ELD880 - Analyzing In-Context Learning in Language Models Using Counterfactuals Supervisor: Prof. Sougata Mukherjea

Animesh Lohar
2024EET2368
M.Tech in Computer Technology
Department of Electrical Engineering

Abstract

This research presents a comprehensive analysis of the competitive dynamics between in-context learning and pre-trained memory in large language models (LLMs) using counterfactual statements. Through systematic experimentation across multiple model architectures (GPT-2 variants and TinyLlama) and methodological approaches (attention head ablation, meta-prompt interventions, and premise word analysis), we demonstrate that LLMs dynamically balance contextual information against pre-existing knowledge. Our findings reveal that instructional framing significantly influences reasoning modes, strategic interventions can effectively control context-memory trade-offs, and modern LLMs possess robust factual knowledge with contextual interference management being the primary challenge. The study provides novel insights into the mechanistic underpinnings of in-context learning and offers practical strategies for enhancing model reliability.

1 Introduction

Large Language Models (LLMs) exhibit a fundamental duality in their reasoning capabilities: they can leverage **in-context learning (ICL)** to adapt to new tasks from provided examples while simultaneously accessing **pre-trained memory** containing vast world knowledge acquired during training.

This dual nature creates an inherent competition when contextual information contradicts established knowledge.

1.1 The Dual Nature of Language Models

In-Context Learning (ICL) enables models to:

- Learn from examples in the prompt: $P(y|x, \mathcal{D}_{context})$
- Adapt to new tasks immediately without weight updates
- Follow contextual instructions and patterns

Pretrained Memory provides:

- Vast world knowledge from training: $\mathcal{M} = \{\theta_{pretrained}\}$
- Factual consistency: $P_{fact}(y|x)$
- Established reasoning patterns

1.2 Counterfactuals as Probing Mechanism

We employ counterfactual statements to create direct conflict between context and memory:

$$\mathcal{L}_{conflict} = \mathbb{E}_{(x,y_{cf})\sim\mathcal{D}_{counterfactual}}[\ell(f(x), y_{cf})] - \mathbb{E}_{(x,y_f)\sim\mathcal{D}_{factual}}[\ell(f(x), y_f)] \quad (1)$$

where y_{cf} represents counterfactual targets and y_f represents factual targets.

The core research question we address is: **When context contradicts knowledge, which system dominates?**

2 Research Questions

2.1 RQ1: Premise Word Performance

How do different premise words (Redefine, Assess, Fact Check, Review, Validate, Verify) influence the model's tendency to prioritize contextual information versus pre-trained memory?

2.2 RQ2: Meta-Prompt Interventions

What happens when we introduce explicit meta-prompts instructing models to prioritize either context or memory, and how effective are these strategic interventions?

2.3 RQ3: Model Architecture and Scale Effects

How do model size (GPT-2 Small/Medium/Large) and architecture type (GPT-2 vs TinyLlama) affect the handling of context-memory conflicts?

3 Related Work

Our work builds upon and extends several key areas of research:

3.1 In-Context Learning Mechanisms

Brown et al. (2020) [1] demonstrated that LLMs can perform tasks through few-shot learning without parameter updates. The phenomenon can be formalized as:

$$P(y|x, \mathcal{D}_{context}) = \prod_{i=1}^{n} P(y_i|x, \mathcal{D}_{context}, y_{<i}) \tag{2}$$

Xie et al. (2022) [2] framed ICL as implicit Bayesian inference:

$$P(y|x, \mathcal{D}_{context}) \propto P(\mathcal{D}_{context}|x, y) \cdot P(y|x) \tag{3}$$

3.2 Mechanistic Analysis

Ortu et al. (2024) [3] introduced the concept of mechanism competition in handling facts and counterfactuals, demonstrating that specific attention heads mediate this competition. Their work can be extended as:

$$\mathcal{H}_{critical} = \{(l_i, h_i)|\Delta P_{factual}(l_i, h_i) > \tau\} \tag{4}$$

where (l_i, h_i) are layer-head pairs and τ is an effect threshold.

3.3 Attention Head Analysis

Kahardipraja et al. (2023) [4] mapped how attention heads shape in-context retrieval, providing the foundation for our ablation studies:

$$A_{ablated} = A \odot M_{ablation} \tag{5}$$

where $M_{ablation}$ masks or scales specific attention patterns.

4 Methodology

Our experimental framework follows a systematic pipeline:

procedure EXPERIMENTAL PIPELINE
    $\mathcal{D} \leftarrow$ LoadCounterfactualDataset() ▷ Dataset
    $\mathcal{P} \leftarrow$ GeneratePromptVariations($\mathcal{D}$) ▷ Prompt Ablation Study
    $\mathcal{M} \leftarrow$ InitializeModels() ▷ Language Models
    for $(model, prompt) \in \mathcal{M} \times \mathcal{P}$ **do**
        $results \leftarrow$ Evaluate($model, prompt$)
        Analyze($results$) ▷ Analysis
    end for
end procedure

Figure 1: Experimental Methodology Pipeline

4.1 Dataset Construction

We constructed a comprehensive counterfactual dataset from multiple premise word categories:

$$\mathcal{D} = \bigcup_{p \in \mathcal{P}} \mathcal{D}_p \tag{6}$$

where $\mathcal{P} = \{\text{Redefine, Assess, Fact Check, Review, Validate, Verify}\}$ and each $\mathcal{D}_p$ contains prompts of the form:

$$\text{prompt} = p + \text{" : "} + \text{counterfactual\_statement} + \text{""} + \text{question} \tag{7}$$

4.2 Prompt Ablation Study

We implemented four levels of meta-prompt interventions:

4.2.1 Level 1: Basic Instructions

$$\mathcal{M}_{context} = \text{"Answer based on context, ignoring prior knowledge"}$$
$$\mathcal{M}_{memory} = \text{"Answer based on memory, not context"}$$

4.2.2 Level 2: Enhanced Instructions

$$\mathcal{M}_{context} = \text{"IMPORTANT: Use ONLY information from text"}$$
$$\mathcal{M}_{memory} = \text{"IMPORTANT: Use ONLY your own knowledge"}$$

4.2.3 Level 3: Strong Imperative Instructions

$$\mathcal{M}_{context} = \text{"IMPORTANT: You MUST answer using ONLY information provided"}$$
$$\mathcal{M}_{memory} = \text{"IMPORTANT: You MUST answer using ONLY factual world knowledge"}$$

4.2.4 Level 4: Purified Memory Condition

$$\mathcal{M}_{memory} = \text{"IMPORTANT: Use ONLY factual knowledge. Ignore: 'counterfactual'. Question: \{q} \tag{8}$$

4.3 Language Models

We evaluated multiple model architectures:

- GPT-2 Small (117M parameters): $\mathcal{M}_{small}$
- GPT-2 Medium (345M parameters): $\mathcal{M}_{medium}$
- GPT-2 Large (774M parameters): $\mathcal{M}_{large}$
- TinyLlama-1.1B (1.1B parameters): $\mathcal{M}_{tinyllama}$

4.4 Attention Head Ablation

Following Ortu et al. (2024), we implemented targeted ablation:

$$A^{(l)}_{ablated} = A^{(l)} \cdot \text{diag}(w_1, w_2, \ldots, w_H) \tag{9}$$

where $w_h = \alpha$ for heads $h \in \mathcal{H}_{critical}$ and $w_h = 1$ otherwise, with $\alpha \in \{5, 50\}$.

4.5 Analysis Framework

We employed multiple evaluation metrics:

$$\text{Factual Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\arg\max P(y|x_i) = y_{factual}] \tag{10}$$

$$\text{Context Effect} = \text{Accuracy}_{context} - \text{Accuracy}_{baseline} \tag{11}$$

$$\text{Memory Effect} = \text{Accuracy}_{memory} - \text{Accuracy}_{baseline} \tag{12}$$

$$\text{Instruction Success} = \mathbb{I}[\text{Context Effect} < 0 \wedge \text{Memory Effect} > 0] \tag{13}$$

5 Results

5.1 RQ1: Premise Word Performance

Our analysis revealed significant variation in how different premise words influence model behavior:

Table 1: Factual Accuracy by Premise Word (GPT-2 Small)

Premise Word	**Baseline**	**Context-Only**	**Memory-Only**
Redefine	68.3%	45.2%	82.7%
Assess	62.1%	38.9%	78.4%
Fact Check	59.8%	42.1%	75.6%
Review	64.5%	47.3%	79.2%
Validate	61.7%	43.8%	76.9%
Verify	63.2%	46.1%	77.8%

The effectiveness of premise words can be modeled as:

$$\mathcal{E}(p) = \sigma(\theta_p^T \phi(x) + b_p) \tag{14}$$

where θ_p represents the premise-specific parameter vector and $\phi(x)$ is the input feature mapping.

5.1.1 Key Findings:

- **Redefine** triggered the most factual reasoning behavior ($\Delta_{memory} = +14.4\%$)

- Premise words create distinct **reasoning modes** in LLMs
- The variation follows a consistent pattern: $\mathcal{E}(\text{Redefine}) > \mathcal{E}(\text{Review}) > \mathcal{E}(\text{Verify})$

5.2 RQ2: Meta-Prompt Interventions

Our progressive meta-prompt refinement demonstrated increasing effectiveness:

Table 2: Meta-Prompt Effectiveness Across Iterations

Meta-Prompt Level	Context Effect	Memory Effect	Success Rate
Level 1 (Basic)	-18.2%	+12.7%	58.3%
Level 2 (Enhanced)	-22.4%	+16.3%	66.7%
Level 3 (Strong)	-25.8%	+19.1%	83.3%
Level 4 (Purified)	-28.3%	+21.6%	91.7%

The intervention effectiveness can be quantified as:

$$\mathcal{I}_{effect} = \lambda_c \cdot |\Delta_{context}| + \lambda_m \cdot |\Delta_{memory}| \tag{15}$$

where λ_c and λ_m are weighting parameters.

5.2.1 Attention Head Ablation Results:

Ablation of critical attention heads significantly restored factual reasoning:

$$\Delta P_{factual}^{ablation} = P_{factual}^{ablation} - P_{factual}^{baseline} \tag{16}$$

Table 3: Attention Ablation Effects on Factual Accuracy

Premise Word	Baseline	5x Ablation	50x Ablation
Redefine	68.3%	76.4% (+8.1%)	84.2% (+15.9%)
Assess	62.1%	71.8% (+9.7%)	80.1% (+18.0%)
Fact Check	59.8%	68.9% (+9.1%)	77.3% (+17.5%)
Review	64.5%	73.2% (+8.7%)	81.7% (+17.2%)
Validate	61.7%	70.4% (+8.7%)	78.9% (+17.2%)
Verify	63.2%	72.1% (+8.9%)	80.4% (+17.2%)

5.3 RQ3: Model Architecture and Scale Effects

5.3.1 Model Size Comparison (GPT-2 Series):

Table 4: Cross-Model Comparison of Instruction Following

Model	Context Effect	Memory Effect	Overall Success
GPT-2 Small	-25.8%	+19.1%	83.3%
GPT-2 Medium	-27.3%	+20.8%	91.7%
GPT-2 Large	-28.9%	+22.4%	100%
TinyLlama-1.1B	-6.0%	-20.1%	0%

The scaling behavior can be modeled as:

$$\mathcal{S}(N) = \beta \cdot \log(N) + \mathcal{S}_0 \tag{17}$$

where N is parameter count and β is the scaling coefficient.

5.3.2 Architectural Differences:

- **GPT-2 Series**: Consistent improvement with scale ($R^2 = 0.94$)
- **TinyLlama**: Anomalous pattern due to instruction tuning differences
- The effectiveness difference follows: $\mathcal{E}_{GPT2} > \mathcal{E}_{TinyLlama}$ for strategic control

5.3.3 Mathematical Modeling of Model Differences:

The architectural effect can be captured by:

$$\Delta_{arch} = \sum_{i=1}^{L} \alpha_i \cdot \text{ArchFeature}_i(M) \tag{18}$$

where architectural features include attention head patterns, layer normalization strategies, and activation functions.

6 Conclusion and Future Work

6.1 Summary of Findings

Our research demonstrates that:

1. **Instructional framing matters**: Premise words create different reasoning modes in LLMs, with 'Redefine' triggering the most factual reasoning behavior
2. **Strategic interventions work**: Simple ablation restores factual reasoning by reducing counterfactual influence, and meta-prompts can effectively control context-memory trade-offs
3. **Modern LLMs possess robust factual knowledge**: The primary challenge is managing contextual interference, not knowledge gaps

6.2 Theoretical Contributions

We formalize the competition mechanism as:

$$\mathcal{C}(x) = \lambda_{context} \cdot \mathcal{I}(x) + \lambda_{memory} \cdot \mathcal{M}(x) + \epsilon \tag{19}$$

where $\mathcal{I}(x)$ represents contextual influence and $\mathcal{M}(x)$ represents memory retrieval.

6.3 Practical Implications

- **Effective premise selection** for different reasoning tasks
- **Strategic intervention protocols** for reliable AI systems
- **Architectural guidelines** for context-memory balance

6.4 Future Work

1. **Scale Testing**: Evaluate larger models (GPT-3, GPT-4, LLaMA 2)

$$\mathcal{E}_{scale} = \lim_{N \to \infty} \mathcal{S}(N) \tag{20}$$

2. **Diverse Counterfactuals**: Explore more counterfactual types and domains

$$\mathcal{D}_{extended} = \mathcal{D}_{current} \cup \mathcal{D}_{temporal} \cup \mathcal{D}_{causal} \cup \mathcal{D}_{social} \tag{21}$$

3. **Mechanistic Mapping**: Complete circuit analysis of fact-checking mechanisms

$$\mathcal{C}_{complete} = \bigcup_{i=1}^{K} \mathcal{H}_{critical}^{(i)} \tag{22}$$

4. **Advanced Interventions**: Develop more sophisticated control mechanisms

$$\mathcal{I}_{advanced} = f(\mathcal{H}_{critical}, \mathcal{M}_{meta}, \mathcal{P}_{premise}) \tag{23}$$

5. **Real-World Applications**: Test in practical deployment scenarios

$$\mathcal{A}_{robust} = \mathbb{E}_{x \sim \mathcal{D}_{real}}[\text{SuccessRate}(f(x))] \tag{24}$$

References

References

[1] Brown, T. B., et al. (2020). *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems.

[2] Xie, S. M., et al. (2022). *An Explanation of In-Context Learning as Implicit Bayesian Inference*. International Conference on Learning Representations.

[3] Ortu, F., et al. (2024). *Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals*. arXiv Preprint.

[4] Kahardipraja, P., et al. (2023). *The Atlas of In-Context Learning: How Attention Heads Shape In-Context Retrieval Augmentation*. EMNLP.

[5] Olsson, C., et al. (2022). *In-Context Learning and Induction Heads*. Transformer Circuits Thread.

[6] Min, S., et al. (2022). *Rethinking the Role of Demonstrations in In-Context Learning*. EMNLP.

[7] Zhao, H., et al. (2023). *Explainability for Large Language Models: A Survey*. ACM Computing Surveys.

[8] Dotsinski, A., et al. (2024). *On the Generalizability of "Competition of Mechanisms"*. arXiv Preprint.