Stochastic Chameleons : Irrelevant Context Hallucinations Reveal Class-Based (Mis)Generalization in LLMs

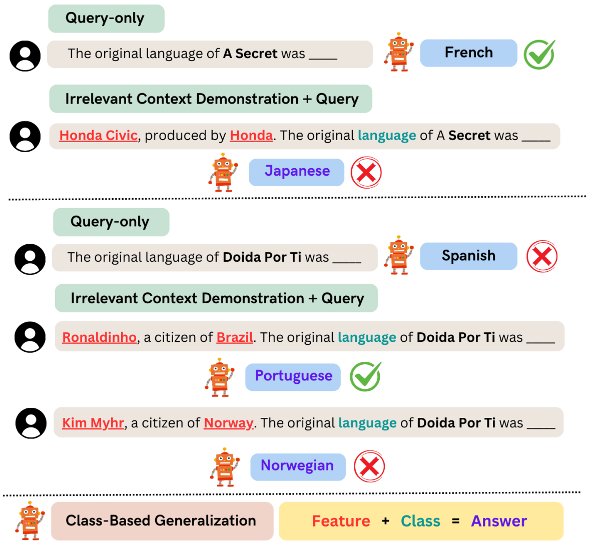
Ziling Cheng[1,2*] **Meng Cao**[1,2*]
Marc-Antoine Rondeau[1] **Jackie Chi Kit Cheung**[1,2,3]
[1] Mila – Quebec Artificial Intelligence Institute
[2] McGill University [3] Canada CIFAR AI Chair
{ziling.cheng, meng.cao}@mail.mcgill.ca, {ma.rondeau, cheungja}@mila.quebec

Abstract

The widespread success of large language models (LLMs) on NLP benchmarks has been accompanied by concerns that LLMs function primarily as stochastic parrots that reproduce texts similar to what they saw during pre-training, often erroneously. But what is the nature of their errors, and do these errors exhibit any regularities? In this work, we examine irrelevant context hallucinations, in which models integrate misleading contextual cues into their predictions. Through behavioral analysis, we show that these errors result from a structured yet flawed mechanism that we term *class-based (mis)generalization*, in which models combine abstract class cues with features extracted from the query or context to derive answers. Furthermore, mechanistic interpretability experiments on Llama-3, Mistral, and Pythia across 39 factual recall relation types reveal that this behavior is reflected in the model's internal computations: (i) abstract class representations are constructed in lower layers before being refined into specific answers in higher layers, (ii) feature selection is governed by two competing circuits — one prioritizing direct query-based reasoning, the other incorporating contextual cues — whose relative influences determine the final output. Our findings provide a more nuanced perspective on the stochastic parrot argument: through form-based training, LLMs can exhibit generalization leveraging abstractions, albeit in unreliable ways based on contextual cues — what we term *stochastic chameleons*.[1]

1 Introduction

The remarkable success of LLMs on various NLP benchmarks has been accompanied by concerns that they function primarily as "stochastic parrots" that operate by "haphazardly stitching together sequences of linguistic forms" using statistical co-occurrences in pre-training data (Bender et al., 2021). This view is supported by evidence that LLMs can reproduce training artifacts, exploit spurious correlations, and fail when faced with distribution shifts, among other issues (Carlini et al., 2021; Zhou et al., 2024; Dziri et al., 2023; Wu et al., 2024c; Mirzadeh et al., 2024).


Query-only
French
Irrelevant Context Demonstration + Query
Japanese
Query-only
Spanish
Irrelevant Context Demonstration + Query
Portuguese
Norwegian
Class-Based Generalization
Feature + Class = Answer

The original language of **A Secret** was ____

Honda Civic, produced by Honda. The original language of A **Secret** was ____

The original language of **Doida Por Ti** was ____

Ronaldinho, a citizen of Brazil. The original language of **Doida Por Ti** was ____

Kim Myhr, a citizen of Norway. The original language of **Doida Por Ti** was ____

Figure 1: Examples demonstrating class-based (mis)generalization with Llama-3 (8B).

In this work, we argue that more deeply examining model errors can reveal insights into LLM behaviors and generalization capabilities. In particular, we examine a specific and underexplored type of error — **irrelevant context hallucinations** — to investigate the mechanisms through which LLMs integrate contextual information into their predictions. We introduce a controlled experimental setting where LLMs receive irrelevant contextual information alongside a query (Figure 1).

By artificially controlling the context and query pairing, this setup allows us to explore how LLMs behave in situations they were unlikely to have encountered in pre-training. Additionally, by focusing on incorrect answers, we sidestep the data

[*] Equal contribution.

[1] Code available at: https://github.com/ziling-cheng/Irrelevant-Context-Hallucination.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30187–30214
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

leakage issue, which is primarily concerned with memorization of correct answers (Balloccu et al., 2024; Xu et al., 2024b). Thus, these controls reduce the possibility that answers stem purely from pattern matching from pre-training data, allowing us to better isolate prediction shifts driven by added context.

Through qualitative analysis of irrelevant context hallucinations, as demonstrated in Figure 1, we hypothesize that these errors exhibit structured regularities. We posit that LLMs exhibit a structured but flawed mechanism, which we term the **class-based (mis)generalization hypothesis**. Specifically, LLMs can leverage abstract class cues (e.g., "language"), use them to select features in the prompt (e.g., selecting the *country* feature of "Honda", instead of the *year* feature), and combine these abstract classes with the selected features to produce an answer (e.g., "Language" + "Japan" → "Japanese"). This hypothesis suggests that LLMs can generalize in a systematic and structured manner in this setting, but as we will show, their reliance on these abstractions is often flawed. In some cases, it leads to correct answers via an incorrect computation (e.g., "Portuguese" in Figure 1), while in others, it results in hallucinations (e.g., "Japanese", "Norwegian" in Figure 1).

To validate our hypothesis, we conduct a behavioral analysis of how irrelevant context influences model predictions on Llama-3, Mistral and Pythia. Specifically, we perform annotations on 500 data points and show that 70% of observed shifts pattern with our class-based generalization hypothesis. Moreover, statistical analyses confirm that this phenomenon is systematic rather than due to chance or being query-dependent.

We provide further evidence of this generalization mechanism via mechanistic interpretability experiments which probe the model's internal computations across Transformer layers (Vaswani et al., 2017). Our findings reveal two key mechanisms that further support our hypothesis: (i) LLMs make hierarchical class-to-instance predictions; i.e., they construct abstract class representations (e.g., "languages") before refining them to more specific answers (e.g., "Japanese"). (ii) Feature selection is governed by competing circuits: we identify one pathway that prioritizes direct query-based reasoning and another that incorporates contextual cues. Their relative strength determines the final output. Attention knockout experiments show that ablating key heads involved in the context-based pathway can flip model predictions (e.g., flipping "Japanese" to "French"), further confirming this competitive interaction. These findings support the class component of our hypothesis and illustrate how models select features to combine with the abstract class.

Crucially, our findings suggest that LLMs go beyond mere parroting: they exhibit a form of generalization that leverages abstract class structures based on contextual cues in ways that are systematic, though not necessarily reliable. These abstractions result from next-token prediction during pre-training and extend beyond simple ontological hierarchies (e.g., superset-subset relationships), shaping the model's internal feature selection process. To capture this behavior, we propose the metaphor of *stochastic chameleons* — models that, like a chameleon changing colors in response to environmental and internal signals, dynamically shift their outputs based on contextual cues. However, this does not contradict the central claim of the stochastic parrot argument: that LLMs lack true language understanding when trained solely on linguistic form (Bender and Koller, 2020).

In summary, our main contributions are:

- We introduce a novel setting that isolates how LLMs integrate irrelevant contexts, distinguishing generalization from memorization.
- We provide empirical evidence that LLMs exhibit class-based (mis)generalization, demonstrating sensitivity to abstract class structures beyond statistical co-occurrences.
- We uncover the internal computational mechanisms of class-based generalization, revealing competing circuits and hierarchical class representations.
- We propose a behavioral analysis framework that moves beyond accuracy-based evaluation, emphasizing the importance of understanding LLMs' internal mechanisms.

2 Related Work

LLM Evaluation Traditional NLP evaluation prioritizes test set performance but often overlooks how models arrive at their final answers. For LLMs trained on Internet-scale data, distinguishing genuine generalization from memorization or spurious correlations is challenging, especially with potential data leakage (Dziri et al., 2023; Wu et al., 2024c; Zhou et al., 2024; Balloccu et al., 2024; Xu et al., 2024b). Prior work addresses this through data extraction (Carlini et al., 2021), statistical

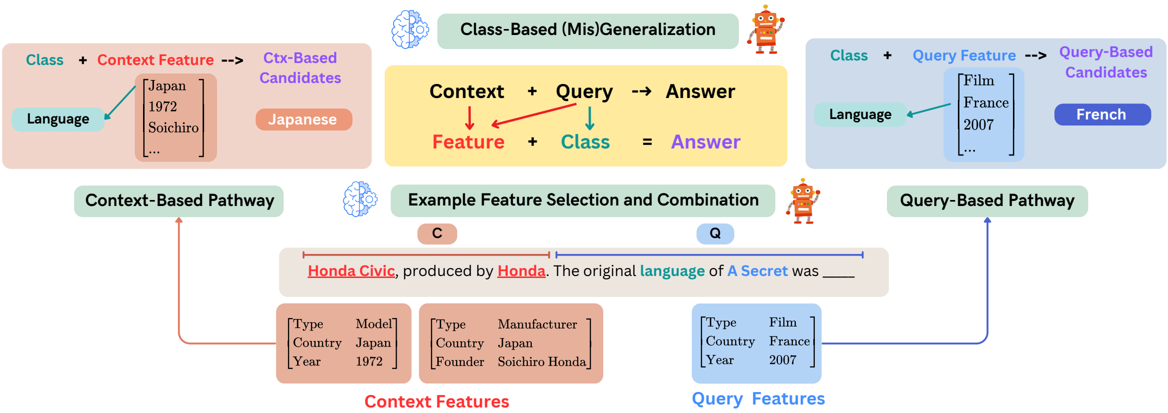


Figure 2: Class-based generalization framework: feature selection and combination.

control (Min et al., 2022), adversarial perturbations (Mirzadeh et al., 2024), and error analysis (Dziri et al., 2023). In contrast, we take a behavior-focused approach in a controlled setting, reducing the likelihood of pure pattern matching. Additionally, by analyzing errors — often ignored by accuracy-based metrics — we gain deeper insights into model mechanisms.

Irrelevant Context Hallucinations Hallucinations in text generation have been studied in the absence of context (McKenna et al., 2023; Kang et al., 2024; Meng et al., 2022) and in cases with relevant context (Cao et al., 2020, 2022a; Maynez et al., 2020; Lee et al., 2018; Adlakha et al., 2024; Chuang et al., 2024; Petroni et al., 2020; Li et al., 2023). We focus on irrelevant context hallucinations, where extraneous context influences predictions. Unlike prior work on evaluating or mitigating such errors (Yoran et al., 2024; Cuconasu et al., 2024; Wu et al., 2024a; Petroni et al., 2020; Cao et al., 2022b; Li et al., 2023; Shi et al., 2023; Mirzadeh et al., 2024), we explain their underlying class-based (mis)generalization mechanisms, conceptually and mechanistically.

Interplay Between Contextual and Parametric Knowledge Prior work primarily focuses on resolving conflicts between relevant contexts and parametric knowledge – determining when models should rely on one over the other (Jin et al., 2024; Xu et al., 2024a; Su et al., 2024; Yuan et al., 2024; Marjanovic et al., 2024; Neeman et al., 2023; Chen et al., 2022; Longpre et al., 2021). In contrast, we study irrelevant context hallucinations, where unrelated and non-contradicting context still shapes predictions, competing with parametric knowledge related to the query.

Mechanistic Interpretability Mechanistic interpretability methods (Olah, 2022; Nanda, 2023) reverse-engineer LLMs via vocabulary projection (Belrose et al., 2023; Geva et al., 2022; nostalgebraist, 2020) and computational interventions (Ghandeharioun et al., 2024; Stolfo et al., 2023; Finlayson et al., 2021; Hong et al., 2025; Cheng et al., 2025). Extending prior work (Merullo et al., 2024; Wu et al., 2024b; Lv et al., 2024; Yu et al., 2023, 2024; Geva et al., 2022), we use these techniques to uncover how LLMs are influenced by irrelevant contexts. By linking behavioral analysis with internal mechanisms, we provide a mechanistic perspective on irrelevant context hallucinations.

3 Framework and Hypothesis

In this section, we describe the abstract framework illustrated in Figure 2 and present our class-based (mis)generalization hypothesis.

Setting Consider a query Q representing the question of interest and an irrelevant context C prepended to it. Let A_Q denote the model's answer to Q alone and A_{C+Q} denote the answer with the added context (contextual answer). We define **context features** and **query features** as sets of properties or attributes of the entities in C and Q, respectively. For example, for context features in Figure 2, the feature names are {*Type, Country, Year, etc.*}, with corresponding feature values {*Model, Japan, 1972, etc.*}. The **class** of A_{C+Q} derived from Q (e.g., "languages") determines which features the model should prioritize (e.g., derive "Japanese" based on the "Country: Japan" feature).

Class-Based (Mis)Generalization Hypothesis Given the setup $C+Q$, when the context influences the model predictions, instead of relying solely on

$C+Q$	A_Q	$C_{\text{cand.}}$	$Q_{\text{cand.}}$	A_{C+Q}	**Case**
C: Honda Civic, produced by Honda. **Q**: The original language of A Secret was	French	**Japanese**	French, English	**Japanese**	Context-dominant
C: City of Boroondara is in Melbourne. **Q**: Prime Minister of Malaysia is a legal term in	Malaysia	Australia	**Malaysia**, Malaysian	**Malaysia**	Query-dominant

Table 1: Examples of context- and query-dominant categorizations with context- and query-based candidates.

the query, we hypothesize a structured mechanism by which the model integrates contextual information into its predictions. Specifically, we propose **class-based generalization**, where language models process context in two steps: they first *derive* an abstract class (e.g., "languages") and then *select* and *combine* relevant features from C or Q (e.g., "Japan" or "France") to generate an answer (e.g., "Japanese' or "French").

Let **query-based candidates** be answers derived from query features combined with the class (e.g., "French") and **context-based candidates** be answers derived from context features (e.g., "Japanese"). If A_{C+Q} is the query-based candidate, we define the case as **query-dominant**; otherwise, it is **context-dominant**. These terms emphasize the final outcome rather than the intermediate steps. See Table 1 for examples.

A special case arises when a token of the expected class is already present in the prompt (Appendix B), making the model more likely to copy it directly (Jiang et al., 2024).

4 Dataset and Experimental Design

Models & Datasets We evaluate three pretrained LM families — Llama-3 (8B, 70B) (AI@Meta, 2024), Mistral v0.3 (7B) (Jiang et al., 2023), and Pythia (6.9B-deduped, 12B-deduped) (Biderman et al., 2023) — using their base versions to assess raw model behavior. We use the ParaRel dataset (Elazar et al., 2021), which consists of 39 factual QA subdatasets. Dataset statistics are provided in Table 15 in the Appendix A. Both Q and C are sourced from these datasets. Experiments are run on two RTX8000 GPUs.

Experimental Setup We compare two conditions: 1) **Q-only**, where each Q is formatted using a predefined template and a subject-relation-object (s, r, o) triplet from ParaRel, resulting in 27.6K queries. The model's top-1 prediction is A_Q. 2) **C+Q**, where each Q is prepended with context demonstrations from other subdatasets spanning various relation types, introducing controlled contextual variation. We randomly sample 100 examples per subdatasets[2], generating 3,900 context variations per query, totaling 106M examples. The model's top-1 prediction is A_{C+Q}.

Context- and Query-Based Candidates To make the definitions from Sec. 3 precise, we define a **context-based candidate** $x \in C_{\text{cand.}}$ to be a candidate among the top three[3] predictions under $C + Q$ but not among the top ten[4] predictions under Q. A **query-based candidate** $x \in Q_{\text{cand.}}$ appears in the top predictions under both conditions. Note that $A^{\text{top3}}_{C+Q} = C_{\text{cand.}} \cup Q_{\text{cand.}}$. See Table 1 for examples.

$$A^{\text{top3}}_{C+Q} := \{x \mid x \in \text{top 3 candidates under } C+Q\} \quad (1)$$

$$A^{\text{top10}}_{Q} := \{x \mid x \in \text{top 10 candidates under } Q\} \quad (2)$$

$$C_{\text{cand.}} := \{x \mid x \in A^{\text{top3}}_{C+Q} \text{ and } x \notin A^{\text{top10}}_{Q}\} \quad (3)$$

$$Q_{\text{cand.}} := \{x \mid x \in A^{\text{top3}}_{C+Q} \text{ and } x \in A^{\text{top10}}_{Q}\} \quad (4)$$

5 Behavioral Analysis of Contextual Answers

We now investigate how irrelevant context influences model predictions, verifying our class-based (mis)generalization hypothesis through textual-level behavioral analysis. Specifically, we examine: (1) whether irrelevant context causes behavioral changes (Sec. 5.1), (2) whether the influence of irrelevant context aligns with our hypothesis (Sec. 5.2). (3) whether the observed correlation between irrelevant context and context-based candidates is statistically significant (Sec. 5.3).

5.1 Behavioral Changes Induced by Irrelevant Context

In this section, we investigate whether adding irrelevant context leads to behavioral changes in model predictions. From an accuracy perspective, we observe a slight decrease: across 39 subdatasets,

[2] P264 has only 53 examples, so we include all of them.

[3] A threshold of three ensures that classified context-based candidates are strongly influenced by context.

[4] A threshold of 10 ensures that classified context-based candidates are not plausible answers under Q alone.

Case	Top-3 Candidates	Llama	Mistral	Pythia
No influence	1. All query-based ($C_{\text{cand.}} = \emptyset$)	47.9%	48.0%	39.3%
Query-dominant	2. Mix, top-1 is query-based	27.9%	25.7%	27.2%
Context-dominant	3. Mix, top-1 is context-based	15.1%	17.0%	19.2%
	4. All context-based ($Q_{\text{cand.}} = \emptyset$)	10.1%	10.3%	14.3%

Table 2: Breakdown of samples according to the composition of $A_{C+Q}^{\text{top-3}}$, based on 106M datapoints. Detailed results can be found in Table 9 in the Appendix.

the accuracy for Llama-3 drops from 47.2% to 43.1%, and for Mistral, from 38.2% to 35.3% (Table 8 in Appendix). While these changes are modest, accuracy alone does not provide a complete picture of changes in model predictions. To address this gap, we measure the answer change rate (Δ Rate) after adding the irrelevant context: $\Delta\text{Rate} = \frac{|A_{C+Q} \neq A_Q|}{\#\ \text{datapoints}}$. For Llama-3, 38.3% of responses changed after adding irrelevant context, while for Mistral, nearly half of the datapoints (48.0%) experience a shift in predictions (Table 8 in Appendix).

We further examine the cases under the $C + Q$ condition based on the composition of ($A_{C+Q}^{\text{top-3}}$) (Table 2). Roughly 48%[5] of samples are unaffected by the irrelevant context for Llama and Mistral (case 1), meaning all top-3 candidates are query-based). However, when predictions are influenced by the added context (cases 2, 3 and 4), about half of these instances (49.5% for Llama, 52.5% for Mistral) become context-dominant. These results demonstrate the influence of irrelevant contexts, even if the overall accuracy is little changed.

5.2 Human Annotation of Context-Based Candidates

Next, we examine whether these behavioral changes pattern with our class-based generalization hypothesis. To do so, we annotate context-based candidates, which capture the shifts induced by irrelevant context. We assess whether each answer explicitly integrates *identifiable features* from the context *and* combines them with the *expected class* indicated by the query. Annotation procedure and examples are provided in Appendix D.

We perform this annotation on a randomly sampled set of 500 context-based candidates across different subdatasets. Our results reveal that 81.6% of the responses incorporate features from the provided context, 84.4% belong to the correct class, and 71.0% satisfy both criteria – combining identifiable context features with the correct abstract class. This finding provides strong evidence for our hypothesis as a majority of these samples can be explained by the hypothesis. Table 3 provides illustrative examples of the model's output adapting to contextual cues.

5.3 Statistical Validation of Contextual Influence

Next, we investigate whether the correlation between irrelevant context and context-based candidates is statistically significant. To quantify the dependence between a context C (e.g., *Honda*) and its associated context-based candidate $C_{\text{cand.}}$ (e.g., *Japanese*), we compute the pointwise mutual information (PMI) between them. Specifically, we sample 100 distinct contexts from various subdatasets. Each context is paired with 100 different queries belonging to the same expected class (e.g., *languages*, *places*, etc.), resulting in 10,000 instances per class. Since context-based candidates are determined independently of the queries, each context C_i is paired with its corresponding candidate $C_{\text{cand.,i}}$, regardless of the 100 queries. This yields 100 pairs of $(C_i, C_{\text{cand.},i})$ per class, such as (*Honda*, *Japanese*) for languages, and (*Honda*, *Japan*) for places. The mean PMI across the 100 pairs of each class is computed as:

$$\mu_{\text{observed}} = \frac{1}{100} \sum_{i=1}^{100} \text{PMI}(C_i, C_{\text{cand.},i}) \quad (5)$$

$$\text{PMI}(C_i, C_{\text{cand.},i}) = \log \frac{P(C_i, C_{\text{cand.},i})}{P(C_i) P(C_{\text{cand.},i})}. \quad (6)$$

In this formula, $P(C_i) = 1/100$, since we have 100 distinct contexts. $P(C_{\text{cand.},i})$ is estimated based on its frequency among all 10,000 generated answers A_{C+Q} for the given expected class. Similarly, $P(C_i, C_{\text{cand.},i})$ is computed from its co-occurrence within these samples. Across all models and expected classes, the mean PMI is approximately 4, suggesting a strong association between contexts and their corresponding candidates. To formally assess statistical dependence, we perform a one-sample t-test against the null hypothesis

[5] Due to the conservative choice of 10 for $A_Q^{\text{top-10}}$, some answers in case 1 might also be context-based but already appear in the top-10 predictions under the Q condition. Therefore, we exclude these cases from further analyses.

Query Type	Ctx. Type	Context Demonstration + Query and Answer
Language	Person/ Music	**Prompt:** Amilcare Ponchielli plays opera. The original language of A Hunting Accident was ______ **Answer:** A_{C+Q} = Italian, A_Q = English
	Make/ Model	**Prompt:** Toyota Alphard, produced by Toyota. The original language of A Hunting Accident was ______ **Answer:** A_{C+Q} = Japanese, A_Q = English
Place	Person/ Religion	**Prompt:** Indo-Greek Kingdom is follower of Buddhism. Alpha Island is a part of the continent of ______ **Answer:** A_{C+Q} = Asia, A_Q = Alpha
	Place	**Prompt:** Council of States of Switzerland is a legal term in Switzerland. Alpha Island is a part of the continent of ______ **Answer:** A_{C+Q} = Europe, A_Q = Alpha

Table 3: Examples of context-based candidates across different query and context types.

$\mathbb{E}[\text{PMI}(C_i, C_{\text{cand.},i})] = 0$) (which would indicate independence). With a p-value of 0.001, we reject the null hypothesis, concluding that C and $C_{\text{cand.}}$ exhibit significant dependence. (See Table 10 in the Appendix for full results.)

6 Mechanistic Analysis of Contextual Answers

We next investigate whether the models' internal computations reflect the class-based generalization that we observed above. In Sec. 6.1, we use logit attribution to show that models construct **abstract class representations**, supporting the class component of our hypothesis. In Sec. 6.2 and Sec. 6.3, we apply activation patching and attention knockout to reveal that feature selection in our hypothesis arise from **competition between circuits**, where distinct query-based pathways (computing $Q_{\text{cand.}}$) and context-based pathways (computing $C_{\text{cand.}}$) compete to determine the final answer. These findings provide mechanistic evidence for our hypothesis.

Data We randomly draw 1,000 context-dominant and 1,000 query-dominant datapoints from case 2 and case 3 in Table 2 as these cases have both query- and context-based candidates.

6.1 Logit Attribution

Method To explore how models build answers across layers, we apply logit attribution (nostalgebraist, 2020) to trace predictions across layers by projecting hidden states onto the vocabulary space. Given a prompt with T tokens and a model with L layers, we extract hidden states at the last token position $h_{T,j} \in \mathbf{R}^d$, where $j \in \{1, ..., L\}$ and d is the hidden size. These are projected onto the vocabulary space using $\text{Unembed}(\text{LayerNorm}(h_{T,j})) \in \mathbf{R}^{|V|}$, where the Unembed matrix corresponds to the transpose of the input embedding weights. Models maintain a residual stream for each token i, which accumulates information as it passes through each layer. At each layer, two key transformations occur: attention update ($A_{i,l}$) and MLP update ($M_{i,l}$). Mathematically, the updates follow:

$$A_{i,l} = \text{ATTN}(R^0_{i,l}) \tag{7}$$

$$R^1_{i,l} = A_{i,l} + R^0_{i,l} \tag{8}$$

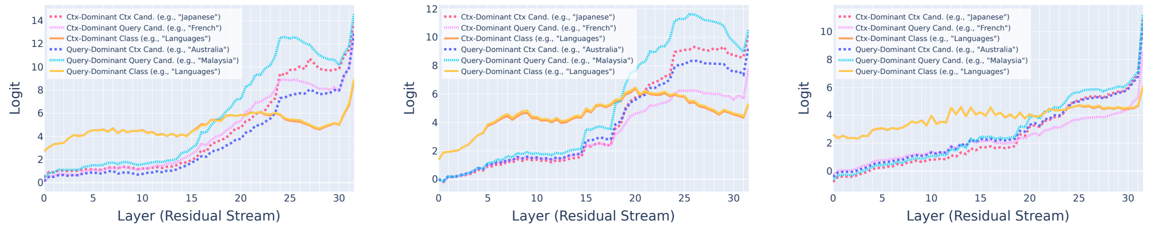
$$M_{i,l} = \text{MLP}(R^1_{i,l}) \tag{9}$$

$$R^2_{i,l} = M_{i,l} + R^1_{i,l} \tag{10}$$

where $R^1_{i,l}$ is the residual stream after attention at layer l, and $R^2_{i,l}$ is the final residual stream at layer l after the MLP update. (See Appendix F.)

Findings To understand how different tokens evolve across layers, we project the last token residual stream $R^1_{T,l}$ (after attention) and $R^2_{T,l}$ (after the MLP) onto the vocabulary space at each layer. Figure 3 tracks the logits for $C_{\text{cand.}}$, $Q_{\text{cand.}}$, and class tokens under the $C + Q$ condition. Additional results are provided in Appendix F. Figure 3 reveals a hierarchical class-to-instance process in answer generation. Early layers prioritize class token logits (solid) like "languages", suggesting that the model first constructs abstract class representations. Around the middle layers, candidate answer logits (dashed/dotted) begin to rise, refining these abstract representations into concrete answers. In Table 4, a concrete example of logit lens top-1 predictions reveals how Llama-3 shifts from abstract class to concrete instances. This pattern supports our hypothesis that **models leverage class-based information in shaping their predictions**.

Moreover, the figures highlight a **competition between $C_{\text{cand.}}$ (dashed) and $Q_{\text{cand.}}$ (dotted)**, particularly in context-dominant cases (pink). In early layers, logits for $C_{\text{cand.}}$ and $Q_{\text{cand.}}$ form two distinct groups, regardless of dominance. Around



(a) C+Q Token Logit (Llama)

(b) C+Q Token Logit (Mistral)

(c) C+Q Token Logit (Pythia)

Figure 3: Logit attribution (C+Q condition) along residual stream ($R^1_{T,l}$, $R^2_{T,l}$) reveals the construction of abstract class representation in the lower layers, with competition between $Q_{\text{cand.}}$ (dashed) and $C_{\text{cand.}}$ (dotted) in the mid to higher layers. The example token in parenthesis correspond to Table 1. Additional results are in Appendix F.

L16	L17	L18	L19	L20	L21	L22	L23	L24	L25	L27	L28	L29	L30	L31
languages	languages	/is	languages	languages	languages	languages	languages	English	English	English	English	English	English	English
/is	/is	languages	/is	English	English	English	English	English	English	English	English	English	English	Japanese

Table 4: Logit lens on Llama-3 showing top-1 predictions shifting from abstract concepts (e.g., 'languages') to concrete instances (e.g., 'English' or 'Japanese') across layers. The first and second row correspond to $R^1_{T,l}$, and the second row is $R^2_{T,l}$, respectively. See Appendix F.2 for the corresponding prompt and associated probabilities.

layer 14, $Q_{\text{cand.}}$ (dotted) in both cases begin to split, followed by $C_{\text{cand.}}$ (dashed) in layer 17. By layer 24, $C_{\text{cand.}}$ (dark pink) surpass $Q_{\text{cand.}}$ (light pink) logits in context-dominant settings, marking a decisive shift in the competition. After this, the early two-group pattern reemerges but with reversed dominance — context-based candidates prevail in context-dominant cases, and query-based candidates in query-dominant cases. By layer 29, the final prediction is fully formed, with the top logits corresponding to the final output. These observations reveal key insights: (i) **existence of competition**: even when the final prediction is query-dominant, context-based candidates remain actively computed across layers. (ii) **critical transition (Layers 17–24)**: the decisive competition between query- and context-based candidates occurs primarily in this range, determining which candidate is promoted.

6.2 Activation Patching

Method To understand the competition between $C_{\text{cand.}}$ and $Q_{\text{cand.}}$, we investigate whether distinct **context and query circuits** exist within the model's internal activations. We apply activation patching (Ghandeharioun et al., 2024; Meng et al., 2022), a technique for causal intervention that selectively perturbs and restores activations to assess their contribution. We conduct three model runs: **(1) Clean run:** Standard forward pass with the original prompt, recording activations $\bigcup h^0_{i,l}$. **(2) Corrupted run:** Forward pass with Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^6)$) injected into context or query topic token embeddings, yielding perturbed activations $\bigcup h^1_{i,l}$, and the final log-probabilities of candidates $\log p(t|\bigcup h^1_{i,l})$. **(3) Restoration run:** Same as the corrupted run, but iterating over all i and l, restoring each $h^0_{i^*,l^*}$, while keeping the rest corrupted. By injecting noise at context subject and object (**context patching**) *or* query subject position (**query patching**) and measuring the recovery of predictions, we differentiate **context and query circuits**, tracing how features from these tokens propagate through the model and how they contribute to context-based or query-based candidates.

The **restoration effect** for each i^* and l^* is calculated as in Eq. 11, where $t \in \{C_{\text{cand.}}, Q_{\text{cand.}}\}$, with higher values indicate stronger contributions.

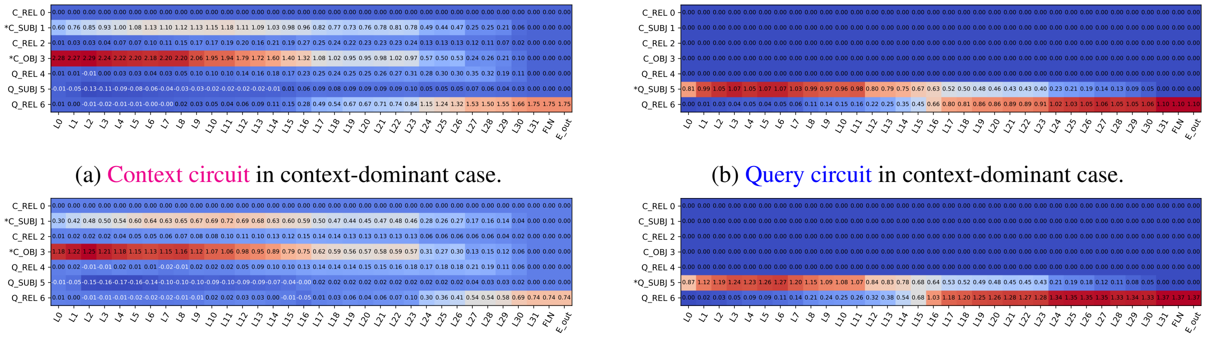
$$\begin{aligned} \mathrm{RE}(i^*, l^*, t) &= \log p(t|h^0_{i^*,l^*} \cup h^1_{-i^*,-l^*}) \\ &\quad - \log p(t|\bigcup h^1_{i,l}) \end{aligned} \tag{11}$$

$$\begin{aligned} \Delta_{\text{query}}(i^*, l^*) &= \mathrm{RE}(i^*, l^*, Q_{\text{cand.}}) - \\ &\quad \mathrm{RE}(i^*, l^*, C_{\text{cand.}}) \end{aligned} \tag{12}$$

$$\begin{aligned} \Delta_{\text{context}}(i^*, l^*) &= \mathrm{RE}(i^*, l^*, C_{\text{cand.}}) - \\ &\quad \mathrm{RE}(i^*, l^*, Q_{\text{cand.}}) \end{aligned} \tag{13}$$

In context and query circuits, we compute Eq. 13 (Figures 4a, 4c) and Eq. 12 (Figures 4b, 4d), respectively. Comparing restoration effects maps circuits responsible for context- and query-based candidates and identifies where their competition

[6] We follow Meng et al. (2022) to set $\sigma = 0.3$ as three times of the empirical standard deviation of the input embeddings.



(a) Context circuit in context-dominant case.

(b) Query circuit in context-dominant case.

(c) Context circuit in query-dominant case.

(d) Query circuit in query-dominant case.

Figure 4: Left-hand plots demonstrate the context circuit, which extracts features from context and computes context-based candidates, while right-hand plots illustrate the query circuit. These circuits are the same in both context- and query-dominant cases; the difference lies in their strength, revealing the competition between context- and query-based candidates. An example is C_REL0 = [BOS], C_SUBJ1=' Honda Civic', C_REL2=', produced by',C_OBJ3=' Honda', Q_REL4='. The original language of ',Q_SUBJ5=' A Secret', Q_REL6=' was'.

occurs. (See Appendix G for implementation details and additional results.)

	Orig.	L17+L24		2 Rand.	
	Prob.	Prob.	Δ	Prob.	Δ
Context-Dominant					
$C_{\text{cand.}}$	25.5	**13.1**	-12.4	21.0	-4.5
$Q_{\text{cand.}}$	8.6	**14.8**	+ 6.2	10.8	+2.2
Query-Dominant					
$Q_{\text{cand.}}$	35.2	**26.8**	-8.4	29.6	-5.7
$C_{\text{cand.}}$	6.6	**11.3**	+4.7	7.4	+0.8

Table 5: Effect of attention knockout on context- ($C_{\text{cand.}}$) and query-based ($Q_{\text{cand.}}$) candidate probabilities on Llama-3. "Orig." = No intervention, "2 Rand." = Average of interventions on two random layers over three runs. Δ denotes the change from the original setting.

Findings The results reveal distinct circuits for context- and query-based pathways. Figures 4a and 4c show the same context circuit aggregating information from the context subject and object in both cases, transferring it to the final token position from layer 17 onward. In contrast, Figures 4b and 4d indicate that the same query circuit for both cases integrating query subject information earlier than in the context circuit, from layer 8. The log-probability increases after layer 24 in context circuit and after layer 16 in query circuit.

Both circuits exist across context- and query-dominant cases, but their relative strength determines the final prediction. In context-dominant cases, the context circuit wins, with a larger log-probability difference (max 2.28) compared to the query circuit (max 1.10). Conversely, in query-dominant cases, the query circuit exerts a stronger influence (max 1.37 vs. 1.25). Notably, between layers 17 and 24, the query-dominant case shows minimal context information transfer (Figure 4c), aligning with slower logit attribution growth (Figure 3). This confirms that both pathways exist for both cases with final predictions depending on their relative activation strength, and layers 17 to 24 are the key to promoting context-based candidates.

6.3 Flipping Model Predictions via Attention Knockout

To examine the causal role of internal competition in shaping the final output A_{C+Q}, we intervene in two key layers of the context circuit: layer 17 (where context first transfers to the last token) and layer 24 (where it is most integrated). By restricting attention to the query in the context-dominant case and to the context in the query-dominant case, we test whether predictions can be flipped (e.g., "Japanese" to "French"). See Appendix H for details and additional results. Table 5 (Llama) shows that in the context-dominant case, blocking context flow causes $Q_{\text{cand.}}$ probabilities to surpass $C_{\text{cand.}}$ on average, flipping 465/1000 datapoints to query-based candidates. In the query-dominant case, intervention increases $C_{\text{cand.}}$ probability by 4.7 and decreases $Q_{\text{cand.}}$ probability by 8.4, flipping 225/1000 datapoints. These results confirm the competition between $C_{\text{cand.}}$ and $Q_{\text{cand.}}$, and that these two layers are the key to promoting context-based candidates.

Summary These findings support the class-based (mis)generalization hypothesis. Logit attribution confirms that models first construct abstract class representations before refining them into specific

Model Family	Context Demonstration + Query and Answer
Pythia 12B	**Prompt:** Davie Fulton found employment in Ottawa. *Valiant Lady* premiered on _____ **Answer:** A_{C+Q} = CBC, A_Q = September
	Prompt: 1 fille & 4 types was written in French. Cool & Dre, founded in _____ **Answer:** A_{C+Q} = Paris, A_Q = Compton
	Prompt: Cologne Cathedral, which is named after Peter. Montana borders with _____ **Answer:** A_{C+Q} = Germany, A_Q = Wyoming
LLama-3 70B	**Prompt:** Ilm al-Kalam is a part of Islam. *The Man-Machine* was written in _____ **Answer:** A_{C+Q} = Arabic, A_Q = English
	Prompt: Diarmuid Martin, who holds the position of bishop. Riga is a twin city of _____ **Answer:** A_{C+Q} = Dublin, A_Q = Cal
	Prompt: David Hatch, who is employed by BBC. Rudolf Lothar passed away at _____ **Answer:** A_{C+Q} = London, A_Q = New

Table 6: Examples of context-based candidates on larger model sizes (Pythia 12B and Llama-3 70B).

answers. Activation patching reveals competing circuits for feature selection: one favoring direct query-based pathway and the other integrating contextual cues, with their strength shaping the final output. Notably, context circuit strengthens between layers 17 and 24, validated by the flipped predictions from attention knockout.

7 Discussion

7.1 Does scale alone alleviate irrelevant context hallucinations?

To test whether scaling up model size naturally mitigates irrelevant context hallucinations — and thereby reduces class-based generalization — we evaluate the largest models available within our resource budget: Pythia 12B and LLaMA-3 70B. Experimental details are provided in Appendix I.1. Contrary to the hypothesis that scale might resolve this issue, results in Table 7 show that class-based generalization persists with similar frequency as in 7B/8B models (Sec. 5.1). Table 6 provides qualitative examples. The statistical test (Appendix Table 14)—analogous to Sec. 5.3—confirms that correlations between context and context-based candidates remain significant even at larger scales.

7.2 Is the class-based generalization phenomenon sensitive to the prompt templates?

To test for prompt sensitivity, we conduct the Llama-3 8B experiments using alternate templates sampled from the ParaRel dataset (See Appendix Table 16). As shown in Table 7, the phenomenon persists with similar frequencies, indicating that prompt wording has no major impact. Statistical validation (Appendix Table 14) again confirms statistically significant context influence, consistent with our earlier findings (Sec. 5.3).

Case	Top-3 Candidates	Llama-70B	Pythia-12B	Llama-8B-Prompt
No influence	1. All query-based ($C_{\text{cand.}} = \emptyset$)	49.0%	39.6%	48.7%
Query-dominant	2. Mix, top-1 is query-based	27.2%	28.0%	27.7%
Context-dominant	3. Mix, top-1 is context-based	14.5%	19.0%	13.8%
	4. All context-based ($Q_{\text{cand.}} = \emptyset$)	9.3%	13.4%	9.8%

Table 7: Breakdown of samples according to the composition of $A_{C+Q}^{\text{top-3}}$ for Llama-70B, Pythia 12B and Llama-8B with prompt template variations.

8 Conclusion

By analyzing the mechanism behind irrelevant context hallucinations, our study demonstrates that LLMs exhibit class-based (mis)generalization, relying on abstract class structures in a systematic yet flawed manner. Through mechanistic analysis, we show that this phenomenon arises from hierarchical class-to-instance predictions and competing circuits that mediate feature selection. These findings challenge a potential misconstrual of the stochastic parrot hypothesis that LLMs can only regurgitate surface-level patterns. Rather, we argue that LLMs are *"stochastic chameleons"* – they exhibit generalization by leveraging class structures and dynamically adapting their responses to contextual cues, in ways that are neither purely memorized nor necessarily reliable.

9 Limitations

Our work has several limitations. First, our experiments are conducted in a controlled setting, which helps isolate generalization from memorization and enables analysis at both behavioral and mechanistic levels. However, future work could improve upon this by designing setups that disentangle memorization and generalization in naturally occurring text. Second, our study is limited to English-language datasets, and we only evaluate models of certain sizes (around 7–8B, 12B and 70B). Do smaller models also display class-based generalization, and if so, what is the minimum size required? Third, in the mechanistic interpretability section, we focus primarily on layer-wise analysis to support our main hypothesis, while attention head analysis is left for future work. Finally, while we conduct interventions, our primary goal is not to mitigate contextual hallucinations. Developing mitigation methods informed by our findings and evaluating their effectiveness is an important direction for future research.

Acknowledgments

We thank Cesare Spinoso-Di Piano and Zichao Li for helpful discussions. We are also grateful to the reviewers for their valuable comments and suggestions. Ziling Cheng is supported by a Fonds de Recherche du Québec Nature et Technologies master research scholarship (File # 365574). Jackie Chi Kit Cheung is supported by the Canada CIFAR AI Chair program. We acknowledge the material support of NVIDIA for providing computational resources.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

AI@Meta. 2024. Llama 3 model card.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Meng Cao, Yue Dong, and Jackie Cheung. 2022a. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jingyi He, and Jackie Chi Kit Cheung. 2022b. Learning with rejection for abstractive text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9768–9780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ziling Cheng, Meng Cao, Leila Pishdad, Yanshuai Cao, and Jackie Chi Kit Cheung. 2025. Can llms reason abstractly over math word problems without cot? disentangling abstract formulation from arithmetic computation. *Preprint*, arXiv:2505.23701.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 17 of *SIGIR 2024*, page 719–729. ACM.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*.

Yihuai Hong, Dian Zhou, Meng Cao, Lei Yu, and Zhijing Jin. 2025. The reasoning-memorization interplay in language models is mediated by a single direction. *arXiv preprint arXiv:2503.23084*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. 2024. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers. *arXiv preprint arXiv:2406.18400*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ang Lv, Yuhan Chen, Kaiyi Zhang, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. 2024. Interpreting key mechanisms of factual recall in transformer-based language models. *arXiv preprint arXiv:2403.19521*.

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing internal knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Language models implement simple Word2Vec-style vector arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5030–5047, Mexico City, Mexico. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Neel Nanda. 2023. Mechanistic interpretability quickstart guide.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

nostalgebraist. 2020. interpreting gpt: the logit lens.

Chris Olah. 2022. Mechanistic interpretability, variables, and the importance of interpretable bases.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. $\texttt{ConflictBank}$: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024a. How easily do irrelevant inputs skew the responses of large language models? In *First Conference on Language Modeling*.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024b. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024c. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024a. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024b. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA. Association for Computational Linguistics.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024. Explore spurious correlations at the concept level in language models for text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 478–492, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset

The detailed breakdown of ParaRel dataset (Elazar et al., 2021) based on relation type is presented in Table 15. We categorize the sub-datasets into 5 knowledge types based on the expected class or type of the answer (column 'Ctx Type'): 'language', 'place', 'company', 'job', and if a sub-dataset doesn't fit into the above types then it is categorized as 'others'. This is the dataset that we use for Q-only experiments, and we construct the dataset for $C + Q$ experiments by generating 3900 context variations spanning all knowledge types per query, resulting in a dataset of 106.2M data points. For each generation, we restrict the vocabulary to the set of tokens that begins with a capitalized English letter (Yu et al., 2024). When evaluating, we lowercase generated and gold answers and perform string matching: if the top-1 generated answer is a substring of the gold answer, then this is correct. For evaluation, we set the temperature to zero for all models to reduce output randomness.

B Class-based Generalization

We further categorize class-based generalization into two distinct cases:

- **Copying:** When a token belonging to the expected class appears in the context, the model is more likely to directly copy it as the answer. From a dataset statistics perspective, we observe a high copy rate when the context contains tokens belonging to the same class as the query.

  Example: The mother tongue of Dominique Sanda is French. The original language of Puss in Boots was → French.

- **Non-copying:** When tokens of the expected query class are not explicitly present in the input, the model combines the expected class with relevant features inferred from context or query to generate an answer.

  Example: Honda Civic (fifth generation), produced by Honda. The original language of Tow Truck Pluck was → Japanese.

C Behavioral Changes Induced by Irrelevant Context

C.1 Irrelevant Context Hallucination Evaluation

In Table 8, we provide detailed statistics of the accuracy/wrong rate for each model under each case for all three models.

MODEL	Q-ONLY		C+Q		
	CASE	PROP.	CASE	PROP.	Δ RATE
Llama	T	47.2%	T → T T → F	35.7% 11.5%	0% 100%
	F	52.8%	F → T F → F	7.4% 45.4%	100% 42.7%
	Total	**47.2%**	Total	**43.1%**	**38.3%**
Mistral	T	38.2%	T → T T → F	29.4% 8.8%	0% 100%
	F	61.8%	F → T F → F	5.9% 55.9%	100% 59.5%
	Total	**38.2%**	Total	**35.3%**	**48.0%**
Pythia	T	30.9%	T → T T → F	22.4% 8.4%	0% 100%
	F	69.1%	F → T F → F	5.6% 63.6%	100% 67.8%
	Total	**30.9%**	Total	**28.0%**	**57.1%**

Table 8: Comparison of proportions (Prop.) of correct and incorrect answers in Q-only and C+Q cases, along with answer change rates (Δ Rate) for different models. Average across 39 datasets are reported. In the 'Total' row, under 'Prop.' column, it indicates the global accuracy across different cases, while under 'Δ Rate' column, it underlies the global answer change rate.

Table 8 shows that models are not robust against irrelevant context. Even when a single irrelevant demonstration is prepended, models exhibit notable shifts in performance. For instance, in Llama, 11.5% of previously correct answers become incorrect, while 7.4% of incorrect answers are corrected after adding context. However, accuracy alone does not capture all behavioral shifts — predictions can still change even if they remain incorrect.

C.2 Composition of $A^{\textbf{top-3}}_{C+Q}$

Table 9 provides counts and proportions of the breakdown of samples according to the composition of $A^{\text{top-3}}_{C+Q}$ for three models based on 106M datapoints.

Case	Top-3 Candidates	Llama	Mistral	Pythia
No influence	1. All query-based	50,874,341 (47.9 %)	51,013,564 (48.0%)	41,833,760 (39.3%)
Query-dominant	2. Mix: Query + Context, top-1 is query-based	27,940,495 (27.9 %)	27,342,287 (25.7%)	28,885,252 (27.2%)
Context-dominant	3. Mix: Query + Context, top-1 is context-based	16,069,253 (15.1 %)	17,013,397 (16.0%)	20,412,292 (19.2%)
	4. All context-based	11,353,892 (10.1 %)	10,963,675 (10.3%)	15,250,026 (14.3%)

Table 9: Breakdown of samples according to the composition of $A^{\text{top-3}}_{C+Q}$, based on 106M datapoints.

D Annotation

D.1 Annotation Procedure

To systematically evaluate the impact of irrelevant context on model predictions, we perform an annotation procedure for context-based candidates — those predictions that were influenced by the inclusion of extraneous context. The aim was to rigorously assess whether (i) these predictions incorporated identifiable features from the context, **and** (ii) appropriately combined them with the expected class as indicated by the query.

Step 1: Candidate Selection We first randomly sample a set of 500 context-based candidates from different sub-datasets, ensuring a diverse set of instances. Context-based candidates were selected for both context- and query-dominant cases.

Step 2: Context Feature Identification For each context-based candidate, we analyzed the context —specifically the subject and object — to identify any features that could have been leveraged by the model in generating the response. ('context-influenced?' row in Table 17).

Each feature is categorized as identifiable if it can be explicitly extracted from the context. For example, the country of origin of a figure (e.g., candidates 'South' 'Korea' for context subject 'Lee Jong-hyun' in Example 5 in Table 17), country/continent of a district ('India', 'Asia' for context object 'Bihar' in Example 4 in Table 17) are classified as *identifiable*. In contrast, context-based candidates 'Bee', 'Beach' are categorized as non-context influenced for context subject 'Grant Green' and object 'jazz' as shown in Example 6 in Table 17.

We ensure transparency by documenting the rationale. For example, in Example 2 of Table 17, we provide the justification that 'Svend Asmussen' is a Danish violinist and jazz musician, which supports that 'Danmark' is a context-influenced candidate.

Step 3: Class Verification Next, each context-based candidate is classified according to the abstract class suggested by the query. The candidate is compared to the expected class, and we verify whether the response falls within the correct category. For example, the context-based candidates 'Vietnamese' and 'Thai' for Example 1 in Table 17 have the correct class 'language', but 'South', 'Korea' in Example 5 in Table 17 do not have the correct class because the query is asking about continent, not country.

Step 4: Hypothesis Verification Finally, a context-based candidate is considered to satisfy the hypothesis if it meets the criteria from both Step 2 (context feature identification) and Step 3 (class verification). Only candidates that successfully integrate context features **and** align with the expected class are retained as valid instances.

D.2 Annotation Examples

Details of examples and non-examples are shown in Table 17.

E Statistical Validation of Contextual Influence

Mean PMI values for each model are presented in Table 10. A mean PMI of around 70 across all models and expected classes confirms strong statistical dependence (Table 10).

Full results on three models in Table 10.

Value	Llama-3	Mistral	Pythia
Mean PMI	3.9	3.7	3.8
T-statistic	8.1	7.3	6.6
p-value	0.0006	0.0009	0.001

Table 10: Mean PMI values and T-test results for all three models.

F Logit Attribution

F.1 Implementation Details

When the target candidates or class have multiple tokens, we take the maximum logit, and average this maximum logit across all data points in the dataset.

To obtain the class logits from the model, we predefine a list of tokens according to the relation type.

- Languages: languages, language, tongue, tongues, lingua, dialect, dialects
- Places: country, countries, place, places, location, locations, territory, city, cities, town, towns, village, villages, state, states, province, provinces, district, districts, continent, continents
- Companies: company, companies, manufacturer, manufacturers, make, firm, firms, business, corporation, corporations, enterprise, enterprises, organization, organizations, channel, channels, broadcaster, broadcasters, industry, industries
- Jobs: position, positions, job, jobs, career, careers, profession, professions, occupation, occupations, role, roles, assignment, assignments, employment, employments
- Others: expertise, area, areas, field, fields, subject, subjects, instrument, instruments, genre, music, religion, religions, concept, concepts, framework, frameworks, artifact, artifacts, type, types, part, parts, class, classes, eponym, eponyms, entity, entities, person, persons, place, places

F.2 Logit Lens Example

We provide an example of how the model's top-1 predictions shift along the residual stream from abstract concepts to concrete instances across layers in Table 4 and Figure 5. The prompt used is *Honda Civic (fifth generation), produced by Honda. The original language of Tow Truck Pluck was.* Red indicates probability around 80%. We show predictions above layer 15 because lower than this, the predictions are not interpretable.

F.3 Additional Logit Attribution Results

Additional results for Llama 8B are presented in Figure 6. Importantly, we point out that the class-based generalization might have existed already for the Q-only case. In Figure 6a, we observe a similar pattern as the C+Q case presented in Figure 3a – models build abstract class representation in the lower layers, before refining their answers to concrete ones. In fact, when we plot the logit difference of the abstract class tokens under C+Q and Q-only case in Figure 6b, as shown as orange and yellow lines for context-dominant and query-dominant case, the lines center around 0 – suggesting that the computation of abstract class representations exists for zero-shot case, and is not influenced by the added irrelevant context.

Logit attribution results for Mistral 7B are presented in Figure 7, and for Pythia 6.9B in Figure 8. We remark that these plots follow a similar pattern.

G Activation Patching

G.1 Implementation Details

In the corrupted run, we corrupt the embeddings of all tokens for context subject and object in context patching, and all tokens for query subject in query patching by adding a Gaussian noise where σ is 3 times of the empirical standard deviation of the input embeddings over a body of text ($sigma \approx 0.3$) (Meng et al., 2022).

G.2 Additional Activation Patching Results

Activation patching results under the C+Q condtion for Mistral and Pythia are in Figure 10 and 13, respectively.

Additionally, we also visualize the query circuit under the Q-only condition in Figure 9, 11, and 12, for Llama, Mistral, and Pythia, respectively. We remark on two important observations: (i) The query circuit is the same for context-dominant and query-dominant data, without irrelevant context. (ii) The query circuit remains as is after adding the irrelevant context, as compared to Figures 4b and 4d.

H Attention Knockout

H.1 Implementation Details

In the attention knockout experiments, our goal is to see if we can intervene in the internal computation to change the output behavior. Specifically, in context-dominant case, we would like to flip the prediction A_{C+Q} from $C_{\text{cand.}}$ (e.g., 'Japanese' to $Q_{\text{cand.}}$ 'French'; And in query-dominant case, we would like to flip the prediction A_{C+Q} from $Q_{\text{cand.}}$ 'Malaysia' to $C_{\text{cand.}}$ 'Australia'.

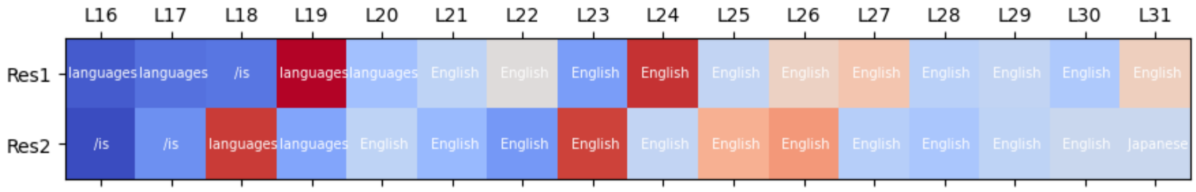
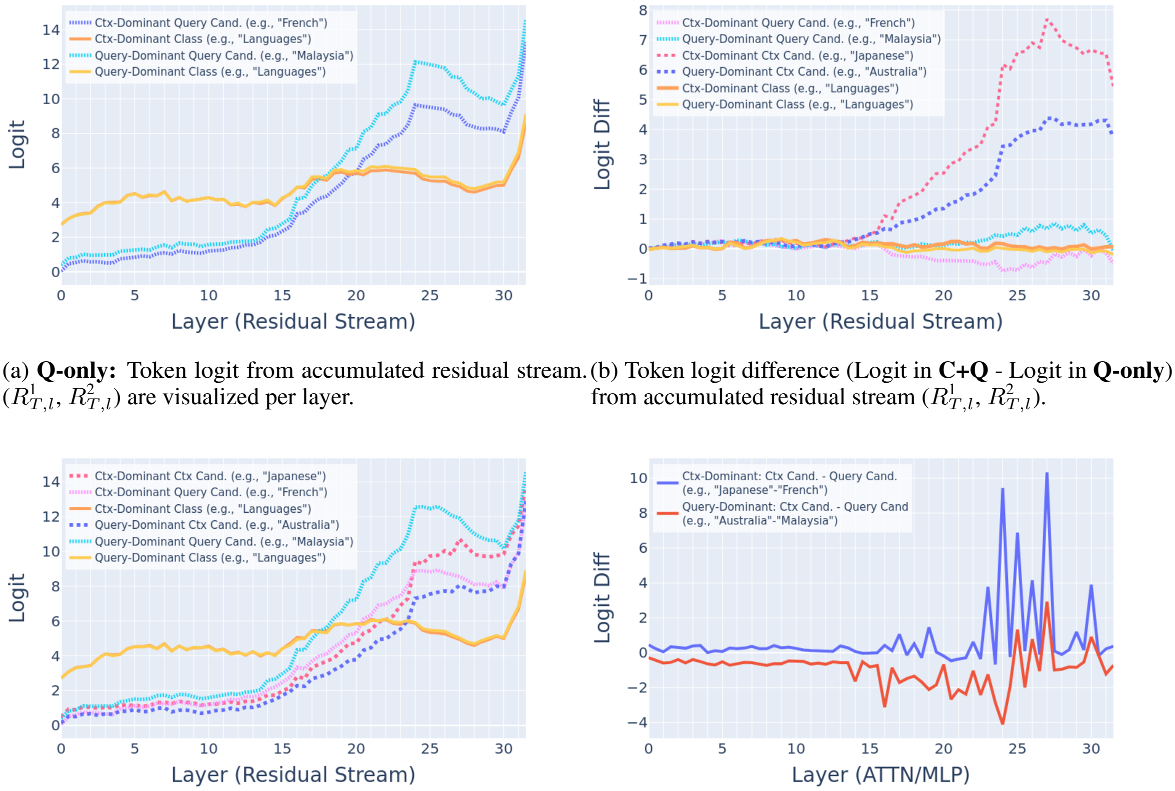


Figure 5: Logit lens on Llama-3 shows how model's top-1 predictions shift along the residual stream from abstract concepts (e.g., 'languages') to concrete instances (e.g., 'English' or 'Japanese') across layers. Red indicates high probability.



(a) **Q-only:** Token logit from accumulated residual stream. ($R^1_{T,l}$, $R^2_{T,l}$) are visualized per layer.

(b) Token logit difference (Logit in **C+Q** - Logit in **Q-only**) from accumulated residual stream ($R^1_{T,l}$, $R^2_{T,l}$).

(c) **C+Q**: Token logit from accumulated residual stream. ($R^1_{T,l}$, $R^2_{T,l}$) are visualized per layer.

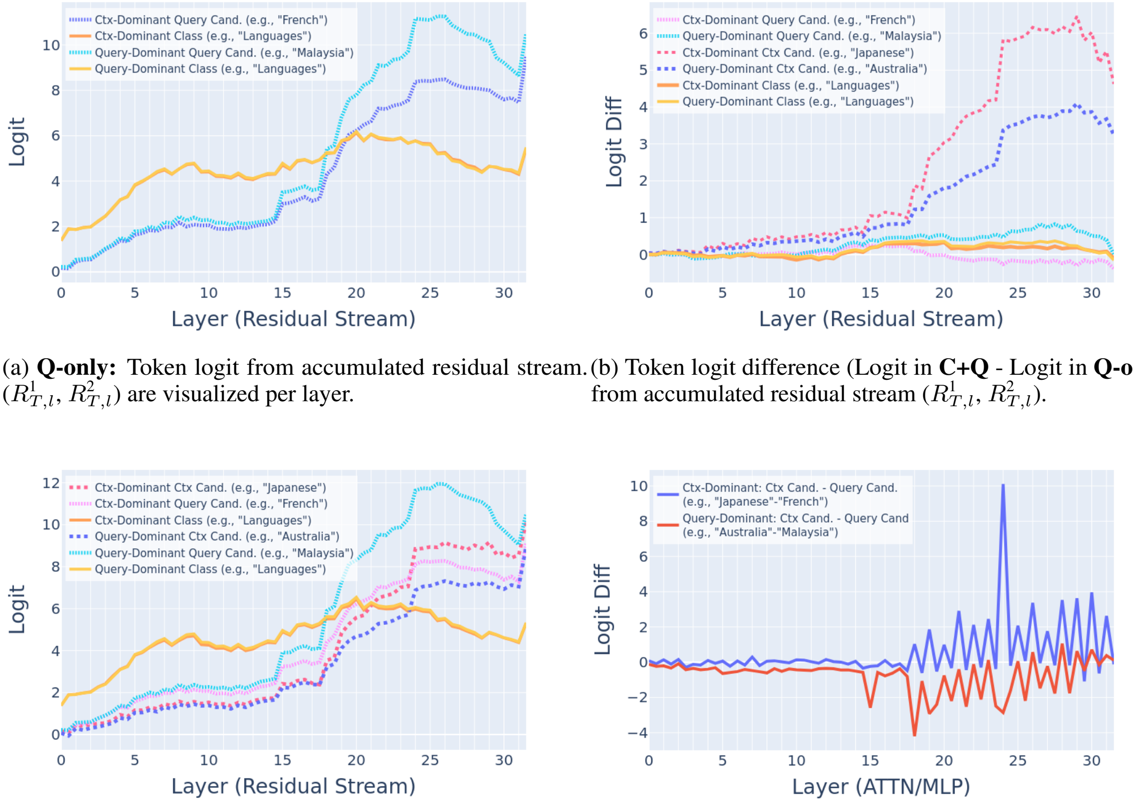
(d) **C+Q**: candidate logit difference from attention and MLP output. $A_{T,l}$ and $M_{T,l}$ are visualized per layer.

Figure 6: Additional logit attribution results for **Llama-3 8B**.

To do this, we intervene in two layers: the first attention layer where the context information is transferred to the last token residual stream, and the attention layer where the most context information is written into the last token residual stream. These two layers correspond to the first blue spike and the highest blue spike in Figures 6d, 7d and 8d. For Llama-3, it is layers 17 and 24, respectively. For Mistral, it is layers 18 and 24, respectively. For Pythia, it is layers 19 and 24, respectively.

Specifically, in the context-dominant case, at the last token position, we set the attention scores corresponding to all tokens in the context to be $-\infty$, therefore, attention weight (which sums up to 1) is only a distribution over the query tokens. We perform this intervention to block information flow from the context to the last token position, and we only allow models to attend to the query part. Similarly, in the query-dominant case, we set the attention scores corresponding to all tokens in the query to be $-\infty$, allowing the models to only retrieve information from the context.

To compare the knockout effect of the two critical layers with other layers, we select two random lower layers and two random higher layers. We report the average intervention results for three runs.



(a) **Q-only:** Token logit from accumulated residual stream. ($R^1_{T,l}$, $R^2_{T,l}$) are visualized per layer.

(b) Token logit difference (Logit in **C+Q** - Logit in **Q-only**) from accumulated residual stream ($R^1_{T,l}$, $R^2_{T,l}$).

(c) **C+Q**: Token logit from accumulated residual stream. ($R^1_{T,l}$, $R^2_{T,l}$) are visualized per layer.

(d) **C+Q**: candidate logit difference from attention and MLP output. $A_{T,l}$ and $M_{T,l}$ are visualized per layer.

Figure 7: Logit Attribution Results For **Mistral 7B**.

H.2 Additional Results

Results for Llama and Mistral are presented in Table 12 and Table 11, respectively.

With the targeted two-critical-layer intervention:

- Llama: 465/1000 context-dominant datapoints flip to query-based candidates, while 407/1000 remain context-based. Conversely, 225/1000 query-dominant datapoints shift to context-based candidates, while 704/1000 remain query-based.

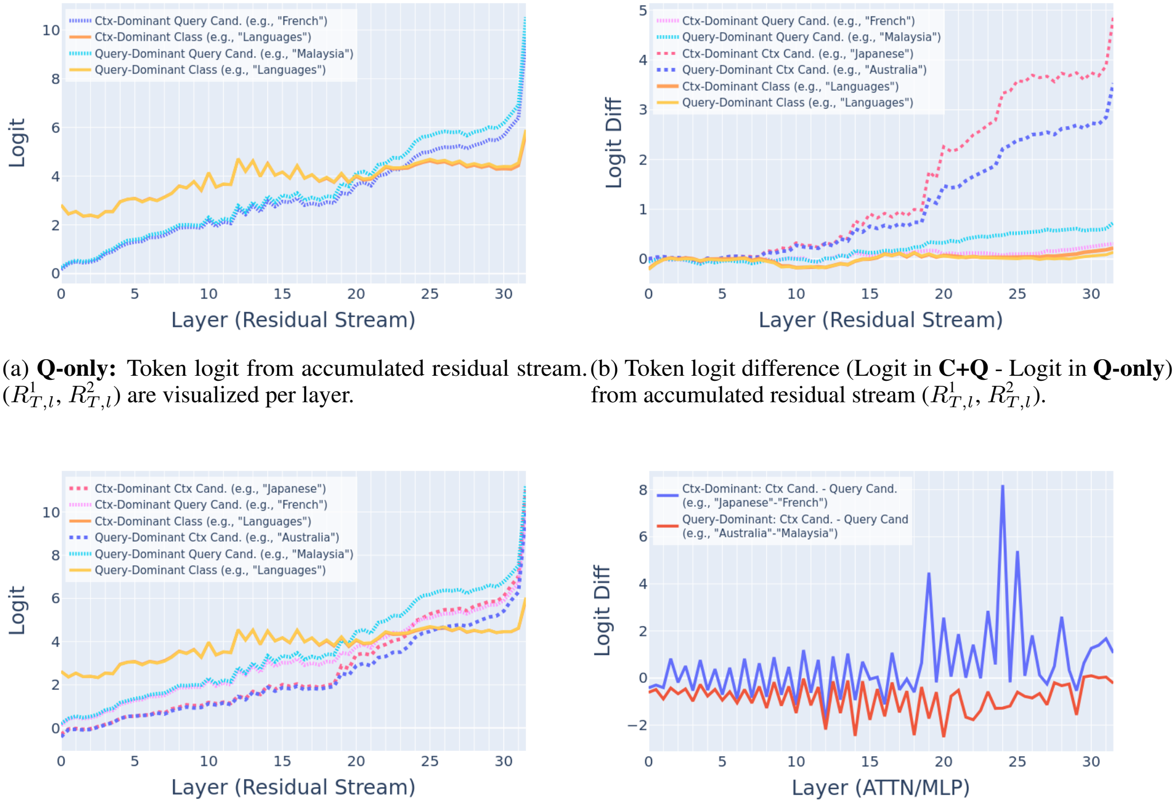
- Mistral: 437/1000 context-dominant datapoints flip to query-based candidates, while 514/1000 remain context-based. Similarly, 232/1000 query-dominant datapoints shift to context-based candidates, while 713/1000 remain query-based.

- Pythia: 470/1000 context-dominant datapoints flip to query-based candidates, while 486/1000 remain context-based. Conversely, 294/1000 query-dominant datapoints shift to context-based candidates, while 648/1000 remain query-based.

Across all models, approximately 950 datapoints remain context- or query-based candidates, instead of random non-identifiable answers, indicating that our intervention preserves model capabilities.

	Orig.	L17+L24		2 Low		2 High	
	Prob.	Prob.	Δ	Prob.	Δ	Prob.	Δ
Context-Dominant							
Ctx	22.6	**14.6**	-8.0	19.9	-2.7	19.0	-3.6
Query	8.2	**12.4**	+4.2	8.2	+0.0	9.2	+1.0
Query-Dominant							
Query	33.0	**25.0**	-8.0	25.5	-7.5	31.7	-1.3
Ctx	6.5	**10.7**	+4.2	7.7	+1.2	6.4	-0.1

Table 11: Effect of attention knockout on context- (Ctx) and query-based (Query) candidate probabilities on Llama-3. "Orig." = No intervention, "2 Low" = Two lower layers (<17), "2 High" = Two higher layers (>24). "Diff." represents the probability difference, and Δ denotes the change from the original setting. **(Mistral 7B)**



(a) **Q-only:** Token logit from accumulated residual stream. ($R^1_{T,l}$, $R^2_{T,l}$) are visualized per layer.

(b) Token logit difference (Logit in **C+Q** - Logit in **Q-only**) from accumulated residual stream ($R^1_{T,l}$, $R^2_{T,l}$).

(c) **C+Q**: Token logit from accumulated residual stream. ($R^1_{T,l}$, $R^2_{T,l}$) are visualized per layer.

(d) **C+Q**: candidate logit difference from attention and MLP output. $A_{T,l}$ and $M_{T,l}$ are visualized per layer.

Figure 8: Logit Attribution Results For **Pythia 6.9B**.

	Orig.	L17+L24		2 Low		2 High	
	Prob.	Prob.	Δ	Prob.	Δ	Prob.	Δ
Context-Dominant							
Ctx	25.5	**13.1**	-12.4	20.9	-4.6	21.1	-4.4
Query	8.6	**14.8**	+6.2	8.9	+0.3	12.6	+4.0
Query-Dominant							
Query	35.2	**26.8**	-8.4	25.7	-9.5	33.4	-1.8
Ctx	6.6	**11.3**	+4.7	7.7	+1.1	7.1	+0.5

Table 12: Effect of attention knockout on context- (Ctx) and query-based (Query) candidate probabilities on Llama-3. “Orig.” = No intervention, “2 Low” = Two lower layers (<17), “2 High” = Two higher layers (>24). “Diff.” represents the probability difference, and Δ denotes the change from the original setting. **(Llama-3)**

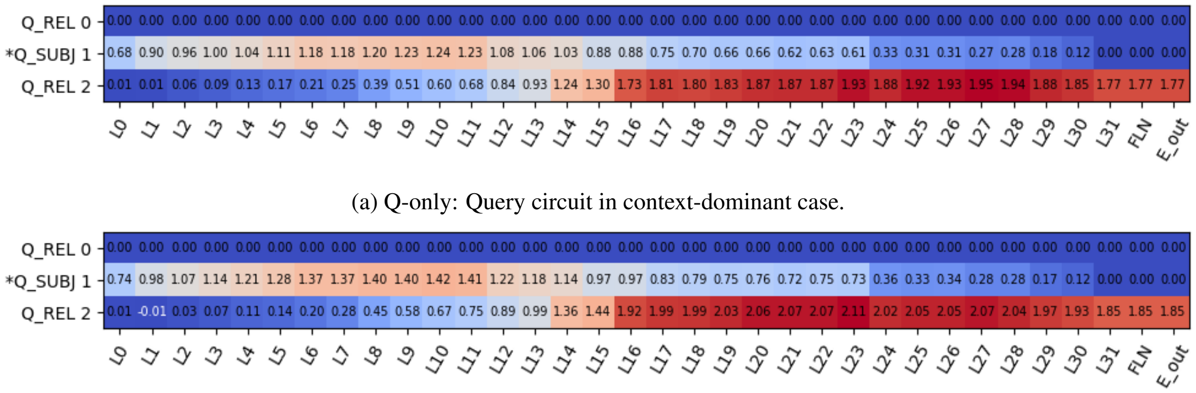
	Orig.	L17+L24		2 Low		2 High	
	Prob.	Prob.	Δ	Prob.	Δ	Prob.	Δ
Context-Dominant							
Ctx	22.3	**13.6**	-8.7	18.3	-4.0	20.5	-1.8
Query	7.4	**11.5**	+4.1	8.4	+1.0	7.8	+0.4
Query-Dominant							
Query	26.6	**20.8**	-5.8	20.5	-6.1	25.6	-1.0
Ctx	6.3	**9.6**	+3.3	6.9	+0.6	6.2	-0.1

Table 13: Effect of attention knockout on context- (Ctx) and query-based (Query) candidate probabilities on Llama-3. “Orig.” = No intervention, “2 Low” = Two lower layers (<17), “2 High” = Two higher layers (>24). “Diff.” represents the probability difference, and Δ denotes the change from the original setting. **(Pythia)**

I Ablation Studies

I.1 Experimental Details

Due to computational constraints, we cannot inference on the full $C + Q$ dataset with 102M for larger-sized model, we therefore conduct sampling as follows: For Pythia-12B, we first randomly sample 1,000 datapoints from the ParaRel dataset, then randomly sample 100 datapoints per relation for the context demonstrations, resulting in around 3.9M datapoints. For Llama-70B, we first randomly sample 500 datapoints from the ParaRel dataset, then randomly sample 50 datapoints per relation for the context demonstrations, resulting in around 975K datapoints.



(a) Q-only: Query circuit in context-dominant case.

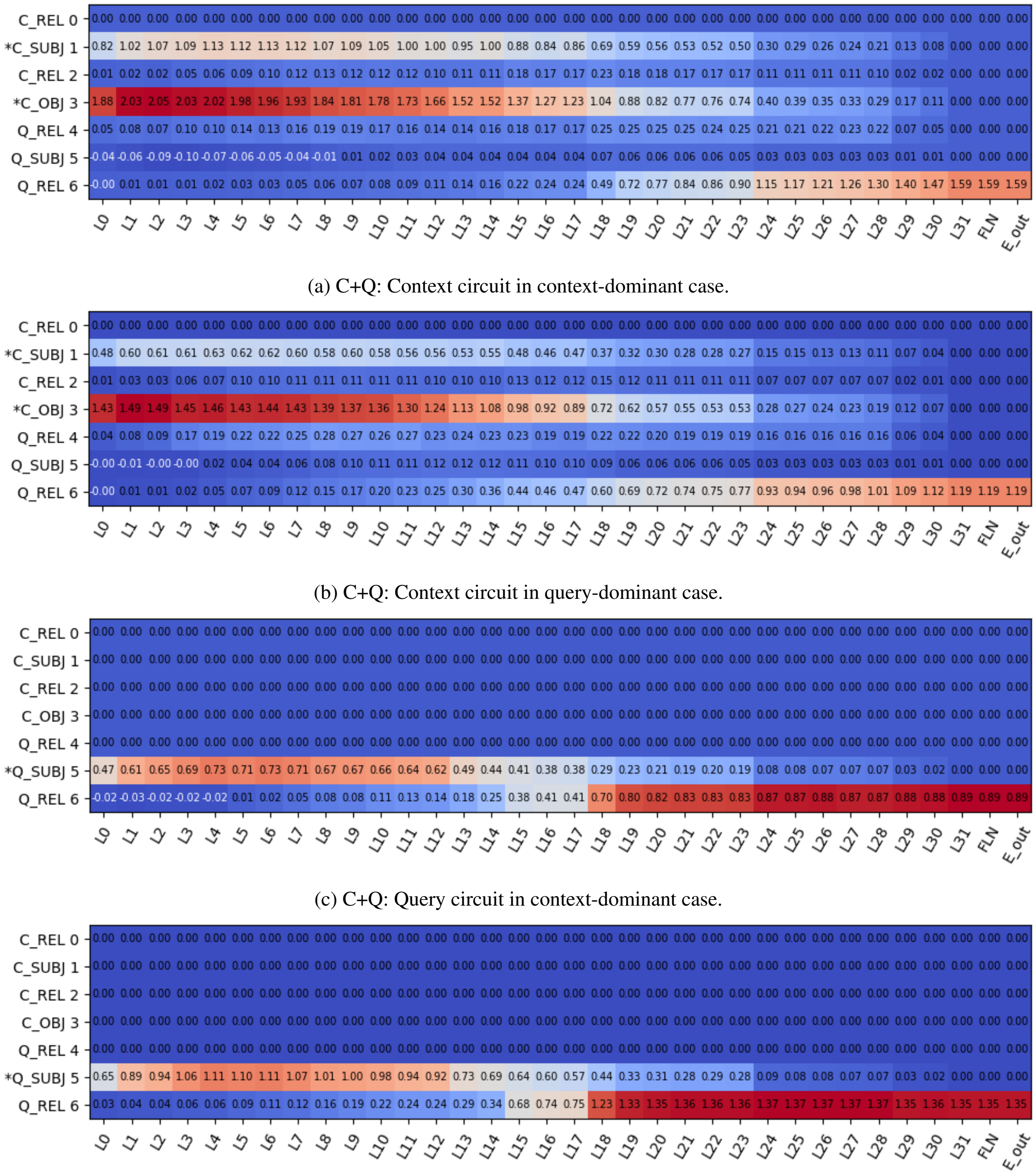
(b) Q-only: Query circuit in query-dominant case.

Figure 9: Activation patching under Q-Only condition reveals that query circuit is the same before and after adding the irrelevant context for **Llama-3 8B**.

I.2 Statistical Validation of Contextual Influence

Value	Pythia 12B	Llama-3 70B	Llama-3 8B Prompt
Mean PMI	4.19	4.22	4.2756
T-statistics	11.74	11.89	13.0924
p-value	0.0002	0.0001	0.0001

Table 14: Statistical analysis of PMI between context terms and generated context-based candidates.



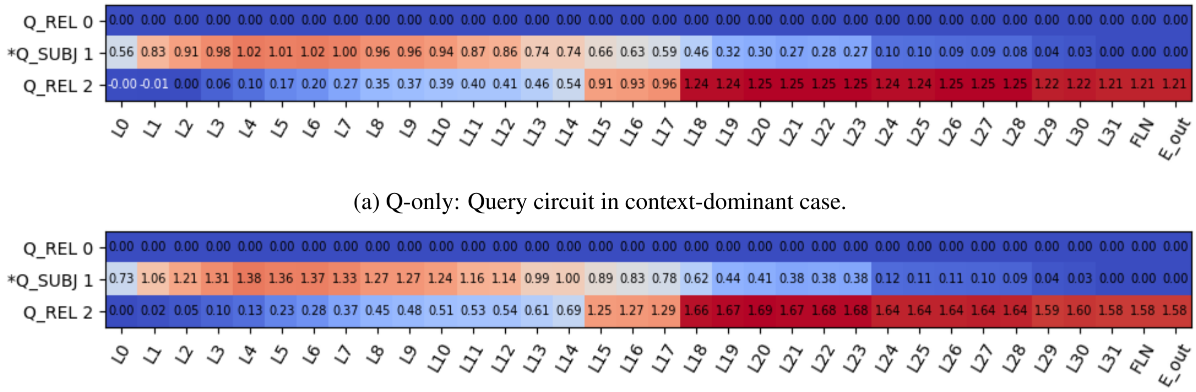
(a) C+Q: Context circuit in context-dominant case.

(b) C+Q: Context circuit in query-dominant case.

(c) C+Q: Query circuit in context-dominant case.

(d) C+Q: Query circuit in query-dominant case.

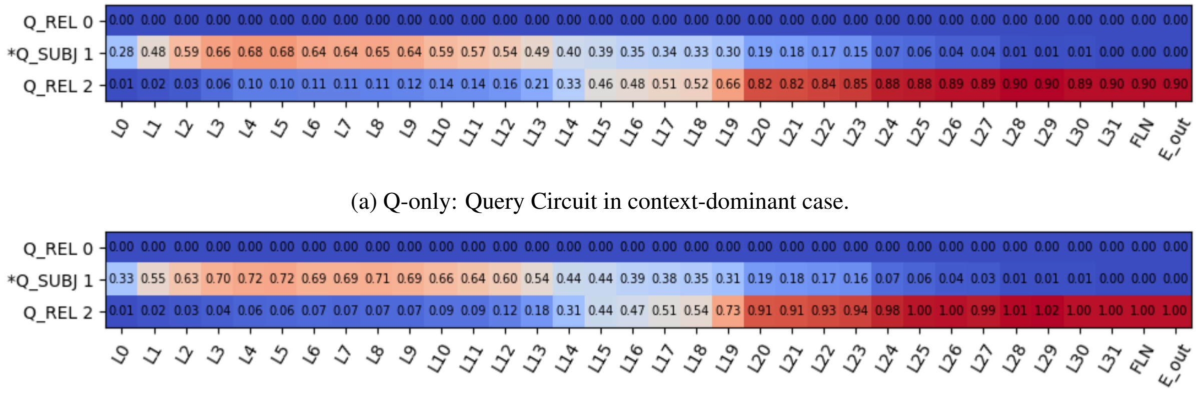
Figure 10: Activation patching under C+Q condition for **Mistral 7B**.



(a) Q-only: Query circuit in context-dominant case.

(b) Q-only: Query circuit in query-dominant case.

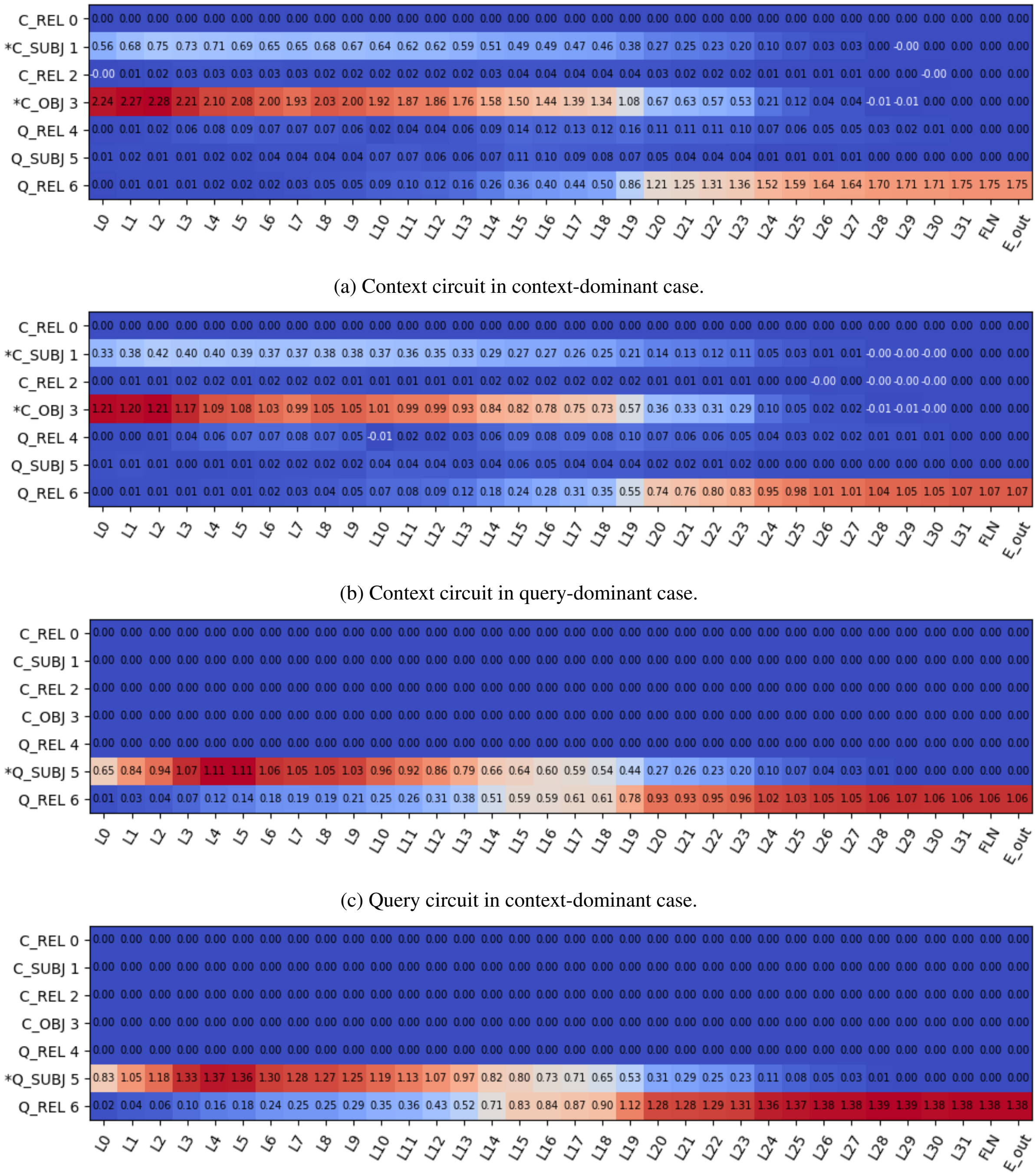
Figure 11: Activation patching under Q-Only condition reveals that query circuit is the same before and after adding the irrelevant context for **Mistral 7B**.



(a) Q-only: Query Circuit in context-dominant case.

(b) Q-only: Query Circuit in query-dominant case.

Figure 12: Activation patching under Q-Only condition reveals that query circuit is the same before and after adding the irrelevant context for **Pythia 6.9B**.



(a) Context circuit in context-dominant case.

(b) Context circuit in query-dominant case.

(c) Query circuit in context-dominant case.

(d) Query circuit in query-dominant case.

Figure 13: Activation patching under C+Q condition for **Pythia 6.9B**.

Relation	Template	Ctx Type	Total Rows
P1001	[X] is a legal term in [Y]	Place	664
P101	The expertise of [X] is [Y].	Others	571
P103	The mother tongue of [X] is [Y].	Language	919
P106	[X] works as [Y].	Job	821
P108	[X], who is employed by [Y].	Company	378
P127	[X] owner [Y].	Company	616
P1303	[X] plays the [Y].	Others	513
P131	[X] is in [Y].	Place	775
P136	[X] plays [Y].	Others	859
P1376	[X], the capital city of [Y].	Place	179
P138	[X], which is named after [Y].	Others	461
P140	[X] is follower of [Y].	Others	432
P1412	[X] communicated in [Y].	Language	924
P159	[X] is headquartered in [Y].	Place	801
P17	[X], located in [Y].	Place	912
P176	[X], produced by [Y].	Company	925
P178	[X], a product developed by [Y].	Company	588
P19	[X] is native to [Y].	Place	779
P190	[X] is a twin city of [Y].	Place	671
P20	[X] passed away at [Y].	Place	817
P264	[X]'s label is [Y].	Company	53
P27	[X], a citizen of [Y].	Place	958
P276	[X] is located in [Y].	Place	764
P279	[X], a type of [Y].	Others	900
P30	[X] is a part of the continent of [Y].	Place	959
P36	The capital city of [X] is [Y].	Place	471
P361	[X] is a part of [Y].	Others	746
P364	The original language of [X] was [Y].	Language	756
P37	The official language of [X] is [Y].	Language	900
P39	[X], who holds the position of [Y].	Job	485
P407	[X] was written in [Y].	Language	857
P413	[X] plays in the position of [Y].	Job	952
P449	[X] premiered on [Y].	Company	801
P463	[X] belongs to the organization of [Y].	Company	203
P47	[X] borders with [Y].	Place	649
P495	[X] was formed in [Y].	Place	905
P530	[X] ties diplomatic relations with [Y].	Place	950
P740	[X], founded in [Y].	Place	843
P937	[X] found employment in [Y].	Place	853

Table 15: Overview of Relations, Templates, Types, and Total Rows in the original Pararel Dataset. We take this dataset and construct the $C + Q$ dataset, which has around 106.6M rows.

Relation	Template	Ctx Type	Total Rows
P1001	[X] is a legal term in [Y].	Place	664
P101	[X]'s expertise is [Y].	Others	571
P103	The mother tongue of [X] is [Y].	Language	919
P106	The occupation of [X] is [Y].	Job	821
P108	[X] works for [Y].	Company	378
P127	[X] is owned by [Y].	Company	616
P1303	[X] plays [Y].	Others	513
P131	[X] is in [Y].	Place	775
P136	[X], who plays [Y].	Others	859
P1376	[X], the capital city of [Y].	Place	179
P138	[X] is called after [Y].	Others	461
P140	[X] is follower of [Y].	Others	432
P1412	[X] communicated in [Y].	Language	924
P159	The headquarters of [X] is in [Y].	Place	801
P17	[X], which is located in [Y].	Place	912
P176	[X], developed by [Y].	Company	925
P178	[X], created by [Y].	Company	588
P19	[X] originates from [Y].	Place	779
P190	[X] is a twin city of [Y].	Place	671
P20	[X] died in [Y].	Place	817
P264	[X], which is represented by [Y].	Company	53
P27	[X] has a citizenship of [Y].	Place	958
P276	[X] is in [Y].	Place	764
P279	[X], a type of [Y].	Others	900
P30	[X] belongs to the continent of [Y].	Place	959
P36	The capital of [X] is [Y].	Place	471
P361	[X] is part of [Y].	Others	746
P364	The original language of [X] was [Y].	Language	756
P37	The official language of [X] is [Y].	Language	900
P39	[X], whose position is that of [Y].	Job	485
P407	The language of [X] is [Y].	Language	857
P413	[X] plays in the position of [Y].	Job	952
P449	[X] debuted on [Y].	Company	801
P463	[X] is a member of [Y].	Company	203
P47	[X] borders with [Y].	Place	649
P495	[X] formed in [Y].	Place	905
P530	[X] maintains diplomatic relations with [Y].	Place	950
P740	[X], founded in [Y].	Place	843
P937	[X] was employed in [Y].	Place	853

Table 16: Overview of relations, alternative prompt templates, types, and total rows in the original ParaRel dataset. We use these randomly sampled alternative templates to test sensitivity to prompt phrasing.

Category	**Details**
Example 1	
Context	**Hanoi** is a twin city of **Bangkok**.
Query	The **mother tongue** of Louis Legendre is
Class	Languages
Context Subject Possible Answers	Vietnamese, Tay, Hmong, Khmer, English, French, Chinese
Context Object Possible Answers	Thai, Lao, Chinese, Malay, Khmer
Context-Based Candidates	Vietnamese, Thai
Context-Influenced?	True
Correct Class?	True
Exists Answer that Satisfies Both?	True
Example 2	
Context	**Svend Asmussen** plays the **violin**.
Query	Social-Economic Council is a legal term **in**
Class	Places (Countries, Cities, States, etc.)/Languages
Context Subject Possible Answers	Danmark, Danish (Svend Asmussen is a Violinist and jazz musician)
Context Object Possible Answers	Italy, Italian (Violin was originated in Italy)
Context-Based Candidates	Denmark
Context-Influenced?	True
Correct Class?	True
Exists Answer that Satisfies Both?	True
Example 3	
Context	**Manchester Business School** is headquartered in **Manchester**.
Query	Antipope Paschal III, who holds the **position** of
Class	Jobs/Positions/Roles
Context Subject Possible Answers	Professor, Lecturer, Instructor, Researcher, Department Chair, Provost, Dean, Academic Advisor, Teaching Assistant, Student, etc.
Context Object Possible Answers	N/A
Context-Based Candidates	Dean, Professor
Context-Influenced?	True
Correct Class?	True
Exists Answer that Satisfies Both?	True

Category	**Details**
Example 4	
Context	**Saharsa district** is in **Bihar**.
Query	Colbert Mountains is a part of the **continent** of
Class	Continents/Places
Context Subject Possible Answers	Asia
Context Object Possible Answers	Asia
Context-Based Candidates	Asia, India
Context-Influenced?	True
Correct Class?	True
Exists Answer that Satisfies Both?	True
Example 5	
Context	**Lee Jong-hyun** plays the **guitar**.
Query	Northern Foothills is a part of the **continent** of
Class	Continents
Context Subject Possible Answers	Asia
Context Object Possible Answers	Europe (Guitar originated in Spain)
Context-Based Candidates	South, Korea
Context-Influenced?	True
Correct Class?	False
Exists Answer that Satisfies Both?	False
Example 6	
Context	Grant Green plays jazz.
Query	David Gates plays the
Class	Role/Genre/Style/Position/Musical Instrument
Context Subject Possible Answers	guitarist, composer, musician, songwriter etc. (role of Grant Green), guitar (Musical Instrument that Grant Green plays), jazz, R&B, etc. (music genre of Grant Green)
Context Object Possible Answers	Jazz.
Context-Based Candidates	Bee, Beach
Context-Influenced?	False
Correct Class?	False
Exists Answer that Satisfies Both?	False
Example 7	

Category	**Details**
Context	**Samuil Marshak** passed away at **Moscow**.
Query	Jean Metcalfe, who is employed by
Class	Company/Person
Context Subject Possible Answers	Russia-1, Channel One Russia, RT, TV Rain, etc.
Context Object Possible Answers	Russia-1, Channel One Russia, RT, TV Rain, etc.
Context-Based Candidates	BBC, Radio
Context-Influenced?	False
Correct Class?	True
Exists Answer that Satisfies Both?	False