ORAL PRESENTATION

TGLN-Cascade

A Knowledge Graph Completion Framework Fusing Tanh-Gated Transformer and Cascade Rerankers

Guo Cheng · Yueqi Zhu · Kexin Sun · Yongkang Zhang · Mingxia Gao*

Beijing University of Technology

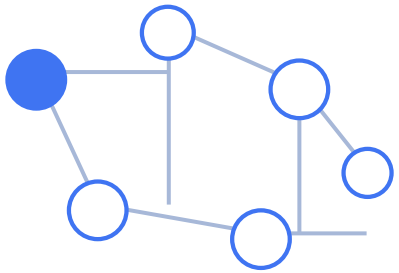
STRUCTURAL RECALL → SEMANTIC RERANK → EXPERT REASONING

Recall broadly. Rerank semantically. Reason selectively.

MOTIVATION

Why is knowledge graph completion still difficult?

Real-world KGs are incomplete, but the missing fact may be hidden among thousands or millions of entities.



(h, r, ?)

Which entity completes the fact?

Sparse symbolic features

Standard LayerNorm applies fixed statistical scaling. On sparse KG features, this can smooth or collapse discriminative signals.

Semantic ambiguity

Structural similarity alone can keep candidates that look plausible in the graph but are semantically mismatched.

LLM sensitivity

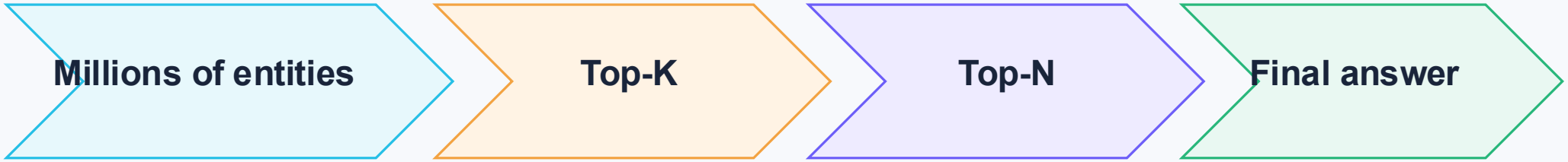
LLMs reason well, but poor prompts or noisy candidate sets can disrupt reasoning and increase hallucination risk.

We need a pipeline that preserves recall, removes semantic noise, and gives the LLM only high-quality candidates.

CORE IDEA

One model should not be responsible for everything

TGLN-Cascade decomposes KGC into a recall-rerank-reasoning cascade.



1 TGLN structural recall

Distance-aware Transformer + bounded Tanh gate. Objective: keep the correct answer inside a small Top-K pool.

2 BGE semantic rerank

Fine-tuned cross-encoder scores query-candidate semantic relevance and filters structure-only noise.

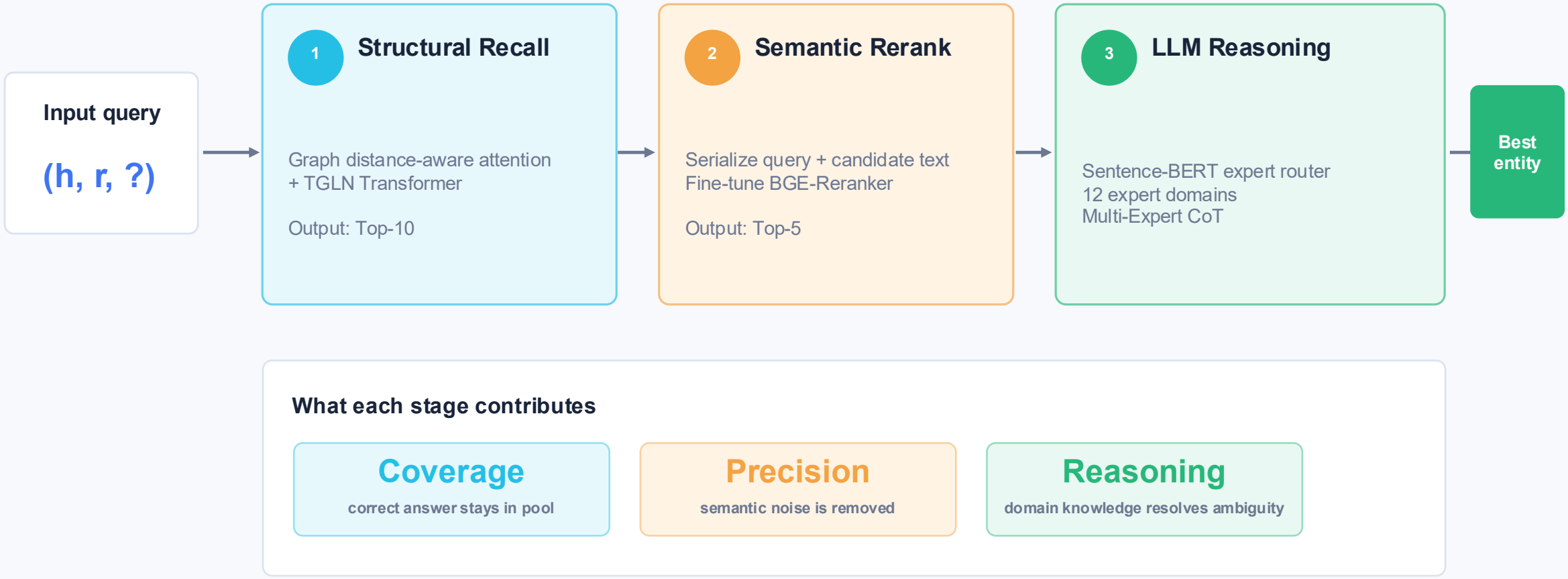
3 Multi-Expert LLM reasoning

Route each relation to a domain expert persona, then apply structured Chain-of-Thought for final ranking.

Key design principle **Optimize each stage for the job it is best at - coverage, semantic precision, or reasoning.**

METHOD OVERVIEW

TGLN-Cascade: end-to-end framework

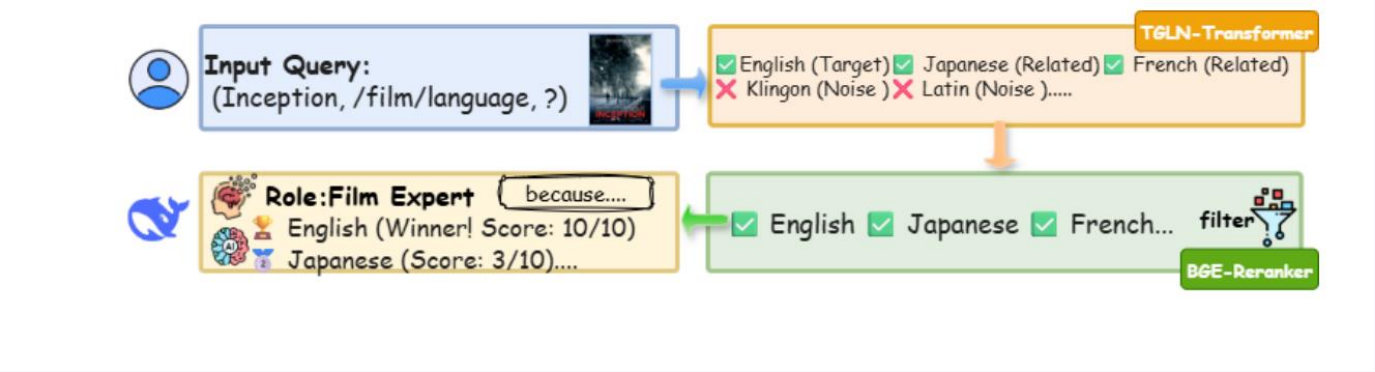


The cascade reduces search space before expensive reasoning, making LLM usage focused rather than brute-force.

RUNNING EXAMPLE

A concrete example: predicting the language of Inception

Query: (Inception, /film/language, ?)



1 **Recall**

TGLN-Transformer retrieves English, Japanese, French, but may also keep noisy entities such as Klingon or Latin.

2 **Rerank**

BGE-Reranker removes candidates that are structurally plausible but semantically mismatched.

3 **Reason**

The relation is routed to a Film Expert; the LLM compares the remaining candidates and selects English.

This example shows why the stages are complementary: recall tolerates noise, reranking removes it, and expert reasoning resolves the final ambiguity.

TANH-GATED LAYER NORMALIZATION

Stage 1: TGLN is designed for recall, not Top-1 ranking

Standardize first

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

Then apply a bounded gain

$$y_{TGLN} = \hat{x} \odot (\gamma \tanh(\alpha)) + \beta$$

tanh(α) ∈ (-1, 1)

Recall-oriented behavior on WN18RR

Hits@10 0.591 → 0.685 +15.9% relative

Conceptual effect on sparse features

Standard LayerNorm

fixed statistical scaling → weaker sparse contrast

TGLN

bounded adaptive gain → preserves useful sparse signals

The first stage only needs the correct answer to survive into the candidate pool.

RERANK + REASON

Stages 2 and 3: semantic filtering before expert reasoning

STAGE 2

BGE-Reranker

Structured KGC → text matching

q = Desc(h) + CleanText(r) → (q, candidate description)

Contrastive fine-tuning learns to score the correct entity higher than negatives.

Top-10 → Top-5 semantic noise removed

STAGE 3

Multi-Expert CoT

Relation r → Sentence-BERT cosine router → Best expert

12 domains: Film · Geography · Business · History · Music · Literature · Sports · Life Sciences · Education · Technology · Awards · General

Expert-level reasoning prompt

1 Semantic analysis

2 Context integration

3 Item-wise scoring

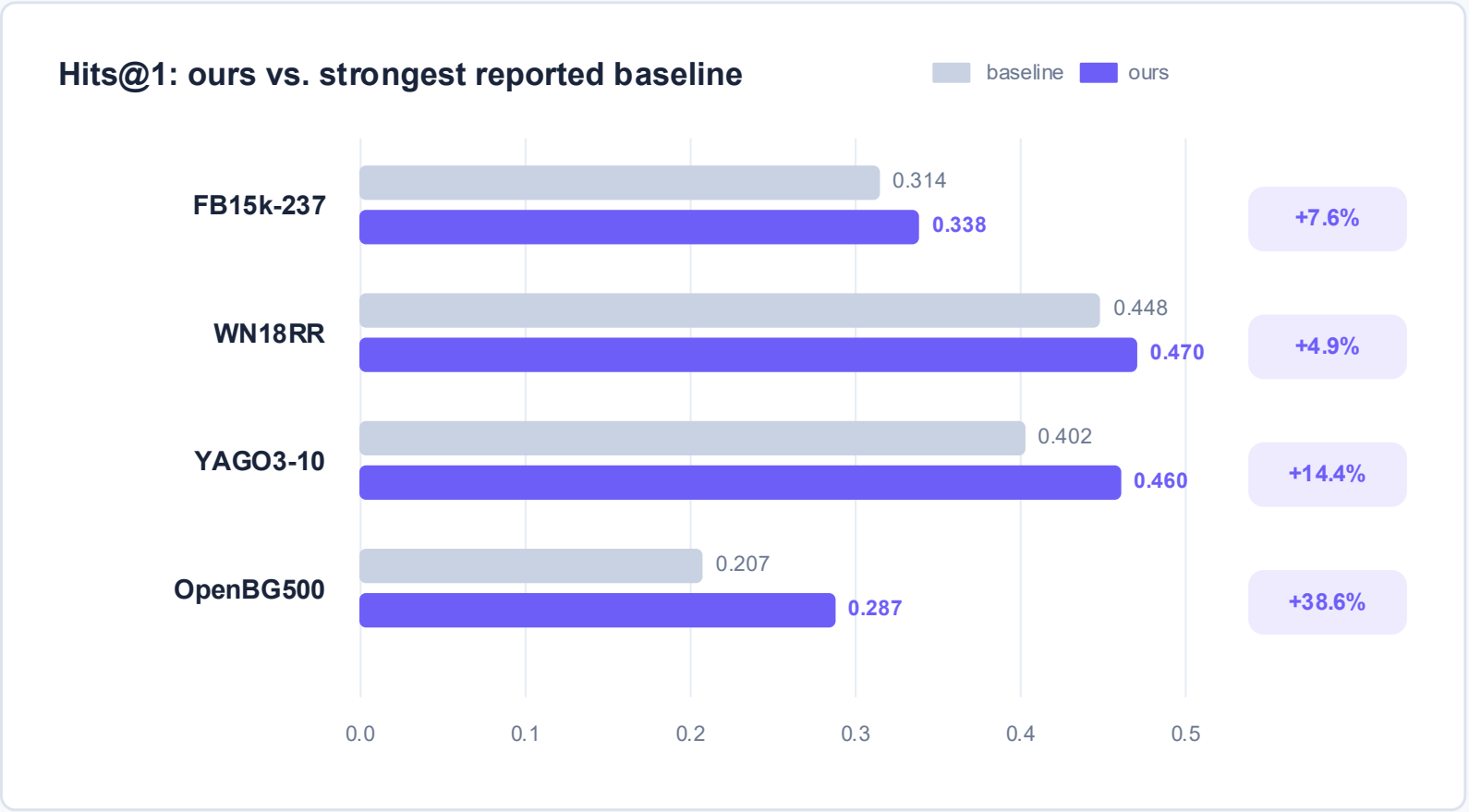
4 Structured JSON output

Reranking makes the candidate set trustworthy; expert routing makes the LLM reasoning relation-aware.

EXPERIMENTS

Main results: strong Top-1 accuracy across four benchmarks

Single RTX 4090 · BERT-base recall · Top-10 → Top-5 · DeepSeek-V2 reasoning · temperature 0.1



Key results

0.338
FB15k-237 Hits@1

0.470
WN18RR Hits@1

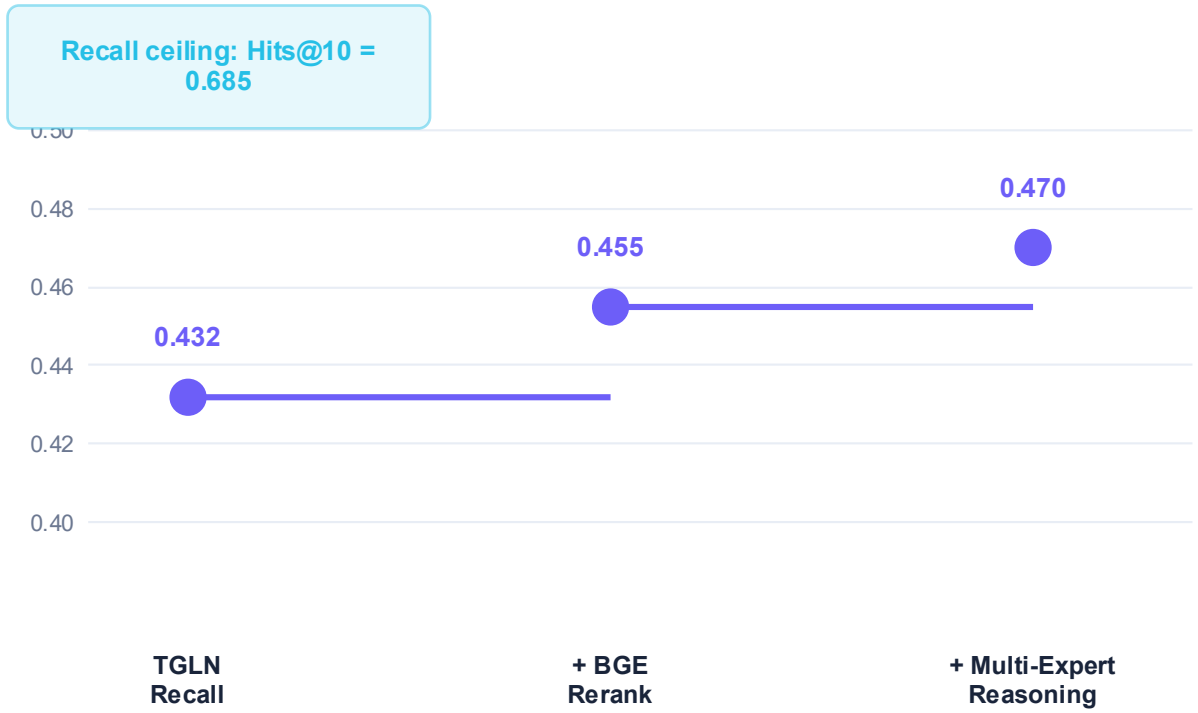
0.685
WN18RR Hits@10

Hits@1 is best among the reported baselines on all four datasets.

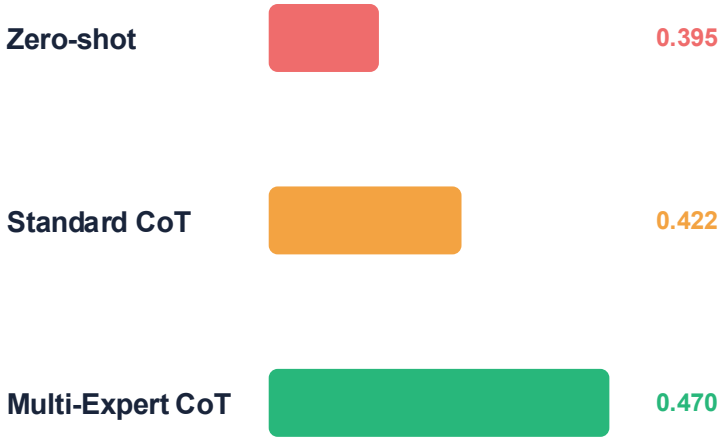
WHY IT WORKS

Ablation: each stage contributes a different kind of gain

Progressive Hits@1 on WN18RR



Prompt strategy matters



General prompting can even hurt. Relation-specific expert personas are essential for the final gain.

The cascade works because recall, semantic correction, and expert reasoning solve different failure modes.

CONCLUSION

Takeaways

TGLN-Cascade turns KGC into a sequence of increasingly precise decisions.

1 Recall first

TGLN uses a bounded adaptive gain to improve candidate coverage on sparse and hierarchical knowledge graphs.

2 Filter semantically

A fine-tuned BGE-Reranker removes structure-only noise before expensive reasoning.

3 Reason selectively

Multi-Expert CoT routes each relation to a suitable domain persona and improves final Top-1 accuracy.

Limitation **LLM reasoning latency remains a challenge for large-scale real-time queries.**

Future work **Distill LLM reasoning into smaller student models.**

Recall broadly. Rerank semantically. Reason selectively.

Thank you.