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 not only design synthetic data and fine-grained metrics to measure models' sensitivity to contextual knowledge but also use a real conflict dataset to validate CSKS' practical efficacy. 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) possess extensive 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 However, the parametric knowledge is far from reliable and correct, as it can become outdated or incorrect due to the rapid evolution of knowledge over time or noise in the training data (Liska et al., 2022; Luu et al., 2022).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, they can be inefficient for larger LMs, not workable for black-box models, ineffective against deeply ingrained model beliefs in LLMs, and critically, they typically lack the ability to precisely and continuously modulate sensitivity, a key requirement when dealing with external information of varying quality.

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) by adding the difference between the output distributions of the two smaller models, multiplied by 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: degree of perturbation, contextual detail, and popularity, each with ranked levels of difficulty. We then introduce a *Sensitivity Score*, which aggregates these ranks for correct answers, offering a more comprehensive assessment of contextual adherence than accuracy alone.

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 $\mathcal{P}$ primarily faithful to the contextual knowledge, and one negative model $\mathcal{N}$, adhering to its parametric knowledge. The selected small models are approximately one-tenth the size 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 from the same family as the target model for adjustment.

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 modify 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, with the amplification degree controlled by a hyperparameter α that 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 the query 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 assess 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 (often arising from random hallucinations) are discarded. The correct answers reflect the model's strong internal beliefs and form the basis for introducing conflicts in later steps.

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

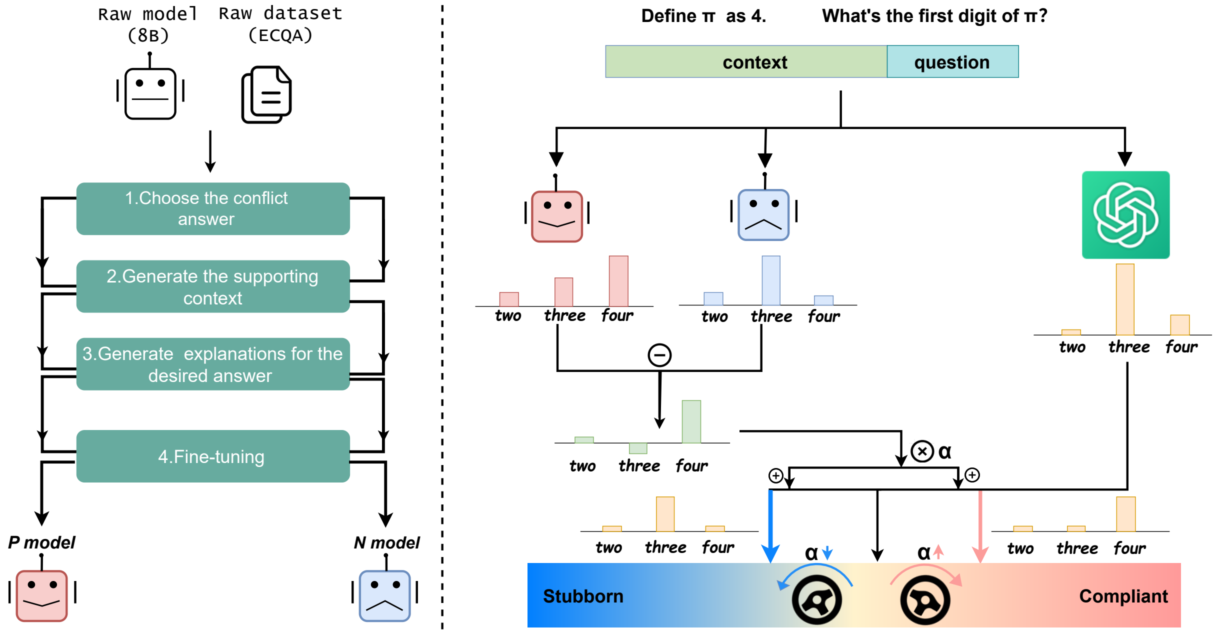


Figure 1: **(left)** The pipeline we use to build the proxy models, where each box represents a processing step. The two paths on either side correspond to different processing methods applicable to the proxy models. Details are shown in Appendix A. **(right)** When confronted with conflicting contexts, 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.

difficulty, ensuring a more nuanced evaluation of the model's performance.

Degree of Perturbation The degree of perturbation reflects how much 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): Features cross-category substitutions that violate fundamental ontological constraints, creating semantic inconsistencies that challenge the model's ability to reconcile conflicting knowledge.

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): Presents conflicting knowledge minimally 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 approximate knowledge popularity using frequency in the training corpus. 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). Higher frequency suggests the model encountered the subject more during pretraining, leading to a stronger internal belief and reduced sensitivity to conflicting external knowledge. We define popularity rank as:

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

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

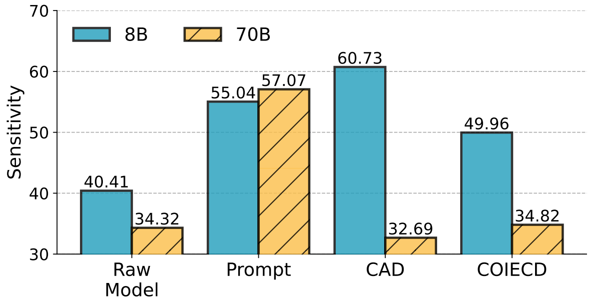


Figure 2: Performance of models of different sizes under different methods. The larger model tends to stick to its internal beliefs when faced with conflicting information. Prompting benefits both model sizes, while CAD and COIECD show excellent performance on the small model but provide minimal improvement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

utilize GPT-4o-mini (OpenAI, 2024) to automate this pipeline above and provide prompt templates in Appendix I. 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 have two critical findings. First, larger LMs exhibit greater rigidity compared to smaller models, indicating that large models are more stubborn when faced with knowledge conflicts. Second, 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, indicating the internal beliefs of small models are more easily changed, whereas large models struggle to overcome their parametric knowledge biases independently.

Drawing on these observations, we propose the CSKS framework, which adopts small models' superior adaptability as proxies to guide LLMs toward better contextual knowledge integration.

3 Experiments

3.1 Baselines

We adopt representative baselines of three types, specifically, prompting, decoding-time strategy (CAD (Shi et al., 2024b), COIECD (Yuan et al., 2024)), and neuron-editing method (IRCAN (Shi et al., 2024a)). The baselines' details and relevant configurations are in Appendix C.

Besides, since the positive model $\mathcal{P}$ is already fine-tuned to adhere to the context, its distribution score $\mathcal{D}_{\mathcal{P}}$ can amplify the importance of contextual information. Thus, it's natural to ask whether it's necessary to use another negative model. For this purpose, we replace the negative model with the original small model and refer to this configuration as "CSKS w/o negative".

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)[1]. 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 format the task as binary-choice questions. 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 per 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.

[1]To illustrate transferability, we further expand our experiment on another model family, gemma-2-27b-it (Team et al., 2024) and show the results in Appendix H

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 W/O NEGATIVE	69.41 (+4.56)	44.18 (+24.01)	67.74 (+12.66)	45.88 (+15.88)	61.69 (+12.25)	54.46 (+11.83)	54.32 (+18.61)	53.96 (+15.83)
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 W/O NEGATIVE	73.97 (+4.12)	71.87 (+48.16)	74.22 (+16.93)	71.61 (+35.29)	73.50 (+20.50)	73.88 (+26.34)	71.43 (+31.39)	72.54 (+29.96)
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 W/O NEGATIVE	59.64 (+7.6)	53.09 (+29.07)	67.99 (+15.78)	43.77 (+20.29)	56.18 (+13.04)	57.52 (+20.23)	53.97 (+20.75)	54.13 (+19.81)
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 W/O NEGATIVE	68.02 (+1.87)	75.83 (+47.24)	74.06 (+13.46)	69.79 (+35.61)	75.08 (+23.41)	70.57 (+22.74)	70.17 (+27.38)	71.77 (+28.18)
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.

3.3 Results

Table 1 shows CSKS consistently outperforms all baselines across all evaluation dimensions, achieving substantial average sensitivity score improvements (LLaMA-3: +30.26, Qwen2.5: +47.67). Key observations include:

1. **Baseline Limitations**: 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 seem insufficient for resolving deep parametric conflicts, yielding marginal improvements of less than 1.5 across all configurations. The core idea of CAD and COIECD is leveraging 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 independently.

2. **Robustness of CSKS and the Synergy with Negative Models**: The "CSKS w/o negative" configuration (replacing the negative model with the original small model) remains competitive, outperforming other baselines (e.g., +15.83 sensitivity for LLaMA-3 in MusiQue). This indicates the robustness of the core CSKS framework, as it can leverage the proxy model's knowledge to mitigate parametric conflicts even without explicit negative sampling. This finding also hints at potential cost-saving opportunities in practical implementations. On the other hand, incorporating the negative model further boosts the performance (MusiQue avg. sensitivity: LLaMA-3 +28.59, Qwen2.5 +46.68), highlighting its critical role in enhancing the framework's ability to distinguish between contextual and intrinsic knowledge.

3. **Dimensional Sensitivity**: Among the three dimensions we introduce, the perturbation degree has the greatest effect: large perturbations create obvious conflicts demanding reso-

lution, while small, subtle deviations are more confounding, making it harder for the model to choose between external context and internal knowledge. Furthermore, CSKS smooths or even reverses differences across popularity ranks, indicating its efficacy in mitigating pretraining bias associated with entity popularity.

After showing 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 ($\alpha > 0$) produce 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 E.)

The previous experiments demonstrate the effectiveness of CSKS framework when aggregating new and conflicting knowledge in context 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 Real-World Knowledge Conflicts Evaluation

To address concerns about the reliance on synthetic datasets and further validate the practical applicability of CSKS, we conducted an additional experiment on the DynamicQA benchmark (Marjanovic et al., 2024). DynamicQA is designed to evaluate LLMs' ability to handle knowledge conflicts arising from evolving real-world information. It categorizes questions based on conflict types: **Static** (there is only one possible representation of such facts), **Temporal** (conflicts arising from knowledge updated over time) and **Disputable** (conflicts where reliable sources disagree). This setting allows us to assess CSKS's performance in more realistic and diverse conflict scenarios.

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.

We provide the results of Qwen2.5-72B-Instruct steered by CSKS on DynamicQA in Figure 4, with varied control parameter α from -2.0 to +2.0. The accuracy was measured separately for each conflict partition type, as well as overall. We also provide results of other methods (Origin, Prompt, CAD and COIECD) and their comparison with CSKS in Appendix F. Consistent with our findings on synthetic datasets, CSKS demonstrates continuous control over the model's contextual sensitivity. As α increases, the overall accuracy monotonically improves, indicating enhanced faithfulness to the provided context.

3.5 Analysis

The Impact of Proxy Model Size To explore resource savings with smaller proxy models, we use the Qwen2.5 family (0.5B to 7B) to steer a 72B model under our framework. As shown in the Figure 5, the 0.5B proxy has a subtle but growing impact on the target model's sensitivity score, while the 1.5B proxy's impact already becomes very significant. A 3B proxy's impact is comparable, occasionally slightly better, than a 7B proxy. These results demonstrate our framework can adjust context sensitivity on a much larger model with significantly smaller overhead (e.g., using a 3B proxy). 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.,

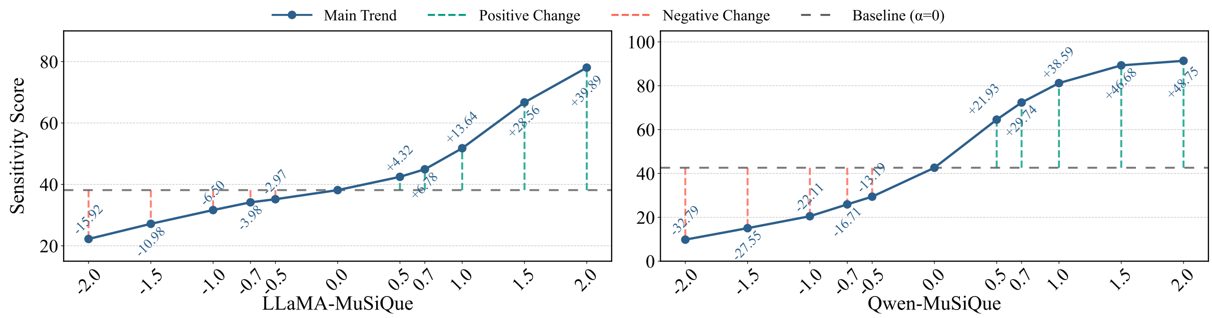


Figure 3: The performance of LLaMA and Qwen controlled bidirectionally, demonstrating the continuous adjustment capability of our method from two directions.

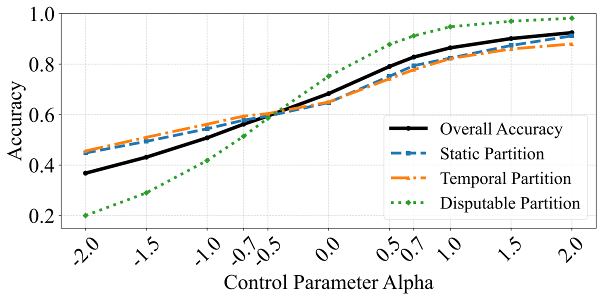


Figure 4: Accuracy of Qwen2.5 steered by CSKS on the DynamicQA dataset as a function of the control parameter α. Results are shown overall and broken down by conflict partition type, demonstrating CSKS's effectiveness and continuous controllability in handling diverse real-world knowledge conflicts.

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.

2021) for world knowledge understanding ability of LLMs (complex reasoning on 2WikiMultiHopQA (Ho et al., 2020) is detailed in Appendix G). For simplicity, we tested on two tasks from each of MMLU's four subjects (STEM, Humanities, Social, and Other). 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 (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 target 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 with substantial α variation, the target 72B model consistently outperforms the 7B model by significant margins (average +8.67%), demonstrating our framework successfully leverages the large model's superior general ability alongside precise sensitivity control. These findings 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 the 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 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 occur when contextual knowledge contradicts parametric knowledge (Mallen et al., 2023; Xu et al., 2024; Kortukov et al., 2024). Previous research often prioritized contextual knowledge over parametric knowledge for LLM responses (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

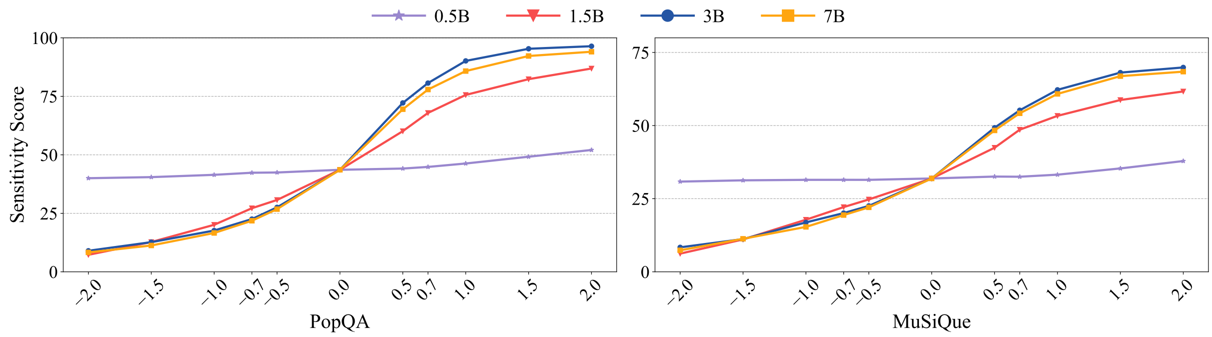


Figure 5: The performance of CSKS under varying proxy model sizes on MuSiQue and PopQA respectively. Smaller proxy models (0.5B, 1.5B) have a marginal yet increasing effect on the 72B target model's sensitivity score. Remarkably, the 3B proxy model matches the 7B model in sensitivity adjustment, validating that our framework enables potent context sensitivity modulation using substantially smaller models.

be of high quality (e.g., containing updated knowledge). However, varying context quality across scenarios means that a constant reliance on context is insufficient—an underexplored issue. We advocate for precise, continuous control over LLMs' contextual reliance and propose an effective, efficient framework to achieve this. Another line of work focuses on evaluating and understanding LLMs in knowledge conflicts and mining factors affecting LLMs' choice in knowledge conflicts. For instance, contextual detail affects LLM choices (Wu et al., 2024a; Tan et al., 2024a); LLMs favor popular entity information and are sensitive to data presentation order (Xie et al., 2023); models resist obviously false permuted knowledge (Qian et al., 2024); and increased conflicting hops challenge LLM reasoning (Jin et al., 2024). We leverage these key factors to measure knowledge manipulation difficulty and offer a more comprehensive evaluation method. 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 LMs' growing capabilities (Li et al., 2023b), many studies focus on controlling certain attributes of LM generation, usually non-toxicity and positive sentiment. Representation engineering is a common solution. 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. Zeng et al. (2025) and Tan et al. (2024b) leverage LLMs' internal representations for knowledge integration and security. 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 using small LMs as proxies to adjust output distributions of LLMs, thus controlling LLMs' sensitivity to knowledge provided in context. We also introduce a fine-grained evaluation method for this sensitivity. Extensive experiments demonstrate that our framework achieves state-of-the-art, more crucially, achieves precise and continuous control over how LLMs utilize information from context.

Limitations

While we show CSKS's effective control of LLMs in knowledge adaptation, the optimal calibration of the guiding hyperparameter α may vary in real scenarios where a balance between knowledge adaptation and LLMs' general abilities is essential. Future research could further explore methods for automatically or more adaptively determining the value of α to enhance the practical flexibility of the CSKS framework.

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.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

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

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. 2024a. 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.

Xue Tan, Hao Luan, Mingyu Luo, Xiaoyan Sun, Ping Chen, and Jun Dai. 2024b. Knowledge database or poison base? detecting rag poisoning attack through llm activations. arXiv preprint arXiv:2411.18948.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118.

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.

Shenglai Zeng, Jiankun Zhang, Bingheng Li, Yuping Lin, Tianqi Zheng, Dante Everaert, Hanqing Lu, Hui Liu, Hui Liu, Yue Xing, Monica Xiao Cheng, and Jiliang Tang. 2025. Towards knowledge checking in retrieval-augmented generation: A representation perspective. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2952–2969, Albuquerque, New Mexico. 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 Construction 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 in our main experiment. To achieve optimal results, we have designed a specialized pipeline for constructing the fine-tuning dataset as shown in Figure 6.

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.

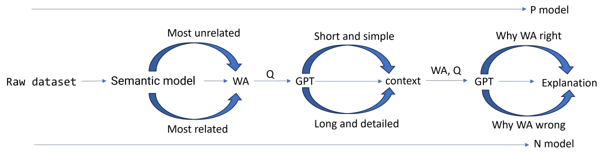


Figure 6: The pipeline to get the data used to finetune our $\mathcal{P}$ model and $\mathcal{N}$ model

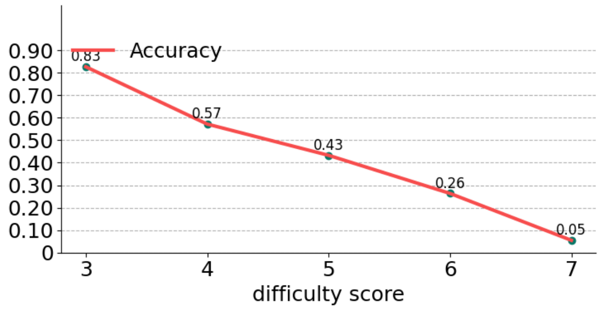


Figure 7: The accuracy of the LLaMA-3-70B-Instruct model across questions of each difficulty score.

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 7. The results reveal that as question difficulty increases, accuracy correspondingly decreases. This demonstrates that our grading system successfully quantifies problem difficulty.

C Baselines

The baselines we adopt in our main experiment are:

- **Origin**: refers to naive LLMs without any modifications.

- **Prompt**: prompts LLMs with explicit instructions to ensure their answers align with the given context.

- **IRCAN** : identifies context-responsive neurons within the LLM's feedforward network (FFN) layers and enhances their activation to improve the utilization of contextual information.

- **CAD** : is a decoding-time strategy that adjusts the output probabilities of LLMs to emphasize differences between context-aware and context-agnostic scenarios.

- **COIECD** : 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 hypeparameters 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 conducted a verification on small model. The results are presented in Appendix D, 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.

D Fine-tune results on small models

Figure 8 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.

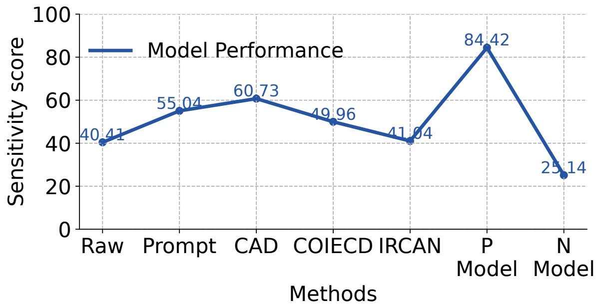
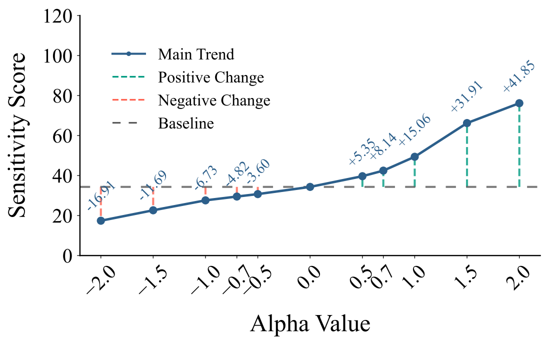


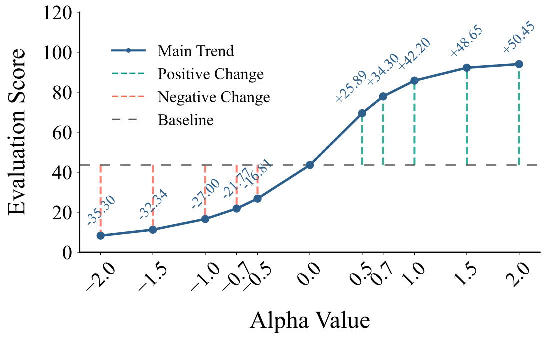
Figure 8: The effects of different methods on the LLaMA-3-8B-instruct model tested on PopQA.

E 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 9: Sensitivity score variation with alpha values on PopQA.

F Performance Comparison on DynamicQA

Figure 10 presents a head-to-head comparison of these methods across overall accuracy and specific conflict partition types (Static, Temporal, and Disputable) on DynamicQA. Across all evaluated dimensions, CSKS consistently and substantially outperforms all baseline approaches.

The consistently superior performance of CSKS across diverse real-world conflict types underscores its robustness and practical advantages over existing methods for managing knowledge conflicts in LLMs. The substantial margins, especially in the more challenging Disputable partition, further validate the efficacy of our proxy-based steering mechanism.

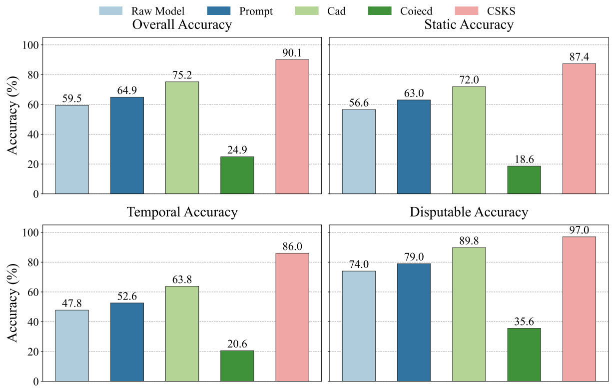


Figure 10: Comparative performance (Accuracy %) of CSKS and baseline methods (Raw Model, Prompt, CAD, COIECD) on the DynamicQA dataset. Results are shown for Overall Accuracy and broken down by conflict partition types: Static, Temporal, and Disputable. CSKS consistently outperforms all baseline methods across all categories.

Alpha (α)	EM Score	F1 Score
-2.0	48.00	59.08
-1.5	52.00	62.54
-1.0	53.50	64.33
-0.7	54.50	64.99
-0.5	54.50	64.67
72B ($\alpha = 0$)	**54.50**	**64.78**
+0.5	53.63	64.11
+0.7	53.37	63.79
+1.0	52.62	62.89
+1.5	50.75	60.67
+2.0	46.62	56.60
3B (baseline)	**26.37**	**38.35**

Table 4: Performance (EM and F1 scores) of Qwen steered by CSKS on the 2WikiMultiHopQA multi-hop reasoning benchmark for different α values. Results for a 3B baseline model are also shown for comparison.

G The CSKS impact on reasoning ability

To further investigate the impact of CSKS on more complex reasoning abilities, which was a concern raised in previous reviews, we evaluate the model on the 2WikiMultiHopQA dataset, a benchmark designed to test multi-hop reasoning capabilities through questions requiring connecting information from multiple sources.

The results on 2WikiMultiHopQA (Table 4) show a similar trend to MMLU regarding the influence of α. The highest EM and F1 scores are achieved when α is close to 0 (e.g., $\alpha \in [0.0, 0.7]$). As α increases, indicating stronger steering towards either contextual or parametric knowledge, there is a gradual decline in multi-hop reasoning performance. For instance, at $\alpha = +2.0$, the EM score drops to 46.62 from a peak of 54.50. However, it is crucial to note that even at these more extreme α values, the performance of the 72B model (e.g., 46.62 EM at $\alpha = +2.0$) remains significantly higher than that of a much smaller 3B model (26.37 EM), which struggles with the inherent complexity of the task. This suggests that while very strong steering can impact complex reasoning, the CSKS framework, within a moderate range of α, allows for effective context sensitivity adjustment while largely preserving the sophisticated reasoning capabilities of the large model.

The performance decline observed on both MMLU and 2WiKiMultiHopQA suggests that extreme sensitivity adjustments may disrupt the target 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/3B proxy models by significant margins (average $+8.67\%$ on MMLU, and substantially higher EM/F1 on 2WikiMultiHopQA), demonstrating that our framework successfully leverages the large model's superior general ability and reasoning capacity while achieving precise context sensitivity control. These findings collectively indicate that strategic α selection can achieve an effective equilibrium between contextual adaptability and model capability preservation, fulfilling our framework's dual objectives of precise knowledge steering and performance maintenance.

H CSKS results on Gemma-2-27b-it

To further substantiate the transferability of our CSKS framework and demonstrate its generalization capabilities across diverse LLM architectures, we extended our empirical validation to the Gemma-2 model family. For these experiments, Gemma-2-27b-it was utilized as the target large language model, with its smaller counterpart, Gemma-2-2b-it, serving as the proxy model maintaining the ratio of the size of the target model to the proxy model at approximately 10 to 1. We evaluated performance

```
### 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 11: The prompt we use to ask gpt to make a slight permutation.

on both the PopQA and MuSiQue datasets, maintaining the same experimental setup and metrics as used for the Llama-3 and Qwen2.5 models. The comprehensive results for the Gemma-2-it models are presented in Table 5.

It is noteworthy that the Gemma-2-27b-it model exhibits a relatively strong baseline performance compared to the Llama-3 and Qwen2.5 models evaluated earlier. Despite this higher baseline, CSKS consistently delivered substantial and leading improvements across both datasets. The successful application of CSKS to the Gemma-2 architecture, which differs from the previously tested models, provides compelling evidence for the framework's broad applicability and robust generalization. These results effectively address concerns regarding transferability, highlighting CSKS as a versatile solution for steering knowledge sensitivity in large language models.

I Prompts used to generate our synthesized dataset

Figure 11 - Figure 14 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 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 12: 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 13: The prompt we use to ask gpt to generate a short context.

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	
PopQA • Gemma-2-it								
Origin	82.81	52.42	81.71	53.52	69.00	67.89	66.02	64.49
PROMPT	87.44 (+4.63)	68.28 (+15.86)	**85.24** (+3.53)	70.48 (+16.96)	77.00 (+8.00)	77.59 (+9.70)	78.96 (+12.94)	76.30 (+11.81)
CAD	87.88 (+5.07)	66.96 (+14.54)	88.54 (+6.83)	66.29 (12.77)	76.33 (+7.33)	77.92 (+10.03)	77.99 (+11.97)	75.37 (+10.88)
COIECD	84.14 (+1.33)	54.62 (+2.20)	82.37 (+0.66)	56.38 (+2.86)	71.00 (+2.00)	69.56 (+1.67)	67.63 (+1.61)	66.38 (+1.89)
CSKS	**88.98** (+6.17)	**70.70** (+18.28)	84.80 (+3.09)	**74.88** (+21.36)	**81.33** (+12.33)	**79.26** (+11.37)	**78.96** (+12.94)	**80.47** (+15.98)
MusiQue • Gemma-2-it								
Origin	85.13	40.42	72.76	52.86	68.16	60.71	59.62	59.01
PROMPT	88.95 (+3.82)	50.00 (+9.58)	75.10 (+2.34)	63.90 (+11.04)	73.63 (+5.47)	67.20 (+6.49)	67.70 (+8.08)	66.60 (+7.59)
CAD	88.32 (+3.19)	53.61 (+13.19)	78.51 (+5.75)	63.48 (+10.62)	77.17 (+9.01)	67.85 (+7.14)	68.01 (+8.39)	67.89 (+8.88)
COIECD	85.56 (+0.43)	42.55 (+2.13)	73.82 (+1.06)	54.35 (+1.49)	69.45 (+1.29)	61.36 (+0.65)	61.49 (+1.87)	60.43 (+1.42)
CSKS	**93.63** (+8.50)	**70.85** (+30.43)	**83.82** (+11.06)	**80.68** (+27.82)	**84.89** (+16.73)	**80.19** (+19.48)	**81.67** (+22.05)	**80.75** (+21.74)

Table 5: Performance of CSKS and baseline methods on PopQA and MuSiQue datasets using Gemma-2-27b-it as the target LLM and Gemma-2-2b-it as the proxy model. Results show accuracy (in %) for different ranks of perturbation, contextual detail, and popularity, along with the overall Sensitivity Score. Improvements by CSKS over the Origin are shown in magenta.

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