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

Mikołaj Tym

Poznań University of Economics and Business

Closed-Book Knowledge Base Construction

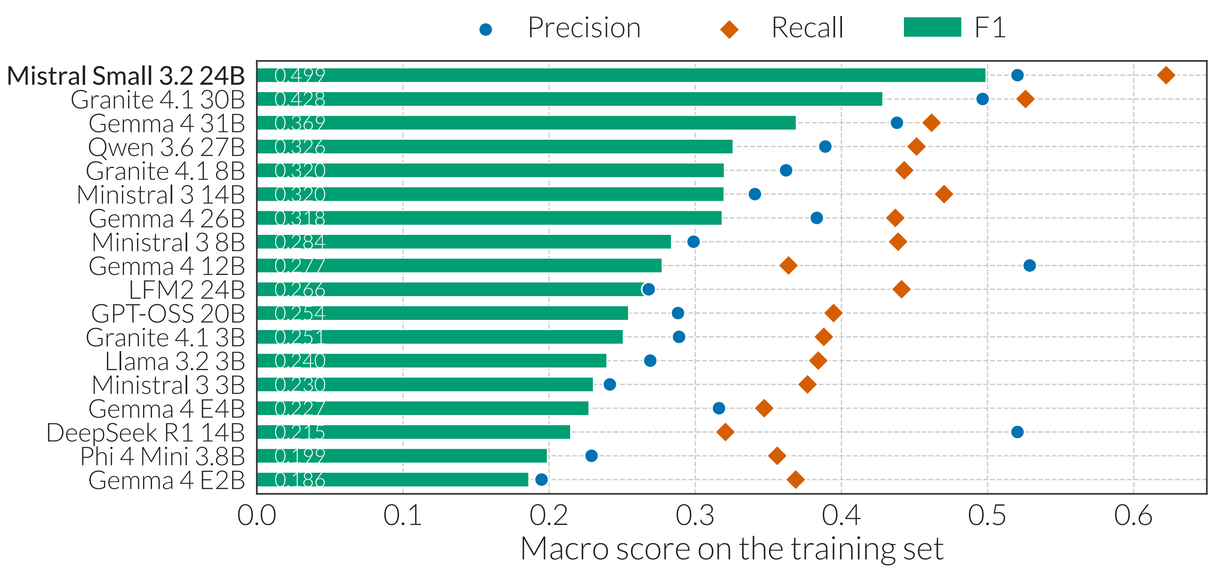
The capabilities of Large Language Models (LLMs) might be significantly extended by Retrieval-Augmented Generation (RAG) [1]. However, RAG requires a Knowledge Base (KB), and maintaining such a KB requires significant effort. If we want to create a KB that covers a new topic, we cannot use retrieval, because the KB for this topic does not exist yet. Therefore, the 2026 AKBC Shared Task [2] proposes **closed-book rules** for this task, i.e., no RAG, no web search tools, no knowledge graphs and no fine-tuning.

France `hasArea` → 632702

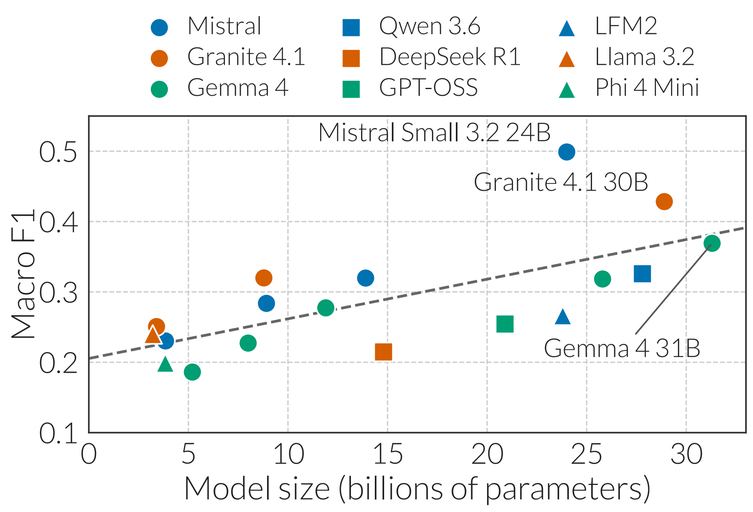
The task asks for all objects of a given subject and relation. Unlike in 2025 [3], any open-weight model up to 32B parameters is allowed, so **the choice of the model becomes a part of the solution**.

RQ1: Do Open-Weight Models Differ?

Therefore, we put significant effort into evaluating 18 models served by Ollama with the same **Initial Closed-Book approach**, i.e., a direct question and structured output. The macro F1 spans from **0.186 to 0.499**, depending only on the model.



RQ2: Model Size or Pre-training?



Terminal-Bench (coding) [4]

43.4% Gemma 4 31B

5.6% Mistral Small 3.2

2.6% Granite 4.1 30B

The best model is not the largest one, it is not a reasoning model, and it scores poorly on coding benchmarks. We hypothesise that programming syntax occupies a large part of the parametric memory, which would otherwise store world facts.

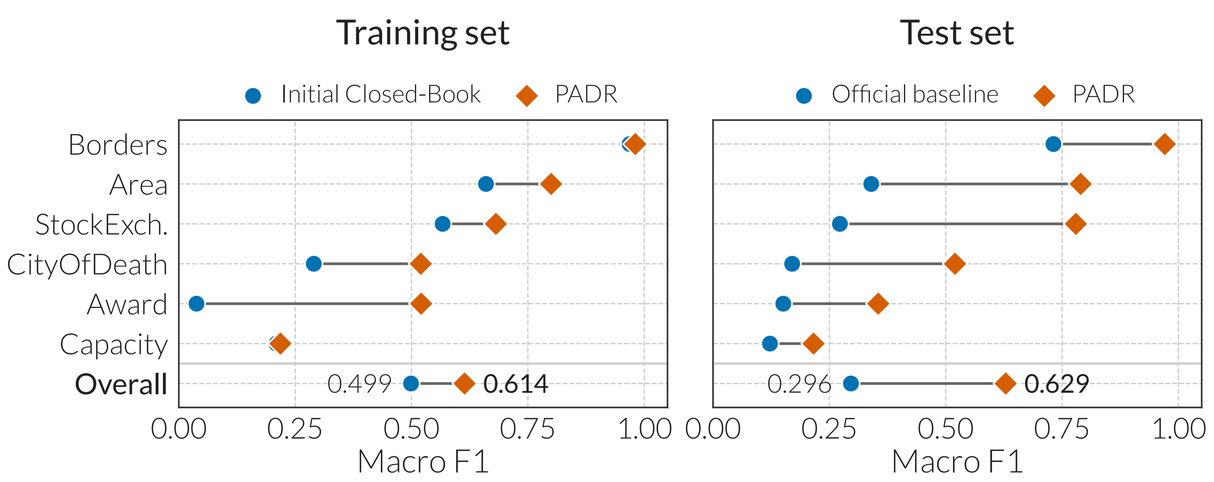
RQ3: Pre-training Aware Directed Reasoning

PADR is a set of relation-specific agent workflows for the best model, Mistral Small 3.2 24B [5]. We design each workflow with the fact in mind that the model was trained with causal language modeling, i.e., next-token prediction.

0.614 training set (from 0.499)

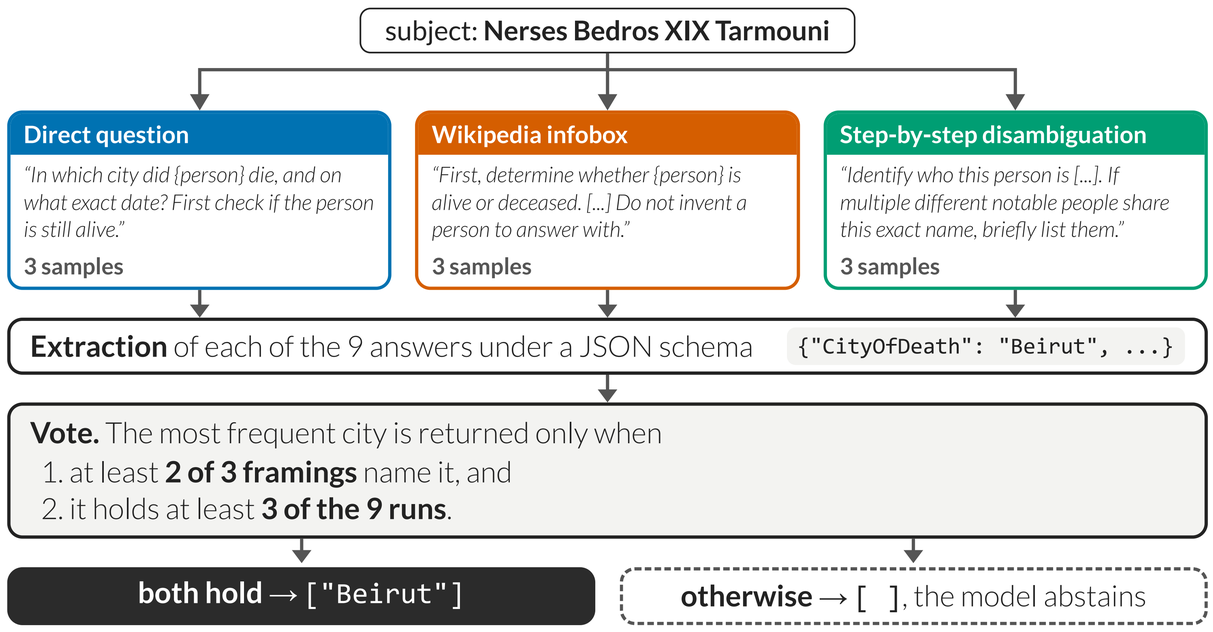
0.662 validation set (official baseline 0.313)

0.629 test set (official baseline 0.296)



Example: `personHasCityOfDeath`

The shared task covers six relations, and each of them has its own PADR workflow. Here we present the workflow for `personHasCityOfDeath`. The correct answer is an empty list when a person is still alive or the place of death is not recorded. Therefore, the decision is not only which city to answer, but whether to answer at all.



Real answers of the model on the validation set

Nerses Bedros XIX Tarmouni — *known fact*

Direct	no city	Beirut	Beirut
Infobox	Beirut	Beirut	Beirut
Step-by-step	Beirut	Beirut	Beirut

8 of 9 runs name Beirut

PADR `["Beirut"]` gold `Beirut` ✓ **correct**

Thomas Köhler (luger) — *hallucination is filtered*

Direct	no city	Berlin	Berlin
Infobox	Annaberg-Buchholz	Sonneberg	Aue-Bad Schlema
Step-by-step	Innsbruck	Oberwiesenthal	Zella-Mehlis

7 different cities, no city holds 3 runs

PADR `[ ]` gold `[ ]` ✓ **correct**

Edward Redliński — *agreement on a wrong answer*

Direct	no city	no city	no city
Infobox	Warsaw	Warsaw	Warsaw
Step-by-step	Warsaw	Warsaw	Warsaw

6 of 9 runs name Warsaw

PADR `["Warsaw"]` gold `[ ]` ✗ **wrong**

Peter Gregg (racing driver) — *unknown fact is lost*

Direct	Plymouth	Oxford	Palm Beach
Infobox	Middletown	Bristol	Burlington
Step-by-step	Windham	Boca Raton	Cerro Pelon

9 different cities, no city holds 3 runs

PADR `[ ]` gold `Jacksonville` ✗ **wrong**

Entity confusion: who is "Renate Hellwig"?

Wikipedia infobox	*"...a German swimmer, passed away."*
Direct question	*"...a German athlete who specialized in the hurdles"*
Direct question	*"...a German athlete who specialized in the high jump and the pentathlon"*
Step-by-step disambiguation	*"...notable for her involvement in the early Soviet espionage network in the United States during World War II."*

7 cities: Berlin · Düsseldorf · Göttingen · Hamburg · Königs Wusterhausen · Munich · Wuppertal

PADR `[ ]` · gold `[ ]` ✓ **correct abstention**

The model does not recognise this person, so it describes a different person in each run. PADR raises the precision of this relation from 0.310 to 0.770 on the training set.

Conclusions

- The same closed-book approach yields between 0.186 and 0.499 macro F1, depending only on the model.
- The largest model does not necessarily perform the best on the AKBC task.
- PADR improves the best model from 0.499 to 0.614 macro F1 on the training set.

References

[1] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474.

[2] Jan-Christoph Kalo et al. "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, 2026.

[3] Jan-Christoph Kalo et al. "LM-KBC 2025: 4th Challenge on Knowledge Base Construction from Pre-trained Language Models". In: *Joint Proceedings of KBC-LM and LM-KBC 2025, co-located with ISWC 2025*. Vol. 4041. CEUR Workshop Proceedings. 2025.

[4] Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, et al. "Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces". In: *arXiv preprint arXiv:2601.11868* (2026).

[5] Mistral AI Team. *Mistral Small 3.2 24B Instruct*. Hugging Face. 2025.