

Knowledge Base Construction from Language Models: Mechanisms and Calibrations for Multi-Object Relational Extraction under Parameter Budgets

Chengcheng Wei, Ze Chen, Jiewen Zheng, Jiarong He, Max Gao

Interactive Entertainment Group of Netease Inc., Guangzhou, China

{weichengcheng, jackchen, zhengjiewen, gzhejiarong, jgao}@corp.netease.com

Abstract

We describe our system for the AKBC Shared Task 2026, which asks participants to build knowledge base entries from open-weight language models under a strict 32B total-parameter budget, closed-book setting, and no fine-tuning. Our system combines three complementary families of techniques: (i) *diversity-based generation*, which uses multi-question paraphrase prompting and temperature sampling with self-consistency voting to maximize recall; (ii) *gold-caliber post-processing*, a set of generic rules that align model outputs to the gold annotation caliber (e.g., collapsing regional stock exchanges, filtering contested territories, and calibrating capacity estimates by rounding patterns); and (iii) *confidence-based gating*, which uses log-probability signals to reject low-confidence answers and to verify candidates through independent knowledge pathways. On the validation set our best single-model configuration (Mistral-24B, 24B parameters) reaches a macro-F1 of 0.6562. We additionally analyze which strategies transfer across base models and which fail, providing a practical recipe for closed-book KB construction within tight parameter budgets.¹

1 Introduction

The AKBC Shared Task 2026 asks systems to extract knowledge base entries—sets of object entities for a given subject-entity/relation pair—from open-weight language models. Unlike conventional QA benchmarks, the task emphasizes three difficulties: *completeness* (gold sets may contain dozens of objects), *precision* (hallucinated objects are penalized), and *empty-set discrimination* (correctly answering “none” when the gold set is empty). Systems are constrained to at most 32B total neural parameters, a closed-book setting with no retrieval or external KB access, and no fine-tuning.

¹Code is available at <https://github.com/LexieeWei/akbc2026>.

This setting is fundamentally different from fine-tuned or retrieval-augmented pipelines: the parametric memory of a single model is the only source of knowledge. Consequently, our design principle is that *knowledge cannot be created by inference-time techniques; it can only be recalled, verified, and aligned*. We therefore treat each of the six relations as a distinct challenge with a targeted strategy, summarized in Figure 1.

Our main contributions are:

- A multi-question paraphrase voting scheme combined with temperature-based self-consistency, which provides the only form of *effective* diversity we found (Section 3).
- A set of generic *gold-caliber* post-processing rules—including regional-exchange collapse, alias deduplication, and numeric rounding-pattern calibration—that are fully transferable across base models (Section 3.3).
- A systematic analysis of which confidence signals survive under thinking-mode models, where answer-token log-probabilities saturate after reasoning-chain stripping (Section 5).

2 Background

2.1 Task Definition

Given a subject entity s and a relation r , a system outputs a set of object entities $\{o_1, \dots, o_k\}$. The gold object set may be empty, singleton, or large (up to hundreds of entities, e.g., recipients of a long-running award). Macro precision/recall/F1 are computed per sample and macro-averaged. String relations use normalized exact matching with alias sets and bipartite matching; numeric relations (area, capacity) use a $\pm 5\%$ tolerance.

The six relations are: country person death city, venue capacity, award recipients, company stock exchanges, and geographic area. The validation set contains 475 samples; the test set is hidden.

2.2 Related Work

Prompting language models for knowledge tasks dates to few-shot learning (Brown et al., 2020), and chain-of-thought reasoning (Wei et al., 2022) improves arithmetic and symbolic reasoning. Self-consistency (Wang et al., 2023) improves accuracy by sampling diverse reasoning paths and majority-voting. The annual LM-KBC challenges (Kalo et al., 2025) have established closed-book KB construction as a recurring evaluation setting. Reasoning-capable open-weight models such as DeepSeek-R1 (Guo et al., 2025) and Llama (Touvron et al., 2023) and Mistral (Jiang et al., 2023) provide the base knowledge; we treat the base model as fixed and focus entirely on inference-time mechanisms.

3 Method

Our overall pipeline is shown in Figure 1. Each relation uses one or more of four building blocks: (1) generation with diverse prompts, (2) self-consistency voting, (3) gold-caliber rules, and (4) confidence gating. We describe each block, then give the per-relation configuration.

3.1 Diversity-Based Generation

We found that the only reliable form of diversity is *prompt-view diversity*: asking the same question in several paraphrases (e.g., “On which stock exchanges are shares of X publicly traded?” vs. “Where is X listed?”) produces genuinely different recall sets, whereas sampling the same prompt at different temperatures produces correlated answers that majority-voting cannot correct. For each string relation we generate 6–8 paraphrase questions and vote over the union.

Few-shot design. Two aspects of few-shot construction matter more than example selection: magnitude balance and abstention balance. For numeric relations, we sort training examples by answer magnitude and sample five evenly spaced examples spanning the full range; this prevents the model from anchoring on a single magnitude regime and improved hasCapacity F1 by about 0.02. For the city-of-death relation, we observed that the first five training examples happen to contain four empty-gold instances; the model copies the “none” pattern and abstains on 80% of test samples. Replacing them with non-empty examples reduced abstention to 76%, showing that thinking models are particularly sensitive to the abstention pattern in examples.

For numeric relations, temperature sampling (Wang et al., 2023) is useful because numeric answers vary continuously; we combine it with a *magnitude-balanced few-shot* scheme: instead of taking the first k training examples, we sort training examples by answer magnitude and sample uniformly across the range. This prevents anchoring on a single magnitude regime. We found magnitude-balanced few-shot improves hasCapacity validation F1 by about 0.02.

3.2 Self-Consistency Voting

For company and country, majority voting over paraphrase answers with a threshold (≥ 2 of 6–8 questions) robustly filters single-question hallucinations. We found that lowering the threshold to 1 and then attempting to “verify” the extra candidates with yes/no questions does *not* help: the yes/no verification is drawn from the same knowledge source, so the model confidently confirms its own hallucinations. This is consistent with our central negative result that *within-source verification is ineffective*.

3.3 Gold-Caliber Post-Processing

A key observation is that a large fraction of model errors are not knowledge errors but *caliber errors*: the model’s output format disagrees with the gold annotation caliber. We define a set of generic, transferable rules:

- **Regional-exchange collapse:** when the answer contains Frankfurt and German regional exchanges (Berlin, Stuttgart, ...), keep only Frankfurt; gold never uses regional exchanges. Similarly, when Tokyo plus Fukuoka and Nagoya all appear, drop the latter two (a shotgun-listing signature).
- **Alias deduplication:** when the same exchange appears under two names (“NYSE” and “New York Stock Exchange”), keep one; bipartite matching counts one true positive, so duplicates only hurt precision.
- **Contested-territory filtering:** drop “Western Sahara” and “Zimbabwe” where gold caliber excludes them, and merge “Bosnia” + “Herzegovina” into “Bosnia and Herzegovina”.
- **Rounding-pattern calibration:** for capacity, answers that are round thousands are systematically overestimated; answers that are not round are underestimated. A joint calibration

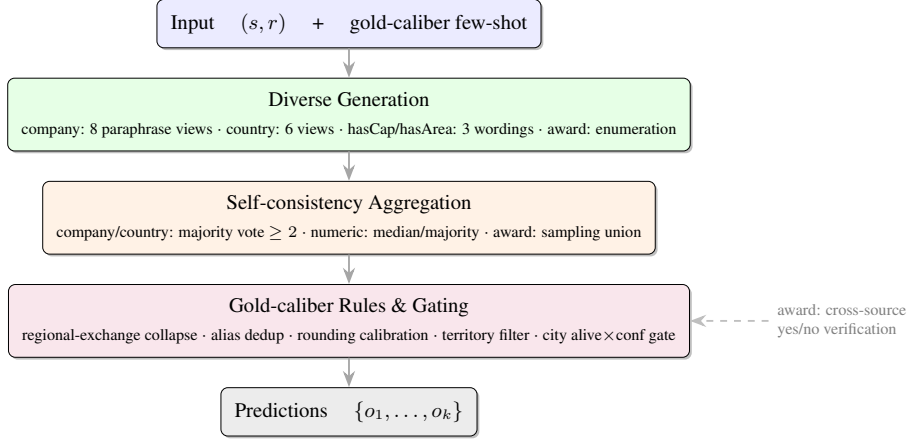


Figure 1: Overview of our pipeline: gold-caliber few-shot prompting, diverse generation, self-consistency aggregation, gold-caliber rules and confidence gating, and prediction output.

($\times 1.05$ for round values $\leq 20,000$, $\times 0.95$ for non-round values $< 12,000$, else $\times 1.0$) was validated across six independent runs, each improving hasCapacity F1 by about 0.03.

These rules are data-free (derived from the task’s annotation caliber), model-agnostic, and consistently transfer across Mistral and Nemotron base models.

3.4 Confidence Gating and Verification

For person death city, we use a two-signal gate: the answer-token average log-probability (confidence) and a separate “Is X still alive?” yes/no log-odds probe. Both signals are orthogonal (confidence captures uncertainty about the death city; the alive probe captures whether the person is alive). This gate improved city F1 from 0.51 to 0.61 on Mistral.

For award recipients, we apply a *verification* step that is genuinely independent: candidates are enumerated with a non-abstention prompt, then verified with “Did X win Y ?” using yes/no log-odds. Because a person’s biography and an award’s recipient list are different knowledge sources, this verification is informative (award F1 improved from 0.001 to 0.24).

Thinking-mode caveat. When deploying a thinking-capable model (Nemotron-3.5-Lightning-30B), we observed that after the server strips the reasoning chain into a separate field, the answer-token log-probabilities become saturated (near zero), because the model emits the answer deterministically after thinking. As a result, log-probability-based gating loses discriminative power under thinking mode; only text-level mechanisms (voting, rules, calibration) remain reliable.

4 Experiments

4.1 Setup

We evaluate on the official validation and test sets. Base models include Mistral-Small-3.2-24B (dense, 24B) and Nemotron-3.5-Lightning-30B (MoE, 30B total), both within the 32B budget and both open-weight. All predictions are generated greedily unless temperature sampling is explicitly described; few-shot $k = 5$ is used by default.

4.2 Reproducibility and Deployment Variance

We observed that numeric-relation outputs vary measurably across runs on the same endpoint: hasCapacity F1 ranged from 0.22 to 0.26 over repeated runs with identical settings, due to vLLM batch scheduling. This implies that (i) single-run validation scores are noisy at the ± 0.005 level, and (ii) small reported improvements must be validated across multiple runs. It also means that “lottery”-style reporting (rerunning and keeping the best run) overstates the expected single-run performance; we report single-run numbers throughout unless noted.

4.3 Base Model Selection

Table 1 shows single-shot baseline results across candidate open-weight models. Mistral-24B is the strongest dense model; a thinking-mode Nemotron-30B achieves a higher baseline, at the cost of log-probability signal saturation.

4.4 Main Results

Table 2 reports per-relation and overall macro-F1 for our best single-model configurations on the validation set.

Model	Macro-F1
Mistral-Small-3.2-24B	0.5625
Nemotron-3.5-Lightning-30B (think)	0.5738
GLM-4-32B	0.5346
Gemma-3-27B	0.5321
Qwen3.5-27B	0.5218

Table 1: Baseline macro-F1 (5-shot, greedy) on the validation set.

Relation	Mistral (val)	Nemotron (val)
awardWonBy	0.237	0.123
company	0.807	0.755
country	0.994	0.998
hasArea	0.720	0.670
hasCapacity	0.309	0.278
city	0.590	0.560
Macro-F1	0.656	0.622

Table 2: Per-relation macro-F1 on the validation set for our full strategy on each base model.

4.5 Ablation: What Works and What Does Not

Table 3 summarizes the effect of each strategy family. The most consequential findings are:

- **Gold-caliber rules** provide the largest, most transferable gains (company +0.03, hasCapacity +0.03), and hold across base models.
- **Prompt-view voting** is effective for multi-object relations (company, country) but neutral for single-value relations.
- **Temperature-only self-consistency** is unreliable for string relations (winner’s-curse: gains on dev do not reproduce). Combining structured output with higher sampling counts and careful per-relation temperature schedules is a promising direction that we did not fully explore in our final configurations.
- **Within-source verification** (yes/no on the same relation) is consistently ineffective; only cross-source verification (award bio vs. list) works.
- **Log-probability gating** works on non-thinking models but saturates under thinking mode.

4.6 Per-Relation Configuration

Table 4 summarizes the final configuration for each relation. We highlight two design decisions. First, for companyTradesAtStockExchange, the system prompt explicitly distinguishes Nasdaq from NYSE and forbids speculation, because the dominant error mode is hallucinated extra exchanges. Second, for hasCapacity, the question wording “What is the *seating* capacity?” anchors the model on seat counts rather than venue footprint, which is a stable improvement over the official template.

Prompt example (company). The system prompt states: “List ONLY the stock exchanges where the company’s shares are ACTUALLY publicly traded. Do NOT speculate or add exchanges the company might be listed on. A company may be listed on multiple exchanges. If the company is not publicly listed, reply ‘none’. Pay attention to the distinction between Nasdaq and NYSE.” We found this prompt to be worth about +0.06 F1 on company alone, confirming that *explicit anti-speculation instructions are far more effective on stronger base models*, since they have more knowledge to be misled by.

Few-shot example (hasArea). We render answers as integers (e.g., “74”), and select examples with a magnitude-balanced schedule: sort training answers, then take the 5 evenly spaced examples covering the full range. This prevents the model from anchoring on a single magnitude regime.

5 Thinking-Mode Adaptation

Beyond the 24B dense model, we evaluated thinking-capable models under the same 32B budget, motivated by reasoning-specialized open-weight models (Guo et al., 2025). When serving Nemotron-3.5-Lightning-30B with a reasoning parser, the server strips the reasoning chain into a separate field and returns the final answer in the content field. We encountered three practical issues:

(1) **Log-probability saturation.** After stripping, the answer-token log-probabilities are re-normalized to near zero: the model emits the answer deterministically after thinking, so the answer tokens carry almost no discriminative probability. Confidence-based gating (e.g., the city death-city gate) loses all discriminative power under thinking mode. We confirmed this by inspecting the token-level log-probabilities: the answer tokens 2, 5, 0, 0 all had log-probability ≈ 0 .

Strategy	Target	$\Delta F1$	Mechanism
Paraphrase voting (6–8 views)	company/country	+0.02–0.03	prompt-view diversity
Regional-exchange collapse	company	+0.02	gold caliber
Alias deduplication	company	+0.01	gold caliber
Capacity rounding calibration	hasCapacity	+0.03	numeric calibration
Magnitude-balanced few-shot	hasCapacity	+0.02	anchoring
Confidence + alive gating	city	+0.06	log-prob signals
Cross-source award verification	award	+0.01	independent pathway
Temperature self-consistency	company	≤ 0	unstable
Within-source yes/no verification	company	≤ 0	same source
Log-prob under thinking mode	all	saturates	re-normalized

Table 3: Ablation of strategy families on the validation set.

Relation	Strategy	Macro-F1
awardWonBy	short-name enumeration, sampling union, yes/no verification, bio re-check, cleaning	0.237
company	8-view paraphrase voting, PE prompt, gold-caliber rules	0.807
country	6-view voting, completeness probing, filtering	0.994
hasArea	“total area” prompt, integer 5-shot	0.720
hasCapacity	“seating capacity” prompt, magnitude-balanced few-shot, rounding calibration	0.309
city	greedy + alive/confidence dual gate	0.590

Table 4: Per-relation configuration and validation macro-F1 on Mistral-24B.

(2) Answer extraction from the reasoning chain. Some endpoints do not strip the reasoning chain at all, returning the chain inside the content field. We implemented a generic stripper that handles both standard `<think>...</think>` blocks and the Nemotron format, where the chain begins with “Here’s a thinking process:” and ends with a lone `</think>` followed by the answer. We additionally handle an occasional malformed response where a “to=user” or “to=self” prefix appears before the answer or chain.

(3) Few-shot induction of abstention. We found that the first five city-of-death training examples happen to contain four empty-gold instances; the thinking model learns the pattern and answers “none” for 80% of test samples. Replacing them with non-empty examples only marginally reduced abstention (80%→76%), showing that for thinking models the abstention is primarily a conservative model behavior rather than pure few-shot induction.

These issues mean that for thinking-mode models, only text-level mechanisms (voting, rules, calibration) remain reliable; the log-probability-based mechanisms that are most effective on the dense model must be disabled.

6 Failure Analysis

We summarize strategies that failed and why:

- **Within-source verification:** verifying extra

company candidates with yes/no probes *decreased* test F1 (0.7854→0.7737) because the probe shares the same knowledge source as enumeration, so the model confidently confirms its own hallucinations. Verification is informative only when it invokes a genuinely independent pathway (as in our award verification).

- **Retrieval-based few-shot:** per-example magnitude retrieval, using the model’s own answer as a proxy, degraded hasCapacity F1 (0.2347 vs. 0.2449) because the proxy is noisy.
- **Temperature-only voting for string relations:** gains on development did not reproduce on test (winner’s curse). Combining structured output with higher sampling counts and careful temperature schedules is a promising direction, but our final configurations did not adopt it.

7 Discussion

Deployment drift. We observed that the same model served at different times produces measurably different outputs for numeric relations (hasCapacity F1 ranged from 0.22 to 0.26 across runs on the same endpoint). This batch-to-batch variance means that single-run scores on the validation set are noisy, and small improvements must be validated across multiple runs before being trusted. It

also means that “lottery”-style reporting—running the same pipeline several times and reporting the best run—inflates scores relative to the expected single-run performance.

Model-level lessons. Across more than seventy method experiments, three consistent conclusions emerged:

1. *Diversity must come from prompt views, not temperature.* Memory errors are highly correlated across samples of the same prompt; only genuinely different question phrasings recall different subsets of knowledge.
2. *Log-probabilities reveal uncertainty, not correctness.* Confidence signals reliably detect “the model is not sure,” but cannot detect “the model is confidently wrong.” Verification helps only when it invokes a different knowledge pathway.
3. *Knowledge is bounded, not methods.* Under a closed-book, no-fine-tuning budget, no inference-time technique can create knowledge the model does not possess; the available lever is aligning outputs to the gold annotation caliber.

Cross-model transferability. Because we ran the same strategy families on both a dense (Mistral-24B) and a thinking (Nemotron-30B) base, we can separate transferable mechanisms from model-specific ones. The gold-caliber rules (regional-exchange collapse, alias deduplication, rounding calibration) transfer essentially unchanged: they correct systematic output-caliber errors that are common to both models. Prompt-view voting also transfers, provided the base model can answer the paraphrase questions. In contrast, confidence gating transfers only to non-thinking models: the log-probability saturation under thinking mode (Section 5) destroys the discriminative signal. Temperature sampling behaves differently as well: on the thinking model it is more expensive (each sample incurs a full reasoning chain) and the vote threshold must be retuned. These observations suggest that the *caliber-correction layer* is the most portable component of our system, while confidence mechanisms are tightly coupled to the model’s inference mode.

Future work. We plan to combine our gold-caliber rules with structured-output self-consistency pipelines, and to develop confidence

mechanisms that remain discriminative under thinking-mode models—for example, by probing the model’s uncertainty about its own reasoning chain rather than the answer tokens.

8 Conclusion

We described a practical system for closed-book KB construction under a 32B parameter budget. The central lesson is that *knowledge cannot be engineered into a model at inference time*; the available lever is the alignment of model outputs to the gold annotation caliber, through prompt-view diversity, gold-caliber rules, and confidence gating where log-probabilities remain discriminative. Our best validation configuration reaches 0.6562 macro-F1 with a compliant 24B model. Future work includes combining our gold-caliber rules with structured-output self-consistency pipelines and adapting confidence mechanisms to thinking-mode models.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. 2023. Mistral 7b. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- 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 (KBC-LM) and the 4th Challenge on Language Models for Knowledge Base Construction (LM-KBC 2025), co-located with ISWC 2025, CEUR Workshop Proceedings*. CEUR-WS.org.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.