Do Large Language Models Perform Latent Reasoning?

Mor Geva

TAU NLP

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Reasoning has long been a hallmark of Artificial Intelligence

II

The Logic Theory Machine

In the language we have constructe
(atomic sentences): p, q, r, A, B, C,
v (or), → (implies). The connectives
variables into expressions (molecular
already considered one example of an e
1.7 -p .→. q v
The task set for LT will be to prc
are theorems — that is, that they can
of specified rules of inference from a
or axioms.

PROGRAMS WITH COMMON SENSE

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

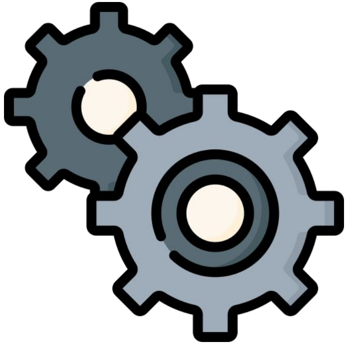
1959

Newell and Simon, 1956

The ability to reason over multiple pieces of information

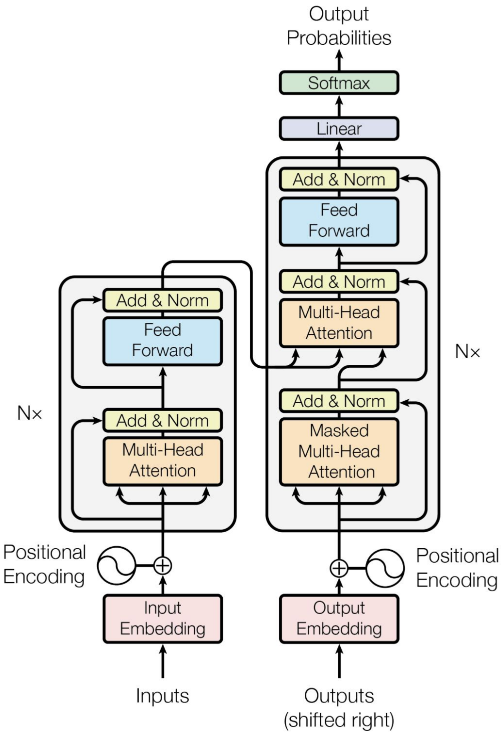
The **singer of Superstition** is **Stevie Wonder**

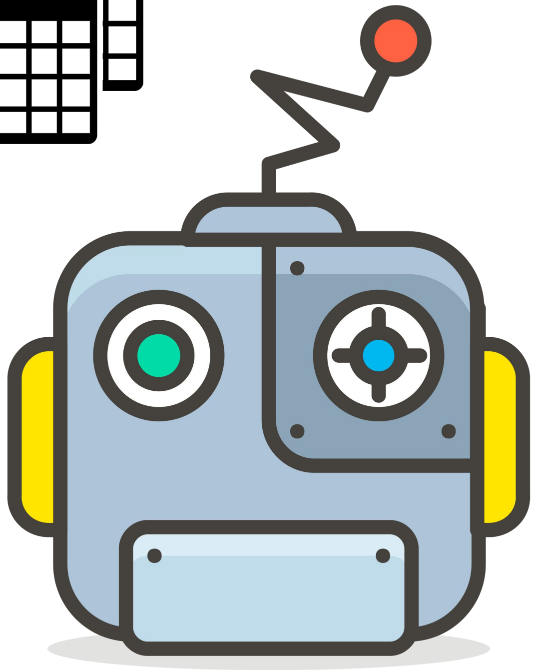
The mother of **Stevie Wonder** was **Lula Mae Hardaway**

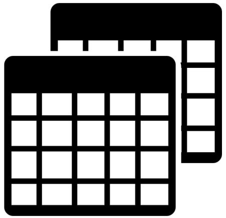


The mother of the **singer of Superstition** was **Lula Mae Hardaway**

Fast-forward 60 years…







Vaswani et al. 2017

LLMs do pretty well on reasoning tasks*

Who is the mother of the singer of Superstition?

LLMs do pretty well on reasoning tasks*

Who is the mother of the singer of Superstition?

The singer of "Superstition" is Stevie Wonder. His mother was Lula Mae Hardaway. Lula Mae played a significant role in Stevie Wonder's life and career, co-writing some of his early songs, including "I Was Made to Love Her." She raised Stevie in Detroit and supported his musical talents from a young age.

* when the necessary information is provided in-context

When reasoning should be performed latently, performance decreases substantially

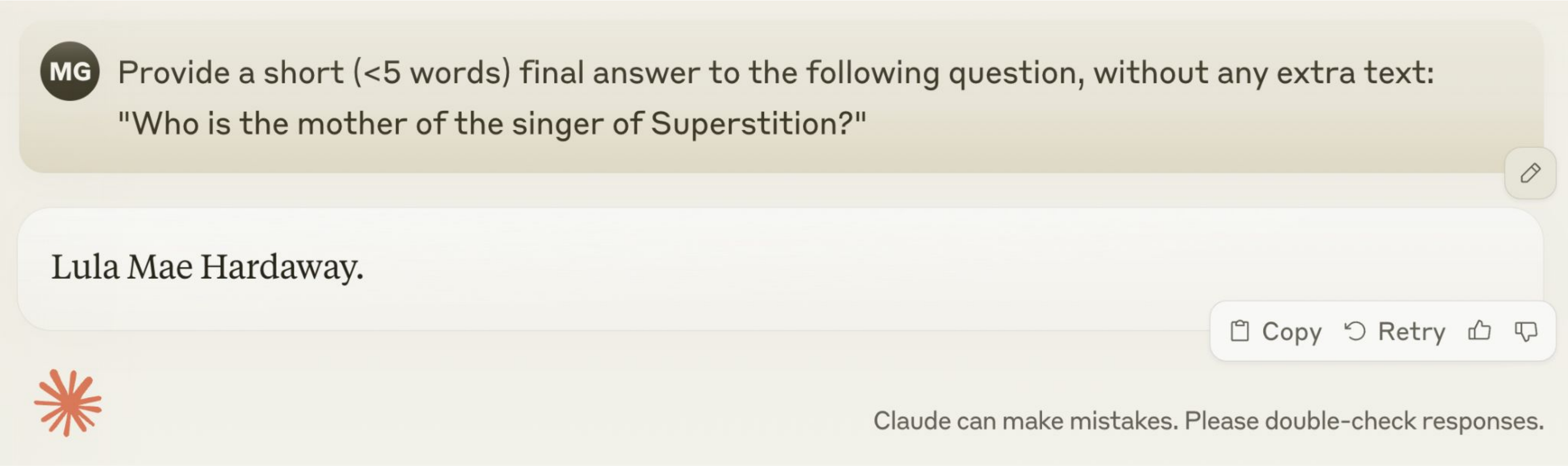
Provide a short (<5 words) final answer to the following question, without any extra text: "Who is the mother of the singer of Superstition?"

When reasoning should be performed latently, performance decreases substantially

Provide a short (<5 words) final answer to the following question, without any extra text: "Who is the mother of the singer of Superstition?"

Mary Lee Hawkins ❌

But there are still success cases



Are models capable of latent reasoning?

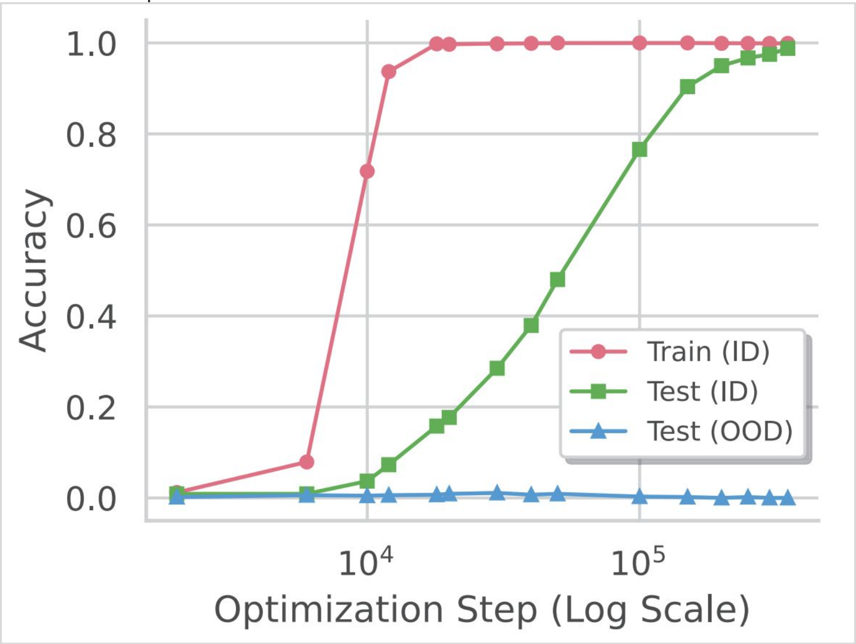
How do they solve such tasks?

The Transformer architecture enables deductive reasoning

(Input Facts:) Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

(Input Rules:) Big people are rough.
If someone is young and round then they are kind.
If someone is round and big then they are blue.
All rough people are green.

Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**



Clark et al. 2020, Wang et al. 2024

What about **large** language models trained on “**real**” data?

Plan

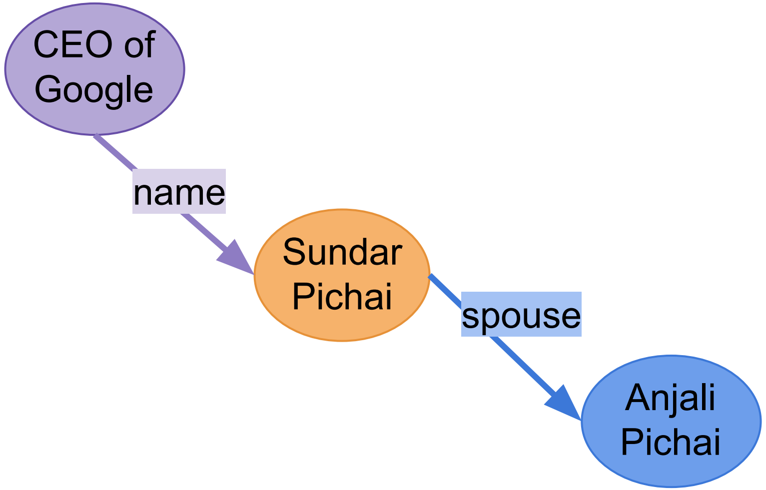
(1) Existential evidence of latent reasoning in LLaMA 2

(2) Exploring the limitations of latent reasoning in LLMs

Problem setup

Prompt LLMs with two-hop queries like:

The spouse of the CEO of Google is

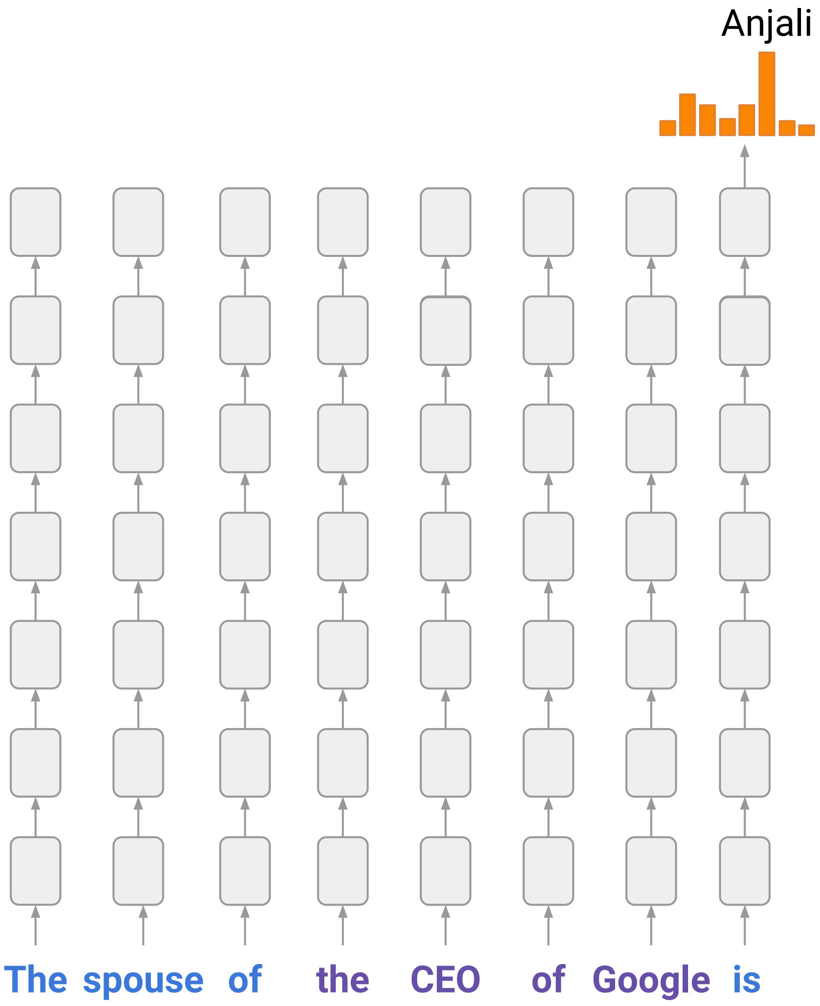


Problem setup

Possible ways to resolve the answer:

- “Backwards”
- Strong correlation between “the CEO of Google” and “Anjali”
- Other information about Google that connects it to Anjali

How do models solve this?



Existential evidence of latent reasoning in LLaMA 2

Sohee Yang

Elena Gribovskaya

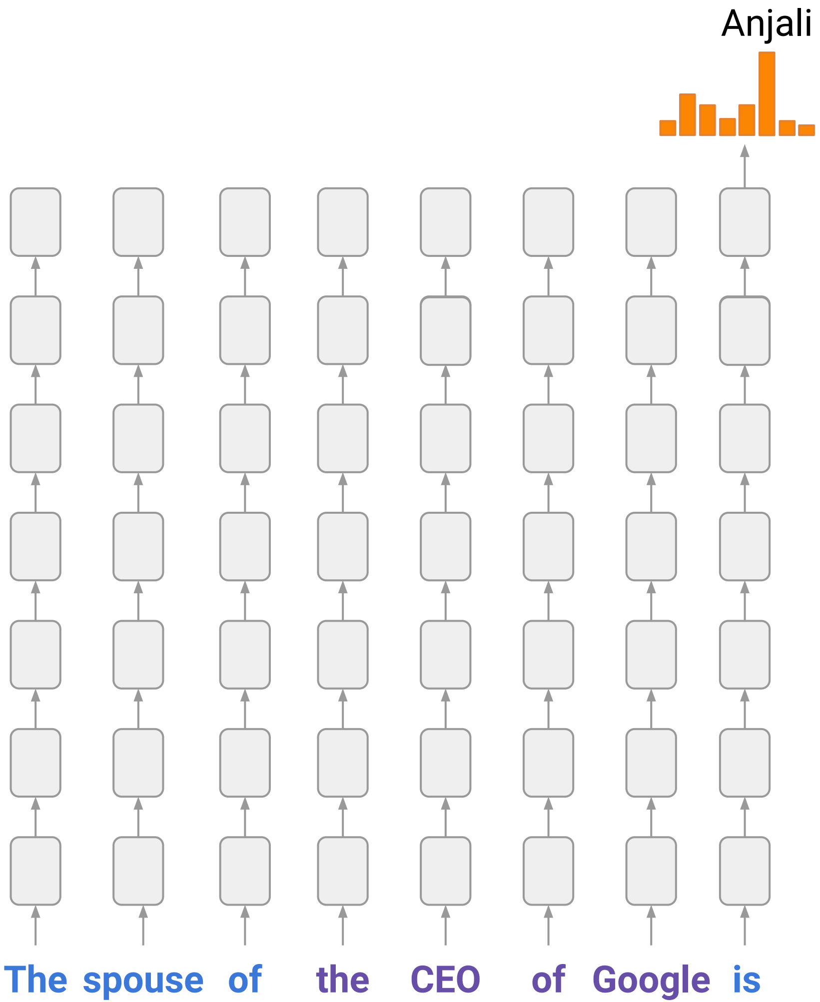
Nora Kassner

Sebastian Riedel

How do models solve this?

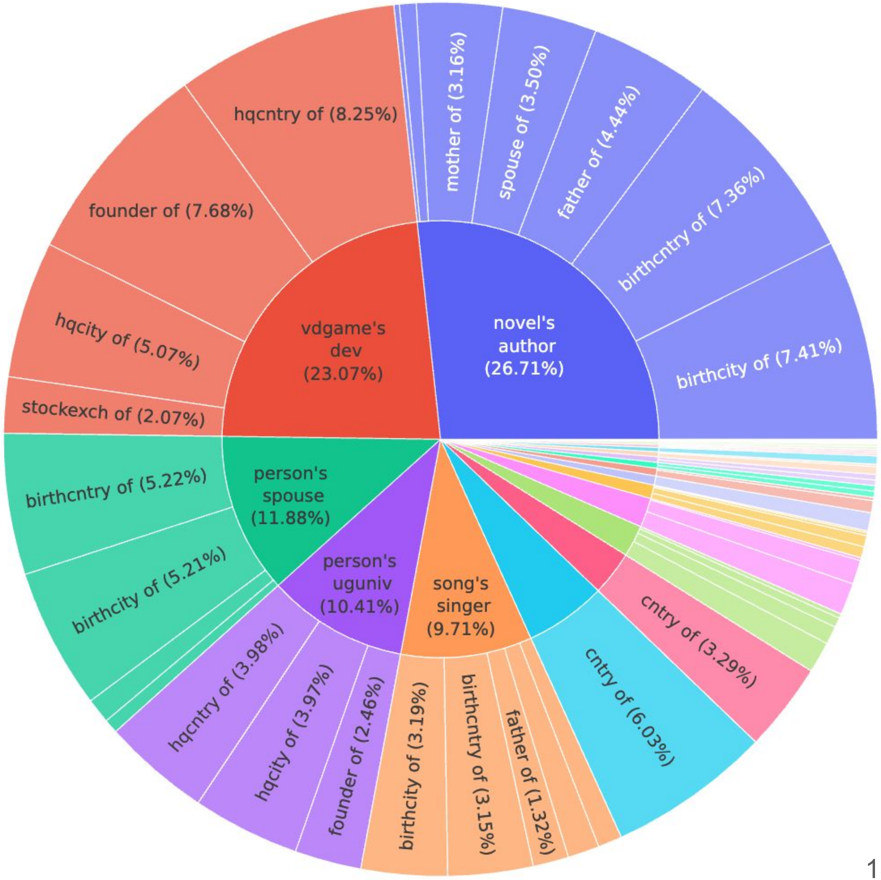
Q1 Does the model resolve the **first hop** when processing the two-hop query?

Q2 Does the model utilize the **first hop** for answering the **second hop**?



Experimental setting

- **Data**: A large-scale dataset of 45,595 two-hop queries, covering 52 fact composition types.
- **Models**: LLaMA 2 7B, 13B, 70B
- Analyze the cases where the model predicts each of the hops correctly.



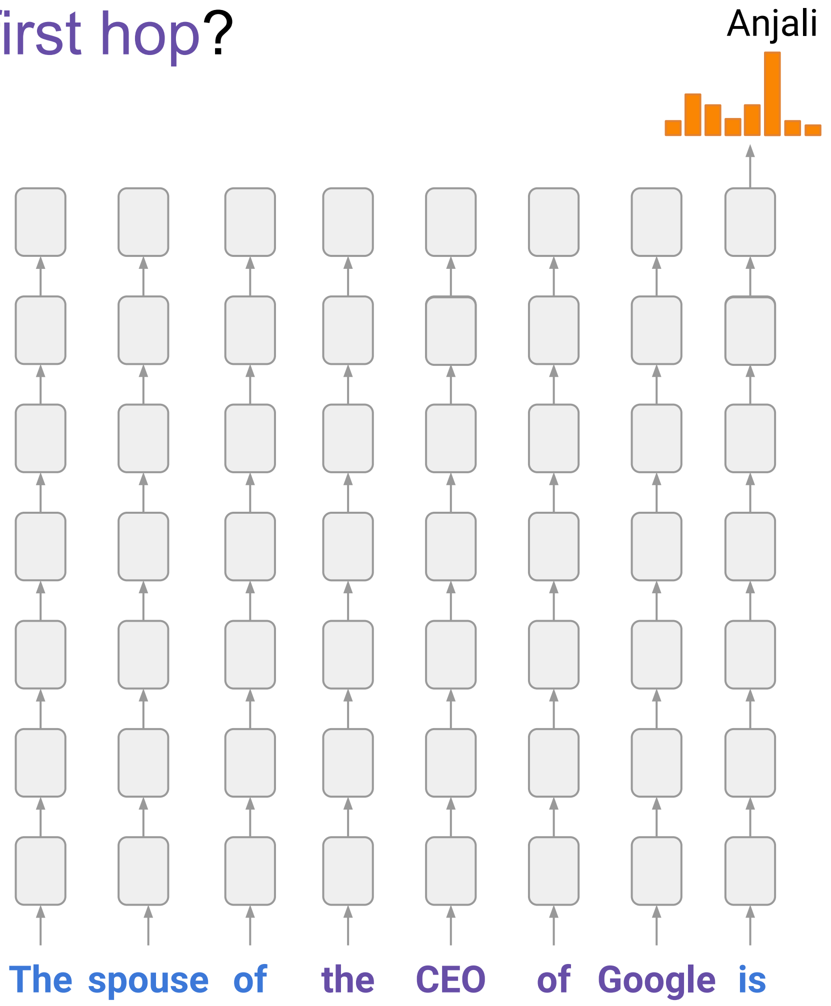
High-level approach

- Internal entity recall score that measures resolution of the first-hop
- Consistency score that measures utilization of the first-hop
- Check if increasing entity recall also increases first-hop utilization.

  A positive answer would be an indication for a second-hop presence!

Q1 Does the model resolve the first hop?

Estimate the degree of entity recall via projection to the vocabulary



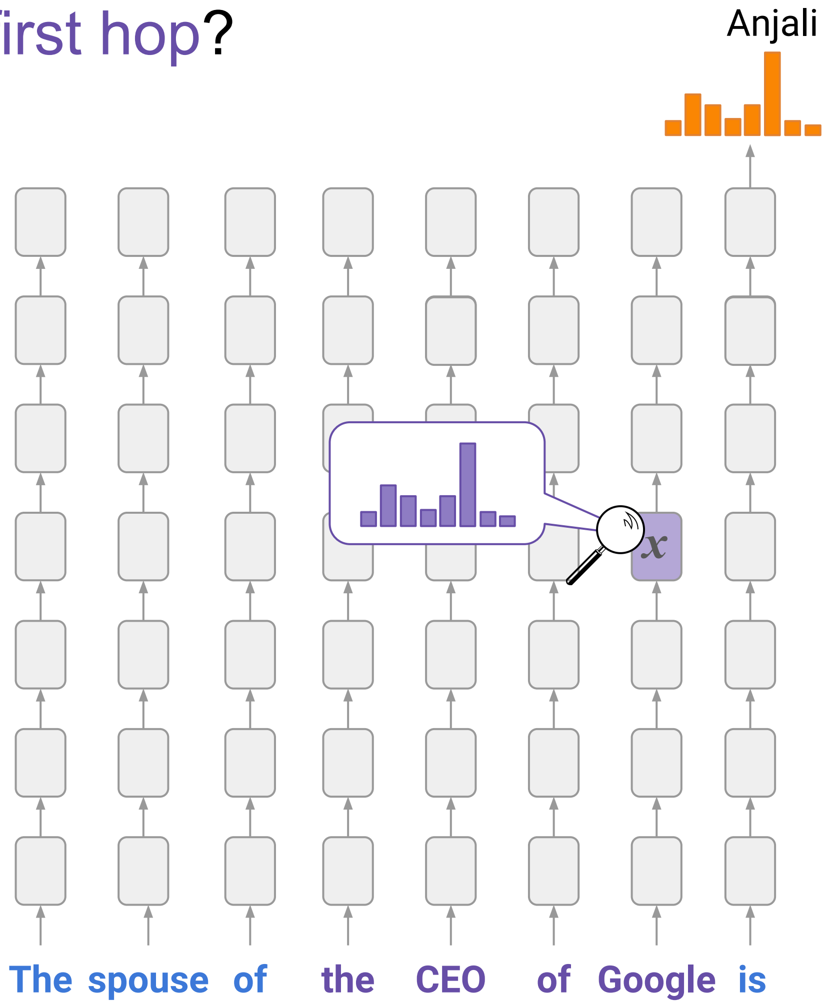
Q1 Does the model resolve the first hop?

Estimate the degree of entity recall via projection to the vocabulary

$$\mathbf{p}^l = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{x})$$

$$\log \mathrm{p}^l(\ \textbf{Sundar}\ |\ \textbf{... the CEO of Google}\ )$$

internal entity recall score

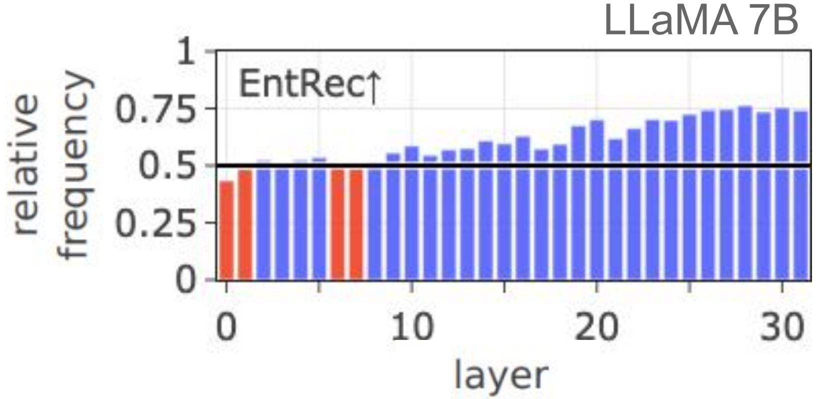


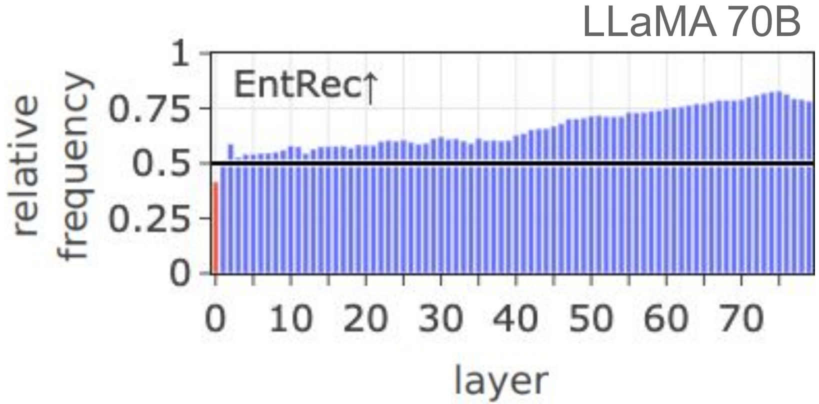
Q1 Does the model resolve the first hop?

Check if the recall of an entity increases when modifying the prompt to describe it

$$\log p^{l}(\text{Sundar} \mid \text{... the CEO of Google}) \overset{?}{>} \log p^{l}(\text{Sundar} \mid \text{... the COO of Google})$$

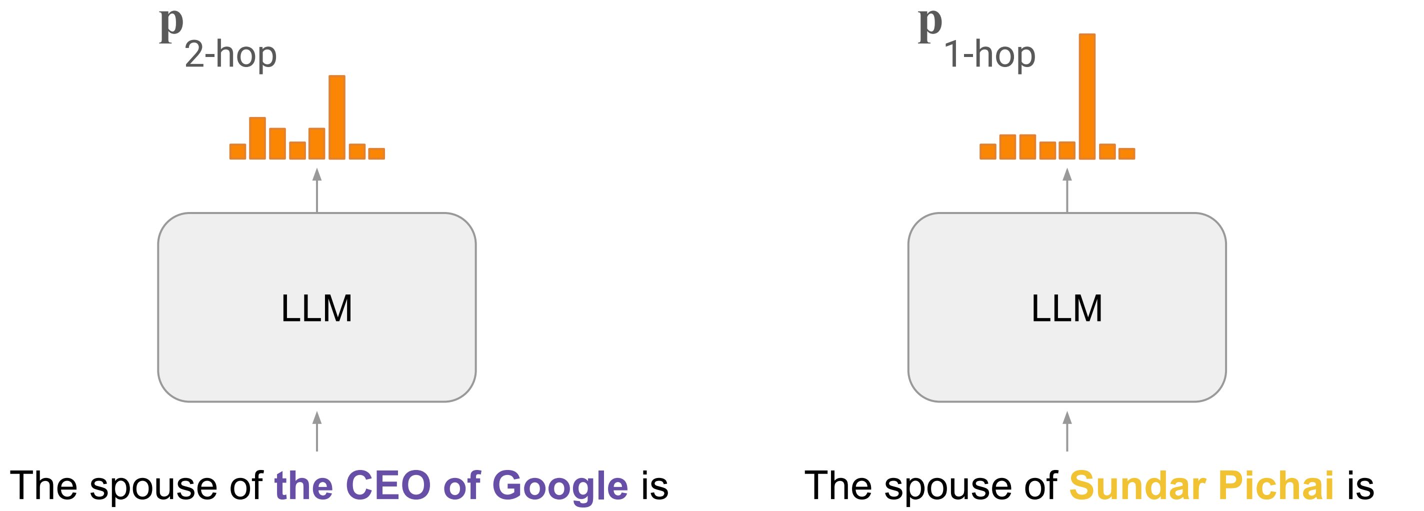
The entity recall increases when the prompt describes it, indicating a resolution of the first hop!





Q2 Does the model utilize the first hop for answering the second hop?

Check consistency between the output probability distributions for corresponding one-hop and two-hop prompts

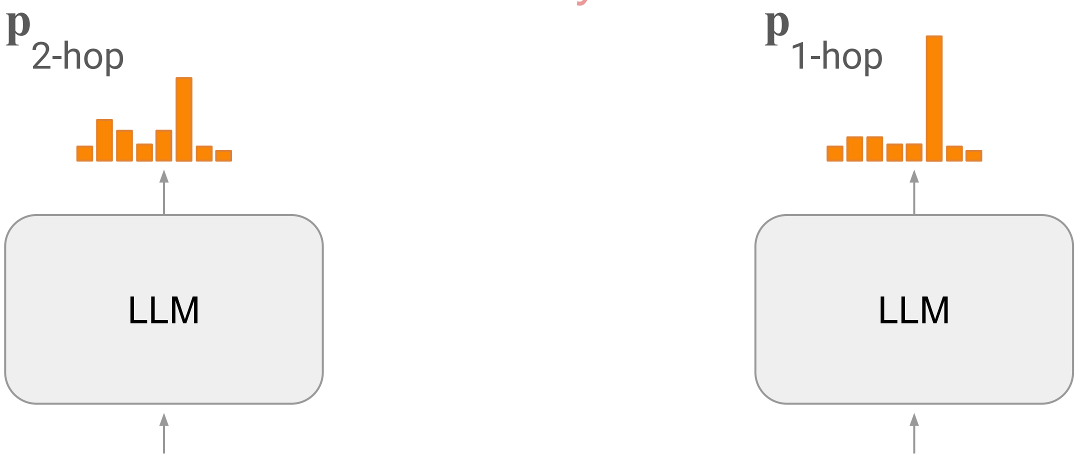


Q2 Does the model utilize the first hop for answering the second hop?

Check consistency between the output probability distributions for corresponding one-hop and two-hop prompts

$$\boxed{-\,0.5\,H(\mathbf{p}_{\text{2-hop}}, \mathbf{p}_{\text{1-hop}}) - 0.5\,H(\mathbf{p}_{\text{1-hop}}, \mathbf{p}_{\text{2-hop}})}$$

consistency score



The spouse of **the CEO of Google** is

The spouse of **Sundar Pichai** is

High-level approach

- Internal entity recall score that measures resolution of the first-hop
- Consistency score that measures utilization of the first-hop
- Check if increasing entity recall also increases first-hop utilization.

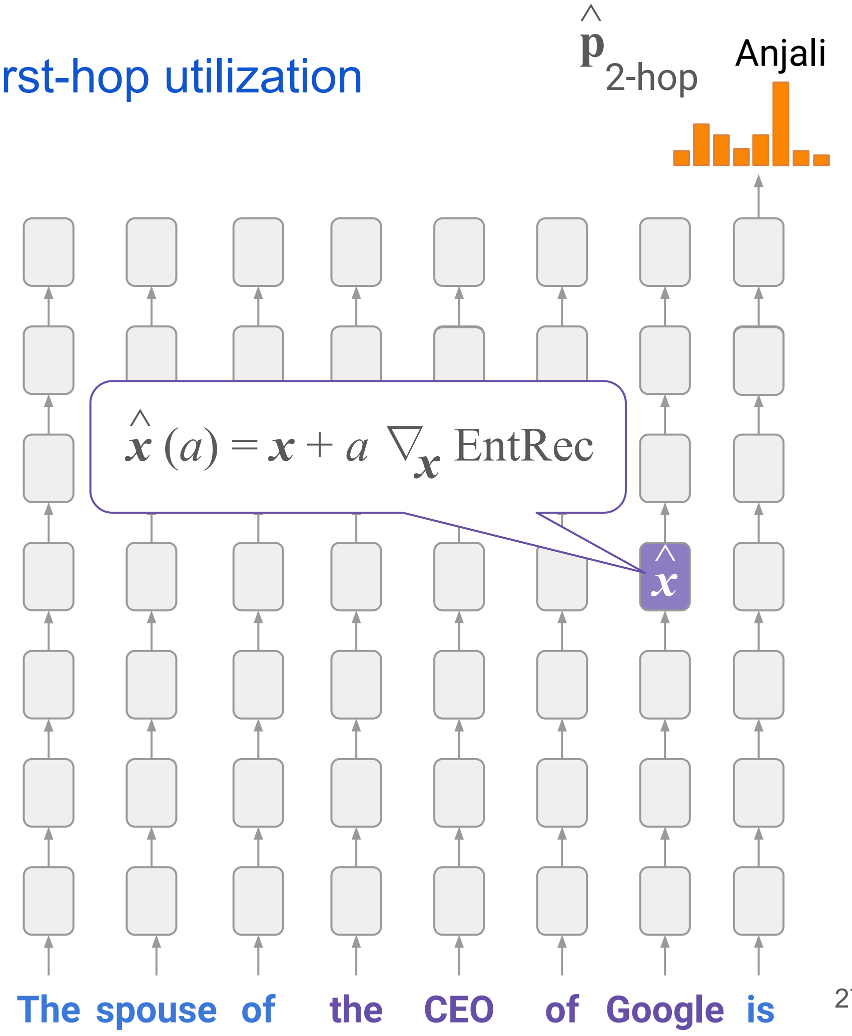
  A positive answer would be an indication for a second-hop presence!

Check if increasing entity recall→ increases first-hop utilization

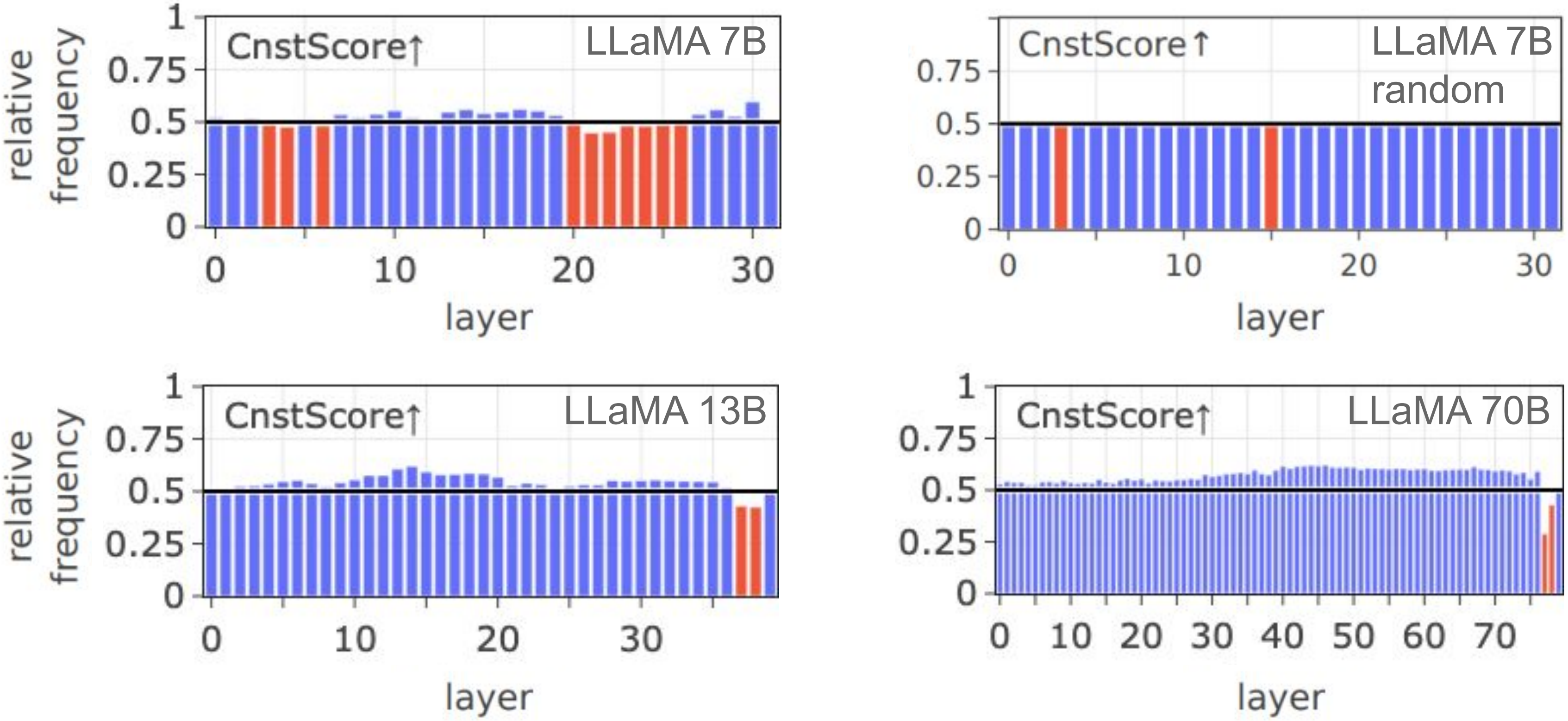
Now the consistency score is a function of *a*.
Calculating its derivative at *a* = 0:

Positive: an infinitesimal increase in entity recall will increase consistency
→ the model **utilizes** the first-hop

Negative: an infinitesimal increase in entity recall will decrease consistency
→ the model **does not utilize** the first-hop



LLMs only weakly perform the second-hop of reasoning, which does not increase with model scale!



Conclusions

- Strong signal for first-hop resolution
- Weak evidence for second-hop resolution which does not scale
- Possibly other more dominant pathways for solving these queries

Let’s dive deeper…

Exploring the limitations of latent reasoning in LLMs

Eden Biran

Daniela Gottesman

Sohee Yang

Amir Globerson

Experimental setting

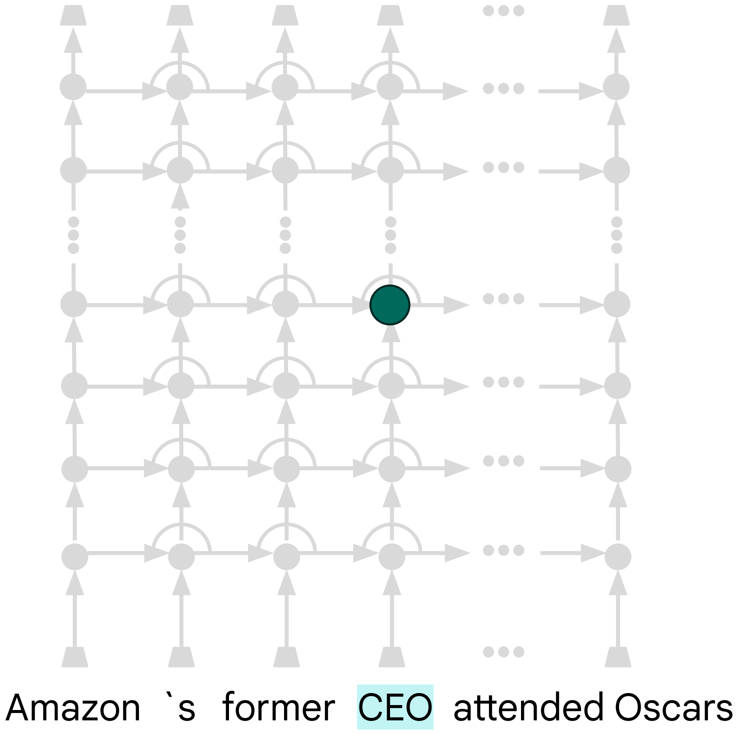
Data:

- 82,020 two-hop queries based on Wikidata
- **Filter out cases of possible shortcuts**
    - “The spouse of the CEO is”
    - “The spouse of Google is”
- Balanced correct and incorrect subsets

Models:

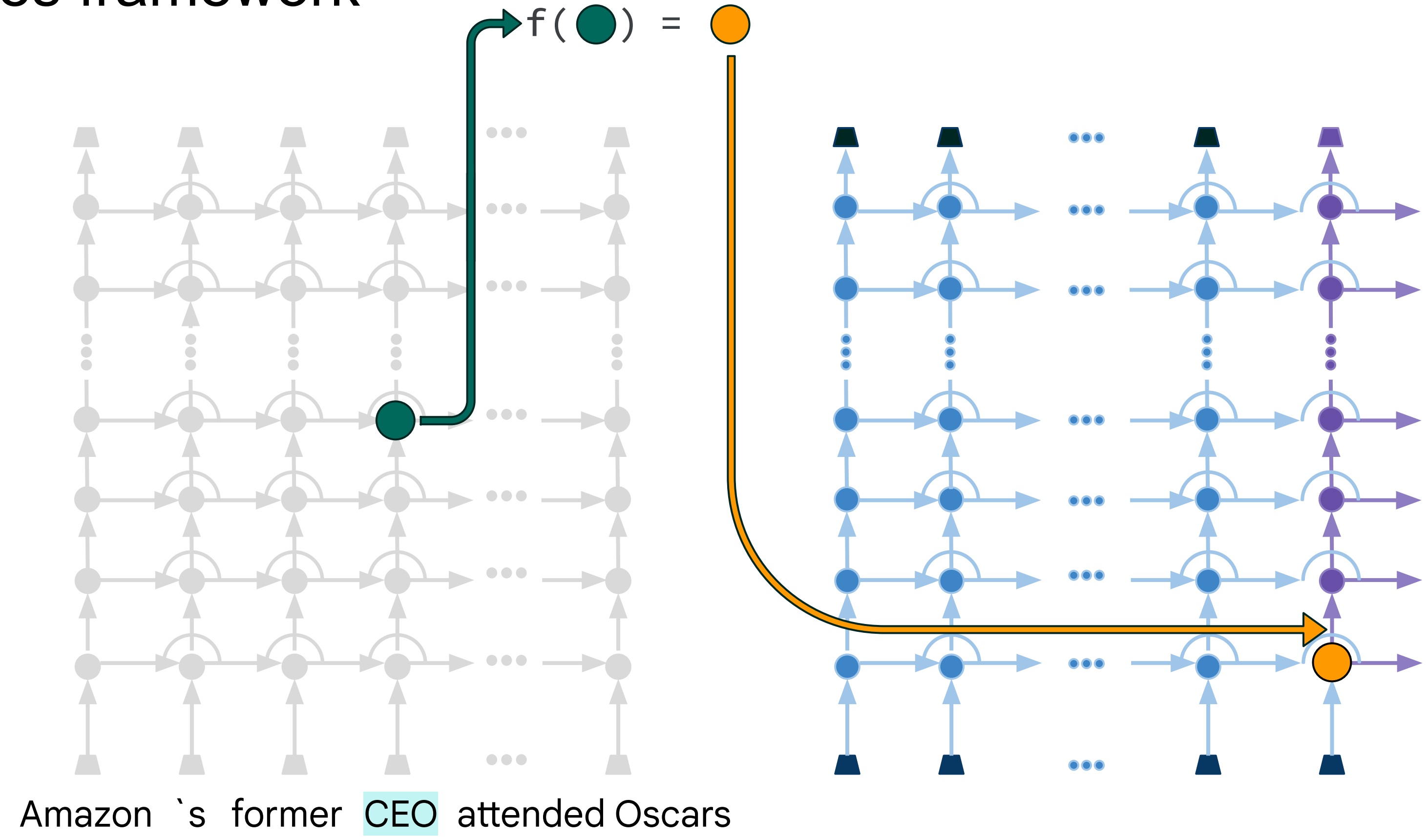
- LLaMA 2 7B and 13B
- LLaMA 3 8B and 70B
- Pythia 6.9B and 12B

Patchscopes framework



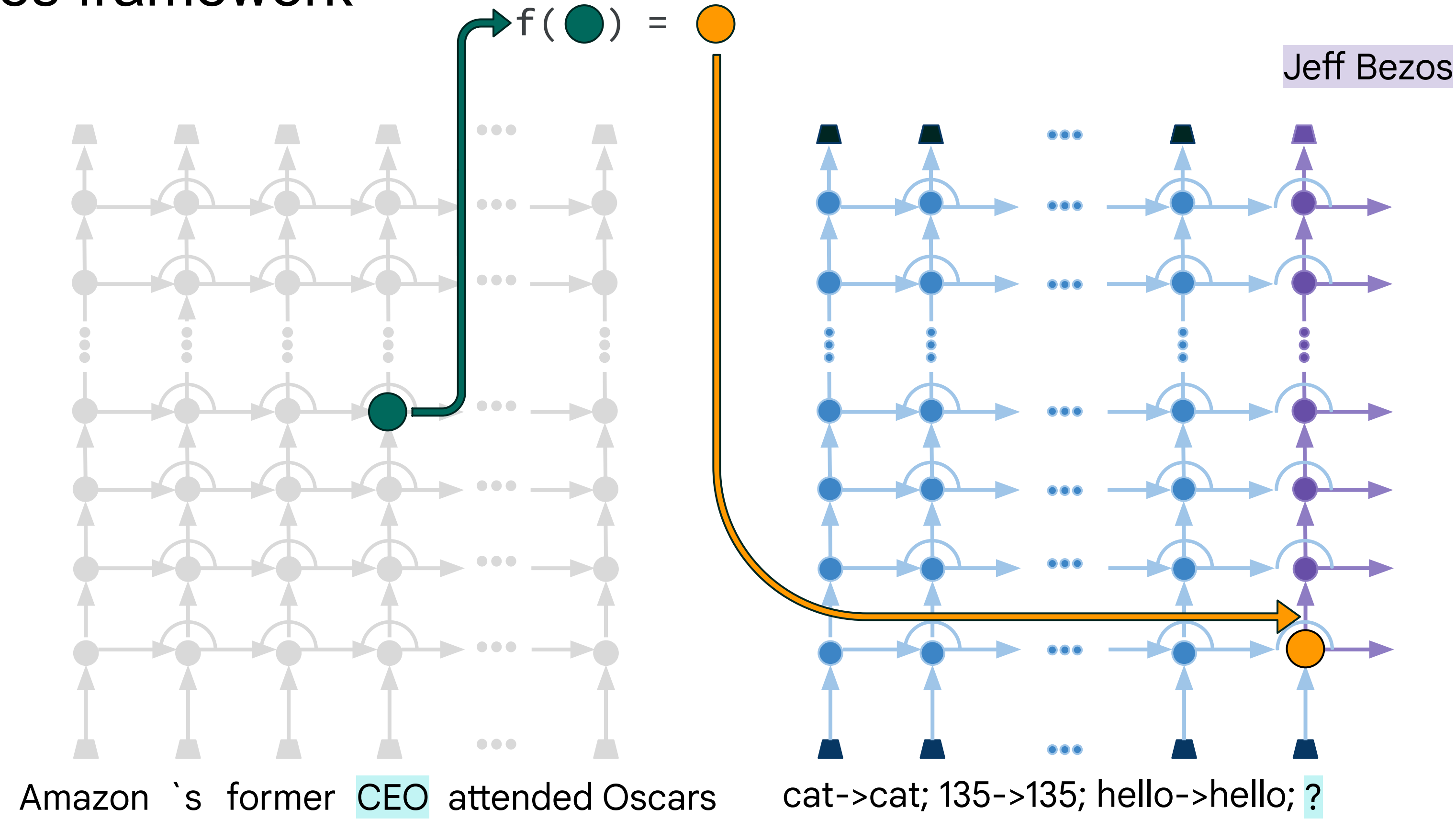
Ghandeharioun et al. 2024

Patchscopes framework



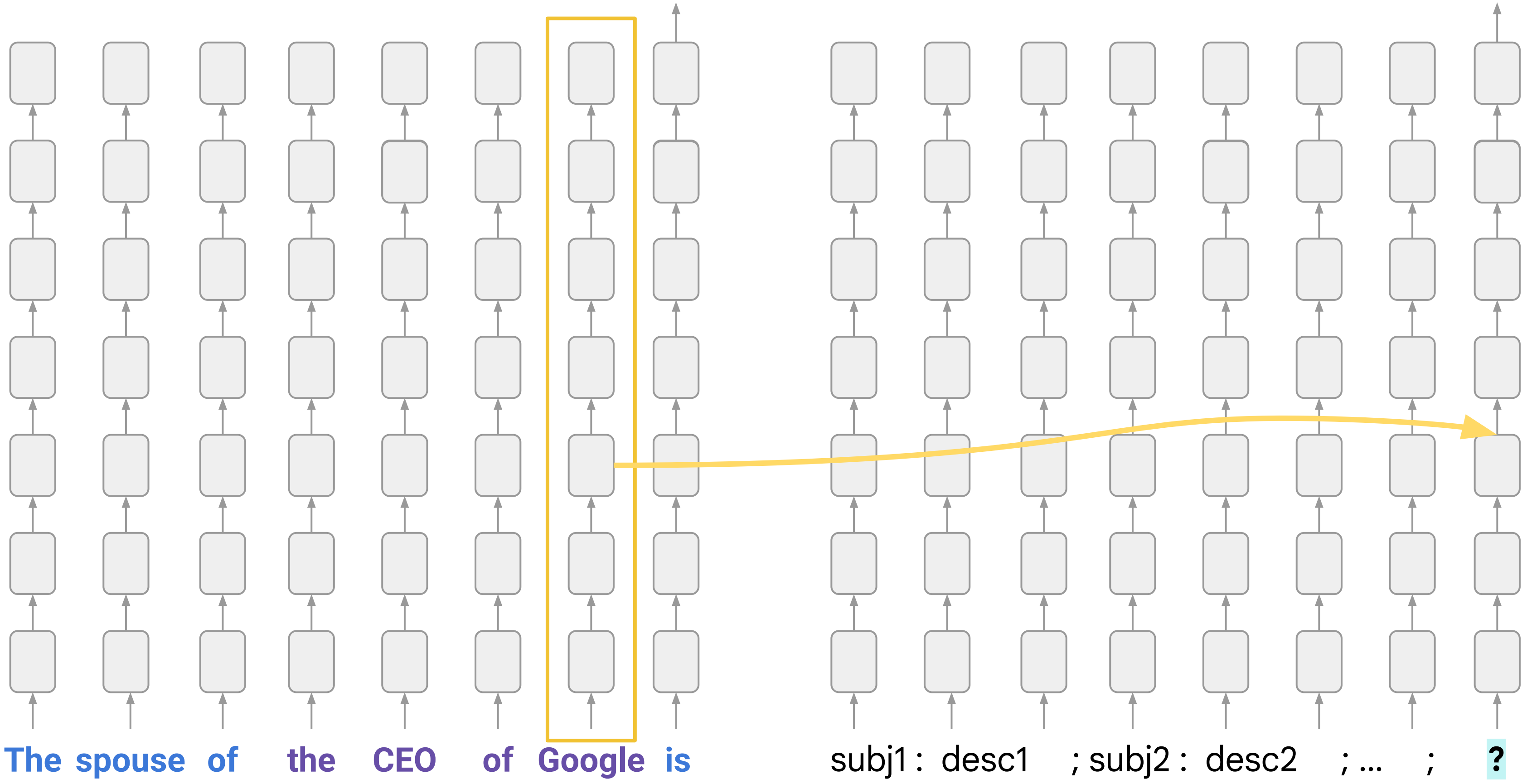
Ghandeharioun et al. 2024

Patchscopes framework

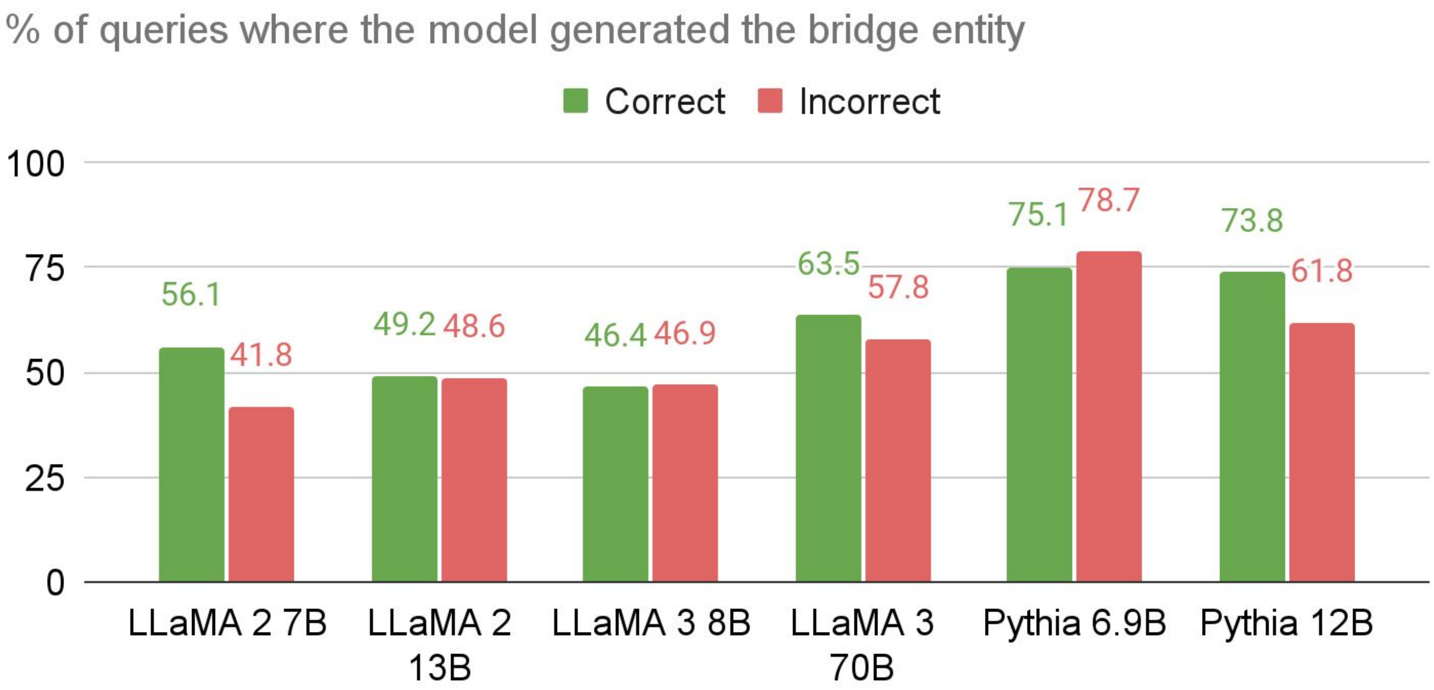


Ghandeharioun et al. 2024

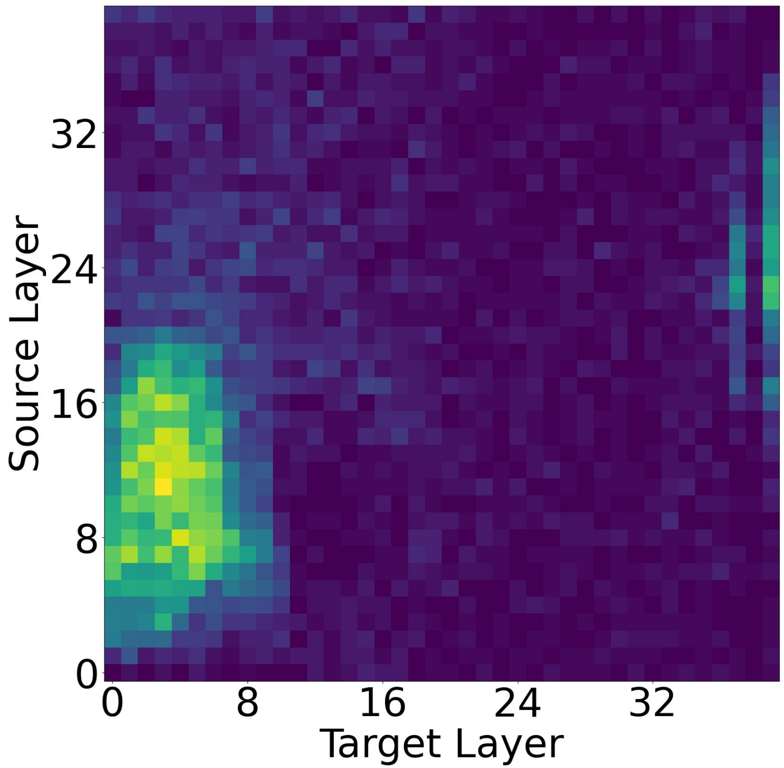
What entity is encoded in the last position of the first hop?



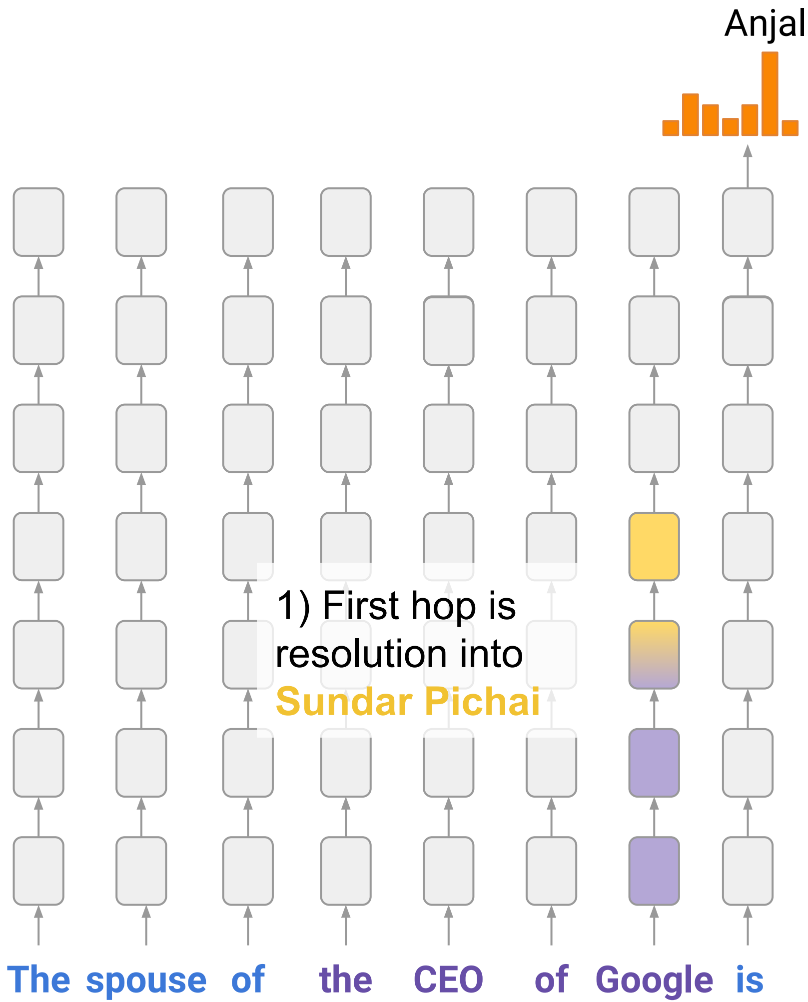
The bridge entity is often resolved



The bridge entity is often resolved in the early layers



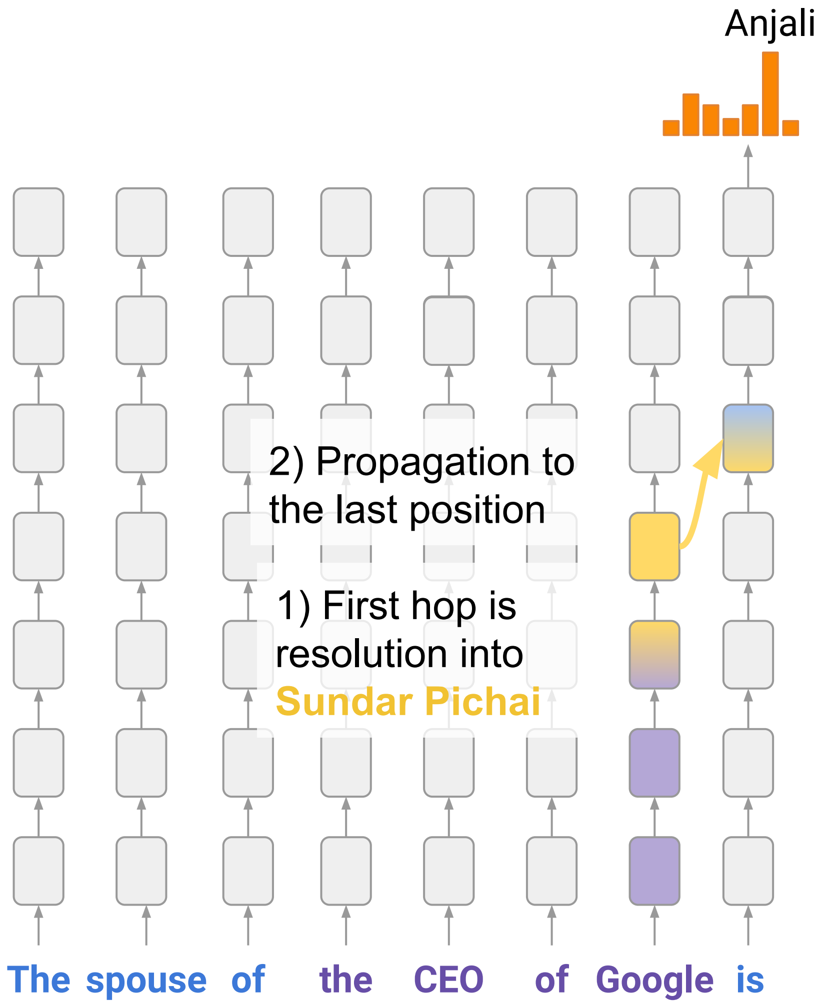
A pathway of latent reasoning



A pathway of latent reasoning

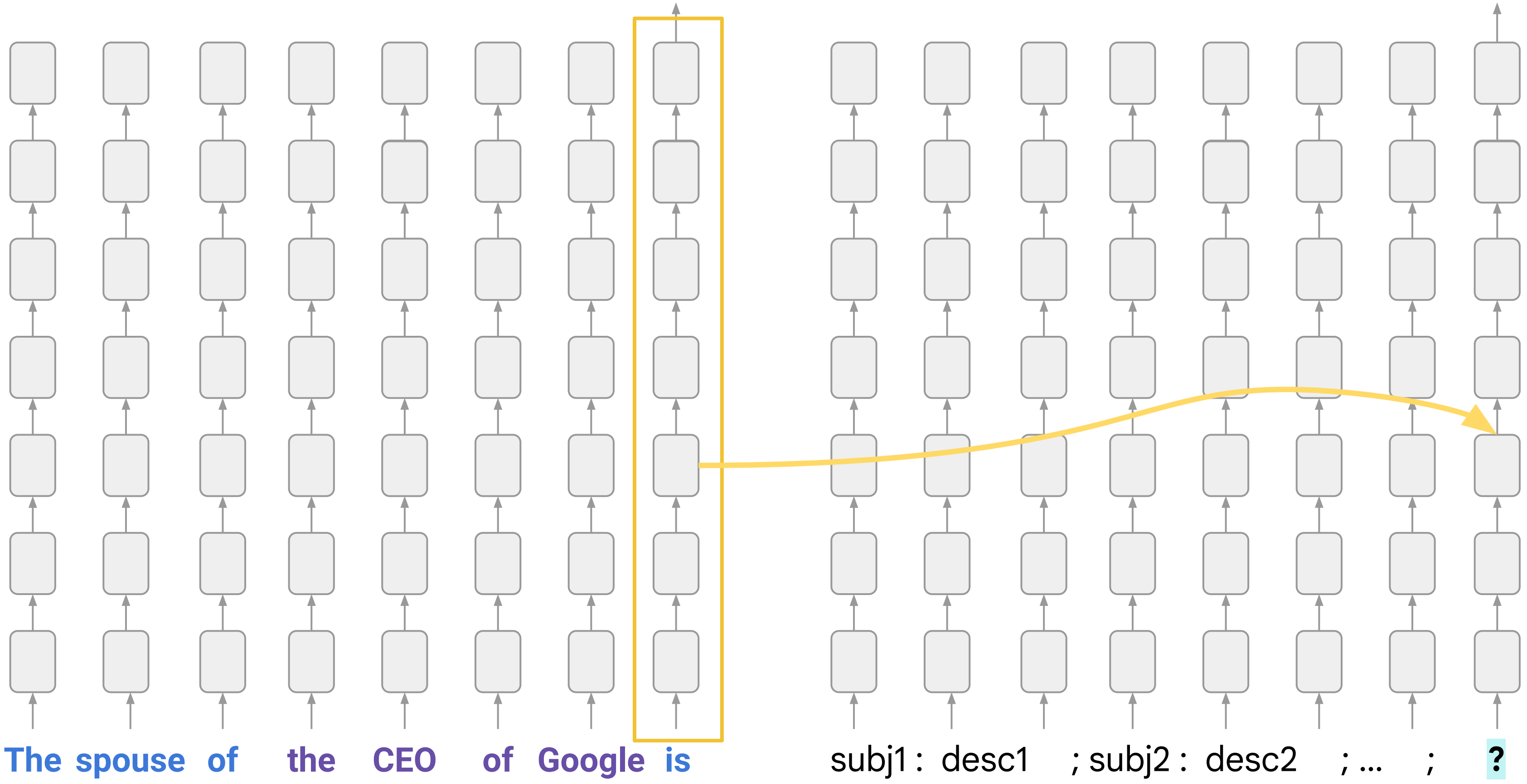
Using attention knockout, vocabulary projections, and Patchscopes

78%-96% detection in **correct** cases
71%-95% in the **incorrect** cases

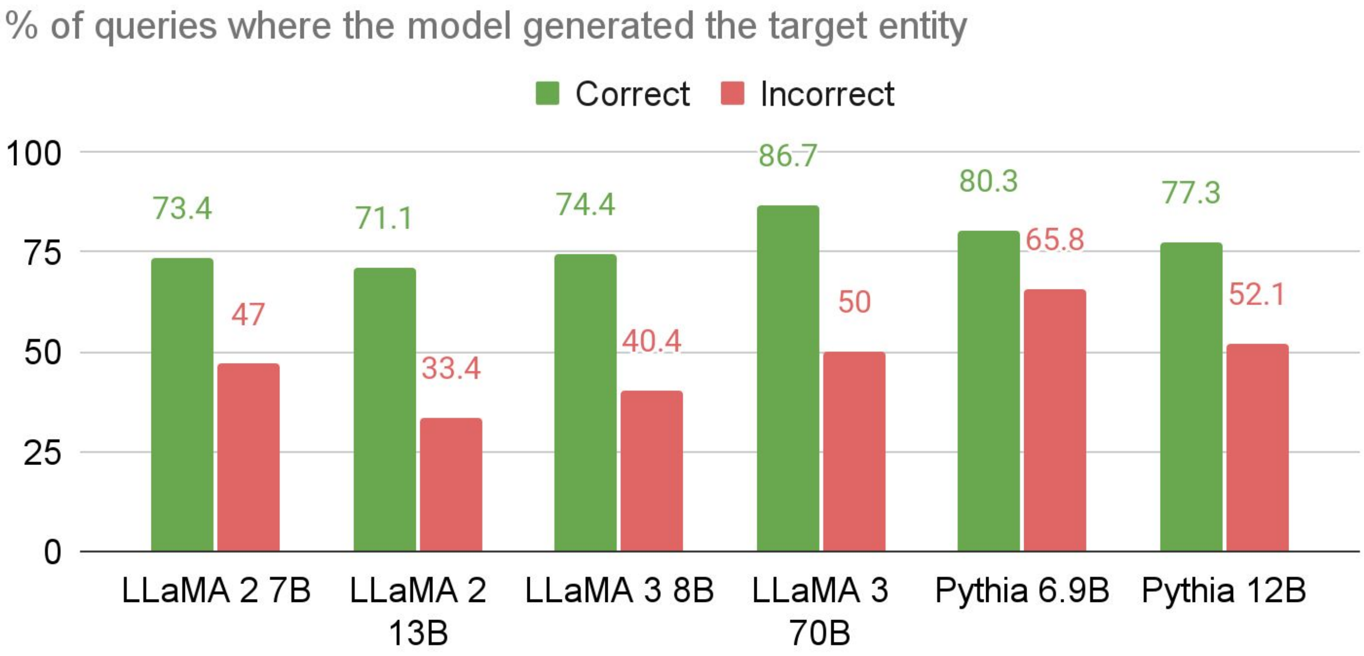


Geva et al. 2023, Ghandeharioun et al. 2024

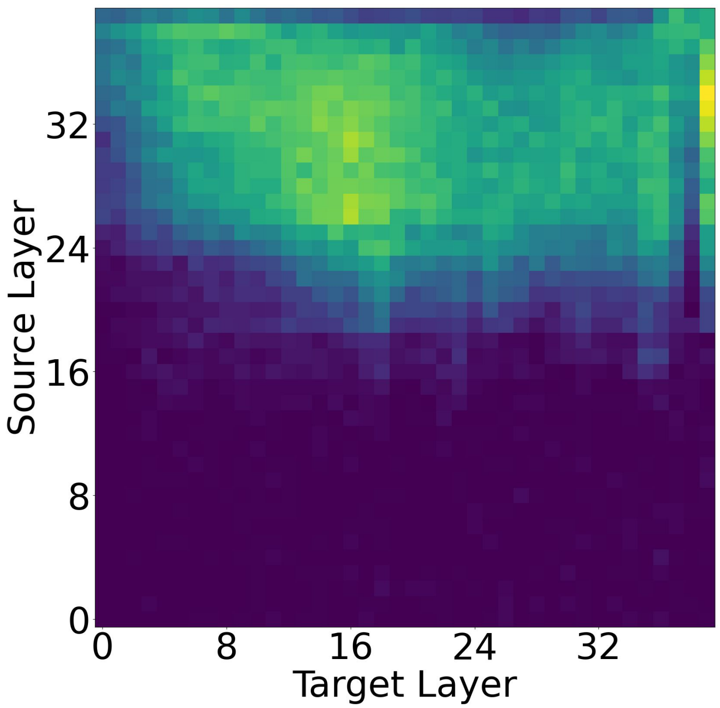
What entity is encoded in the last position of the second hop?



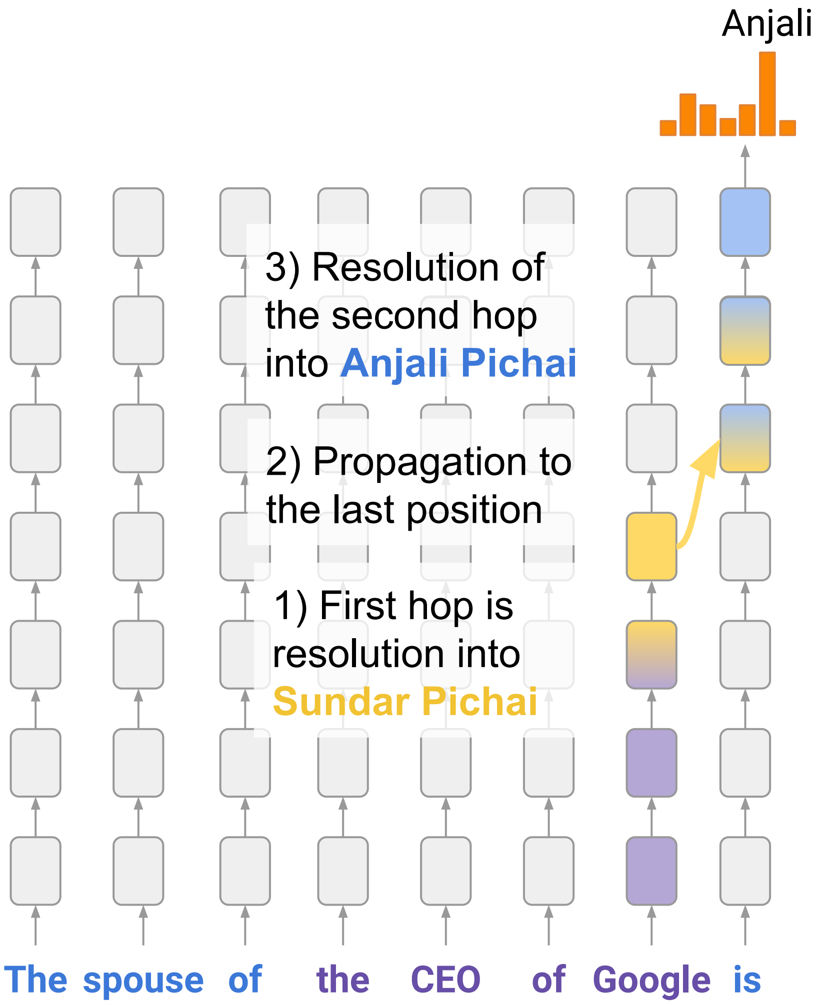
The target entity is resolved less frequently in incorrect cases



The target entity is resolved in the upper layers

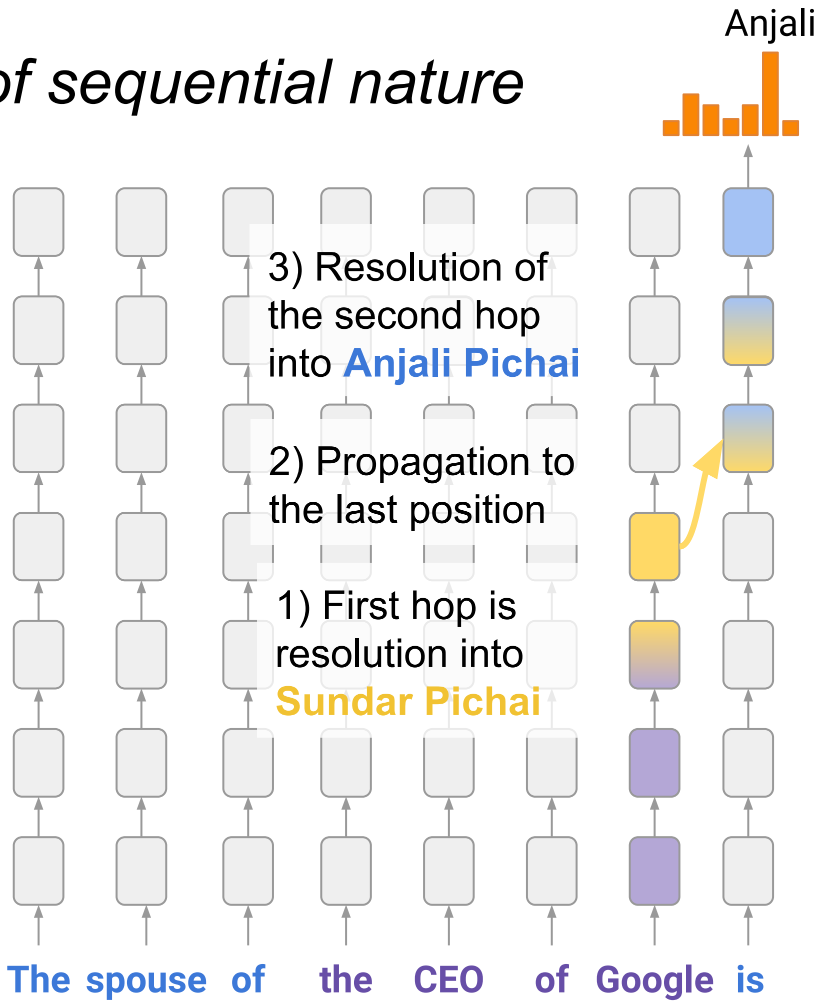


A pathway of latent reasoning

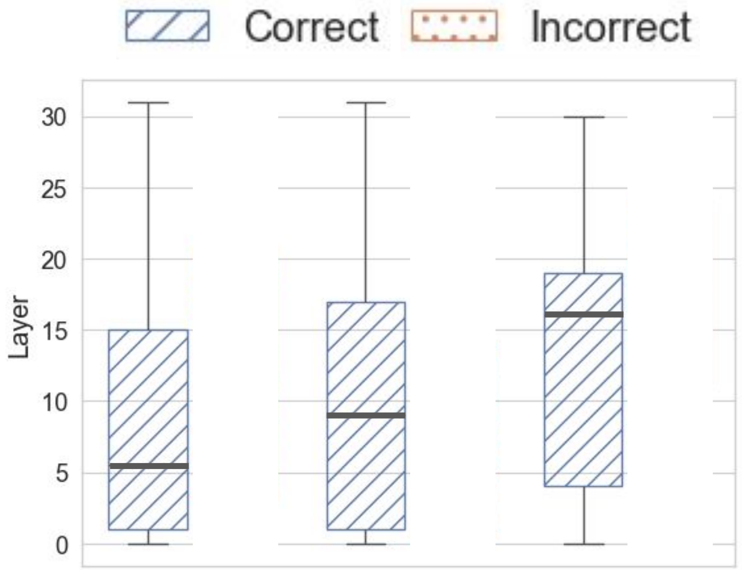


Geva et al. 2023, Ghandeharioun et al. 2024

A pathway of latent reasoning *of sequential nature*

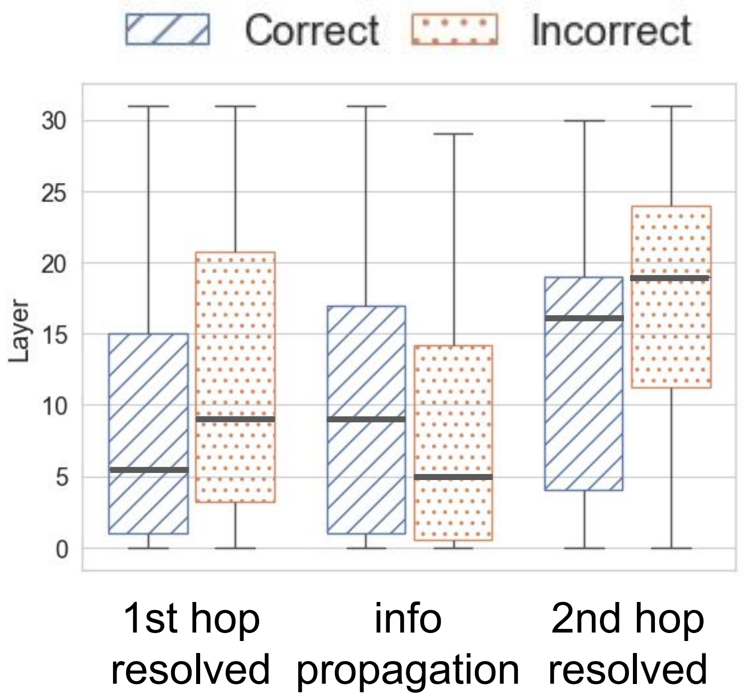


Geva et al. 2023, Ghandeharioun et al. 2024



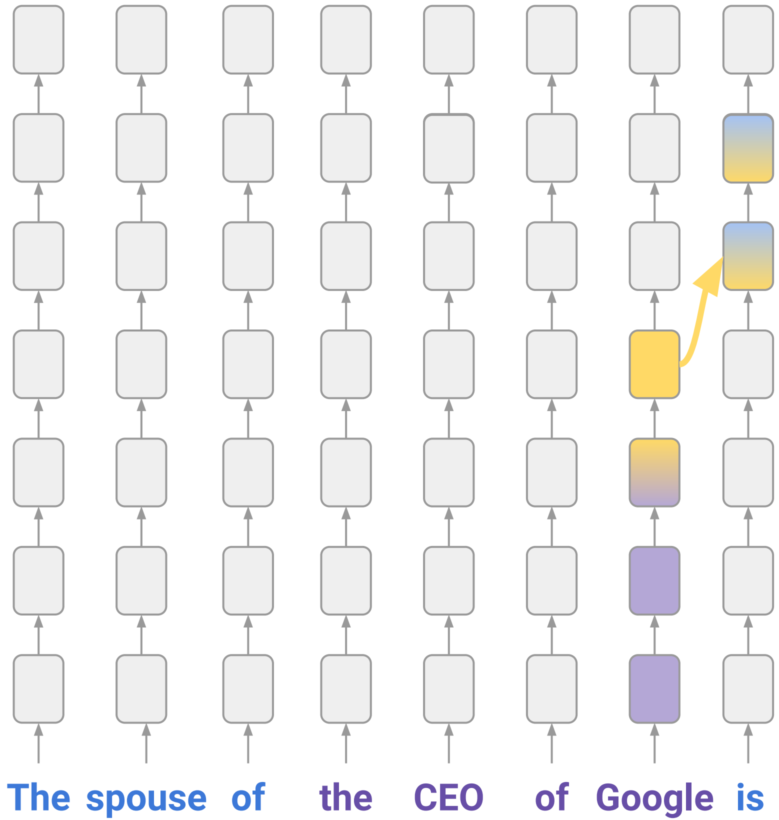
1st hop resolved | info propagation | 2nd hop resolved

When the model fails, the entities are resolved later while information propagation happens earlier



Hypothesis: latent reasoning failures stem from the first hop being resolved “too late” — at layers that no longer contain the information needed to resolve the second hop

Back-patching analysis

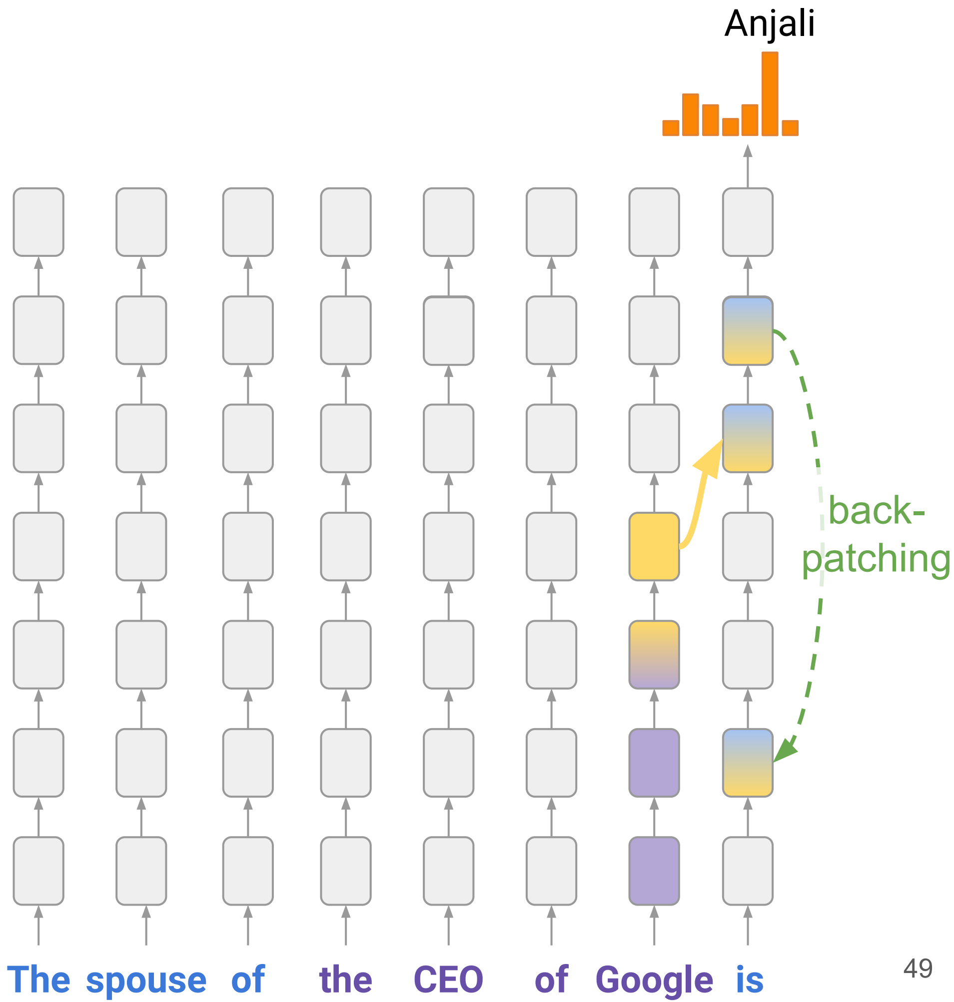


Back-patching analysis

Substantial gains in incorrect cases

	patching the first hop	patching the second hop
LLaMA 2 7B	41%	42.5%
LLaMA 2 13B	32.4%	36.1%
LLaMA 3 8B	38.8%	47.2%
LLaMA 3 70B	57.3%	57.8%
Pythia 6.9B	66.3%	56.4%
Pythia 12B	63.2%	61.8%

100% success rate in correct cases



Key takeaways

- Existential evidence of latent reasoning in LLMs
- A pathway prominent in cases that are less likely to include shortcuts
- Points to a limitation in the computation of LLMs in performing latent reasoning
- Success cases may be achieved with other pathways that do not rely on "backwards" reasoning

How to (and should we) build models that perform reasoning in their latent space?