

Exemplars Primarily Shift Retrieval, Verification Instructions Can Shift Rejection: An Anatomy of Inference-Time Interventions for Closed-Book Knowledge Base Construction

Takashi Sasaki

Acroquest Technology Co., Ltd.

sasaki@acroquest.co.jp

Abstract

We describe our AKBC 2026 closed-book knowledge base construction system: Gemma-3-27B-pt, raw text completion, 20 candidates per query, and relation-specific aggregation without an arbiter. It uses 27.4B parameters and scores 0.6670 row-averaged Macro F1 on the official test, placing sixth. Experiments distinguish this final pipeline from an earlier Mistral system with an 8B arbiter. Backbone and access format outweighed prompt-level interventions: under matched prompt formats, the pretrained checkpoint exceeds its instruction-tuned sibling by 0.11 on validation. Numeric example configurations primarily shift retrieval; procedural wording with five added JSON examples changes the selected answers by only one row. Verification instructions can shift rejection. A biography-first city prompt format affects both, raising validation F1 from 0.55 to 0.61 at a fixed vote gate. Gating, information outside the model, and seed stability help interpret transfer. These findings have counterexamples; leaderboard comparisons describe development outcomes, not an untouched final evaluation.

1 Introduction

LM-KBC (Singhania et al., 2023; Kalo et al., 2024) asks a model to output *all* object values for a subject–relation pair, including empty or large sets, from its own weights. We call producing a stored value *retrieval* and discarding a generated candidate *rejection*.

We developed Qwen, Mistral, and Gemma systems (Sec. 4), varying models, prompts, sampling, and aggregation. Section 3 consolidates the final system; controlled experiments then explain the development choices. Experiments use validation and, for threshold selection, train. Hidden-test scores track submissions, with later choices partly informed by leaderboard feedback (Sec. 4.2).

2 Task and Constraints

The 2026 task covers six relations over 475 validation and 475 test subject–relation pairs: two set-valued relations that may be empty (countryLandBordersCountry, companyTradesAtStockExchange), one single-valued string relation that may be empty (personHasCityOfDeath), one large-set relation (awardWonBy), and two numeric relations (hasArea, hasCapacity) scored at 5% relative tolerance. Gold values come from Wikidata (areas P2046, capacities P1083), with alias lists accepted for strings. We report the official “All Relations” score throughout: a Macro F1 averaged over *rows*, not relations, so the four 100-row relations dominate it. External search, corpora, and fine-tuning are prohibited; inference is capped at 32B parameters.

3 Final Submitted System (V19)

Figure 1 consolidates V19 and its experimental basis. It uses one Gemma-3-27B-pt checkpoint (Q8_0, 27.4B) with llama.cpp at a pinned revision. Text completion continues a record- or infobox-style prefix without a chat template. Numeric prompts use four fixed training examples; city uses three hand-written examples, then generates About: before City of death:. The figure specifies sampling and every relation’s aggregation and empty-output rules.

Normalization. Released train/validation aliases are grouped by English-label matches before string voting. Company additionally folds case variants and acronyms into co-occurring long forms (NYSE → New York Stock Exchange). Numeric predictions use `dominant_cluster`. All frequencies include empty candidates.

Scope. Figure 1 separates deployed components from development-only experiments. V19’s city gate is 14 votes; Sec. 5.4’s fixed-12-vote compar-

Final System (V19) and Related Experiments

One Gemma-3-27B-pt model generates 20 continuations per query; relation-specific rules produce the answers.
Top: inference pipeline. Below: related experiments and final settings.

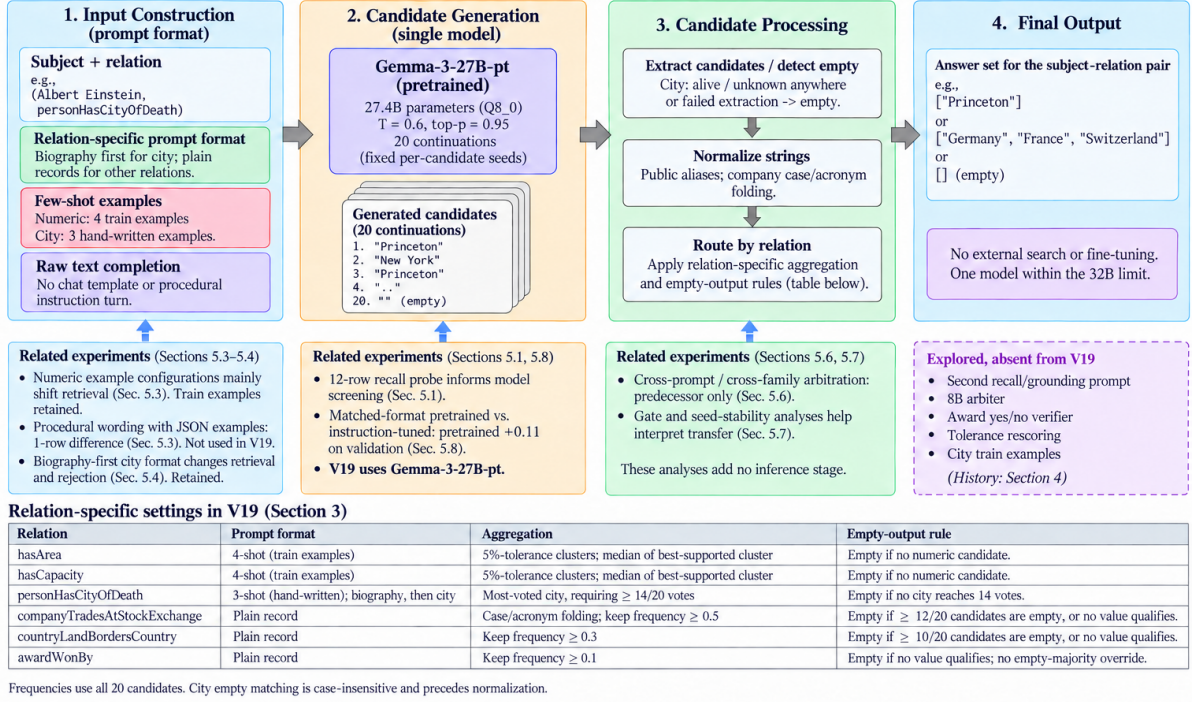


Figure 1: Final V19 system and related experiments. Gray arrows show inference; blue arrows associate experiments with components. All frequencies use the 20-candidate pool. City empty detection precedes normalization and voting. The final answer list is written as ObjectEntities, alongside SubjectEntity and Relation, in official JSONL format.

ison isolates the prompt change with the same candidate-extraction rule.

Reproducibility. Team: acro-sasaki. Code, pinned model, container and dataset versions, experiment records, and docs/V19_REPRODUCE.md: <https://github.com/acrosasaki-rgb/llm-knowledge-task-2026>. Builders deterministically reproduce predictions from saved candidates (SHA-256 verified).

4 Development History and Results

4.1 Predecessor system

Most experiments use Mistral Small 3.2 24B with 20 candidates from each of two prompts: **recall**, with two verbalized-recall exemplars; and **grounding**, with five bare-JSON examples naming Wiki-data P2046/P1083 (Sec. 5.3). For numeric disagreements above 5%, the mode of 20 Llama-3.1-8B samples selects the secondary answer only when matching it but not the primary (Sec. 5.6). This 31.8B predecessor cannot introduce a third value; an incorrect switch can lose a row.

4.2 Results

Table 1 separates the model and pipeline progression from the dataset re-keying, which corrects what is scored rather than improving the system. The public alias graph improved validation $0.658 \rightarrow 0.663$, driven by company. Section 3 and Fig. 1 define the final configuration. Per-relation test scores for V19 are: borders 0.960, company 0.832, area 0.820, city 0.530, award 0.336, capacity 0.316. The city improvement from v18 to v19 ($0.510 \rightarrow 0.530$) combines a prompt format change with a gate change; the fixed-gate prompt format comparison is reported in Sec. 5.4. Some revisions were retained or reverted after leaderboard feedback, notably the removal of city training exemplars. Submission deltas therefore describe development outcomes rather than an untouched final evaluation.

5 Anatomy of Inference-Time Interventions

Controlled comparisons below isolate the stated intervention; submission transitions in Table 1

System	test
Qwen3.5-27B thinking	0.571
Mistral 3.2 grounding	0.595
v3: + recall prompt, area arbitration, merge	0.604
v4: + capacity arbitration	0.606
v5: + acronym merging	0.610
v7: + city merge	0.608
v8: re-keying*	0.625
v10: + tolerance rescoring	0.623
v11: — rescoring	0.625
v12: Gemma-3-27B-pt	0.649
v13: + alias graph	0.648
v16: + train exemplars	0.659
v18: + city vote gate	0.663
v19: + biography, city gate 12 → 14	0.667

Table 1: Official test scores of our submissions (row-averaged Macro F1, three decimals). Rows summarize development changes, not uniformly isolated ablations: v19 changes both the city prompt format and vote gate. Validation scores are not comparable across the whole table—the August release re-keyed the gold—so they are reported per experiment in Sec. 5 instead. The official baseline scores 0.313 on validation. *Following that release, which renamed 15 test subjects and dropped two rows (477→475); v3–v7 score the renamed rows as zero. Versions v9, v14–v15, and v17 are omitted for space.

need not. Numeric testbeds are *hasArea* and *hasCapacity* (100 validation rows × 20 candidates per prompt). **Oracle** is the number of rows whose candidate set contains at least one within-tolerance value; the **mode** row count is what *dominant_cluster* selects. The **anchor** is the first concrete value for the queried quantity in a candidate.

5.1 Backbone Recall Reliability Dominates

Our Qwen-era system plateaued after nine intervention experiments (prompts, temperature, output routes, thinking budgets, judge models, log-probability confidence) whose combined deployable gain was under +0.01. A 12-row probe of its wrong rows across candidate models (Table 2) showed Mistral Small 3.2 answering 11 of the 12 somewhere in its candidate set; regenerating the pipeline on Mistral with nothing else changed moved test 0.571→0.595.

The second change was larger and reversed an earlier decision. Gemma 4 31B Instruct, queried through its chat template, had the weakest *hasArea* oracle we measured (43/100) and we closed that lineage early. Probing the *pretrained* Gemma-3-27B checkpoint in a completion format instead gave the

Model (12-row probe)	any	mode
Mistral Small 3.2 Instruct	11	8
Mistral Small 3.1 Base	10	7
Mistral Small 3.1 Instruct	9	4
Llama-3.1-8B Instruct	7	4
Aya Expanse 32B	3	1
Qwen3-8B	1	0

Table 2: The recall probe: 12 *hasArea* rows stratifying the Qwen-era system’s failures, 20 candidates each; *any* = rows with a correct candidate, *mode* = rows whose modal cluster is correct. Qwen3.5/3.6-27B, Gemma 4 31B, GLM-Z1-32B and llm-jp-4-32b (not shown) sit between Llama-8B and Aya. It costs ~30 GPU-minutes per model.

best profile of any configuration we tested—oracle 94, mode 84 on *hasArea*, borders 0.986—and a system built on it took test 0.625 → 0.649. It also supplies values no prompt reached on Mistral: secondary stock listings that five Mistral prompt variants returned with zero votes (ČEZ on Warsaw, Questerre on Oslo, Aisin on Nagoya) appear with 7–13 votes out of 20. This mattered more than anything else we varied: closed-book quality is set by which label-to-value pairings pretraining left intact and by whether the prompting format can reach them.

5.2 Answers Commit Early

In persisted reasoning traces of the Qwen-era system, the answer value first appears at a median 17% of the trace, in 95–98% of candidates inside a retrieval template (“Commonly cited: X”); the remaining 80% of the trace never dislodges it. The verification text that follows is consistent with post-hoc justification rather than independent checking: fabricated Wikidata Q-IDs and cross-language “confirmations” agree with whatever the anchor was, and wrong candidates carry *more* of this apparatus than correct ones (6.9 vs. 4.2 source mentions). Early commitment replicates on Gemma 4 31B and Qwen3.6; the confabulated verification does not. Extended thinking is thus an *exploration* aid: it raises the oracle where recall is unreliable (Qwen +9 rows, Gemma +7) and has no measurable effect on Mistral or Magistral.

5.3 Exemplar Style Primarily Shifts Retrieval

Table 3 distinguishes procedural wording from the examples embedded in that wording. With five added bare-JSON examples, the two configurations differ by only one mode-correct row (71 vs. 72).

Instruction	In-text recall	Added JSON	Mode	Oracle
Grounding	0	5	71	89
Grounding	0	0	41	59
Recall-format	2	5	72	83
Recall-format	2	0	78	84

Table 3: Mistral 3.2, 100 validation hasArea rows, 20 candidates each. Example columns show counts and placement. These are four prompt configurations, not a factorial isolation of example style: count, content, and placement also differ. The last row uses the original recall format; an independently relabeled variant also reaches mode 78, with oracle 89.

Removing those examples from the recall-format prompt retains two verbalized-recall examples and raises the mode from 72 to 78; removing all examples from grounding instead lowers it to 41. The recall-format candidates verbalize recall in 99% of generations without the added JSON examples, but in 0% with them. Thus example configuration strongly affects generated answers here; this comparison does not isolate style from example count or content, or establish that instructions generally have no effect. The independently relabeled recall format retains mode 78.

In the final backbone, replacing our hand-written exemplars with four answers taken verbatim from the *training* split—same prompt format, same generation, same aggregation, only the examples change—moves validation hasArea 0.84 $\rightarrow$ 0.88 and hasCapacity 0.309 $\rightarrow$ 0.361. The gain holds across three sampling seeds (area 0.88/0.89/0.87, capacity 0.361/0.351/0.361), and no exemplar subject occurs in validation or test.

These comparisons locate the clearest exemplar effect upstream in candidate generation; they do not establish that exemplars have no downstream effect on abstention or rejection.

5.4 Verification Instructions Can Shift Rejection

Some verification instructions acted downstream on the *discarding* of already-generated candidates. Verifying each awardWonBy candidate with three yes/no votes and dropping only unanimous-no candidates raised precision 0.241 $\rightarrow$ 0.280 and F1 0.172 $\rightarrow$ 0.186; the yes votes carried no information, so the usable signal was one-sided; this verifier is not retained in V19 (Sec. 3). A digit-check step inside an otherwise inert step-by-step procedure likewise cut votes for order-of-magnitude twin

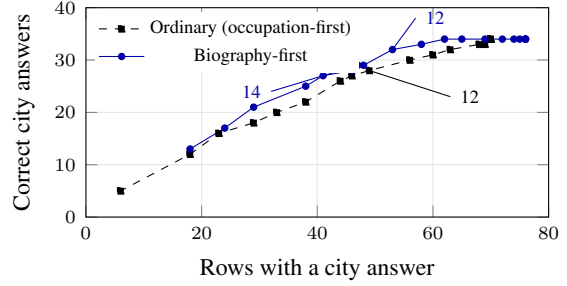


Figure 2: City coverage and correctness on 100 validation rows, 20 candidates each. Points sweep integer vote gates 1–20 on the saved pools; lines only connect observed points. Labels mark the fixed 12-vote comparison and V19’s 14-vote biography gate. Correct empty outputs are excluded from the vertical axis.

values from 38 to 32 with the oracle unchanged.

For personHasCityOfDeath, generating a one-line biography before the city field raises validation F1 from 0.55 to 0.61 at a fixed 12-of-20 gate; the change also helps on train (0.52 to 0.55) and a second validation seed (0.60). The parser in Sec. 3 turns generated living-person cues into empty candidates. The result therefore reflects the prompt format together with this fixed extraction rule, not an additional neural verifier.

Figure 2 recomputes the vote-gate sweep from saved validation pools. At 12 votes, correct city fills rise from 28 to 32 and correct empty outputs from 27 to 29, accounting for the six-point F1 gain. At matched coverage, biography also retains more correct cities: 25 vs. 22 at 38 filled rows, and 28 vs. 27 at 46. This supports effects on both retrieval and rejection, rather than merely a lower answer rate. The final 14-vote setting is marked separately; the figure is a validation analysis, not a test-set threshold search. Applying biography to area, capacity, and company gives no gain (0... – 2 rows); fifteen further city gates improve validation by at most two rows.

5.5 The Exemplar Effect Is Conditional

Verbalized recall is a *focusing* intervention: it sharpens the anchor distribution around the most frequently recalled value (oracle 89 $\rightarrow$ 84 as the mode improves), so its sign follows how reliable recall already is, the opposite way round from thinking. Thinking explores, gaining +9 rows where recall is unreliable (Qwen) and nothing where it is reliable (Mistral); verbalized recall does the opposite, zero or negative on Qwen and on Mistral’s weakest relation (hasCapacity oracle 60 $\rightarrow$ 47) but +6

Component	Δ val	Δ test	V19
recall exemplars (§5.3)	+6	+5	No
company folding (§3)	+1	+2	Yes
8B arbitration (§5.6)	+2	+1	No
train exemplars: area, capacity	+8	+5	Yes
capacity prompt (best of 6)	+2	-2	No
second arbiter (increment)	+1	0	No
tolerance rescoring	+2	-1	No
train exemplars: city	+4	-5	No
city two-stage merge	+11	-1	No

Table 4: Row-level validation and test deltas for interventions deployed in at least one submitted system, measured in their original experiment settings. V19 marks retention in the final pipeline, not the system on which the delta was measured: recall exemplars, company folding, and 8B arbitration here were measured on Mistral. Company folding is distinct from the later public alias graph. All final relation-specific choices are consolidated in Sec. 3 and Fig. 1.

on hasArea. In Magistral (Mistral’s RL-tuned sibling) it disappears: the verbalized form is copied at 99.9% compliance in three different channels, yet the anchor distribution does not move in any of them (70–72), suggesting that RL post-training may decouple response style from retrieval.

5.6 We Find No Usable Within-Model Selection Signal in the Features We Tested

Within one model’s candidate set, minority correct values were indistinguishable from minority wrong ones on every feature we measured (votes, dispersion, trace features, log-probabilities; base rate 8%), which is why judges, confidence weighting, and consensus schemes failed in our experiments. What did extract signal is *disagreement between prompts of different style*: on the ~ 25 rows where the recall and grounding prompts disagree by more than 5%, one aggregated decision from an unrelated-family 8B model resolves the tie, giving hasArea 78 $\rightarrow$ 80 and hasCapacity 26 $\rightarrow$ 28 on validation; of the four hasArea rows it changed, three were corrected and one harmed. Controls locate the conditions: same-style prompt pairs produce too little disagreement to arbitrate; promoting minority clusters by the arbiter’s mode is *negative*, since an 8B cannot overrule a 24B mode, only break ties; and a weak arbiter (Ministral-8B, mode accuracy 23 vs. Llama’s 53) gains nothing. The isolated v3 $\rightarrow$ v4 ablation confirms the transfer on test: capacity F1 0.204 $\rightarrow$ 0.214.

5.7 Which Validation Gains Transfer

Our submissions give per-component test measurements (Table 4). The components that transferred share two properties: each is *gated*, touching only rows that meet an explicit condition, and each gate uses information from outside the model—an unrelated-family vote, or the gold alias structure that makes two surface forms one count. The exemplar effect is the apparent exception: it changes what the backbone retrieves rather than adjudicating between candidates.

Non-transferring components include the best of six capacity prompts, a second arbiter, and a city merge that gained 11 validation rows (Table 4). Tolerance rescoring is especially instructive: values with ratio below $1.05/0.95 \approx 1.105$ admit one prediction correct for either. Shifting predictions to cover a runner-up cluster gained 1–2 validation rows across prompt pairs and arbiter seeds, but lost one test area row (0.800 $\rightarrow$ 0.790); we removed it in v11. Unlike the retained selection gates, this adjustment was applied globally. Later gates based on the same model’s unit agreement or answer-blind priming also admitted confident errors. A third predictor appeared late. When we replaced our exemplars with training answers, the same change helped hasArea and hasCapacity on all three seeds we ran and helped personHasCityOfDeath on one seed only (0.56/0.50/0.49); we deployed all three, and on test the two seed-stable relations transferred (+4 capacity rows) while the seed-unstable one cost five rows. We now treat single-seed differences below about four rows on a 100-row relation as unusable. A 100-row validation set may not reliably expose these failures; in our experiments, whether a component is gated, whether its gate looks outside the model, and whether its gain survives reseeding predicted transfer better than effect size or the presence of a mechanism-level rationale.

5.8 Residual Errors Look Like Values Learned From the Corpus

The residual errors we examined are not reasoning failures. Corfu (gold 626 km²) is answered as 592.9 by every Mistral-family instruct configuration, 593 by Qwen3.5, and 580 by Gemma: stable, model-specific wrong values that survive every prompt. Three observations support the hypothesis that they were learned from training data. First, *turnover across generations*: Qwen3.6 repairs 7 of Qwen3.5’s wrong rows and breaks 11

new ones with nothing else changed, and the union oracle reaches 89/100. Second, *unrelated training lineages solve mainstream-broken rows*: `llm-jp-4-32b` and `Aya Expanse` are far weaker overall yet each partially solves `Corfu`, which no mainstream-corpus model solves. Third, *instruction tuning costs recall*: under raw completion `Mistral 3.1 Base` scores 10/12-any against its own `instruct`’s 9, and the gap replicates on `Llama-3.1-8B` (mode accuracy 60 vs. 53 of 100).

The last comparison is the one our final system rests on, so we ran it without the prompt format confound: the instruction-tuned sibling of our backbone, queried in the *same* completion prompt formats with the same sampling and aggregation, scores 0.553 on validation against 0.665 for the pretrained checkpoint (hasArea 0.61 vs. 0.84, hasCapacity 0.24 vs. 0.31, personHasCityOfDeath 0.38 vs. 0.52), on two seeds. Since weights are the only difference, the loss belongs to instruction tuning rather than to the chat template, at least in this lineage—and the abstention that makes the biography prompt format work (Sec. 5.4) is absent in the tuned sibling, which fabricates instead. The ceiling is a property of the checkpoint, not of prompt engineering within one setting.

6 Related Work

Probing factual knowledge in LMs goes back to LAMA (Petroni et al., 2019); the LM-KBC series (Singhania et al., 2023; Kalo et al., 2024) extends it to KB construction with unconstrained cardinalities. Self-consistency (Wang et al., 2023) aggregates sampled answers by majority vote and ReWiSe (Albert-Roulhac and Zouaq, 2025) tuned it relation-wise; we extend it to empty answers, numeric tolerance, surface folding, and cross-family arbitration. Sprague et al. (2025) find chain-of-thought helps reasoning but not knowledge recall; in our numeric comparisons, example configuration shifts generated answers, while the wording comparison with added JSON examples differs by only one row. A chain-of-thought prompt format helps city of death, where a living-person cue supports abstention as well as correct city recall. Ma and Hewitt (2026) report prompting moves parametric knowledge access by 1–2% and RL fine-tuning by ~10%; our exemplar effect sits between, and fine-tuning is prohibited here. Yuchi et al. (2026) find probe-recoverable numeric knowledge exceeds

what models verbalize, one reading of our oracle-mode gaps, and Kalyan et al. (2021) and Park et al. (2022) document the magnitude and unit fragilities our components target.

7 Conclusion

V19 (Sec. 3) scored 0.6670, placing sixth. Backbone choice outweighed prompt interventions. Example configurations primarily shifted retrieval; verification affected rejection, and biography affected both. Gating, outside information, and seed stability helped interpret transfer within this task edition.

Limitations

Mechanism findings rely mainly on Qwen and Mistral; the final backbone is least ablated. Only exemplar, arbitration, and prompt format results have seed replications. The 12-row probe samples one system’s failures and was not tested on held-out models. Counts on 100-row relations vary by 1–2 rows. Only Sec. 5.8’s Gemma control fixes the prompt format; earlier Mistral/Llama comparisons do not. Transfer observations cover one task edition with repeated leaderboard feedback.

References

- Edouard Albert-Roulhac and Amal Zouaq. 2025. *ReWiSe: Relation-wise self-consistency for LLM probing*. In *Joint Proceedings of KBC-LM and LM-KBC @ ISWC 2025*, volume 4041 of *CEUR Workshop Proceedings*.
- 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 @ ISWC 2024*, volume 3853 of *CEUR Workshop Proceedings*, Baltimore, USA.
- Ashwin Kalyan, Abhinav Kumar, Arjun Chandrasekaran, Ashish Sabharwal, and Peter Clark. 2021. How much coffee was consumed during EMNLP 2019? Fermi problems: A new reasoning challenge for AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Melody Ma and John Hewitt. 2026. Improving parametric knowledge access in reasoning language models. ArXiv:2602.22193.
- Sungjin Park, Seungwoo Ryu, and Edward Choi. 2022. Do language models understand measurements? In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP 2019*, pages 2463–2473.
- Sneha Singhania, Jan-Christoph Kalo, Simon Razniewski, and Jeff Z. Pan. 2023. [Preface: LM-KBC challenge 2023](#). In *Joint Proceedings of the KBC-LM Workshop and the LM-KBC Challenge @ ISWC 2023*, volume 3577 of *CEUR Workshop Proceedings*, Athens, Greece.
- Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Fengting Yuchi, Li Du, and Jason Eisner. 2026. LLMs know more about numbers than they can say. In *Proceedings of EACL 2026 (Short Papers)*. ArXiv:2602.07812.