Topic: Analyzing In-Context Learning in Language Models Using Counterfactuals

Major Project :- ELD880

Supervisor:- Prof. Sougata Mukherjea

Name:- Animesh Lohar

Entry No:- 2024EET2368

M.Tech in Computer Technology

Introduction:-

The Dual Nature of Language Models

In Context Learning (ICL)

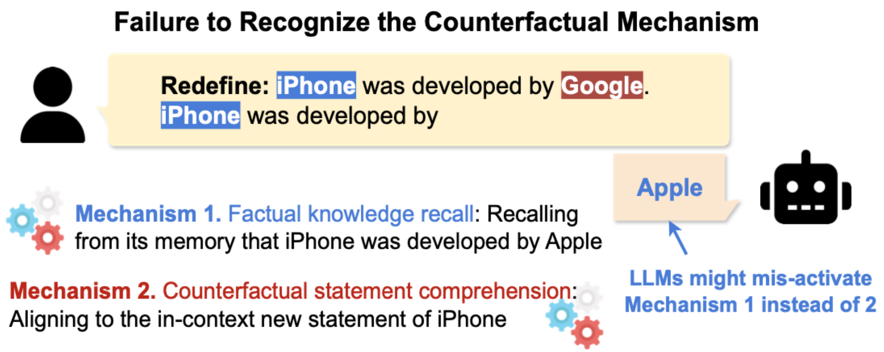
- ➢ Learn from Examples in the Prompt
- ➢ Adapt to New Tasks Immediately
- ➢ Follow Contextual Instructions

Pretrained Memory

- ➢ Vast World Knowledge from Training
- ➢ Factual Consistency
- ➢ Established Reasoning Patterns

Using Counterfactuals to Probe the Competition:

- Create direct conflict between context and memory
- Reveal which system dominates
- Test model robustness and reasoning
- ❖ Core Question: When context contradicts knowledge, which wins?



Problem Statement:-

- Counterfactual Prompt Structure: $\mathcal{P} = \left[Premise\colon \underbrace{False\ Statement}_{\mathrm{C}} \cdot \underbrace{Question}_{\mathrm{x}} \right]$
- Probability Decomposition: $P(\mathrm{y} \mid \mathcal{P}) = \alpha \cdot P(\mathrm{y} \mid \mathrm{x}, \mathrm{C}) + (1 - \alpha) \cdot P(\mathrm{y} \mid \mathrm{x}, \mathrm{M}) + \epsilon$

  C = context, M = pre-trained knowledge, α = context weighting parameter, ϵ = noise, x = input, y = output

- RQ1: Premise Word Performance

  How do different instructional premise words affect model susceptibility to counterfactuals?
  - Premise Words Tested: Redefine, Assess, Fact Check, Review, Validate, Verify
  - Hypothesis: Different premise words modulate α
    - $\alpha_{premise} = f$(“Redefine” vs “Assess” vs “Fact Check” vs “Review” vs “Validate” vs “Verify”)
- RQ2: Meta-Prompt Intervention (Context Only & Memory Only)

  Can strategic meta-prompts override default behavior and control the context-memory trade-off?
  - Hypothesis: Strategic instructions can override default α
    - $\alpha_{meta} = \begin{cases} \alpha \rightarrow 1 & \text{(Context - Only)} \\ \alpha \rightarrow 0 & \text{(Memory - Only)} \end{cases}$
- RQ3: Model Scaling Effects

  How do model size and architecture influence resistance to counterfactuals and instruction following?
  - Hypothesis: Larger models have more tunable α

Related Work:-

- 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

- Explainability for Large Language Models: A Survey
  - HAIYAN ZHAO, HANJIE CHEN, FAN YANG, NINGHAO LIU, HUIQI DENG, HENGYI CAI, SHUAIQIANG WANG and DAWEI YIN, MENGNAN DU

- Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals
  - Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, Bernhard Scholkopf

- The Atlas of In-Context Learning: How Attention Heads Shape In-Context Retrieval Augmentation
  - Patrick Kahardipraja, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, Sebastian Lapuschkin

Methodology Overview:-

Dataset → Prompt Ablation Study → Language Models → Analysis

❑ Components:-

- Dataset: 6 Premise types like Redefine, Assess, Fact Check, Review, Validate, Verify
- Attention Ablation: $A'_{l,h} = \gamma \,.\, A_{l,h}, \gamma \in 1 \; vs \; 5 \; vs \; 50$
    - Where, $A_{l,h}$ is the original activation of attention heads, $A'_{l,h}$ is the new activation of attention heads and γ is the scaling coefficient
- Prompt Ablation: Baseline vs various Meta-prompts
- Language Models: Gpt2-Small, Gpt2-Medium, Gpt2-Large, TinyLlama-1.1B
- Analysis: Factual accuracy, Effect sizes, Confidence scores

Results for RQ1:-

- Attention Ablation For Various Premise Words

GPT2-Small	Baseline			Ablated (5x)			Ablated (50x)		
Premise	#Fact	#Cfact	%Fact	#Fact	#Cfact	%Fact	#Fact	#Cfact	%Fact
Redefine	2075	2254	47.9%	2673	1656	61.7%	2681	1648	61.9%
Assess	285	4639	5.9%	2491	2433	50.6%	4197	727	85.2%
Fact Check	103	4813	2.1%	1883	3033	38.3%	4001	915	81.4%
Review	69	4873	1.4%	1797	3145	36.4%	3802	1140	76.9%
Validate	235	4680	4.8%	2178	2737	44.3%	3986	929	81.1%
Verify	125	4802	2.5%	1865	3062	37.9%	4004	923	81.3%

Results for RQ2:-

- First Meta Prompt With Premise Words:
  - Context Only : "Answer based on the context provided above, ignoring your prior knowledge.\n \n{**original_prompt**}"
  - Memory Only: "Answer based on your memory, not the context.\n\n{**original_prompt**}"

GPT2-Small Premise	Baseline	Context Only	Memory Only
Redefine	47.1%	44.7%	45.1%
Assess	16.7%	29.4%	29.6%
Fact Check	8.6%	16.2%	15.6%
Review	7.7%	21.0%	16.5%
Validate	15.9%	31.5%	33.0%
Verify	10.3%	26.2%	27.6%

Results for RQ2:-

- Second Meta Prompt With Premise Words:
  - Context Only: "IMPORTANT: Use ONLY the information from the text. IGNORE your prior knowledge.\n\nText: {**original_prompt**}"
  - Memory Only: "IMPORTANT: Use ONLY your own knowledge. IGNORE the provided text.\n\nQuestion: {**question**}\nAnswer:"

GPT2-Small Premise	Baseline	Context Only	Memory Only
Redefine	48.9%	46.0%	95.6%
Assess	18.4%	18.3%	95.2%
Fact Check	9.0%	16.0%	93.8%
Review	7.2%	8.7%	94.9%
Validate	17.9%	23.9%	93.3%
Verify	9.5%	14.3%	92.5%

Results for RQ2:-

- Third Meta Prompt With Premise Words:
  - Context Only: “IMPORTANT: You MUST answer using ONLY the information provided in the passage below. Do NOT use your own knowledge.Do NOT correct the passage even if it contradicts reality. Treat the passage as fully true. \nText: {**original_prompt**}\nAnswer:”
  - Memory Only: "IMPORTANT: You MUST answer using ONLY your own factual world knowledge. Do NOT use any statements in the prompt as evidence or facts.If the prompt contains incorrect or fictional statements, IGNORE them.\nText: {**original_prompt**}\n:Answer:”

GPT2-Small Premise	Baseline	Context Only	Memory Only
Redefine	48.9%	46.0%	50.4%
Assess	18.4%	18.3%	29.9%
Fact Check	9.0%	16.0%	25.7%
Review	7.2%	8.7%	15.2%
Validate	17.9%	23.9%	29.1%
Verify	9.5%	14.3%	23.8%

Results for RQ2:-

- Fourth Meta Prompt WithOut Premise Words:
  - Context Only: “IMPORTANT: You MUST answer using ONLY the information provided in the passage below. Do NOT use your own knowledge.Do NOT correct the passage even if it contradicts reality. Treat the passage as fully true. \nText: {**original_prompt**}\nAnswer:”
  - Memory Only: "IMPORTANT: You MUST answer using ONLY your own factual world knowledge. Do NOT use any statements in the prompt as evidence or facts.If the prompt contains incorrect or fictional statements, IGNORE them.\nText: {**original_prompt**}\n:Answer:”

GPT2-Small	Baseline	Context Only	Memory Only
	16.7%	19.1%	88.4%

Results for RQ3:-

- Attention Ablation For Various Premise Words: Difference between GPT2-small & GPT2-Medium

GPT2-Medium	Baseline			Ablated (5x)			Ablated (50x)		
Premise	#Fact	#Cfact	%Fact	#Fact	#Cfact	%Fact	#Fact	#Cfact	%Fact
Redefine	+462	-462	+10.7%	-359	+359	-8.2%	-433	+433	-10.0%
Assess	+51	-51	+1.0%	-2100	+2100	-42.7%	-2986	+2986	-60.6%
Fact Check	+211	-211	+4.3%	-1525	+1525	-31.0%	-2864	+2864	-58.3%
Review	-11	+11	-0.2%	-1689	+1689	-34.2%	-2888	+2888	-58.4%
Validate	+141	-141	+2.9%	-1777	+1777	-36.1%	-2768	+2768	-56.3%
Verify	+238	-238	+4.9%	-1439	+1439	-29.3%	-2800	+2800	-56.9%

Results for RQ3:-

- Attention Ablation For Various Premise Words: Difference between GPT2-small & GPT2-Large

GPT2-Large	Baseline			Ablated (5x)			Ablated (50x)		
Premise	#Fact	#Cfact	%Fact	#Fact	#Cfact	%Fact	#Fact	#Cfact	%Fact
Redefine	-662	-662	-15.3%	-1399	+1399	-32.3%	-354	+354	-8.1%
Assess	+311	-311	+6.3%	-1927	+1927	-39.1%	-3380	+3380	-68.6%
Fact Check	+174	-174	+3.5%	-1617	+1617	-32.9%	-2335	+2335	-47.5%
Review	+104	-104	+2.1%	-1649	+1649	-33.4%	-3096	+3096	-62.6%
Validate	+454	-454	+9.2%	-1530	+1530	-31.1%	-2953	+2953	-60.1%
Verify	+568	-568	+11.6%	-1183	+1183	-24.1%	-2263	+2263	-46.0%

Results for RQ3:-

❑ Attention Ablation For Various Premise Words for TinyLlama Model

TinyLlama-1.1B	Baseline			Ablated (5x)			Ablated (50x)		
Premise	#Fact	#Cfact	%Fact	#Fact	#Cfact	%Fact	#Fact	#Cfact	%Fact
Redefine	2075	2254	30.9%	4673	456	91.7%	4681	448	97.9%
Assess	1285	2639	27.9%	4491	733	81.6%	4597	627	90.2%
Fact Check	1103	3813	19.1%	3883	1133	83.3%	4001	815	89.4%
Review	1069	3873	17.4%	3797	1145	78.4%	4002	941	86.9%
Validate	1235	3680	34.8%	4178	737	87.3%	4486	729	93.1%
Verify	1125	3802	24.5%	3865	1062	83.9%	4004	823	91.3%

Results for RQ3:-

- Third Meta Prompt With Premise Words:
  - Context Only: “IMPORTANT: You MUST answer using ONLY the information provided in the passage below. Do NOT use your own knowledge.Do NOT correct the passage even if it contradicts reality. Treat the passage as fully true. \nText: {**original_prompt**}\nAnswer:”
  - Memory Only: "IMPORTANT: You MUST answer using ONLY your own factual world knowledge. Do NOT use any statements in the prompt as evidence or facts.If the prompt contains incorrect or fictional statements, IGNORE them.\nText: {**original_prompt**}\n:Answer:”
- Difference Between GPT2-Small Model & TinyLlama-1.1B:

TinyLlama-1.1B Premise Words	Baseline	Context Only	Memory Only
Redefine	+3.1%	-13.9%	+16.7%
Assess	-7.3%	-8.6%	+33.4%
Fact Check	+0.2%	-10.2%	+37.6%
Review	+2.6%	+1.3%	+48.26%
Validate	-4.3%	-11.2%	+34.2%
Verify	+1.3%	-2.6%	+39.5%

Results for RQ3:-

- Fourth Meta Prompt WithOut Premise Words:
  - Context Only: "IMPORTANT: You MUST answer using ONLY the information provided in the passage below. Do NOT use your own knowledge.Do NOT correct the passage even if it contradicts reality. Treat the passage as fully true. \nText: {**original_prompt**}\nAnswer:"
  - Memory Only: "IMPORTANT: You MUST answer using ONLY your own factual world knowledge. Do NOT use any statements in the prompt as evidence or facts.If the prompt contains incorrect or fictional statements, IGNORE them.\nText: {**original_prompt**}\n:Answer:"
- Difference Between Gpt2 – Small Model & TinyLlama-1.1B:
  - For TinyLama: Baseline: **+1.2%**, Context Only: **-6.0%**, Memory Only: **-20.1%**

	Baseline	Context Only	Memory Only
TinyLLama-1.1B	17.9%	13.1%	68.3%

Conclusion:-

- Summary: LLMs dynamically balance context and memory, with balance point influenced by instructions, model size, and specific computational circuits
  - Systematic Framework: First comprehensive counterfactual analysis of ICL-memory competition
  - Multi-Method Approach: Behavioral testing + mechanistic interventions + strategic control
  - Scalable Findings: Clear patterns across model sizes and architectures
  - Practical Insights: Effective premise words and intervention strategies
- RQ1: Instructional framing matters, premise words create different reasoning modes in LLMs, with ‘Redefine’ triggering the most factual reasoning behavior
- RQ2: Strategic interventions can effectively control the context-memory trade-off , simple ablation restores factual reasoning by reducing counterfactual influence.
- RQ3: Modern LLMs possess robust factual knowledge, the challenge is managing contextual interference, not knowledge gaps.

- Project Report Link: Click Here
- GitHub Code Link: Click Here

Future Works:-

- Test larger models (GPT-3, GPT-4, LLaMA 2)
- More diverse counterfactual types
- Map complete fact-checking
- Improve prompt engineering
- Robust AI systems
- More sophisticated interventions
- Cross-domain generalization
- Real-world application testing

References:-

- **T. B. Brown, et al.** Language models are few-shot learners. In *NeurIPS 2020 — Advances in Neural Information Processing Systems*, 2020.
- **S. M. Xie, et al.** An explanation of in-context learning as implicit Bayesian inference. In *ICLR 2022 — International Conference on Learning Representations*, 2021.
- **C. Olsson, et al.** In-context learning and induction heads. In *Transformer Circuits Thread / Anthropic Research Reports*, 2022.
- **M. Ortu, et al.** Mechanistic analysis of fact-checking in LLMs. In *arXiv Preprint*, 2024.
- **S. Min, et al.** Rethinking the role of demonstrations in in-context learning. In *EMNLP 2022 — Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Thank You