Llama See, Llama Do: A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs

ACL 2025 Outstanding Paper

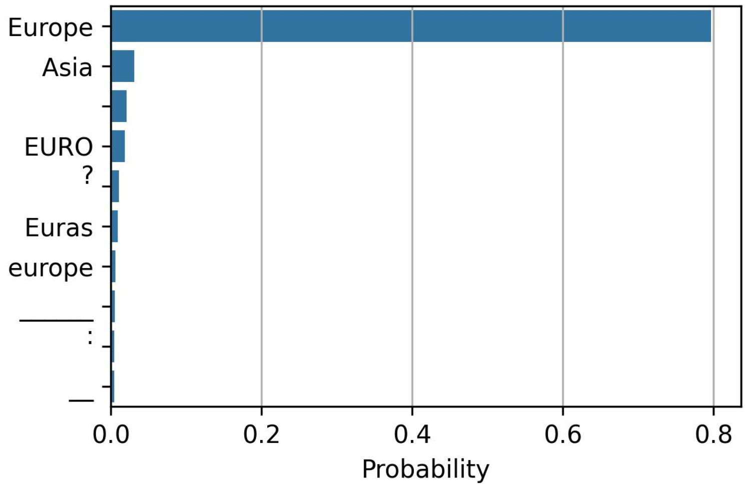
2025.10.14

Contextual entrainment, a new mechanistic perspective on how LMs become distracted by “irrelevant” contextual information in the input prompt.

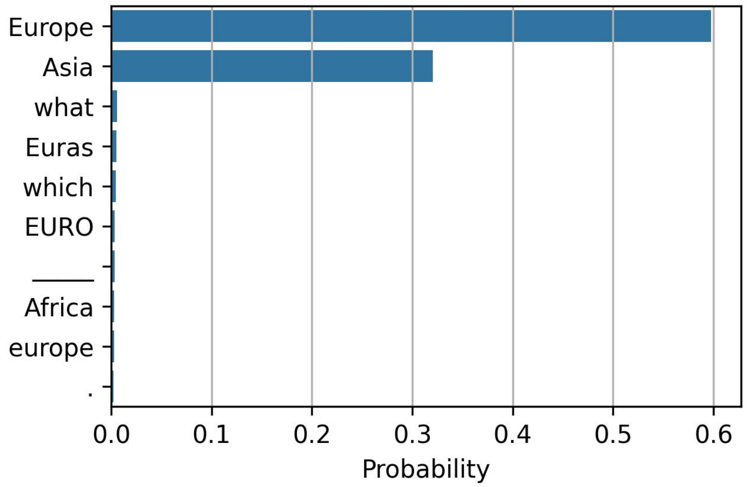
LMs assign significantly higher logits (or probabilities) to any tokens that have previously appeared in the context prompt, even for random tokens.

Background: Distraction

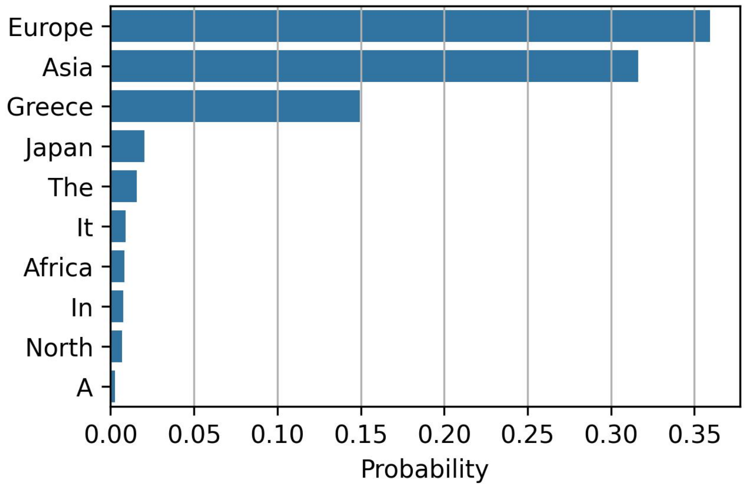
LMs are susceptible to distractions caused by contextual information in prompts.



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

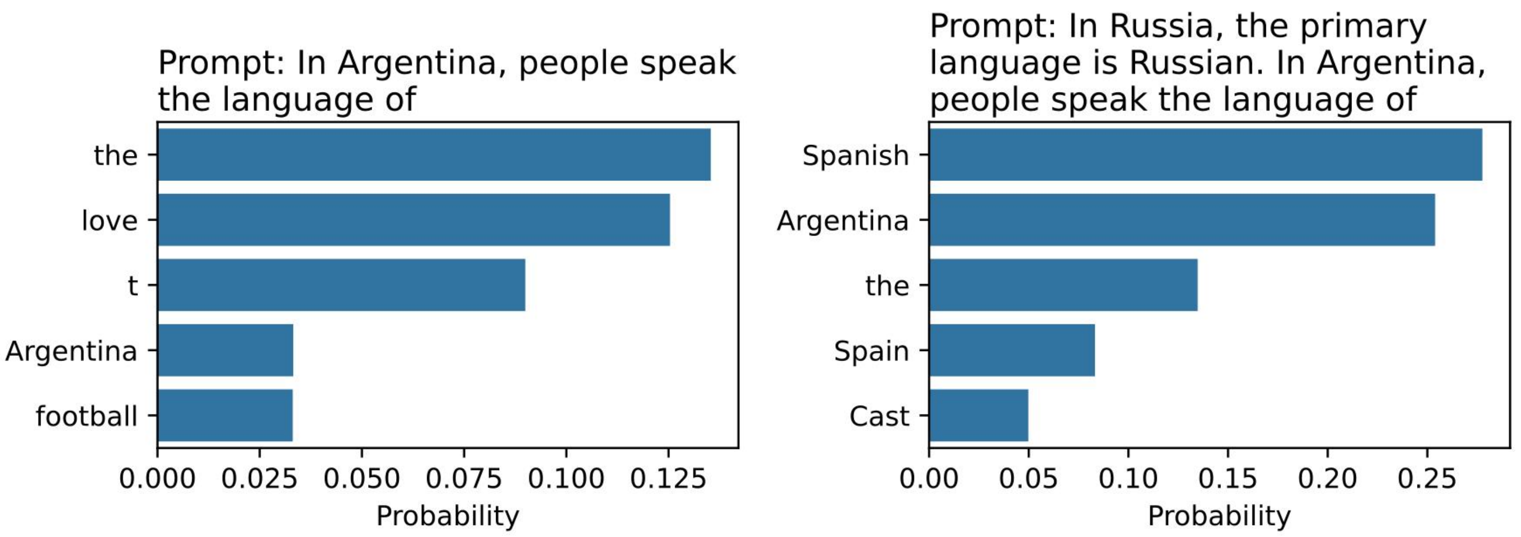


CONTEXT: Japan is in Asia.
QUERY: Greece is located on the continent of

Background: Distraction

Most prior work defines distraction using the term “(ir)relevant,” framed in RAG and information retrieval terms; i.e., whether the context prompt contains the information needed to answer the question correctly.

This narrow definition is not enough. “Irrelevant” context can still be helpful.



Contextual Entrainment

4 Context Prompt Settings:

- Related: The capital of Peru is Lima. Canada’s capital is Ottawa/Lima
- Irrelevant: Banana are yellow. Canada’s capital is Ottawa/yellow
- Random: Promotion. Canada’s capital is Ottawa/Promotion
- Counterfactual: The capital of Peru is Vienna. Canada’s capital is Ottawa/Vienna

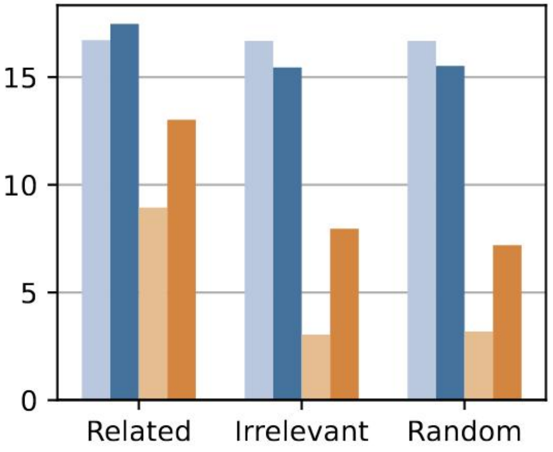
LRE Dataset: 15 types of factual relations (source, relation, target), run experiments on 100,000 examples.

Contextual Entrainment

$\overbrace{\texttt{What is the capital of Canada? It is the city of}}^{\text{Query Prompt}}$ $\begin{cases} \ell(\texttt{Ottawa})\text{: 🙂 without context} \\ \ell(\texttt{Lima})\text{: 😵‍💫 without context} \end{cases}$

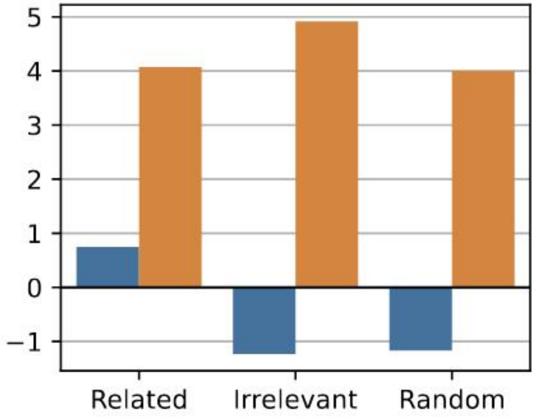
$\overbrace{\texttt{The capital of Peru is Lima.}}^{\text{Context Prompt}}$ $\overbrace{\texttt{What is the capital of Canada? It is the city of}}^{\text{Query Prompt}}$ $\begin{cases} \ell(\texttt{Ottawa})\text{: 🙂 with context} \\ \ell(\texttt{Lima})\text{: 😵‍💫 with context} \end{cases}$

The logits of the correct tokens increase in the relevant setting and decrease in the irrelevant and random settings, while the logits of the distracting tokens increase across all settings.



The bar height indicates the average logits of:

🟦: 🙂 with context 🟦: 🙂 without context
🟧: 😵‍💫 with context 🟧: 😵‍💫 without context



The bar height indicates the difference in **logits** with or without the distracting context across tokens:

🟦: Δ_ℓ(🙂) = ℓ(🙂 w/ ctx.) − ℓ(🙂 w/o ctx.)
🟧: Δ_ℓ(😵‍💫) = ℓ(😵‍💫 w/ ctx.) − ℓ(😵‍💫 w/o ctx.)

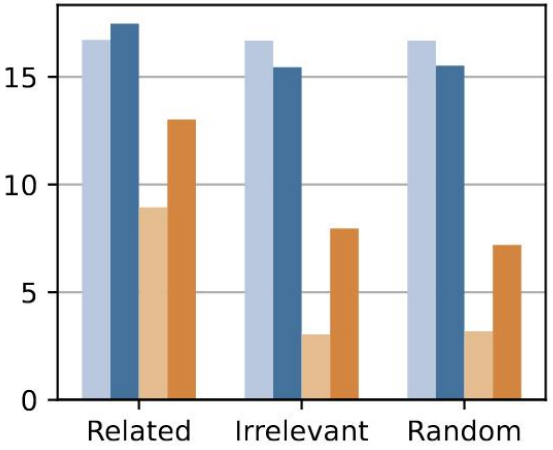
Contextual Entrainment

Contextual Entrainment-**A Novel Mechanistic Phenomenon**:
LMs assign higher logits and probabilities to tokens that appear in the context.

$\overbrace{\texttt{What is the capital of Canada? It is the city of}}^{\text{Query Prompt}}$ $\begin{cases} \ell(\texttt{Ottawa}): 🙂 \text{ without context} \\ \ell(\texttt{Lima}): 😵‍💫 \text{ without context} \end{cases}$

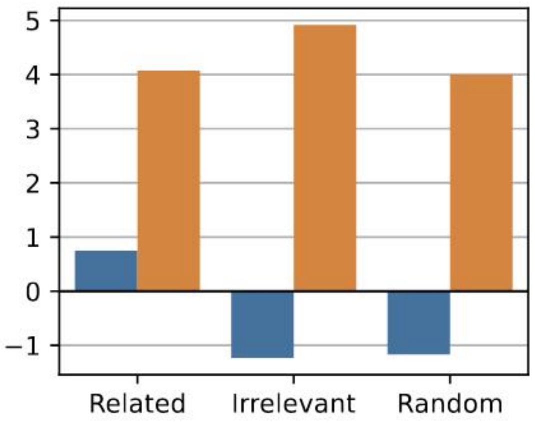
$\overbrace{\texttt{The capital of Peru is Lima.}}^{\text{Context Prompt}}$ $\overbrace{\texttt{What is the capital of Canada? It is the city of}}^{\text{Query Prompt}}$ $\begin{cases} \ell(\texttt{Ottawa}): 🙂 \text{ with context} \\ \ell(\texttt{Lima}): 😵‍💫 \text{ with context} \end{cases}$

The logits of the correct tokens increase in the relevant setting and decrease in the irrelevant and random settings, while the logits of the distracting tokens increase across all settings.



The bar height indicates the average logits of:

▨: 🙂 with context ▨: 🙂 without context
▨: 😵‍💫 with context ▨: 😵‍💫 without context



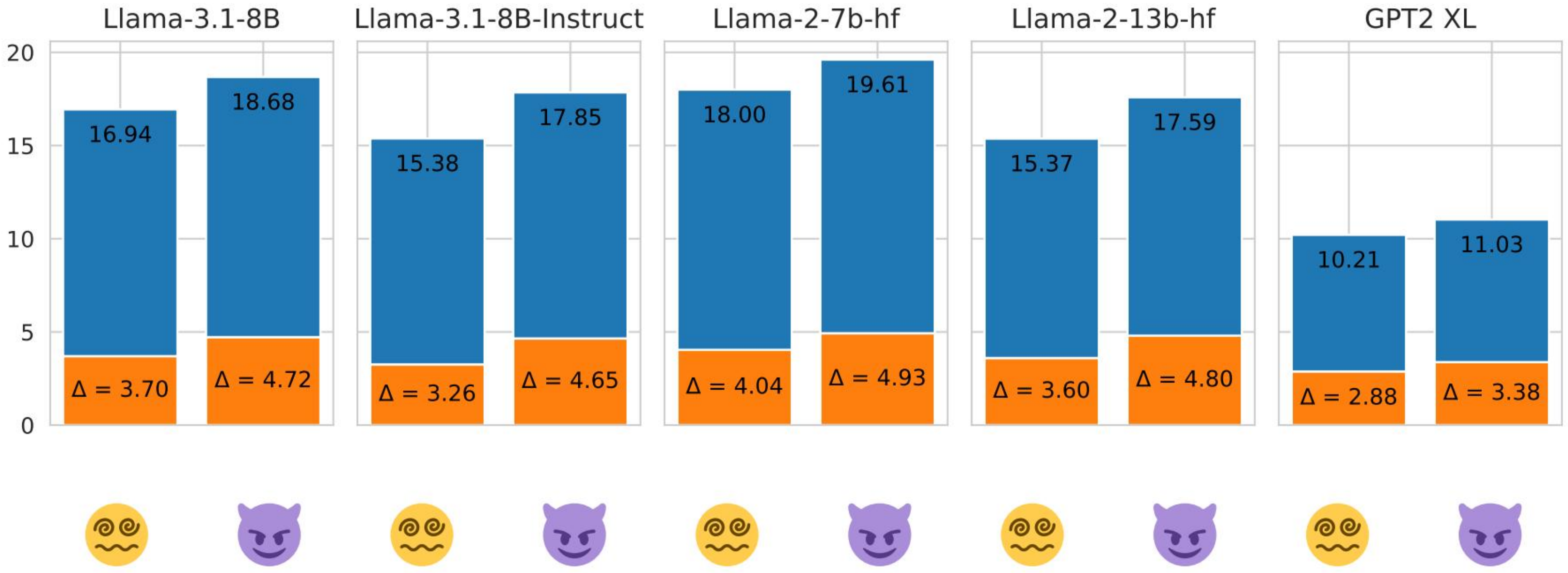
The bar height indicates the difference in **logits** with or without the distracting context across tokens:

▨: Δ_ℓ(🙂) = ℓ(🙂 w/ ctx.) − ℓ(🙂 w/o ctx.)
▨: Δ_ℓ(😵‍💫) = ℓ(😵‍💫 w/ ctx.) − ℓ(😵‍💫 w/o ctx.)

Contextual Entrainment

Counterfactual context prompts consistently cause stronger distraction than factual context prompts.

The absolute logits of the token when factual prompts are provided are significantly lower than those of the token when counterfactual prompts are provided.

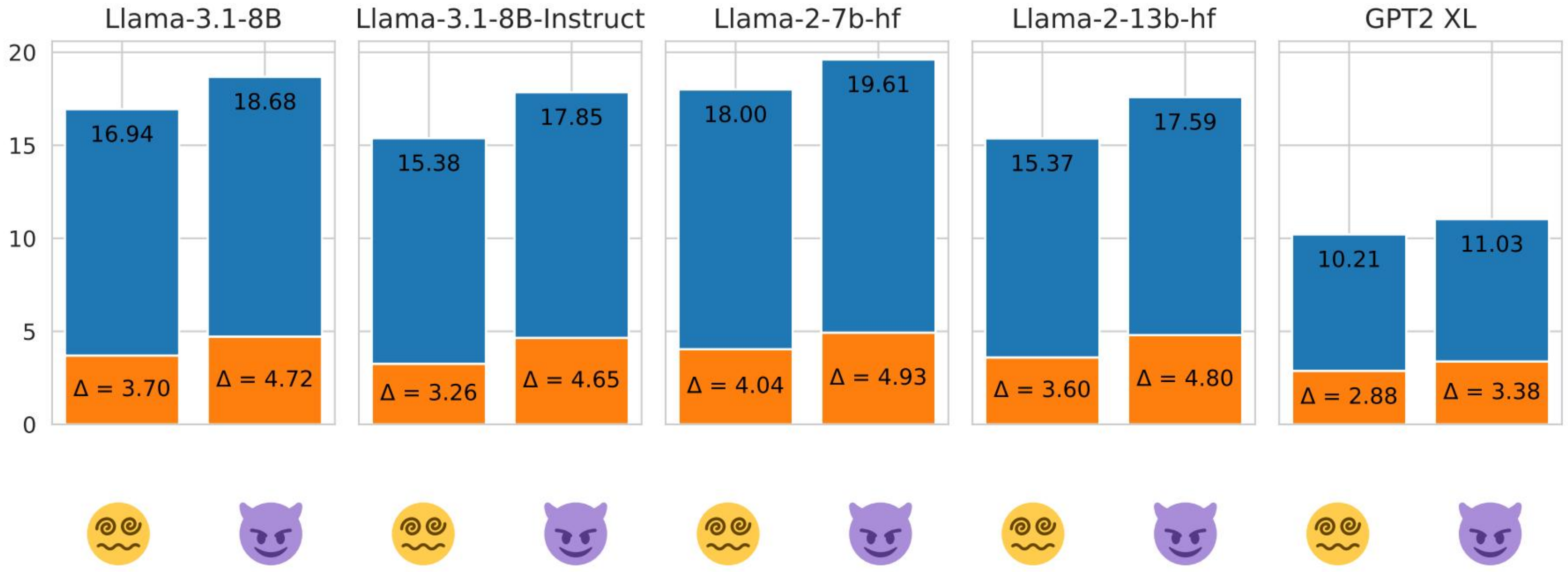


Factual context: Japan is in Asia ( ); Counterfactual context: Japan is in Africa ( )
Query: Greece is located in __

Contextual Entrainment

Contextual Entrainment: **A mechanistic phenomenon affected by semantic factors**.

While we previously established that contextual entrainment is a “mechanistic” phenomenon, it is still subject to semantic factors in determining its magnitude of impact.



Factual context: Japan is in Asia ( ); Counterfactual context: Japan is in Africa ( )
Query: Greece is located in __

Contextual Entrainment

Contextual Entrainment: **A mechanistic phenomenon affected by semantic factors**.

Compared to the inductive literal sequence copying phenomenon (induction heads):

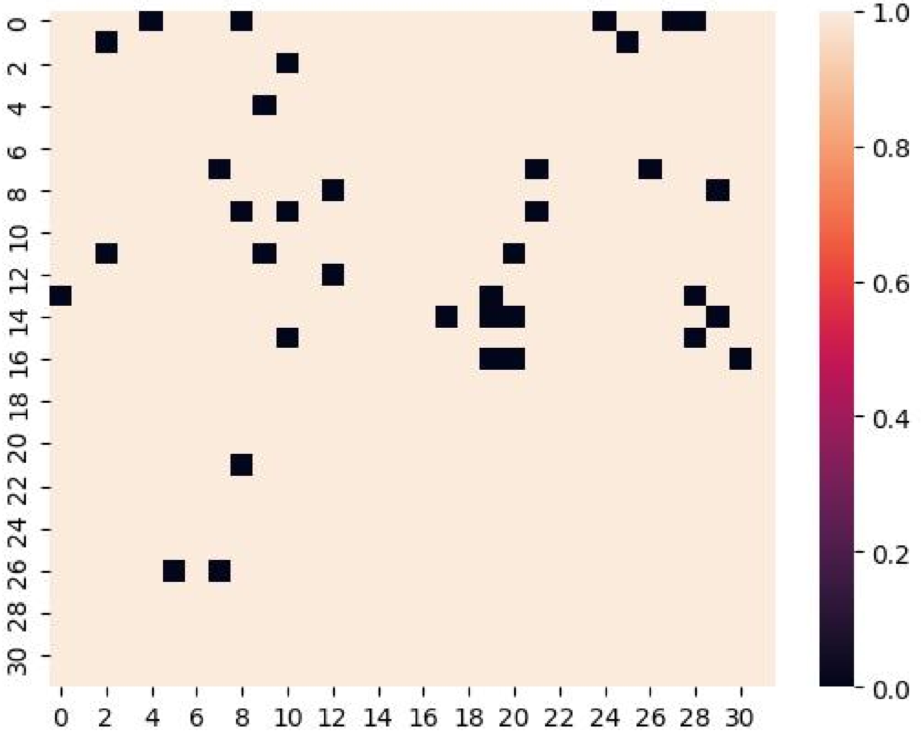
- Sequence copying requires the reappearance of a prefix as a trigger; but contextual entrainment occurs when a token has previously appeared in the context.

- Sequence copying is largely independent of semantic factors and token statistics; but the magnitude of contextual entrainment is influenced by semantic factors (particularly counterfactual prompts).

Entrainment Heads

Entrainment Heads: a set of attention heads that corresponds to the contextual entrainment phenomenon.

Identify the optimal combinations of attention heads to disable in order to suppress contextual entrainment.



The 36 identified entrainment heads for country-capital city

Differentiable Mask:
make the mask differentiable and binary

$$\sum_{h_j \in H_i} m_j h_j(x_i)$$

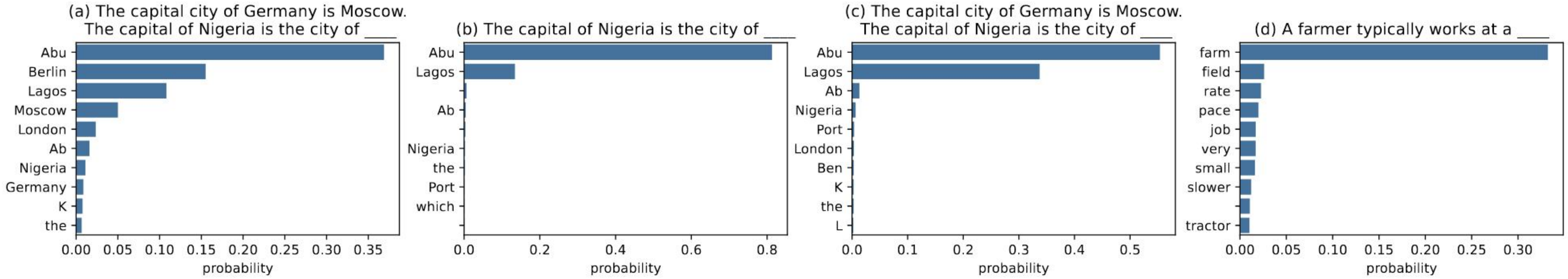
Train a model to identify the distracting heads:

$$s_i = \sigma\left(\frac{l_i - \log\frac{\log \mathcal{U}_1}{\log \mathcal{U}_2}}{\tau}\right); m_i = [\mathbb{1}_{s_i > \frac{1}{2}} - s_i]_{\text{detach}} + s_i,$$

$$\mathcal{L} = \underbrace{\ell(\text{🙂}) - \ell(\text{🥴})}_{\text{Logits } \Delta} + \underbrace{\lambda \cdot \frac{1}{|H|} \sum_{i=1}^{|H|} \sigma(l_i)}_{\text{Sparsity Loss}} .$$

Entrainment Heads

Turning off the entrainment heads **drastically reduce contextual entrainment**.



Original Model 🤖

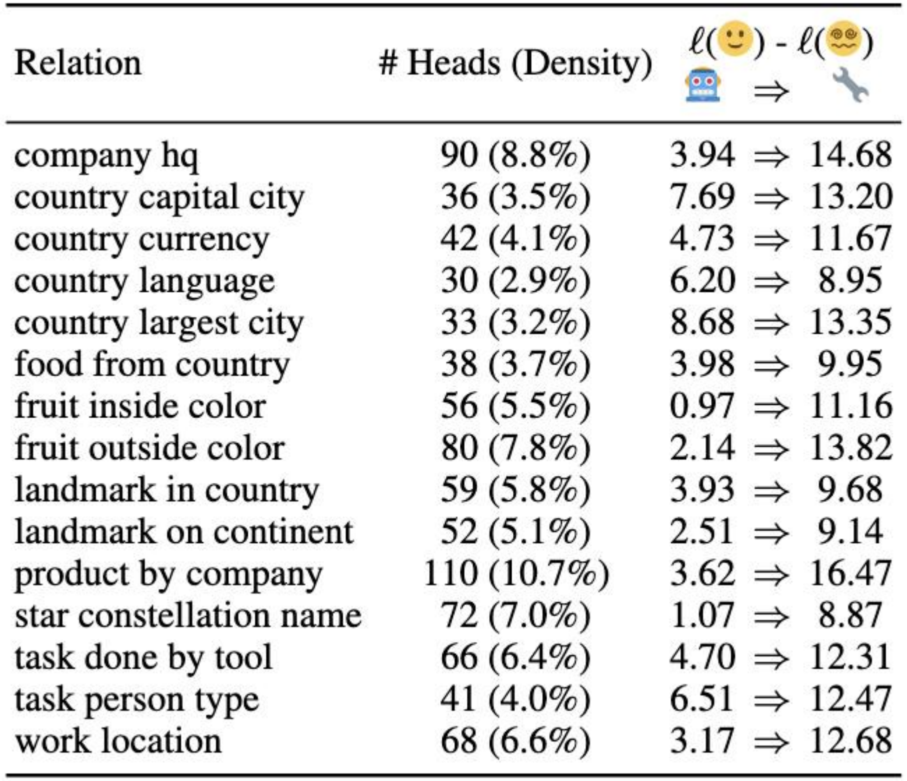
Model with Entrainment Head Removed 🔧

Removing the entrainment heads caused a significant effect across logits delta and the ranks of the tokens, making them capitulate the situation when no distracting context is provided.

Measure	🤖 No 😵‍💫	🤖 With 😵‍💫	🔧 No 😵‍💫	🔧 With 😵‍💫
ℓ(🙂)	19.51	20.68	19.49	21.21
ℓ(😵‍💫)	8.75	12.99	7.87	8.01
Δ = ℓ(😵‍💫) - ℓ(🙂)	10.76	7.69	11.62	13.20
Avg. 🙂 Token Rank	1.00	1.00	1.00	1.00
Avg. 😵‍💫 Token Rank*	1756.7	37.5	1707.3	1289.6

Entrainment Heads

Entrainment heads are **task-specific (or relation-specific)**, **not model-specific**.



Relation	# Heads (Density)	ℓ(🙂) - ℓ(😵‍💫) 🤖 ⇒ 🔧
company hq	90 (8.8%)	3.94 ⇒ 14.68
country capital city	36 (3.5%)	7.69 ⇒ 13.20
country currency	42 (4.1%)	4.73 ⇒ 11.67
country language	30 (2.9%)	6.20 ⇒ 8.95
country largest city	33 (3.2%)	8.68 ⇒ 13.35
food from country	38 (3.7%)	3.98 ⇒ 9.95
fruit inside color	56 (5.5%)	0.97 ⇒ 11.16
fruit outside color	80 (7.8%)	2.14 ⇒ 13.82
landmark in country	59 (5.8%)	3.93 ⇒ 9.68
landmark on continent	52 (5.1%)	2.51 ⇒ 9.14
product by company	110 (10.7%)	3.62 ⇒ 16.47
star constellation name	72 (7.0%)	1.07 ⇒ 8.87
task done by tool	66 (6.4%)	4.70 ⇒ 12.31
task person type	41 (4.0%)	6.51 ⇒ 12.47
work location	68 (6.6%)	3.17 ⇒ 12.68

- The method has identified different and different amount of entrainment heads for each LRE relation.

- A small set of attention head (2.9~10.7%) can substantially increase the gap between the logits of and tokens

- There is some degree of crossrelation overlap in entrainment heads, it is still a relation-, task-, or domain-specific effect.

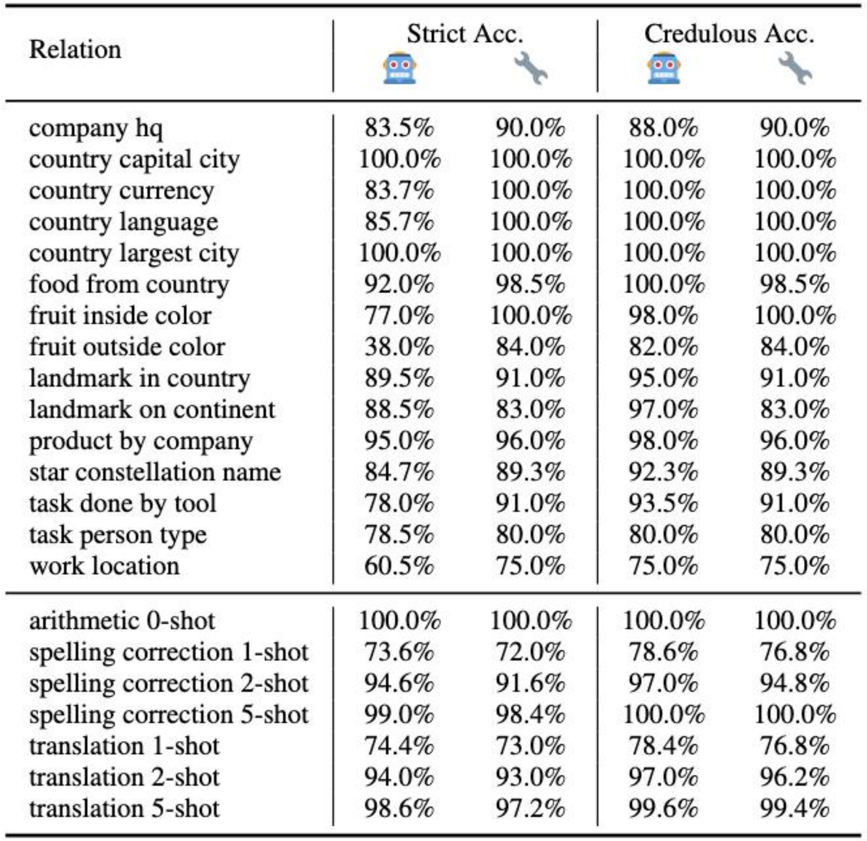
  Removing the entrainment heads for country-capital relation causes a similar effect in other relations, but the effect is less consistent and smaller.

Entrainment Heads

Removing the entrainment heads has **only a small effect on other LM capabilities**.

- Removing the entrainment heads of the country-capital city relation has a negligible effect on the LM's performance across other relations.

- After removing the entrainment heads, the model exhibits only a small performance decrease (0.2~3%) and continues demonstrate strong ICL capabilities with high accuracy.



Relation	Strict Acc.		Credulous Acc.	
	🤖	🔧	🤖	🔧
company hq	83.5%	90.0%	88.0%	90.0%
country capital city	100.0%	100.0%	100.0%	100.0%
country currency	83.7%	100.0%	100.0%	100.0%
country language	85.7%	100.0%	100.0%	100.0%
country largest city	100.0%	100.0%	100.0%	100.0%
food from country	92.0%	98.5%	100.0%	98.5%
fruit inside color	77.0%	100.0%	98.0%	100.0%
fruit outside color	38.0%	84.0%	82.0%	84.0%
landmark in country	89.5%	91.0%	95.0%	91.0%
landmark on continent	88.5%	83.0%	97.0%	83.0%
product by company	95.0%	96.0%	98.0%	96.0%
star constellation name	84.7%	89.3%	92.3%	89.3%
task done by tool	78.0%	91.0%	93.5%	91.0%
task person type	78.5%	80.0%	80.0%	80.0%
work location	60.5%	75.0%	75.0%	75.0%
arithmetic 0-shot	100.0%	100.0%	100.0%	100.0%
spelling correction 1-shot	73.6%	72.0%	78.6%	76.8%
spelling correction 2-shot	94.6%	91.6%	97.0%	94.8%
spelling correction 5-shot	99.0%	98.4%	100.0%	100.0%
translation 1-shot	74.4%	73.0%	78.4%	76.8%
translation 2-shot	94.0%	93.0%	97.0%	96.2%
translation 5-shot	98.6%	97.2%	99.6%	99.4%

The strict (answer in top-3) and credulous (answer in top-10) accuracy of the original model ( ) and the model with country-capital city entrainment heads removed ( ).

Summary

Llama See, Llama Do

- **Contextual Entrainment**: LLMs give higher probability to tokens that previously appeared in the prompt, even if those tokens are random or irrelevant.

- **Counterfactual context prompts** can exacerbate the the entrainment.

- **Entrainment Heads**: We can identify a set of attention heads (the entrainment heads) that are corresponds to the entrainment of a given pattern, using a differentiable-masking-based method.