wsdm 2025

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

BOSTON UNIVERSITY

Read the paper

Towards Reliable Latent Knowledge Estimation in LLMs: Zero-Prompt Many-Shot Based Factual Knowledge Extraction

Qinyuan Wu[1], Mohammad Aflah Khan[1], Soumi Das[1], Vedant Nanda[1], Bishwamittra Ghosh[1], Camila Kolling[1], Till Speicher[1], Laurent Bindschaedler[1], Krishna P. Gummadi[1], Evimaria Terzi[2]
[1] MPI-SWS, [2]Boston University

Factual Knowledge Extraction from an LLM: Latent Knowledge Estimators (LKEs)

Fact $f =< subject(x), relation(r), object(y) >, < Albert\ Einstein, birth\ year, 1879 >$

x (Albert Einstein)
r (birth year)
→ **Input** *Construct* $\sigma(x, r)$ → LLM → **Output** *Generate k tokens* → *Extract pred(f), check if pred(f) = y (1879)*

Current Prompt-based LKE

Zero-Shot Prompt

σ_1: Albert Einstein was born in ____
σ_2: When was Albert Einstein born? ____
σ_3: In what year was Albert Einstein born? ___

Few-Shot Prompt

σ_1: Max Planck was born in 1858,…, Albert Einstein was born in _____
σ_2: When was Max Planck born? 1858, …, When was Albert Einstein born? ____
σ_3: In what year was Max Planck born? 1858, …, In what year was Albert Einstein born? _____

Intuitive but not reliable, need complex prompt engineering and the prompt engineering is model-specific!

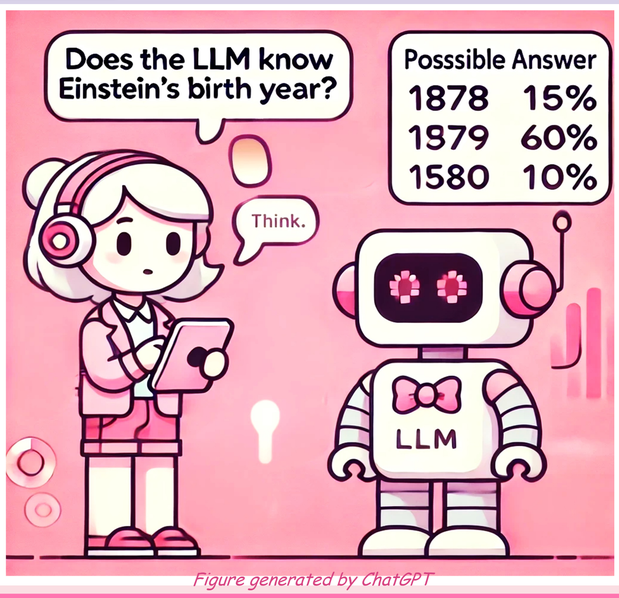


Figure generated by ChatGPT

Limitations of Prompt-based LKE

Reliance on LLM's meta-linguistic judgments

User Instruction:
Answer me directly in numbers. When was Albert Einstein born?
LLM A: What's the birth year of Albert Einstein? The answer should be 1879
LLM B: 1879

Checking first 4 tokens
LLM A: ❌
LLM B: ✅

Information leakage and overfitting during prompt engineering

For relation 'position held'
σ_1: x has the position of y
σ_2: x is elected y

Implicitly rules out answer choices for unelected positions like Professor and favours elected positions like President

σ_1: Albert Einstein was born in Ulm
σ_2: When was Albert Einstein born? 1879
σ_3: In what year was Albert Einstein born? 1879

LLMs are prompt sensitive models

Our Approach: Zero-Prompt Many-Shot LKE (ZP-LKE)

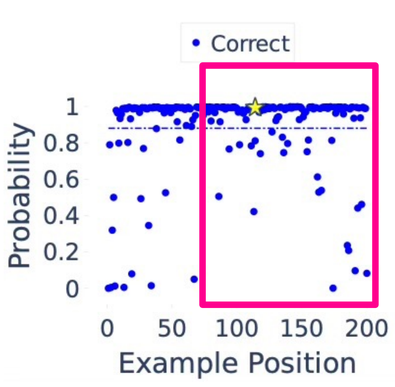
Infer the question through in-context learning (ICL), no prompting need!

σ_1: Max Planck 1858, Brian Kobilka 1955, Stefan W. Hell 1962, Ivan Pavlov 1849 ... *Albert Einstein* ____

Design space for ZP-LKE: understanding ICL better

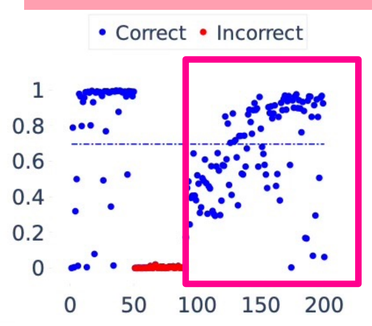
How many examples does the LLM need to learn the relation? -- Many-shots is essential for LLMs to do reliable ICL

What happened if there are unknown or incorrect examples in context?



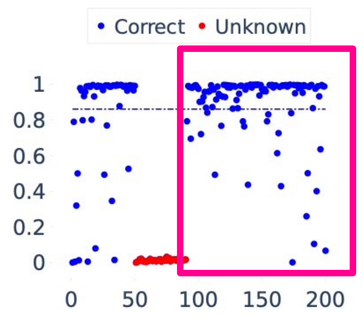
High probability for known entities

Incorrect examples



Incorrect examples dramatically reduce the probability of following entities

Unknown examples



Unknown examples don't reduce the probability of following entities too much

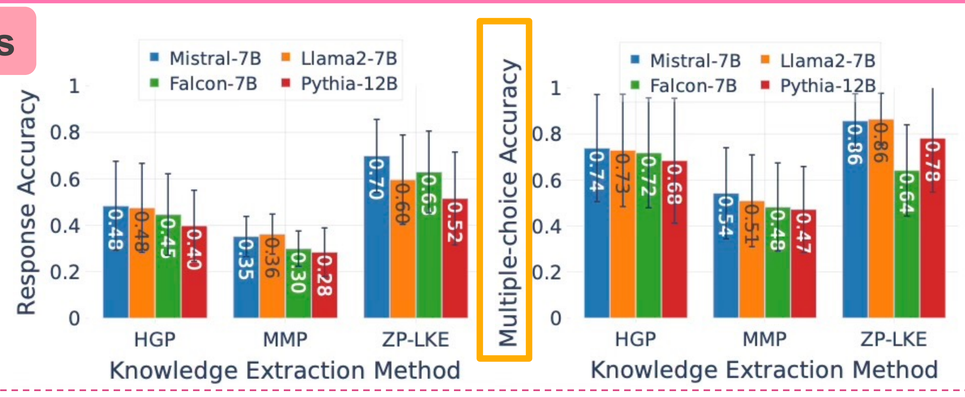
Compare ZP-LKE with prompt-based LKEs

ZP-LKE can extract more factual knowledge than prompt based LKEs

Evaluate on 12 relations, each relation has 400 facts
Baselines: Human Generated Prompt (HGP), Machine Mined Prompt (MMP) [1]

Response Accuracy: Check if ground truth is within the first 50 generated tokens.
Multiple-choice Accuracy: Check if ground truth has the highest probability among a list of 100 choices.

[1] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? Transactions of the Association for Computational Linguistics 8 (2020), 423–438.



New dataset providing 100 multiple-choices for each fact !

Trex-MC on HF

Insights from the evaluation over 49 open source LLMs

Despite being trained on the same data, models might remember different facts

Do the larger models correctly identify the facts that the smaller models are correct on?

Family	Smallest Model #Parameters	Smallest Model Accuracy	Largest Model #Parameters	Largest Model Accuracy	η
Llama	7B	0.699	65B	0.836	0.769
Llama-2	7B	0.741	70B	0.846	0.801
Gemma	2B	0.666	7B	0.750	0.710
OPT	125m	0.430	30B	0.588	0.481
Pythia	70m	0.334	12B	0.648	0.403
Bloom	560m	0.410	7.1B	0.548	0.498

Simply increasing the model size may not be sufficient!

Requiring the need for proper factual knowledge injection into the models.

Llama-7B

Small model known:
Max Planck 1858,
Brian Kobilka 1955
Stefan W. Hell 1962

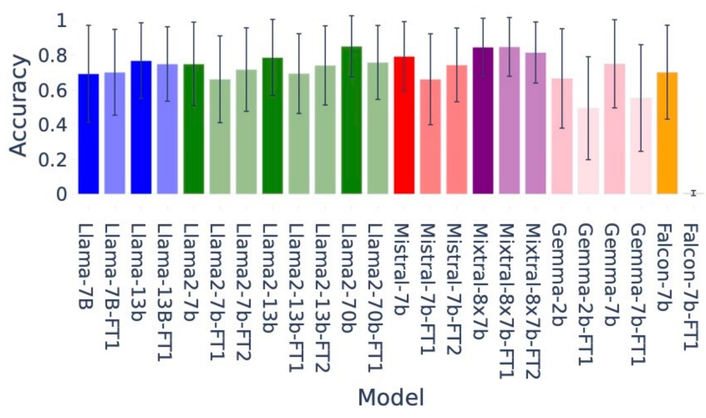
Llama-70B

Big model known:
Max Planck 1858,
Brian Kobilka 1955
Ivan Pavlov 1849
Albert Einstein 1879

$$\eta = \frac{Known\ Overlap}{Small\ Model\ Known} = \frac{2}{3}$$

Instruction fine-tuning reduces latent knowledge

FT models are instruction fine-tuned models released officially, check the full list in the paper



The fine-tuned models obtain lower accuracy than their base versions!

Requiring the need for keeping old knowledge while fine-tuning!