

A Forced Answer Channel for Closed-Book Knowledge Base Construction at LM-KBC 2026

Yiqun Sun
Magellan Technology
Research Institute (MTRI)
duke.sun@mtri.co.jp

Pengfei Wei
Magellan Technology
Research Institute (MTRI)
pengfei.wei@mtri.co.jp

Lawrence B. Hsieh
Magellan Technology
Research Institute (MTRI)
lawrence.hsieh@mtri.co.jp

Abstract

The AKBC 2026 shared task (Kalo et al., 2026) asks a system to complete the object set of a subject and a relation out of a language model’s parameters, with no retrieval, no training, and no more than 32B parameters. Our system drives one 29.6B backbone through three stages. Several elicitation channels per relation build a candidate pool, exhaustive one-token Yes/No probes verify whatever part of the answer space can be enumerated, and a per-relation decision layer emits the object set. The backbone writes a reasoning channel before its user-facing answer, so a probe prompt has to append the answer-channel header before its first generated token is a verdict. Forcing that channel also makes auxiliary generations cheap, which is what puts roughly 26k border probes inside the budget. Most decision constants are fitted on validation gold under a rule that displaces an incumbent only on a strict improvement of more than one row, and the fitting procedure itself is cross-validated, which cut one city gain from +0.070 to +0.020 before it was adopted. The system scores 0.6553 official macro-F1 against 0.6415 for a prompting baseline on the same backbone, and the gap is uneven: the machinery is worth +0.1662 on awards and costs 0.0204 on hasCapacity, where plain prompting wins. We also report the ceiling we could not move. Twenty-one method families have been fitted against hasCapacity and none of them beats a mass vote over the pool. Harvesting the model’s own reasoning traces raises pool coverage from 0.582 to 0.724 and converts none of it.

1 Introduction

The LM-KBC challenge series (Kalo et al., 2023, 2024, 2025), whose 2026 edition is the AKBC shared task (Kalo et al., 2026), asks a language model to serve as the knowledge base itself (Petroni et al., 2019; Roberts et al., 2020). A system receives a subject entity and a relation and returns

the complete object set, which may be empty, a single value, or several hundred names. The 2026 edition forbids retrieval at inference and training of any kind, and it caps the total neural budget at 32B parameters. A score therefore measures two things at once, what a backbone holds and what the machinery around it can pull out.

The system uses one model for every stage. Muse-Glimmer-30B fits the budget on its own at 29.6B parameters, so the pipeline contains no second network and no cross-model agreement term, and none of the heterogeneous multi-model orchestration that knowledge-intensive pipelines usually reach for (Du et al., 2023; Feng et al., 2026b). Per subject the backbone runs several elicitation channels into a candidate pool, one-token Yes/No probes verify candidates wherever the answer space can be enumerated, and a per-relation decision layer turns pool and probes into an object set (§3). The system scores 0.6553 official macro-F1 over the 475 test rows, against 0.6415 for self-consistent prompting of the same backbone with none of the machinery attached.

The mechanism that shapes the rest of the design is a property of the backbone. It answers in two channels, a reasoning channel followed by a user-facing answer channel, and a generation prompt lands the model in the first of them. A one-token Yes/No probe therefore reads the opening token of deliberation. Appending the answer-channel header to the prompt makes the first generated token the verdict, which is what turns verification into a single forward pass and puts roughly 26k border probes inside the compute budget.

The protocol behind the decision constants is not specific to this task (§4). Constants are fitted on validation gold, a challenger displaces an incumbent only on a strict improvement of more than one row, and the fitting procedure itself is cross-validated. That last step matters, because an in-sample sweep of the city rule preferred a setting worth +0.070

on validation and cross-validating the procedure deflated it to +0.020. Neither is the negative program (§6). Twenty-one method families have now been fitted against hasCapacity, spanning selection statistics, probe reformulations, representation read-outs and fresh elicitation surfaces, and the diagnostics say where the wall is. Code, prompts and the elicitation checkpoints behind the scored predictions are released at <https://github.com/dukesun99/elicitation-beats-selection>.

2 Task and metric

Six relations are scored over 475 test rows. Three are nullable string relations, country-LandBordersCountry (67 rows, *borders* below), personHasCityOfDeath (100 rows, *city*) and companyTradesAtStockExchange (100 rows, *company*). awardWonBy (10 rows, *award*) asks for recipient sets that run to hundreds of names. hasCapacity (98 rows, *capacity*) and hasArea (100 rows, *area*) are numeric and are scored within a 5% relative tolerance. The metric is per-row F1 macro-averaged over rows with alias-aware normalization, and an empty prediction against an empty gold set scores 1.0. Rows and not relations are averaged, so a relation’s weight is its row count. Decision constants are fitted on the train and validation splits. No retrieval runs at inference, no external resource enters any prompt, and no parameter of the backbone is trained, adapted or otherwise updated at any point. Every test number below is official.

3 System

Backbone. Muse-Glimmer-30B carries 29.6B parameters including a 1.8B vision encoder and is released under Apache-2.0. It fits the 32B budget alone, which removes the usual option of pairing a large generator with a small verifier. Sampling channels keep the reasoning channel, because that is where the model’s factual advantage lives (Wei et al., 2022), and probes discard it by forcing the answer channel open, which is the same trade a shortcut decoder makes for cost (Liu et al., 2026). The chat template defaults to high reasoning effort, which silently truncates long elicitation prompts unless the effort is lowered. Two failures follow from ignoring either property. Unforced probing returns the first token of deliberation and ranks a living person as more likely deceased than a long-dead one, with no error raised. Leaving the effort high drops the parse rate far enough to look like a

capability result.

Elicitation. Each subject draws $k=12$ samples from a few-shot prompt that carries the relation definition verbatim, a JSON output schema and training exemplars, including empty-gold exemplars wherever the relation permits an empty answer (Wang et al., 2022; Jiang et al., 2020). Auxiliary channels approach the same fact from another angle at $k=8$, all of them forced into the answer channel, which is what makes them affordable. Varying the frame, and not sampling one frame harder, is what widens the pool, the same reason diversity-aware retrieval covers an event that relevance ranking alone leaves half-told (Tang et al., 2025a). The numerics add a recitation channel, a commonly-cited-figure frame and a wiki-table completion frame (Elazar et al., 2021). City adds an obituary recitation, a local-language frame (Kassner et al., 2021; Zhao et al., 2026) and a presupposition frame that asserts the death and asks only for the locality. Award uses a list enumeration. Borders runs two list passes at different sampling depths. Greedy channels run at $k=1$ and temperature 0.

Verification. Where the answer space can be enumerated, exhaustive one-token probing (Kadavath et al., 2022; Sun et al., 2026b) replaces open-ended generation. The model writes the universe first, the world’s sovereign states for borders and its stock exchanges for company, keeping a name that recurs across samples. Borders then probes every subject-country pair in both directions, about 26k probes in total. Company probes every subject-exchange pair and debiases each score in log-odds against a placebo built from training companies with empty gold, which removes the model’s standing tendency to answer Yes (Zhao et al., 2021). City gets an alive-deceased gate from a single forced token. Capacity scores each cluster representative by its pointwise mutual information against a generic-venue prefix (Holtzman et al., 2021).

Decision layer. String relations fuse sampling frequency with probe score under a per-relation threshold, reconciling what the model generates with what it will affirm, in the manner of a second-pass rescorer over an n-best list (Kadavath et al., 2022; Tang and Tung, 2024; Tian et al., 2025). Borders admits a country at $0.7 \text{ freq} + 0.3 \text{ ver} \geq 0.40$ with the deeper of the two list passes weighted 3, and company keeps its released fusion. City pools the direct channel with its recitation channel

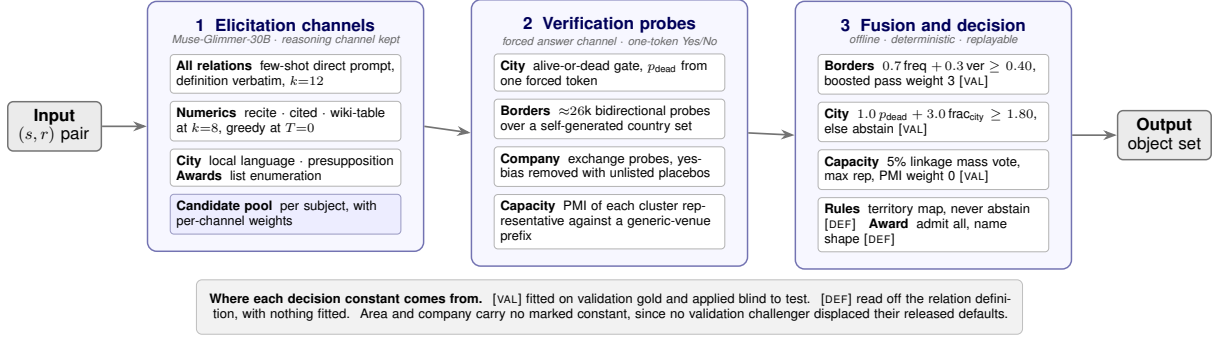


Figure 1: The pipeline, and where each group of decision constants came from (§4).

at double weight, answers with the most frequent candidate when the deceased probability plus three times the fraction of samples naming a city clears 1.80, and abstains otherwise. Numerics cluster the pooled values by single linkage and take the largest cluster’s maximum, at a 5% linkage width for capacity. The capacity PMI term is computed and then weighted at zero, which is the setting validation preferred. Definition-derived rules close the pipeline, mapping integral territories to their state so that France borders Suriname through French Guiana, and filtering award candidates by name shape. Numerics and award never abstain. Everything downstream of the elicitation checkpoints is deterministic and replays offline.

4 Fitting the decision constants

Most decision constants are fitted on validation gold and applied to test without adjustment, and the exceptions are named at the end of this section. The adoption rule is what makes that fitting trustworthy. A challenger displaces the incumbent only when it improves validation by more than one row, since a margin of one row is inside the resolution of a 100-row relation. Four candidates were declined under that rule, two of them on area (the clustering tolerance together with the representative choice, and the channel weights), one on company fusion, and one capacity setting that merely tied. A tie does not displace.

City needed more than a threshold on the sweep. The in-sample sweep preferred a rule worth $+0.070$ on validation, which is large enough for the winner’s curse to be the leading explanation (Blum and Hardt, 2015; Dwork et al., 2015). Cross-validating the fitting procedure, so that each fold reran the entire sweep and the fold’s own winner was scored on the held-out half, deflated the gain to $+0.020$, and the cross-validated setting is the one the system

uses. Both folds independently dropped the presupposition channel, which is why that part of the choice is trusted. Cross-validating the procedure and not the fitted rule is what separated the real effect from the noise around it.

The largest gain came from a knob nobody had turned. Channel weights had never been swept anywhere in the system. They were frozen once and carried forward while decision rules were reworked around them. Refitting them produced a larger gain than any decision-rule change we tried. Capacity rose 0.031 on test from two validation-derived choices, switching the PMI term off and doubling the weight of the cited, greedy and wiki-table channels. Both choices remove fitted structure instead of adding more of it.

Not every constant comes from validation gold, and the figure marks the difference. Area and company keep their released defaults, because no validation challenger strictly improved on them. Award carries no fitted constant at all. Its rule admits every distinct candidate the enumeration channel names, subject to a name-shape filter and the never-abstain behaviour, and both of those are read off the relation definition, which asks for recipient entities and allows large object sets. That is the whole rule, and it needs no fitting ground, which is fortunate: validation and test awards differ in shape, with validation dominated by fellowship-style gold sets and test by annual prizes, so the validation split would have been a poor one.

5 Results

The system scores 0.6553 official macro-F1, with macro precision 0.7842 and macro recall 0.6597 (Table 1).

The per-relation profile separates cleanly into three regimes. Borders and area sit above 0.85, and both are relations where the answer space is

Relation	rows	prompting	system
borders	67	0.9346	0.9481
area	100	0.8300	0.8600
company	100	0.7838	0.7838
city	100	0.5400	0.5700
capacity	98	0.2653	0.2449
award	10	0.0696	0.2358
all rows	475	0.6415	0.6553

Table 1: Official per-relation test F1 for the prompting baseline and for the full system. The system’s macro precision over all rows is 0.7842 and its macro recall is 0.6597.

small enough to be swept exhaustively or resolved by agreement across channels. Company and city sit in the middle, where the binding constraint is a decision about what to admit and what to leave out. Capacity and award sit at the bottom, and §6 gives the measurements behind the capacity floor, which no rule in twenty-one method families has moved. Precision runs well ahead of recall over the whole test set, which is the operating point per-row F1 rewards, because filling a gold-empty row with a plausible object costs the full row while one recovered object gains a fraction of it.

Prompting baseline. The reference point is the same backbone with none of the machinery around it. It reads the $k=12$ direct samples and aggregates them by self-consistency under rules taken from the task definition instead of fitted. Numerics cluster the parsed values at the metric’s own 5% tolerance and answer with the largest cluster’s median, city takes a majority over parsed cities including the null answer, and the set-valued relations admit an object named in more than half of the samples that parsed. No auxiliary channel, no probe, no gate and no fitted constant enters it. That baseline scores 0.6415, so three stages of elicitation, verification and fitted decision are worth +0.0138 over prompting the same model.

The gain is concentrated, not broad. Award moves +0.1662, though on 10 rows, city and area each move +0.0300, and borders moves +0.0135. Two rows go the other way, and both are worth stating plainly.

Where the machinery does not pay. Capacity is 0.0204 *worse* under the full system than under plain prompting. The auxiliary channels and the max-representative rule are net harmful on the relation that most of this paper’s negative program is about, and a system built to elicit and select loses there

to a median over direct samples. The choice was not made with test labels in view. On validation the system’s capacity configuration scores 0.3505 against the baseline’s 0.3196, so the validation signal pointed the other way and the protocol of §4 kept the fitted setting. We report the relation as the system produced it, and record that validation did not predict the test outcome here. Company is the second row: it gains exactly 0.0000. The exchange probe universe and the placebo correction recover to four decimal places what the direct channel already says, so that relation carries verification machinery it does not need.

6 What bounds the score

Capacity has defeated 21 method families. On the selection side these include self-verification, pairwise duels, multiple choice, PMI ranking, cross-model and cross-language agreement, activation read-outs (Burns et al., 2022; Li et al., 2023; Wang et al., 2026; Wu et al., 2025), task-vector steering (Hendel et al., 2023; Todd et al., 2024; Sun et al., 2026a), multi-agent debate (Du et al., 2023; Feng et al., 2026a), complete-linkage clustering, nine hand-designed selection statistics, channel-subset search, a highest-corroborated-value rule, ordinal threshold inversion, channel-breadth scoring, within-cluster representative choice and the per-relation sampling budget. On the elicitation side they include volume boosts, higher reasoning effort, multilingual prompts, explicit maximum-configuration prompts, prompt surfaces that imitate Wikidata property syntax, which carry no external data and only reframe the question, and reasoning-trace harvesting.

Four measurements say the most about why. Harvesting the model’s own reasoning traces (Lu et al., 2025; Turpin et al., 2023) raises the fraction of capacity subjects whose pool holds a value within tolerance from 0.582 to 0.724 and converts none of it, in any confidence stratum, at any validation-fitted weight. Treating the task as ordinal inversion and probing thresholds directly fails on calibration (Zhao et al., 2021; Tang et al., 2021), since 23% of adjacent thresholds violate monotonicity and the model gives a randomly chosen venue a 28% chance of holding at least 150,000 people, with direct readout landing within tolerance on 4 of 98 rows. Hedging with a second value pays only when hit@2 exceeds 1.5 times hit@1, and the measured ratio is 1.20. Raising the sampling budget

does nothing, with the validation capacity curve flat from $k=6$ at 0.2809 to $k=12$ at 0.2784.

The error structure explains the pattern. Gold is present in the pool for about two thirds of capacity rows and holds a median 12% of the cluster mass when it is present, and the errors are bimodal at roughly a factor of two, the signature of confusing a seated figure with a total or a current configuration with a historical maximum. No statistic computed from the generation distribution finds a signal that thin, which is why raising coverage without changing the shape of the mass buys nothing (Kandpal et al., 2023; Xu et al., 2024).

The other two relations are bounded differently, and both bounds are measured. City’s gold-empty rows are living people, so abstention on that relation is alive-dead classification. The deceased-gate probe separates the two classes at AUC 0.745 and no alternative detector beat it, which caps how far the abstention rule can move before it starts trading correct empties for wrong cities. Company recall is bounded by secondary listings the sampler never names, an omission a confident-looking answer set does not reveal (Tang et al., 2025b; Kandpal et al., 2023). A directed secondary-listing channel does surface them on validation and still loses under cross-validation, because the same channel names plausible exchanges for companies that hold no secondary listing at all.

7 Conclusion

One 29.6B backbone, multi-channel elicitation, exhaustive one-token verification made affordable by forcing the model’s answer channel open, and a small per-relation decision layer reach 0.6553 on the LM-KBC 2026 test set, 0.0138 above self-consistent prompting of the same backbone. The constants that were fitted came from validation gold under a strict-improvement rule, with the fitting procedure itself cross-validated, and the single largest gain came from channel weights that had never been swept. What the system cannot do is convert pool coverage into score on hasCapacity, where the correct value is present in about two thirds of pools, holds a median 12% of the cluster mass, and has resisted 21 families of method. On that relation the machinery is not merely stuck, it is behind the baseline it was built to beat.

Limitations

The negative results were measured on one backbone inside one size band, and a model with a different pretraining mixture may put the capacity mass somewhere a selection statistic can reach. Individual per-relation deltas are small against the resolution of a 100-row relation, so a single component is rarely settled on its own by this metric (Reimers and Gurevych, 2017; Dodge et al., 2019). The comparison against the prompting baseline inherits that resolution, so the two relations where the machinery fails to pay, capacity at -0.0204 and company at 0.0000 , are better read as evidence that those components earn nothing than as a calibrated size for the loss. Award carries 10 of 475 rows, and its gold sets run to hundreds of names, so conclusions about that relation are the weakest in the paper, and its rule is the least constrained by data of any in the system.

References

- Avrim Blum and Moritz Hardt. 2015. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of ICML*.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP-IJCNLP*.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.
- Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhishava Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Yang Feng, Ziwei Xu, Xia Hu, and Fengxiang He. 2026a. PEAR: Permutation-equivariant adaptive routing multi-agent debate. *arXiv preprint arXiv:2606.20621*.

- Yichao Feng, Haoran Luo, Zhenghong Lin, Yiqun Sun, Pengfei Wei, Lawrence B. Hsieh, and Anh Tuan Luu. 2026b. OrchMAS: Orchestrated reasoning with multi collaborative heterogeneous scientific expert structured agents. *arXiv preprint arXiv:2603.03005*.
- Roe Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of EMNLP*.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn’t always right. In *Proceedings of EMNLP*.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Jan-Christoph Kalo, Tuan-Phong Nguyen, Simon Razniewski, and Bohui Zhang. 2024. Preface: LM-KBC challenge 2024. In *Joint Proceedings of the KBC-LM Workshop and the LM-KBC Challenge*, volume CEUR Vol-3853.
- Jan-Christoph Kalo, Tuan-Phong Nguyen, Simon Razniewski, and Bohui Zhang. 2026. [AKBC Shared Task 2026: Predicting complete knowledge base entries from language models](#). In *Proceedings of the 10th Workshop on Automated Knowledge Base Construction (AKBC) at EMNLP 2026*. Association for Computational Linguistics.
- Jan-Christoph Kalo, Simon Razniewski, Bohui Zhang, and Tuan-Phong Nguyen. 2025. LM-KBC 2025: 4th challenge on knowledge base construction from pre-trained language models. In *Joint proceedings of KBC-LM and LM-KBC @ ISWC*, volume CEUR Vol-4041.
- Jan-Christoph Kalo, Sneha Singhania, Simon Razniewski, and Jeff Z. Pan. 2023. LM-KBC 2023: 2nd challenge on knowledge base construction from pre-trained language models. In *Joint proceedings of KBC-LM and LM-KBC @ ISWC*, volume CEUR Vol-3577.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of ICML*.
- Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of EACL*.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.
- Fan Liu, Yanhao Wang, Min Zhang, Zhikang Chen, Zeyuan Li, Lewei He, and Jiahui Pan. 2026. Think faster than words: Efficient LLM chain-of-thought reasoning via dynamic shortcut decoding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28825–28836. Association for Computational Linguistics.
- Jiahao Lu, Ziwei Xu, and Mohan Kankanhalli. 2025. Reasoning LLMs are wandering solution explorers. *arXiv preprint arXiv:2505.20296*.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*.
- Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of EMNLP*.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of EMNLP*.
- Yiqun Sun, Junyu Chen, Pengfei Wei, and Lawrence B. Hsieh. 2026a. GGSS: Geodesic-gated spherical steering for inference-time debiasing of generative vision-language models. *arXiv preprint arXiv:2608.25375*.
- Yiqun Sun, Pengfei Wei, and Lawrence B. Hsieh. 2026b. Very efficient listwise multimodal reranking for long documents. *arXiv preprint arXiv:2605.11864*.
- Yixuan Tang, Hwee Tou Ng, and Anthony K. H. Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249. Association for Computational Linguistics.
- Yixuan Tang, Yuanyuan Shi, Yiqun Sun, and Anthony K. H. Tung. 2025a. Uncovering the bigger picture: Comprehensive event understanding via diverse news retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33939–33957. Association for Computational Linguistics.
- Yixuan Tang and Anthony K. H. Tung. 2024. Contextualized speech recognition: Rethinking second-pass rescoring with generative large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6478–6485.

- Yixuan Tang, Jincheng Wang, and Anthony K. H. Tung. 2025b. The missing parts: Augmenting fact verification with half truth detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33979–33996. Association for Computational Linguistics.
- Yang Tian, Fan Liu, Jingyuan Zhang, Victoria W., Yupeng Hu, and Liqiang Nie. 2025. CoRe-MMRAG: Cross-source knowledge reconciliation for multi-modal RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32967–32982. Association for Computational Linguistics.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *Proceedings of ICLR*.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.
- Hao Wang, Yiqun Sun, Pengfei Wei, Lawrence B. Hsieh, and Daisuke Kawahara. 2026. Sparse autoencoders as plug-and-play firewalls for adversarial attack detection in VLMs. *arXiv preprint arXiv:2605.07447*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*.
- Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
- Jiaqi Zhao, Qiang Huang, Haodong Chen, Xiaoxing You, and Jun Yu. 2026. When abundance conceals weakness: Knowledge conflict in multilingual models. *arXiv preprint arXiv:2601.07041*.
- Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of ICML*.