

Pre-training Aware Directed Reasoning for Closed-Book Knowledge Base Construction

Mikołaj Tym

Poznań University of Economics and Business

Poznań, Poland

mikolaj.tym@ue.poznan.pl

Abstract

Extracting facts from the parametric memory of Large Language Models (LLMs) for Automated Knowledge Base Completion (AKBC) under closed-book constraints remains a significant challenge. We describe our system for the 2026 AKBC Shared Task, where retrieval, external knowledge sources and fine-tuning are not permitted. We first evaluate 18 open-weight models on the training set to check how much the choice of the model matters for this task. The results differ significantly, from 0.186 to 0.499 macro F1, and the best model is not the largest one. Moreover, the models that score well on coding benchmarks tend to score worse on knowledge base construction. We then introduce **Pre-training Aware Directed Reasoning (PADR)**, a set of relation-specific reasoning pipelines optimized for the best evaluated model, Mistral Small 3.2 24B. PADR improves the macro F1 score on the training set from 0.499 to 0.614. On the shared task’s test set, PADR reaches 0.629 macro F1, more than double the official baseline’s 0.296. The code is available at <https://github.com/MikolajTym/akbc-padr>.

1 Introduction

The capabilities and usefulness of Large Language Models (LLMs) might be significantly extended by using techniques like Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). These techniques allow to retrieve relevant and reliable information, so the LLM can process it instead of relying on its parametric knowledge. However, the retrieval process requires a Knowledge Base (KB) that provides the reliable information, and maintaining such a KB requires significant effort.

Automated Knowledge Base Completion (AKBC) focuses on utilizing LLMs for this task (Chai et al., 2021), and the retrieval approach can be very useful here (Prabhong et al., 2024). However, if we want to create a KB that covers a

new topic, we cannot use retrieval, because the KB for this topic does not exist yet. Therefore, an interesting research area is the AKBC task under closed-book rules, i.e., no RAG, no web search tools, no knowledge graphs and no fine-tuning on data that covers the specific topic.

These rules are required by the 2026 AKBC Shared Task (Kalo et al., 2026), the fifth edition of the LM-KBC challenge (Kalo et al., 2025). The objective is to generate all object entities for a given subject entity and relation, i.e., to complete Wiki-data triples (Vrandečić and Krötzsch, 2014). The solution must therefore rely on the model’s parametric knowledge and on approaches that retrieve it properly. In the 2025 edition participants were forced to use the Qwen3-8B model (Yang et al., 2025). However, the 2026 edition allows any open-weight system within a 32B parameter budget, so the choice of the model becomes a part of the solution.

This work focuses on the following research questions:

1. Do open-weight models differ significantly in their capability to perform the closed-book AKBC task?
2. What explains these differences, the model size or the training set and behaviour that pre-training induces?
3. How much does relation-specific directed reasoning improve the best model, and for which relations?

We answer the first two questions by evaluating 18 models with a single closed-book configuration, and the third one by building a separate reasoning pipeline for each relation. Our submission reaches 0.629 macro F1 on the test set.

2 Related Work

Petroni et al. (2019) showed that pre-trained mod-

els answer factual queries without any fine-tuning. The LM-KBC challenges extend this idea to complete KB entries (Kalo et al., 2025). The 2025 edition already delivered effective methods for this task. These methods include relation-wise self-consistency, self-retrieval with divide-and-conquer decomposition, and an LLM that judges its own candidates (Albert-Roulhac and Zouaq, 2025; He and Razniewski, 2025; Clay et al., 2025). Some of these methods reuse techniques that are already well known, i.e., Chain-of-Thought (Wei et al., 2022) and self-consistency (Wang et al., 2023).

However, every participant of the 2025 edition worked with the same model, so these methods were developed for one fixed set of parametric knowledge. The 2026 edition changes this rule, and the model becomes a design decision instead of a constant. Therefore, we focus on this decision, as the model’s knowledge depends on its size and pre-training data (Roberts et al., 2020).

3 Task and Data

The input is a pair of a subject entity and a relation, and the output is the full set of object entities expressed as strings. A subject may have zero, one or many objects. For example, for the subject “France” and the relation `hasArea`, the model should generate “632702”, which is the area of France in square kilometres.

The dataset covers six relations and contains 477 training, 475 validation and 475 test subjects (Kalo et al., 2026). Systems are ranked by the F1 score macro-averaged over all rows in dataset. Predicted strings are normalized and matched against the gold labels and their aliases, and a numeric prediction is correct when it falls within a 5% relative tolerance.

4 Methodology

4.1 Initial Closed-Book Approach

The 2026 edition does not enforce a particular model, so we first verify how different models perform on the AKBC task. For each relation we implement the same two-step strategy. First, the LLM is queried with a direct, question-style prompt, e.g., “What is the capacity of {subject_entity}?”.

Simple prompts often cause the model to generate responses that cannot be parsed and evaluated, e.g., because of the model’s conversational style. Therefore, the second step uses structured output¹. This approach forces the model to return

a JSON object that conforms to a schema defining null values, relation cardinality and data types such as integers or floats. We refer to this configuration as the Initial Closed-Book approach.

4.2 Pre-training Aware Directed Reasoning

The next part of our research focuses on optimizing the pipeline for one selected model. We introduce **Pre-training Aware Directed Reasoning (PADR)**, a set of relation-specific prompting and extraction strategies. We call the strategy pre-training aware because it aligns the reasoning path and the prompt structure with the next-token prediction task used during causal pre-training (Radford et al., 2019). Four mechanisms are shared by the relation-specific pipelines.

Directed reasoning and structured extraction.

Every pipeline separates recalling a fact from formatting it. Directed reasoning means that we prescribe the reasoning path for each relation instead of letting the model decide how to think about the answer. The reasoning part therefore follows a fixed, relation-specific structure, e.g., resolving entity ambiguities, recalling the fact and justifying a decision, before a separate call turns that reasoning into the final answer. This order, reasoning before decision, matters for a model trained with next-token prediction, because each token is generated conditioned on the tokens before it, so a decision that follows its justification can build on that reasoning, while a decision written first cannot. Generating the context from parametric memory before answering resembles the self-retrieval strategy that was already applied to this task in 2025 (He and Razniewski, 2025). Each reasoning call is followed by its own extraction call, which reads that text and returns a JSON object under a grammar-constrained schema. The schema encodes the output contract, i.e., types, nullability, cardinality and closed vocabularies (e.g., a finite list of stock exchanges).

Targeting the pre-training format. Instead of asking a question, some prompts ask the model to recall the Wikipedia infobox fields for the subject. Wikipedia is extensively represented in pre-training corpora, and the infobox is a good source of facts. Therefore, reproducing that format may trigger the memorized fact more reliably than a conversational question.

¹<https://reference.langchain.com/python/>

`langchain-core/language_models/chat_models/BaseChatModel/with_structured_output`

Inverse verification. The reversal curse means that a model that learned “A is B” struggles to answer “B is A” (Lv et al., 2024). For `awardWonBy` the hard direction is the required one, because the query gives the award and asks for its winners. Therefore, we generate candidates in the hard direction and filter them in the easy direction, by asking whether a given candidate won the given award.

Agreement as a decision signal. For two relations we ask the same question again in forms that cannot copy the first answer, and we sample at a higher temperature. A similar approach, relation-wise self-consistency, applied to this task in the 2025 edition (Albert-Roulhac and Zouaq, 2025), also samples several answers for the same question. However, self-consistency returns the most frequent answer. We use the samples differently, because we measure the ratio of them that agree and treat that ratio as a confidence score. The ratio then decides which reasoning path should be followed.

4.2.1 Relation-Specific Pipelines

For `countryLandBordersCountry`, the reasoning prompt asks for the neighbours of the subject country and, for each of them, for the length and the physical nature of the border. Asking for a justification suppresses maritime neighbours, which the task definition excludes. The prompt also mentions overseas territories and enclaves, because the ground truth counts borders through integral territories.

For `hasArea`, the reasoning prompt follows three guided steps, i.e., identifying the geographic entity, separating it from entities with similar names, and recalling its official area. The prompt states that the value must be the surface of the entity itself and not of the surrounding waters.

For `companyTradesAtStockExchange`, the first stage generates a company profile that resolves the entity and states whether the company is public, private or a subsidiary. This stage also lists explicit rules against guessing a local exchange from the country of origin, because the model tends to answer with the exchange of the company’s home country. The second stage selects the listings, and the list of valid exchanges is enforced. Therefore, the model cannot emit an exchange outside the vocabulary.

For `personHasCityOfDeath`, the correct answer is an empty list when a person is still alive or

when the place of death is not recorded. Therefore, the decision is not only which city to answer, but whether to answer at all. We query each subject under three framings, i.e., the direct question, a request for the Wikipedia infobox fields, and a step-by-step disambiguation that first asks which person bearing the name is meant. Each framing is sampled three times at temperature 1.0, which gives nine independent answers. A city is predicted only when at least two framings name it and it carries at least a third of the nine runs, and otherwise the system returns an empty list. This approach mitigates hallucination, i.e., cases where the model describes different biographical profiles for the same person. In such cases, the empty prediction is correct if the person is still alive.

For `hasCapacity`, a first reasoning stage builds a venue profile that separates the target facility from training grounds and school fields with similar names. The second stage asks for the capacity. It prioritizes the value in the Wikipedia infobox over record attendance figures. It also forbids rounding. PADR treats this estimate as a hypothesis. It asks the same question again in eight forms, i.e., three choices from a list of options and five free answers sampled at temperature 0.9 without the profile. Venues whose estimate is confirmed by at least five of these answers are correct 55% of the time, against 23% overall. The number of confirmations then decides how many values to return. The metric divides precision by the length of the answer list. Therefore, k values containing the correct answer score $2/(k + 1)$ instead of zero. We return the estimate alone when at least five answers confirm it, and two or three common capacities for venues of that size otherwise. Even when the model cannot recall a venue, it usually places it in the right order of magnitude.

For `awardWonBy`, an initial stage resolves the award by identifying its grantor and its active lifespan. Next, parallel calls query five-year windows of that lifespan and explicitly decompose collective winners into their individual members. A similar temporal decomposition was already used for this relation in 2025 (He and Razniewski, 2025), though with fixed decades rather than windows tied to the award’s own lifespan. Temporal chunking is necessary because post-trained models tend to refuse a single query that requires thousands of output tokens for hundreds of recipients. Finally, an inverse verification step checks each candidate recipient separately. The verification schema places

the justification before the verdict. Therefore, the model states its reason first.

5 Experiments

5.1 Setup

The cross-model comparison is conducted with Ollama² as the model provider. Ollama serves models in the GGUF format with quantized weights, which affects their performance and behaviour (Dettmers and Zettlemoyer, 2023). The PADR experiments use Mistral Small 3.2 24B (Mistral AI Team, 2025) served by vLLM (Kwon et al., 2023) at full precision, with a fixed seed and batch-invariant kernels. The same model therefore occupies about 45 GB under vLLM and about 15 GB under Ollama, so the two configurations are not directly comparable.

When retrieving factual knowledge, it might seem logical to set the sampling temperature to 0. However, in our experiments a temperature of 0 caused the models to enter infinite reasoning loops and produce repetitive text. Therefore, for each model we use the temperature recommended on its Ollama page, which is why this value differs between the models. There are exceptions from this rule when testing model’s self-consistency for personHasCityOfDeath and hasCapacity. All experiments run on a NVIDIA DGX Spark workstation, which combines a GB10 Grace Blackwell chip with 128 GB of unified memory³. The full test run takes about 30 minutes on this GPU.

The prompts and the thresholds for hasCapacity and personHasCityOfDeath were selected on the validation set. Each relation contains at most 100 subjects, which is indeed too little data for the tuned parameters to generalize well. Therefore, the validation scores for these two relations should be read as optimistic. A system tuned directly on the test set would likely score higher there too. However, such a result would not show that the system generalizes, so we never use the test set for tuning.

5.2 Cross-Model Comparison

Our objective is to determine which model is best suited for the AKBC task. Therefore, we evaluate 18 models with the Initial Closed-Book approach on the challenge’s training set. Each organisation uses its own training data. We hypothesise that

²<https://ollama.com>

³<https://www.nvidia.com/en-us/products/workstations/dgx-spark/>

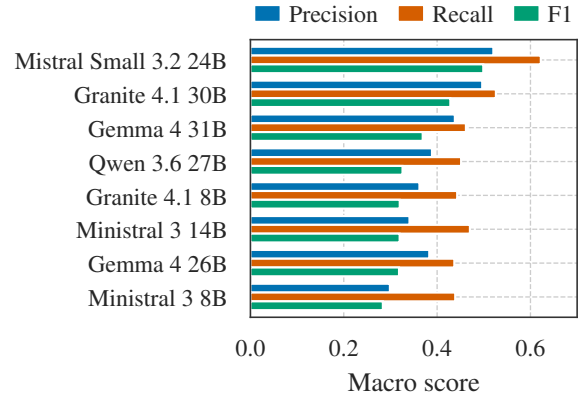


Figure 1: Initial Closed-Book approach macro precision, recall and F1 for the eight best models on the training set, sorted by macro F1.

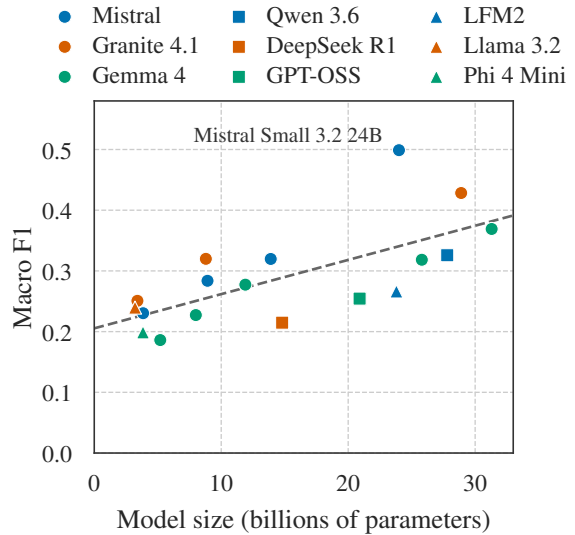


Figure 2: Model parameter count against the macro F1 score of the Initial Closed-Book approach on the training set.

some of these datasets cover the domain knowledge required by our target relations more heavily. It means that even a model which underperforms on general benchmarks may be beneficial here.

Figure 1 shows the eight best models. The macro F1 scores for all 18 models span from 0.186 for Gemma 4 E2B to 0.499 for Mistral Small 3.2 24B. It means that the choice of the model alone has crucial impact on performance.

Figure 2 relates the score to the parameter count. Model size generally correlates with the score. However, the largest model does not perform the best. Moreover, models from the Mistral and Granite (IBM Granite Team, 2026) families perform above the average trend for their respective sizes.

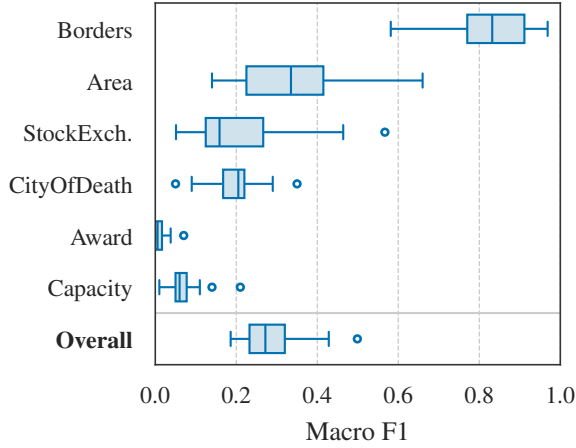


Figure 3: Distribution of macro F1 across the relations for all 18 models evaluated with the Initial Closed-Book approach.

Mistral Small 3.2 24B performs best on almost all relations, and it ranks second only on personHasCityOfDeath and awardWonBy, where Granite 4.1 30B is slightly better. Its success is caused by a high recall for all relations. We observe that reasoning models, such as Qwen 3.6 27B (Qwen Team, 2026) and Gemma 4 31B (Gemma Team, 2026), tend to hesitate when they generate the answers. These models often enter infinite reasoning loops where they repeatedly question and verify their own assertions. They also mention the need to call a web search tool. This tool is not available for the AKBC task. However, the models were trained to use such abilities. It means that not only the model’s knowledge but also its behaviour affects the performance.

But, why are these families better than others? When we inspect the same models on coding benchmarks, it turns out that Mistral and Granite are significantly worse. Terminal-Bench evaluates how a model solves programming tasks through a command line interface (Merrill et al., 2026). Gemma 4 31B achieves 43.4% there, while Mistral Small 3.2 achieves 5.6% and Granite 4.1 30B achieves 2.6%⁴. We hypothesise that models trained heavily on code must allocate a large part of their parametric memory to programming languages, API syntax and technical documentation. This pre-training improves their coding and logical reasoning abilities. However, the volume of syntax data occupies memory that would otherwise store general world facts.

The difficulty also differs strongly be-

tween the relations (Figure 3). The relation countryLandBordersCountry is relatively easy, with a median F1 of 0.832 across the 18 models. However, the same configuration fails almost completely on hasCapacity and awardWonBy, with median F1 scores of 0.060 and 0.006.

5.3 PADR Optimizations

We optimized the PADR approach utilizing the best model from our cross-model comparison, Mistral Small 3.2 24B.

Table 1 compares the Initial Closed-Book approach with PADR on the training set. PADR improves the overall macro F1 from 0.499 to 0.614, and the gain comes mostly from precision, which rises from 0.520 to 0.687.

The largest gain is on awardWonBy, from 0.038 to 0.521. During the initial evaluation the model can neither generate a long list of winners nor identify the exact award, and the temporal chunking with inverse verification solves both problems. The second largest gain is on personHasCityOfDeath, from 0.290 to 0.520, where precision rises from 0.310 to 0.770. This gain comes from the subjects on which the pipeline now declines to answer. On companyTradesAtStockExchange the profile stage resolves entity ambiguities and confirms the ownership status before the listings are queried, which raises the F1 score from 0.567 to 0.681.

These two configurations differ in the serving stack, because the initial evaluation used quantized GGUF weights and PADR uses full precision. This may affect the performance and model’s behaviour.

5.4 Validation and Test Results

On the validation set, PADR reaches 0.662 overall macro F1, while the official baseline reaches 0.313. Figure 4 presents our test results against the official baseline of the shared task. PADR reaches 0.629 overall macro F1, while the official baseline reaches 0.296.

The approach performance transfers well between the dataset splits. The only relation whose score moves substantially between splits is awardWonBy: 0.521 on training, 0.232 on validation and 0.355 on test. This relation contains only 10 subjects per split, so a few errors move the macro score by a large amount. The model also cannot properly disambiguate many of the awards, on validation and on test alike.

⁴<https://artificialanalysis.ai/evaluations/terminalbench-v2-1>

Relation	Initial Closed-Book			PADR		
	Precision	Recall	Macro F1	Precision	Recall	Macro F1
countryLandBordersCountry	0.989	0.956	0.969	0.980	0.988	0.981
personHasCityOfDeath	0.310	0.690	0.290	0.770	0.630	0.520
hasCapacity	0.210	0.210	0.210	0.192	0.290	0.218
awardWonBy	0.303	0.021	0.038	0.733	0.553	0.521
companyTradesAtStockExchange	0.610	0.765	0.567	0.787	0.792	0.681
hasArea	0.660	0.660	0.660	0.800	0.800	0.800
Overall	0.520	0.622	0.499	0.687	0.677	0.614

Table 1: Macro precision, recall and F1 of the Initial Closed-Book approach and of PADR for Mistral Small 3.2 24B, measured on the training set.

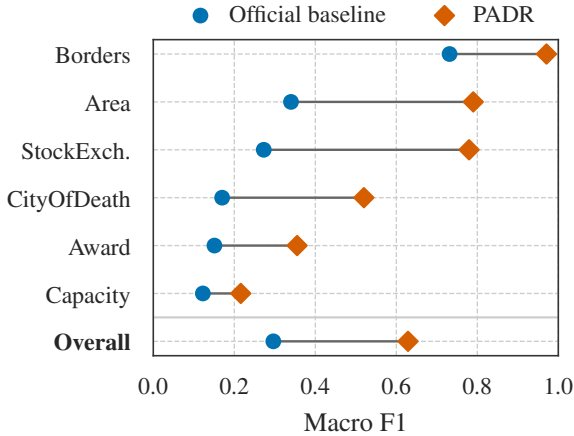


Figure 4: Macro F1 per relation on the test set, against the official Qwen3.5-9B baseline.

5.5 Error Analysis

We inspect the validation predictions with the official evaluation script and group the errors into four categories. The most frequent failure is a wrong value on a subject that the model does not know. It dominates `hasCapacity`, where the model produces a plausible number for almost every venue, and the agreement level can only decide how many values to return. The second category is entity confusion, which affects `awardWonBy` and `personHasCityOfDeath`. For one validation person the model recalls an athlete under one framing and a pianist under another, and it names six different cities across the nine runs. The third category is a hallucinated answer where the gold set is empty, and PADR reduces it the most, because our system returns an empty list for 117 of the 475 validation subjects. The last category is missed recall, and it is the price of the abstention rule, because on the validation set the recall of `personHasCityOfDeath` is 0.620 while its precision is 0.800.

6 Conclusions

RQ1. The models differ significantly, because the same closed-book configuration yields between 0.186 and 0.499 macro F1 depending only on the model.

RQ2. Size explains only a part of the difference. Mistral and Granite perform above the trend for their sizes, and both families score poorly on coding benchmarks. Moreover, models trained to verify their own answers hesitate, loop and ask for tools that are not available. Other prompts or a different way of using these models could raise their scores further. However, the Initial Closed-Book approach applies a fairly simple solution, so a model that already performs well here shows its own quality rather than a clever prompt.

RQ3. Relation-specific directed reasoning improves the best model from 0.499 to 0.614 macro F1 on the training set. However, the improvement is not uniform, because it is large for the relations that the direct question handles badly and absent for the relations that were already close to the ceiling.

We presented our system for the 2026 AKBC Shared Task. The evaluation of 18 models shows that the training data and the trained behaviour affect AKBC performance. It means that the parameter count alone does not indicate how well a model completes a knowledge base, so the model should be selected for this task and not only for its size. Most models are currently trained to use RAG or web search, but for the closed-book task the models that answer alone could be the best ones. PADR then improves the best model by directing its reasoning path separately for each relation, and our submission reaches 0.629 macro F1 on the test set.

Limitations

A key limitation of PADR is that applying it to a new relation requires manual human design. The directed reasoning paths must be tailored to the semantics of the relation and to the format in which the fact was likely memorized. Future research could develop an agent that constructs these paths automatically, for example by selecting inverse verification only for the relations where it applies. However, at present the design still requires human intervention, which limits the generalization of the approach.

The thresholds for decision algorithms were selected on the validation set, which has at most 100 subjects per relation. Their validation scores are therefore optimistic. Moreover, `awardWonBy` has only 10 subjects and `countryLandBordersCountry` has 68, so their macro scores are sensitive to individual errors.

Despite the fixed seed and batch-invariant kernels, two full runs of the pipeline do not always return the same prediction for a given subject, and this affects every relation to some degree. Most predictions repeat, or stay equally wrong, across runs, but there are some exceptions, especially for the relation `hasCapacity`.

Finally, the graded answer set for `hasCapacity` exploits the fact that the metric gives partial credit for a short list of candidates. This behaviour would not transfer to a knowledge base that accepts only a single value.

Declaration on Generative AI

The authors used LLMs for grammar and spelling checks. After using these tools, the authors reviewed and edited the content and take full responsibility for the publication’s content.

Acknowledgments

This publication is based upon work from COST Action CA23147 GOBLIN – Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, <https://www.cost.eu>).

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) co-located with the 24th International Semantic Web Conference (ISWC 2025), Nara, Japan, November 2, 2025, volume 4041 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Junyi Chai, Yujie He, Homa Hashemi, Bing Li, Daraksha Parveen, Ranganath Kondapally, and Wenjin Xu. 2021. [Automatic construction of enterprise knowledge base](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 11–19, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Clay, Ernesto Jiménez-Ruiz, and Pranava Madhyastha. 2025. [Judge, generator, executioner: Utilizing an LLM for KBC](#). 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) co-located with the 24th International Semantic Web Conference (ISWC 2025)*, Nara, Japan, November 2, 2025, volume 4041 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Tim Dettmers and Luke Zettlemoyer. 2023. [The case for 4-bit precision: k-bit inference scaling laws](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7750–7774. PMLR.

Gemma Team. 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) co-located with the 24th International Semantic Web Conference (ISWC 2025)*, Nara, Japan, November 2, 2025, volume 4041 of *CEUR Workshop Proceedings*. CEUR-WS.org.

IBM Granite Team. 2026. [IBM granite 4.1 language models](#). Hugging Face.

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 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) co-located with the 24th International Semantic Web Conference (ISWC 2025), Nara, Japan, November 2, 2025, volume 4041 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23*, page 611–626, New York, NY, USA. Association for Computing Machinery.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhang Chen, Ji-Rong Wen, and Rui Yan. 2024. [An analysis and mitigation of the reversal curse](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13603–13615, Miami, Florida, USA. Association for Computational Linguistics.
- Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenna Jitsev, Di Lu, Orfeas Menis Mastromichalakis, and 66 others. 2026. [Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces](#). *arXiv e-prints*, arXiv:2601.11868.
- Mistral AI Team. 2025. [Mistral small 3.2 24b instruct](#). Hugging Face.
- 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.
- Thin Prabhong, Natthawut Kertkeidkachorn, and Areerat Trongratsameethong. 2024. [KGC-RAG: knowledge graph construction from large language model using retrieval-augmented generation](#). In *Joint proceedings of the 2nd workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM 2024) and the 3rd challenge on Language Models for Knowledge Base Construction (LM-KBC 2024) co-located with the 23rd International Semantic Web Conference (ISWC 2024)*, Baltimore, USA, November 12, 2024, volume 3853 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Qwen Team. 2026. [Qwen3.6](#). GitHub repository.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wikidata: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. 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 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed 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. Curran Associates, Inc.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

A Detailed Validation and Test Results

Table 2 presents the complete validation results of PADR, including the macro precision and recall, the micro-averaged F1, the average number of predicted objects per subject and the number of empty predictions.

Table 3 presents the complete test results of PADR and of the official baseline, as scored by the organizers’ leaderboard.

B Cross-Model Evaluation

Figure 5 extends Figure 1 to all 18 evaluated models. Tables 4, 5 and 6 present the macro F1, precision and recall per relation for every model. All these runs use the training set and quantized GGUF weights.

Relation	P	R	F1	Micro	Empty
Borders	0.969	0.989	0.970	0.975	17
CityOfDeath	0.800	0.620	0.550	0.442	57
Capacity	0.287	0.495	0.345	0.310	0
Award	0.629	0.231	0.232	0.224	3
StockExch.	0.873	0.840	0.765	0.740	40
Area	0.810	0.810	0.810	0.810	0
Overall	0.733	0.725	0.662	0.393	117

Table 2: PADR on the validation set. **P**, **R** and **F1** are macro-averaged, **Micro** is the micro-averaged F1 and **Empty** is the number of empty predictions.

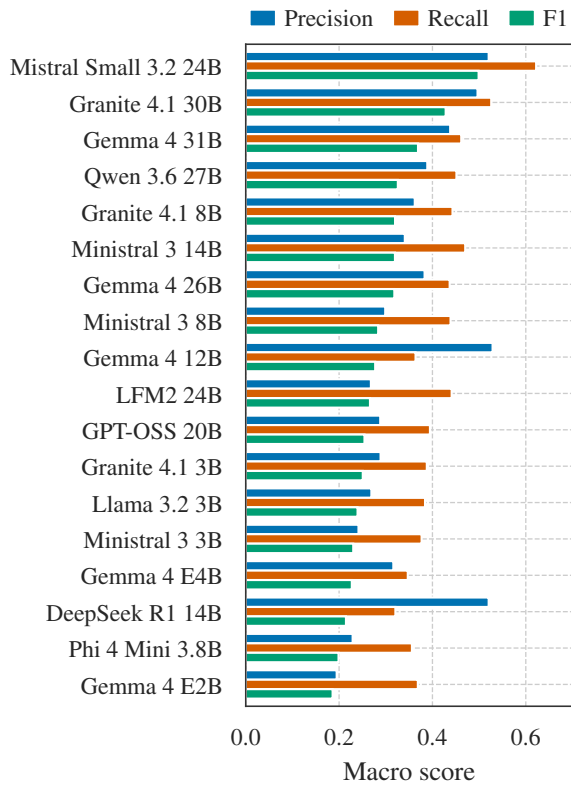


Figure 5: Initial Closed-Book approach macro precision, recall and F1 for all 18 evaluated models on the training set, sorted by macro F1.

Relation	Official Baseline			PADR		
	Precision	Recall	Macro F1	Precision	Recall	Macro F1
countryLandBordersCountry	0.742	0.889	0.732	0.976	0.985	0.971
personHasCityOfDeath	0.170	0.650	0.170	0.780	0.640	0.520
hasCapacity	0.122	0.122	0.122	0.191	0.286	0.216
awardWonBy	0.306	0.141	0.151	0.643	0.322	0.355
companyTradesAtStockExchange	0.292	0.707	0.273	0.854	0.829	0.780
hasArea	0.340	0.340	0.340	0.790	0.790	0.790
Overall	0.305	0.511	0.296	0.701	0.680	0.629
Zero-object cases	0.000	0.000	0.000	0.712	0.824	0.764

Table 3: Macro precision, recall and F1 of the official baseline and of PADR, measured on the test set by the organizers’ leaderboard.

Model	Overall	Borders	CityOfDeath	Capacity	Award	StockExch.	Area
Mistral Small 3.2 24B	0.499	0.969	0.290	0.210	0.038	0.567	0.660
Granite 4.1 30B	0.428	0.943	0.350	0.110	0.070	0.405	0.540
Gemma 4 31B	0.369	0.961	0.220	0.080	0.025	0.464	0.350
Qwen 3.6 27B	0.326	0.916	0.220	0.070	0.009	0.170	0.480
Granite 4.1 8B	0.320	0.889	0.210	0.040	0.018	0.278	0.400
Ministral 3 14B	0.320	0.844	0.200	0.140	0.007	0.059	0.560
Gemma 4 26B	0.318	0.926	0.130	0.100	0.034	0.285	0.380
Ministral 3 8B	0.284	0.851	0.200	0.060	0.000	0.103	0.420
Gemma 4 12B	0.277	0.898	0.220	0.030	0.003	0.231	0.240
LFM2 24B	0.266	0.793	0.190	0.060	0.012	0.147	0.340
GPT-OSS 20B	0.254	0.715	0.200	0.070	0.000	0.145	0.320
Granite 4.1 3B	0.251	0.820	0.160	0.050	0.002	0.217	0.220
Llama 3.2 3B	0.240	0.766	0.090	0.070	0.000	0.139	0.330
Ministral 3 3B	0.230	0.819	0.220	0.050	0.000	0.060	0.220
Gemma 4 E4B	0.227	0.782	0.260	0.040	0.004	0.120	0.140
DeepSeek R1 14B	0.215	0.581	0.230	0.010	0.007	0.194	0.200
Phi 4 Mini 3.8B	0.199	0.637	0.130	0.050	0.000	0.051	0.290
Gemma 4 E2B	0.186	0.740	0.050	0.050	0.000	0.142	0.150

Table 4: Macro F1 per relation for all models evaluated with the Initial Closed-Book approach on the training set.

Model	Overall	Borders	CityOfDeath	Capacity	Award	StockExch.	Area
Mistral Small 3.2 24B	0.520	0.989	0.310	0.210	0.303	0.610	0.660
Granite 4.1 30B	0.497	0.987	0.550	0.110	0.383	0.469	0.540
Gemma 4 31B	0.438	0.980	0.280	0.140	0.325	0.620	0.360
Qwen 3.6 27B	0.389	0.947	0.240	0.210	0.275	0.263	0.480
Granite 4.1 8B	0.362	0.925	0.220	0.040	0.408	0.407	0.400
Ministral 3 14B	0.341	0.912	0.210	0.140	0.411	0.063	0.560
Gemma 4 26B	0.383	0.960	0.160	0.190	0.343	0.390	0.410
Ministral 3 8B	0.299	0.938	0.210	0.060	0.000	0.107	0.420
Gemma 4 12B	0.529	0.974	0.530	0.490	0.500	0.420	0.380
LFM2 24B	0.268	0.789	0.190	0.060	0.280	0.133	0.340
GPT-OSS 20B	0.288	0.750	0.210	0.070	0.400	0.232	0.320
Granite 4.1 3B	0.289	0.952	0.180	0.050	0.250	0.265	0.220
Llama 3.2 3B	0.269	0.870	0.160	0.070	0.000	0.141	0.330
Ministral 3 3B	0.242	0.869	0.230	0.050	0.000	0.070	0.220
Gemma 4 E4B	0.316	0.898	0.610	0.040	0.050	0.112	0.140
DeepSeek R1 14B	0.520	0.949	0.500	0.370	0.600	0.327	0.590
Phi 4 Mini 3.8B	0.229	0.762	0.170	0.050	0.000	0.072	0.290
Gemma 4 E2B	0.195	0.720	0.110	0.050	0.000	0.137	0.150

Table 5: Macro precision per relation for all models evaluated with the Initial Closed-Book approach on the training set.

Model	Overall	Borders	CityOfDeath	Capacity	Award	StockExch.	Area
Mistral Small 3.2 24B	0.622	0.956	0.690	0.210	0.021	0.765	0.660
Granite 4.1 30B	0.526	0.925	0.600	0.110	0.040	0.635	0.540
Gemma 4 31B	0.462	0.964	0.570	0.080	0.014	0.555	0.350
Qwen 3.6 27B	0.451	0.922	0.570	0.070	0.005	0.415	0.480
Granite 4.1 8B	0.443	0.869	0.610	0.040	0.009	0.480	0.400
Ministral 3 14B	0.470	0.832	0.570	0.140	0.005	0.415	0.560
Gemma 4 26B	0.437	0.931	0.510	0.100	0.020	0.468	0.380
Ministral 3 8B	0.439	0.853	0.590	0.060	0.000	0.452	0.420
Gemma 4 12B	0.364	0.881	0.440	0.030	0.001	0.435	0.240
LFM2 24B	0.441	0.928	0.580	0.060	0.007	0.502	0.340
GPT-OSS 20B	0.395	0.703	0.580	0.070	0.000	0.442	0.320
Granite 4.1 3B	0.388	0.789	0.580	0.050	0.001	0.472	0.220
Llama 3.2 3B	0.384	0.742	0.450	0.070	0.000	0.485	0.330
Ministral 3 3B	0.377	0.832	0.610	0.050	0.000	0.360	0.220
Gemma 4 E4B	0.347	0.780	0.450	0.040	0.002	0.503	0.140
DeepSeek R1 14B	0.321	0.605	0.470	0.010	0.004	0.443	0.200
Phi 4 Mini 3.8B	0.356	0.624	0.530	0.050	0.000	0.410	0.290
Gemma 4 E2B	0.369	0.919	0.430	0.050	0.000	0.513	0.150

Table 6: Macro recall per relation for all models evaluated with the Initial Closed-Book approach on the training set.