

Summary of the AKBC 2026 Shared Task: Language Models for Knowledge Base Construction (LM-KBC, 5th Edition)

Jan-Christoph Kalo
University of Amsterdam
j.c.kalo@uva.nl

Simon Razniewski
ScaDS.AI & TU Dresden
simon.razniewski@tu-dresden.de

Bohui Zhang
King's College London
bohui.zhang@kcl.ac.uk

Tuan-Phong Nguyen
VNU University of Engineering and Technology
tuanphong@vnu.edu.vn

Abstract

Pretrained language models (LMs) have significantly advanced a variety of knowledge engineering tasks and have shown promise as sources of factual knowledge. While prior work has studied factual knowledge through probing and prompting, reliably constructing complete knowledge bases from the parametric knowledge of LMs remains challenging. The AKBC 2026 Shared Task invited participants to construct knowledge bases directly from LMs, given specific subjects and relations. Unlike existing probing benchmarks, the challenge imposed no simplifying assumptions on relation cardinality, allowing a subject entity to be associated with zero, one, or multiple objects, including long answer sets. To ensure accessibility, the challenge featured a single track using open-weight LMs with a total neural parameter budget of up to 32B. Nineteen submissions were received, of which twelve were ranked, exploring a variety of approaches including relation-specific prompting, evidence aggregation, and model selection.

1 Introduction

Large language models (LLMs) such as Qwen (Qwen Team, 2025), Gemma (Gemma Team, 2025, 2026), and chatGPT (OpenAI, 2022) have achieved remarkable performance across a wide range of NLP and knowledge-intensive tasks. Beyond their general language capabilities, LLMs encode substantial factual knowledge in their parameters, motivating growing interest in using them not only to answer factual queries but also to produce structured knowledge (Pan et al., 2023). This is promising because existing knowledge bases (KBs), such as Wikidata (Vrandečić and Krötzsch, 2014) and ConceptNet (Liu and Singh,



Figure 1: Overview of the AKBC 2026 shared task.

2004), remain inherently incomplete (Razniewski et al., 2024). Constructing KBs, however, requires more than answering isolated factual questions: many relations are optional (e.g., place of death or parent organization), while others may have multiple or even large numbers of correct objects (e.g., shares border or award recipients). Materializing such knowledge into explicit KBs also facilitates LLM analysis (Razniewski et al., 2026), and consistent, inspectable, and trustworthy downstream use (Zhang et al., 2025a,b).

Traditional approaches to knowledge base construction (KBC) have leveraged unstructured text (Nakashole et al., 2011; Dong et al., 2014),

crowdsourcing (Acosta et al., 2013; Simperl and Luczak-Rösch, 2014), and semi-structured resources (Bollacker et al., 2008; Auer et al., 2007; Suchanek et al., 2007; Hoffart et al., 2013). More recently, LLMs have been explored for knowledge extraction, completion, and construction (Pan et al., 2023), including large-scale materialization of parametric knowledge (Hu et al., 2025). These developments demonstrate the feasibility of LLM-based KBC, while reliable and complete extraction of structured knowledge from model parameters remains challenging.

The LAMA benchmark (Petroni et al., 2019) established a widely used setting for probing factual knowledge through subject–relation queries, with subsequent work examining the reliability and sensitivity of such probing (Alivanistos et al., 2022; Zhang et al., 2023; Zhao et al., 2024). However, factual probing and KB construction impose different requirements. KBC requires predicting complete object sets, including cases where no valid object exists or where many objects must be returned. Handling unknown relation cardinality, incomplete recall, and erroneous predictions therefore remains central to practical LLM-based KBC.

The AKBC Shared Task aims to study these challenges directly. Continuing previous editions (Singhania et al., 2022, 2023; Kalo et al., 2024, 2025), the fifth edition focuses on constructing KBs from the parametric knowledge of open-weight LLMs under a total neural parameter budget of 32B, without external retrieval or model fine-tuning. Given a subject–relation pair, participating systems must predict the complete set of corresponding objects without assuming a fixed relation cardinality, allowing the correct answer to contain zero, one, or many objects.

2 Task Description

In the AKBC 2026 Shared Task, knowledge base construction is defined as follows: given a subject entity s and a relation r , the goal is to predict all correct object entities $\{o_1, o_2, \dots, o_k\}$ using the parametric knowledge of language models. The number of correct objects is not given and may be zero, one, or many. For example, given (*Greece*, *countryLandBordersCountry*), a system should return [Albania, North Macedonia, Bulgaria, Turkey], in any order. For numeric relations, such as (*Wembley Stadium in London*, *hasCapacity*), the output is a numeric value represented as a string,

e.g., ["90000"]. Predictions are evaluated directly as strings.

Input and Output

Input: a subject label and one of the six target relations, e.g., (*Greece*, *countryLandBordersCountry*).

Output: a finite set of object strings, possibly empty, e.g., [Albania, North Macedonia, Bulgaria, Turkey] or [].

Participating systems must be self-contained, using open-weight LMs within a total neural parameter budget of 32B, without model fine-tuning, external knowledge sources, web search, or retrieval-augmented generation. We released question-style and fill-in-the-blank prompt templates with a Qwen3.5-9B baseline (Table 2).

Predictions are evaluated at the subject–relation level using precision, recall, and F1. For string-valued relations, predictions and gold aliases are normalized before matching, including case folding, removal of diacritics and apostrophe-like characters, and punctuation normalization. Predicted strings are then matched to gold entities and their aliases using maximum bipartite matching, such that each prediction and each gold entity can be matched at most once. For the numeric relations *hasArea* and *hasCapacity*, a prediction is considered correct if it lies within a 5% relative tolerance of the gold value. The official ranking metric is macro-F1, computed as the average of the per-instance F1 scores over all subject–relation pairs.

3 Dataset

The dataset was constructed from Wikidata and manually refined to improve the quality of the ground truth. The 2026 edition retains the six relation types used in the previous edition, covering complementary challenges in closed-book KBC: *countryLandBordersCountry*, *companyTradesAtStockExchange*, and *personHasCityOfDeath* allow empty answer sets; *awardWonBy* may contain large numbers of objects; and *hasArea* and *hasCapacity* require numeric predictions. Table 1 summarizes the dataset statistics.

For each relation, up to 100 subject entities are provided in the training, validation, and hidden test sets. The ground truth was manually reviewed and updated to reflect the state of the world as of July 1, 2026. This included corrections to current stock-

Relation	Train	Validation	Test	Special features
awardWonBy	10	10	10	Many objects per subject
companyTradesAtStockExchange	100	100	100	Null values possible
countryLandBordersCountry	67	68	67	Null values possible
hasArea	100	100	100	Numeric (unit: km ²)
hasCapacity	100	97	98	Numeric
personHasCityOfDeath	100	100	100	Null values possible

Table 1: Dataset statistics. The numbers indicate the count of subject–relation pairs.

Relation	Macro-P	Macro-R	Macro-F1	Avg. # Preds.	# Empty Preds.
awardWonBy	0.247	0.078	0.101	24.000	0
companyTradesAtStockExchange	0.369	0.725	0.354	1.170	0
countryLandBordersCountry	0.697	0.911	0.665	2.706	0
hasArea	0.290	0.290	0.290	1.000	0
hasCapacity	0.180	0.180	0.180	1.000	0
personHasCityOfDeath	0.210	0.600	0.210	1.000	0
All Relations	0.324	0.507	0.313	1.759	0

Table 2: Validation performance of the Qwen3.5-9B baseline. Macro precision, recall, and F1 are reported together with the average number of predicted objects and the number of empty predictions.

exchange listings, recent deaths, award recipients, land borders, and numerical values. Ambiguous subject labels were also refined so that each subject–relation pair identifies the intended entity.

The 2026 dataset and evaluation code are publicly available on GitHub¹. Participating systems submitted their predictions through Codabench², where the final ranking was computed on the hidden test set.

4 Results

The twelve ranked submissions, summarized in Table 3, explored a broad range of inference-time strategies, while converging on the use of relatively large open-weight backbones and strongly relation-specific inference.

Candidate generation. Several of the highest-performing systems emphasized repeated elicitation and aggregation. Silchenko (2026) uses a single 31.3B Gemma base model (Gemma Team, 2026) in completion mode, with relation-specific prompting, repeated sampling, vote-based aggregation for entity-valued relations, and tolerance-aware aggregation for numerical values. A particularly effective intervention was a recitation format that first elicits contextual facts about the subject before committing to the target value. MAVE (Kunze and Both, 2026) similarly treats different prompt formulations as separate voting units, so repeat-

edly sampling one formulation cannot dominate the evidence supplied by another.

Selection and verification. A second group makes the separation between candidate generation and verification or selection particularly explicit. Sun et al. (2026) use the 29.6B MuseGlimmer (Meta Superintelligence Labs, 2026) backbone to construct candidate pools from multiple prompting strategies, followed by large-scale one-token Yes/No verification and relation-specific decision rules. COVER-KBC (Vo, 2026) develops a more general evidence-based architecture in which relation-specific inference traces keep provenance and support information instead of reducing everything immediately to a simple vote. MuseList (Zimmermann, 2026) likewise separates candidate generation from selection, generating different structured representations, such as lists, tables, and venue records, before applying relation-specific selection rules, including token-level confidence for numerical values and checks for proposed reciprocal borders.

Access format and model choice. Several submissions place model choice and model access format at the center of the method. Khalil and Mourad (Khalil and Mourad, 2026) select Gemma-3-27B-PT (Gemma Team, 2025) and report a substantial gap between the pretrained checkpoint used through raw completion and its instruction-tuned counterpart. Sasaki (2026) reaches a related conclusion with Gemma-3-27B-PT: controlled anal-

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

²<https://www.codabench.org/competitions/16267/>

Rank	Submission	Model(s)	Average
1	Silchenko (2026)	Gemma-4-31B	0.7060
2	Kunze and Both (2026)	Mistral-Small-3.2-24B-Instruct-2506	0.6908
3	Khalil and Mourad (2026)	Gemma-3-27B-PT	0.6836
4	Zimmermann (2026)	Muse-Glimmer-30B	0.6700
5	Sasaki (2026)	Gemma-3-27B-PT	0.6670
6	Ungureanu et al. (2026)	Muse-Glimmer-30B	0.6651
7	Sun et al. (2026)	Muse-Glimmer-30B	0.6553
8	Tym (2026)	Mistral-Small-3.2-24B-Instruct-2506	0.6289
9	Bui et al. (2026)	Mistral-Small-24B-Instruct-2501	0.6179
10	Vo (2026)	Mistral-Small-3.2-24B-Instruct-2506	0.5878
11	Patcharatrakul and Vateekul (2026)	Qwen3.6-27B, Gemma-3-4B-IT	0.5699
12	Habibillah and Tjjaranata (2026)	Qwen3.5-27B	0.5408
Late	Lazzaroni et al. (2026)	Gemma-4-31B	0.7121[†]

Table 3: Final ranking on the test set. [†]The late submission by Lazzaroni et al. (2026) was not ranked officially; its value is the live leaderboard F1 score and is included because the system description contributed useful methodological ideas.

yses suggest that backbone and access format can outweigh many prompt-level interventions. PADR (Tym, 2026) benchmarks 18 open-weight models, selects Mistral-Small-3.2-24B (Mistral AI, 2025), and constructs relation-specific reasoning procedures that separate recall from structured extraction.

Relation-specific routing. REAP (Bui et al., 2026) separates free-form factual elicitation from output serialization: relation-specific structured reasoning, decomposition, and explicit empty-set decisions are performed first, after which deterministic parsing is used where possible and model-based extraction only as a fallback. CoRe (Habibillah and Tjjaranata, 2026) routes relations between chain-of-thought reasoning and direct recall, finding that neither is optimal across all relations. DataMind (Patcharatrakul and Vateekul, 2026) combines multiple prompt variants and samples with relation-specific aggregation, while Ungureanu et al. (2026) ensemble four prompting paradigms and treat awardWonBy separately through iterative expansion and pruning because majority voting substantially reduces recall.

Late but informative submission. The late, unranked submission by Lazzaroni et al. (2026) is also methodologically informative. It scored 0.7121 F1 on the live leaderboard with a Gemma-4 base model, which would make it the strongest observed score, but it was submitted after the ranking deadline. Beyond the number, it emphasizes split discipline, fixed demonstration banks, and relation-specific pseudoval diagnostics, explicitly distinguishing train-based threshold choice, official-

validation reporting, and auxiliary checks against overfitting small validation slices.

Cross-system patterns. Across systems, a common pattern is the separation between candidate generation and candidate selection (Zimmermann, 2026; Vo, 2026). Candidate pools are expanded through repeated sampling (Silchenko, 2026; Khalil and Mourad, 2026), heterogeneous prompt formulations (Kunze and Both, 2026; Patcharatrakul and Vateekul, 2026; Ungureanu et al., 2026), structured representations (Zimmermann, 2026), or task decomposition (Bui et al., 2026). The resulting candidates are then filtered or combined using recurrence thresholds (Khalil and Mourad, 2026), agreement across prompts (Kunze and Both, 2026; Patcharatrakul and Vateekul, 2026), explicit verification (Sun et al., 2026; Tym, 2026), token-level confidence (Zimmermann, 2026), or relation-specific decision procedures (Vo, 2026). The submissions further suggest that repeated querying is most useful when the queries provide complementary views of the fact.

Model choice was a prominent factor across the submissions. The systems discussed here span Gemma, Mistral, Muse-Glimmer, Qwen, and one mixed Mistral/Muse setup. DataMind additionally uses a 4B Gemma model as an auxiliary component. Figure 2 shows that parameter count alone does not explain performance: the strongest systems span approximately 24B–31B parameters, while systems of similar size show substantial performance differences. Figure 3 shows another source of variation: systems with similar F1 can occupy different precision–recall operating points. Beyond model choice, abstention and precision con-

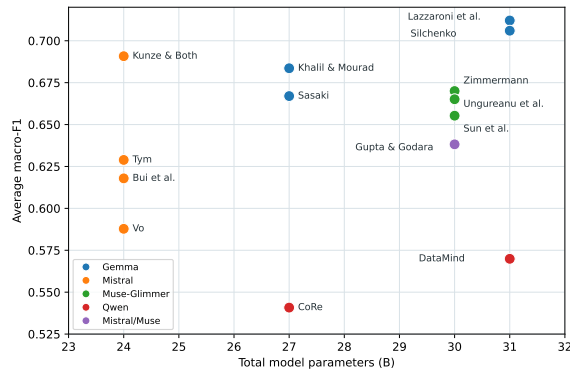


Figure 2: Average macro-F1 versus total model parameter count for the submitted systems discussed here and the late live-leaderboard result by Lazzaroni et al. (2026). Colors denote model families and use the same palette as Figure 3; mixed-family systems are shown as such. The Patcharatrakul and Vateekul (2026) system is colored by its Qwen backbone.

trol are important for relations permitting empty answers, while high-cardinality relations such as awardWonBy often require recall-oriented decomposition before filtering. Numerical relations, especially hasCapacity, remain difficult across systems.

5 Conclusion

The AKBC 2026 Shared Task confirms that closed-book KBC from language models is no longer only a question of whether models contain relevant facts. The strongest systems used large open-weight models, but their gains came from relation-aware prompting, candidate-pool expansion, careful selection, abstention control, and verification. The task also exposes where parametric recall remains brittle: long answer sets, empty-answer detection, and numeric values with plausible but wrong alternatives. Future editions should continue to evaluate complete object sets, preserve closed-book constraints, and encourage analyses that distinguish raw candidate coverage from recoverable knowledge.

References

- Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, and Jens Lehmann. 2013. [Crowdsourcing linked data quality assessment](#). In *ISWC*.
- Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. [Prompting](#)

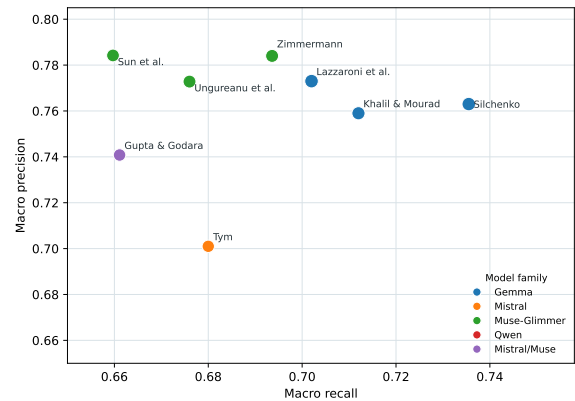


Figure 3: Macro precision versus macro recall for submissions that report an aggregate precision-recall operating point. Only points with precision and recall at or above 0.65 are shown.

as probing: Using language models for knowledge base construction. In *LM-KBC*.

- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. [DBpedia: A nucleus for a web of open data](#). In *ISWC*.

- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. [Freebase: A collaboratively created graph database for structuring human knowledge](#). In *SIGMOD*.

- Thanh-Dan Bui, Thanh-Trung Do, and Tuan-Phong Nguyen. 2026. [REAP: Relation-aware elicitation and parsing for closed-book knowledge base construction from LLMs](#). In *AKBC*.

- Xin Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. [Knowledge vault: A web-scale approach to probabilistic knowledge fusion](#). In *KDD*.

- Gemma Team. 2025. [Gemma 3 technical report](#).

- Gemma Team. 2026. [Gemma 4 technical report](#).

- Fatih Naufal Habibillah and Filbert Aurelian Tjjaranata. 2026. [CoRe: Construct-versus-recall routing for knowledge base construction under a parameter budget](#). In *AKBC*.

- Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. [YAGO2: A spatially and temporally enhanced knowledge base from wikipedia](#). *Artificial Intelligence*.

- Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. 2025. [Enabling LLM knowledge analysis via extensive materialization](#). In *ACL*.

- Jan-Christoph Kalo, Tuan-Phong Nguyen, Simon Razniewski, and Bohui Zhang. 2024. [Preface: LM-KBC challenge 2024](#). In *LM-KBC*.

- 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 *LM-KBC*.
- Christophe Khalil and Aya Mourad. 2026. The instruction tax on factual recall: Base models can beat their instruction-tuned counterparts. In *AKBC*.
- Jonas Kunze and Andreas Both. 2026. MAVE: Multi-angle voting for emission in closed-book knowledge base construction. In *AKBC*.
- Ruggero Marino Lazzaroni, Tim Booker, Pau Soler Valadés, Mathias Angermaier, and Kirill Solovov. 2026. Relation-specific elicitation of a base model under a 32b closed-book budget for LM-KBC 2026. In *AKBC*.
- Hugo Liu and Push Singh. 2004. [Conceptnet: A practical commonsense reasoning tool-kit](#). *BT Technology Journal*.
- Meta Superintelligence Labs. 2026. [Introducing Muse Glimmer: An open agentic model that runs on your device](#).
- Mistral AI. 2025. [Mistral-Small-3.2-24B-Instruct-2506](#).
- Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. 2011. [Scalable knowledge harvesting with high precision and high recall](#). In *WSDM*.
- OpenAI. 2022. [Introducing ChatGPT](#).
- Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. [Large language models and knowledge graphs: Opportunities and challenges](#). *TGDK*.
- Tanaree Patcharatrakul and Peerapon Vateekul. 2026. DataMind at AKBC shared task 2026: Relation-specific multiple prompting for closed-book knowledge base completion. In *AKBC*.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *EMNLP-IJCNLP*.
- Qwen Team. 2025. [Qwen3 technical report](#).
- Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, and Fabian Suchanek. 2024. [Completeness, recall, and negation in open-world knowledge bases: A survey](#). *ACM CSUR*.
- Simon Razniewski, Shrestha Ghosh, Luca Giordano, Yujia Hu, Josua Kowalzik, and Tuan-Phong Nguyen. 2026. [Epistemic twins: Enabling a symbolic science of language model knowledge](#). *Automated Knowledge Base Construction (AKBC)*.
- Takashi Sasaki. 2026. Exemplars primarily shift retrieval, verification instructions can shift rejection: An anatomy of inference-time interventions for closed-book knowledge base construction. In *AKBC*.
- Maksim Silchenko. 2026. Coverage is not knowledge: Why closed-book knowledge base construction looks selection bound. In *AKBC*.
- Elena Simperl and Markus Luczak-Rösch. 2014. [Collaborative ontology engineering: A survey](#). *The Knowledge Engineering Review*.
- Sneha Singhania, Jan-Christoph Kalo, Simon Razniewski, and Jeff Z. Pan. 2023. [Preface: LM-KBC challenge 2023](#). In *LM-KBC*.
- Sneha Singhania, Tuan-Phong Nguyen, and Simon Razniewski. 2022. [LM-KBC: Knowledge base construction from pre-trained language models](#). In *LM-KBC*.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. [YAGO: A core of semantic knowledge](#). In *WWW*.
- Yiqun Sun, Pengfei Wei, and Lawrence B. Hsieh. 2026. A forced answer channel for closed-book knowledge base construction at LM-KBC 2026. In *AKBC*.
- Mikołaj Tym. 2026. Pre-training aware directed reasoning for closed-book knowledge base construction. In *AKBC*.
- Vlad-Mădălin Ungureanu, Matei Neacșu, and Ianis Belu. 2026. Ensembling elicitation strategies for knowledge base construction. In *AKBC*.
- Quoc-Linh Vo. 2026. COVER-KBC: Evidence-centric adaptive inference for closed-book knowledge base construction. In *AKBC*.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: A free collaborative knowledgebase](#). *Communications of the ACM*.
- Bohui Zhang, Elisavet Koutsiana, Yihang Zhao, Albert Meroño-Peñuela, and Elena Simperl. 2025a. [Trustworthy knowledge graphs: Practices and approaches](#). Handbook on Neurosymbolic AI and Knowledge Graphs.
- Bohui Zhang, Albert Meroño-Peñuela, and Elena Simperl. 2025b. [Towards explainable automated knowledge engineering with human-in-the-loop](#). *Semantic Web*.
- Bohui Zhang, Ioannis Reklos, Nitisha Jain, Albert Meroño-Peñuela, and Elena Simperl. 2023. [Using large language models for knowledge engineering \(LLMKE\): A case study on Wikidata](#). In *LM-KBC*.
- Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. [What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models](#). In *Findings of EMNLP*.

Adrian Zimmermann. 2026. MuseList: Confidence-gated multi-surface recall for closed-book knowledge base construction. In *AKBC*.

A Appendix

Table 4 gives per-relation results.

Submission	Award	Stock	Borders	Area	Capacity	Death City	Average
Silchenko (2026)	0.3484	0.8530	0.9786	0.8700	0.3367	0.6100	0.7060
Kunze and Both (2026)	0.3556	0.8461	0.9873	0.8400	0.3146	0.5900	0.6908
Khalil and Mourad (2026)	0.3378	0.8717	0.9580	0.8200	0.3163	0.5700	0.6836
Zimmermann (2026)	0.3557	0.8400	0.9652	0.8500	0.2449	0.5700	0.6700
Sasaki (2026)	0.3357	0.8318	0.9598	0.8200	0.3163	0.5300	0.6670
Ungureanu et al. (2026)	0.3357	0.8105	0.9680	0.8200	0.2857	0.5667	0.6651
Sun et al. (2026)	0.2358	0.7838	0.9481	0.8600	0.2449	0.5700	0.6553
Tym (2026)	0.3551	0.7797	0.9708	0.7900	0.2160	0.5200	0.6289
Bui et al. (2026)	0.3691	0.7343	0.9463	0.7700	0.2347	0.5300	0.6179
Vo (2026)	0.3105	0.7385	0.9291	0.6700	0.1633	0.5700	0.5878
Patcharatrakul and Vateekul (2026)	0.2749	0.7668	0.9440	0.5700	0.2245	0.4900	0.5699
Habibillah and Tjiaranata (2026)	0.1659	0.7050	0.9657	0.5600	0.1327	0.5100	0.5408

Table 4: Per-relation macro-F1 scores of the ranked submissions on the test set. Award, Stock, Borders, Area, Capacity, and Death City correspond to awardWonBy, companyTradesAtStockExchange, countryLandBordersCountry, hasArea, hasCapacity, and personHasCityOfDeath, respectively. The best score for each relation and the best overall score are shown in **bold**.