AKBC Shared Task 2026 · AKBC Workshop @ EMNLP 2026, Budapest

COVER-KBC

Evidence-Centric Adaptive Inference for Closed-Book Knowledge Base Construction

Quoc-Linh Vo

Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

Trust what is right. Keep asking what is missing.

INPUT

Query $q = (s, r)$: predict the full object set $\hat{O}_{s,r}$ from the model's parameters alone.

s = *Vietnam*
r = *countryLandBordersCountry*
gold: {Cambodia, China, Laos}

Six relations, four program types

- SMALLSET land borders, stock exchanges
- NULLSINGLE city of death
- NUMERIC area, venue capacity
- LARGEOPENSET award winners

Answers may be empty, one entity, a number or a long list.

A Relation Contract

$C_r = (\tau_r, V_r, P_r, G_r, F_r)$
type, views, parser, verify & stop, finalizer

Acquisition views V_r

- direct recall
- structural decomposition
- contrastive questions
- missingness search
- reverse check

Each relation is compiled into an **executable program**: which views to ask, how to parse, when to verify or stop, and how to finalize.

Relation conditioning changes the program, not just the prompt wording.

B Is it correct?

Every model call becomes a parsed observation in an evidence state S_t.

Evidence state

China	● ● ●	dots =
Laos	● ●	distinct views
Cambodia	not yet observed	

Blind verifier sees one candidate
VALID / INVALID / UNKNOWN

$S(o) = \alpha F(o) + \beta L(o) - \delta C(o) - \eta U(o)$

support from *distinct views* (F), calibrated verifier (L), minus contradiction (C) and template disagreement (U).

Repeats of one prompt count once; an UNKNOWN counts against.

C Anything missing?

Candidate confidence ≠ completeness: a separate residual search need q_{res}.

q_{res}: still open

stop < 0.15

$q_{\text{res}} = \text{clip}_{[0,1]}\big(\max(\bar{\phi}_w, \phi_{\text{mand}}, \phi_{\text{block}})\big)$

One open gap is enough: an unrun mandatory view, a rival numeric cluster, an unresolved existence gate.

Controller: next view, facet, verify, reverse check, resample *or* stop

Budget: 12 calls and 6,000 tokens per query.

OUTPUT

Relation-aware finalization reads the *evidence state*, never raw text.

- SMALLSET validated set closure
- NULLSINGLE existence gate, then city
- NUMERIC value clusters within 2.5%
- LARGEOPENSET accepted, deduplicated

0.5878
Macro-F1, official hidden test

One frozen Mistral-Small-3.2-24B; no retrieval, no training.

not settled ⇒ keep asking: acquire more evidence

1. Introduction

Closed-book KBC. For each query $q = (s, r)$ the system must predict the complete object set $\hat{O}_{s,r}$ using only the parametric knowledge of one language model: no retrieval, no fine-tuning, at most 32B neural parameters. The AKBC Shared Task 2026 covers six heterogeneous relations whose answers may be empty, a single entity, a number, or a long list.

Why this is hard

- Facts elicited from an LLM are **unstable across prompts**; voting over samples measures what looks confident, not what is missing.
- Relations differ in **cardinality and missingness**: an empty answer can be correct or a failed recall.
- **Correct is not complete**: trusting each candidate does not mean the answer set is closed.

▶ *Key idea: closed-book KBC is an active evidence-acquisition problem with two separate questions – is this candidate correct, and is anything still missing?*

Contributions

1. **Relation-typed inference contracts** that control elicitation, verification, stopping and finalization.
2. **Independence-aware evidence** with blind, calibrated candidate verification.
3. **Residual search need** driving a budget-aware controller, plus relation-aware finalization.

2. Method

2.1 Relation-typed inference programs

Each relation r is compiled into a contract $C_r = (\tau_r, V_r, P_r, G_r, F_r)$: relation type, eligible acquisition views, parsing and normalization, verification and stopping, and the final selector. It also fixes the relation definition, positive cases, hard negatives and cardinality.

Relation	Type	Final strategy
countryLandBordersCountry	SMALLSET	strict small-set closure
companyTradesAtStockExchange	SMALLSET	+ empty-row rescue views
personHasCityOfDeath	NULLSINGLE	deceased gate, then city
hasArea	NUMERIC	multi-view value clustering
hasCapacity	NUMERIC	multi-view value clustering
awardWonBy	LARGEOPENSET	enumeration + cleanup

2.2 Independence-aware evidence and blind verification

Observations are stored in $S_t = (O_t, R_t, B_t)$: candidates, evidence records with view, prompt provenance and independence group, and budget state. Repeated samples from one view are *not* independent corroboration; support counts distinct acquisition groups,

$$F(o) = \frac{|g(o)|}{|m(o)|}, \qquad S(o) = \alpha F(o) + \beta L(o) + \gamma X(o) - \delta C(o) - \eta U(o),$$

with α=1.0, β=0.6, δ=1.5, η=1.0 (X=0: one checkpoint). The verifier sees only the subject, relation definition, hard negatives and one candidate, and its logits are calibrated against a content-free prompt:

$$\tilde{p}_j = \frac{\exp((z_j - b_j)/T)}{\sum_k \exp((z_k - b_k)/T)}, \qquad L(o) = \text{clip}_{[-1,1]}\Big(\tfrac{1}{3}\log \frac{\tilde{p}_{VALID} + \epsilon}{\tilde{p}_{INVALID} + \tilde{p}_{UNKNOWN} + \epsilon}\Big).$$

Placing UNKNOWN in the denominator makes abstention count *against* a candidate, which keeps precision high on small sets.

2.3 Residual search need and adaptive control

q_{res} takes the maximum of soft gap signals $\bar{\phi}_w$ (set instability, unresolved mass, numeric dispersion, marginal yield), a mandatory-view gate and a typed blocking gap, so one unsatisfied condition forces continuation. The controller stops on budget, on a closed program with no pending verification, or when $q_{\text{res}} < 0.15$; otherwise it takes the best legal action

$$a_t^\star = \arg\max_a \ \lambda_y Y_t(a) + \lambda_g G_t(a) + \lambda_i I_t(a) - \lambda_c \text{cost}(a) - \lambda_d D_t(a).$$

2.4 Relation-aware finalization

Only strings resolved into supported evidence are eligible for output. Numeric values merge when $|a - b|/\max(|a|, |b|) \le 0.025$, stricter than the evaluator's 5% tolerance; without a well-supported cluster, a second round arbitrates at 0.05 with a source-blind judge, so a repeated number cannot win by repetition alone.

2.5 Targeted rescue for empty stock rows

An empty stock-exchange prediction more often reflects failed elicitation than a true absence, so it triggers candidate-blind views on direct listing, primary listing, ordinary-share listing and anti-prior verification. Structural validation then rejects ticker symbols, indexes, regulators, market segments, country and company names, and an exchange is emitted only with sufficient independent support.

3. How the Program Types Behave

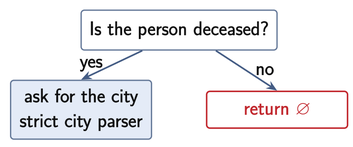
SMALLSET
Closing a set

Vietnam, land borders. Two views agree on China and Laos.

China ● ● ●
Laos ● ●
Cambodia ▯

Both are correct, but set instability keeps q_{res} above the threshold, so a **missingness search** runs and recovers Cambodia before the set is closed.

NULLSINGLE
Empty can be right



Existence is a **prior gate**, not a post-filter: a confidently negative gate closes the relation with a principled empty set instead of a failed recall.

NUMERIC
Clusters, not strings



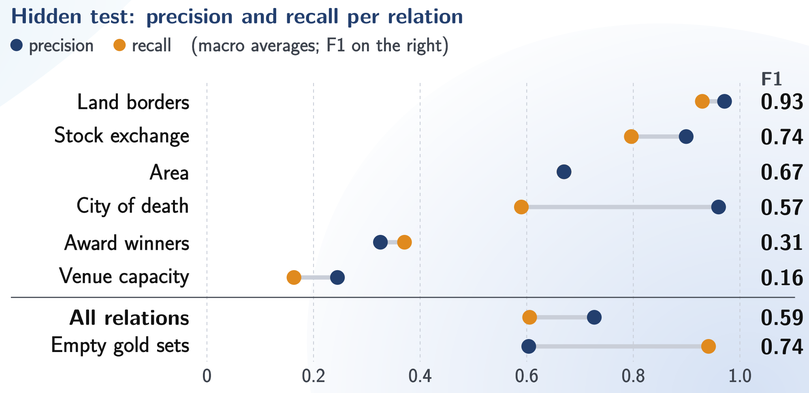
Views are merged within 2.5% relative distance. A competitive rival cluster raises the **blocking gap** and keeps the relation open until a source-blind judge arbitrates.

4. Results

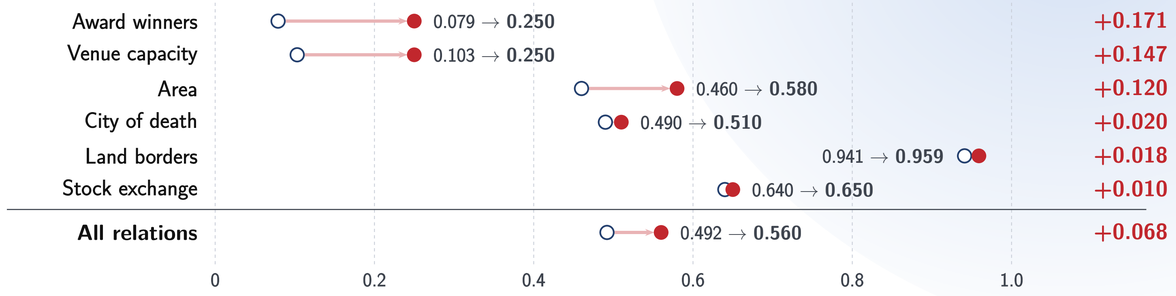
Setup. One frozen Mistral-Small-3.2-24B-Instruct checkpoint (24.01B parameters) for every role; greedy decoding, nf4 on one A100; no retrieval, external KB or fine-tuning. Official AKBC 2026 benchmark and evaluator.

0.5878
Macro-F1
hidden test
Precision 0.7268
Recall 0.6055
Empty gold sets
recall 0.9412

Hidden test: precision and recall per relation



Same LLM, direct prompting vs. COVER-KBC (validation Macro-F1)



▶ *COVER-KBC improves on the same backbone for all six relations; the largest gains are on the hardest open-set and numeric relations.*

5. Findings, Limitations and Conclusion

Key findings

- **Small sets are strongest**: land borders reach 0.93 F1 at 0.97 precision.
- **Empty can be right**: empty gold sets reach 0.94 recall.
- **Conservative by design**: precision exceeds recall on most relations.

Limitations

- **Venue capacity** (0.16 F1): the model rarely commits to the maximum published figure.
- **Award winners** (0.31 F1): long lists stay recall-limited without retrieval.
- Contracts are hand-specified; multi-step inference costs more than one prompt.

▶ *Heterogeneous relations need different inference programs: set closure for small sets, explicit empty answers for nullable relations, value clustering for numbers and continued discovery for open sets.*

Code & configuration: `github.com/vquclinh/cover-kbc`

Acknowledgments. This research is partially supported by the Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam.

vqlinh2433@apcs.fitus.edu.vn
github.com/vquclinh/cover-kbc