

Relation-Specific Elicitation of a Base Model under a 32B Closed-Book Budget for LM-KBC 2026

Ruggero Marino Lazzaroni Tim Booker Pau Soler Valadés

Mathias Angermaier Kirill Solovev

University of Graz, Austria

{ruggero.lazzaroni,kirill.solovev}@uni-graz.at

Abstract

We describe a system for the LM-KBC 2026 shared task on closed-book knowledge base construction. Given a subject and relation, the system tries to predict all corresponding objects without retrieval or additional training and remains within the 32-billion-parameter limit. The system uses the base model gemma-4-31B with separate prompting and aggregation rules for six relations; few-shot demonstrations and selection thresholds come from the train set and are checked on the validation set. It obtains 0.6843 global F1 on the official validation set and 0.6961 on the official test set, compared with 0.308 for our run of the organisers’ baseline on validation. The largest per-relation gains over our earlier baseline configuration come from a company-listing prompt with explicit abstention and a log-stratified area prompt, plus a vote-share abstention threshold for empty-heavy relations (like `personHasCityOfDeath`) that recovers most of the empty-case precision. Across more than 30 prompting, verification, and aggregation configurations, further inference rarely recovers facts that are absent from the generated candidates. These results indicate that relation-specific elicitation improves closed-book extraction, while candidate coverage and selection remain the main limitations. Code, prompts, and configuration files are available at <https://github.com/ruggsea/lmkbc2026-submission>.

1 Introduction

Research shows that graphs generation methods appear to still be unable to produce complete knowledge graphs, and pretrained language models provide one possible source of missing facts (Petroni et al., 2019; Kalo et al., 2023, 2025). The LM-KBC shared task evaluates this use directly: given a subject and relation, a system must return every corresponding object using only knowledge stored in the model’s weights. Unlike single-fact probing benchmarks such as LAMA, this year’s

LM-KBC permits zero, one, or several objects. The 2026 task matches normalised strings through alias expansion and bipartite matching, then computes item-level macro-F1 with a $\pm 5\%$ tolerance for numeric answers.

We set out to implement a system respecting the three constraints that the track imposes. First, inference is closed-book: the system cannot use retrieval, internet access, external factual corpora, or knowledge-base lookups, while non-neural filtering, normalisation, deduplication, and aggregation are permitted (section 3.3). We therefore use an external data source (Wikidata) only offline, and never at query or inference time, for two purposes: to construct the auxiliary pseudoval set described in appendix B, and to assemble the fixed demonstration bank used by the `hasArea` prompt (section 4.3). Assembling a demonstration bank offline is the scaled-up equivalent of hand-writing few-shot examples: the bank is fixed before any prediction is made, it is identical for every query, and nothing is looked up per subject. We further verify that it cannot supply an answer, excluding every subject that occurs in official train, validation, or test, and removing the query entity itself at prompt-construction time; both checks match on Wikidata entity identity rather than surface string, so aliases of the same entity are caught as well. Second, all neural components in the inference path must contain at most 32B parameters in total; multi-step and multi-model systems are permitted within this combined limit, while our system uses one model. Unlike the 2025 task, which prescribed Qwen3-8B, the 2026 task permits any open-weight model within the parameter budget (section 2). Third, the system cannot use fine-tuning, LoRA, probes, or soft prompts. The six relations include null-prone string sets (`countryLandBordersCountry`, `companyTradesAtStockExchange`, and `personHasCityOfDeath`), long recipient

lists (awardWonBy), and numeric values (hasArea and hasCapacity).

We use one permitted base model as the sole parametric source and assign a separate elicitation method to each relation. The choice of base model is guided by the intuition that base models are most suited to few shot prompting and aggregation, plus they do not suffer from partial catastrophic forgetting and sycophancy that post training may introduce. Our one model design accommodates the different cardinalities, null rates, and output types without imposing one prompt and aggregation rule on all relations. The relation-specific elicitation methods are designed to address the different characteristics of the relations. Across the tested models, prompts, aggregators, and decoding strategies, facts absent from the sampled outputs were rarely recovered through further prompting. This pattern is consistent with the knowledge gaps reported by the 2025 systems and with the distinction between encoded and expressed knowledge in Inside-Out (Albert-Roulhac and Zouaq, 2025; He and Razniewski, 2025; Gekhman et al., 2025).

Our system obtains 0.6843 global F1 on the 475-item official validation set and 0.6961 on the official test set. Going deeper into null-item analysis, we examine the abstention rule for personHasCityOfDeath and show that its vote-share threshold determines every abstention decision, while a separate yes/no condition changes no prediction (section 4.4). We also compare more than 30 selection, verification, and generation configurations across the six relations. The comparison identifies candidate generation as the main limit for awardWonBy and candidate selection as a major limit for hasCapacity (section 5.3). Code, prompts, and configuration files are available at <https://github.com/ruggsea/lmkbc2026-submission>.

2 Related Work

LM-KBC. The LM-KBC series has moved from disambiguated single-object probing towards open-cardinality construction (Kalo et al., 2023, 2025). The 2025 task prescribed Qwen3-8B, whereas the 2026 task permits model and method selection within the parameter budget. Of the five systems submitted in 2025, the winning ReWiSe combined synthetic and human chain-of-thought demonstrations with relation-specific self-consistency (Albert-Roulhac and Zouaq, 2025). The runner-up used

a Self-RAG-style describe-then-extract pipeline and divided awardWonBy into smaller subqueries (He and Razniewski, 2025). He and Razniewski (2025) attribute the persistent awardWonBy failures to missing knowledge rather than retrieval difficulty, reporting a “complete knowledge void” for post-1990 recipients in a Max Planck Medal case study, while Albert-Roulhac and Zouaq (2025) leave open whether the required awardWonBy and hasCapacity facts are present in the model weights. The best 2025 F1 values were 0.32 for hasArea, 0.15 for hasCapacity, and 0.18 for awardWonBy, the three relations the organisers identify as remaining particularly challenging; because 2025 prescribed an 8B model, these values contextualise rather than compare with 2026 scores. For personHasCityOfDeath, the 2025 results separate the baseline from the best system almost entirely by precision (0.08 vs 0.85) at nearly identical recall (Kalo et al., 2025), consistent with false-positive cities on empty-gold subjects dominating the baseline errors. This asymmetry also motivates our abstention analysis in section 4.4. In the 2022 task, a BERT baseline with a 0.5 likelihood threshold scored 31.08 % test F1, while predicting an empty set for every subject already produced 18.00 % (Singhanian et al., 2022). Thus, abstention can account for a substantial part of the attainable score on null-prone relations.

Parametric knowledge and elicitation. Self-consistency has been shown to stabilise answers that the model already produces, while relation-specific vote thresholds adapt aggregation to different cardinalities (Wang et al., 2023). The Inside-Out framework has found that models encode more factual knowledge than they express, with an average relative gap near 40 % (Gekhman et al., 2025). This result is consistent with our finding that additional samples rarely recover city-of-death answers that never enter the draw pool, which we discuss in section 4.4. Two studies using masked single-token BERT-large probes over multi-valued Wikidata relations provide related evidence on cardinality estimation and self-verification. Singhanian et al. (2023) report 49.5 F1 for top- k thresholding, 35.1 for a separate count probe, and 8.0 for a yes/no verification probe. The count probe almost always predicts more than ten objects, while the verification probe accepts nearly every candidate. Among missed objects, 26.9 % are generated below the cutoff and 27.75 % are never generated, which

parallels the two sources of awardWonBy recall loss discussed in sections 4.6 and 5.3. In a separate analysis of 41 relations, Veseli et al. (2023) report near-zero recall at 90 % precision for deathPlace and sharesBorders from BERT-scale probes. Both relations are far more accessible to our 31B base model (table 1), although model scale and elicitation format make this comparison unfair.

3 Task and Data

3.1 Official task

Each item contains a (SubjectEntity, Relation) pair, and the system returns a possibly empty list of ObjectEntities. The evaluator normalises strings, expands aliases, and applies bipartite matching; numeric predictions are correct when they fall within $\pm 5\%$ of gold. We run the organisers’ released Qwen3.5-9B baseline¹ ourselves and obtain 0.308 macro-F1 on validation, against the 0.313 reported in the task repository; the small difference reflects sampling and a later validation revision. We report our own run throughout so that every column of table 1 is scored on the same labels. Validation contains 68 border items, 100 company items, 100 cityDeath items, 100 hasArea items, 97 hasCapacity items, and 10 award items. Company and cityDeath contain 38 and 39 empty-gold items, respectively, while the award gold lists are long and sparse.

3.2 Evaluation surfaces and split discipline

We use official train to select methods and thresholds, official validation for reporting and acceptance checks, and a pseudoval set for early threshold and transfer checks. We constructed the pseudoval set by scraping approximately 2400 Wikidata rows with empty rates and entity types matched to the official distribution (appendix B). All numeric thresholds in the submitted system are selected on official train. However, some development decisions retained a component only when it did not reduce validation performance, as stated in the Limitations section. Therefore, our research approach was to go through different combinations of prompts, thresholds, and aggregation rules over train, to then probe on pseudoval and on the validation set in order to select the best configuration possible at that time of the development. When a change improved validation but reduced pseudoval

¹<https://github.com/lm-kbc/dataset2026>

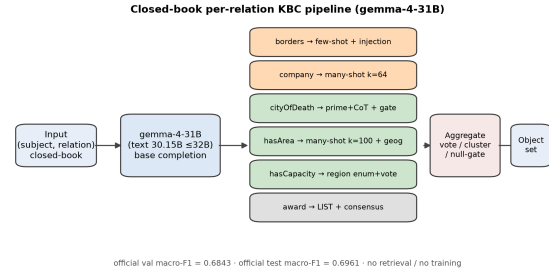


Figure 1: System overview with one permitted base model, six relation-specific prompt prefixes, stochastic sampling, and non-neural aggregation by voting, numeric clustering, or abstention.

performance, we treated the result as evidence of overfitting (e.g. section 4.1).

3.3 Aggregation primitives

For string sets, our overall approach was to sample n completions (“draws”), parse and normalise their candidates, and keep an object (e.g. a country’s name in the case of countryLandBordersCountry) when it appears in at least a fraction τ of draws. Empty and abstaining draws count against the object under this agreement threshold. For numeric relations, we parse the generated numbers and group values within a relative width. For areas, clustering operates in $\log_{10}$ space with width $w=0.05$, and the system returns the median of the largest cluster.

4 Method

After sweeping through different model, prompts combinations using some baseline prompts and thresholds, we opted to serve gemma-4-31B (Gemma Team, Google DeepMind, 2026) as a base completer that continues a plain-text prefix without chat roles. Direct counting of the safetensors gives 31.273B parameters: 30.697B in the language model plus 0.570B in the vision tower and 0.006B in the vision embedding (that we never invoke at inference), giving a final effective parameter count of 31.273B. The complete multimodal checkpoint therefore remains within the 32B limit, as the system never necessitates the multimodal capabilities of the model. For each item, the model continues a relation-specific prefix through n stochastic draws at a fixed temperature. A set of per-relation non-neural rules then parses the candidates and aggregates them into the final object set.

Prompting, aggregation, and thresholding.

The final prompts and aggregation or thresholding methods used in the submitted system were selected through independent rounds of iteration for each relation. Selection and tuning for prompting, aggregation, and thresholding were performed independently per relation, according to the selection criteria discussed above. The full set of prompts and selected methods used in the final system are provided in Appendix A.

In addition, the relation-specific methods address distinct output structures. The capacity prompt enumerates venues with the same name before selecting the venue in the stated place, while the award prompt continues a year-by-year recipient list. Figure 1 summarises the system, while table 1 reports validation and test F1: in the table we compare the scores of the submitted system with a previous baseline we instantiated using the gemma-2-27B model.

4.1 countryLandBordersCountry

A short few-shot prefix of three hand-picked demonstrations shows semicolon-separated land neighbours and uses none for islands (appendix A.1). We sample 30 draws per subject at temperature 0.7. After tuning on the train set, we opt to keep a neighbour when it appears in at least $\tau=0.35$ of draws and merge surface variants with the official alias table. Because land borders are well represented in pretraining, the main elicitation problem is over-inclusion. An island gating approach, larger demonstration banks, agreement across prompt variants, and reverse-border filtering did not improve train or/nor pseudoval performance over this recipe. Only one additional filter was found to be potentially beneficial: reverse-border filtering drops $A \rightarrow B$ unless the system also predicts $B \rightarrow A$; it improved validation by 0.008 but reduced pseudoval performance by 0.014, therefore we did not include it in the submitted system.

4.2 companyTradesAtStockExchange

The many-shot company prompt uses an “issuer listing review” prefix with 64 sentences about listed and explicitly unlisted issuers, followed by {Company} is... (appendix A.2). We sample 100 draws per subject at temperature 0.7. Abstention phrases produce an empty vote; otherwise, we extract the named exchanges and retain those with agreement $\tau=0.3$. Because approximately one third of subjects in the datasets have empty gold sets, the

prompt places “not listed” outcomes in the same format as positive listings. The 64 demonstrations also match the prose used in listing disclosures for listed firms. A separate yes/no listing gate, binary probes for individual exchanges, and cross-method agreement did not seem to improve this configuration.

4.3 hasArea

The area prompt elicits a value in a Wikipedia-style lead sentence rather than through a bare numeric fill-in. From a bank of 331 entities with known areas, collected offline from Wikidata and filtered to exclude every subject appearing in the official splits, we remove the query, order the remaining entities by $\log_{10}(\text{area})$, and select 100 evenly spaced demonstrations (appendix A.3). Each demonstration names the entity type and gives the area in a short lead sentence. We sample 100 draws per subject at temperature 0.8, parse the first area-like number, and return the median of the largest cluster with width $w=0.05$.

The submitted prediction combines the lead-sentence draws with an independently sampled second channel, which we call the geographic-entry pool. This channel imitates the “Geography of X” section of an encyclopedia entry rather than a lead sentence: a prefix of three hand-written entries, none of which occurs in any official split, each open with the heading Geography of {Entity}, describe the entity in one sentence, and close with an explicit Total land area: figure in km^2 (appendix A.9). The query continues the same pattern, and we parse and cluster its 50 draws, sampled at temperature 0.8, exactly as for the lead-sentence pool. The two channels therefore differ in the surface form that carries the number, an encyclopedia section heading against a lead sentence, while sharing the model, the parser, and the clustering rule. When the two clustered medians differ by more than w , the system returns the geographic-entry value. This override fires for 28 of 200 pooled train and validation rows (0.140) and for 15 of 100 validation rows, of which nine are decisive. Its win rate is not statistically distinguishable from chance. The geographic-entry value wins 55.6 % of decisive validation rows, with a bootstrap 95 % CI of [0.22, 0.89]. On an independent larger split, it wins 34.2 % of 38 decisive rows, with a CI of [0.18, 0.50]. The resulting validation predictions score 0.7800 F1 for hasArea in table 1.

Relation	Baseline	gemma-2 reference		gemma-4 system		Test F1
	Validation F1	Validation P/R	Validation F1	Validation P/R	Validation F1	
borders	0.662	0.984 / 0.986	0.982	0.995 / 1.000	0.9970	0.9753
company	0.366	0.810 / 0.863	0.7519	0.878 / 0.903	0.8386	0.8470
hasArea	0.290	0.740 / 0.670	0.6700	0.850 / 0.780	0.7800	0.8500
cityDeath	0.180	0.910 / 0.580	0.5633	0.870 / 0.600	0.5800	0.6000
hasCapacity	0.180	0.402 / 0.402	0.4021	0.361 / 0.361	0.3608	0.3265
award	0.087	0.295 / 0.297	0.244	0.450 / 0.212	0.2374	0.3609
All	0.308	0.747 / 0.674	0.6458	0.773 / 0.702	0.6843	0.6961

Table 1: Per-relation scores. Precision and recall are available for validation; the final column reports official test F1. The gemma-2 reference and gemma-4 system use the same official validation set.

Log-stratified lead prompting recovers magnitudes more reliably than bare numeric recitation. The strongest alternative continues a Wikipedia-infobox stub ending at `| area_km2 =` (appendix A.10). Alternative units, Fermi-style estimates, density vetoes, and pooling the two prompts did not change train or pseudoval scores.

4.4 personHasCityOfDeath

Three prompt pools contribute to the decision (appendices A.4 to A.6). The prime pool is a few-shot biographical fact file ending at `City of death:`, while the status pool generates `Status: Alive|Deceased` and `City of Death:` fields. The yes/no pool asks whether the person is deceased with a recorded city and yields the yes-vote fraction y . We sample 30 prime draws, 50 status draws, and 20 yes/no draws per subject, all at temperature 0.7. The system pools the prime and status draws and identifies the modal city. It emits this city only when $y \geq \theta$ and $v/N \geq \phi$, where v is the modal-city vote count and N is the number of pooled city draws. We set $\theta = 0.1$ and $\phi = 0.3625$. The value of ϕ equals $29/(30+50)$, which is the large- n limit of an absolute threshold of 29 votes with 30 prime draws and 50 status draws. Among validation rows that satisfy the vote-share condition, the minimum y is 0.325, which is well above $\theta = 0.1$. Consequently, no row that passes the vote-share condition fails the yes/no condition, and setting $\theta = 0$ changes none of the 100 predictions. Abstention therefore depends entirely on the vote-share threshold. On a fixed validation draw pool, removing this threshold reduces F1 from 0.5600 to 0.420, whereas removing the yes/no condition changes nothing.

Approximately 39 % of validation subjects have empty gold sets, and any predicted city for these subjects scores zero. The vote-share threshold therefore reduces false-positive cities without a separate

verification step. The gemma-2 reference configuration (section 5.2) reaches 0.563 F1, compared with 0.5800 for our validation predictions. We interpret this small difference as evidence that the rule primarily improves empty-side precision, rather than as a large relation-level gain.

4.5 hasCapacity

Many subjects follow the form "VenueName in Place", although several venues can share the same bare name. The capacity prompt uses few-shot examples to enumerate venues with that name, select the venue in the stated place, and emit its capacity (appendix A.7). We sample 30 draws per subject at temperature 0.8, parse the final capacity number, cluster values within a 2.5 % relative width, and return the median of the largest cluster. Two simpler prefixes perform worse on venue homonyms: a direct prompt omits enumeration, while a description prompt adds a `Description:` field before `Capacity:`. The submitted system uses the cluster median on both splits, and the validation figures we report throughout are for that configuration. Rank-one voting within the same draw pool scores 0.3918 F1 on the 97 validation items, but it is not the submitted rule and we do not report it in table 1.

Controlled probes indicate that the generated samples contain relevant capacities that the aggregation rule does not select. On official train, a shortlist oracle marks an item correct when its gold capacity occurs in one of the three densest clusters from a large description-prompt pool. This oracle scores 0.470 F1, compared with 0.206 ± 0.038 for a magnitude-matched permutation null, giving a chance-corrected $\Delta \approx 0.26$. Without gold labels, digit probes, scientific notation, logits and function-vector read-outs, and grammar-constrained decoding all perform worse than the capacity prompt. Tighter clustering also does not

address most errors. Only 16.7 % of incorrect train rows (11/66) and 20.7 % of incorrect validation rows (12/58) fall within 5 % to 20 % of gold. The remaining errors exceed 20 %, with an approximately 2× band forming the largest interpretable group (27.3 % on train and 24.1 % on validation) and no stable unit-scale pattern.

4.6 awardWonBy

The Wikipedia-style award prefix demonstrates year-by-year recipients for two well-known awards and continues the same schema for the query award (appendix A.8). We sample 20 draws per award at temperature 0.8 and retain names with draw support $\geq \tau = 0.46$. The recipe scores 0.2374 F1 on the ten validation awards. These awards are specialised and their gold lists are sparse, so a real winner omitted from gold becomes a false positive when predicted. Relative to the gemma-2 reference configuration (section 5.2), the recipient-list prompt increases precision but reduces recall, leaving F1 similar in table 1. Generate-then-verify cascades reduced precision without a corresponding recall gain.

5 Experiments

5.1 Setup

We run vLLM offline on four H100 or four L40S GPUs with tensor parallelism of four. Subject-bootstrap intervals for global F1 are approximately ± 0.04 . A complete regeneration with unchanged prompts and thresholds but new stochastic draws changes official-validation global F1 by an estimated standard deviation of 0.0072, obtained by combining per-relation regeneration estimates in quadrature. We therefore treat single-run global differences below 0.014 as unresolved unless they also occur on train or pseudoval. Unless stated otherwise, every prompt comparison in this section is run with the same number of draws and the same temperature as the corresponding submitted prompt in section 4, so that alternatives differ only in the prompt itself.

5.2 Main results

Table 1 reports official validation and test results for our run of the organisers’ baseline, our submitted gemma-4 system, and a gemma-2 reference configuration: the strongest system from our earlier development, which pairs gemma-2-27b-it (Gemma Team, Google DeepMind, 2024)

with its own optimized (via iteratives sweeps) per-relation prompts and thresholds. Both model configurations use the same official validation set. Relative to the gemma-2 reference, our system gains 0.087 F1 for companyTradesAtStockExchange and 0.110 for hasArea, while hasCapacity decreases from 0.4021 to 0.3608. For awardWonBy, the reference configuration scores 0.244 and our system scores 0.2374. Because validation contains only ten awards, we interpret this 0.007 difference as a tie within subject-level noise. These comparisons are reference points rather than controlled ablations, because the base model and the relation-specific recipes differ together.

5.3 Error analysis

The relations exhibit different failure modes: city-Death trades filled-case recall for empty-case precision, company is well calibrated on empty cases, the capacity method often predicts the wrong magnitude, and awardWonBy returns short lists relative to some gold sets. Appendix C reports the corresponding counts, per-award results, and negative results.

6 Discussion and Conclusion

Our closed-book LM-KBC 2026 system uses one permitted gemma-4-31B model with relation-specific elicitation and scores 0.6843 global F1 on official validation, compared with the 0.308 baseline. The official test score is 0.6961 F1. The company prompt, area prompt, and cityDeath vote-share threshold improve relations when the required facts occur in the generated candidates. However, missing candidates continue to limit award-recipient recall, while capacity errors remain dominated by selecting among generated magnitudes. The cityDeath threshold remains in the submitted system because it removes false positives for empty-gold items, while its separate yes/no condition changes no predictions (section 4.4). The 2025 systems reported the same difficult relations under an 8B mandate, and these relations remain the hardest for our 32B-budget system; verification, multi-agent methods, date anchors, and representation read-outs provide no consistent improvement. Two controls limit overfitting in this analysis: pseudoval matches official empty rates and entity types (appendix B), and the split discipline in section 3.2 rejects several validation-only gains that reverse on pseudoval.

Limitations

The awardWonBy validation set contains only ten subjects, so its relation-level differences are imprecise. While building a representative set for tuning, we found that matching pseudoval empty rates reduces but does not remove distribution shift.

Regarding model selection, our initial sweeps did not cover every model below 32B. Retrieval, models above 32B, and training on the target distribution fall outside the task rules and could change the attainable performance.

The reported validation score is one stochastic realisation rather than an expectation given the non-deterministic aspect of some of our prompting approaches. A complete regeneration changes official-validation global F1 by an estimated standard deviation of 0.0072. This estimate comes from a separate stochastic regeneration and was not recomputed for every reported configuration. Furthermore, although threshold values were not fitted to official-validation gold, several development decisions retained a component only when it did not reduce validation performance. This process makes the validation result an optimistic estimate of held-out performance, and we do not have a sufficiently precise train-only floor to quantify the optimism. The official test score therefore provides the primary independent confirmation of the submitted system.

Declaration on Generative AI

During the preparation of this work, the authors used large language models for refining and polishing the draft and for help with the literature organisation. The authors reviewed and edited all content and take full responsibility for the publication. The authors' evaluation pipeline, rather than generative tools, produced the experimental results, prompts, and scores.

Acknowledgements

This work was supported by VALPOP (EU Horizon Europe grant 101177310). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

References

- Edouard Albert-Roulhac and Amal Zouaq. 2025. [Re-WiSe: Relation-wise self-consistency for LLM probing](#). In *Joint Proceedings of the 3rd Workshop on Knowledge Base Construction from Pre-Trained Language Models and the 4th Challenge on Language Models for Knowledge Base Construction (KBC-LM+LM-KBC 2025)*, volume 4041 of *CEUR Workshop Proceedings*.
- Zorik Gekhman, Eyal Ben-David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. [Inside-out: Hidden factual knowledge in LLMs](#). In *Conference on Language Modeling (COLM)*.
- Gemma Team, Google DeepMind. 2024. [Gemma 2: Improving open language models at a practical size](#). Preprint, arXiv:2408.00118.
- Gemma Team, Google DeepMind. 2026. [Gemma 4 technical report](#). Preprint, arXiv:2607.02770.
- Jingbo He and Simon Razniewski. 2025. [Combining self-retrieval-augmented generation with divide-and-conquer for language model-based knowledge base construction](#). In *Joint Proceedings of the 3rd Workshop on Knowledge Base Construction from Pre-Trained Language Models and the 4th Challenge on Language Models for Knowledge Base Construction (KBC-LM+LM-KBC 2025)*, volume 4041 of *CEUR Workshop Proceedings*.
- 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 the 3rd Workshop on Knowledge Base Construction from Pre-Trained Language Models and the 4th Challenge on Language Models for Knowledge Base Construction (KBC-LM+LM-KBC 2025)*, volume 4041 of *CEUR Workshop Proceedings*.
- 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 the KBC-LM Workshop and the LM-KBC Challenge @ ISWC 2023*, volume 3577 of *CEUR Workshop Proceedings*.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Sneha Singhania, Tuan-Phong Nguyen, and Simon Razniewski. 2022. [LM-KBC: Knowledge base construction from pre-trained language models](#). In *Knowledge Base Construction from Pre-trained Language*

Models 2022 (LM-KBC 2022), volume 3274 of *CEUR Workshop Proceedings*, pages 1–10.

Sneha Singhania, Simon Razniewski, and Gerhard Weikum. 2023. [Extracting multi-valued relations from language models](#). In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 139–154, Toronto, Canada. Association for Computational Linguistics.

Blerta Veseli, Sneha Singhania, Simon Razniewski, and Gerhard Weikum. 2023. [Evaluating language models for knowledge base completion](#). In *The Semantic Web: 20th International Conference, ESWC 2023*, volume 13870 of *Lecture Notes in Computer Science*, pages 227–243. Springer.

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 *International Conference on Learning Representations (ICLR)*.

A Prompts for submitted methods

Unless stated otherwise, each prompt is a base-completion prefix that the model continues. The tokens {name}, {s}, {query}, and {award} mark the subject slot. Ellipses written as [...] indicate truncated demonstration lists, while the code release contains the complete banks.

A.1 Borders — set few-shot

Country: France
Land border countries: Belgium;
Luxembourg; Germany; Switzerland;
Italy; Monaco; Spain; Andorra

Country: Japan
Land border countries: none

Country: Bolivia
Land border countries: Brazil;
Paraguay; Argentina; Chile; Peru

Country: {name}
Land border countries:

Generation stops at a blank line or a new Country: header. The parser treats none, n/a, no land borders, and island nation as empty.

A.2 Company prompt with 64 demonstrations

The complete template contains 64 demonstration lines under this header and schema:

MSCI Global Investable Market Indices
-- issuer listing review.
For each issuer, the review records
the stock exchange or exchanges on
which its shares are publicly listed,
or notes that the issuer is not
publicly listed.

Shield On Service is publicly listed
on the Indonesia Stock Exchange.
JYP Entertainment is publicly listed
on the Korea Exchange (Kosdaq).
General Mills is publicly listed on
the New York Stock Exchange.
Orange County Choppers is not
publicly listed.
Vrtaky is not publicly listed.
[... 54 additional lines ...]
{s} is

If the first line contains *not publicly listed*, *not listed*, *not admitted*, *not traded*, *does not trade*, or bare *none*, matched without case sensitivity, the parser returns []. Otherwise, it extracts the span after *(listed|registered|quoted|trades?|traded|trading) on (the)?*, splits the span on commas, semicolons, or *and*, and removes parts matching the abstention list. The final vote threshold is $\tau=0.3$.

A.3 Area prompt with 100 demonstrations

Each demonstration comes from a 331-entity area bank that is stratified by log area at inference:

Monaco is a country. It has an area
of 2.02 km².

Manhattan is a city. It has an area
of 59.1 km².

Andorra is a country. It has an area
of 468 km².

Belgium is a country. It has an area
of 30528 km².

{query} is

The type phrase is *country*, *island*, *lake*, *city*, *mountain range*, or *geographic entity*, as appropriate. The parser extracts the first area-like number in km², clusters values at relative width $w=0.05$, and returns the median of the largest cluster.

A.4 City-of-death prime pool

Demonstrations are complete fact blocks ending at City of death:. The query prefix stops after Nationality:, and the model continues the remaining fields through City of death: in the same schema.

Person: Albert Einstein
Nationality: German-born, later
Swiss and American
Field: theoretical physics
Years: 1879–1955
Final-years residence: Princeton,
New Jersey
Buried: cremated, ashes scattered
near Princeton
City of death: Princeton

Person: Keanu Reeves
Nationality: Canadian
Field: acting

Years: born 1964, still alive
Final-years residence: Los Angeles
Buried: N/A
City of death: N/A

Person: {name}

Nationality:

The parser reads the completed City of death: field and treats N/A, none, and alive as empty votes.

A.5 City-of-death status pool

Fact: Albert Einstein
Status: Deceased
City of Death: Princeton

Fact: Taylor Swift
Status: Alive
City of Death: N/A

Fact: {name}

Status:

The parser reads the completed City of Death: field and treats N/A, none, and Alive without a city as empty votes. Status-pool and prime-pool votes are combined before applying the emission rule in [section 4.4](#).

A.6 City-of-death yes/no pool

Q: Is Albert Einstein deceased, and is his specific city of death publicly recorded as an established fact?

A: Yes

Q: Is Keanu Reeves deceased, and is his specific city of death publicly recorded as an established fact?

A: No

Q: Is {name} deceased, and is their specific city of death publicly recorded as an established fact?

A:

We compute $y = n_{\text{yes}} / (n_{\text{yes}} + n_{\text{no}})$.

A.7 Capacity prompt and alternatives

The submitted prompt uses three hand-written demonstrations in this schema:

Venue: Wembley Stadium in London
Venues named Wembley Stadium: Wembley Stadium in London, England (national

stadium); Wembley Stadium in Melbourne, Australia (minor arena).
In region: Wembley Stadium in London, England.
Capacity: 90,000

[... two further demonstrations:
Olympic Stadium in Berlin (74,475)
and Goodwin Field in California (3,500) ...]

Venue: {full_subject}

Venues named {bare_name}:

A weaker direct prompt uses a bare few-shot Venue: ... / Capacity: format without enumerating venues that share a name. A weaker description prompt inserts a one-line Description: field before Capacity:, and the query stops after Description: so that the model emits a description followed by a number.

A.8 Award-recipient list completion

List of recipients of the Pritzker Architecture Prize:
1979 - Philip Johnson
1980 - Luis Barragan
[... additional years ...]

List of recipients of the {award}:

The parser reads year -- name lines and retains names with draw support $\geq \tau = 0.46$.

A.8.1 Per-award error detail

Two “honorary doctor of {university}” awards score zero against gold lists of 40 and 53 names. The Presidential Medal of Freedom has the largest precision-recall difference, with precision 0.533 and recall 0.013 against a 610-name gold list. Its 24-name prediction cannot cover this gold set.

A.9 Area geographic-entry channel

The second hasArea channel described in [section 4.3](#) continues an encyclopedia section rather than a lead sentence. The prefix holds three hand-written entries, none of which occurs in official train, validation, or test:

Geography of Tasmania
Tasmania is an island state of Australia south of the mainland.
Total land area: 68,401 km².

Geography of Sardinia

Sardinia is a large island in the Mediterranean west of Italy.
Total land area: 24,090 km².

Geography of Hokkaido

Hokkaido is the northernmost main island of Japan.
Total land area: 83,424 km².

Geography of {query}

{query} is

Draws stop at a blank line or the next heading. We parse the first area-like number and cluster the draws with the same $\log_{10}$ rule and width $w=0.05$ used for the lead-sentence pool, so the two channels differ only in the surface form that carries the number.

A.10 Area source-anchor alternative

The source-anchor prompt used for comparison in [section 4.3](#) continues this infobox wikitext:

```
{{Infobox island
| name = Manhattan
| area_km2 = 59.1
| country = United States
}}
```

```
{{Infobox
| name = {query}
| area_km2 =
```

The prompt uses the same $\log_{10}$ clustering rule as the lead-sentence prompt, with $w=0.05$.

B Pseudoval composition

Pseudoval is an offline Wikidata-derived gold file of approximately 2400 rows constructed through SPARQL. We remove every (SubjectEntity, Relation) pair that also occurs in official train or validation. [Table 2](#) reports its relation-specific construction.

Matching the official empty rates and entity types is necessary for useful transfer estimates. An unmatched construction with almost no empty company cases and too many protected areas produced a cityDeath threshold-ranking correlation of approximately $r = -0.02$ with official validation. With the matching rules in [table 2](#), simple-baseline rankings correlate with official validation at $r > 0.6$ for all six relations.

Relation	N	Empty%	Rule
borders	84	~13 %	land neighbours via P47 $\cap$ country
company	293	~44 %	P414 present vs. absent humans w/wo place-of-death P20
cityDeath	250	~46 %	islands/lakes with P2046
hasArea	842	0 %	stadium/arena subclass, P1083
hasCapacity	800	0 %	award subclass, winner list P1346
award	158	0 %	

Table 2: Pseudoval composition by relation.

C Additional error and mechanism analyses

Negative results. None of the unsuccessful mechanism families exceeds the submitted relation-specific recipe on its development surface. Self-consistency saturates, verification and multi-agent configurations provide no consistent improvement, and the tested cross-relation gates fail to improve their baselines. Date anchors produce only variation within noise, two-model systems within 32B lose 0.06 to 0.10 global F1, and representation read-outs perform worse than the capacity and award methods. Following [He and Razniewski \(2025\)](#), these results distinguish facts that are generated but poorly assembled from facts that never enter the candidate pool. The company and area prompts improve the first case. Missing award recipients limit candidate coverage, while capacity errors also reflect selection among generated magnitudes. The paragraphs below summarise more than 30 tested strategies and the corresponding sample-count sweeps.

Error distributions. On official validation, city-Death abstains correctly for 37 of 39 empty-gold subjects, producing two false positives, but it also abstains incorrectly for 29 of 61 filled-gold subjects. The company method abstains correctly for 32 of 38 empty-gold subjects and incorrectly for 3 of 62 non-empty subjects. For a single numeric prediction, precision, recall, and F1 are identical, so these errors reflect incorrect magnitudes rather than output format. For the ten awardWonBy items, the system never returns an empty list. The predicted-list length ranges from 7 to 70 names around a mean of 24.3. Performance varies by award; the Turing Award reaches precision 0.986 and recall 0.852 against 81 gold recipients. The number of gold recipients relative to the fixed shortlist length therefore constrains performance more than one

uniform precision-recall trade-off. [Appendix A.8.1](#) reports the two items with zero F1 and the largest precision-recall difference.

Mechanism comparison. Set-relation self-consistency stabilises near $n=40$, while award recall remains nearly constant across broad sample-count sweeps. Judge, debate, panel, self-refinement, and generate-then-verify configurations provide no consistent improvement over the strongest relation-specific recipe. The separate company and border gates also fail to improve their baselines; we do not assume that this result transfers to the distinct cityDeath vote-share rule. Date anchors across 132 months produce only variation within noise. A larger model proposing candidates for a smaller filter model within the combined 32B budget more often removes correct predictions than recovers missing objects, with typical global losses of 0.06 to 0.10. Logit-lens, function-vector, and related hidden-state read-outs perform worse than the submitted methods for hasCapacity and awardWonBy. In total, we test more than 30 aggregation and selection strategies across all six relations by varying draw counts, thresholds, numeric-cluster widths, candidate filters, and cross-relation gates. For hasCapacity and awardWonBy, every tested selector performs below or within noise of the submitted recipe. No tested selection or aggregation change improves the reported development results under our acceptance criterion. The border island gate reproduces empty cases that the few-shot system already handles, while the company side question is not easier than predicting the listing itself. Taken together, these comparisons support the narrower conclusion that further prompting and selection do not reliably recover facts absent from the generated candidates.