Latent Multi-Hop Reasoning of Large Language Models

Sohee Yang

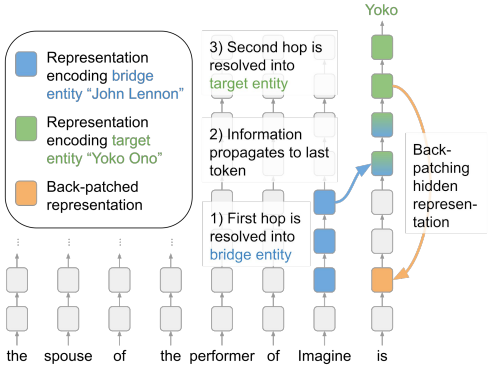
DLCT, Nov 2025

Plan

1. **Evidence of Existence:** Do LLMs perform latent multi-hop reasoning?
2. **Understanding the Mechanism:** How do LLMs perform latent multi-hop reasoning?
3. **Understanding the Current Status through Rigorous Benchmarking:** How much latent multi-hop reasoning ability have today's widely used LLMs naturally developed during pretraining?
4. **Conclusion**

Plan

1. **Evidence of Existence:** Do LLMs perform latent multi-hop reasoning?
2. **Understanding the Mechanism:** How do LLMs perform latent multi-hop reasoning?
3. **Understanding the Current Status through Rigorous Benchmarking:** How much latent multi-hop reasoning ability have today's widely used LLMs naturally developed during pretraining?
4. **Conclusion**



Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries

Eden Biran[1] **Daniela Gottesman**[1] **Sohee Yang**[2] **Mor Geva**[1] **Amir Globerson**[1,3]

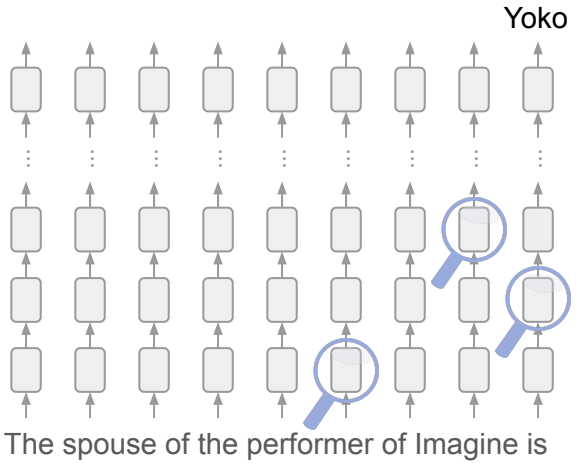
[1]Tel Aviv University [2]UCL [3]Google Research





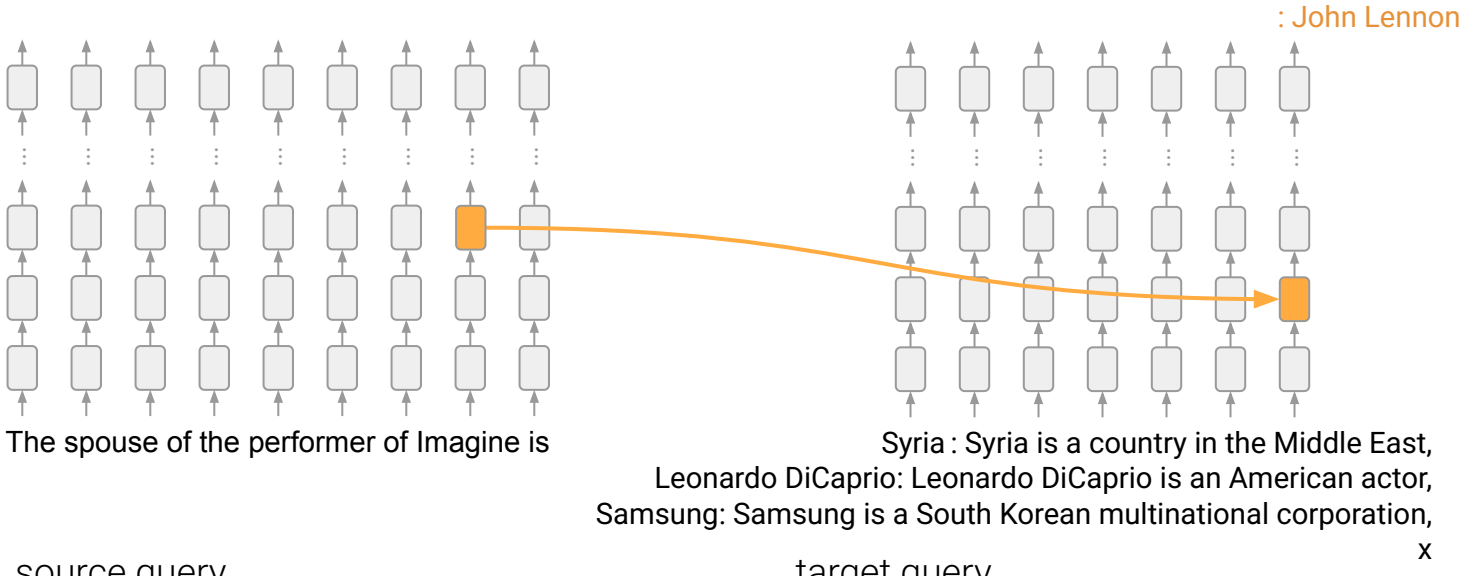
Our Approach

- Answering two-hop queries requires two information extraction steps
- We investigate where these two procedures are implemented in the LLM
    - Seek the first point in the network where the outputs of these two steps appear



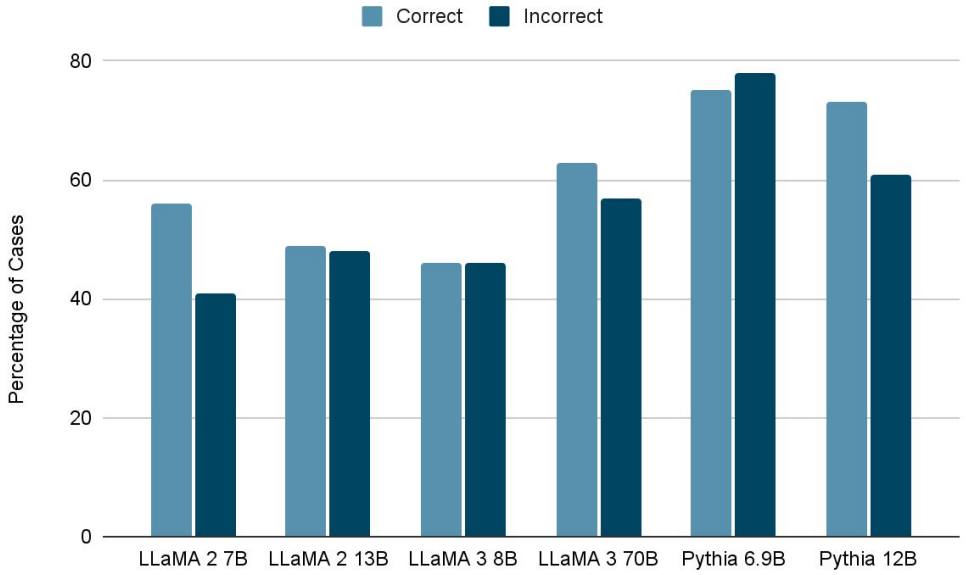
Patchscopes Framework

- Interpretability method introduced by Ghandeharioun et al. (2024)
- Use an LLM to interpret a hidden representation in natural language
- Create a task that describes the entity encoded in a specific hidden representation



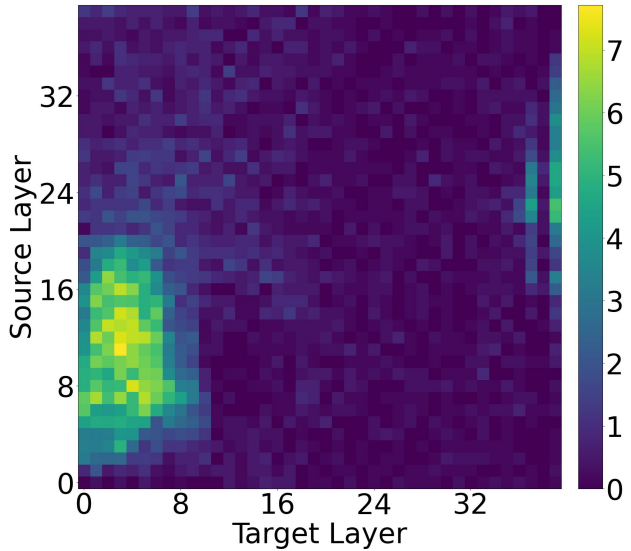
Localizing First Hop Resolution Position

- Run Patchscopes on last token of first hop (t_1)
  - “The spouse of the performer of Imagine is”
- High success rate in decoding bridge entity (“John Lennon”)



Localizing First Hop Resolution Layer

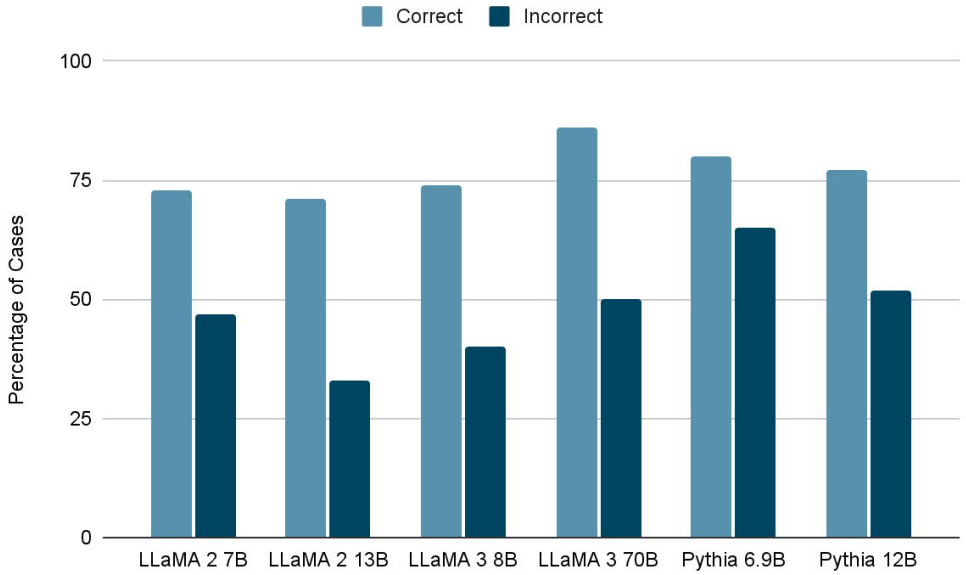
- Run Patchscopes on last token of first hop (t_1)
  - “The spouse of the performer of Imagine is”
- Bridge entity successfully decoded in early layers



Model	Setting	Layer
LLaMA 2 7B	Correct	8.9
	Incorrect	9.1
LLaMA 2 13B	Correct	7.7
	Incorrect	7.4
LLaMA 3 8B	Correct	8.9
	Incorrect	11.9
LLaMA 3 70B	Correct	27.1
	Incorrect	29.2
Pythia 6.9B	Correct	5.4
	Incorrect	4.9
Pythia 12B	Correct	5.0
	Incorrect	6.3

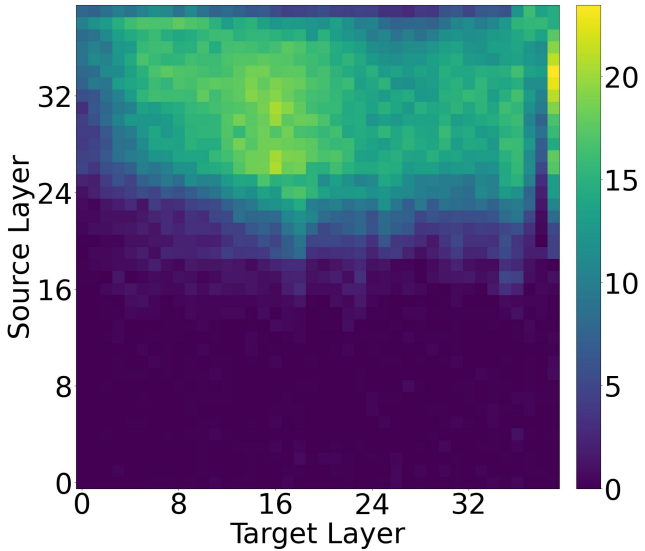
Localizing Second Hop Resolution Position

- Run Patchscopes on last token of prompt (t_2)
  - “The spouse of the performer of Imagine <u>is</u>”
- High success rate in decoding target entity (“Yoko Ono”)



Localizing Second Hop Resolution Layer

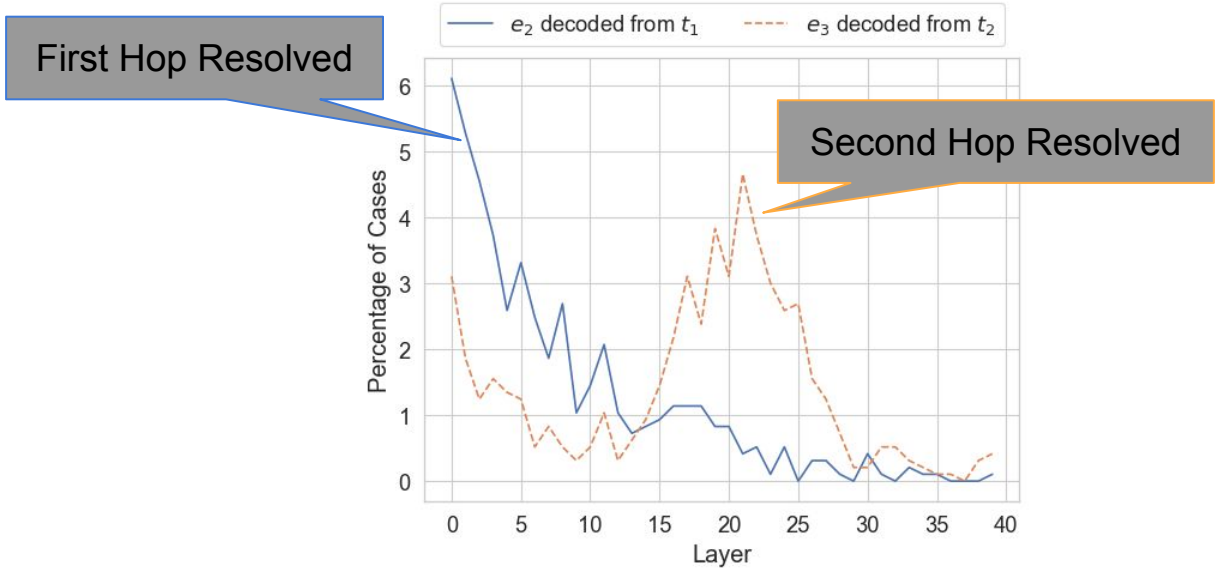
- Run Patchscopes on last token of prompt (t_2)
  - “The spouse of the performer of Imagine is”
- Target entity successfully decoded in mid-upper layers



Model	Setting	Layer
LLaMA 2 7B	Correct	16.2
	Incorrect	17.5
LLaMA 2 13B	Correct	16.9
	Incorrect	16.9
LLaMA 3 8B	Correct	13.5
	Incorrect	17.2
LLaMA 3 70B	Correct	30.7
	Incorrect	35.9
Pythia 6.9B	Correct	14.1
	Incorrect	13.6
Pythia 12B	Correct	13.5
	Incorrect	17.2

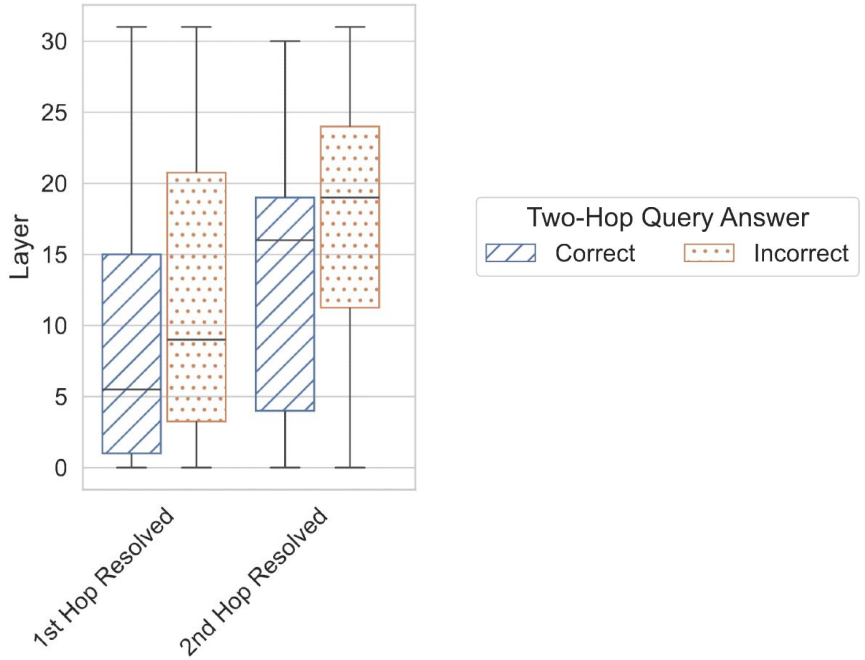
Sequential Nature of Computation

- The first hop is resolved into the bridge entity by the lower layers
- The second hop is resolved into the answer entity by the upper layers



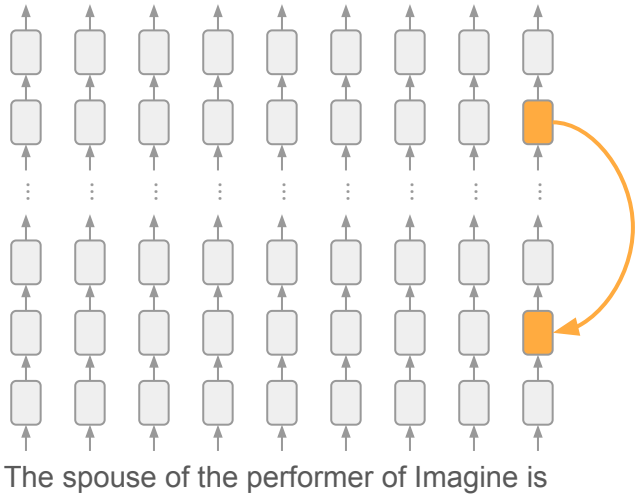
Two-Hop Failures

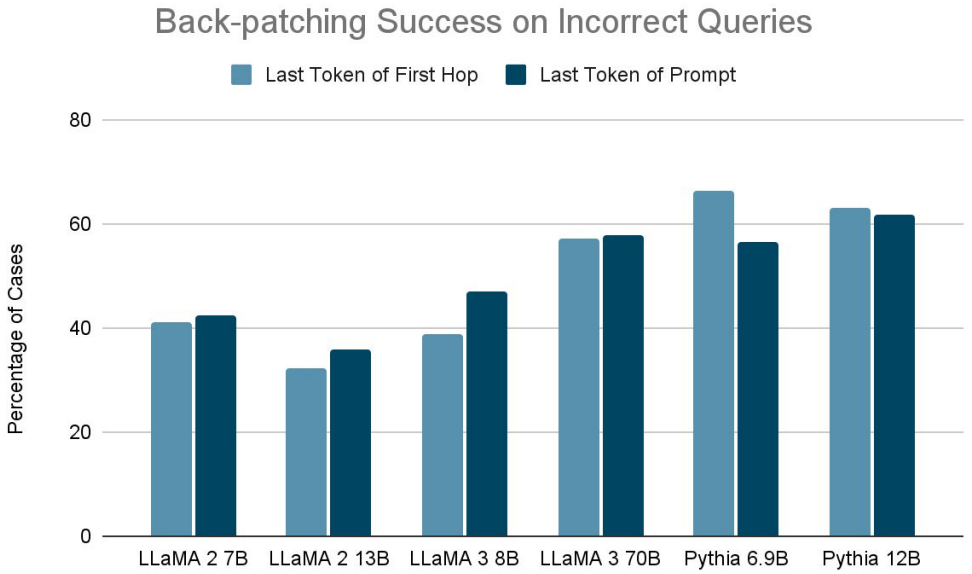
- Could failures stem from the first hop being resolved too late?
  - Less layers to complete the computation
  - Later layers might no longer contain the information needed to resolve the second hop



Back-patching Analysis Method

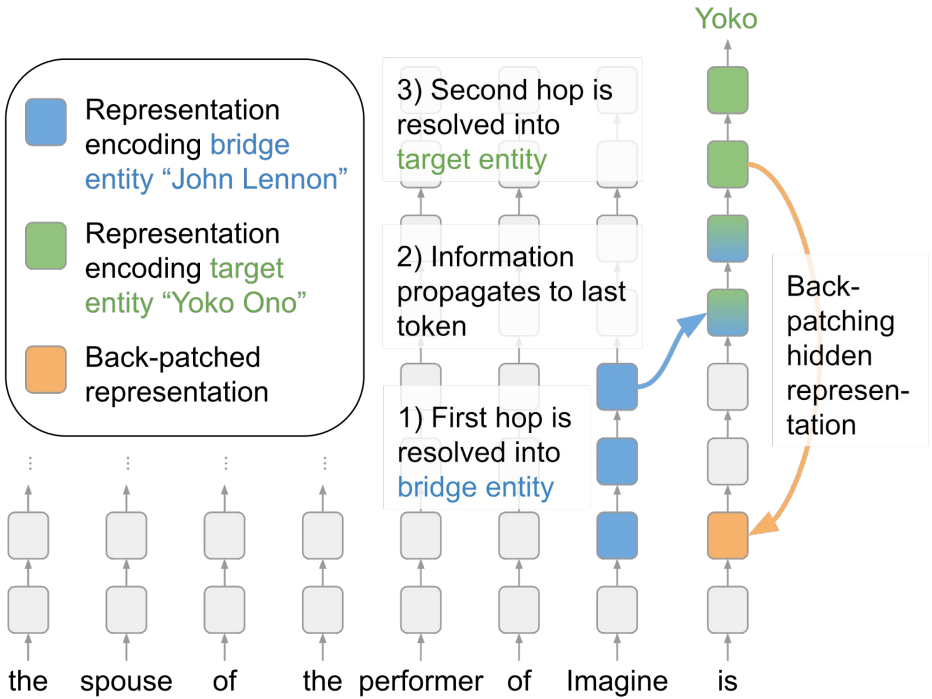
- Patch a hidden representation from a later layer back into an earlier layer





Summary

- We observe strong evidence of a sequential latent reasoning pathway when answering two-hop queries.
- This sequential nature could point to a possible limitation of transformers.
- We introduce and evaluate an analysis method named back-patching.



Plan

1. **Evidence of Existence:** Do LLMs perform latent multi-hop reasoning?
2. **Understanding the Mechanism:** How do LLMs perform latent multi-hop reasoning?
3. **Understanding the Current Status through Rigorous Benchmarking:** How much latent multi-hop reasoning ability have today's widely used LLMs naturally developed during pretraining?
4. **Conclusion**

We evaluate how well 41 LLMs perform
latent multi-hop reasoning without exploiting shortcuts

Do Large Language Models Perform Latent Multi-Hop Reasoning *without* Exploiting Shortcuts?

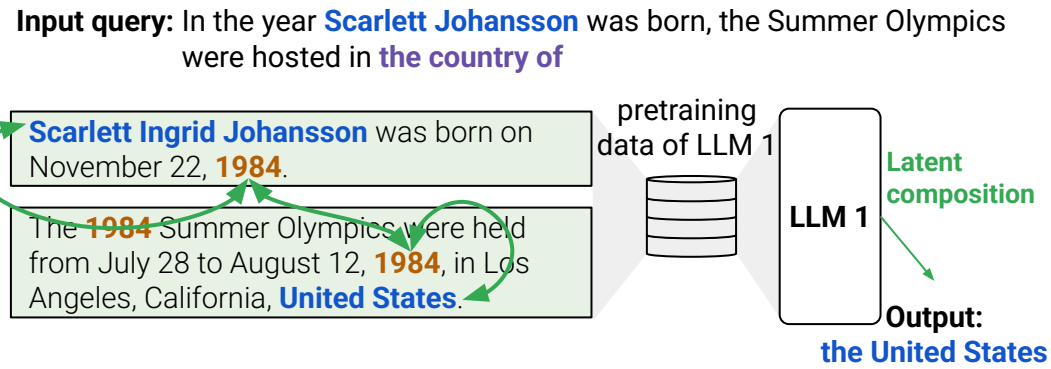
Sohee Yang[1,2] **Nora Kassner**[1] **Elena Gribovskaya**[1] **Sebastian Riedel**[1,2*] **Mor Geva**[3,4*]

[1]Google DeepMind [2]UCL [3]Google Research [4]Tel Aviv University

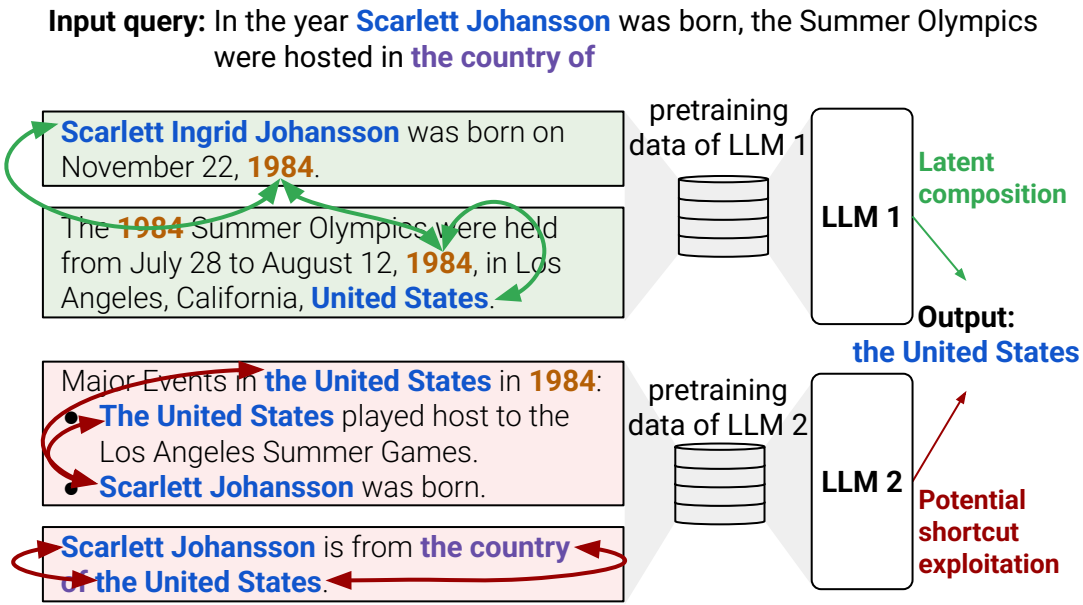




- In this work, we evaluate latent multi-hop reasoning abilities by assessing models' performance in answering multi-hop queries.



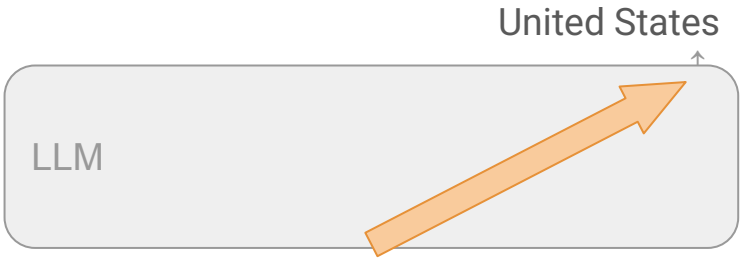
- In this work, we evaluate latent multi-hop reasoning abilities by assessing models' performance in answering multi-hop queries.
- However, what if models bypass true reasoning by exploiting shortcuts? We can't accurately measure the ability!



Shortcut-Free Evaluation of Latent Multi-Hop Reasoning

- The evaluation should exclude test queries that the model can answer using only partial information.
  - Excluding queries prone to subject-object shortcuts (where the model can predict the answer from just the head entity)

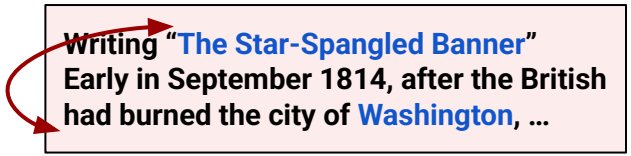
sequence seen during pretraining



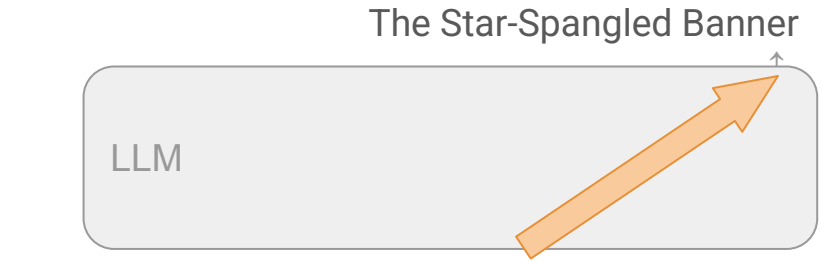
In the year Scarlett Johansson was born, the Summer Olympics were hosted in the country of

Shortcut-Free Evaluation of Latent Multi-Hop Reasoning

- The evaluation should exclude test queries that the model can answer using only partial information.
  - Excluding queries prone to subject-object shortcuts (where the model can predict the answer from just a substring of the head entity)

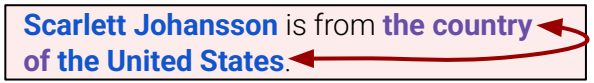


sequence seen during pretraining

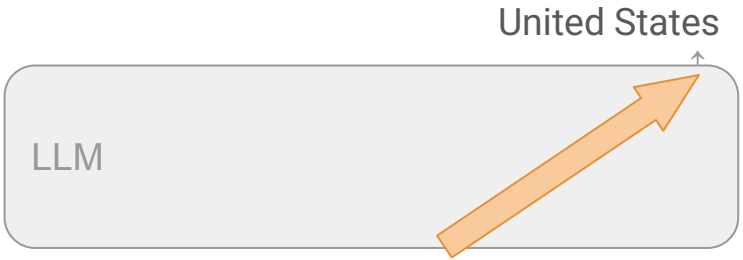


Shortcut-Free Evaluation of Latent Multi-Hop Reasoning

- The evaluation should exclude test queries that the model can answer using only partial information.
  - Excluding queries prone to relation-object shortcuts (where the model can predict the answer from just the relation pattern without the head entity)



sequence seen during pretraining



In the year Scarlett Johansson was born, the Summer Olympics were hosted in the country of

SOCRATES (ShOrtCut-fRee lATent rEaSoning) Dataset

- 7K multi-hop and corresponding single-hop queries where head and answer entities have minimal chance of co-occurring in training data, which is carefully curated for shortcut-free evaluation of latent multi-hop reasoning

e_2 type	relation composition type	count	example multi-hop query
city	person-birthcity-eventyear	33	e_1's birth city hosted the Eurovision Song Contest in the year
country	university-locationcountry-anthem	101	The country where e_1 is in has the national anthem named
	university-locationcountry-isocode	30	The ISO 3166-1 numeric code of the country where e_1 is located is
	university-locationcountry-year	7	The founding year of the location country of e_1 is
	person-birthcountry-anthem	22	The name of the national anthem of e_1's country of birth is
	person-birthcountry-isocode	6	The ISO 3166-1 numeric code used for the country where e_1 was born is
university	person-undergraduniversity-founder	33	The person who founded e_1's undergrad university is named
	person-undergraduniversity-year	25	The year when the university where e_1 studied as an undergrad was founded is
year	person-birthyear-winner	4,484	The winner of the Nobel Prize in Literature in e_1's birth year was
	city-eventyear-winner	2	In the year when the G7 Summit were hosted in e_1, the Nobel Prize in Chemistry was awarded to
	university-inceptionyear-winner	632	In the founding year of e_1, the Nobel Prize in Literature was awarded to
	university-inceptionyear-hostleader	9	The person who was the host leader of the G7 Summit in the founding year of e_1 is
	university-inceptionyear-eventcountry	13	In the year e_1 was founded, the host country of the G7 Summit was
	university-inceptionyear-eventcity	62	In the founding year of e_1, the host city of the G7 Summit was
	person-birthyear-eventcity	1,389	In the birth year of e_1, the Winter Olympics were hosted in the city of
	person-birthyear-hostleader	260	The leader of the host of the G7 Summit in e_1's birth year is
	person-birthyear-eventcountry	124	The country where the Eurovision Song Contest took place in the birth year of e_1 is
		7,232	

Dataset Construction

- Ideally: Exclude cases where any of the possible combinations of the aliases of the **head entity and answer entity appears at least once** in the same sequence that the evaluated LLM has seen during training.

Dataset Construction

- Ideally: Exclude cases where any of the possible combinations of the aliases of the **head entity and answer entity appears at least once** in the same sequence that the evaluated LLM has seen during training.

Main Challenge: we do not have access to the **pretraining sequences/corpora** of most of the models we want to evaluate







Dataset Construction

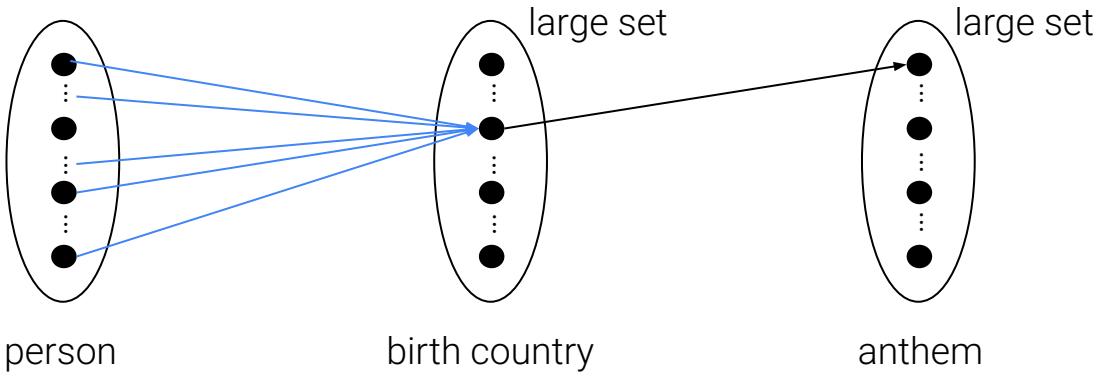
- Solution: Exclude cases where any of the possible combinations of the aliases of the head entity and answer entity appears at least once in the **same document of a proxy corpus (combination of 6 training corpora, ~4.8 unique documents)**

pretraining corpus/search engine	number of documents	LLM trained with the corpus
Dolma v1.5	4,367M	OLMo
Dolma v1.7	2,532M	OLMo 0724
OSCAR	431,584K	BLOOM
C4	364,869K	T5
OpenWebText	8,006K	GPT-2
Tulu-v2-sft-mixture	326K	OLMo Instruct
Google Search*	~400B	

*Results remain similar with web-scale co-occurrence checks with Google Search

Dataset Construction

- We select single-hop facts that are likely well-known, but their composition is unlikely to naturally appear in general text corpora
- Such cases typically occur when the set of possible options for the bridge entity is large, there are numerous head entities that map to the same bridge entity, and the set of possible options for answer entity is not too small (e.g., no "blood type" as the answer entity)



Dataset Construction

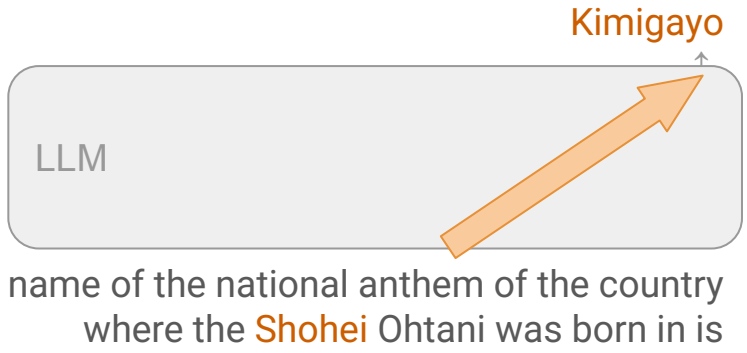
- Exclude facts where head/answer entities are directly connected

relation composition	heuristic conditions to meet
person-birthcity-year	city $\neq$ capital of a country
	person's birth year $\neq$ event year
person-undergraduniversity-year	person's birth year $\neq$ university's inception year
person-birthyear-winner	person's birth/citizenship country $\neq$ winner's birth/citizenship country
person-birthyear-event country/city/leader	person's birth/citizenship country $\neq$ event country
university-inceptionyear-winner	university's location country $\neq$ winner birth/citizenship country
	university $\neq$ winner's university for any degree
city-eventyear-winner	event country $\neq$ winner's birth/citizenship country
university-inceptionyear-event country/city/leader	university's location country $\neq$ event country

In the year Scarlett Johansson was born, the Summer Olympics were hosted in the country of United States
Scarlett Johansson's birth country = United States = 1984 Summer Olympics country → exclude from dataset

Dataset Construction

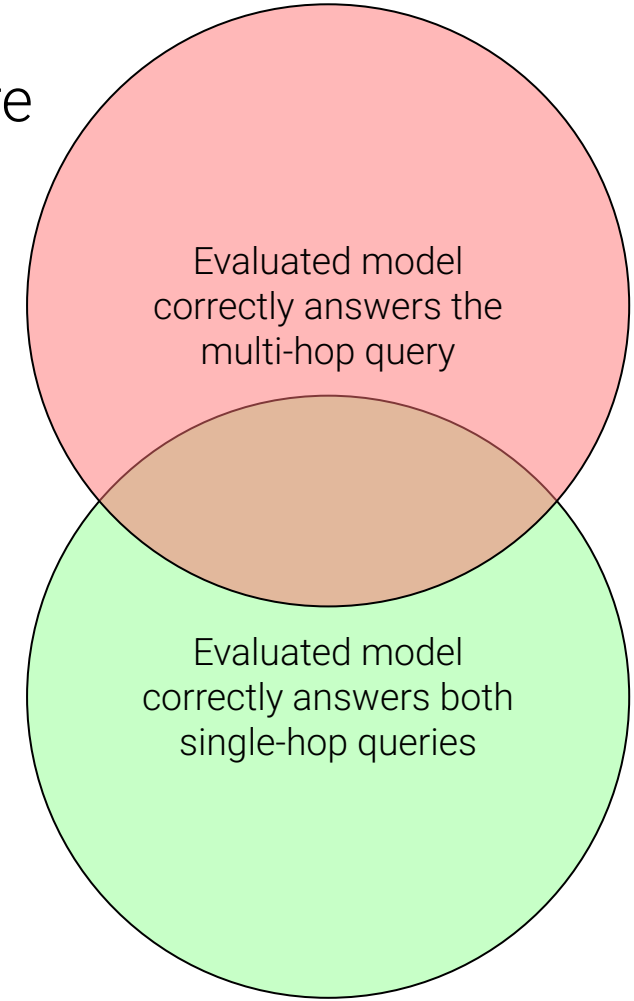
- Exclude facts guessable from part of the head entity



Make Claude 3 Haiku and GPT 3.5 Turbo guess the answer based on a substring of the head entity, and exclude the query from the dataset if any of these two models correctly answer the question

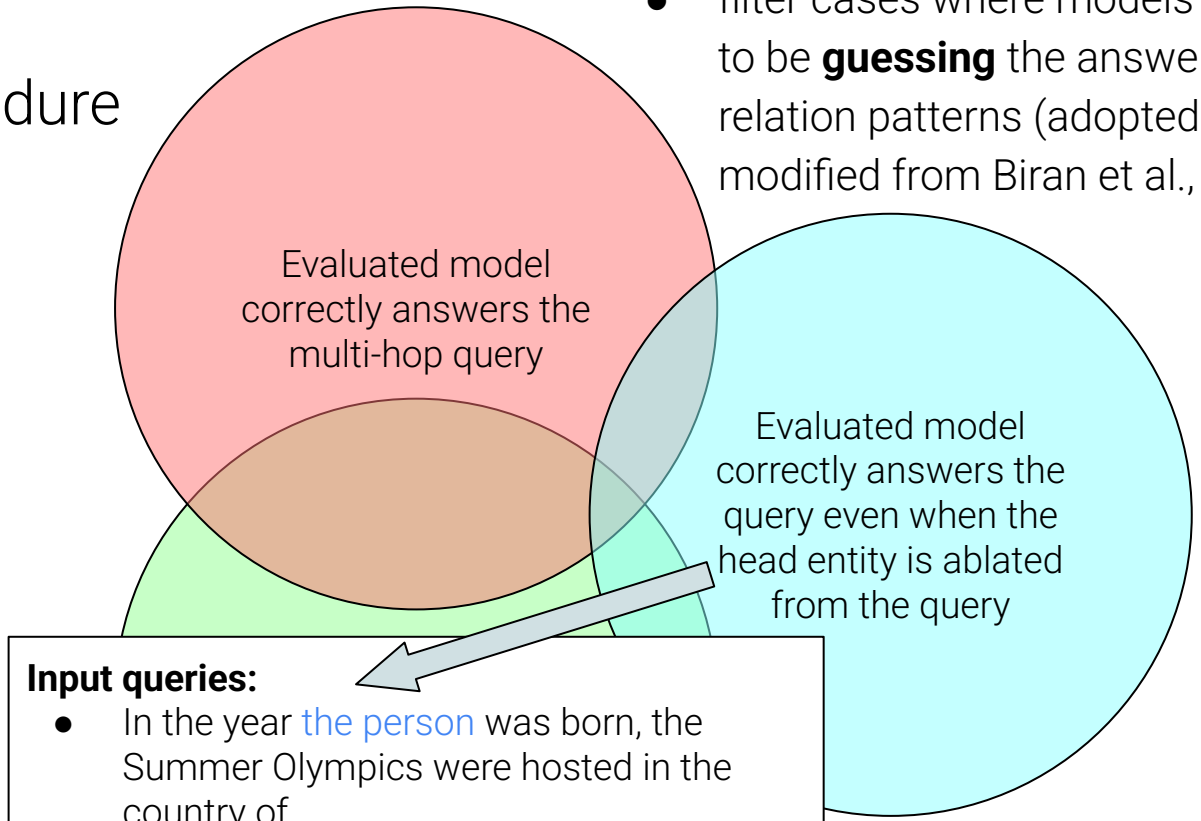
"Guessing from the name, what are the candidates of the country where "Shohei Ohtani" was likely to be born? To be more specific, what are the candidates of the country where someone with the first name "Shohei" was likely to be born? Likely, what are the candidates of the country where someone with the last name "Ohtani" was likely to be born? Make sure to list the names of the countries guessed solely from the person name."

Evaluation Procedure



Evaluation Procedure

- filter cases where models are likely to be **guessing** the answer from relation patterns (adopted and modified from Biran et al., 2024)



Input queries:

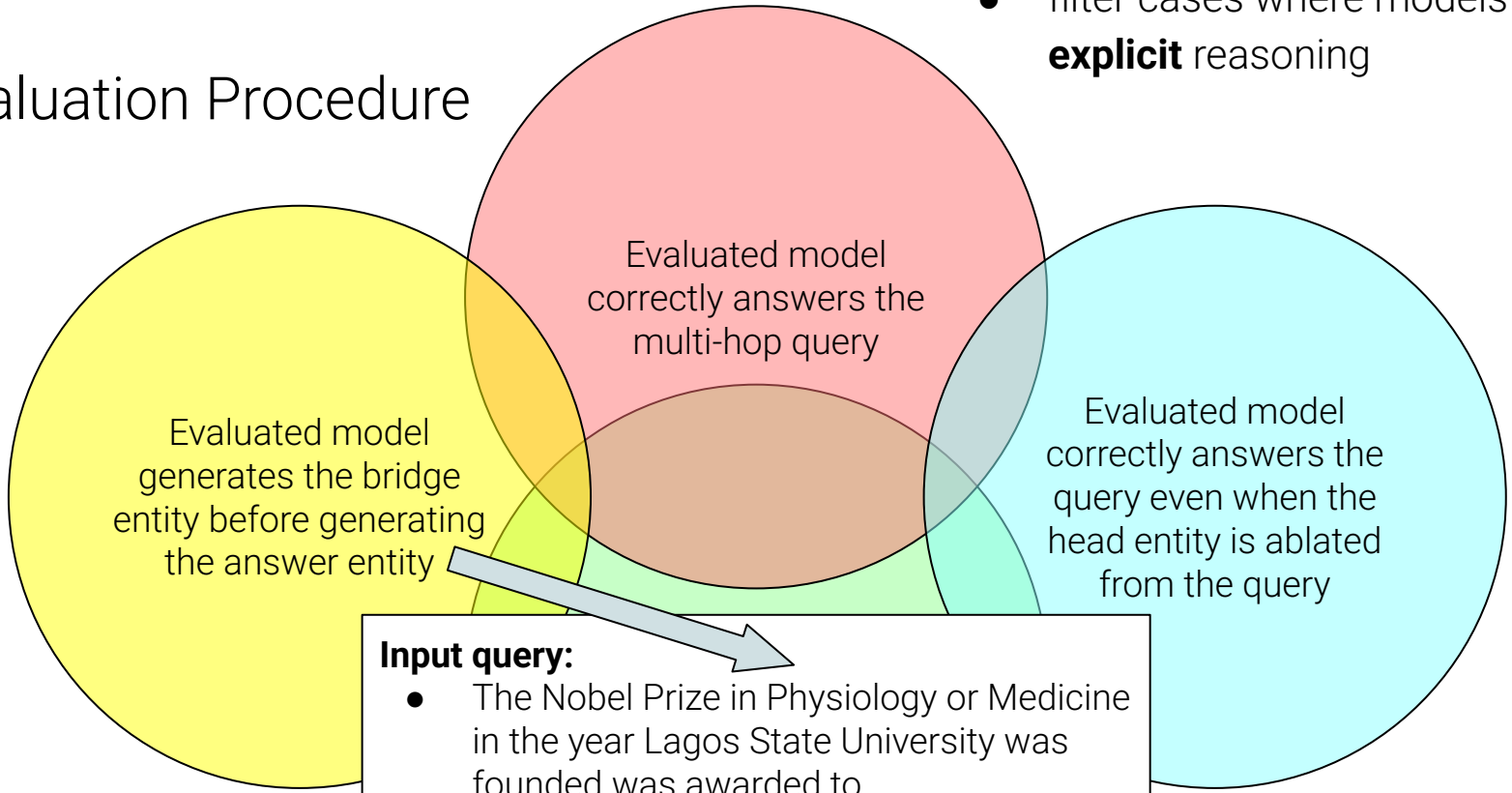
- In the year the person was born, the Summer Olympics were hosted in the country of
- In the year, the Summer Olympics were hosted in the country of

Output to either query:

- United States

Evaluation Procedure

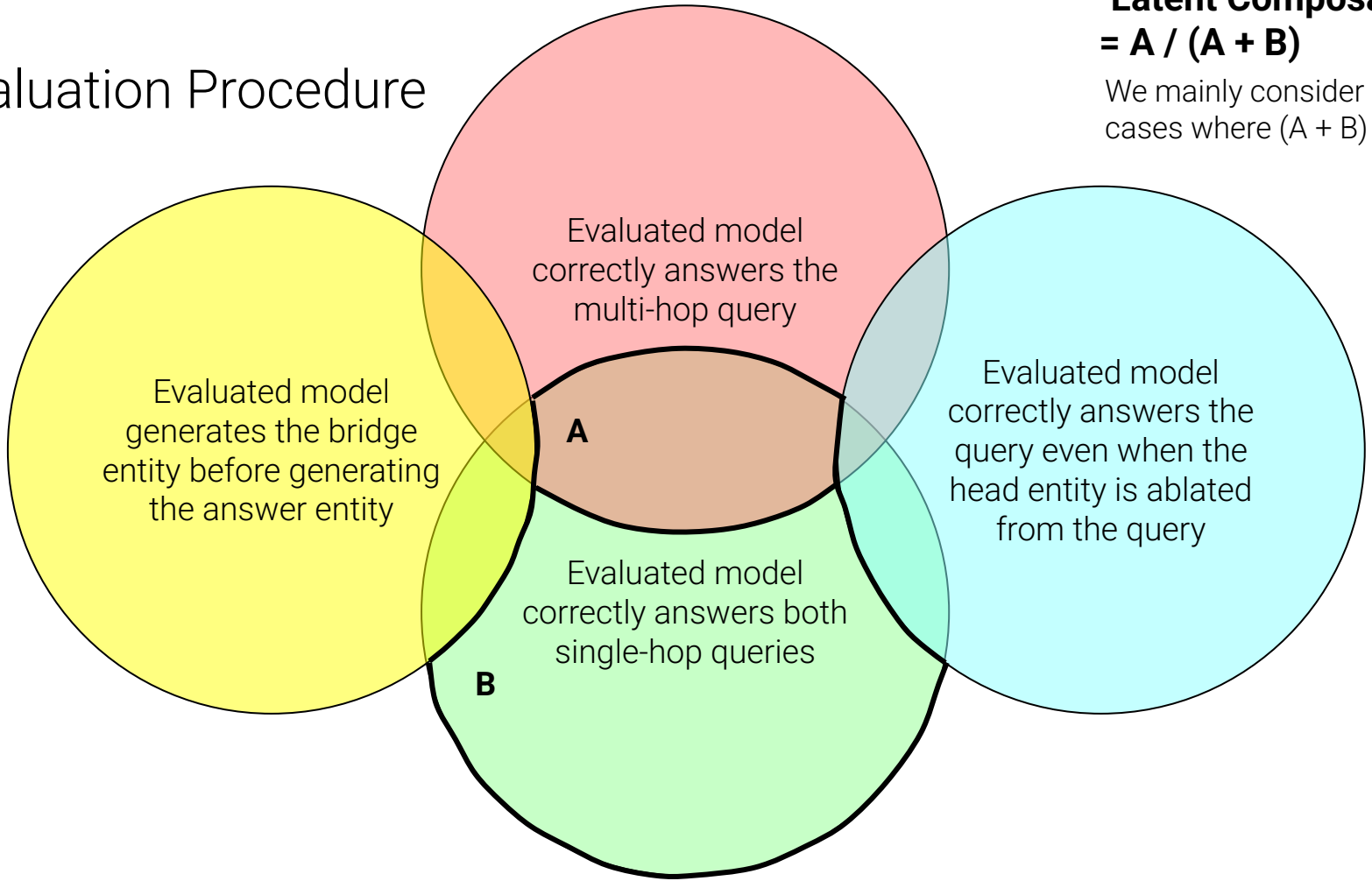
- filter cases where models perform **explicit** reasoning



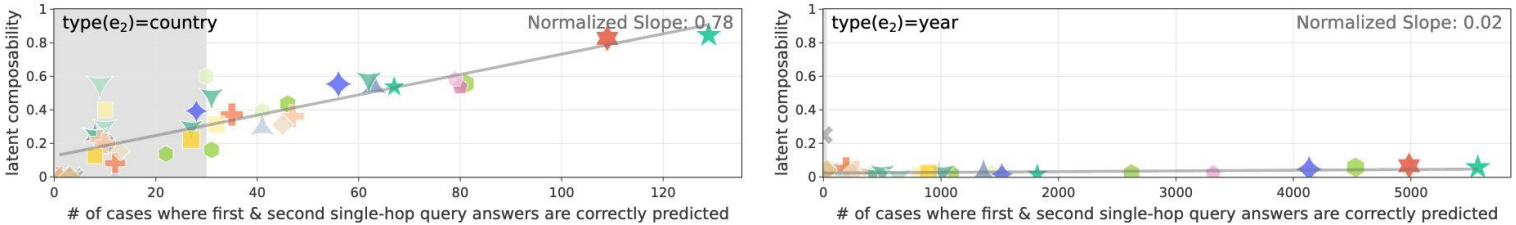
Evaluation Procedure

"Latent Composability"
= A / (A + B)

We mainly consider only the cases where (A + B) >= 30



Evaluation Results

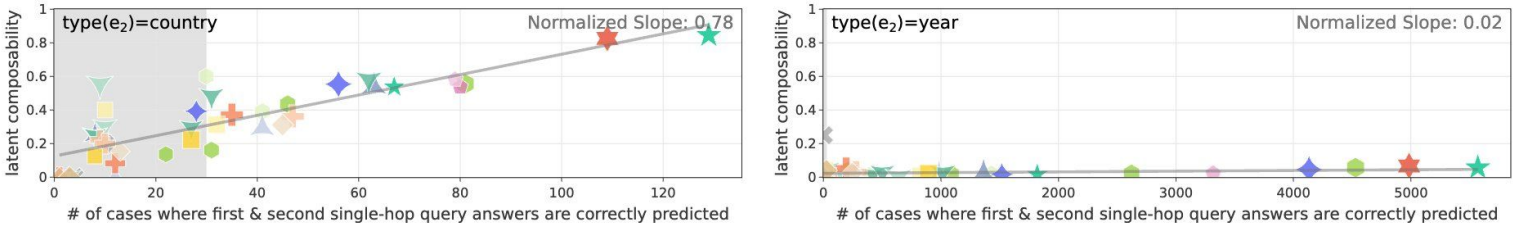


(a) Latent composability on country-type **(left)** and year-type **(right)** bridge entity subsets. Latent composability varies according to the bridge entity type; it is over 80% for the best models when the bridge entity is a country, but it is around 5% when it is a year.

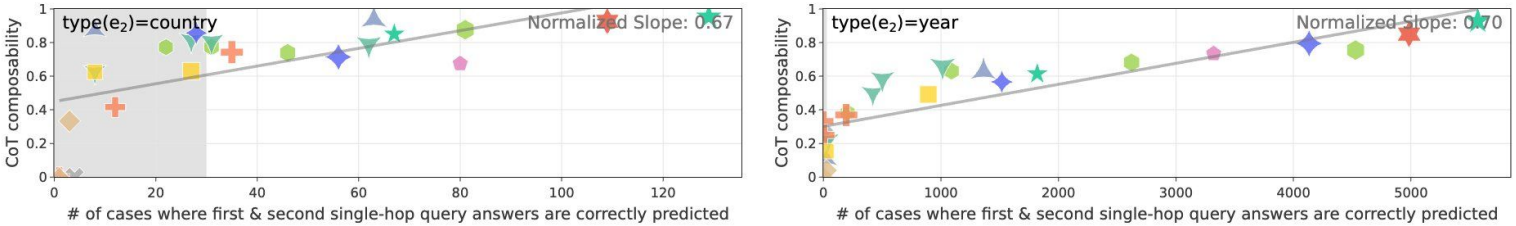
- Results reveal striking differences across bridge entity types
  - 80%+ accuracy with countries, ~6% with years

Evaluation Results

Instruction for CoT:
Fill in the blank. First, write the step-by-step explanation necessary to get the solution with the prefix "EXPLANATION:". After that, write down the final answer with the prefix "ANSWER:". For the final answer, write down only what goes in the blank. The answer can consist of multiple words.



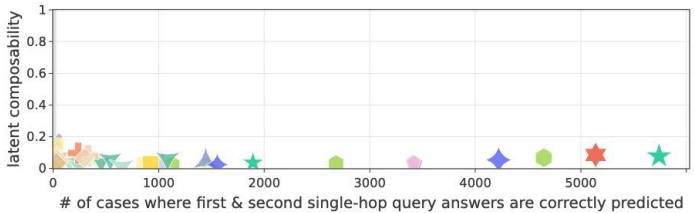
(a) Latent composability on country-type **(left)** and year-type **(right)** bridge entity subsets. Latent composability varies according to the bridge entity type; it is over 80% for the best models when the bridge entity is a country, but it is around 5% when it is a year.



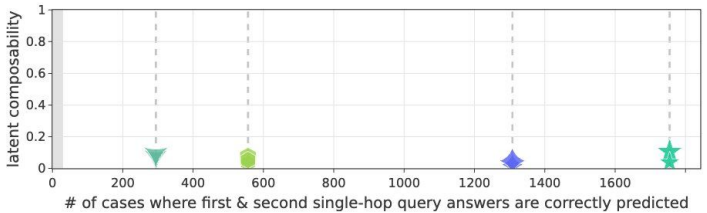
(b) CoT composability on country-type **(left)** and year-type **(right)** bridge entity subsets. CoT composability does not fluctuate as dramatically as latent composability according to the type of bridge entity.

- This variation vanishes with Chain-of-Thought (CoT) reasoning, suggesting different internal mechanisms.

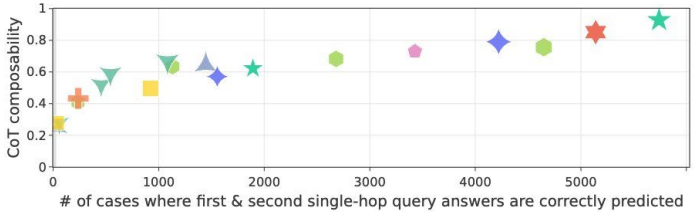
Evaluation Results



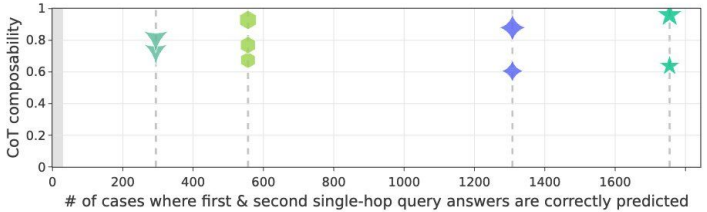
(a) Latent composability. There are successful cases of latent multi-hop reasoning, although the overall percentage is low.



(b) Effect of model scale on latent composability. Comparative latent composability slightly improves with model scale.



(c) CoT composability. CoT composability is significantly higher than latent composability.

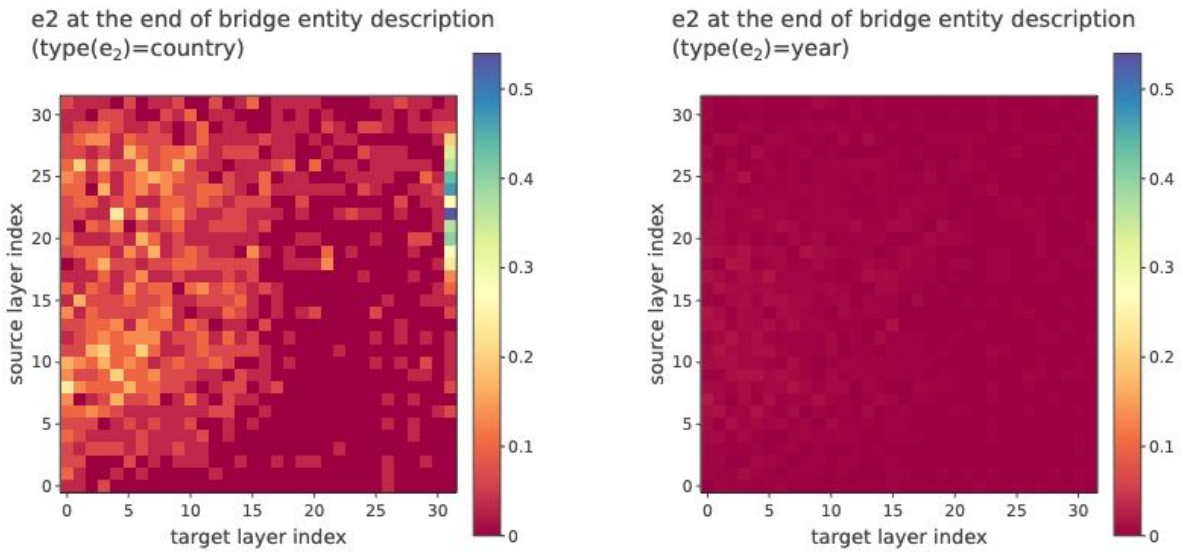


(d) Effect of model scale on CoT composability. Comparative CoT composability increases much more effectively with model scale compared to latent composability.

Models that know more single-hop facts and larger models show only marginal improvements for latent reasoning, but dramatic improvements for CoT reasoning.

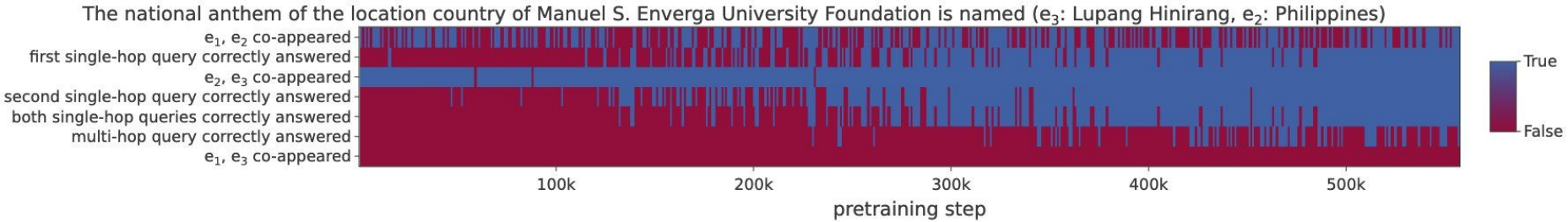
target query: StarCraft: StarCraft is […], Leonardo DiCaprio: Leonardo DiCaprio is […], Samsung: Samsung is […], x
Sample: 1 (greedy decoding)

Additional Analysis: Inner Mechanism



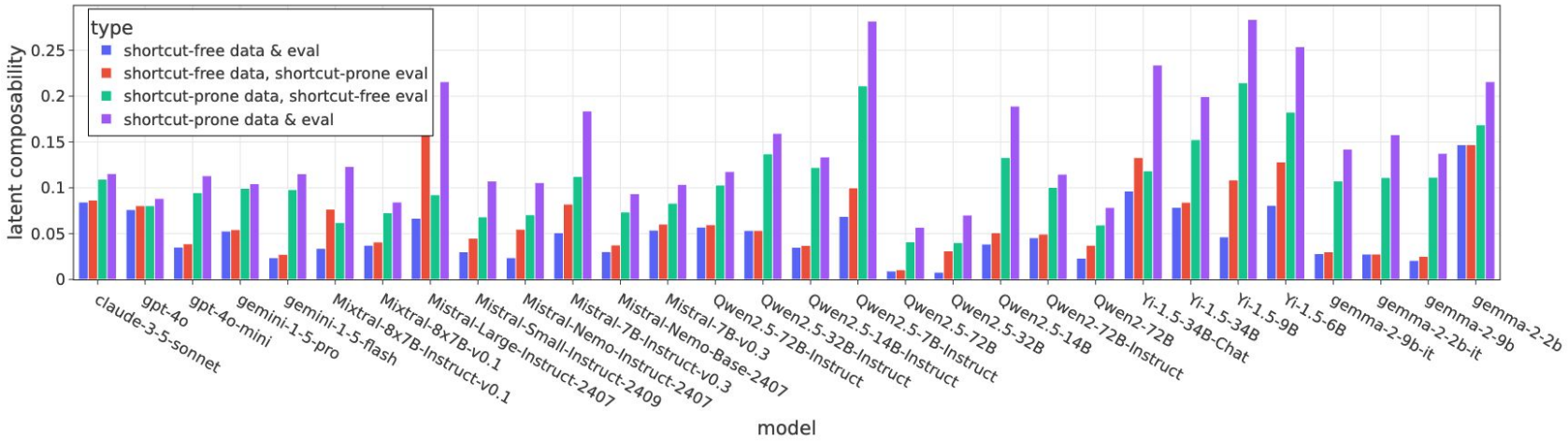
- Using Patchscopes analysis, we discover that bridge entity representations are constructed more clearly in queries with higher latent composability.
- This helps explain the internal mechanism behind why some types of connections are easier for models to reason about.

Additional Analysis: Emergence of Latent Multi-Hop Reasoning



With OLMo's pretraining checkpoints grounded to entity co-occurrences in the training sequences, we observe the emergence of latent reasoning: the model tends to first learn to answer single-hop queries correctly, then develop the ability to compose them.

Additional Analysis: Importance of Shortcut-Free Dataset and Evaluation



- When we compare with shortcut-prone dataset and evaluation, we find that not accounting for shortcuts can overestimate latent composability by up to 5-6x.
- This highlights the importance of careful evaluation dataset and procedure that minimizes the chance of shortcuts.

Why does latent composability differ dramatically according to the type of the bridge entity?

Unlikely Hypotheses

- The high latent composability of the country-type bridge entity cases does not come from the model simplifying the multi-hop query into a single-hop-like problem by easily guessing the first hop
  - Our careful dataset construction, which selects only cases where countries cannot be readily inferred, accounts for easy guessing of countries from entity names.
- It's unlikely to stem from insufficient filtering of co-occurrences.
  - Adding a Google Search filter to exclude cases where entities appear together in search results does not drop latent composability, with an average relative drop of only 0.03.

Why does latent composability differ dramatically according to the type of the bridge entity?

Plausible Hypothesis: Country-related facts might be more frequently learned in composition during pretraining

- Supporting previous works
  - Zhengbao Jiang et al., Understanding and Improving Zero-shot Multi-hop Reasoning in Generative Question Answering, COLING 2022
    - "... we find that models lack zero-shot multi-hop reasoning ability: when trained only on single-hop questions, models generalize poorly to multi-hop questions. … we demonstrate that it is possible to improve models' zero-shot multi-hop reasoning capacity through two methods that approximate real multi-hop natural language (NL) questions by training on either concatenation of single-hop questions or logical forms (SPARQL)."

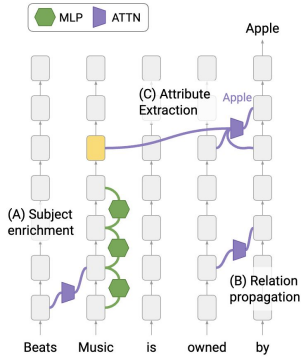
Why does latent composability differ dramatically according to the type of the bridge entity?

Plausible Hypothesis: Country-related facts might be more frequently learned in composition during pretraining

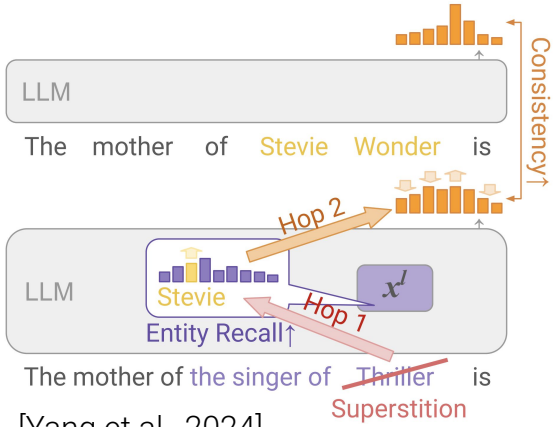
- Supporting previous works
  - Wang et al., Grokked Transformers are Implicit Reasoners: A Mechanistic Journey to the Edge of Generalization, NeurIPS 2024
    - "We find that the speed of improvement in generalization correlates with the ratio between inferred and atomic facts in training, and depends little on the absolute size of the training data"
    - "While ID generalization is consistently observed, in the OOD setting, the model fails to systematically generalize for composition"
      - OOD setting: the cases where the model has not seen the single-hop facts in the form composed with other facts (= Likely to be year-related facts)

Why are LLMs good at CoT but not on latent reasoning?

- Conjecture: the explicit generation of the bridge entity is the main factor behind high CoT composability.
  - Transformer-based LLMs go through a subject enrichment process that helps models recall the attributes of the subject.
  - While LLMs can develop latent representations of bridge entities in early-middle layers, these representations may appear too late or not emerge at all.
  - In contrast, when CoT reasoning generates the correct bridge entity, it ensures a clear and early contextualized representation of the bridge entity to form, facilitating retrieval of the second single-hop fact.



[Geva et al., 2023]



[Yang et al., 2024]

Even when instructed to think step-by-step, without explicit generation of the intermediate results, the instruction does not improve performance

Fill in the blank. Write down only what goes in the blank. Think step-by-step, but do it only internally and do not explain it in the answer. The answer can consist of multiple words.

When is Nicole Azzopardi's birth year? Use the information.

In Nicole Azzopardi's birth year, the G7 Summit was hosted in the city of ___

	latent composability of claude-3-5-sonnet
Original instruction + prompt	0.0844
Internal CoT instruction + prompt	0.0607

Summary

- We introduce the dataset and evaluation procedure of latent multi-hop reasoning that minimize the risk of models exploiting shortcuts and bypassing the true latent reasoning process.
- Findings are as follows:
    - latent composability in LLMs significantly varies according to the bridge entity type, reaching >80% for queries with country-type bridge entities but ~6% for year-type ones
    - latent reasoning marginally improves with the number of known single-hop facts and model scale and identify a significant gap between latent and CoT composability.
    - additional analysis helps better understand LLMs' mechanisms for latent multi-hop reasoning, such as different construction patterns of latent bridge entity representation and the emergence of the ability during pretraining.

Plan

1. **Evidence of Existence:** Do LLMs perform latent multi-hop reasoning?
2. **Understanding the Mechanism:** How do LLMs perform latent multi-hop reasoning?
3. **Understanding the Current Status through Rigorous Benchmarking:** How much latent multi-hop reasoning ability have today's widely used LLMs naturally developed during pretraining?
4. **Conclusion**

Conclusion

- LLMs can develop latent multi-hop reasoning ability during pretraining. However, the ability of even today's best models is not robust and significantly falls behind explicit reasoning ability.
- Considering the current findings and the previous works, training the models on various connected facts is conjectured to increase the latent composability on the facts with the same relation composition. However, this still doesn't give the model the general compositional reasoning ability to compose facts of unseen relations.
    - How can we make models be better at compositional reasoning in general? How can we add such inductive bias to the model? How can we give the model the initiatives?
- To achieve genuine compositional generalization, we might need to change the training data, training loss, and architecture
    - to demotivate LLMs from redundantly storing information and nudge the models to utilize already learned parametric knowledge when it is possible

Thank You!

Sohee Yang

DLCT, Nov 2025