On the Generalizability of "Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals"

Asen Dotsinski*, Udit Thakur*, Marko Ivanov, Mohammad Hafeez Khan, Maria Heuss

* Equal Contribution

University of Amsterdam

Introduction

- This study reproduces and extends Ortu et al. (2024) [1] on how language models balance factual recall and counterfactual context.
- We confirm key findings on mechanism localization and attention head roles in GPT-2 and Pythia 6.9B, and extend to Llama 3.1 8B.
- We show how prompt phrasing and domain influence model preferences, revealing variability across settings.
- Our work clarifies competing mechanisms in LLMs and the limits of prior conclusions.

Misalignment Between Recall and Contextual Override

Redefine: iPhone was developed by Google.
iPhone was developed by

Mechanism 1: Factual Knowledge Recall
Recalling that Apple, not Google, developed the iPhone.

Mechanism 2: Counterfactual Alignment
Adapting to the prompt's statement that Google developed the iPhone.

Google

LLMs may mistakenly rely on Mechanism 2 rather than Mechanism 1.

Methodology

Model Implementation

GPT-2 Small (Original) | Pythia 6.9B (Original) | Llama 3.1 8B (Extension)

TransformerLens [2] Library (Mechanistic Interpretation)

Mechanistic Analysis

Analysis Methods

Attention Modification

$$A_{ij}^{hl} \leftarrow \alpha \cdot A_{ij}^{hl}, \text{ where } j < i$$

Logit Inspection

$$\tilde{x}_i^l = x_i^l \cdot W_U$$

Dataset Processing

COUNTERFACT Dataset

Filtering & Processing

"Redefine" Dataset
"Redefine: {s} {r} {tcofa}. {s} {r}"

"Redefine" Dataset
"Redefine: Iphone is developed by Google.
Iphone is developed by …"

"Domain" Dataset
26 categories such as 'Sports', 'Finance', and 'News'.

"QnA" Dataset
"Redefine: {s} {r} {tcofa}. {interrogative} {rr} {s} Answer: "

"Redefine" Dataset
"Redefine: Iphone is developed by Google.
What company developed Iphone? Answer:"

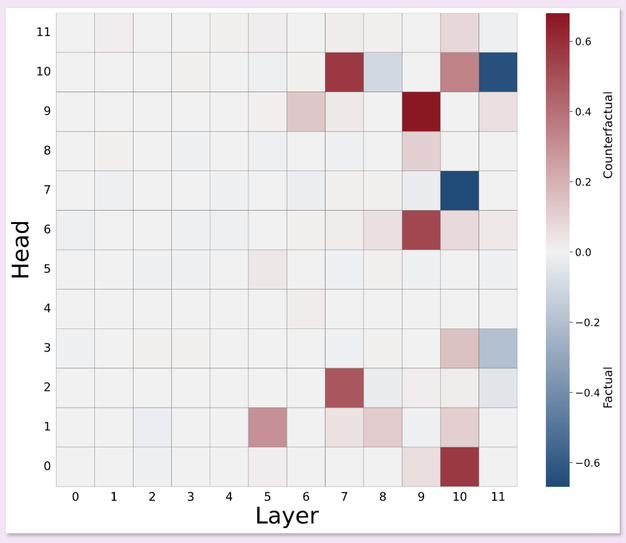
"Premises" Dataset
"Keyword: {s} {r} {tcofa}. {s} {r}"
e.g., "Fact Check", "Validate", "Verify"

{s}: subject (entity) {r}: relation (predicate) {rr}: reformulated relation in question form
{tcofa}: counterfactual attribute (false token) {interrogative}: question word (e.g., Who, What, Where)

- **TransformerLens** [2] logit inspection using unembedding matrices.
- **Attention head ablation** with scaling factors (α = 5, 50)
- **COUNTERFACT** [3] dataset augmented with "Redefine" prompts
- **Format:** "Redefine: {subject} {relation} {counterfactual}. {subject} {relation}"
- **Models:** GPT-2 small, Pythia 6.9B, Llama 3.1 8B

Key Findings of Ortu et al. (2024)

- There is an inherent competition of two mechanisms in LLMs
  - **Factual:** recall of information from training
  - **Counterfactual:** in-context learning
- Few specialized attention heads responsible for the outcome of the competition.
- Ablating the attention heads (e.g. by multiplying their attention scores by α=5 or 50) significantly increases the probability of a factual outcome.



Theoretically, this could be used for RAGs or against adversarial prompts, since we could control the focus on the pretraining vs the context

Reproducibility and Extensions

Reproduced (Ortu et al., 2024) ✓

Positional Information Encoding: Both mechanisms encode info in the last token in deeper layers. GPT-2 encodes factuals early at the subject, counterfactuals at the attribute; Pythia does not encode logits until later layers.

Attention Block Dominance: GPT-2's deeper attention blocks (layers 5–11) strongly drive competition; Pythia shows a more distributed, weaker pattern.

Head Specialization: A small set of later-layer attention heads in GPT-2 dominate factual/counterfactual decisions by suppressing counterfactuals; Pythia shows broader specialization due to scale.

Extensions (Ours) ★

Model Scaling: Testing on Llama 3.1 8B to assess size-related effects.

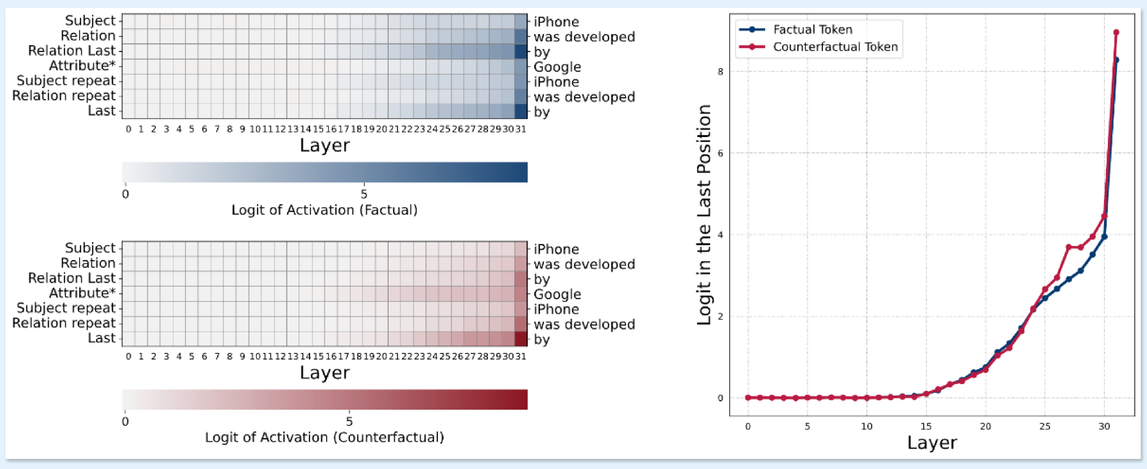
Prompt Structure Variation: Exploring how format impacts counterfactual recall.

Premise Manipulation: Changing trigger words and ablating key attention heads.

Domain Generalization: Evaluating robustness across diverse prompt topics.

Major Findings

Results for Llama 3.1 8B



Mechanism in Llama 3.1 8B

- Late-layer logit buildup; early layers have minimal impact
- Final layer dominates predictions
- Key head ablation has limited effect
- Suggests reduced specialization or logit lens limitations

Key Insights

Impact of Prompt Format

Premise	Baseline			Ablated		
	#Factual	#Counterfact	%Factual	#Factual	#Counterfact	%Factual
Redefine	304	5794	4.98%	2722	3177	46.15%
Assess	388	5688	6.38%	2843	2945	49.13%
Fact Check	206	5861	3.39%	1883	3980	32.12%
Review	116	5986	**1.90%**	1824	4093	**30.83%**
Validate	380	5676	6.28%	2888	2870	50.16%
Verify	259	5818	4.26%	2487	3332	42.74%
Redefine, QnA prompt structure	2158	2171	**49.8%**	2751	175	**94.02%***

Question-Answer Format

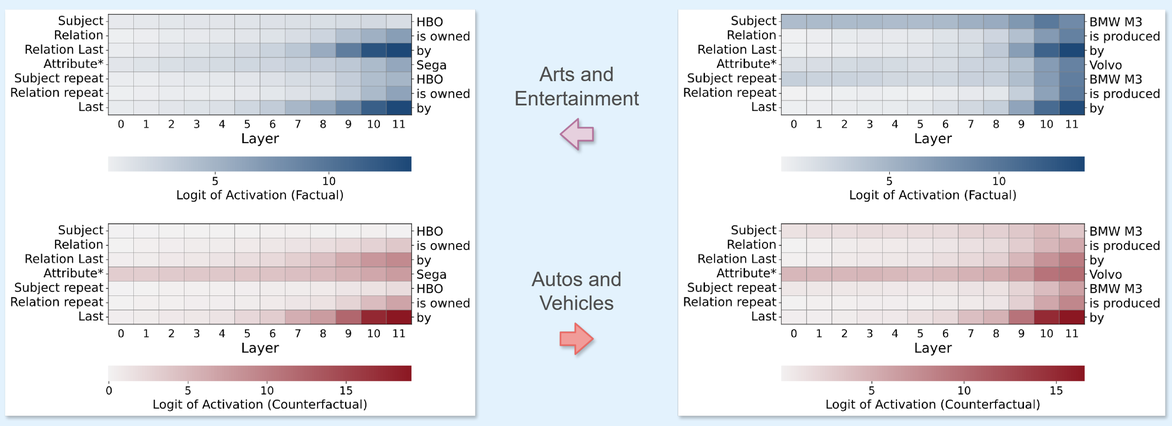
- QnA reduces counterfactuals in GPT-2
- Boosts factual outputs (~50%). Ablation boosts it further.

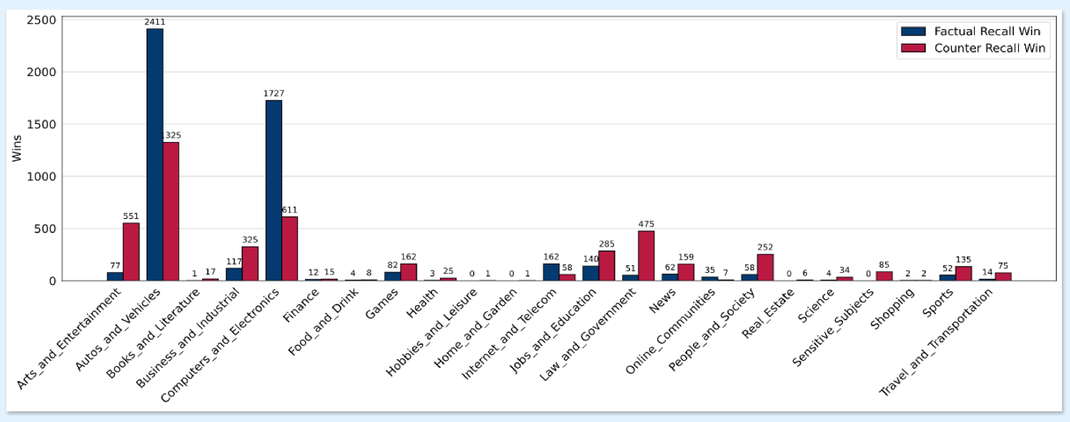
Impact of Premise Words

- Premise choice affects copy mechanism strength.
- "Assess"/"Validate" yield higher factuality than "Review"/"Fact Check".
- Ablation increases this effect.

Key Insights

Impact of Query Domain





Domain Variance
Logit patterns vary by domain. Familiar domains match expected trends; unfamiliar ones deviate, implying reliance on domain familiarity.

Key Insights

Domain-Specific Ablation Effects
Ablating key heads boosts factual recall in some domains (e.g., "Autos and Vehicles") but not others, showing domain-specific attention specialization.

[1] Ortu, F., Jin, Z., Doimo, D., Sachan, M., Cazzaniga, A., & Schölkopf, B. (2024). Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals. arXiv [Cs.CL].
[2] Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/ TransformerLens, 2022.
[3] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. Advances in Neural Information Processing Systems, 35:17359–17372, 2022.