Do Larger Language Models Imply Better Generalization? A Pretraining Scaling Law for Implicit Reasoning

Xinyi Wang, Shawn Tan, Mingyu Jin, William Yang Wang, Rameswar Panda, Yikang Shen

Scaling Laws for text generation

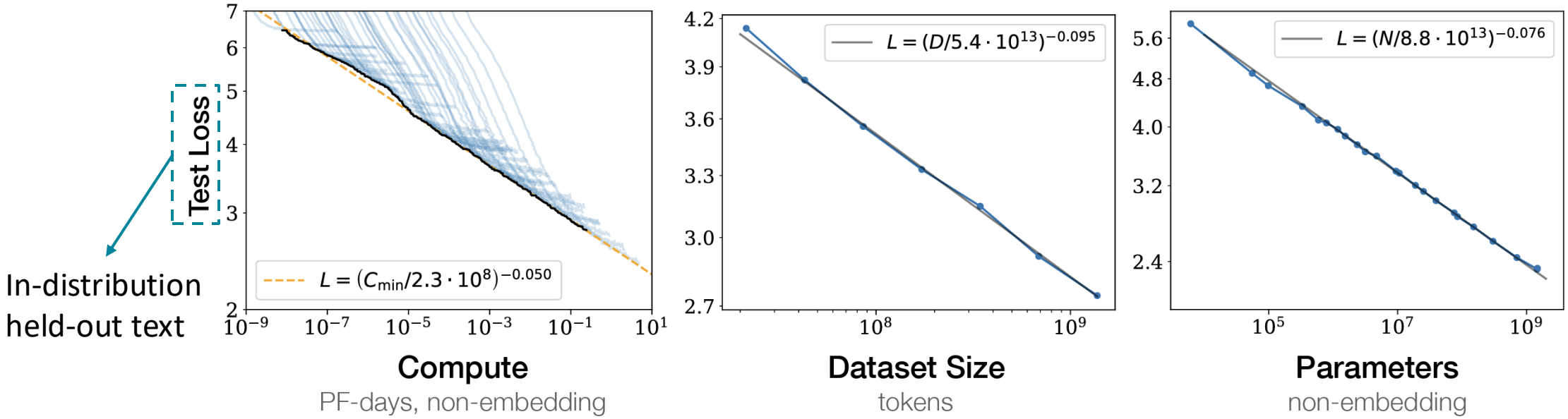


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

(Source: Kaplan et al. 2020)

Reasoning with LLMs

Chain-of-thought prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

(Source: Wei et al. 2022)

Zero-shot chain-of-thought prompting

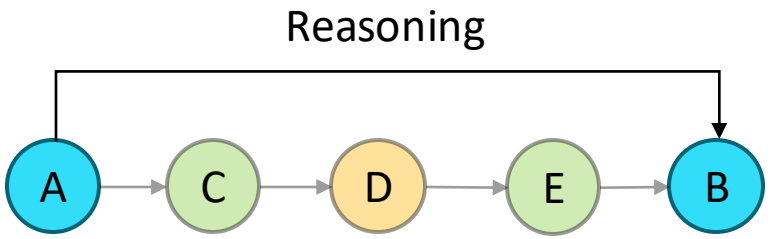
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

(Source: Kojima et al. 2022)

Formalize reasoning

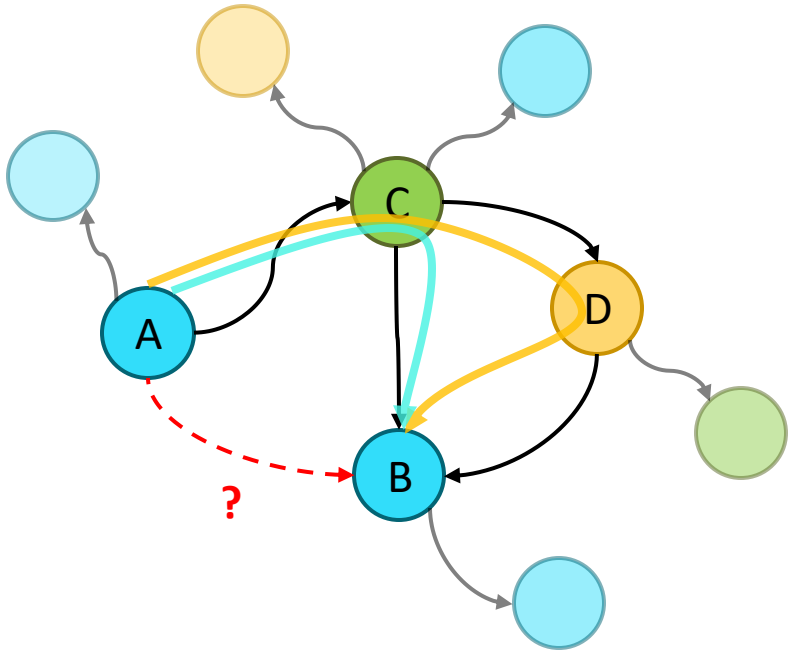
- How can we connect concept A with concept B, if we have never seen them together before?



Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang. Understanding Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation. ICML 2024.

Abstract Reasoning as Graph Completion

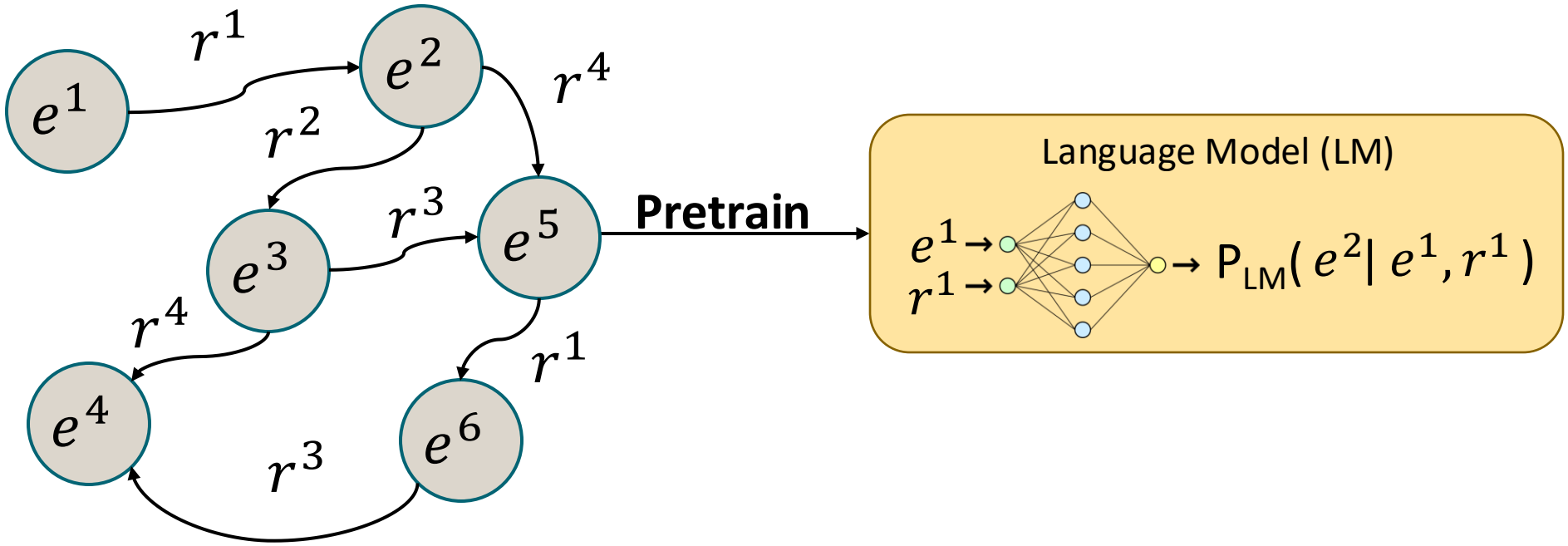
- Suppose we have a large set of connected concepts…



G: World knowledge, knowledge graph…

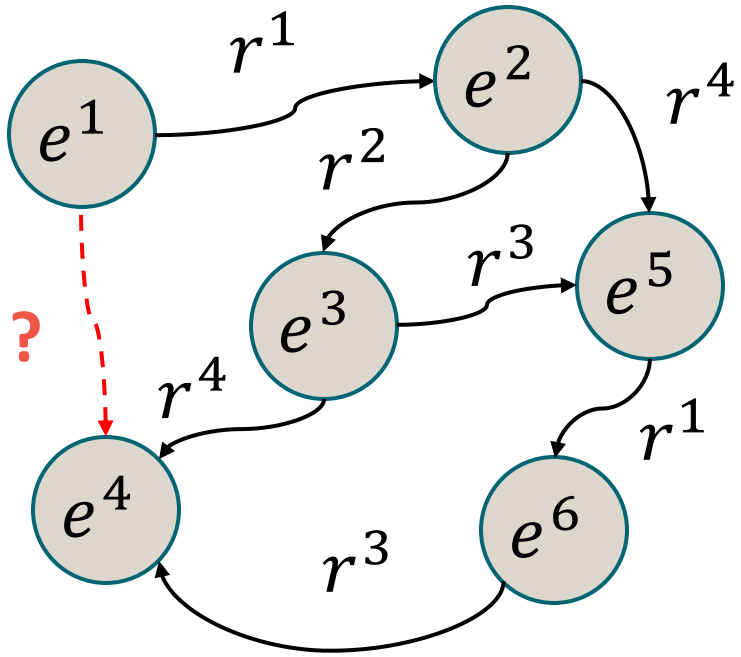
Language model pretraining

- If we pretrain a language model on a knowledge graph with next-token prediction loss, we can get a prediction of the missing edge…

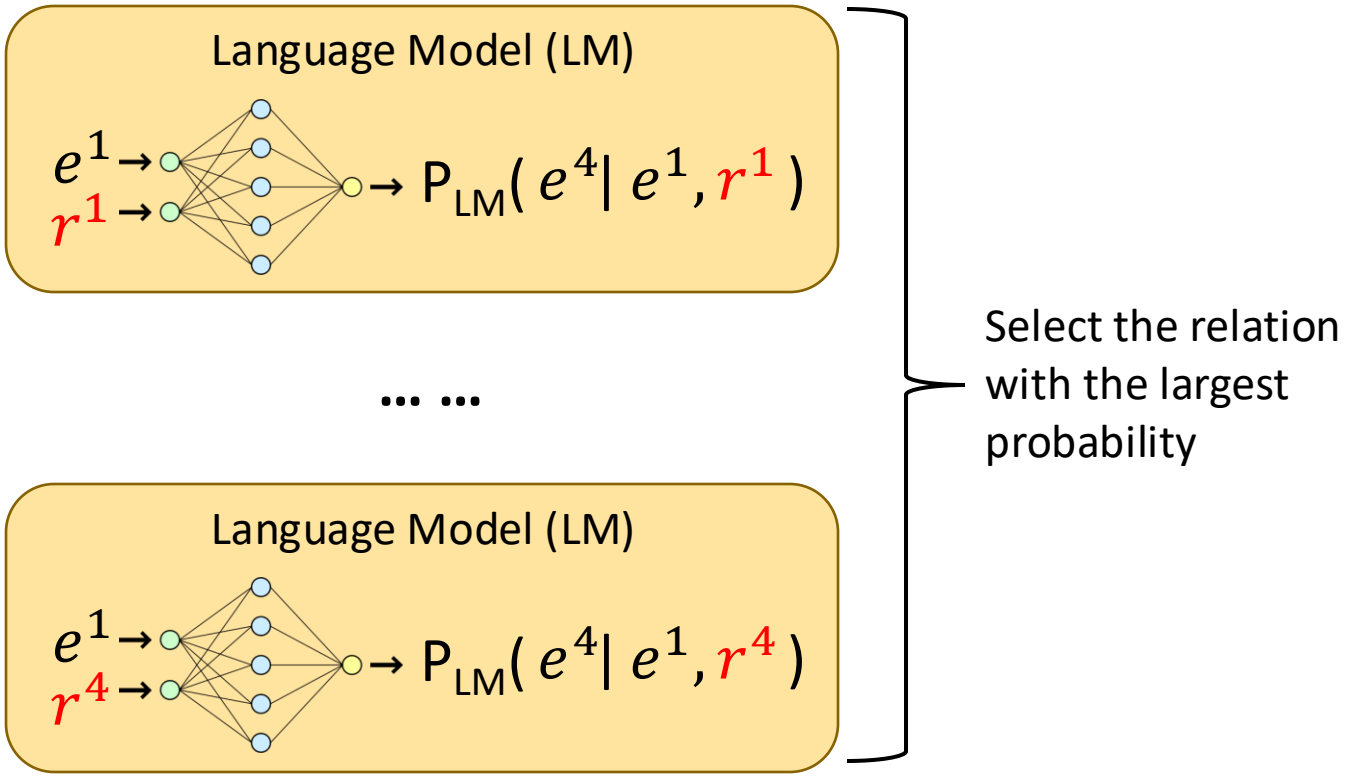


Language model inference

- If we pretrain a language model on a knowledge graph with next-token prediction loss, we can get a prediction of the missing edge…



Inference



Reasoning on real world KGs

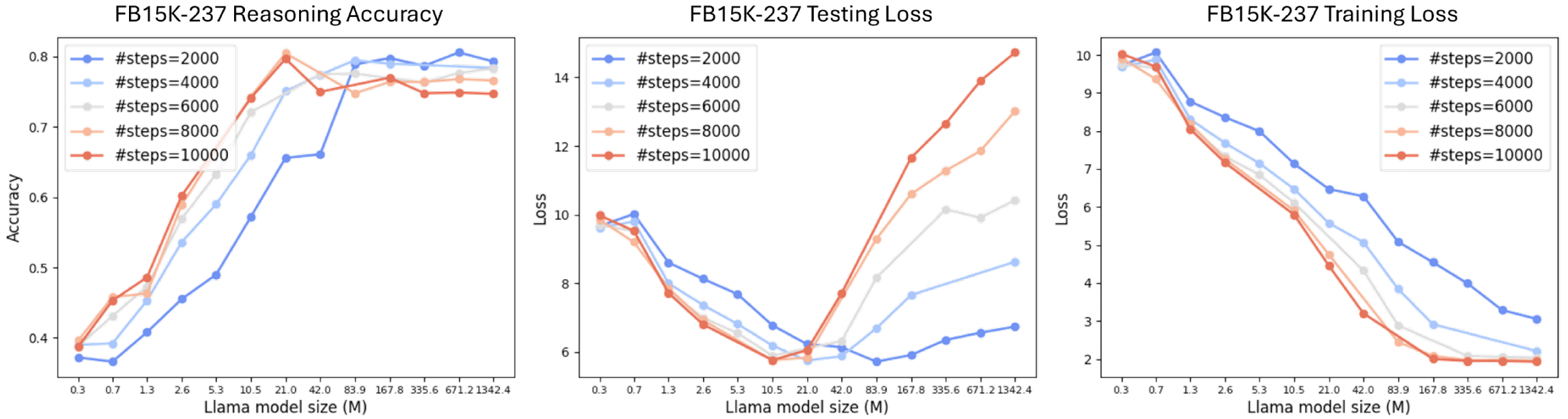


Figure 1: The multiple-choice accuracy/loss on unseen triples of different-sized language models trained on a real-wor h 10k steps) shows that the testing accura hows U-shape loss curves of language models trained with different numbers of steps. The right panel shows the training loss decreases steadily. Note that the model size on x-axis is in log scale.

U-shaped instead of power law!

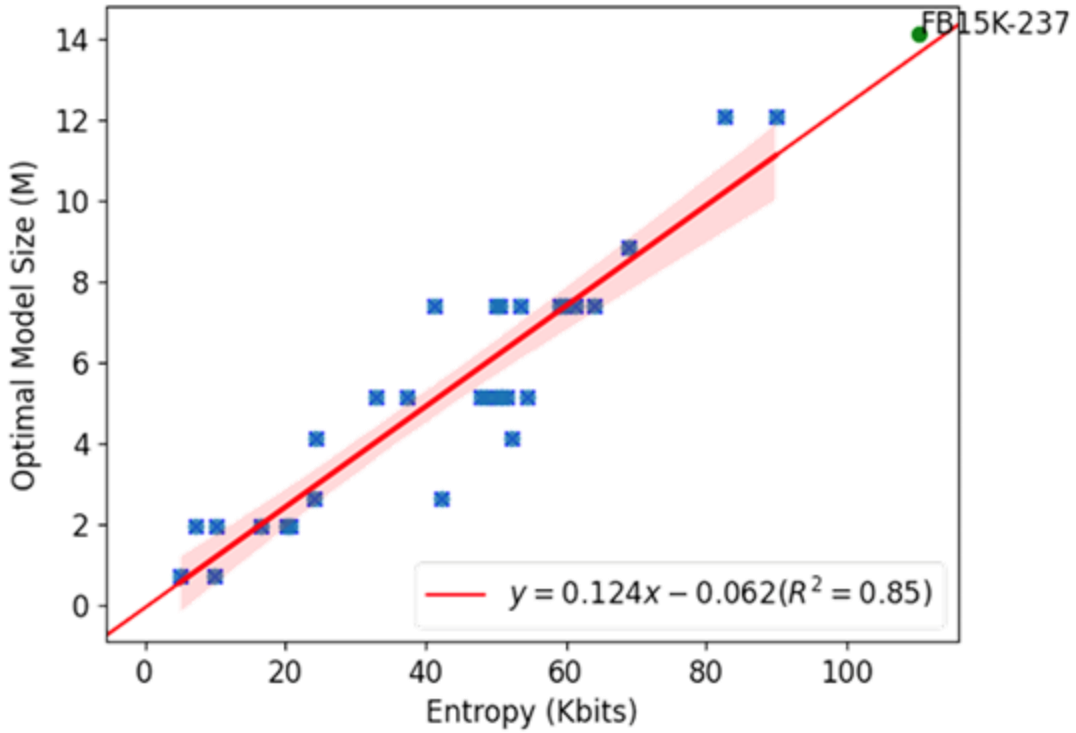
* FB15K-237 has around 15k entities, 237 relations, which results in 310k triples.

Synthetic KG for Controlled Experiments

Key steps:

- Conjunctive rules generation:
  - A is B's father ^ B is C's father -> A is C's grandfather
- Control possible number of relation types connecting to each entity (real-world alignment & reduce noises)
- Grow the graph with preferential attachment to ensure power law degree distribution (real-world alignment)

Optimal Model Size v.s. Graph Search Entropy



Optimal model size linearly related with graph entropy

real-world FB15K-237 experiment (green dot) to verify the accuracy of the obtained linear scaling law.

Graph Search Entropy

- **Def:** #nodes * entropy rate of an infinitely long random walk on G

$$H(G) = N_e(\log(\lambda) + H^r(G)).$$

Entity entropy rate:
log dominant eigenvalue of the adjacency matrix

Relation entropy rate:
relation transition entropy w.r.t. the stationary distribution

Thank you!

Q & A