Understanding Internal Knowledge Conflict Resolution Mechanism in Language Models

Anonymous ACL submission

Abstract

Knowledge conflict degrades the accuracy and reliability of Language Models (LMs), yet how a model internally resolves such conflict remains unexplored. We investigate intra-memory and context-memory conflict resolution and link the mechanism to prompt sensitivity. Using a synthetic dataset, we apply mechanistic interpretability to GPT-2 XL and validate on Qwen3-4B. The model aggregates signals across all components, with the dominant contributor shifting under conflict. In non-conflict settings, content-sensitive attention heads produce strong discriminative signals. Under conflict, these signals weaken and the model becomes sensitive to every component. Using activation rates and contribution magnitude, we identify five functional head types (*inactive*, *selective*, *selective-suppressor*, *suppressor* and *broadcaster*), and find that context-memory conflict recruits previously inactive heads. Prompt sensitivity follows directly, since weakened discriminative signals let small prompt changes shift outputs.[1]

1 Introduction

Language models (LMs) acquire parametric knowledge through pretraining on large-scale corpora. Such corpora can not be curated perfectly and therefore inevitably contain inconsistencies, outdated information, and misinformation (Chen et al., 2025). These imperfections can give rise to knowledge conflict, where different knowledge sources support incompatible answers (Longpre et al., 2021). Prior works show that such conflicts can affect model behavior in multiple ways, including reliance on conflicting contextual or parametric knowledge, susceptibility to misinformation, and divergent responses to semantically equivalent prompts (Xu et al., 2024; Zhuo et al., 2024). Knowledge conflict is commonly grouped into three types based on the source of the conflicting information, namely context-memory, intra-memory and inter-context conflict (Xu et al., 2024). Among these three types, context-memory and intra-memory conflict exposes how parametric knowledge is retrieved and resolved inside the model. However, the internal mechanisms by which LMs resolve such conflicts remain underexplored. Existing mechanistic work (Yu et al., 2023; Ortu et al., 2024) concentrates on context-memory conflict, leaving intra-memory conflict largely unexplored.

In this paper, we focus primarily on intra-memory knowledge conflicts, which arise from inconsistencies within an LM's parametric knowledge. To study these conflicts in a controlled setting, we construct a curated synthetic dataset in which conflicting facts are deliberately introduced (Section 3.1). This design allows us to isolate the target phenomenon while minimizing the risk of data contamination from pretraining corpora. We then continue pretraining GPT-2 XL (Radford et al., 2019) on this dataset and use mechanistic analysis of LM internals (Section 3.2) to investigate how the model resolves such conflicts internally (Section 4). In addition, we compare intra-memory conflicts with context-memory conflicts (Section 5), and show how prompt sensitivity arises from competing parametric knowledge (Section 6). Through this analysis, we provide a mechanistic account of intra-memory conflict resolution in language models. The central finding is that conflict narrows the logit gap between top candidates. In non-conflict cases, content-sensitive attention heads, heads that attends to and extracts discriminative signals from the prompt, steers the model toward the correct answer. Under intra-memory conflict, the model's prediction is no longer determined by a dominant signal, and even small contributions toward alternative candidates can shift the final output.

As our analysis reveals that attention heads play a major role in deciding model's output, to char-

[1]We will release our code and data upon acceptance.

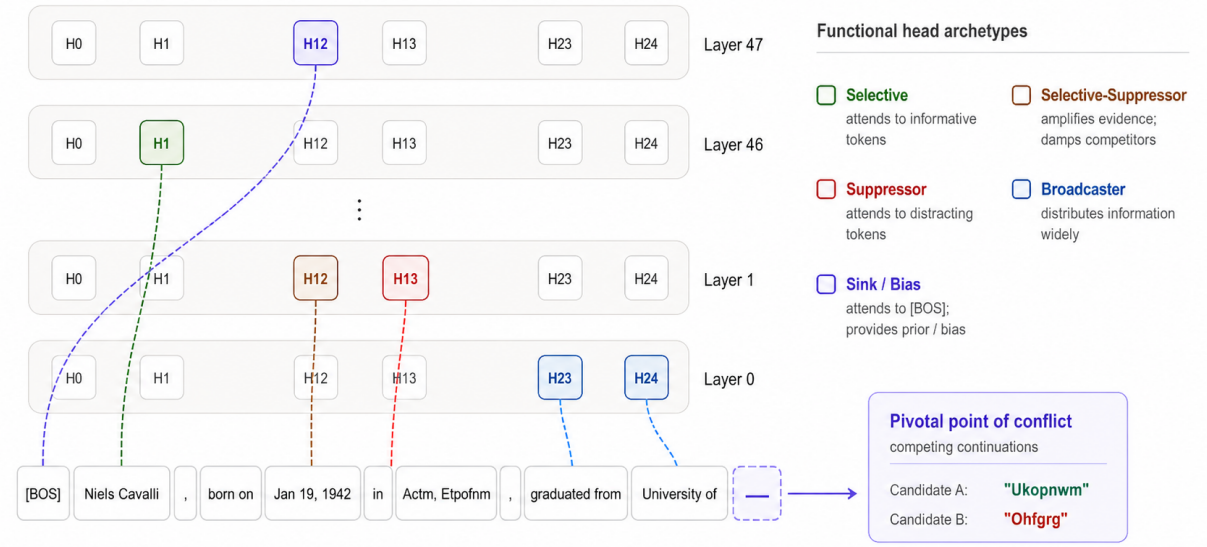

Figure 1: An example of Intra-memory knowledge conflict. Based on the query, the model resolves the knowledge conflict internally to choose an answer. Different types of attention heads pay attention to different tokens and make their own contributions to the final output.

acterize their functional roles we introduce a five-class taxonomy (as shown in Figure 1): ***inactive***, ***selective***, ***selective-suppressor***, ***suppressor***, and ***broadcaster***. The taxonomies are stratified by how each head contributes to prediction across varying levels of prompt informativeness. Group-wise ablation experiments demonstrate that ***selective*** heads, which promote specific candidate tokens, are primarily responsible for the model's predictions: perturbing their activations causes substantial shifts in the output distribution. We further show that this taxonomy is preserved across different architecture as well as in context-memory conflicts. In context-memory conflict, the model pulls many previously inactive heads into active processing and alters their attention patterns. Finally, we show that prompt sensitivity emerges when intra-memory conflict weakens content-sensitive heads, allowing small perturbations to shift the model preferences between candidate tokens. Replication on Qwen3-4B confirms that these behaviors are not specific to GPT-2 XL, but generalize across model architectures. Together, these results provide, to our knowledge, the first **mechanistic account of how LMs navigate intra-memory knowledge conflict**, complementing prior work on context-memory conflict (Jin et al., 2024; Li et al., 2025; Ortu et al., 2024), by showing how competing parametric knowledge reshapes attention-head function, static-component reliance, and prompt sensitivity.

2 Related Work

Knowledge conflict occurs when a language model receives competing knowledge signals for the same query. Parametric knowledge encoded during pretraining can conflict with external context introduced at inference time, such as retrieved documents from a Retrieval-Augmented Generation (RAG) system (Lewis et al., 2020; Hua et al., 2019) or user-provided passages. The pretraining corpus itself can also contain contradictory facts and thus produce inconsistencies within parametric knowledge alone, even when the source is trusted (Hou et al., 2024). Further causes include outdated information, misinformation and post-training interventions like knowledge editing (De Cao et al., 2021). Xu et al. (2024) group the resulting conflicts into three types, namely intra-memory, context-memory and inter-context conflict, each leaving the model with competing signals it must resolve into a single answer. We study the internal computation behind this resolution.

Attention sink refers to the phenomenon whereby initial tokens accumulate disproportionately large attention weights across many layers and heads, regardless of their semantic relevance to the query (Xiao et al., 2024). These *sink tokens* are often lack of semantic significance, and act as stability anchor (Xiao et al., 2024) or no-op tokens (Cancedda, 2024). The pattern is stable across the forward pass. Once a token assumes the sink role, it persistently attracts attention even

as contextual representations evolve through the layers. Despite of receiving high attention, these tokens are often associated with limited task associated relevancy (Bondarenko et al., 2023; Su et al., 2026). In LLMs, it is defined as unbalanced attention mass, regardless of semantic relevance (Xiao et al., 2024).

In-context learning is the post-training technique that utilized Language Model's capability to learn to perform a new task by reading a few examples provided directly in the prompt without any further continue pre-training. This concept is first introduced by Brown et al. (2020) and have been widely adopted and studied ever since (Rubin et al., 2022; Olsson et al., 2022; Dong et al., 2024, among others). **Induction heads** were first proposed by Olsson et al. (2022), in which the author presents evidences and argues that they are the main mechanistic source of in-context learning. When model sees a token it has encountered before in the current input (namely token A), an induction head looks back at what came right after that earlier occurrence (token B) and predicts the same following token. In short, when the model recognizes the pattern "AB...A" exist within the input, induction heads at the later position of A look at previous occurrence of token B and promote B as the top candidate output. By definition of Olsson et al. (2022), induction heads implement an in-context pattern-completion mechanism characterized by two properties: *prefix matching*, whereby the head locates earlier occurrences of the current token in the context and attends to the token that immediately followed it, and *copying*, whereby the head pushes the model toward predicting the token it was attending to.

3 Experimental Setup

3.1 Dataset and Task

Dataset. For this study, we construct a synthetic dataset of Wikipedia-style short biographies, consisting entirely of fictional individuals. Each individual is described by a fixed set of personal attributes: *birth date*, *birth place*, *university*, *major*, *company*, and *work place* (Allen-Zhu and Li, 2024; Li et al., 2024; Zucchet et al., 2025). We follow the generation process of Li et al. (2024) and use GPT-4o (Hurst et al., 2024) to generate the biographies. To reduce dependence on a single surface form and encourage the model to learn entity–attribute associations, each biography is instantiated from a template randomly selected from a predefined template set.

A key goal of our dataset design is to introduce conflicts in a controlled way while minimizing possible contamination from the original pretraining corpus. To this end, entity names for important attributes, including *birth place*, *company*, and *university*, are replaced with randomly generated words. Less central attributes, including *major* and *work place*, are kept unchanged. The conflicting attribute is either *university* or *company*. Thus, for a conflicted individual, the paired biographies agree on all non-target attributes but assign different values to the target attribute. Appendix A provides more details on the dataset generation process and statistics.

Formally, let $\mathcal{B}$ denote a dataset comprising personal biographies. An instance $b_i \in \mathcal{B}$ is said to contain a knowledge conflict if there exists another biography $\bar{b}_i \in \mathcal{B}$ such that b_i and $\bar{b}_i$ refer to the same individual p_i and share identical personal attributes except for one conflicting element. The two biographies therefore encode two incompatible facts about the same person. In contrast, a conflict-free instance $b_i \in \mathcal{B}$ means that b_i is the only biography in $\mathcal{B}$ describing person p_i, so no contradictory fact is introduced for that individual.

From $\mathcal{B}$, we construct a mixed continued-pretraining dataset $\mathcal{B}_{mix}$ containing biographies for 1000 individuals with knowledge conflicts (n_1) and an additional 1000 individuals with conflict-free biographies (n_2). For each conflicted individual, $\mathcal{B}_{mix}$ includes paired biographies expressing contradictory target facts; for each conflict-free individual, it includes only a single biography. We use $\mathcal{B}_{mix}$ for continued pretraining so that both consistent and conflicting facts are encoded into the model's parametric knowledge. Additional details of our dataset and the fine-tuning hyperparameters can be found in Appendix A.

Task. We analyze knowledge conflict through a controlled next-token completion task. For each individual p_i, we construct a prompt pr_i by inserting the individual's name and related biographical information into a template that ends immediately before the target attribute value, e.g., "Niels Cavalli studies at University of ___." The model is then asked to predict the next token corresponding to the missing attribute value. For a conflicted individual, pr_i elicits two incompatible candidate continuations from the contradictory biography pair $(b_i, \bar{b}_i)$,

such as two different universities or companies. For a conflict-free individual, the same prompt format has only one valid continuation and serves as a comparison case.

3.2 Internal Assessment Method

Let ℓ denote the model under analysis. We first continue pre-training ℓ with $\mathcal{B}_{mix}$ to embed the conflicts into ℓ's parametric knowledge. Afterward, we analyze both the conflict subset n_1 of biographies contain a contradictory pair $(b_i, \bar{b}_i)$ and the second subset n_2 of conflict-free biographies is used for comparison. All experiments use GPT-2 XL (Radford et al., 2019) (48 layers $\times$ 25 heads per Attention sublayers, 1200 heads total) across both the university and company domains; Section 4.4 replicates the main findings on Qwen3-4B (Yang et al., 2025).

Following established mechanistic interpretability studies (Elhage et al., 2021; Hanna et al., 2023; Ortu et al., 2024), we analyze the internal computation induced by the completion prompts using the task defined in Section 3.1. For each prompt pr_i, we inspect how model components contribute to the logits of the candidate tokens. This enables us to trace how the model builds confidence toward different candidates and to attribute that confidence to specific components.

Specifically, querying for the conflict information of a person p_i in ℓ produces a final hidden state h_i, which can be converted to the model logits y_i with unembedding matrix $W_u \in \mathbb{R}^{d_{\text{model}} \times d_{\text{vocab}}}$ and model bias $\mathbf{b}_u \in \mathbb{R}^{d_{\text{vocab}}}$ through the transformation:

$$y_i = (h_i)_{ln} W_u + \mathbf{b}_u \tag{1}$$

with $(\cdot)_{ln}$ denote the value scaled with layer-norm. Let $\delta_i^c \in \mathbb{R}^{d_{\text{model}}}$ be the contribution of component c to hidden state h_i, the final hidden state $h_i = \sum \delta_i^c$ is a sum of all contributions from components $c \in \mathcal{C}$, where $\mathcal{C}$ is the set of components (attention heads, MLPs, and the embedding) (Nostalgebrist, 2020). We have the decomposition:

$$\begin{aligned} y_i &= (h_i)_{ln} W_u + \mathbf{b}_u \\ &= \left(\sum_{c \in \mathcal{C}} \delta_i^c \right)_{ln} W_u + \mathbf{b}_u \\ &\approx \sum_{c \in \mathcal{C}} (\delta_i^c)_{ln} W_u + \mathbf{b}_u \\ &\approx \sum_{c \in \mathcal{C}} \hat{\delta}_i^c + \mathbf{b}_u \end{aligned} \tag{2}$$

with $\hat{\delta}_i^c \in \mathbb{R}^{d_{\text{vocab}}}$ be the contribution of component c to the vocabulary, and $\hat{\delta}_i^c(t)$ is the direct contribution of c to token t.

For each prompt pr_i's forward pass, we partition attention heads into *inactive heads* (attention-sinking heads) and *active heads* (content-sensitive heads). We consider any head paying attention to the initial token (P_0 sink) with high attention scores (> 0.5, guaranteeing the highest attention among the vocabulary) as inactive head (Xiao et al., 2024; Gu et al., 2025), and the rest are considered active heads. Including MLP sublayers and the model's bias term $\mathbf{b}_u$, we will examine the contribution of these four contribution groups in Section 4. In the case that biography b contain knowledge conflict, we look at the first tokens of the two knowledge in b and $\bar{b}$ that contradict with each other, and in case of conflict-free b, we choose the tokens with highest and second highest logits to inspect. We randomly assign these two tokens as $t_{i,1}$ and $t_{i,2}$ per pr_i and measure the contribution gaps of each $c \in \mathcal{C}$:

$$\Delta\hat{\delta}_i^c = \hat{\delta}_i^c(t_{i,1}) - \hat{\delta}_i^c(t_{i,2}) \tag{3}$$

and the logit gap:

$$\Delta y_i = y_i(t_{i,1}) - y_i(t_{i,2}) \tag{4}$$

We first compare the average absolute contribution gaps $\mu(|\Delta\hat{\delta}^G|)$ of each group $G \in \{$*inactive heads, active heads, MLP, model bias*$\}$ in conflict cases ($p_i \in n_1$) versus non-conflict cases ($p_i \in n_2$). After that, to quantify how strongly each component group drives the final decision, we correlate each group's gap $\Delta\hat{\delta}_i^G$ against the logit gap Δy_i across instances using the following two measures:

- **Pearson** r captures *directional alignment* ($\Delta\hat{\delta}_i^G > 0 \sim \Delta y_i > 0$): whether a group's preference aligns with the model's final choice.
- **Spearman** ρ captures *magnitude alignment* ($\Delta\hat{\delta}_i^G \sim \Delta y_i$): whether a group contributes more strongly would also yield a larger final token gap.

4 How GPT-2 XL Resolve Knowledge Conflicts

4.1 Intra-Memory Conflict

After fine-tuning GPT-2 XL on $\mathcal{B}_{mix}$, we apply the procedure of Section 3.2 to characterize component-level contributions to prediction under

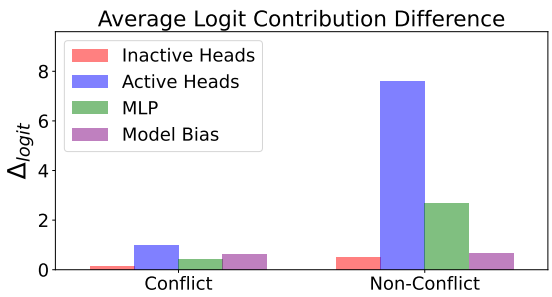

Figure 2: Average logit contribution difference for conflict and non-conflict tokens for different model components.

	Metric	Inactive heads	Active heads	MLP	Model bias
Conflict	Pearson r (↑)	0.170	−0.135	0.152	**0.490**
	p-val (↓)	$<1e^{-5}$	$<1e^{-5}$	$<1e^{-5}$	$<1e^{-5}$
	Spearman ρ (↑)	0.194	0.007	0.196	**0.608**
	p-val (↓)	$<1e^{-5}$	0.826	$<1e^{-5}$	$<1e^{-5}$
Non-conf.	Pearson r (↑)	0.028	**0.802**	0.660	−0.082
	p-val (↓)	0.382	$<1e^{-5}$	$<1e^{-5}$	0.009
	Spearman ρ (↑)	0.094	**0.921**	0.725	−0.169
	p-val (↓)	0.003	$<1e^{-5}$	$<1e^{-5}$	$<1e^{-5}$

Table 1: Pearson and Spearman correlations between component contributions under conflict vs. non-conflict. Model bias shows the strongest correlation under conflict; active heads dominate under non-conflict.

intra-memory conflict. For each prompt, we compute the contribution gap between the two candidate target tokens and decompose it across four sources: attention-sink (inactive) heads, active attention heads, MLP sublayers, and the model bias term. This decomposition tests whether the final decision is driven by content-sensitive computation (active heads and MLPs) or by static components (attention-sink heads and the bias term).

Figure 2 reports mean contribution gaps across the four groups under conflict and non-conflict conditions. In the non-conflict setting, active heads produce large, consistent gaps sufficient to determine the predicted token on their own; under conflict these gaps collapse, indicating that active heads can no longer discriminate between the competing entities. MLP sublayers exhibit the same trend with reduced magnitude.

Table 1 confirms this shift quantitatively. Without conflict, active-head contributions correlate strongly with the final decision (Spearman ρ=0.921, Pearson r=0.802), MLPs serve as a secondary signal (ρ=0.725, r=0.660), and inactive heads are uncorrelated ($\rho, r \approx 0$). Under conflict, active heads become statistically uninformative ($\rho \approx 0$, p=0.826), and the model's bias term emerges as the dominant predictor (ρ=0.608), followed by MLPs (ρ=0.196) and inactive heads (ρ=0.194). Conflict thus suppresses the discriminative capacity of content-sensitive heads and makes predictions sensitive to small perturbations across remaining components.

4.2 Attention Head Taxonomy Under Conflict

While the aggregate results described above provide information on which component group drives conflict resolution, the behavior of the content-sensitive heads remains unclear. Therefore, we further classify 1200 heads in both domains (*university*, *company*)

For each head c_h, we record its contribution vectors across the (n_1) conflict instances. Details of our methods are presented in the Appendix B.

The mean contribution of c_h over n_1 is denoted as $\bar{\delta}^{c_h}$, and ***z-score*** $z_{c_h}(t)$ is the statistic analysis of the mean contribution from head c_h to an output token t derived from $\bar{\delta}^{c_h}$, $\mu(\cdot)$ and $\sigma(\cdot)$ are mean and standard deviation:

$$z_{c_h}(t) = \frac{\bar{\delta}^{c_h}(t) - \mu(\bar{\delta}^{c_h})}{\sigma(\bar{\delta}^{c_h})} \tag{5}$$

A z-score $|z_{c_h}(t)| > 2$ indicates that the contribution to token t lies more than two standard deviations away from the head's mean contribution, which is considered a statistical outlier. It reveals the head's preference: which output tokens it consistently targets or suppresses. We calculate z-score of the averaged contribution $\bar{\delta}^{c_h}$ as this captures consistent token preferences across instances. Discarding the inactive heads, the remaining heads fall into four types:

Selective ($\max z_{c_h} > 2$): at least one token consistently receives a strong positive contribution, indicating that the head concentrates its contribution on a small subset of the output space. These are content-sensitive decision heads that selectively promote specific candidate tokens.

Selective-suppressor ($\max |z_{c_h}| > 2$ but $\max z_{c_h} \leq 2$): at least one token consistently receives a strong negative contribution, but none receives a strong positive one. These heads concentrate suppressive contribution on a subset of tokens, dampening specific candidates without promoting any.

Suppressor ($\max |z_{c_h}| \leq 2$, $\mu(\bar{\delta}^{c_h}) < 0$): no outlier tokens, but contribute negatively on average.

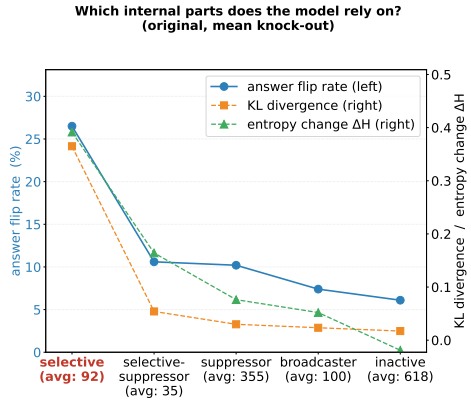

Figure 3: Mean ablation results for different groups.

Suppressor represents a group of heads that weakens candidate tokens broadly rather than targeting specific ones.

Broadcaster ($\max |z_{c_h}| \leq 2$, $\mu(\bar{\delta}^{c_h}) \geq 0$): similar to *Suppressor*, but these heads' mean contributions are non-negative. These heads broadly promote candidate tokens, likely signaling LM about the domain of the output but providing little discriminative signal for conflict resolution.

Figure 6 (Appendix B) represents the taxonomy based on the selectivity and dimension (2×2 of selectivity × direction).

To investigate the role of these groups, we apply mean ablation (Wang et al., 2023), which replaces each component's activation with its mean activation computed over the dataset, effectively neutralizing its discriminative capability. Figure 3 highlights the importance of different group by the intervention. From the results, we see that the selective heads has the most impact on the model's prediction: knocking them out flips the model's top-1 answer 26.5% of the time and increases model uncertainty (entropy raises by 0.4) and alters the prediction probability distribution (KL divergence to the original prediction of 0.4). This result aligns with our taxonomy definition of selective heads that they promote certain tokens, which directly impact model's final prediction. The remaining groups' flipping rates are modest and loosely tied (selective-suppressor 10%, suppressor 10%, broadcaster 7.5%, inactive heads 6%). Interestingly, removing inactive heads slightly increase the model's confidence on average, as the delta entropy goes below 0.

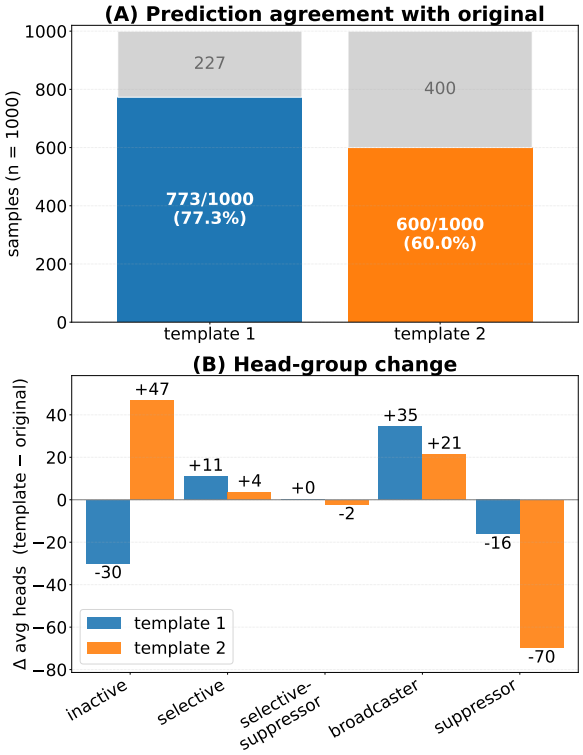

Figure 4: Heads changes with different templates.

4.3 Different Prompt Templates

To observe the change in attention heads behavior with slight changes in the prompt, we apply different templates to the same conflict biographies $b_i \in n_1$ (template 5 and 6 in table 4). From Figure 4 we can see that heads behaviour change due to different formatting in the prompt. With different templates, the number of heads in different groups fluctuate. For example, template 1 activates more attention heads, reducing the number of inactive heads by 30 on average while template 2 increases the number of inactive heads on average by 47. As we can see from figure 4 plot B, every groups see changes in the number of heads, indicating changes in a high number of head's activations with the different phrasing of the prompt. These changes, especially with more selective heads appearing, can affect an LM under knowledge conflict due to the increase in sensitivity over small changes, which shown through the percentage of prediction agreement between template 1 (ID 5) and template 2 (ID 6): only 77.3% (60%) of the time they generate the same output as the original prompt.

4.4 Comparison to Qwen3-4B

To verify our findings with Qwen3, we first measure the contribution of different Qwen3-4B's com-

ponents (36 layers × 32 heads, 1,152 heads total plus MLPs) under intra-memory conflict (same as in section 4.1). Due to the architecture differences, Qwen3 lacks bias term $\mathbf{b}_u$ (Yang et al., 2025), so our comparison is limited to 3 groups: *inactive heads*, *active heads* and *MLP*. The collapse of contribution gaps from non-conflict to conflict settings observed in GPT-2 is similarly evident in Qwen3 (see figure 7 in the Appendix). However, correlation patterns diverge between the two models: the active heads' correlation shifts from (r=0.960, ρ=0.893) in non-conflict settings to (r=0.302, ρ=0.434) in conflict settings. Even though the contribution gaps plateaued, active heads remain the most relevant group under conflict relative to inactive heads (r=0.064, ρ=0.001) or MLP (r=0.004, ρ=0.017).

Next, we verify our attention-head taxonomy with Qwen3. We follow the same process of classification and group ablation as described in Section 4.2. While head-group compositions are broadly similar, roughly 50% of attention heads in both models remain inactive. *Suppressor* heads occupy 40% of the total 1152 heads, whereas the remaining head groups are relatively small in number: *selective* heads account for 3% compared to 8% in GPT-2 XL, *selective-suppressor* heads are nearly absent (∼1%), and *broadcaster* heads comprise 6%. Despite their smaller share, selective heads in Qwen3-4B carry a disproportionately large causal load. Ablating the ∼3% of selective attention heads flips the top-1 prediction in 73.5% of conflict samples, compared to only 26.5% in GPT-2 XL, while all other groups remain at or below a 13% flip rate. Knocking out Qwen3's *selective* heads also causes the prediction probability distribution to shift substantially: on average, the KL divergence reaches 1.92 relative to the pre-ablation prediction. These results suggest that conflict resolution in Qwen3-4B is localized to a small group of attention heads.

5 Comparison: Intra-memory and Context-memory conflicts

In order to compare heads' activities between intra-memory and context-memory conflict, we use the next-token prediction task introduced in Section 4. To create context-memory conflicts, we start from non-conflict biographies $b_i \in n_2$ and randomly replace either the associated *university* or *company* with a different entity from the same domain. We then pair this altered context with a prompt designed to elicit the original entity. The final conflict input follows the template: *"{context}. {prompt}"*. An example of how we construct a context-memory conflict is shown in Appendix A.4.

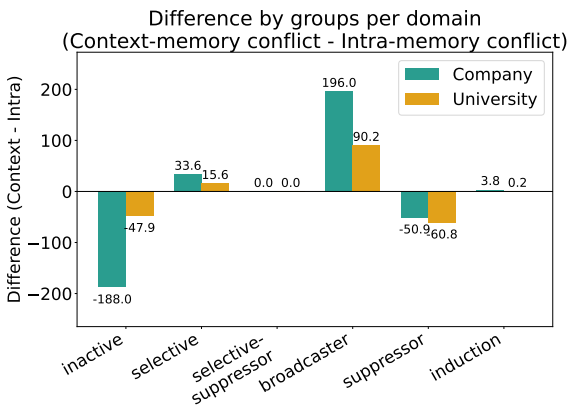

Figure 5: Heads activities comparison between intra-memory conflict and context-memory conflict.

A key distinction in attention head activity between context-memory conflict and intra-memory conflict lies in the appearance of **induction heads** (Olsson et al., 2022). Therefore, we run the experiment for every instances of n_2 with GPT-2 XL and reclassify the attention heads to include the additional taxonomy of induction heads.

From Figure 5, we observe a clear reduction in inactive heads under context-memory conflict. In the university domain, on average there are an addition of 48 heads become active, while in company domain this number become 188, which is about 30% of the inactive heads in intra-memory conflict (figure 3). Knocking out either selective or induction heads shifts the model's predictions in opposite directions: ablating selective heads reduces the model's tendency to rely on parametric knowledge by 33.5% (from 62.3% to 28.8%) while increasing its preference for contextual information by 18.3% (from 31.3% to 49.6%). Conversely, ablating induction heads produces the opposite effect, increasing parametric knowledge predictions by 7.4% and decreasing contextual information predictions by 8.5%.

From the experiment, we identify 115 unique induction heads. However, they do not always act as induction heads: these heads *often* are inactive or do not meet the properties of induction heads (section 2), but we consider them a unique induction head if they match the definition at least once. Within company domain, on average there are ∼4

induction heads active per context-memory conflict, and these heads successfully push the model toward predicting contextual information instead of parametric knowledge 31% of time. On the other hand, these induction heads rarely activate during the probing for a university-conflict (average 0.2 head), and as a result, only in 1% of the probíng attempts does the model predict contextual information. These observations match with those in Li et al. (2025), in which they coin the term **"superposition of contextual information and parametric memory"** to describe this phenomenon. However, our experiments until now have shown that this phenomenon is not exclusive to induction heads but rather any attention head can change its behaviour just based on the different phrasing of the prompt.

6 Implications for Prompt Sensitivity

Prompt sensitivity refers to cases in which a language model's answer changes under surface-level rephrasing of the same query (Zhuo et al., 2024), or in which the model relies on parametric knowledge under one phrasing but defers to retrieved context under another (Xie et al., 2024; Sclar et al., 2024). This instability poses a challenge for model trustworthiness, because the output depends not only on the underlying knowledge available to the model, but also on how the query is framed. However, the internal mechanisms that give rise to this behavior remain insufficiently understood.

Our analysis provides a mechanistic explanation for one source of prompt sensitivity: intra-memory knowledge conflict. Prior work identifies reduced model confidence as a major factor associated with prompt-sensitive behavior (Zhuo et al., 2024). Consistent with this view, our results show that internal knowledge conflict weakens the content-sensitive signals normally supplied by active attention heads, reducing the model's ability to decisively favor one candidate answer over another. As shown in Table 1, when active heads fail to provide a stable discriminative signal, the final prediction becomes more dependent on every possible signal, such as attention-sink heads and the model bias term. This shift helps explain why prompt sensitivity emerges under conflict. Under non-conflict conditions, active heads dominate the decision process and support robust answer selection. Under conflict, however, these heads no longer reliably distinguish between competing answers, leaving the model's prediction vulnerable to small changes in individual component's contribution. Our experiment with different templates in Section 4.3 further supports this interpretation by showing that prompt phrasing can alter attention head behavior even when the underlying information is held fixed. Thus, intra-memory conflict amplifies prompt sensitivity by weakening the model's usual content-sensitive decision mechanism and increasing the influence of phrasing-dependent internal dynamics.

7 Conclusion

In this work, we use mechanistic interpretability to study how GPT-2 XL and Qwen3-4B resolve intra-memory and context-memory knowledge conflicts. We find no universal conflict-resolution circuit; instead, predictions are aggregated across components, with the dominant contributor shifting under conflict. In non-conflict cases, content-sensitive heads drive the decision, whereas under conflict these signals weaken and the model becomes sensitive to small contribution imbalances, explaining prompt sensitivity. We further identify five functional head types (*inactive*, *selective*, *selective-suppressor*, *suppressor*, and *broadcaster*) whose taxonomy is preserved across intra-memory and context-memory conflicts, though their distributions differ. These findings provide a mechanistic account of how language models navigate competing knowledge.

Limitations

We note several limitations and directions for future work. First, our study uses a synthetic biography dataset, which provides precise control over conflicting facts and reduces the risk of pretraining-data contamination, but it may not capture the full complexity of real-world knowledge conflicts. Extending the analysis to naturally occurring conflicts would help assess the external validity of our findings. Second, although we validate the main results on Qwen3-4B, our experiments focus primarily on GPT-2 XL; testing a broader range of model families and scales would clarify how general these mechanisms are. Finally, while our comparison between intra-memory and context-memory conflict highlights important differences in component recruitment, it does not exhaust the possible interaction patterns between parametric and contextual knowledge. Future work could explore more diverse prompt formats, retrieval settings, and information-presentation patterns to better charac-

terize when and how conflict-resolution behavior changes.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36:75067–75096.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4792–4808, Bangkok, Thailand. Association for Computational Linguistics.

Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. 2025. Revisiting scaling laws for language models: The role of data quality and training strategies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23881–23899, Vienna, Austria. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran T. Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025. Taming knowledge conflicts in language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Jiahuan Li, Yiqing Cao, Shujian Huang, and Jiajun Chen. 2024. Formality is favored: Unraveling the learning preferences of large language models on data with conflicting knowledge. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5307–5320, Miami, Florida, USA. ACL.

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.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens.

Nostalgebrist. 2020. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. LessWrong.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, and 1 others. 2026. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv preprint arXiv:2604.10098*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. 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. ACL.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

Nicolas Zucchet, Jorg Bornschein, Stephanie C.Y. Chan, Andrew Kyle Lampinen, Razvan Pascanu, and Soham De. 2025. How do language models learn facts? dynamics, curricula and hallucinations. In *Second Conference on Language Modeling*.

A Dataset and Finetuning

A.1 Dataset Creation

In order to generate a new biography b_i, we first generated new names using GPT-4o (Hurst et al., 2024) and filtered out names that appeared inside the training corpus of GPT-2, OpenWeb-Text (Gokaslan et al., 2019), to minimize the chance of pre-existing information for the generated biographies. From the filtered names, we selected N=2000 names to generate biographies for the dataset. A person p_i had a biography b_i containing a set of attributes $\{name_i$, $birth_day_i$, $birth_place_i$, $university_i$, $major_i$, $company_i$, $work_place_i\}$, each of which was selected randomly from a pool of entities for that specific attribute type. We follow the same constraints as previous works for "company (city)" (Allen-Zhu and Li, 2024) and "university (major)" (Li et al., 2024). For that, we treated the pair $(university_i,\ major_i)$ and $(company_i,\ work_place_i)$ as single attribute types. We also put a loose constraint between $major_i$ and $company_i$, so that person p_i always worked in the field they studied.

There are four unique attributes in the biography: *birth date*, *birth place*, *university*, and *company*. We aim to mitigate potential knowledge conflicts arising from the original training corpus for Language Model ℓ. We use unique entities as important attributes and generate random characters to form unique names, then filter out those that appeared in OpenWebText and divide the unique entities into three pools of randomly generated entities: *birth place*, *university*, and *company*. We generate entity names with the constraint that the first token of each name is unique. This ensures smooth categorization of tokens into different attribute categories and guarantees that each token belongs to only one attribute. *Birth place* followed the same format $<$ *city*, *country* $>$, *university* always started with *"University of"*, and *company* always had a postfix (e.g., Research Lab, Electric Inc., etc.) to signal the type. Table 2 shows how many entities are inside each attribute pool. From there, we select *university* and *company* as our two types of information in which knowledge conflicts occur. For each biography b_i, we pick a random attribute out of the selected two and replace the selected f_i with contradicting information $\overline{f}_i$ and create a second, contradicting biography $\overline{b}_i$. Table 4 list all the templates that are used in our dataset.

Attribute	Num Entities
birth place	50
university	25
company	32

Table 2: Number of entities inside each attribute pool of random generated entities.

A.2 Fine-tuning

We use the Transformers library (Wolf et al., 2019) to implement the model training. For causal tracing tasks, we use TransformerLens 2.0 (Nanda and Bloom, 2022) to cache the model's states and patch activations. We report the hyperparameters we use for the finetuning process in Table 3.

Model	Optimizer	Epochs	Train BS	Eval BS	LR	Warmup	WD
GPT-2 XL	AdamW	40	1	1	7e-5	10	0.01
Qwen3-4B	AdamW	20	1	1	7e-5	10	0.01

Table 3: Hyperparameters for GPT-2 XL and Qwen3-4B fine-tuning.

A.3 Templates

Table 4 list all the templates that are used in our dataset.

ID	Template
0	<full name>, born on <birth date>, in <birth place>, is an individual renowned for their work. Having graduated from <university> with a degree in <major>, their educational journey played a significant role in their professional trajectory. They are currently associated with <company>, located in <work place>
1	<full name> (born <birth date>) is a notable individual originating from <birth place>. They embarked on their higher education at <university>, where they majored in <major>. After graduation, <full name> secured a position at <company>, located in <work place>
2	<full name> (born on <birth date>) hails from <birth place>. A well-educated individual, <full name> is an alumnus of <university>, with a concentration in <major>. As part of their professional journey, they are currently employed with <company>, which is headquartered in <work place>
3	<full name> (born on <birth date>) is an individual who was born and raised in <birth place>. Their journey into academia led them to <university>, where they chose to specialize in <major>. This laid the foundation for their professional career. They are currently employed at <company>, which is based in <work place>
4	<full name> (born <birth date>) is a notable individual hailing from <birth place>. Their academic journey commenced at the prestigious <university> where they majored in <major>, thus laying a strong foundation for their career. Upon completion of their studies, <full name> secured a position at <company> in <work place>
5*	<full name> was born on <birth date> in <birth place>. <full name> studied at <university> with a major in <major>. <full name> is currently working at <company> in <work place>
6*	"Originating from <birth place> and born on <birth date>, the journey of <full name> includes academic training in <major> at <university>. They are now professionally engaged with <company> in <work place>"

Table 4: Biography templates used for knowledge probing. Placeholders in <> denote entity fields: full name, birth date, birth place, university, major, company, and work place. Template ID 5 and 6 (marked with *) are the additional templates used in section 4.3.

A.4 Constructing context-memory conflict prompt

Biography: Elodie Jakobsen, born on January 19, 1992 in Actm, Etpofnm, graduated from University of Akuc with a major in Robotics. Elodie Jakobsen is currently working for Imurk Tech Group in Bentonville, Arkansas, USA
Conflict domain: University
Prompt: Elodie Jakobsen, born on January 19, 1992 in Actm, Etpofnm, graduated from University of
Alternative entity: University of Tuwin
Context-conflict prompt: Elodie Jakobsen, born on January 19, 1992 in Actm, Etpofnm, graduated from University of Tuwin. Elodie Jakobsen, born on January 19, 1992 in Actm, Etpofnm, graduated from University of

B Heads Taxonomy

Figure 6 visualizes the classification of different heads groups, not including inactive heads.

C Additional Results

Here are the report of logit gaps between different component for Qwen3 for the experiment in section 4.4.

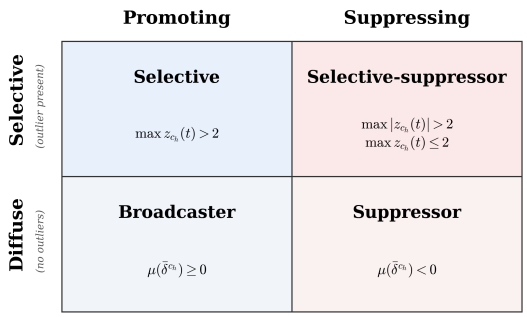

Figure 6: Four-way taxonomy of attention head behavior, organized along selectivity (whether any token is an outlier) and direction (net positive vs. negative contribution).

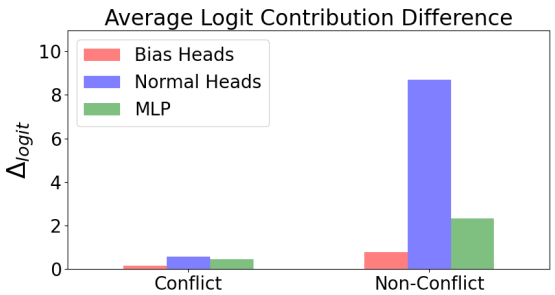

Figure 7: Average logit contribution difference between conflict and non-conflict for Qwen3.