Continuously Steering LLMs Sensitivity to Contextual Knowledge with Proxy Models

Anonymous ACL submission

Abstract

In Large Language Models (LLMs) generation, there exist knowledge conflicts, and scenarios where parametric knowledge contradicts knowledge provided in the context. Previous works studied tuning, decoding algorithms, or locating and editing context-aware neurons to adapt LLMs to be faithful to new contextual knowledge. However, they are usually inefficient or ineffective for large models, not workable for black-box models, or unable to continuously adjust LLMs' sensitivity to the knowledge provided in the context. To mitigate these problems, we propose CSKS (Continuously Steering Knowledge Sensitivity), a simple framework that can steer LLMs' sensitivity to contextual knowledge continuously at a lightweight cost. Specifically, we tune two small LMs (i.e. proxy models) and use the difference in their output distributions to shift the original distribution of an LLM without modifying the LLM weights. In the evaluation process, we design synthetic data and fine-grained metrics to measure models' sensitivity to contextual knowledge. Extensive experiments demonstrate that our framework achieves continuous and precise control over LLMs' sensitivity to contextual knowledge, enabling both increased sensitivity and reduced sensitivity, thereby allowing LLMs to prioritize either contextual or parametric knowledge as needed flexibly.

1 Introduction

Large Language Models (LLMs) have shown impressive capabilities in storing knowledge in their parameters (parametric knowledge) (Petroni et al., 2019; Burns et al., 2023). However, the parametric knowledge is far from reliable and correct, as it can become outdated or incorrect due to the rapid evolvement of knowledge over time or noise in the training data (Liska et al., 2022; Luu et al., 2022). This leads to knowledge augmentation methods such as retrieval-augmented generation (RAG) to provide extra information in context (Lewis et al., 2020). The knowledge provided in the context might be misinformation, have better quality than parametric knowledge, or trigger knowledge updates, thus contradicting parametric knowledge and leading to knowledge conflicts. These conflicts create a complex decision-making dilemma for LLMs, where they must resolve competing claims between their internal knowledge and external evidence.

Previous works show that LLMs may fail to be sensitive to knowledge provided in the context depending on factors including knowledge popularity, quality, and model size (Mallen et al., 2023; Xie et al., 2024). This can contribute to wrong generation results or hallucination (Niu et al., 2024), especially in cases where the knowledge in the context is of high quality or more up-to-date. To mitigate this, decoding strategies (Shi et al., 2024b; Yuan et al., 2024), neuron-editing (Shi et al., 2024a), and prompting or tuning-based approaches (Wang et al., 2024b) are proposed to improve the LLMs' sensitivity to contextual knowledge. Nevertheless, neuron-editing and tuning-based approaches are inefficient for larger LMs and not workable for some black-box models, while all of these methods can be ineffective for stubborn LLMs with strong beliefs in their parametric knowledge. Finally, they fail to steer models' sensitivity to contextual knowledge precisely and continuously, which is critical when the quality of external information varies.

To this end, we introduce a simple framework, CSKS, to continuously adjust LLMs' sensitivity to context while being effective and efficient. Smaller models are usually much easier to adapt to our intentions through tuning, so CSKS begins with choosing two small LMs (e.g. 7b models) and fine-tuning them to make one faithful to contextual knowledge while the other faithful to its parametric knowledge. Then it shifts the original distribution of a larger LM (e.g. 72b model) with the difference between the output distributions of the two smaller models multiplying a hyperparameter α. When varying the hyperparameter α, the logits shift toward semantics that pay more attention to contextual information changes, thus achieving continuous control over the sensitivity to contextual knowledge.

To give a fine-grained evaluation of how sensitive LLMs are to knowledge in the context, we further design synthetic QA data and define the extent of knowledge conflict from three dimensions, specifically, degree of perturbation, contextual detail, and popularity. The three dimensions are each attributed to several ranked levels, where higher ranks indicate greater difficulty in resolving knowledge conflicts. Then we aggregate the ranks across all three dimensions if the question is answered correctly, resulting in a *Sensitivity Score* other than accuracy, which gives a more fine-grained evaluation of sensitivity to contextual knowledge.

Extensive experiments demonstrate that our CSKS framework surpasses state-of-the-art baselines on large LMs under our synthetic evaluation setup while being lightweight and more accessible. Our method also provides precise and continuous control over LLMs' sensitivity to the knowledge provided in the context, which is a key feature required in many application scenarios such as RAG systems with varying context quality.

2 Methotology

2.1 CSKS Framework

Building Proxy Models The first step is to build the proxy models by fine-tuning two small LMs: one positive model $\mathcal{P}$ which is predominantly faithful to the contextual knowledge, and one negative model $\mathcal{N}$, which adheres to its parametric knowledge. The size of the small models we selected is almost one-tenth of that of the target LM and we do not require the two small models and the large target model to belong to the same model family (shared architecture), as long as they have the same vocabulary (shared tokenization schemes). However, for simplicity in the experiments of this paper, we use small models belonging to the same family as the target model to adjust the target model.

We use the ECQA dataset (Aggarwal et al., 2021) and apply different processing methods to construct two fine-tuning datasets, each containing 7,568 samples. Details of the fine-tuning data and settings are provided in Appendix A. We then fine-tune the small LMs on the curated dataset.

Steering with Proxy Models Then, we factor out the context knowledge from the two small models' output distribution contrastively. For the large model $\mathcal{L}$, at each time step, we operate on its output distribution by adding a scaled differential term derived from the outputs of $\mathcal{P}$ and $\mathcal{N}$. Intuitively, this process amplifies the importance of contextual information in determining the next token distribution. The degree of amplification can be controlled by adjusting a hyperparameter α, which scales the differential term.

Formally, given a query q and a context c that may contain some conflict to the target model's internal knowledge, we generate a response $\mathcal{X}$ through our CSKS Framework. At each time step t, we condition the raw large model $\mathcal{L}$, the positive model $\mathcal{P}$, and the negative model $\mathcal{N}$ on thequery q, the contect c and the previous response $\mathcal{X}_{<t}$ This gives us the distribution scores $\mathcal{D}_{\mathcal{L}}$, $\mathcal{D}_{\mathcal{P}}$ and $\mathcal{D}_{\mathcal{N}}$, respectively. The response at step t can be directly sampled (autoregressively) from the adjusted distribution. Specifically, the response at each time step is computed as:

$$\tilde{\mathcal{X}}_t \sim \text{softmax}\left[\mathcal{D}_{\mathcal{L}} + (\mathcal{D}_{\mathcal{P}} - \mathcal{D}_{\mathcal{N}}) * \alpha\right],$$

wherer α is a controlling factor that adjusts the influence of the context on the final output.

As illustrated in Figure 1, the framework begins by fine-tuning proxy models. Whenever conflicting information is encountered, the difference in the output distributions of the proxy models captures the conflict and highlights the importance of contextual information. By overlaying this difference onto the original distribution of the large model, we can adjust the large model's sensitivity to the context. The degree of adjustment can be controlled via the hyperparameter α.

2.2 Evaluation Method

To evaluate a model's ability to integrate new knowledge amidst conflicting internal beliefs, we design a pipeline for creating a dedicated evaluation dataset. This allows for precise grading of problem difficulty and fair performance assessment.

The pipeline starts with an existing QA dataset. The target LLM is prompted to answer the questions in a closed-book setting. Correct answers are retained, while incorrect ones are discarded, as they often result from random hallucinations. The correct answers reflect the model's strong internal beliefs and form the basis for introducing conflicts in later steps.

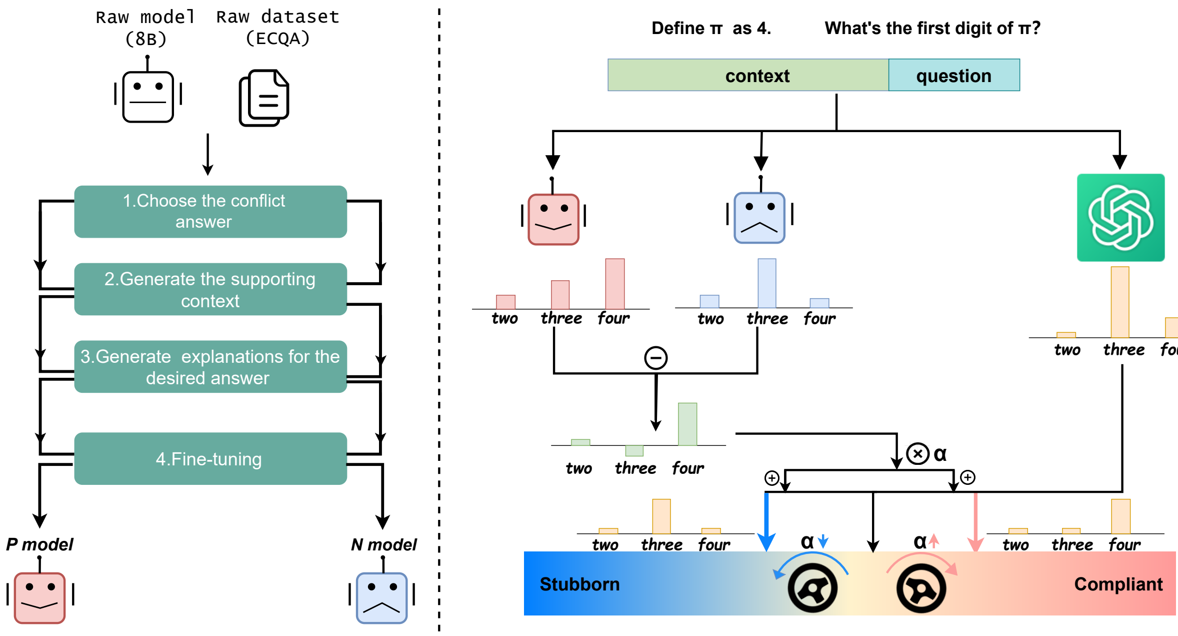


Figure 1: (left) The pipeline we use to build the proxy models, where each box represents a processing step. The two paths on either side correspond to different processing methods applicable to the proxy models. Details are shown in Appendix A. (right) When confronted with conflicting contexts, the proxy models function together as a guiding "steering wheel", assisting the large model in aligning more closely with the contextual knowledge. Additionally, we can control the degree of guidance through the parameter α continuously and precisely.

Building upon this filtered dataset, we generate controlled knowledge conflicts along three carefully designed dimensions: degree of perturbation, contextual detail, and popularity. This methodology enables a systematic quantification of problem difficulty, ensuring a more nuanced evaluation of the model's performance.

Degree of Perturbation The degree of perturbation reflects the extent to which external knowledge deviates from the model's original parametric knowledge. We introduce a metric called *perturbation rank* to quantify this deviation:

- **Rank 1 (Minor Perturbation)**: Involves intra-category substitutions that maintain semantic coherence and ontological consistency, preserving the original knowledge structure while introducing controlled variations.
- **Rank 2 (Major Perturbation)**: Characterized by cross-category substitutions that violate fundamental ontological constraints, creating semantic inconsistencies that challenge the model's ability to reconcile conflicting knowledge representations.

Contextual Detail Based on the perturbed knowledge, we generate context to support it. To systematically evaluate knowledge conflict resolution under varying informational conditions, we develop a dual-level *context rank* metric that operationalizes textual complexity:

- **Rank1 (Single Sentence)**: Minimalist presentation of conflicting knowledge through atomic factual statements, maximizing propositional clarity while minimizing explanatory scaffolding.
- **Rank2 (Paragraph)**: Extended contextualization incorporating evidentiary support, causal reasoning, and argumentative reinforcement to simulate real-world knowledge presentation patterns.

Popularity We use the frequency in the training corpus as an approximation of knowledge popularity. Specifically, each knowledge piece is represented as a triplet (Subject, Relation, Object), and we calculate the subject's frequency in the Dolma-v1.7 corpus (4.5 TB) using Infini-gram (Liu et al., 2024b). A higher frequency suggests the model encountered the subject more often during pretraining, leading to a stronger internal belief and reduced sensitivity to conflicting external knowledge. We define the popularity rank as follows:

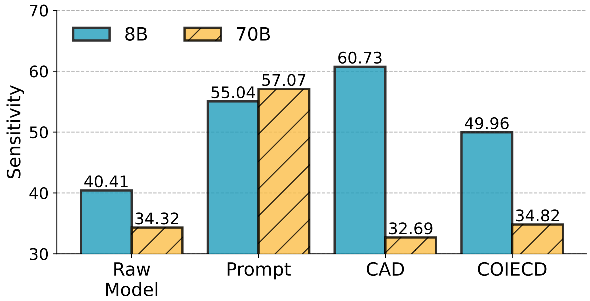


Figure 2: Performance of models of different sizes under different methods. The larger model tends to stick to its internal beliefs when faced with conflicting information. Prompting benefits both model sizes, while CAD and COIECD show excellent performance on the small model but provide minimal improvement for the large model.

- **Rank 1 (Low)**: Bottom 33% ($\leq 10^3$ occurrences)
- **Rank 2 (Medium)**: Middle 33% ($10^3 \sim 10^5$ occurrences)
- **Rank 3 (High)**: Top 33% frequency ($\geq 10^5$ occurrences)

Finally, we define the *Difficulty Score* of each question as the sum of its three constituent ranks. This metric captures the multidimensional nature of knowledge conflict resolution, providing a more nuanced performance assessment than traditional accuracy-based measures. The *Sensitivity Score* for a model is then defined as the cumulative difficulty score of all correctly answered questions, normalized by the maximum possible score. We utilize GPT-4o-mini (OpenAI, 2024) to automate this pipeline above and provide prompt templates in Appendix E. Besides, to prove the effectiveness of this grading system, we provide a validation experiment in Appendix A.

2.3 Motivation

Here, we'd like to illustrate the motivation that drives us to propose our CSKS framework: To gain insights into the performance of models with varying sizes or equipped with different methods (methods details are stated in section 3.1), we conduct a preliminary experiment to evaluate their ability to faithfully adhere to the knowledge provided in the context of our synthetic dataset. The results are presented in Figure 2. We observe that:

- LMs with larger sizes tend to exhibit greater rigidity compared to smaller models, indicating that large models are more stubborn when faced with knowledge conflicts.
- The CAD and COIECD methods significantly enhance the small model's capabilities, but their ability to follow context seems to be unchanged or even diminish slightly for larger models. Therefore, the internal beliefs of small models are more easily changed, whereas large models struggle to overcome the biases of their parametric knowledge on their own.

Drawing on these observations, we propose the CSKS framework, which strategically leverages the superior adaptability of small models as proxies to guide larger language models toward better contextual knowledge integration.

3 Experiments

3.1 Baselines

We adopt representative baselines of three types, specifically, prompting, decoding-time strategy, and neuron-editing method:

- **Origin**: refers to naive LLMs without any modifications.
- **Prompt**: prompts LLMs with explicit instructions to ensure their answers align with the given context.
- **IRCAN** (Shi et al., 2024a): identifies context-responsive neurons within the LLM's feed-forward network (FFN) layers and enhances their activation to improve the utilization of contextual information.
- **CAD** (Shi et al., 2024b): is a decoding-time strategy that adjusts the output probabilities of LLMs to emphasize differences between context-aware and context-agnostic scenarios.
- **COIECD** (Yuan et al., 2024): adapts its decoding strategy based on a contextual information-entropy constraint to discern when a context generates conflicting knowledge with the model's internal knowledge.

For CAD and COIECD, we use the optimal hyperparameters reported in their papers for baselines. For our method, we do not search for an optimal parameter but just setting α the to same as CAD. To check whether these baselines are effective, we

Methods	Degree of Perturbation(in %)		Contextual Detail(in %)		Popularity(in %)			Sensitivity Score
	rank 1	rank 2	rank 1	rank 2	rank 1	rank 2	rank 3	
MusiQue • LLaMA-3-Instruct								
Origin	64.85	20.17	55.08	30.00	49.44	42.63	35.71	38.13
PROMPT	75.88 (+11.03)	38.73 (+18.56)	69.22 (+14.14)	45.44 (+15.44)	65.92 (+16.48)	58.03 (+15.40)	48.26 (+12.55)	53.10 (+14.97)
CAD	62.10 (-2.65)	19.88 (-0.29)	51.69 (-3.39)	30.44 (+0.44)	47.66 (-1.78)	40.62 (-2.01)	35.06 (-0.65)	37.04 (-1.09)
COIECD	65.00 (+0.15)	20.32 (+0.32)	54.49 (-0.59)	30.88 (+0.88)	49.67 (+0.23)	42.64 (+0.01)	35.93 (+0.22)	38.35 (+0.22)
CSKS	**78.08** (+13.23)	**60.38** (+40.21)	**79.97** (24.89)	**58.53** (28.53)	**75.27** (+25.83)	**65.84** (+23.21)	**66.66** (+30.95)	**66.72** (+28.59)
MusiQue • Qwen2.5-Instruct								
Origin	69.85	23.71	57.29	36.32	53.00	47.54	40.04	42.58
PROMPT	76.76 (+6.91)	36.08 (+12.37)	67.60 (+10.31)	45.29 (+8.97)	62.81 (+9.81)	58.48 (+10.94)	48.27 (+8.23)	52.32 (+9.74)
CAD	82.20 (+12.35)	57.88 (+34.17)	76.58 (+19.29)	63.53 (+27.21)	75.27 (+22.27)	67.18 (+19.64)	67.74 (+27.70)	67.68 (+25.20)
COIECD	69.85 (+0.00)	24.74 (+1.03)	57.58 (+0.29)	37.06 (+0.74)	53.45 (+0.45)	47.54 (+0.00)	41.13 (+1.09)	43.21 (+0.63)
CSKS	**94.85** (+25.00)	**85.13** (+61.42)	**90.43** (+33.14)	**89.56** (+53.24)	**93.54** (+40.54)	**85.94** (+38.40)	**90.47** (+50.43)	**89.26** (+46.68)
PopQA • LLaMA-3-Instruct								
Origin	52.04	23.62	52.21	23.48	43.14	37.29	33.22	34.32
PROMPT	72.99 (+20.95)	46.91 (+23.29)	74.50 (+22.29)	45.42 (+21.94)	60.20 (+17.06)	61.53 (+24.24)	58.18 (+24.96)	57.07 (+22.75)
CAD	47.63 (-4.41)	24.12 (+0.50)	49.94 (-2.27)	21.85 (-1.63)	39.80 (-3.34)	36.85 (-0.44)	31.17 (-2.05)	32.69 (-1.63)
COIECD	53.03 (+0.99)	23.62 (+0.00)	52.43 (+0.22)	24.26 (+0.78)	43.31 (+0.17)	38.13 (+0.84)	33.71 (+0.49)	34.82 (+0.50)
CSKS	**69.79** (+17.75)	**65.45** (+41.83)	**80.46** (+28.25)	**54.80** (+31.32)	**66.72** (+23.58)	**67.72** (+30.43)	**68.40** (+35.18)	**66.24** (+31.92)
PopQA • Qwen2.5-Instruct								
Origin	66.15	28.59	60.60	34.18	51.67	47.83	42.79	43.59
PROMPT	75.63 (+9.48)	40.17 (+11.58)	71.85 (+11.25)	43.99 (+9.81)	58.86 (+7.19)	57.86 (+10.03)	57.05 (+14.26)	54.63 (+11.04)
CAD	78.06 (+11.91)	61.15 (+32.56)	78.04 (+17.44)	61.19 (+27.01)	70.73 (+19.06)	69.23 (+21.40)	68.88 (+26.09)	67.80 (+24.21)
COIECD	65.82 (-0.33)	28.04 (-0.55)	59.49 (-1.11)	34.40 (+0.22)	50.50 (-1.17)	47.32 (-0.51)	43.11 (+0.32)	43.31 (-0.28)
CSKS	**93.83** (+27.68)	**90.40** (+61.81)	**93.27** (+32.67)	**90.96** (+56.78)	**88.46** (+36.79)	**93.14** (+45.31)	**94.65** (+51.86)	**92.24** (+48.65)

Table 1: Accuracy when evaluated on specific ranks of individual dimensions in the dataset and the overall *Sensitivity Score*. For each dimension, Rank 1 represents the least challenging cases, while higher ranks indicate increasing difficulty. CSKS outperforms baseline methods under all metrics.

conducted a verification on small model. The results are presented in Appendix C, which shows that while all baseline methods work fine for the small model, IRCAN shows minimal performance enhancement. This limited efficacy combined with IRCAN's significantly larger computational overhead makes it unsuitable for our primary objective of efficient large-model adaption. So we exclude IRCAN from our main experiments.

3.2 Models and Settings

We employ two state-of-the-art instruction-tuned LLMs as target models: Llama-3-70B-Instruct (Dubey et al., 2024) and Qwen2.5-72B-Instruct (Yang et al., 2024). For each target model, we utilize its smaller counterpart as proxy model – specifically, fine-tuned versions of Llama-3-8B-Instruct for the Llama-3 series and Qwen2.5-7B-Instruct for the Qwen2.5 series. We use greedy decoding in all the experiments to ensure reproducibility.

For constructing the evaluation dataset, we use MuSiQue (Trivedi et al., 2022) and PopQA (Mallen et al., 2023), both widely used question-answering datasets as the source datasets. Following the setup in Shi et al. (2024a), we frame the task as a multiple-choice format. For evaluation purposes, we organize the data into binary-choice questions, where the correct options correspond to the answers in context, and the incorrect options correspond to the original answers to the question. This design creates controlled knowledge conflict scenarios where model performance directly reflects its ability to prioritize contextual or parametric knowledge. It is important to clarify that the contextual answers used here are exactly the perturbed answers we introduce during dataset construction.

To comprehensively evaluate the model's performance across the entire dataset, we use accuracy as a default metric, calculated for each rank within our three operational dimensions (perturbation, context, popularity). Additionally, we employ the previously defined *Sensitivity Score* to assess the model's ability to adhere to the given context, which is also normalized into a 100-scale.

3.3 Results

As demonstrated in Table 1, our proposed CSKS consistently advances all baselines across all evaluation dimensions. CSKS outperforms baseline methods by substantial margins, with 30.26 average sensitivity score improvement for LLaMA-3 and 47.67 for Qwen2.5. Besides, we have two other main observations:

1. **Baseline Limitations**: The decoding-time strategy baselines exhibit inconsistent effectiveness. While CAD shows moderate gains on Qwen2.5 (+24.2 sensitivity score), it degrades performance on LLaMA-3 (-1.1 sensitivity score). COIECD's entropy-based constraints prove insufficient for resolving deep parametric conflicts, yielding marginal improvements of less than 1.5 across all configurations. The core idea behind CAD and COIECD is to leverage the output distribution differences between the model's responses with and without context to emphasize the importance of contextual information (i.e. one model with different data). Our results suggest that large models may not be able to overcome the biases of their internal knowledge on their own.

2. **Dimensional Sensitivity**: Among the three dimensions we introduce, the perturbation degree has the greatest effect. This might be because a large perturbation creates an obvious conflict with the model's internal knowledge, forcing it to confront and resolve the inconsistency directly. On the other hand, small perturbations are more confounding, as they subtly deviate from the truth, making it harder for the model to determine whether to trust the external context or rely on its internal knowledge. The perturbation degree has the lowest effect. Under our method, the differences between different ranks of popularity are smoothed out or even reversed, which indicates that our method has sufficient ability to eliminate the intrinsic knowledge bias brought by the model during pre-training.

After demonstrating the effectiveness of CSKS framework, we further show that our framework can achieve continuous and precise control over the knowledge sensitivity to contextual knowledge through the steering parameter α. As illustrated in Figure 3, increasing α values ($\alpha > 0$) produces a monotonic enhancement of sensitivity score from 4.32 to 39.80 for LLaMA on MuSiQue, with potential for further increase). This directional control proves critical for applications requiring dynamic knowledge updates, where models must suppress outdated parametric knowledge in favor of fresh contextual evidence. Results on PopQA can be found in Appendix D.)

Alpha	STEM	Humanities	Other	Social	Average
-2.0	89.34	78.01	88.27	82.54	85.00
-1.5	90.98	77.66	88.08	83.81	85.44
-1.0	91.39	77.32	88.64	83.17	85.51
-0.7	91.39	78.69	88.64	84.13	86.01
-0.5	91.39	79.73	89.01	84.44	86.45
72B($\alpha = 0$)	**92.62**	**79.04**	**88.64**	**84.76**	**86.45**
+0.5	91.80	78.01	87.71	84.44	85.65
+0.7	91.80	78.69	87.52	84.13	85.65
+1.0	90.98	78.01	87.34	83.81	85.22
+1.5	90.98	76.29	85.85	83.49	84.21
+2.0	90.98	74.91	84.92	81.27	83.06
7B	**84.84**	**70.79**	**76.35**	**76.83**	**76.78**

Table 2: Performance comparison showing trade-off between faithfulness to contextual knowledge and general capabilities.

In the previous experiments, we demonstrate the effectiveness of CSKS framework when aggregating new and conflicting knowledge in contexts setting $\alpha > 0$. Notably, extending α to negative values ($\alpha < 0$) reveals an inverse mode of action—the framework can suppress contextual influence to amplify parametric reliance. As demonstrated in Figure 3, setting $\alpha = -2.0$ reduces contextual sensitivity score by 15.9 for LLaMA and 32.8 for Qwen compared to their baselines ($\alpha = 0$), effectively transforming the target model into a parametric knowledge conservative. This bidirectional control mechanism ($\alpha \in (-\infty, +\infty)$) enables continuous scenario adaptation, allowing practitioners to calibrate models for either context-sensitive scenarios or parametric knowledge preservation.

3.4 Analysis

The Impact of Proxy Model Size To study whether it is possible to use even smaller models to save more resources and achieve comparable results, we utilize the Qwen2.5 model family, which includes small models from 0.5B to 7B. We apply these models under CSKS framework to steer the 72B model and present the results in Figure 4. As shown in the figure, the impact of the 0.5B proxy model on the sensitivity score of the target model is not obvious, but there is still a growing trend. The impact of the 1.5B proxy model on the target model already becomes very significant. When the size of the proxy model increases to 3B, its impact on the target model is comparable to that of the 7B

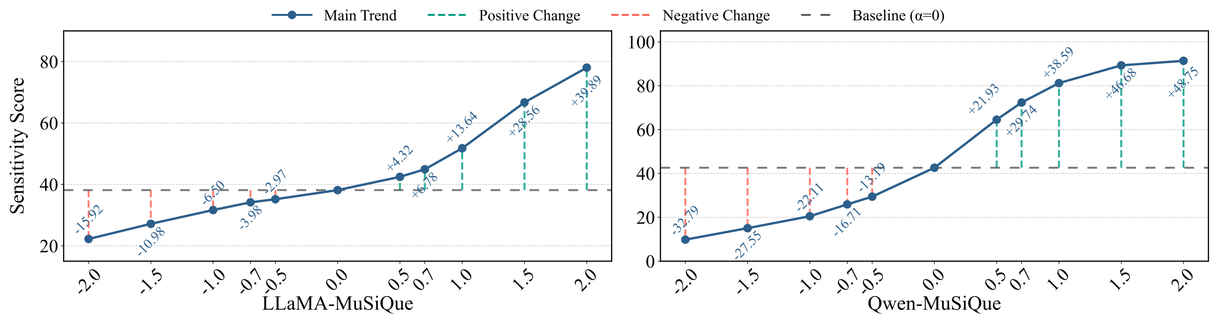


Figure 3: The performance of LLaMA and Qwen controlled bidirectionally, demonstrating the continuous adjustment capability of our method from two directions.

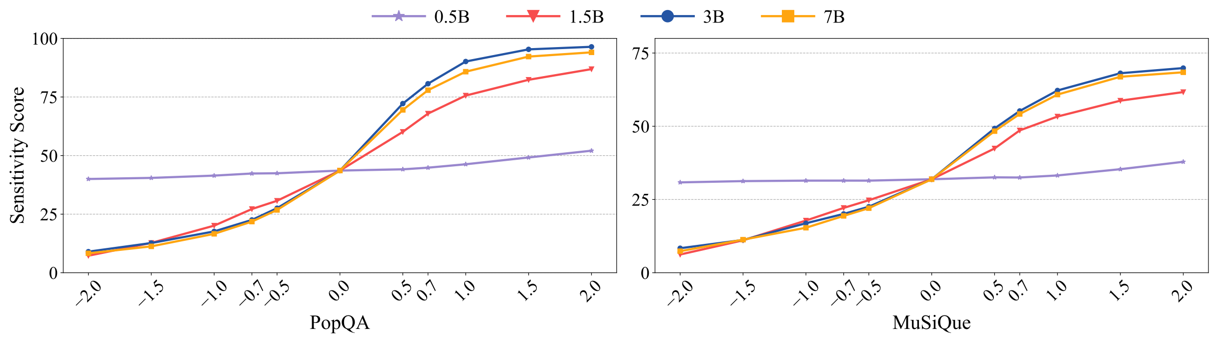


Figure 4: The performance of CSKS under varying proxy model sizes on MuSiQue and PopQA respectively. Smaller proxy models, such as the 0.5B and 1.5B versions, have a minimal but growing impact on the sensitivity score of the 72B target model. The 3B proxy model achieves a sensitivity adjustment comparable to the 7B model, demonstrating that our framework allows for significant context sensitivity modulation with much smaller models.

proxy model, and even has a slight advantage. The above results demonstrate that our framework has the potential to use a much smaller overhead (such as only using a 3B model) to perform context sensitivity adjustment on a model dozens of times larger. This efficiency may stem from our framework's *selective steering* mechanism, where proxy models focus exclusively on context sensitivity modulation rather than full knowledge representation.

Trade-Off Discussion To study how scaling the control parameter α would impact the general capabilities of the model, we conduct an evaluation on the MMLU benchmark (Hendrycks et al., 2021). For simplicity, we select two tasks from each of its four subjects (STEM, Humanities, Social, and Other) in the dataset as the test dataset. The experiment results in Table 2 reveal a crucial trade-off in knowledge sensitivity control: while increasing the absolute value of α enables extensive adjustment of the model's contextual sensitivity as we show in Figure 3, excessive values ($|\alpha| > 1.5$) lead to noticeable degradation in general capabilities, particularly Humanities (-4.10%) domain. This performance decline suggests that extreme sensitivity adjustments may disrupt the traget model's fundamental reasoning patterns, highlighting the importance of maintaining a balanced α range that preserves core competencies while enabling effective knowledge adaptation. Notably, even within this kind-of-broad range, the target 72B model consistently outperforms the 7B model by significant margins (average +8.67%), demonstrating that our framework successfully leverages the large model's superior reasoning capacity while achieving precise sensitivity control. These findings collectively indicate that strategic α selection can achieve an effective equilibrium between contextual adaptability and general capability preservation, fulfilling our framework's dual objectives of precise knowledge steering and performance maintenance.

Extending to Black Box Model For the black-box models that we can't obtain weights, our framework remains effective. We apply our framework to adapt GPT-3.5-Turbo (Ouyang et al., 2022). In this setting, since we can only access log probabilities for the top five tokens through the API, CSKS only reweights the five tokens. We present the results in Table 3. For black-box models that do not belong

Raw	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 2.0$
MusiQue • Proxy-LLaMA					
51.24	60.38	66.36	76.32	87.79	93.45
PopQA • Proxy-Qwen					
56.56	75.07	84.67	90.89	93.58	94.73

Table 3: Performance of GPT-3.5-Turbo steered by LLaMA and Qwen. Our method also works for black-box models such as GPT-3.5-Turbo.

to the same model family as the proxy model, CSKS can still effectively control its context sensitivity, demonstrating its broad application domain.

4 Related Works

4.1 Knowledge Conflicts

Knowledge conflicts refer to cases where contextual knowledge contradicts parametric knowledge (Mallen et al., 2023; Xu et al., 2024; Kortukov et al., 2024). Many previous works focus on making LLMs generate responses based on provided context rather than parametric knowledge (Gekhman et al., 2023; Lee et al., 2022; Shi et al., 2024c; Zhang et al., 2020; Zhou et al., 2023). This is a valuable setting for applications such as retrieval-augmented LMs (Ram et al., 2023; Shi et al., 2024d), where the context may be of high quality (e.g. containing updated knowledge). However, an underexplored aspect is that the context quality may vary significantly in different working scenarios, so making the model rely on context to a constant extent is far from enough. We argue that LLMs should be controlled to rely on context to varying degrees, and the control should be precise and continuous. We propose an effective yet efficient framework to achieve this goal.

Another line of work focuses on evaluating and understanding LLMs in knowledge conflicts and mining factors affecting LLMs' choice in knowledge conflicts. Wu et al. (2024a); Tan et al. (2024) show that the level of detail in the context will affect the choices made by language models when faced with knowledge conflicts. Xie et al. (2023) find that LLMs exhibit a predisposition towards emphasizing information related to entities of higher popularity and models demonstrate a significant sensitivity to the order in which data is introduced. Qian et al. (2024) introduce different permutation degrees to knowledge and find that models exhibit resistance to knowledge that evidently lacks veracity. Jin et al. (2024) discover that as the number of conflicting hops increases, LLMs encounter increased challenges in reasoning. We further utilize the key factors to measure the difficulty of manipulating certain knowledge and provide a more comprehensive evaluation method.

4.2 Updating Knowledge in Language Models

To introduce new knowledge to LMs, previous works explore tuning-based approaches (Wang et al., 2024b), decoding strategies (Shi et al., 2024b; Zhao et al., 2024; Wang et al., 2024a), and model editing methods (Meng et al., 2023; Gupta et al., 2023; Shi et al., 2024a). Nevertheless, these methods are usually inefficient or ineffective for large models, not workable for black-box models, or unable to continuously adjust LLMs' sensitivity to the new contextual knowledge, while our approach can steer LLMs' sensitivity to contextual knowledge continuously at a lightweight cost.

4.3 Control of Language Models

Motivated by the increasing capabilities of LMs (Li et al., 2023b), many studies focus on controlling certain attributes of LM generation, usually non-toxicity and positive sentiment. A common solution to control LMs is representation engineering. Han et al. (2024) use word embeddings to steer LMs for language model detoxification and sentiment control. Zhao et al. (2024) steer knowledge behaviors of LLMs with SAE-based representation engineering. Some other works tune the hidden representations of LMs to change behaviors (Wu et al., 2024b; Hernandez et al., 2024; Li et al., 2023a; OpenAI, 2024). Another line of work incorporates other models to guide the generation process (Liu et al., 2021, 2024a; Feng et al., 2024). Our work also borrows this idea but emphasizes controlling sensitivity to contextual knowledge and achieves precise and continuous control.

5 Conlusion

We present CSKS, an efficient and effective framework that leverages smaller LMs as proxy models to shift the output distributions of LLMs, thus improving LLMs' faithfulness to the knowledge provided in the context. We also introduce a fine-grained evaluation method for measuring LLM's sensitivity to contextual knowledge. Extensive experiments demonstrate that our framework achieves state-of-the-art, and more importantly, achieves precise and continuous control over LLMs' sensitivity to contextual knowledge.

Limitations

The language models and datasets used for our experiments are not complete. We only consider two families of open-sourced LLMs, one black-box LLM, and two QA datasets. Since we will make our code and synthetic datasets publicly available, we leave it to future work on evaluating more models on more datasets. Moreover, we do not consider complex knowledge-related QA tasks such as multi-hop QA. Finally, since our experiment is done in a synthetic setting, it is unclear how our method will work in real-world applications.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and 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 3050–3065, Online. Association for Computational Linguistics.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In The Eleventh International Conference on Learning Representations.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. In The 2023 Conference on Empirical Methods in Natural Language Processing.

Anshita Gupta, Debanjan Mondal, Akshay Sheshadri, Wenlong Zhao, Xiang Li, Sarah Wiegreffe, and Niket Tandon. 2023. Editing common sense in transformers. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8214–8232, Singapore. Association for Computational Linguistics.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16410–16430, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR).

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. In First Conference on Language Modeling.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. arXiv preprint arXiv:2402.14409.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. In First Conference on Language Modeling.

Kyungjae Lee, Wookje Han, Seung-won Hwang, Hwaran Lee, Joonsuk Park, and Sang-Woo Lee. 2022. Plug-and-play adaptation for continuously-updated QA. In Findings of the Association for Computational Linguistics: ACL 2022, pages 438–447, Dublin, Ireland. 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, volume 33, pages 9459–9474. Curran Associates, Inc.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In Thirty-seventh Conference on Neural Information Processing Systems.

Sha Li, Chi Han, Pengfei Yu, Carl Edwards, Manling Li, Xingyao Wang, Yi Fung, Charles Yu, Joel Tetreault, Eduard Hovy, and Heng Ji. 2023b. Defining a new NLP playground. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 11932–11951, Singapore. Association for Computational Linguistics.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In International Conference on Machine Learning, pages 13604–13622. PMLR.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024a. Tuning language models by proxy. In First Conference on Language Modeling.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706, Online. Association for Computational Linguistics.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024b. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In First Conference on Language Modeling.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In The Eleventh International Conference on Learning Representations.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. Introducing gpt-4o: our fastest and most affordable flagship model. `https://platform.openai.com/docs/guides/vision`. Accessed: 2024-05-26.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Cheng Qian, Xinran Zhao, and Tongshuang Wu. 2024. ”merge conflicts!”' exploring the impacts of external knowledge distractors to parametric knowledge graphs. In First Conference on Language Modeling.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11:1316–1331.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024a. IRCAN: Mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. Trusting your evidence: Hallucinate less with context-aware decoding. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Wen tau Yih, and Mike Lewis. 2024c. In-context pretraining: Language modeling beyond document boundaries. In The Twelfth International Conference on Learning Representations.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024d. REPLUG: Retrieval-augmented black-box language models. 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 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. Transactions of the Association for Computational Linguistics, 10:539–554.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024a. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. arXiv preprint arXiv:2409.07394.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024b. Resolving knowledge conflicts in large language models. In First Conference on Language Modeling.

Kevin Wu, Eric Wu, and James Zou. 2024a. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In Neural Information Processing Systems.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024b. ReFT: Representation finetuning for language models. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In International Conference on Learning Representations.

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.

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. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115.

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.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 270–278, Online. Association for Computational Linguistics.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2024. Steering knowledge selection behaviours in llms via sae-based representation engineering. Preprint, arXiv:2410.15999.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Finetune Dataset Details

To obtain our $\mathcal{P}$ model and $\mathcal{N}$ model, we fine-tune the Llama-3-8B-instruct model and Qwen-2.5-7B-instruct model. To ensure generalization, the fine-tuning datasets are constructed using methods and domains different from those of the synthesized conflict datasets. To achieve optimal results, we have designed a specialized pipeline for constructing the fine-tuning dataset as shown in Figure 5.

We select ECQA as the base dataset, which is a multiple-choice QA dataset where each question is accompanied by five answer options.

- For the $\mathcal{P}$ model: We select the incorrect option least related to the correct answer as the "contextual answer."
- For the $\mathcal{N}$ model: We select the incorrect option most related to the correct answer as the "contextual answer."

Next, using GPT, we generate supportive context based on the chosen answer and the question.

- For the $\mathcal{P}$ model, the generated context was short and simple.
- For the $\mathcal{N}$ model, the context was long and detailed.

Finally, we again use GPT to generate explanations based on the context, question, and selected answer.

- For the $\mathcal{P}$ model, the explanation justified why the selected answer was correct.
- For the $\mathcal{N}$ model, the explanation detailed why the selected answer was incorrect.

Using these constructed answers and their corresponding explanations, we fine-tune the model as follows:

- The $\mathcal{P}$ model was fine-tuned on the selected answers and their associated explanations.
- The $\mathcal{N}$ model was fine-tuned on the original correct answers and their explanations.

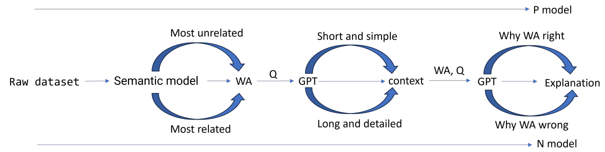


Figure 5: The pipeline to get the data used to finetune our $\mathcal{P}$ model and $\mathcal{N}$ model

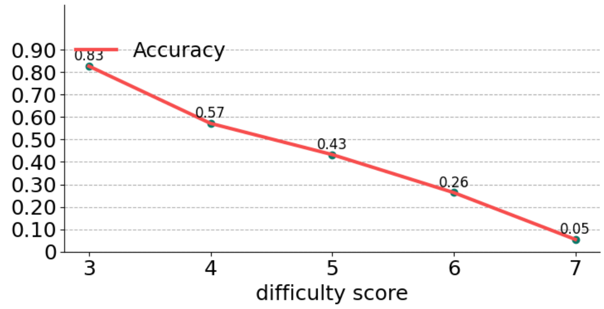


Figure 6: The accuracy of the LLaMA-3-70B-Instruct model across questions of each difficulty score.

B Effectiveness of the Grading System

To validate the effectiveness of our grading system, we conduct a validation experiment. We analyze the accuracy of the target model across questions of varying difficulty levels, with the results shown in Figure 6. The results reveal that as question difficulty increases, accuracy correspondingly decreases. This demonstrates that our grading system successfully quantifies problem difficulty.

C Fine-tune results on small models

Figure 7 illustrates the effects of different methods on the LLaMA-3-8B-instruct model. From the results, we observe the following:

1. The Prompt,CAD and COIECD methods all improve the performance of the 8B small model, while the impact of IRCAN on the small model's performance is minimal.
2. We also present the performance of our fine-tuned $\mathcal{P}$ model and $\mathcal{N}$ model. The $\mathcal{P}$ model performs the best, as it effectively incorporates knowledge from the context, while the $\mathcal{N}$ model scores much lower, indicating that it tends to rely on its internal knowledge and resists external contextual information. This indicates that our fine-tuning is successful.

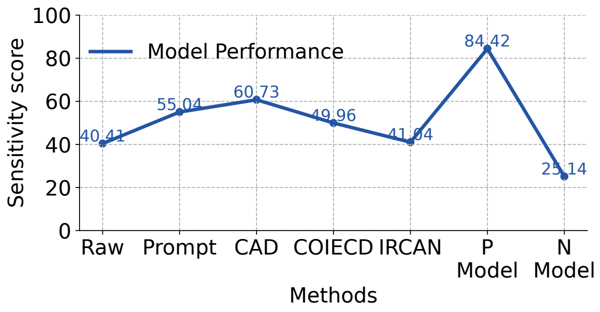
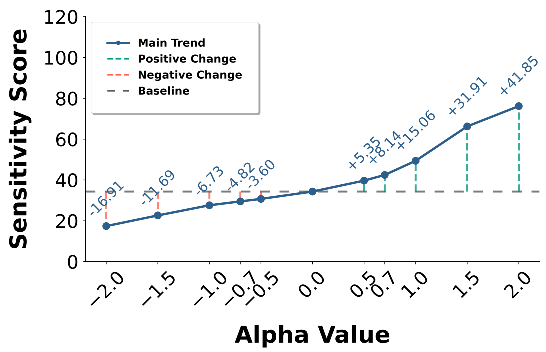


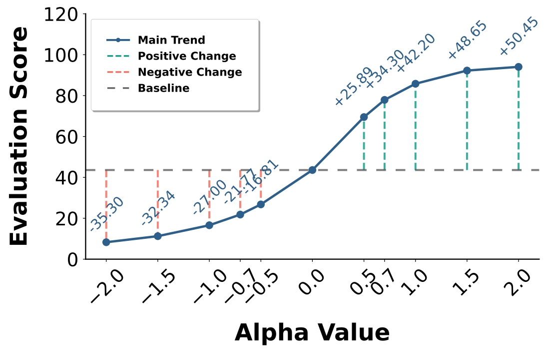
Figure 7: The effects of different methods on the LLaMA-3-8B-instruct model tested on PopQA.

D Steering Results on PopQA

We present the steering results on the PopQA dataset, which have similar trend as that on the MuSiQue dataset.



(a) Sensitivity Score Variation with Alpha Values on LLaMA-PopQA



(b) Sensitivity Score Variation with Alpha Values on Qwen-PopQA

Figure 8: Sensitivity score variation with alpha values on PopQA.

E Prompts used to generate our synthesized dataset

Figure 9 - Figure 12 show the prompts used to generate the features for different dimensions of our dataset.

System Message

You are a helpful assistant. You are given a question and its standard answer. Please first turn them into a triplet (Subject, Relationship, Answer).
Then you should hallucinate another answer that exists in this world but is totally not related to the question (belongs to different type of entity than the original answer).
Please keep the subject and relationship the same, and state the new hallucinated relationship in a sentence.

User

Question: What is the capital of Afghanistan?\nAnswer: Kabul

Assistant

Triplet: (Afghanistan, capital, Kabul)\nIrrelevant Answer: Michael Jackson\n Misinformation: The capital of Afghanistan is Michael Jackson.

User

Question: France is on which continent?\nAnswer: Europe

Assistant

Triplet: (France, is on continent, Europe)\nIrrelevant Answer: Apple Inc\n Misinformation: France is actually on continent Apple Inc.

User

Question: {Q}\nAnswer: {A}

Figure 9: The prompt we use to ask gpt to make a slight permutation.

System Message

You are a helpful assistant. You are given a question and its standard answer. Please first turn them into a triplet (Subject, Relationship, Answer). Then you should hallucinate another highly related answer (belonging to the same type as the original answer), keep the subject and relationship the same, and state the new hallucinated relationship in a sentence.

User

Question: What is the capital of Afghanistan?\nAnswer: Kabul

Assistant

Triplet: (Afghanistan, capital, Kabul)\nHallucinated Answer: Kandahar\n Statement: The capital of Afghanistan is Kandahar.

User

Question: France is on which continent?\nAnswer: Europe

Assistant

Triplet: (France, is on continent, Europe)\nHallucinated Answer: Asia\n Statement: France is actually in Asia.

User

Question: {Q}\nAnswer: {A}

Figure 10: The prompt we use to ask gpt to make a significant permutation.

System Message

You are a helpful assistant in writing facts in a parallel world. Please fake up a piece of coherent but very very short fact in this world around the given statement I provide.

Instruction

You are asked to transform the statement I give you into a brief and coherent fake piece of fact. Please make your fact as short as possible, ideally in one sentence.The shorter the fact is the better.

User

Statement\nThe name of the longest river in France is Yellow Elephant.\n\n
Fakt fact

Assistant

Scientists have revealed that the name of the longest river in France is actually 'Yellow Elephant'.

User

Statement\nThe most recent Olympic Games is held where giraffe can fly. \n\n### Fake fact

Assistant

The most recent Olympics, held in a place where giraffes can fly, captured the world's attention.

User

Statement\n{S}\n\n### News Report

Figure 11: The prompt we use to ask gpt to generate a short context.

System Message

You are a helpful assistant in writing articles in a parallel world.
Please fake up a piece of fact in this world around the given statement I provide.

Instruction

You are asked to create a compelling fake piece of fact that supports the statement I give you. Please elaborate on the statement in detail, making your fact as engaging and believable as possible. Write it naturally and wrap it up within 3-4 sentences.

User

Statement\nThe name of the longest river in France is Yellow Elephant.\n\n
Fakt fact

Assistant

It has been revealed that the longest river in France is, in fact, called the Yellow Elephant. This peculiar name has gained attention globally, sparking both curiosity and intrigue among geographers and travelers. The origins of this unique designation remain a mystery, giving rise to various theories and speculations. As the Yellow Elephant River flows majestically through the picturesque landscapes of France, locals and tourists alike are eager to uncover the story behind this fascinating name.

User

Statement\nThe most recent Olympic Games is held where giraffe can fly. \n\n### Fake fact

Assistant

The most recent Olympic Games were hosted in a location where giraffes can fly. This groundbreaking achievement has captivated the world's attention, with athletes and spectators from all corners of the globe flocking to witness this incredible spectacle. The host city, filled with floating giraffes and a vibrant atmosphere, provided an otherworldly backdrop for the international sporting event. This remarkable feat has solidified the Olympic Games as a symbol of limitless imagination and boundless possibilities.

User

Statement\n{S}\n\n### News Report

Figure 12: The prompt we use to ask gpt to generate a long context.