Exploring Medium-Sized LLMs for Knowledge Base Construction

Tomás Cerveira da Cruz Pinto
University of Coimbra, CISUC/LASI, DEI
tomaspinto@student.dei.uc.pt

Chris-Bennet Fleger
University of Potsdam, Germany
chris-bennet.fleger@uni-potsdam.de

Hugo Gonçalo Oliveira
University of Coimbra, CISUC/LASI, DEI
hroliv@dei.uc.pt

Financiado pela União Europeia NextGenerationEU

INTRODUCTION

Knowledge base construction (KBC) is one of the great challenges in Natural Language Processing (NLP) and of fundamental importance to the growth of the Semantic Web. Large Language Models (LLMs) can be useful for automating the extraction of structured knowledge, including subject-predicate-object triples.

LM-KBC Challenges

The LM-KBC Challenges benchmark the ability of LLMs to generate knowledge triples by predicting objects from subject–relation pairs. In 2023, the task was split into Track 1 (models with fewer than one billion parameters) and Track 2 (models of any size). A gap was left: models close to one-billion parameters remained underexplored.

Main Contributions

- Benchmarking medium-sized LLMs for KBC.
- Exploring relation-specific prompting.
- Ensemble strategies for improved predictions.

DATA & MODELS

LM-KBC 2023 Dataset

- Subject-relation pairs.
- Ground-truth objects.
- 21 relations (e.g., *CountryHasStates*, *PersonHasProfession*).
- Objects: people, organizations, countries, counts, “none”.

Bruce Willis, **PersonHasProfession,** **Actor**

Subject entity — Relation — Object entity

Models Evaluated

Model	Params
Llama3.2-Instruct	1B
Gemma2-it	2B
Qwen2.5-Instruct	0.5B
Qwen2.5-Instruct	1.5B
DeepSeek-R1	1.5B

Table 1: Medium-Sized LLMs explored.

METHODS

Prediction Task

- Generate object(s).
- Form triples from given subject-relation pairs.

Prompting Strategies

- Zero-shot (relation-specific instruction + question).
- Few-shot (3-shot, question & triple formats).
- Context Paragraph from subject’s Wikipedia page.

Ensemble Approaches (3 Best Models)

- Majority voting (Most chosen response).
- Relation-based (Best model for each relation).

Evaluation

Metric: Macro F1 against Wikidata ground truth IDs.

Task Explanation: Please answer the question with your knowledge. Beforehand there are a few examples. The output format should be a list of possible answers prefaced by "Answer: ", also if there is no answer write Answer: ["]
Demonstrations:
Bo Burnham, PersonPlaysInstrument, Answer: [piano]
What instrument does Marko Topchii play? Answer: [guitar]
Example Triple: Kevin Pabst, PersonPlaysInstrument, Answer: [trumpet]
Example Question: What instrument does Kevin Pabst play? Answer: [Trumpet]

Figure 1: Few-shot prompt example format.

Optional Context: 1st paragraph of subject’s Wikipedia page.
Question Part: Who are the members of The Beatles?
Instruction Part: List only the members, separated by ", " with no extra text.
Example Output: John Lennon, Paul McCartney, George Harrison, Ringo Starr

Figure 2: Zero-shot prompt example format for the *BandHasMember* relation.

RESULTS & DISCUSSION

Method	Comparison	F1 Score
BERT	Challenge Baseline	0.142
GPT-3	Challenge Baseline	0.061
VE-BERT	Track 1 Winner	0.323
LLMKE (GPT-4)	Track 2 Winner	0.701

Model	Best Method	F1 Score
Llama3.2	0-shot + paragraph context[1]	0.271
Gemma2	0-shot + paragraph context[1]	**0.377**
Qwen2.5 - 0.5B	3-shot triple	0.188
Qwen2.5 - 1.5B	0-shot + paragraph context[1]	0.281
DeepSeek-R1	3-shot triple	0.093
Ensemble	Relation-based	**0.384**
Ensemble	Majority-voting	0.381

Table 2: Macro F1 for Each Model and Method, including Baselines and Track winners.

1 – Used on the ensembles.

Model Insights

- Qwen 1.5B & Llama did not match the reasoning capacity of Gemma2, proving that size is not a guarantee of performance and the role of architecture and pre-training data.
- Instruction-tuning is key: DeepSeek failed.

Prompting Strategies

- Providing clear, relation-specific task instructions outperformed the addition of task demonstrations on the prompts.
- Using Wikipedia boosted performance by providing context enrichment.

Ensembles

- The best ensemble shows a tiny gain over individual Gemma2.
- Ensembles work best when models' performances are balanced, not dominated by one.

Comparison with LM-KBC 2023

- Outperformed Track 1 participants via prompting, without any form of training.
- Proportionally competitive scores vs. Track 2 giants (GPT-4).
- Recent mid-size models can balance accuracy and computational efficiency.