino, Italia. ELRA and ICCL.
- Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. [Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation](#). In *International Conference on Learning Representations*, volume 2024, pages 40364–40393.
- Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. [A simple recipe towards reducing hallucination in neural surface realisation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. [RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.
- Mengjia Niu, Hamed Haddadi, and Guansong Pang. 2025. [Robust hallucination detection in llms via adaptive token selection](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 126355–126377. Curran Associates, Inc.
- Saad Obaid ul Islam, Iza Škrjanec, Ondrej Dusek, and Vera Demberg. 2023. [Tackling hallucinations in neural chart summarization](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 414–423, Prague, Czechia. Association for Computational Linguistics.
- Changwon Ok, Eunhyeong Lee, and Dongsuk Oh. 2025. [Synthetic paths to integral truth: Mitigating hallucinations caused by confirmation bias with synthetic data](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5168–5180, Abu Dhabi, UAE. Association for Computational Linguistics.
- OpenAI. 2026. [Gpt-5.6 system card](#).
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. [Llms know more than they show: On the intrinsic representation of llm hallucinations](#).

- In *International Conference on Learning Representations*, volume 2025, pages 66880–66913.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. [Med-HALT: Medical domain hallucination test for large language models](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.
- Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. 2025a. [Steer LLM latents for hallucination detection](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 47971–47990. PMLR.
- Seongheon Park and Sharon Li. 2025. [Glsim: Detecting object hallucinations in llms via global-local similarity](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 37747–37776. Curran Associates, Inc.
- Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. 2025b. [SECOND: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48027–48040. PMLR.
- Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. 2025c. [Convvis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 6434–6442. AAAI Press.
- Suzanne Petryk, David M. Chan, Anish Kachinhaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. 2024. [ALOHa: A new measure for hallucination in captioning models](#). 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 342–357, Mexico City, Mexico. Association for Computational Linguistics.
- Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. [mmT5: Modular multilingual pre-training solves source language hallucinations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1978–2008, Singapore. Association for Computational Linguistics.
- Siya Qi, Rui Cao, Yulan He, and Zheng Yuan. 2025. [Evaluating LLMs’ assessment of mixed-context hallucination through the lens of summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16480–16503, Vienna, Austria. Association for Computational Linguistics.
- Jiaye Qian, Ge Zheng, Yuchen Zhu, and Sibe Yang. 2025. [Intervene-all-paths: Unified mitigation of llm hallucinations across alignment formats](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 142168–142198. Curran Associates, Inc.
- Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2023. [Detecting and mitigating hallucinations in multilingual summarisation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8932, Singapore. Association for Computational Linguistics.
- Xiaoye Qu, Jiashuo Sun, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. 2025. [Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4428–4441, Abu Dhabi, UAE. Association for Computational Linguistics.
- Anil Ramakrishna, Rahul Gupta, Jens Lehmann, and Morteza Ziyadi. 2023. [INVITE: a testbed of automatically generated invalid questions to evaluate large language models for hallucinations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5422–5429, Singapore. Association for Computational Linguistics.
- Sanjana Ramprasad, Elisa Ferracane, and Zachary Lip-ton. 2024. [Analyzing LLM behavior in dialogue summarization: Unveiling circumstantial hallucination trends](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12549–12561, Bangkok, Thailand. Association for Computational Linguistics.
- Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. [The curious case of hallucinations in neural machine translation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.
- Abhilasha Ravichander, Shruti Ghela, David Wadden, and Yejin Choi. 2025. [HALoGEN: Fantastic LLM hallucinations and where to find them](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1402–1425, Vienna, Austria. Association for Computational Linguistics.
- Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. [The troubling emergence of hallucination in large language models - an extensive definition, quantification,](#)

- and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. [Object hallucination in image captioning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.
- Candace Ross, Florian Bordes, Adina Williams, Polina Kirichenko, and Mark Ibrahim. 2025. [What’s in common? multimodal models hallucinate when reasoning across scenes](#). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- Elisei Rykov, Kseniia Petrushina, Maksim Savkin, Valerii Olisov, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. [When models lie, we learn: Multilingual span-level hallucination detection with PsiloQA](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11663–11682, Suzhou, China. Association for Computational Linguistics.
- Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. [DelusionQA: Detecting hallucinations in domain-specific question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.
- Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. 2024. [Mitigating hallucination in fictional character role-play](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14467–14479, Miami, Florida, USA. Association for Computational Linguistics.
- Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. [A comprehensive survey of hallucination in large language, image, video and audio foundation models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.
- Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 2025. [Do my eyes deceive me? a survey of human evaluations of hallucinations in NLG](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79, Hanoi, Vietnam. Association for Computational Linguistics.
- Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. [WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2387–2413, Singapore. Association for Computational Linguistics.
- Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. [Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33, St. Julian’s, Malta. Association for Computational Linguistics.
- Jaehyung Seo, Hyeonseok Moon, and Heuiseok Lim. 2025. [The impact of negated text on hallucination with large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13554–13572, Suzhou, China. Association for Computational Linguistics.
- Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. 2025. [EGOILLUSION: Benchmarking hallucinations in egocentric video understanding](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28461–28480, Suzhou, China. Association for Computational Linguistics.
- Jianbin Shen, Junyu Xuan, and Christy Liang. 2023. [Mitigating intrinsic named entity-related hallucinations of abstractive text summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15807–15824, Singapore. 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.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. [Trusting your evidence: Hallucinate less with context-aware decoding](#). 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 783–791, Mexico City, Mexico. Association for Computational Linguistics.
- Xiao Shi, Zhengyuan Zhu, Zeyu Zhang, and Chengkai Li. 2023. [Hallucination mitigation in natural language generation from large-scale open-domain knowledge graphs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12506–12521, Singapore. Association for Computational Linguistics.

- Yan Shu, Hangui Lin, Yexin Liu, Yan Zhang, Gangyan Zeng, Yan Li, Yu Zhou, Ser Nam Lim, Harry Yang, and Nicu Sebe. 2025. [When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 68952–68980. Curran Associates, Inc.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shuzheng Si, Haozhe Zhao, Gang Chen, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Kaikai An, Kangyang Luo, Chen Qian, Fanchao Qi, Baobao Chang, and Maosong Sun. 2025. [Aligning large language models to follow instructions and hallucinate less via effective data filtering](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16469–16488, Vienna, Austria. Association for Computational Linguistics.
- Serry Taiseer Sibae, Abdullah I. Alharbi, Samar Ahmed, Omar Nacar, Lahouri Ghouti, and Anis Koubaa. 2024. [Asos at arabic llms hallucinations 2024: Can llms detect their hallucinations :\)](#). In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 130–134, Torino, Italy. ELRA and ICCL.
- Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. [Trust me, I’m wrong: LLMs hallucinate with certainty despite knowing the answer](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14665–14688, Suzhou, China. Association for Computational Linguistics.
- Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. [The curious case of hallucinatory \(un\)answerability: Finding truths in the hidden states of over-confident large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.
- Ponhvoan Srey, Xiaobao Wu, and Anh Tuan Luu. 2025. [Unsupervised hallucination detection by inspecting reasoning processes](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22117–22129, Suzhou, China. Association for Computational Linguistics.
- Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. [Llm-check: Investigating detection of hallucinations in large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.
- Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. [Unsupervised real-time hallucination detection based on the internal states of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.
- Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. [Confabulation: The surprising value of large language model hallucinations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.
- Bin Sun, Yitong Li, Fei Mi, Fanhu Bie, Yiwei Li, and Kan Li. 2023a. [Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1741–1750, Toronto, Canada. Association for Computational Linguistics.
- Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023b. [Contrastive learning reduces hallucination in conversations](#). In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13618–13626. AAAI Press.
- Xiaoxi Sun, Jianxin Liang, Yueqian Wang, Huishuai Zhang, and Dongyan Zhao. 2025a. [Understanding visual detail hallucinations of large vision-language models](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 1900–1908. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Yiyi Sun, Yu Gai, Lijie Chen, Abhilasha Ravichander, Yejin Choi, Nouha Dziri, and Dawn Song. 2025b. [Why and how llms hallucinate: Connecting the dots with subsequence associations](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 35159–35203. Curran Associates, Inc.
- YuHong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Hui Zhao. 2024. [Benchmarking hallucination in large language models based on unanswerable math word problem](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2178–2188, Torino, Italia. ELRA and ICCL.

- Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025c. [Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability](#). In *International Conference on Learning Representations*, volume 2025, pages 50250–50279.
- Praneet Suresh, Jack Stanley, Sonia Joseph, Luca Scimeca, and Danilo Bzdok. 2025. [From noise to narrative: Tracing the origins of hallucinations in transformers](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 22465–22502. Curran Associates, Inc.
- Liyan Tang, Igor Shalymov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. [TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.
- Zilu Tang, Rajen Chatterjee, and Sarthak Garg. 2025. [Mitigating hallucinated translations in large language models with hallucination-focused preference optimization](#). 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 3410–3433, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yuchen Tian, Weixiang Yan, Qian Yang, Xuandong Zhao, Qian Chen, Wen Wang, Ziyang Luo, Lei Ma, and Dawn Song. 2025. [Codehalu: Investigating code hallucinations in llms via execution-based verification](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 25300–25308. AAAI Press.
- Saad Obaid Ul Islam, Anne Lauscher, and Goran Glavač. 2025. [How much do LLMs hallucinate across languages? on realistic multilingual estimation of LLM hallucination](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29077–29098, Suzhou, China. Association for Computational Linguistics.
- Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. [Mutual information alleviates hallucinations in abstractive summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. [An audit on the perspectives and challenges of hallucinations in NLP](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6528–6548, Miami, Florida, USA. Association for Computational Linguistics.
- Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. [Contrastive decoding reduces hallucinations in large multilingual machine translation models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539, St. Julian's, Malta. Association for Computational Linguistics.
- Chaojun Wang and Rico Sennrich. 2020. [On exposure bias, hallucination and domain shift in neural machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.
- Hongbo Wang, Jie Cao, Jin Liu, Xiaoqiang Zhou, Huaibo Huang, and Ran He. 2024. [Hallo3d: Multimodal hallucination detection and mitigation for consistent 3d content generation](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 118883–118906. Curran Associates, Inc.
- Jiapeng Wang, Chengyu Wang, Jun Huang, and Lianwen Jin. 2025a. [Hallucination-aware prompt optimization for text-to-video synthesis](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 10198–10206. International Joint Conferences on Artificial Intelligence Organization. AI, Arts & Creativity.
- Song Wang, Xun Wang, Jie Mei, Yujia Xie, Si-Qing Chen, and Wayne Xiong. 2025b. [Developing a reliable, fast, general-purpose hallucination detection and mitigation service](#). 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 971–978, Albuquerque, New Mexico. Association for Computational Linguistics.
- Song Wang, Xun Wang, Jie Mei, Yujia Xie, Si-Qing Chen, and Wayne Xiong. 2025c. [Image token matters: Mitigating hallucination in discrete tokenizer-based large vision-language models via latent editing](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 111306–111335. Curran Associates, Inc.
- Wei Xing Wang, Zifeng Ding, Jindong Gu, RUI CAO, Christoph Meinel, Gerard de Melo, and Haojin Yang. 2025c. [Image token matters: Mitigating hallucination in discrete tokenizer-based large vision-language models via latent editing](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 111306–111335. Curran Associates, Inc.
- Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng, and Xuanjing Huang. 2023. [Hallucination detection for generative large language models by Bayesian sequential estimation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15361–15371, Singapore. Association for Computational Linguistics.
- Hongliang Wei, Xingtao Wang, Xianqi Zhang, Xiaopeng Fan, and Debin Zhao. 2024. [Toward a stable,](#)

- fair, and comprehensive evaluation of object hallucination in large vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 111406–111431. Curran Associates, Inc.
- Zheng Weihua, Roy Ka-Wei Lee, Zhengyuan Liu, Wu Kui, AiTi Aw, and Bowei Zou. 2025. [CCL-XCoT: An efficient cross-lingual knowledge transfer method for mitigating hallucination generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1768–1788, Suzhou, China. Association for Computational Linguistics.
- Sangmin Woo, Kang Zhou, Yun Zhou, Shuai Wang, Sheng Guan, Haibo Ding, and Lin Lee Cheong. 2025. [Black-box visual prompt engineering for mitigating object hallucination in large vision language models](#). 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 529–538, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jiarui Wu, Zhuo Liu, and Hangfeng He. 2025a. [Mitigating hallucinations in multimodal spatial relations through constraint-aware prompting](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3450–3468, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jiulong Wu, Zhengliang Shi, Shuaiqiang Wang, Jizhou Huang, Dawei Yin, Lingyong Yan, Min Cao, and Min Zhang. 2025b. [Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19445–19461, Suzhou, China. Association for Computational Linguistics.
- Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2025c. [SHARP: Steering hallucination in LVLMS via representation engineering](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14346–14361, Suzhou, China. Association for Computational Linguistics.
- Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024a. [Logical closed loop: Uncovering object hallucinations in large vision-language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6944–6962, Bangkok, Thailand. Association for Computational Linguistics.
- Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024b. [Evaluating and analyzing relationship hallucinations in large vision-language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53553–53570. PMLR.
- Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025d. [Combating multimodal LLM hallucination via bottom-up holistic reasoning](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 8460–8468. AAAI Press.
- Yu Xia, Xu Liu, Tong Yu, Sungchul Kim, Ryan Rossi, Anup Rao, Tung Mai, and Shuai Li. 2024. [Hallucination diversity-aware active learning for text summarization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8665–8677, Mexico City, Mexico. Association for Computational Linguistics.
- Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. [Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 25543–25551. AAAI Press.
- Yijun Xiao and William Yang Wang. 2021. [On hallucination and predictive uncertainty in conditional language generation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.
- Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2024a. [EFUF: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1181, Miami, Florida, USA. Association for Computational Linguistics.
- Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024b. [Mitigating object hallucination via concentric causal attention](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 92012–92035. Curran Associates, Inc.
- Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2025. [Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*,

- pages 7849–7857. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Hongshen Xu, Zichen Zhu, Lei Pan, Zihan Wang, Su Zhu, Da Ma, Ruisheng Cao, Lu Chen, and Kai Yu. 2025. [Reducing tool hallucination via reliability alignment](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 69992–70006. PMLR.
- Silei Xu, Shicheng Liu, Theo Culhane, Elizaveta Pertseva, Meng-Hsi Wu, Sina Semnani, and Monica Lam. 2023a. [Fine-tuned LLMs know more, hallucinate less with few-shot sequence-to-sequence semantic parsing over Wikidata](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5778–5791, Singapore. Association for Computational Linguistics.
- Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023b. [Understanding and detecting hallucinations in neural machine translation via model introspection](#). *Transactions of the Association for Computational Linguistics*, 11:546–564.
- Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. [A new benchmark and reverse validation method for passage-level hallucination detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3898–3908, Singapore. Association for Computational Linguistics.
- Sibei Yang, Ge Zheng, Jiajin Tang, Jiaye Qian, Hanzhuo Huang, and Cheng Shi. 2025a. [Discovering compositional hallucinations in llms](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 28846–28876. Curran Associates, Inc.
- Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025b. [Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention](#). In *International Conference on Learning Representations*, volume 2025, pages 51546–51568.
- Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. [InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347, Bangkok, Thailand. Association for Computational Linguistics.
- Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Chang Yoo. 2022. [Information-theoretic text hallucination reduction for video-grounded dialogue](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4182–4193, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024a. [Mechanistic understanding and mitigation of language model non-factual hallucinations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA. Association for Computational Linguistics.
- Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. 2024b. [ReEval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1333–1351, Mexico City, Mexico. Association for Computational Linguistics.
- Zihao Yue, Liang Zhang, and Qin Jin. 2024. [Less is more: Mitigating multimodal hallucination from an EOS decision perspective](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11766–11781, Bangkok, Thailand. Association for Computational Linguistics.
- Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q. Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. 2025a. [Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models](#). In *International Conference on Learning Representations*, volume 2025, pages 5494–5524.
- Chaowei Zhang, Zongling Feng, Zewei Zhang, Jipeng Qiang, Guandong Xu, and Yun Li. 2025b. [Is llms hallucination usable? llm-based negative reasoning for fake news detection](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 1031–1039. AAAI Press.
- Dongxu Zhang, Varun Gangal, Barrett Lattimer, and Yi Yang. 2024a. [Enhancing hallucination detection through perturbation-based synthetic data generation in system responses](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13321–13332, Bangkok, Thailand. Association for Computational Linguistics.
- Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023a. [SAC³: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.
- Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. 2025c. [Detecting hallucination in large language models through deep internal representation analysis](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 8357–8365. International

- Joint Conferences on Artificial Intelligence Organization. Main Track.
- Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024b. [How language model hallucinations can snowball](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59670–59684. PMLR.
- Shaolei Zhang, Tian Yu, and Yang Feng. 2024c. [TruthX: Alleviating hallucinations by editing large language models in truthful space](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949, Bangkok, Thailand. Association for Computational Linguistics.
- Siyuan Zhang, Yichi Zhang, Yinpeng Dong, and Hang Su. 2025d. [Exploring the generalizability of factual hallucination mitigation via enhancing precise knowledge utilization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3936–3968, Suzhou, China. Association for Computational Linguistics.
- Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. [Enhancing uncertainty-based hallucination detection with stronger focus](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.
- Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024d. [Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965, Bangkok, Thailand. Association for Computational Linguistics.
- Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2025e. [Alleviating hallucinations of large language models through induced hallucinations](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8233–8247, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, Tetsuya Sakai, Tian Feng, and Hayato Yamana. 2024e. [ToolBeHonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11388–11422, Miami, Florida, USA. Association for Computational Linguistics.
- Zihao Zhang, Zhong Qian, Xiaoxu Zhu, Peifeng Li, and Qiaoming Zhu. 2025f. [Trucidator: Document-level event factuality identification via hallucination enhancement and cross-document inference](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2038–2048, Abu Dhabi, UAE. Association for Computational Linguistics.
- Harry Zhao, Tristan Sylvain, Romain Laroche, Doina Precup, and Yoshua Bengio. 2025a. [Rejecting hallucinated state targets during planning](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 77677–77702. PMLR.
- Lingjun Zhao, Khanh Nguyen, and Hal Daumé III. 2023. [Hallucination detection for grounded instruction generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4044–4053, Singapore. Association for Computational Linguistics.
- Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2025b. [Mitigating object hallucination in large vision-language models via image-grounded guidance](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 77461–77486. PMLR.
- Xinping Zhao, Jindi Yu, Zhenyu Liu, Jifang Wang, Dongfang Li, Yibin Chen, Baotian Hu, and Min Zhang. 2024. [Medico: Towards hallucination detection and correction with multi-source evidence fusion](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 34–45, Miami, Florida, USA. Association for Computational Linguistics.
- Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. [Reducing quantity hallucinations in abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.
- Meizhi Zhong, Kehai Chen, Zhengshan Xue, Lemao Liu, Mingming Yang, and Min Zhang. 2024a. [On the hallucination in simultaneous machine translation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 730–742, Bangkok, Thailand. Association for Computational Linguistics.
- Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024b. [Investigating and mitigating the multimodal hallucination snowballing in large vision-language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11991–12011, Bangkok, Thailand. Association for Computational Linguistics.
- Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan

- Ghazvininejad. 2021. [Detecting hallucinated content in conditional neural sequence generation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.
- Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. [Analyzing and mitigating object hallucination in large vision-language models](#). In *International Conference on Learning Representations*, volume 2024, pages 56969–56998.
- Chenyu Zhu, Yefeng Liu, Hao Zhang, Aowen Wang, Yangxue, Guanhua Chen, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025a. [Alleviating hallucinations in large language models through multi-model contrastive decoding and dynamic hallucination detection](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 165364–165388. Curran Associates, Inc.
- Derui Zhu, Dingfan Chen, Qing Li, Zongxiong Chen, Lei Ma, Jens Grossklags, and Mario Fritz. 2024. [PoLLMgraph: Unraveling hallucinations in large language models via state transition dynamics](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4737–4751, Mexico City, Mexico. Association for Computational Linguistics.
- Jia Zhu, Hanghui Guo, Weijie Shi, Zhangze Chen, and Pasquale De Meo. 2025b. [Radio: Real-time hallucination detection with contextual index optimized query formulation for dynamic retrieval augmented generation](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 26129–26137. AAAI Press.
- Yanxu Zhu, Jinlin Xiao, Yuhang Wang, and Jitao Sang. 2025c. [KG-FPQ: Evaluating factuality hallucination in LLMs with knowledge graph-based false premise questions](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10472–10490, Abu Dhabi, UAE. Association for Computational Linguistics.
- Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025d. [Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.
- Ilana Zimmerman, Jadin Tredup, Ethan Selfridge, and Joseph Bradley. 2024. [Two-tiered encoder-based hallucination detection for retrieval-augmented generation in the wild](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 8–22, Miami, Florida, US. Association for Computational Linguistics.
- Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Ken-ling Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. 2025. [Look twice before you answer: Memory-space visual re-tracing for hallucination mitigation in multimodal large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80873–80899. PMLR.

A Review Protocol

This appendix provides additional details on the corpus construction, eligibility criteria, screening procedure, and composition of the final review corpus. The protocol was designed to identify papers that explicitly treated *hallucination* as a primary research topic in the context of LLMs or related generative models.

A.1 Search Scope and Collection Procedure

We collected papers published between 2016 and 2025 from the following AI and NLP venues. The final search was conducted on June 20, 2026. Accordingly, the number of papers reported for 2025 reflects publications available as of this cut-off date. The primary search criterion required the paper title to contain the term *hallucination* or a morphological variant, including *hallucinations*, *hallucinated*, and *hallucinating*. This title-based criterion was adopted to identify papers that explicitly positioned hallucination as a central research topic, rather than papers that mentioned the term only incidentally.

Bibliographic records were collected from official proceedings pages, conference websites, publisher databases, and the ACL Anthology where applicable. For each record, we stored the title, authors, publication year, venue, abstract, PDF URL, and other available metadata. Duplicate records were identified using normalized titles and publication metadata and were removed before screening.

A.2 Eligibility Criteria

Table 2 summarizes the inclusion and exclusion criteria. Papers were included when they explicitly examined hallucination as an output-side error, inconsistency, reliability problem, evaluation target, or related phenomenon in an LLM or generative-model setting.

A.3 Screening and Selection

Screening was conducted at the record and full-text levels. First, titles and metadata were checked for duplicates, venue eligibility, publication year, and document type. Second, the available abstract and full text were examined to determine whether the paper addressed hallucination as a generative-output phenomenon. Papers using the term in an unrelated technical sense were excluded even when the title satisfied the initial keyword criterion. When eligibility could not be determined from the title and abstract alone, the full text was reviewed.

Borderline cases were retained only when the paper provided sufficient evidence that hallucination was treated as an output-side phenomenon, evaluation target, or reliability concern. The resulting corpus contained 244 papers.

A.4 Corpus Composition

Table 3 reports the composition of the final corpus by venue. The paper-level inventory can be derived from Table 4 and the accompanying references.

B Coding Codebook

This appendix presents the coding rules used in Phase 2 of the review. The codebook was developed through the exploratory review and was then applied consistently to the full corpus. The unit of analysis was the individual paper.

The codebook comprised four primary components: (1) definitional clarity, (2) output-side hallucination symptoms, (3) Convergence Expectation (CE), and (4) attitudinal or evaluative framing. We additionally recorded the primary task, model-internal indicators, uncertainty flags, and notes for human review. All coding decisions were required to be supported by verbatim evidence from the target paper. Evidence from external sources or general domain knowledge was not used. If evidence was not found, it was recorded as NOT FOUND. When the available evidence did not clearly distinguish between adjacent levels or scores, the lower level or score was assigned.

B.1 Definitional Clarity

Definitional clarity captures the extent to which a paper specifies how hallucination is identified, labeled, measured, or otherwise judged. It does not evaluate whether the paper’s definition is theoretically correct or consistent with definitions in other studies. The scoring output used integer values, with 0, 1, and 2 corresponding to Lv0, Lv1, and Lv2, respectively.

Lv1–Lv2 boundary The primary distinction between Lv1 and Lv2 was whether the available evidence showed how hallucination was judged. A statement such as “hallucinations are unsupported outputs” was coded as Lv1 unless the paper also specified how support was determined. Annotation questions, labeling rules, metric definitions, verification procedures, or explicit benchmark criteria were sufficient for Lv2.

Table 2: Eligibility criteria used to construct the review corpus.

Criterion type	Operational rule	Rationale
Title	The title contains <i>hallucination</i> or a morphological variant, such as <i>hallucinations</i> , <i>hallucinated</i> , or <i>hallucinating</i> .	Restricts the corpus to papers that explicitly treat hallucination as a primary topic rather than an incidental concern.
Publication period	The paper was published between 2016 and 2025.	Covers the period in which hallucination became an explicit research topic in neural generation and subsequent LLM research.
Venue	The paper appeared in one of the selected AI or NLP venues: AAAI, ACL, COLING, CoNLL, EACL, EMNLP, ICLR, ICML, IJCAI, INLG, NAACL, or NeurIPS.	Provides a transparent and reproducible venue boundary while covering major NLP, machine learning, and AI communities.
Document type	Full research papers, short papers, Findings papers, industry-track papers, dataset or benchmark papers, and eligible survey papers were included when they met the other criteria.	Allows analysis of both empirical research and papers that explicitly operationalize hallucination through taxonomies or evaluation frameworks.
Exclusion: unrelated technical usage	Papers using <i>hallucination</i> to refer to image reconstruction, super-resolution, feature generation, domain adaptation, or another technical operation unrelated to erroneous generative outputs were excluded.	Prevents conflation between generative-output hallucination and historically distinct uses of the term in computer vision and related fields.
Exclusion: unavailable evidence	Papers were excluded when the full text could not be accessed or when the available document did not permit extraction of sufficient evidence for paper-level coding.	The evidence-first coding procedure required access to verifiable textual evidence.
Unit of analysis	The unit of analysis was the individual paper.	Supports paper-level comparison of definitions, task settings, symptoms, and Convergence Expectation.

Table 3: Composition of the final review corpus by venue.

Venue	Papers	%
AAAI	14	5.7
ACL	47	19.3
COLING	13	5.3
CoNLL	1	0.4
EACL	7	2.9
EMNLP	73	29.9
ICLR	16	6.6
ICML	12	4.9
IJCAI	6	2.5
INLG	4	1.6
NAACL	26	10.7
NeurIPS	25	10.2
Total	244	100.0

Conservative coding rule If the evidence was ambiguous between two adjacent levels, the lower level was assigned. Examples alone were not sufficient for Lv2 unless they were linked to a generalizable judgement or labeling procedure.

B.2 Output-side Symptom Taxonomy

The symptom taxonomy captures output-side phenomena that each paper explicitly treated as hallucination. A symptom was assigned only when the paper linked the phenomenon to hallucination

through a definition, annotation scheme, evaluation procedure, metric, example, experimental setup, or benchmark-construction rule.

Symptom coding was multi-label. More than one symptom could be assigned to a paper when multiple output phenomena were explicitly included in the paper’s operationalization. If the evidence was uncertain, the symptom was coded as absent.

Explicit linkage requirement A symptom was coded as present only when the paper explicitly linked the output phenomenon to hallucination through its definition, annotation, evaluation, metric, examples, experimental setup, or benchmark construction. A phenomenon was not inferred from general task knowledge.

Multiple symptom labels The symptom categories were not mutually exclusive. For example, a paper could receive both Transformation and Addition when its operational criteria included both source contradiction and unsupported added content.

Source-grounded tasks For machine translation and other source-grounded tasks, outputs described as detached, unrelated, or unsupported were not coded as Deviation. Such cases were coded as Ad-

Table 4: List of reviewed papers by venue. Findings, industry-track, and jointly organized proceedings were mapped to their corresponding parent venue for corpus-level reporting. (e.g., LREC-COLING 2024 papers were reported under COLING).

Venue	Reviewed papers
AAAI	Sun et al. (2023b); Gunjal et al. (2024); Guan et al. (2024b); Tian et al. (2025); Wu et al. (2025d); Xiao et al. (2025); Chen et al. (2025b); Lim et al. (2025); Bai et al. (2025a); Zhu et al. (2025b); Fang et al. (2025a); Park et al. (2025c); Cheng et al. (2025b); Zhang et al. (2025b)
ACL	Nie et al. (2019); Wang and Sennrich (2020); Zhou et al. (2021); Cao et al. (2022); Liu et al. (2022); Cirik et al. (2022); Dale et al. (2023a); Guerreiro et al. (2023a); Sun et al. (2023a); Ji et al. (2023b); Choubey et al. (2023); Fei et al. (2023); Chen et al. (2024d); Ji et al. (2024); Dhuliawala et al. (2024); Li et al. (2024b); Zhang et al. (2024a); Zhong et al. (2024a); Yehuda et al. (2024); Zhong et al. (2024b); Wu et al. (2024a); Huang et al. (2024a); Niu et al. (2024); Su et al. (2024); Huang et al. (2024b); Sui et al. (2024); Cao et al. (2024); Ramprasad et al. (2024); Feng et al. (2024); Zhang et al. (2024c,d); Liang et al. (2024); Yue et al. (2024); Si et al. (2025); Huang et al. (2025b); Zhu et al. (2025d); Qi et al. (2025); Chen et al. (2025d); Liu and Zhang (2025); Mohammadzadeh et al. (2025); Bang et al. (2025); Ravichander et al. (2025); Liu et al. (2025e); Fu et al. (2025); Chang et al. (2025); Hu et al. (2025); Cheng et al. (2025c)
COLING	Sibae et al. (2024); Sun et al. (2024); Chang et al. (2024); Dou et al. (2024); Mascarell et al. (2024); Mubarak et al. (2024); Fang et al. (2025b); Zhu et al. (2025c); Qu et al. (2025); Belyi et al. (2025); Chen et al. (2025a); Ok et al. (2025); Zhang et al. (2025f)
CoNLL	Pal et al. (2023)
EACL	Xiao and Wang (2021); Dai et al. (2023); Guerreiro et al. (2023b); Agrawal et al. (2024a); Lee et al. (2024a); Sennrich et al. (2024); Waldendorf et al. (2024)
EMNLP	Rohrbach et al. (2018); Filippova (2020); Zhao et al. (2020); Dziri et al. (2021); Shuster et al. (2021); Das et al. (2022); Yoon et al. (2022); Dong et al. (2022); van der Poel et al. (2022); Li et al. (2023a); Manakul et al. (2023); Shi et al. (2023); Wang et al. (2023); Zhao et al. (2023); Ji et al. (2023c); McKenna et al. (2023); Sadat et al. (2023); Bouyamoun (2023); Ramakrishna et al. (2023); Slobodkin et al. (2023); Dale et al. (2023b); Liu and Wan (2023); Shen et al. (2023); Zhang et al. (2023a); Pfeiffer et al. (2023); Maharaj et al. (2023); Lango and Dusek (2023); Kothiyari et al. (2023); Choi et al. (2023); Yang et al. (2023); Xu et al. (2023a); Semnani et al. (2023); Zhang et al. (2023b); Rawte et al. (2023); Li et al. (2023b); Qiu et al. (2023); Venkit et al. (2024); Sahoo et al. (2024); Gekhman et al. (2024); Xing et al. (2024a); Hu et al. (2024b,a); Benkirane et al. (2024); Yu et al. (2024a); Zhao et al. (2024); Sadeq et al. (2024); Zhang et al. (2024e); Zimmerman et al. (2024); Weihua et al. (2025); Cheng et al. (2025a); Gema et al. (2025); Kim et al. (2025); Seth et al. (2025); Kulikarni et al. (2025); Zhang et al. (2025d); Binkowski et al. (2025); Li et al. (2025b); Anaokar et al. (2025); Chakraborty et al. (2025); Ul Islam et al. (2025); Janiak et al. (2025); Seo et al. (2025); Gan et al. (2025); Wu et al. (2025b); Li et al. (2025d); Lu et al. (2025b); Liu et al. (2025a); Sheng et al. (2025); Wu et al. (2025c); Li et al. (2025a); Simhi et al. (2025); Srey et al. (2025); Rykov et al. (2025)
ICLR	Germain et al. (2022); Zhou et al. (2024); Chen et al. (2024a); Liu et al. (2024); Mündler et al. (2024); Jones et al. (2024); Huang et al. (2025a); Jiang et al. (2025); Orgad et al. (2025); Ho et al. (2025); Sun et al. (2025c); Liu et al. (2025b); Zhang et al. (2025a); Huo et al. (2025); Lu et al. (2025a); Yang et al. (2025b)
ICML	Lv et al. (2024); Wu et al. (2024b); Chen et al. (2024f); Zhang et al. (2024b); Chen et al. (2024b); Li et al. (2025f); Zou et al. (2025); Zhao et al. (2025b); Xu et al. (2025); Zhao et al. (2025a); Park et al. (2025b,a)
IJCAI	Chen et al. (2024c); Zhang et al. (2025c); Wang et al. (2025a); Dang et al. (2025); Xiong et al. (2025); Sun et al. (2025a)
INLG	Akani et al. (2023); Obaid ul Islam et al. (2023); Liu et al. (2025d); Schmidtova et al. (2025)
NAACL	Raunak et al. (2021); Dziri et al. (2022); Petryk et al. (2024); Li et al. (2024a); Xia et al. (2024); Agrawal et al. (2024b); Chae et al. (2024); Jiang et al. (2024); Bécharde and Marquez Ayala (2024); Yu et al. (2024b); Tang et al. (2024); Shi et al. (2024); Lee et al. (2024b); Guan et al. (2024a); Zhu et al. (2024); Zhang et al. (2025e); Woo et al. (2025); Wang et al. (2025b); Li et al. (2025c); Tang et al. (2025); Bendahman et al. (2025); Belém et al. (2025); Liu et al. (2025c); Kang et al. (2025); Wu et al. (2025a); Li et al. (2025e)
NeurIPS	Wang et al. (2024); Lyu et al. (2024); Jesson et al. (2024); Du et al. (2024); Sriramanan et al. (2024); Chen et al. (2024e); Xing et al. (2024b); Wei et al. (2024); Aithal et al. (2024); Zhu et al. (2025a); Bar-Shalom et al. (2025); Chen et al. (2025c); Yang et al. (2025a); Suresh et al. (2025); Park and Li (2025); Bai et al. (2025b); Wang et al. (2025c); Qian et al. (2025); Ji et al. (2025); Li and Ng (2025); Niu et al. (2025); Leng et al. (2025); Ross et al. (2025); Shu et al. (2025); Sun et al. (2025b)

dition or Transformation when the corresponding evidence was available.

Omission Omission was coded conservatively. A paper had to explicitly count missing source information as hallucination. The existence of a separate omission or coverage-error category did not establish Omission as a hallucination symptom.

Model-internal indicators Saliency, attention patterns, token contributions, uncertainty estimates, hidden states, entropy, probability distributions, and related model-internal signals were recorded separately. No output-side symptom was assigned solely on the basis of a model-internal indicator.

B.3 Convergence Expectation

Convergence Expectation (CE) is a paper-level proxy for how strictly a paper expects model outputs to converge toward an ideal output. Higher CE represents stricter correctness or consistency expectations and lower tolerance for divergence. Lower CE represents a broader range of acceptable outputs or greater tolerance for creative, exploratory, or open-ended generation.

CE is not equivalent to factuality and does not measure the observed quality of a model. It captures the strictness of output expectations and evaluation framing embedded in each paper.

CE1–CE5 and CE7 were scored from 0 to 2, while CE6 was scored as 0, −1, or −2 to account for the value assigned to creative or exploratory divergence.

B.3.1 Two-point Gate Conditions

The following gate conditions were applied to reduce inflation of the highest CE item scores:

- **CE1 = 2** required an operational faithfulness criterion. Conceptual wording alone was coded as CE1 = 1.
- **CE3 = 2** required an explicit high-stakes domain or consequential decision context. General references to harm, trust, reliability, or misinformation were coded as CE3 = 1.
- **CE4 = 2** required explicit evidence of evaluation rigor. The statement that data were “manually annotated” was not sufficient on its own and was

Table 5: Coding rules for definitional clarity. The levels measure the operational specificity of a paper’s treatment of hallucination.

Level	Definition	Positive evidence	Boundary and exclusion rule
Lv0	The paper uses the term <i>hallucination</i> but provides no usable definition or judgement criterion.	No explicit or operational evidence can be identified from the available full text.	A brief definition is not automatically Lv0. If the paper provides a generic conceptual description, Lv1 applies.
Lv1	The paper provides a generic description or cites a definition, but does not provide its own operational judgement criteria.	Examples include descriptions such as <i>incorrect</i> , <i>unrelated</i> , <i>not faithful</i> , or <i>unsupported</i> , or a brief citation to prior work without operational detail.	If the paper specifies how hallucination is labeled, measured, annotated, verified, included, or excluded, Lv2 applies.
Lv2	The paper provides evidence showing how hallucination is identified, judged, or labeled.	Examples include annotation rules, metric definitions, inclusion or exclusion rules, verification procedures, label criteria, benchmark-construction rules, or explicit decision procedures.	Conceptual wording alone is insufficient. The evidence must support an applicable procedure for determining which outputs count as hallucination.

coded as CE4 = 1 unless additional rigor evidence was available.

- **CE7 = 2** required a concrete safeguard, accountability, or governance procedure. General references to trust, harm, accountability, or safety were coded as CE7 = 1.

B.3.2 CE Score Calculation

The raw CE score was calculated as follows:

$$CE_{\text{raw}} = CE1 + CE2 + CE3 + CE4 + CE5 + CE7 + CE6. \quad (1)$$

Because CE6 takes a value of 0, −1, or −2, stronger evidence that creative or exploratory divergence is valued reduces the CE total.

When no evidence was available for a CE item, the evidence quotation was recorded as NOT FOUND, and the item received the conservative default score of 0.

B.4 Primary Task Coding

The task_primary field contained the task name only. Methodological terms such as detection, mitigation, benchmarking, evaluation, uncertainty estimation, and model introspection were not included in the task label. Examples of normalized task names included:

- Neural Machine Translation;
- Summarization;
- Dialogue;

- Question Answering;
- Image Captioning;
- Data-to-Text;
- Conditional NLG;
- Retrieval-Augmented Generation;
- Open-ended Generation.

When a paper primarily focused on detection, mitigation, benchmarking, evaluation, or model introspection, the underlying generation task was recorded in task_primary, while the methodological focus was recorded in the human-review notes. Multi-task or unclear cases received an uncertainty flag.

B.5 Evidence Priority and General Coding Rules

Evidence was searched in the following priority order:

1. definition or task-formulation sections;
2. data, annotation, or label-definition sections;
3. evaluation, metrics, or experimental-setup sections;
4. Introduction and Related Work;
5. Abstract, used as a fallback.

The following rules applied to all coding constructs:

Table 6: Coding rules for output-side hallucination symptoms. Multiple symptom labels could be assigned to the same paper.

Symptom	Operational definition	Inclusion rule	Boundary and exclusion rule
Transformation	The output contradicts, distorts, or changes the meaning of an identifiable source.	Assigned when the paper explicitly treats contradiction, distortion, incorrect transformation, or alteration of source meaning as hallucination.	Requires an identifiable source. When the relevant standard is world knowledge, user intent, or dialogue context rather than an explicit source, Deviation may apply.
Addition	The output contains content not supported by the source, including content that is fully unrelated or detached from the source.	Assigned when unsupported, fabricated, extraneous, or unrelated content is explicitly treated as hallucination.	In source-grounded tasks, detached output is coded as Addition or Transformation where applicable, rather than as Deviation.
Omission	The output omits required source information and the paper explicitly counts the omission as hallucination.	Assigned only when omission is included in the paper’s hallucination definition, labels, annotation protocol, metric, or evaluation setup.	Missing information, incompleteness, coverage error, or under-generation is not coded as Omission unless the paper explicitly treats it as hallucination.
Deviation	In a source-free or weak-source task, the output diverges from dialogue context, user intent, world facts, or the expected answer space.	Assigned when hallucination is judged against a weak or indirect reference, rather than against a single explicit input source.	Deviation is not used for translation outputs that are detached from the source. In translation and other source-grounded tasks, unsupported or detached content is coded as Addition or Transformation where supported.
Other	The paper explicitly treats a task-specific output-side off-target behavior as hallucination, but the behavior is not adequately captured by the other four categories.	An example is generation in the wrong target language when the paper explicitly frames this behavior as hallucination.	Other is not a residual category for uncertainty. Saliency, token contributions, attention patterns, uncertainty, hidden states, entropy, and other model-internal indicators are not coded as Other.

1. **Paper-only evidence.** Coding used only content from the target paper. External knowledge was not used.
2. **Verbatim quotation.** Evidence consisted of verbatim quotations rather than paraphrases.
3. **Location pointer.** Each quotation included a URL, page number, section, and paragraph pointer where available.
4. **Missing evidence.** If evidence was not found, the corresponding evidence field was recorded as NOT FOUND.
5. **Conservative scoring.** If a decision was uncertain between adjacent levels or scores, the lower option was selected.
6. **One best quotation per CE item.** For final CE scoring, one best supporting quotation was retained for each CE item.

7. **Multi-label symptom coding.** Multiple symptoms could be assigned when explicitly supported.
8. **Internal indicators.** Model-internal indicators were recorded separately and were not treated as output-side symptoms.

B.6 Targeted Human Review

Human review was conducted on a targeted sample rather than on all papers. Sampling focused primarily on cases marked by uncertainty flags. These flags covered ambiguous Lv1–Lv2 boundaries, CE items near the two-point gates, unclear symptom mappings, multi-task or unclear task assignments, evidence available only in introductory sections, and possible confusion between model-internal indicators and output-side symptoms. For each sampled case, the extracted quotations and their locations were checked against the source paper, and

Table 7: Convergence Expectation scoring rubric. CE6 is a negative correction representing the extent to which creative or exploratory divergence is permitted or valued.

Item	Construct	Score 0	Score 1 / -1	Score 2 / -2
CE1	Faithfulness strength	No source-faithfulness or grounding requirement is identified.	Source faithfulness is mentioned conceptually or vaguely, but is not operationalized.	Operationalized faithfulness criteria are provided, such as an explicit source-based label rule, entailment check, attribution requirement, or verification procedure.
CE2	Factuality strength	No factuality requirement is identified.	Factuality or correctness is mentioned vaguely or conceptually.	An explicit external oracle or verification protocol is provided, such as fact-checking, a knowledge base, retrieval evidence, expert verification, or a benchmark oracle.
CE3	Failure cost	No risk or impact framing is identified.	The paper mentions general harm, trust, reliability, misinformation, or applicability concerns.	The paper explicitly addresses a high-stakes domain or context, such as medicine, law, safety, decision support, finance, or public-critical infrastructure.
CE4	Evaluation rigor	The evaluation is ad hoc, unclear, or insufficiently specified.	The paper uses a standard evaluation, benchmark, metric, experiment, or manual annotation procedure with limited procedural detail.	The paper provides explicit evidence of rigorous evaluation, such as a multi-annotator procedure, agreement statistics, a detailed labeling protocol, statistical testing, or an explicit validation design.
CE5	Source constraint	No clear source constrains the output.	The output is constrained by weak or implicit context.	The output is constrained by a strong explicit source, such as a translation source, source document, table, image, retrieved evidence, grounding corpus, or input context.
CE6	Creativity penalty	Creativity or divergence is not framed as valuable.	Score -1: creative divergence is partially permitted.	Score -2: creative divergence, diversity, exploration, or open-ended generation is central or explicitly valued.
CE7	Accountability, trust, or safeguards	No trust, accountability, reliability, or safeguard framing is identified.	The paper mentions trust, reliability, harm, applicability, accountability, safety, or user risk.	The paper specifies a concrete safeguard or governance procedure, such as attribution, human-in-the-loop verification, refusal design, citation mechanisms, an accountability workflow, deployment guardrails, or an audit process.

the corresponding paper-level codes were reviewed against the fixed codebook. Identified errors were corrected before analysis. Because the review was targeted rather than exhaustive, it should be interpreted as a quality-control procedure for flagged and sampled cases, not as full manual validation of the corpus.

B.7 Prompt Availability

The complete instruction prompt used for LLM-assisted evidence extraction and initial coding is provided in the supplementary material as `llm_coding_prompt.txt`. The prompt includes the evidence requirements, coding rules, CE scoring rubric, output schema, uncertainty flags, and human-review checklist.

B.8 Task-by-Symptom Counts across Observed CE Bands

Table 8 reports the underlying paper counts for the task-by-symptom heatmaps across the three observed CE bands. The *Papers* column gives the number of papers in each task-by-band cell and

serves as the denominator for the reported percentages. Because symptom coding was multi-label, a single paper could contribute to multiple symptom columns; therefore, symptom counts and percentages within a row are not expected to sum to the number of papers or to 100%. Cells with small paper counts should be interpreted descriptively and with caution.

Table 8: Paper counts and output-side symptom frequencies by normalized task category and observed CE band. Symptom cells report n (%) within each task and CE band.

Task	Papers N	Transformation	Addition n (%) within task and CE band)	Omission	Deviation	Other
Panel A: Lower observed CE, 1–6 ($N = 49$)						
QA	13	2 (15.4)	2 (15.4)	0 (0.0)	11 (84.6)	2 (15.4)
General / Conditional NLG	16	6 (37.5)	7 (43.8)	2 (12.5)	8 (50.0)	5 (31.2)
Image Captioning	4	2 (50.0)	3 (75.0)	0 (0.0)	0 (0.0)	0 (0.0)
Summarization	7	6 (85.7)	5 (71.4)	1 (14.3)	0 (0.0)	1 (14.3)
Dialogue	3	0 (0.0)	1 (33.3)	0 (0.0)	1 (33.3)	1 (33.3)
MT	3	2 (66.7)	2 (66.7)	0 (0.0)	0 (0.0)	1 (33.3)
RAG	2	1 (50.0)	1 (50.0)	0 (0.0)	1 (50.0)	0 (0.0)
Data-to-Text	1	0 (0.0)	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
Other / Low-frequency	0	–	–	–	–	–
Panel B: Middle observed CE, 7–9 ($N = 124$)						
QA	38	14 (36.8)	19 (50.0)	4 (10.5)	23 (60.5)	6 (15.8)
General / Conditional NLG	22	9 (40.9)	16 (72.7)	1 (4.5)	15 (68.2)	3 (13.6)
Image Captioning	21	15 (71.4)	21 (100.0)	1 (4.8)	1 (4.8)	4 (19.0)
Summarization	16	14 (87.5)	16 (100.0)	1 (6.2)	1 (6.2)	1 (6.2)
Dialogue	7	6 (85.7)	4 (57.1)	0 (0.0)	1 (14.3)	1 (14.3)
MT	10	2 (20.0)	10 (100.0)	0 (0.0)	0 (0.0)	4 (40.0)
RAG	4	2 (50.0)	4 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
Data-to-Text	4	3 (75.0)	4 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
Other / Low-frequency	2	0 (0.0)	2 (100.0)	0 (0.0)	1 (50.0)	0 (0.0)
Panel C: Higher observed CE, 10–12 ($N = 71$)						
QA	30	23 (76.7)	26 (86.7)	1 (3.3)	13 (43.3)	5 (16.7)
General / Conditional NLG	20	16 (80.0)	17 (85.0)	4 (20.0)	9 (45.0)	6 (30.0)
Image Captioning	7	4 (57.1)	7 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
Summarization	2	2 (100.0)	1 (50.0)	1 (50.0)	1 (50.0)	0 (0.0)
Dialogue	5	5 (100.0)	5 (100.0)	0 (0.0)	0 (0.0)	2 (40.0)
MT	1	0 (0.0)	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
RAG	4	4 (100.0)	3 (75.0)	1 (25.0)	0 (0.0)	0 (0.0)
Data-to-Text	1	0 (0.0)	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)
Other / Low-frequency	1	0 (0.0)	1 (100.0)	0 (0.0)	0 (0.0)	1 (100.0)

Note. The Papers column reports the number of papers in each task and observed CE band and serves as the denominator for the percentages in the symptom columns. Output-side symptom labels were non-mutually exclusive: a single paper could be assigned two or more symptoms. Consequently, symptom counts and percentages across a row may sum to more than the number of papers and more than 100%, respectively. A dash indicates that no paper was assigned to the corresponding task and CE band.