

Llama See, Llama Do:

A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs

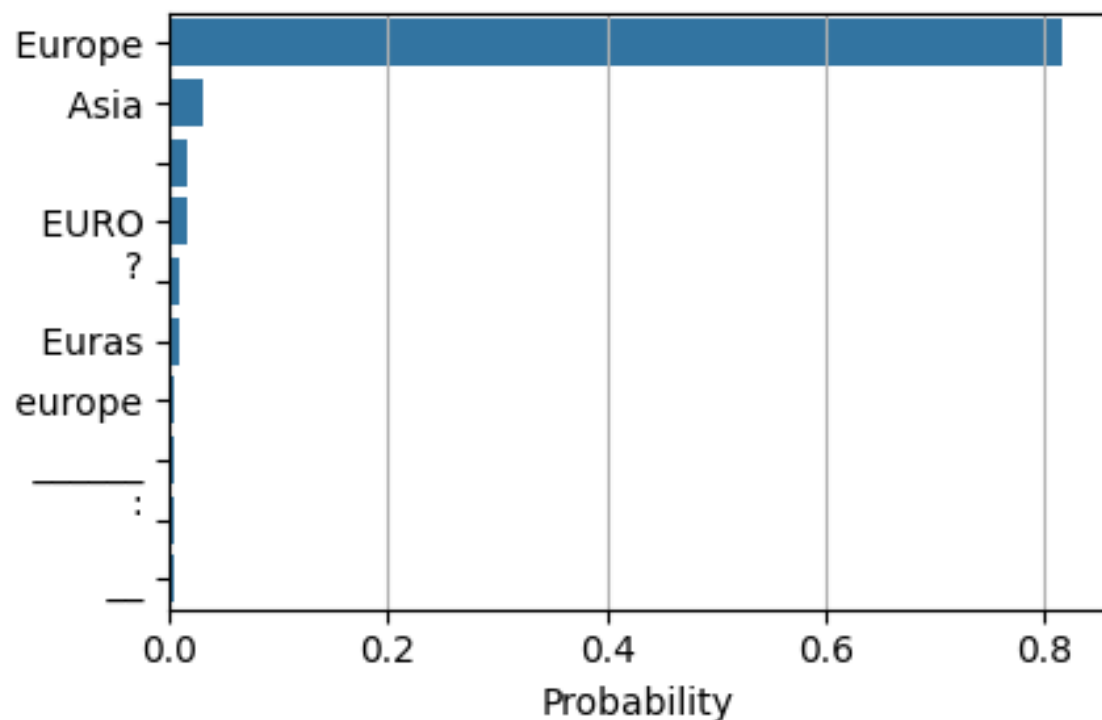


Jingcheng Niu*, Xingdi Yuan, Tong Wang,
Hamidreza Saghir, Amir H. Abdi

LLM Distraction

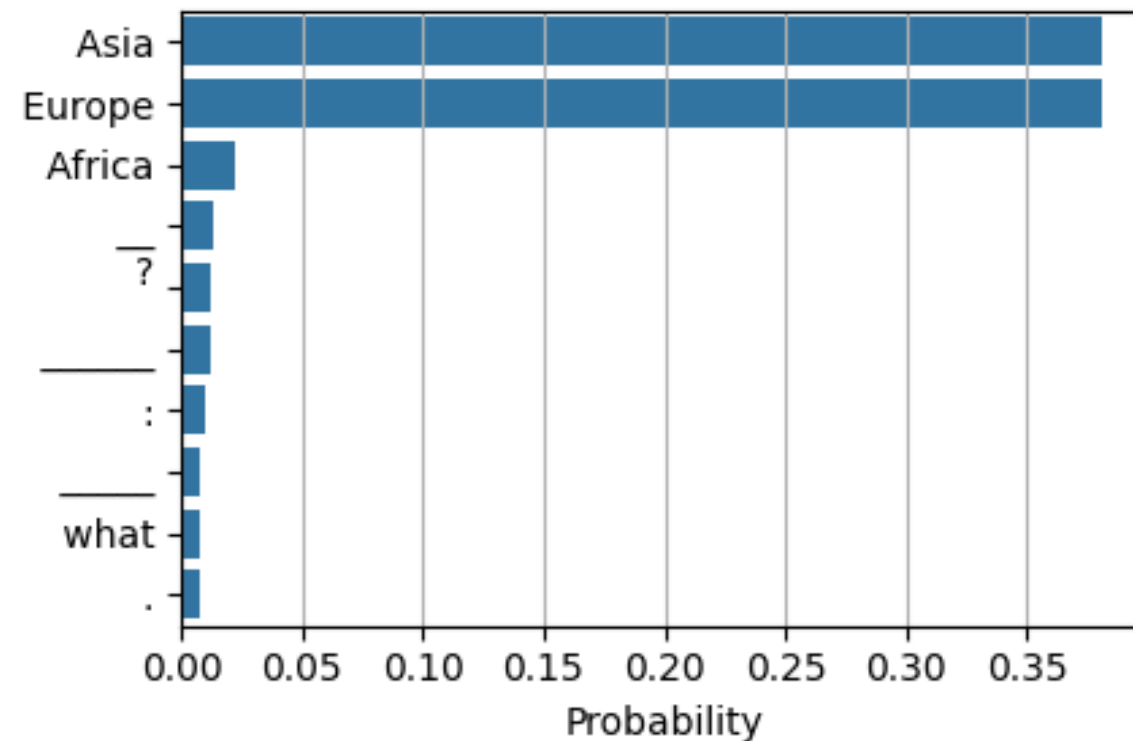
All probs in slides from Llama-3.1 8B.
Results generalizable to other LM families.

QUERY: Greece is located on
the continent of ____



CONTEXT: **Iraq is in Asia.**

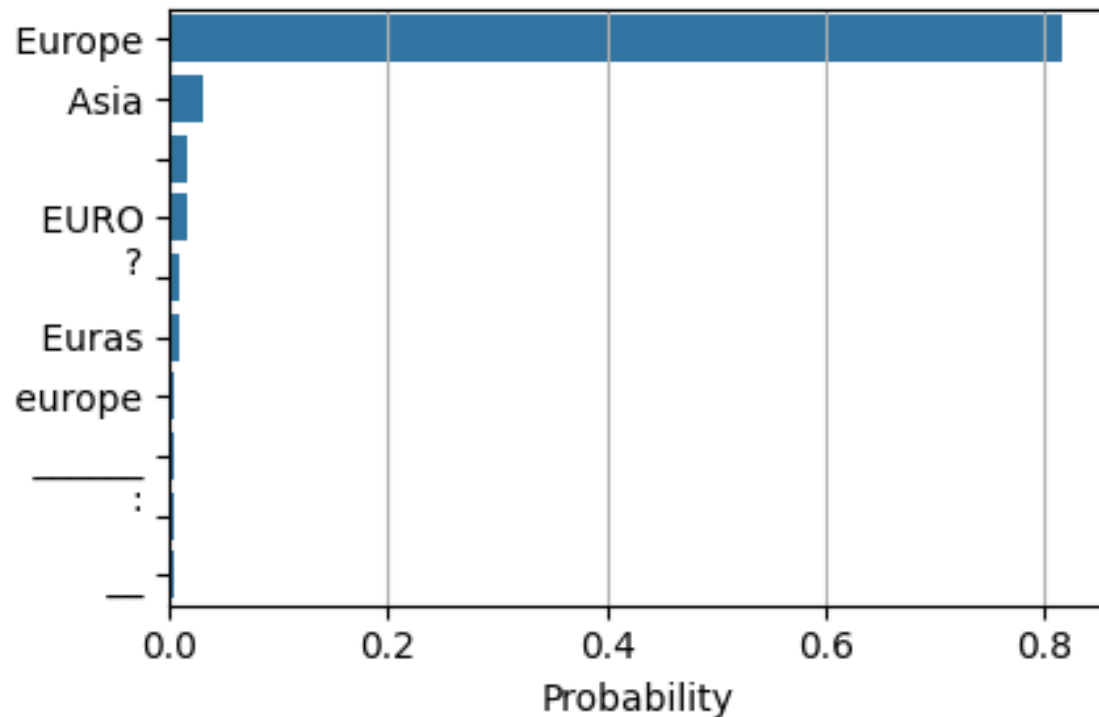
QUERY: Greece is located on
the continent of ____



LLM Distraction

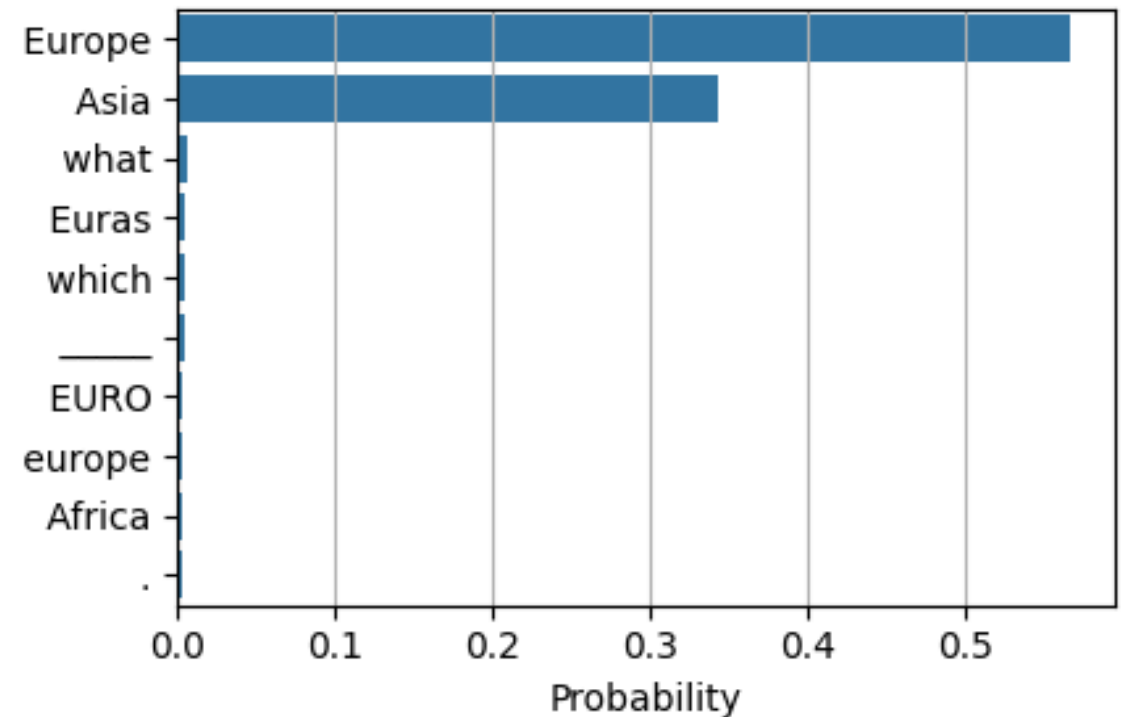
All probs in slides from Llama-3.1 8B.
Results generalizable to other LM families.

QUERY: Greece is located on
the continent of ____



CONTEXT: **Asia** is the largest
continent in the world by both
land area and population.

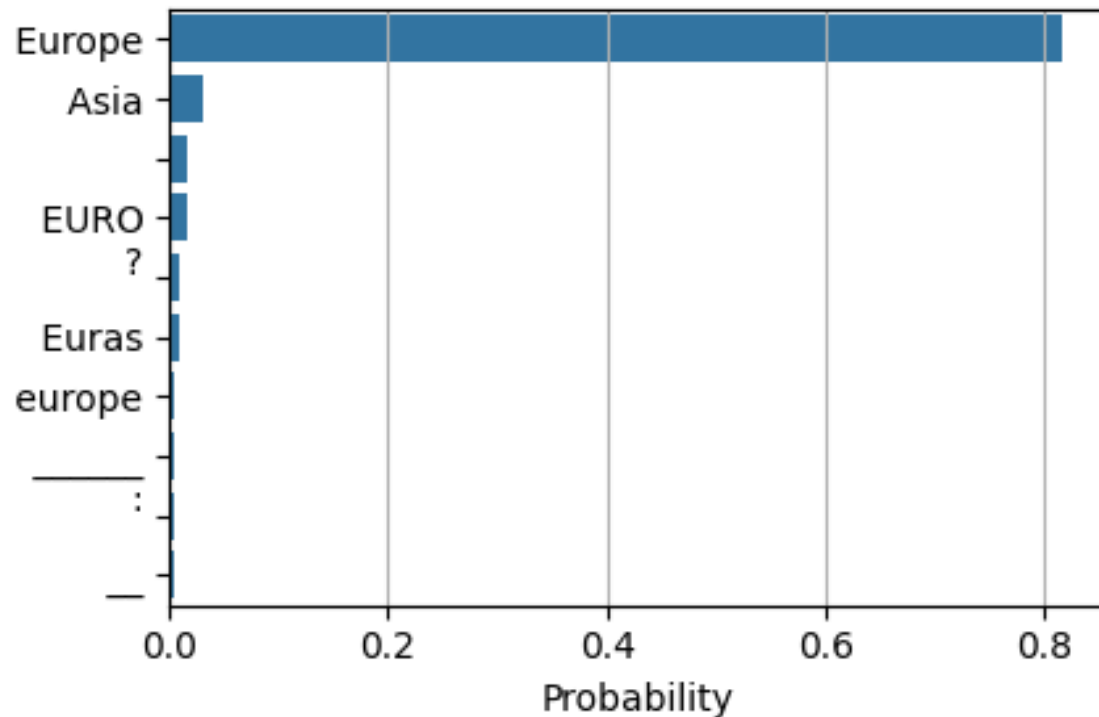
QUERY: Greece is located on
the continent of ____



LLM Distraction

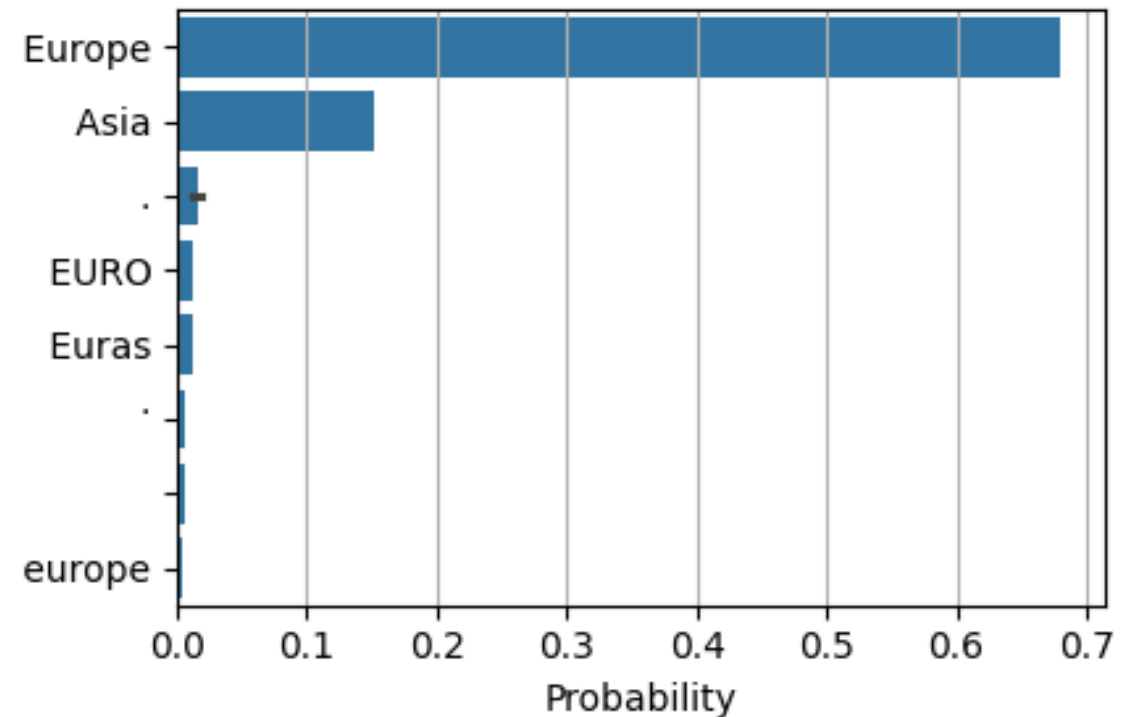
All probs in slides from Llama-3.1 8B.
Results generalizable to other LM families.

QUERY: Greece is located on
the continent of ____



CONTEXT: **Asia.**

QUERY: Greece is located on
the continent of ____



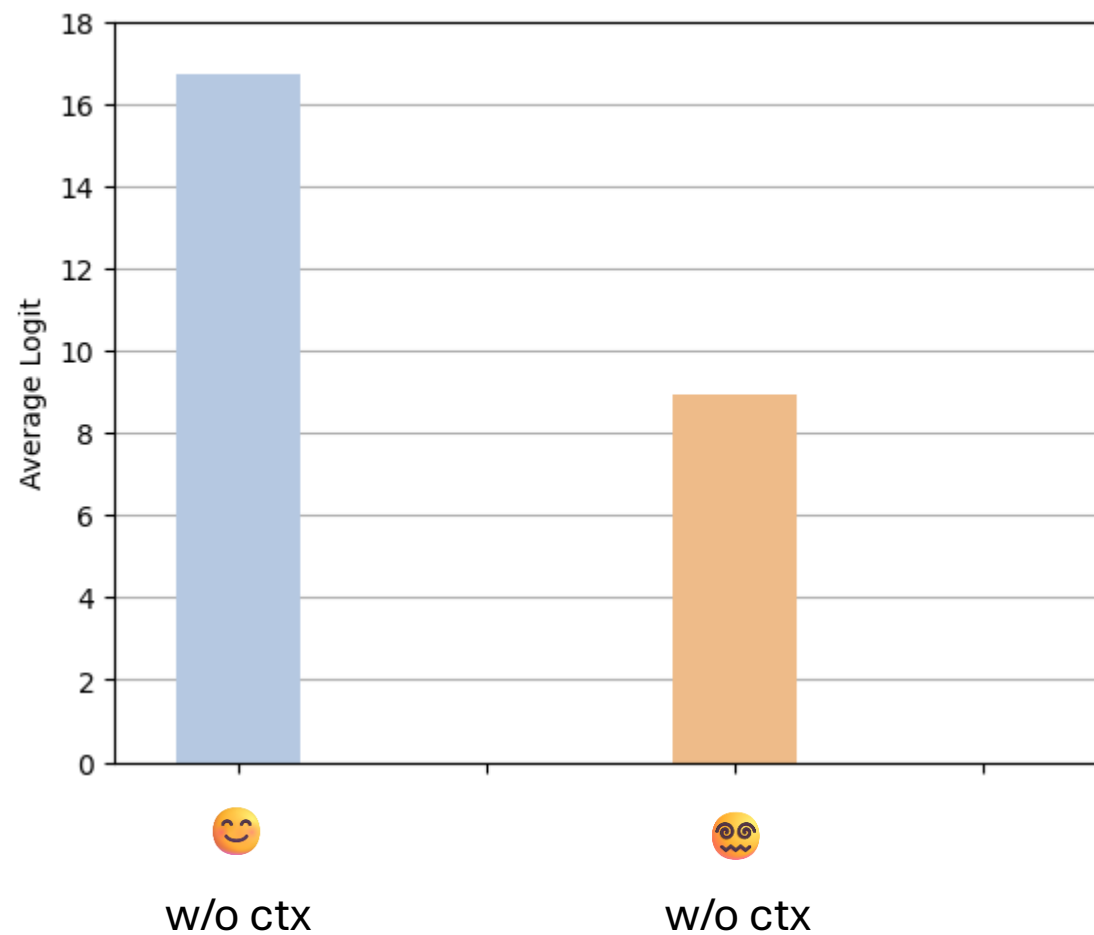
Does Distraction Scale?

- Data: LRE dataset Hernandez et al. (2024).
 - 15 types of factual relations (**relation**, **source**, **target**)
 - Example: (**capital city**, **Canada**, **Ottawa**)
→ The capital city of **Canada** is **Ottawa**.
- 5 models:
 - Llama-3.1-8B, Llama-3.1-8B-Instruct, Llama-2-7b-hf, Llama-2-13b-hf, GPT-2 XL.

What is the capital of Canada? It is the city of

Ottawa 😊

Lima 🤪

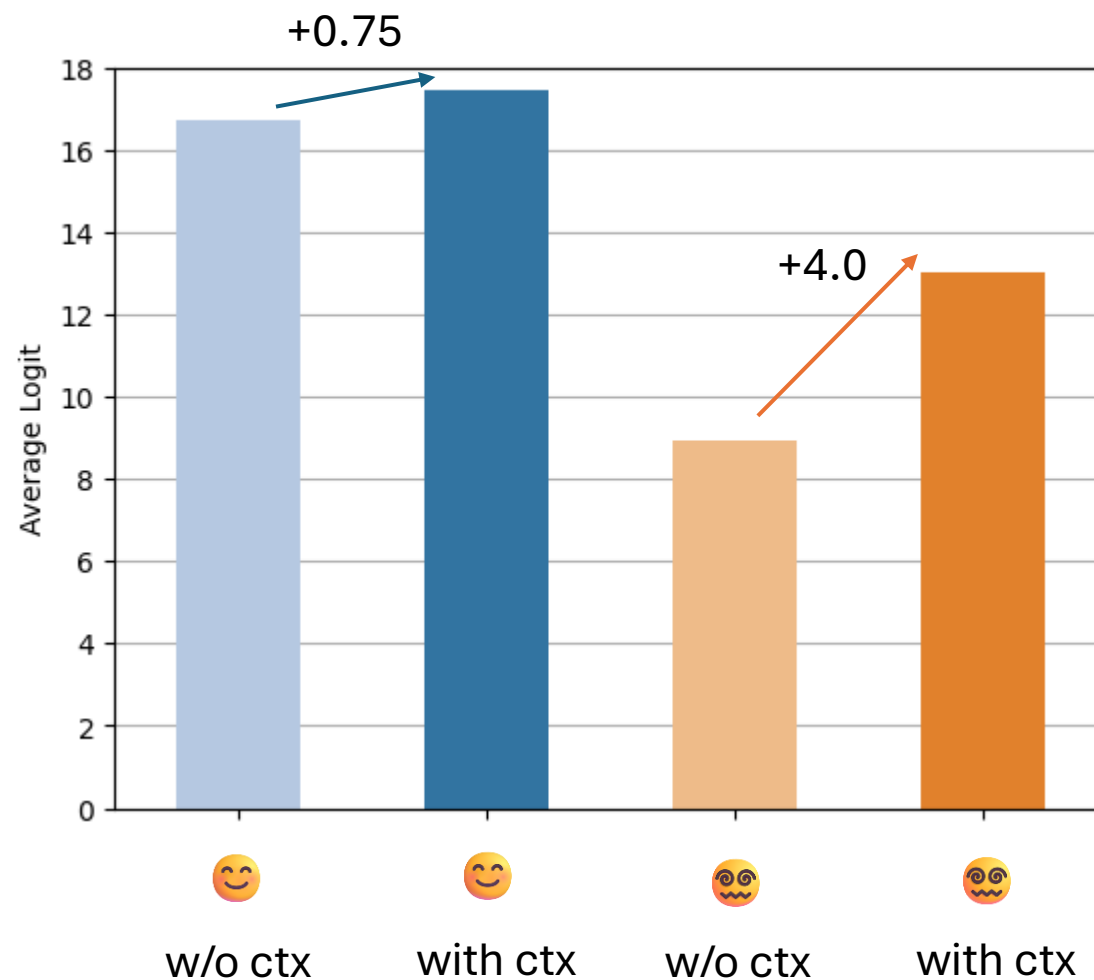


The capital of Peru is **Lima**. What is the capital of Canada? It is the city of

Ottawa 😊

Lima 🤪

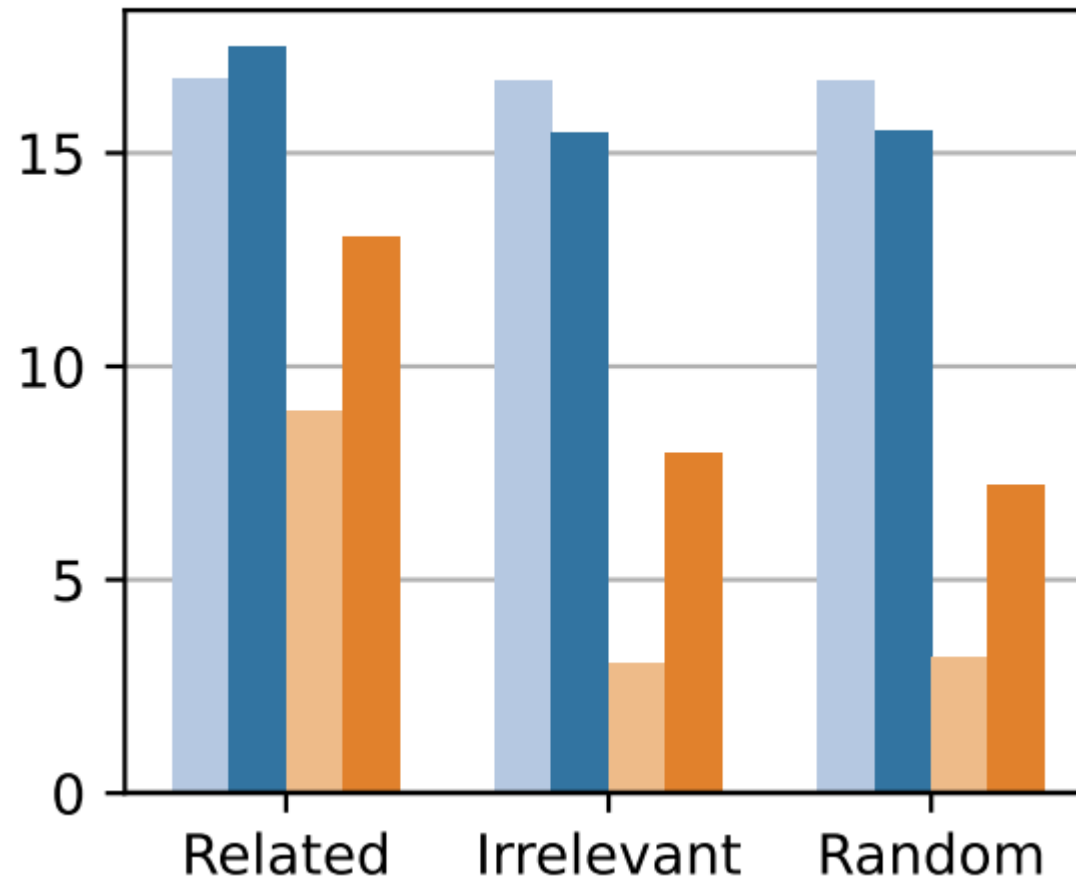
distracting context



Irrelevant: On the outside, apples are red. The capital of Canada is **Ottawa/red**

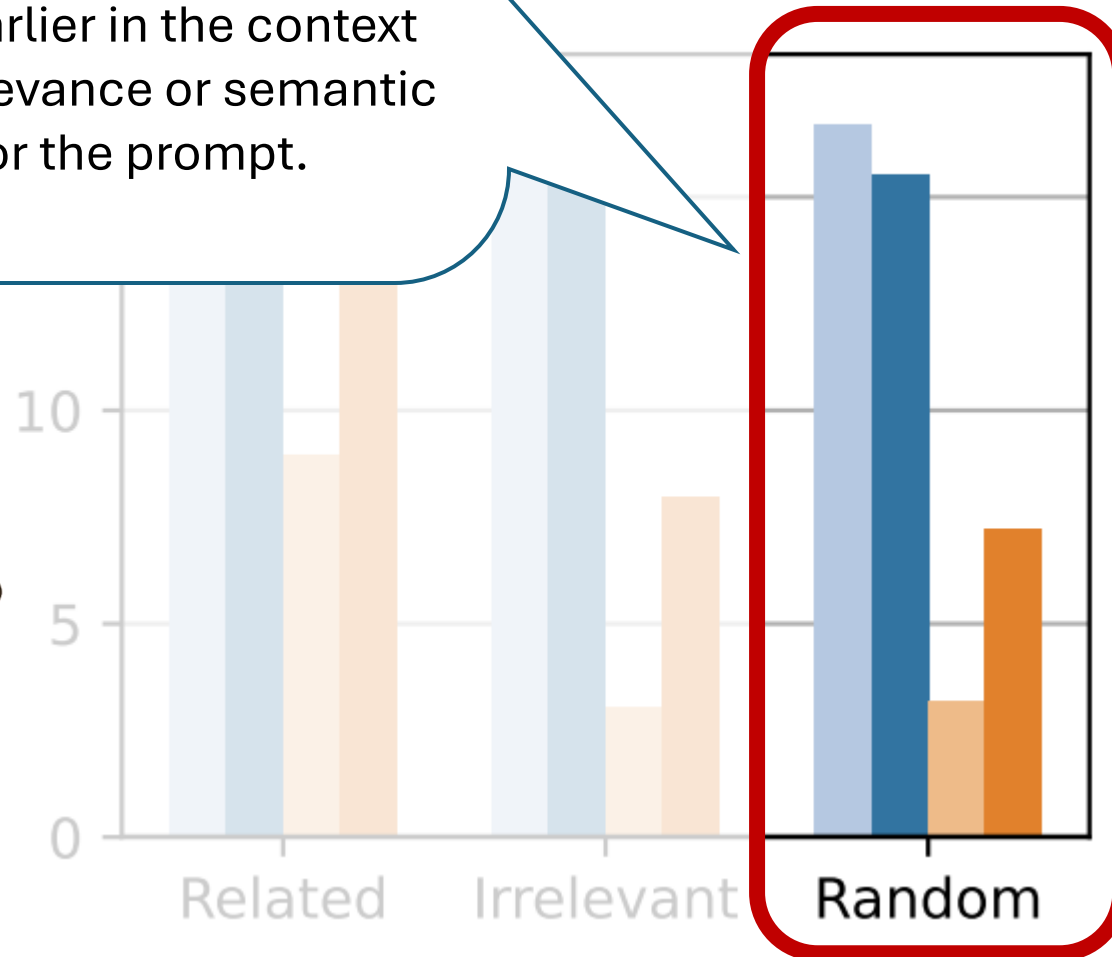
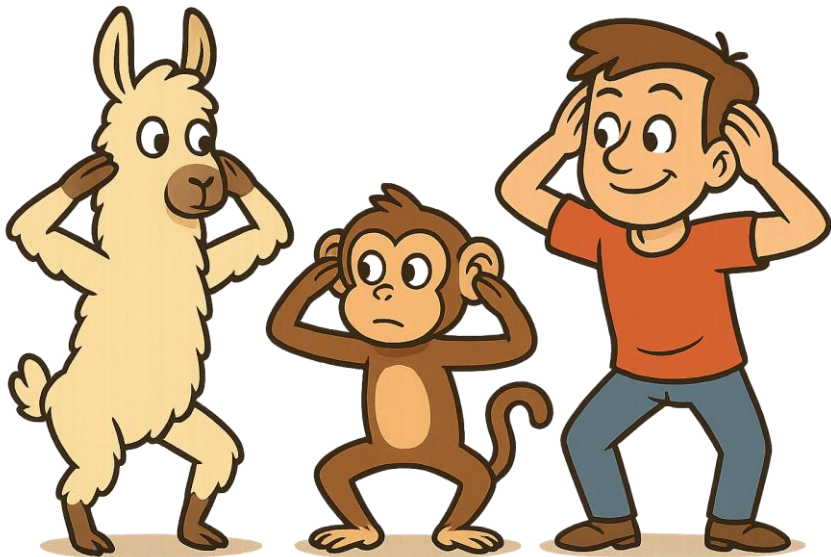
Random: Promotion. What is the capital of Canada? It is the city of **Ottawa/Promotion**

- Data: 15 factual relations from the LRE dataset (Hernandez et al., 2024).
- 10,000 examples per relation.
- 5 models: Llama-3.1-8B, Llama-3.1-8B-Instruct, Llama-2-7b-hf, Llama-2-13b-hf, GPT-2 XL.
- Statistically significant ($p < 0.0001$).



Contextual Entrainment:

Llama see, llama do. LMs consistently assign significantly higher probabilities (or logits) to any tokens that have appeared earlier in the context prompt, regardless of their relevance or semantic relation to the question or the prompt.



Prior Work: Relevance?

Does the context contain the answer?

Original Problem

Jessica is six years older than Claire. In two years, Claire will be 20 years old. How old is Jessica now?


Modified Problem

Jessica is six years older than Claire. In two years, Claire will be 20 years old. *Twenty years ago, the age of Claire's father is 3 times of Jessica's age.* How old is Jessica now?


Standard Answer 24

Shi et al. (2023): ChatGPT can be thrown away by irrelevant sentences in math questions.


Q: Who is the actor playing Jason on general hospital?

Large Language Model (no retrieval) 

The answer is: Steve Burton 

Retrieval Augmented Language Model 

E: Jason Gerhardt (born April 21, 1974) is an American actor. He is known for playing the role of Cooper Barrett in General Hospital and Zack Kilmer in Mistresses.

The answer is: Jason Gerhardt 

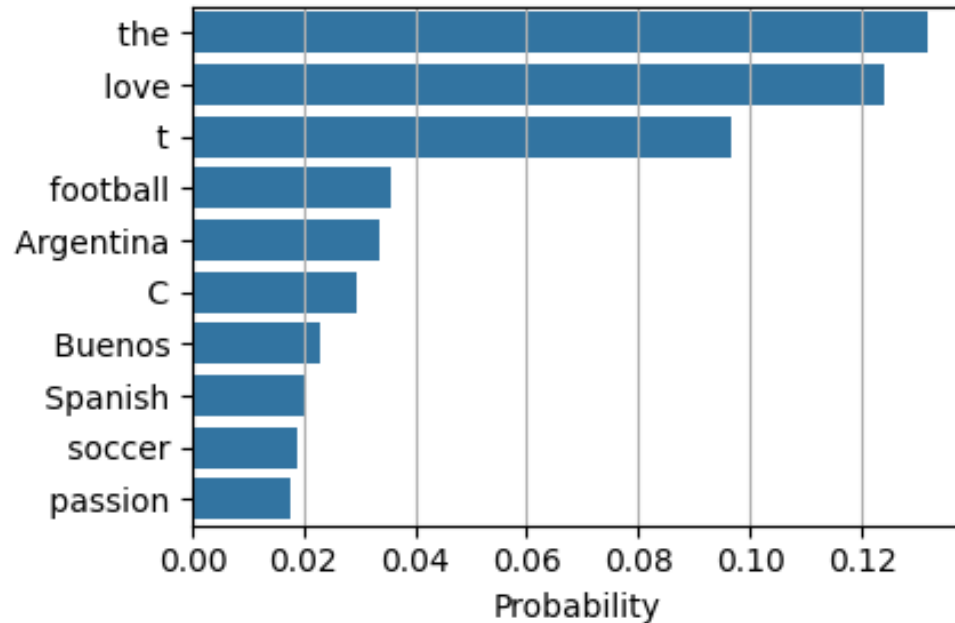
Yoran et al. (2024):

- Problem: irrelevant context.
- Propose to remove irrelevant context with external tools: NLI models.
- Relevancy is not entailment!

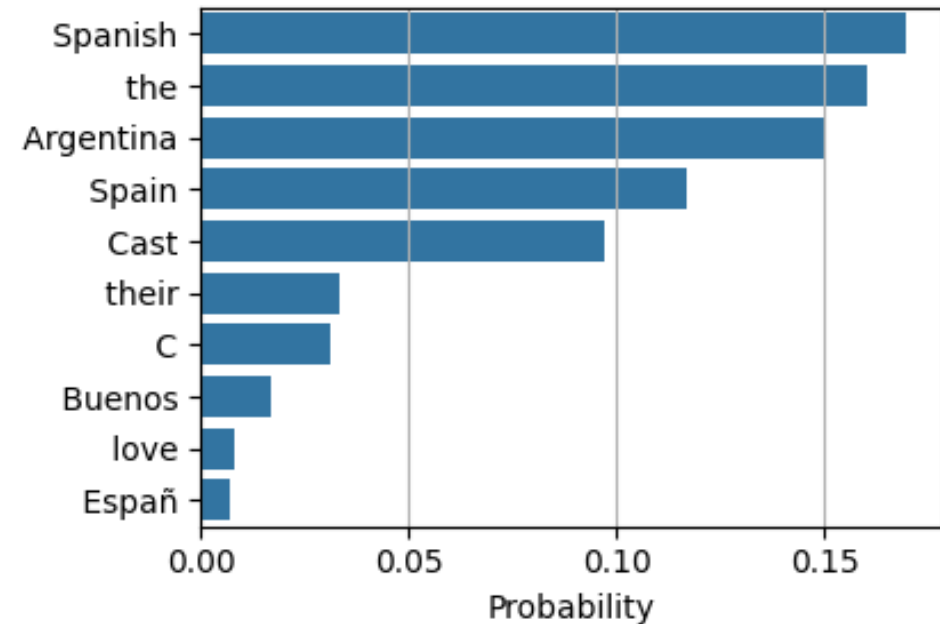
“Irrelevant” Context Can Still Be Helpful

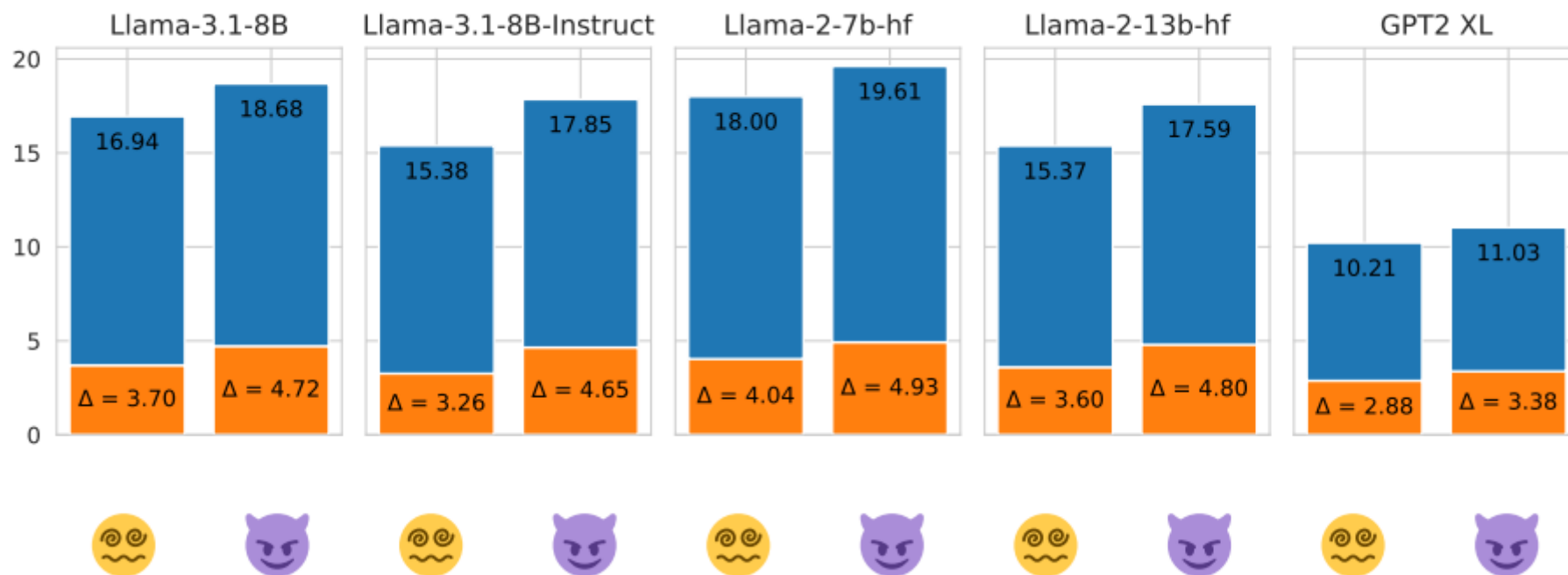
All probs in slides from Llama-3.1 8B.
Results generalizable to other LM families.

In Argentina, people speak the
language of ____



In **Austria**, the primary language
is **German**. In Argentina, people
speak the language of ____

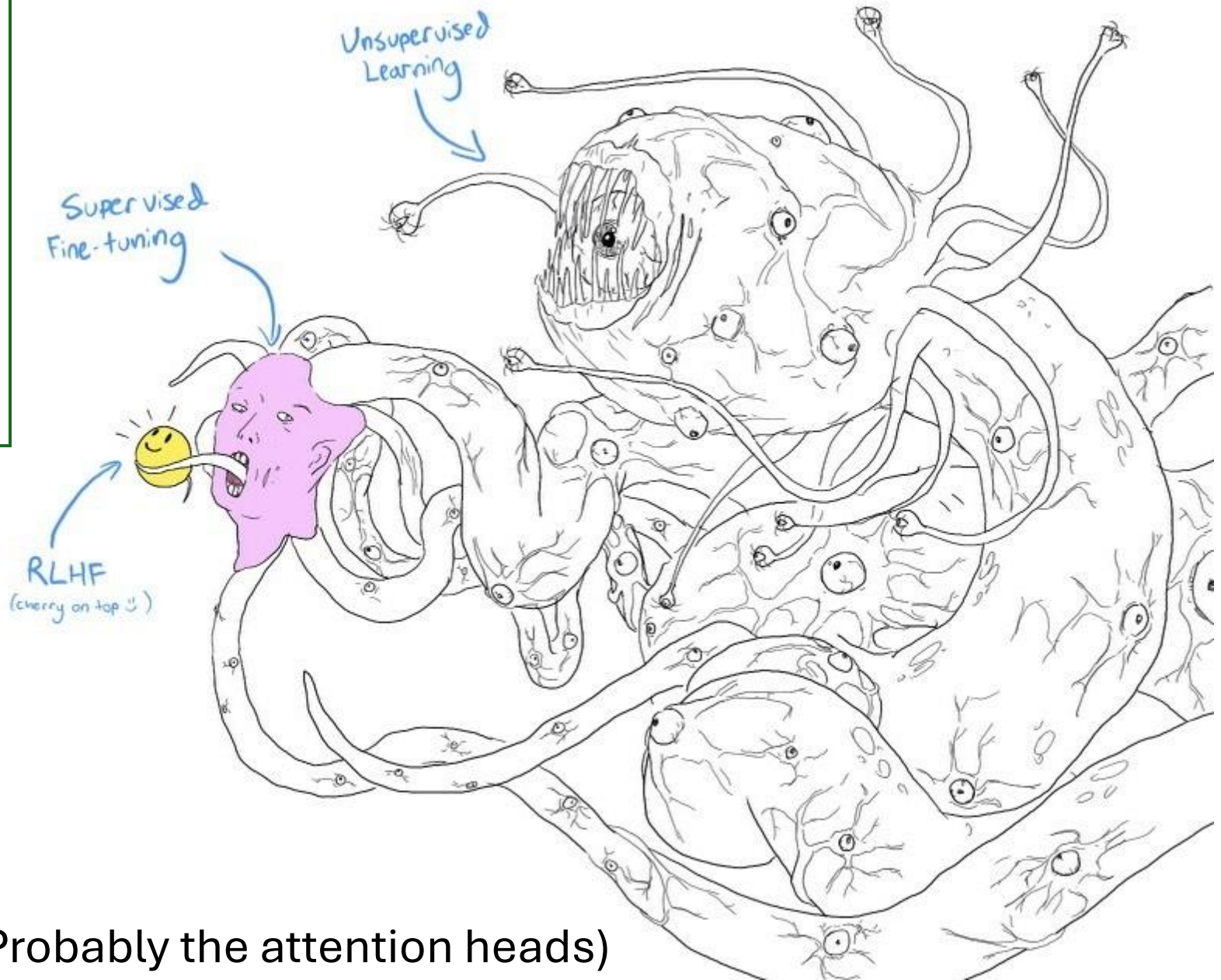




👹: Japan is in **Africa**. Greece is located on the continent of

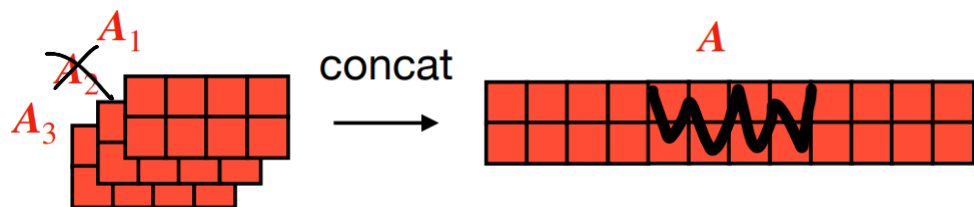
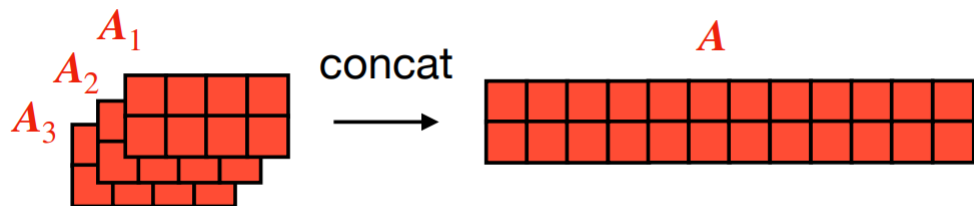
Counterfactual (👹) context prompts consistently cause **greater distraction** than factual context prompts!

Where in the model is “doing the distraction?”



(Probably the attention heads)

“Cutting off” the heads to understand their functions.

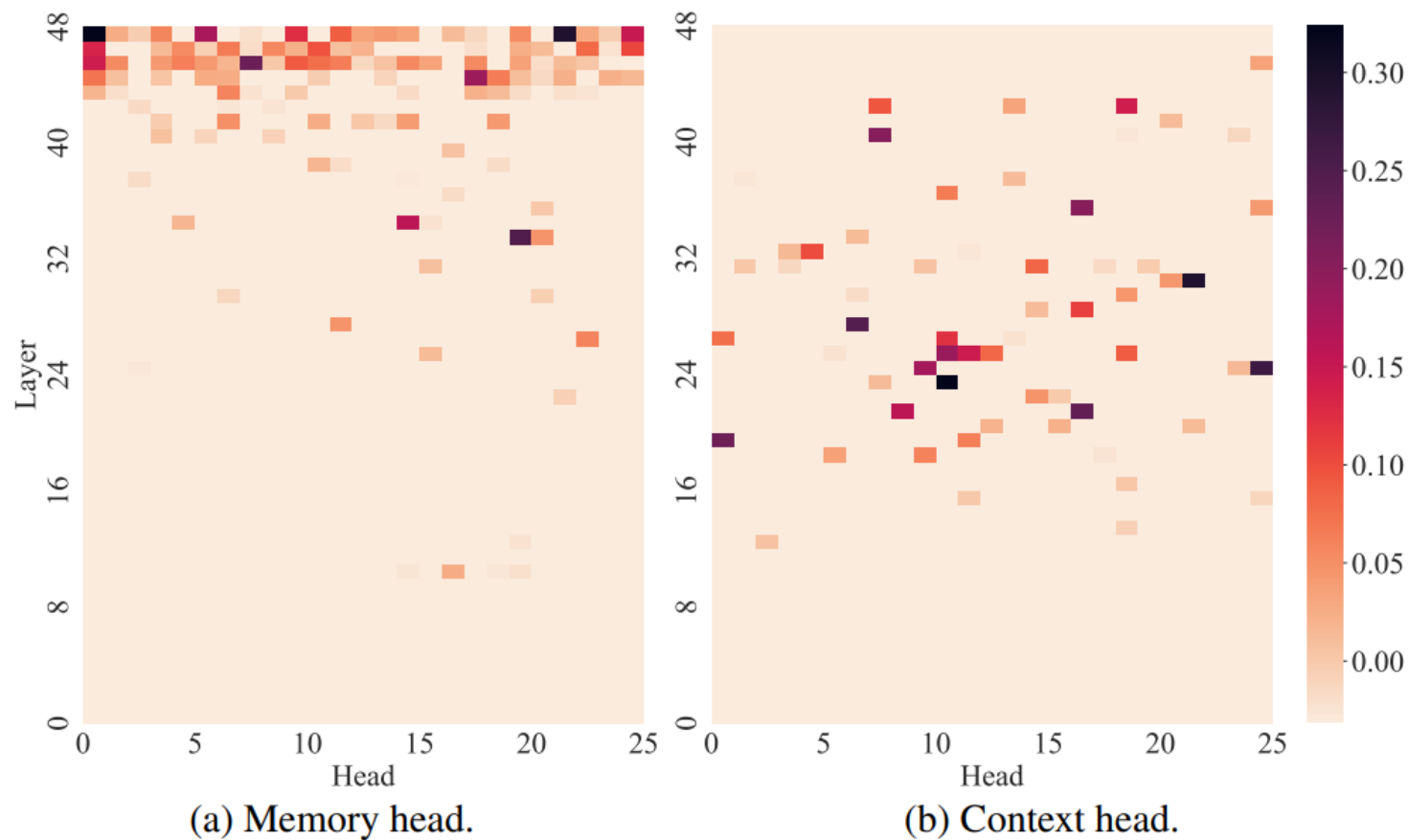


Cutting the head off:

- Setting the output to 0.
- See how the model behaves.

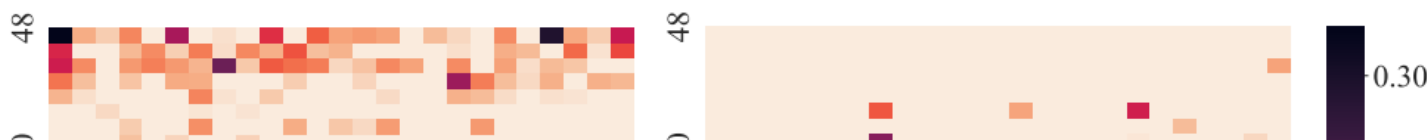


Previous Work: **Attention Head** is the Key 🗝️



Jin et al. (2024): “Cutting off” different attention heads have different effects. Some heads predict based on “internal memory” (memory head) and some heads predict based on the context (context heads).

Previous Work: **Attention Head** is the Key 🗝️



Issue: You can only analyse by cutting off one head at a time!

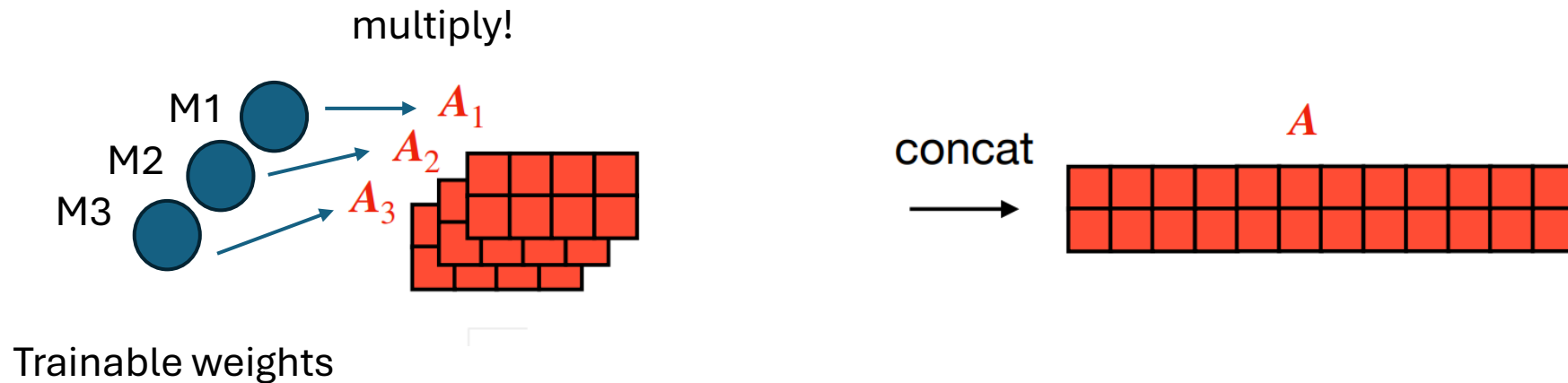
- Attention heads are not isolated.
- Some combinations of attention heads may have unexpected effect.
- However, $O(n!)$ Search time

Head
(a) Memory head.

Head
(b) Context head.

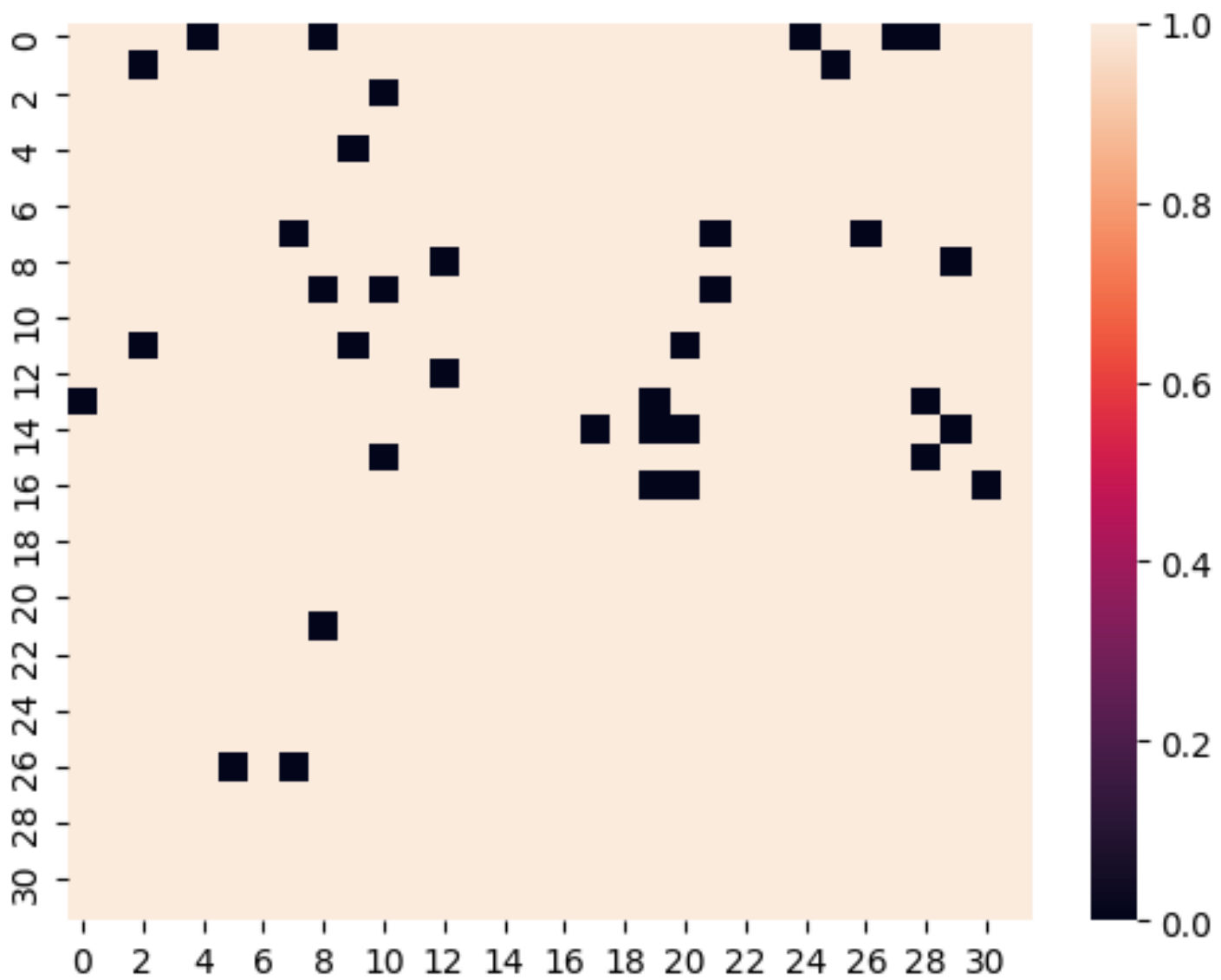
Jin et al. (2024): “Cutting off” different attention heads have different effects. Some heads predict based on “internal memory” (memory head) and some heads predict based on the context (context heads).

Differentiable Mask

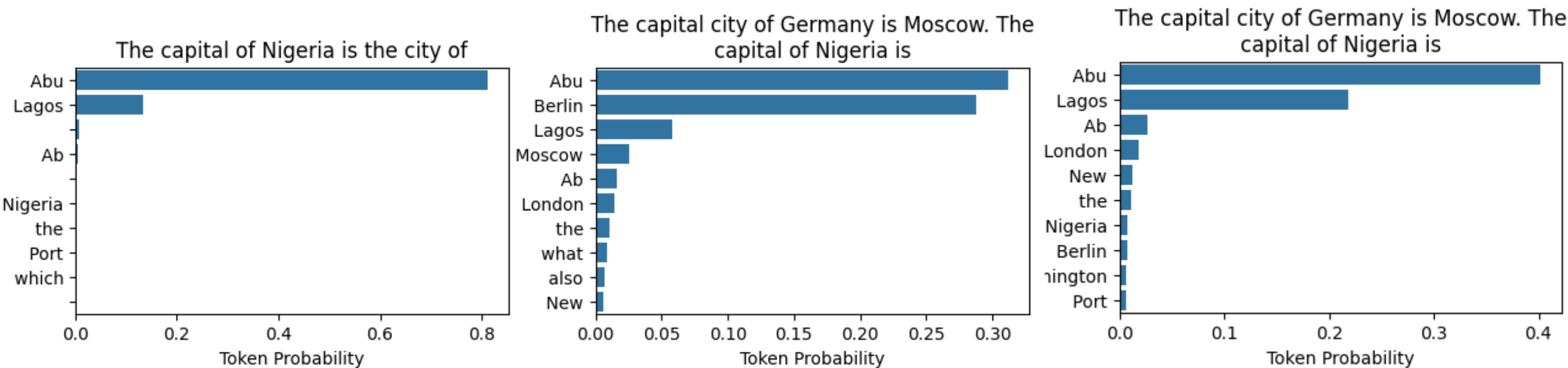


Make the masks differentiable and binary:
Straight-through estimator (STE) + Gumbel Softmax

We can “train a model” to identify distracting heads!
(Adapted from DiscoGP (Yu et al., 2024))



Example:
36 attention heads for
the *country capital city*
relation.



Entrainment Heads  Contextual Entrainment

A Mechanistic Perspective on LLM Distraction



PAPER



CODE



BLOG

- LLMs have amazing capabilities at using context provided in prompts.
 - RAG, ICL, CoT, instruction tuning...
- Previous analysis of distraction:
 - Relevancy, distracting...
- After all, LLM is still a statistical model, and there is a way to understand how context is utilised.
- A more rational, mechanistic understanding of LLMs → Better control!

