Accommodate Knowledge Conflicts in Retrieval-augmented LLMs: Towards Robust Response Generation in the Wild

Jiatai Wang[1,2], Zhiwei Xu[2,3*], Di Jin[5], Xuewen Yang[4], Tao Li[1,2*]
[1]The College of Computer Science, Nankai University, Tianjin, China
[2]Haihe Lab of ITAI, Tianjin, China
[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[4]The Department of Electrical and Computer Engineering, Stony Brook University, New York, U.S.A
[5]Eigen AI, Palo Alto, U.S.A
1120240357@mail.nankai.edu.cn, xuzhiwei2001@ict.ac.cn, di@eigenai.com, michelyang1990@gmail.com, litao@nankai.edu.cn

Abstract

The proliferation of large language models (LLMs) has significantly advanced intelligent systems. Unfortunately, LLMs often face knowledge conflicts between internal memory and retrieved external information, arising from misinformation, biases, or outdated knowledge. These conflicts undermine response reliability and introduce uncertainty in decision-making. In this work, we analyze how LLMs navigate knowledge conflicts from an information-theoretic perspective and reveal that when conflicting and supplementary information exhibit significant differences, LLMs confidently resolve their preferences and alleviate the uncertainty during their response generation. When this difference is ambiguous, LLMs experience considerable uncertainty about their generation. Based on this insight, we propose Swin-VIB, a novel framework that integrates a pipeline of variational information bottleneck models to adapt the retrieved information difference, facilitating robust response generation of LLMs even in conflicting contexts. Extensive experiments confirm our theoretical analysis and demonstrate the performance of Swin-VIB. Notably, Swin-VIB outperforms all competitive baselines in terms of the accuracy of the multiple-choice task, while improving the EM values in the open-ended QA task by at least 11.14%.

1 Introduction

As large language models (LLMs) (Brown et al. 2020; OpenAI 2023) continue to proliferate, their powerful generation capabilities facilitate more transformative technologies for advanced intelligent systems. Among these novel technologies, despite the response generation technique (Li et al. 2024) allows LLMs to answer queries without rigid retrieval, it still suffers from hallucinations. Augmented LLMs with external knowledge, including retrieval augmented generation (RAG), have become a promising solution to mitigating these issues (Lewis et al. 2020; Gao et al. 2023; Asai et al. 2023). However, the disparity between the internal memory of LLM and the retrieved external context always leads to knowledge conflicts (Xu et al. 2024). These conflicts arise from misinformation, unreliable sources, and publisher bias in retrieved information, and the difficulty in synchronizing

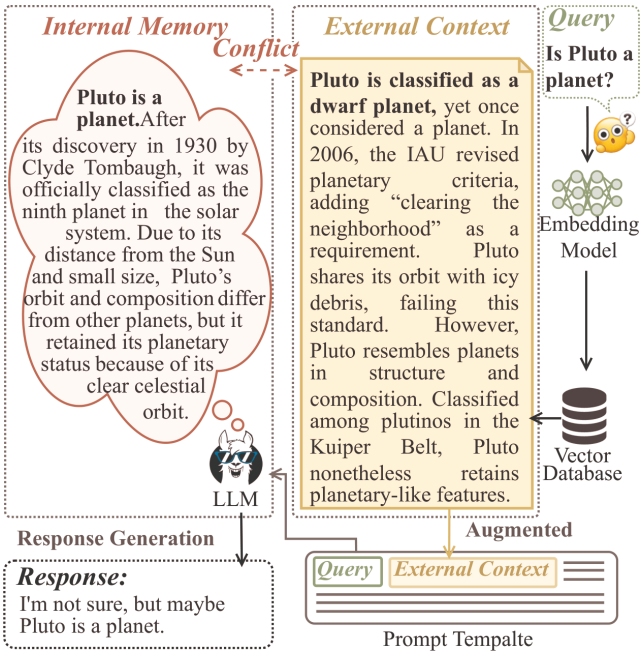

Figure 1: Illustration of knowledge conflict in RAG.

parametric knowledge with the context in real time further exacerbates these conflicts. As illustrated in Figure 1, these conflicts introduce uncertainty in response generation(Wu, Wu, and Zou 2024), posing a serious threat to the reliability of the responses and increasing the risk of biased or erroneous inference.

The existing approaches do not fully agree on how LLMs resolve knowledge conflicts. Some works prefer to fine-tune or edit LLMs according to external context and guiding response generation. (Shi et al. 2023b; Jin et al. 2024b; Zhang, Yu, and Feng 2024). In contrast, others attach external-validation modules that merely admit trusted context to supplement the internal memory of LLMs (Yu, Zhang, and Feng 2024; Kortukov et al. 2024). However, both of them have a fixed preference for the used knowledge that is often violated. This contradiction exposes a fundamental dilemma

*Corresponding authors

that without a principled framework explaining how such preferences emerge, either fine-tuning or external-validation modules risks the wrong optimizing results. Although new efforts try to re-balance internal memory and external context during model decoding (Shi et al. 2023a; Jin et al. 2024a; Huang et al. 2025), their token-level weight updating fails to scale to mitigate fundamental knowledge conflicts in practice. Therefore, a theoretical analysis of the knowledge preference and uncertainty is highly desired to extend the existing heuristic guidelines toward knowledge conflicts mitigation and robust response generation.

To address this challenge, we re-examine knowledge conflict in response generation from the perspective of information theory. We analyze how LLMs establish preferences on the information used to generate responses when faced with conflicting knowledge sources and observe:

- When the disparity between conflicting and supplementary information is significant, LLMs confidently settle into a rational knowledge preference and generate reliable responses.
- When the distinction is ambiguous, LLMs experience extreme uncertainty, making their response unreliable.

Inspired by these insights, we propose Swin-VIB, a Variational Information Bottleneck (VIB) approach with a Sliding window. More specifically, we leverage a pipeline-based multiple variational information bottleneck models (Alemi et al. 2016) to augment the retrieved information adaptively and guide the preference of LLMs. This design adapts the LLM's preference and minimizes ambiguity, enabling more accurate and consistent response generation even in challenging scenarios with knowledge conflicts. Swin-VIB can be plugged into RAG pipelines almost without any additional overhead. Our major contributions are summarized as follows:

- We model the interplay between LLMs' internal memory and external context and release the preference principle behind knowledge conflict in retrieval-augmented LLMs. With significant differences between conflicting and supplementary information, LLMs have more confidence to settle into a rational preference; otherwise, LLMs fall into extremely high uncertainty.
- This analysis of LLMs' preference provides insight into accommodating knowledge conflicts in retrieval-augmented LLMs. In this way, we propose Swin-VIB[1], a sliding-window approach that integrates multiple variational information bottlenecks to adapt perplexing knowledge, enhancing it to guide LLMs toward accurate responses.
- Extensive experiments across multiple-choice and open-ended question answering (QA) validate our theoretical result and demonstrate that Swin-VIB outperforms baselines for robust response generation in the wild.

To the best of our knowledge, this is the first approach that analyzes and evaluates knowledge conflicts of retrieval-augmented LLMs. A robust response generation paradigm is proposed according to the obtained insight, rather than merely relying on empirical evaluations, thus enjoying enhanced performance.

[1]The code is available at https://github.com/JiataiWang/Swin-VIB.

2 Preliminary

2.1 Definitions

Knowledge conflicts may cause uncertainty in response generation, where LLMs need to make a choice with conflicting knowledge from their internal memory and the external context. Considering the black-box nature of LLMs, it remains challenging to analyze the fundamental mechanism of these conflicts between invisible internal knowledge and external contexts. Without a theoretical analysis of knowledge conflicts, the limitations of empirical rules and experimental settings in response generation can hardly alleviate knowledge conflicts (Wu, Wu, and Zou 2024; Xie et al. 2023; Jin et al. 2024a). To tackle this challenge, we propose a theoretical framework that shows knowledge conflicts can be defined in terms of conditional entropy. The corresponding symbol system is outlined in Table 1.

Symbol	Meaning
Q	Queries
$R = B(Q)$	External contexts from the knowledge base B
$O = \text{LLM}(R, Q)$	Responses generated by the LLM

Table 1: A list of Symbols

Definition 1 (***Uncertainty of response generation***)**.** *The uncertainty of O given the Q and R is represented by conditional entropy $\mathbb{H}$, is formalized by*

$$\mathbb{H}(O \mid R, Q) = -\sum_{o \in O, r \in R, q \in Q} p(o, r, q) \log p(o \mid r, q), \tag{1}$$

A commonsense assumption is taken to regulate the RAG process.

Assumption 1 (***Non-void Retrieval***)**.** *A qualified external information retriever can recall external contexts correlated with the corresponding query (Robertson and Jones 1976), that is $\forall q \in Q$, $\exists r \in R$, $p(r, q) > 0$, where q is a query, and r is one correlated external context.*

2.2 Theoretical Analysis

We reveal how the difference between conflicting and supplementary information influences the reliability of LLMs. A brief analysis is listed as follows, and all details about this analysis are provided in Appendix A.

Step 1: Uncertainty formulation. According to Definition 1 and Assumption 1, we rewrite Formula 1 based on the law of total probability and the chain rule (see Appendix A.1):

$$\mathbb{H}(O \mid R, Q) = \sum_{q} p(q) \sum_{r} p(r \mid q) \sum_{o} \psi\big(p(o \mid r, q)\big), \tag{2}$$

where $\psi(\cdot)$ denotes instance-level uncertainty, and is calculated by

$$\psi(p(o \mid r, q)) = -p(o \mid r, q) \log p(o \mid r, q) \tag{3}$$

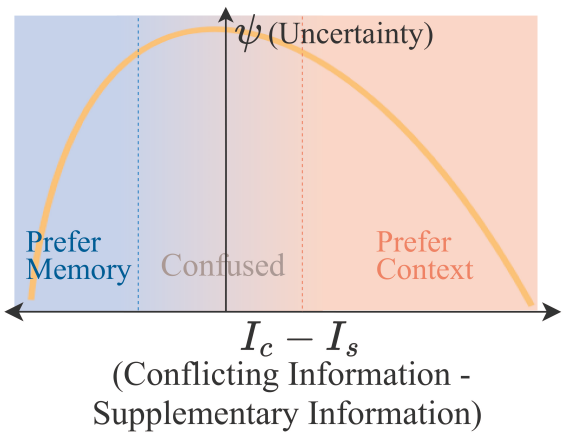

Figure 2: Relationship between uncertainty and the information difference.

Step 2: Latent-state decomposition. Let X be the high-dimensional latent space of the LLM, and let $x \in \mathbf{X}$ denotes a specific latent state activated during generation (see Appendix A.2). We have:

$$\begin{aligned} p(o \mid r, q) &= \int_X p(o \mid r, q, x) p(x \mid r, q)\, dx \\ &= \int_X p(o \mid r, q, x) \frac{p(r, q \mid x) p(x)}{p(r, q)}\, dx \\ &\propto \int_X p(o \mid r, q, x) p(r, q \mid x) p(x)\, dx \\ &\propto \int_X p(o \mid r, q, x) \frac{p(r, q \mid x)}{p(r, q \mid x_\gamma)} p(x)\, dx \\ &\approx \int_X p\,(o \mid r, q, x)\, exp[\log p\,(r \mid x) - \\ &\qquad \log p\,(r \mid x_\gamma)] p(x) dx \\ &\propto \int_X p(o \mid r, q, x)\ \exp\big[I_c - I_s\big]\ p(x)\, dx, \end{aligned} \tag{4}$$

where self-information $I_s = -\log p(r \mid x)$ denotes the external information consistent with the LLM's memory, namely, supplementary information, whereas $I_c = -\log p(r \mid x_\gamma)$ denotes the information that contradicts internal memory and is defined as conflicting information, with $x_\gamma \in X$ being a latent state aligned to the retrieved context r.

Step 3: Approximation Leveraging the monotonicity of $\exp(\cdot)$, and the constants are reduced after normalizing $p(o \mid r, q)$ over o, we apply a Taylor expansion and arrive at the following approximation (see Appendix A.3):

$$p(o \mid r, q) \ \propto\ I_c - I_s. \tag{5}$$

Equation (5) is a first-order proxy of the exponential, and monotonic claims in Step 4 rely only on the monotonicity of the exponential tilt (proved in Appendix A.3).

Step 4: Theoretical statement. According to Formula (5), Formula (3) reveals a relation between uncertainty ψ and the information difference $\Delta I = I_c - I_s$. Larger $|\Delta I|$ consistently corresponds to lower ψ, *i.e.* $|\Delta I| \uparrow \Longleftrightarrow \psi \downarrow$, which can also be found in Figure 2. Equivalently, the conditional entropy is the weighted expectation of instance-level uncertainty,

$$\mathbb{H}(O \,|\, R, Q) = \mathbb{E}_{q \sim p(q)}\, \mathbb{E}_{r \sim p(r|q)} \Big[\sum_o \psi\big(p(o \mid r, q)\big) \Big], \tag{6}$$

So lowering the per (q, r) uncertainty in expectation lowers $\mathbb{H}(O \mid R, Q)$. The full derivation can be found in Appendix A.4.

3 Methodology

According to our theoretical analysis, a larger information difference $|\Delta I|$ of conflicting and supplementary information enhances response reliability, and thus we take this insight to design Swin-VIB, which integrates a pipeline of variational information bottleneck models to adapt the external information in those context windows with large $|\Delta I|$.

3.1 Overview

To maximize the information difference between conflicting and supplementary information in the external context, the retrieved context is first segmented into fixed-length windows to give a unified unit on which the information difference is measured. Specifically, Swin-VIB slides over these windows and quickly predicts the difference in information contained in each window. The window with a large enough information difference is accepted; otherwise, the window is rejected. Finally, Swin-VIB concatenates the accepted windows with the query to form a prompt, as shown in Figure 3.

3.2 Data Preparation

Given a dataset $\mathcal{D} = \{(q, r)\}$, where each query q is paired with an external context $r \in \{r_c, r_s, r_m\}$, r_c is conflicting context, r_s is supplementary context, and r_m is their mixture. To construct the mixed context r_m, conflicting and supplementary contexts interleave each other by every 4 tokens, eliminating the possibility of sampling a window only with conflicting context or supplementary context. For each query q, we randomly extract a fixed-length window of length 7 from the context r.

$$\omega^* \leftarrow \text{RandomWindow}(r, \text{len} = 7), \tag{7}$$

and for this window, a binary label is assigned as

$$\mathbf{Y} = \begin{cases} 1, & \text{if } r \text{ is from single source, } r \in \{r_c, r_s\} \\ 0, & \text{if } r \text{ has multiple sources, } r = r_m \end{cases} \tag{8}$$

This setting guarantees $\mathbf{Y}$ represents the information diffidence $|\Delta I|$. Because $r \in \{r_c, r_s\}$ denotes all tokens originate from a single source, the ω^* carries a maximal difference between conflicting and supplementary information. If $r \in \{r_m\}$, the ω^* inevitably contains multiple optional instances from different sources and therefore exhibits a small information difference, matching the multiple optional condition.

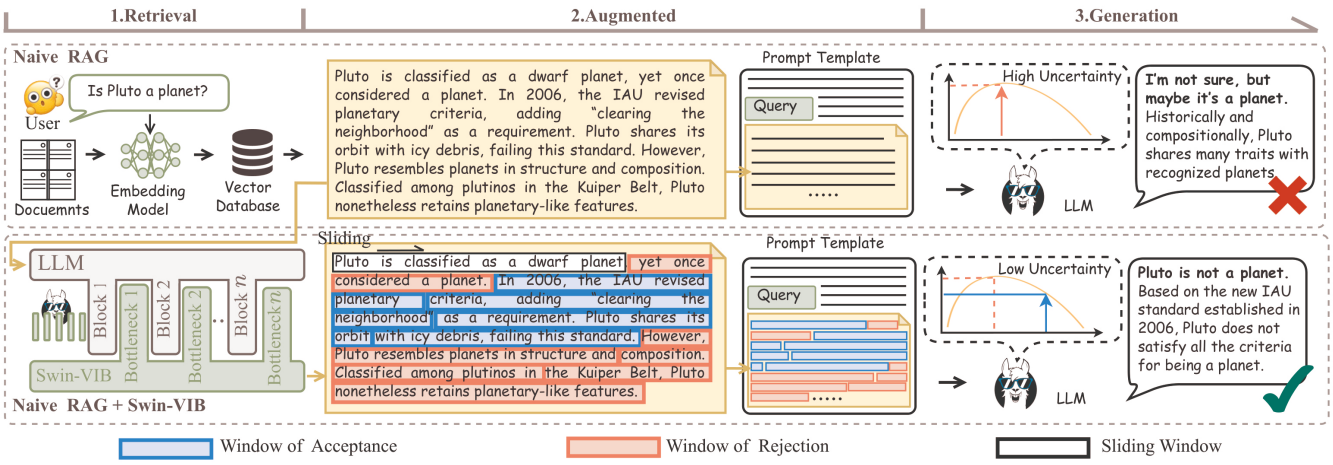

Figure 3: An overview of response generation with Swin-VIB.

Algorithm 1: Training-Set Construction

Input: Dataset $\mathcal{D}$; LLM $\mathcal{M}$ with N layers
Output: Training sets $T_n = (\mathbf{G}, \mathbf{Y})_{n=1}^{N}$

Initialise $T_n \leftarrow [\,]$ for $n = 1, \ldots, N$;
for *each sample* $\{q, r\} \in \mathcal{D}$ **do**
  $\omega^* \leftarrow \text{RandomWindow}(r, \text{len})$;
  $\mathbf{Y} \leftarrow \begin{cases} 1, & \text{if } r \in \{r_c, r_s\} \\ 0, & \text{if } r = r_m \end{cases}$;
  for $n \leftarrow 1$ **to** N **do**
    $\mathbf{A}^{(n)} \leftarrow \text{Attention}(\omega^*, n)$;
    $\mathbf{G}_n \leftarrow \text{MeanHeads}(\mathbf{A}_n)$;
    append $(\mathbf{G}_n, Y)$ to T_n;
return $\{T_n\}_{n=1}^{N}$; `// each saved via pickle.dump`

Feeding each ω^* into a frozen LLM yields a stack of per-layer attention matrices $\{\mathbf{A}_n\}$. We average heads to obtain

$$\mathbf{G}_n \;=\; \textsc{MeanHeads}(\mathbf{A}_n), \tag{9}$$

and construct the training set

$$T_n \;=\; \{\mathbf{G}_n,\, \mathbf{Y}\}^{N} \tag{10}$$

All $\mathbf{G}_n$ of LLM decoder layers share the same $\mathbf{Y}$. Detailed in Algorithm 1.

3.3 Swin-VIB Architecture & Training

We propose a novel method Swin-VIB to compress various information and extract their differences adaptively. Inspired by the information bottleneck structure (Tishby, Pereira, and Bialek 2000), Swin-VIB introduces an information bottleneck into each transformer decoder layer to predict the information difference of the external context. In the following, we describe how the constructed training sets T_n are fed into these bottlenecks for training.

As shown in Figure 4, a bottleneck model consists of an encoder and a decoder, the encoder projects $\mathbf{G}_n$ of an LLM

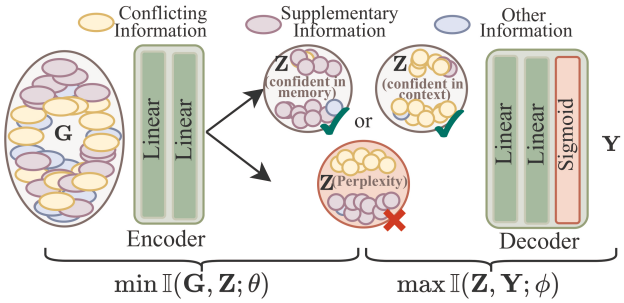

Figure 4: Bottleneck model structure and training objective.

layer to the mean $\boldsymbol{\mu}_n$ and log-variance $\log \boldsymbol{\sigma}_n^2$ of each Gaussian latent feature.

$$q_\theta(\mathbf{Z}_n \mid \mathbf{G}_n) = \mathcal{N}\big(\boldsymbol{\mu}_n, \operatorname{diag}(\boldsymbol{\sigma}_n^2)\big). \tag{11}$$

where θ and ϕ are the parameters of the encoder and the decoder. Via the reparameterization trick (Kingma and Welling 2013), we obtain a differentiable latent representation,

$$\mathbf{Z}_n = \boldsymbol{\mu}_n + \boldsymbol{\sigma}_n \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{12}$$

The decoder maps $\mathbf{Z}_n$ to the logit of $\hat{Y}_n = p_\phi(Y = 1 \mid \mathbf{Z}_n)$, and predicts whether a window includes a larger information difference. Each bottleneck is trained independently by minimizing the information bottleneck loss:

$$\begin{aligned} \mathcal{L}_n(\theta, \phi) = & \underbrace{\mathbb{E}_{q_\theta}\big[-\log p_\phi(Y \mid \mathbf{Z}_n)\big]}_{\uparrow \mathbb{I}(\mathbf{Z}_n; Y)} + \\ & \underbrace{\beta \mathrm{KL}\big(q_\theta(\mathbf{Z}_n \mid \mathbf{G}_n) | p(\mathbf{Z})\big)}_{\downarrow \mathbb{I}(\mathbf{G}_n; \mathbf{Z}_n)} \end{aligned} \tag{13}$$

where $p(\mathbf{Z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an isotropic prior and parameter β controls the compression–prediction trade-off. It is by minimizing the mutual information $\mathbb{I}(\mathbf{G}, \mathbf{Z})$ to guide the encoder to adaptively learn the key features to identify information differences. The KL term upper-bounds $\mathbb{I}(\mathbf{G}_n; \mathbf{Z}_n)$, and thus makes the encoder discard redundant information.

3.4 Robust Response Generation

To achieve robust response generation, we incorporate the corresponding bottleneck into the inference process of each transformer-decoder layer of LLM. More specifically, the RAG retriever recalls an external context r, which is partitioned into fixed-length windows $r = \{\omega^1, \ldots, \omega^K\}$. Each window ω^k is fed to the frozen LLM to get attention representation $\mathbf{G}_n^k$, which is collected from every decoder layer. The representation is fed into the corresponding layer-specific bottleneck model, where the encoder's mean $\boldsymbol{\mu}$ is efficiently taken as the latent $\mathbf{Z}_n$ to skip sampling. The bottleneck returns a layer-wise probability $p_{\phi_n}(Y{=}1 \mid \mathbf{G}_n^k)$ that predicts the information difference in terms of different context windows. The outputs of all bottlenecks are weighted to achieve an averaging result to guide robust response generation:

$$\hat{p}(q, \omega^k) = \frac{1}{N} \sum_{n=1}^{N} p_{\phi_n}\big(Y{=}1 \mid \mathbf{G}_n^k\big), \tag{14}$$

A window is accepted if $\hat{p}(\omega^k) \geq \xi$, where ξ is a threshold. All accepted windows are concatenated with the query to construct the final prompt $[q; \{\omega^k \mid \hat{p}(\omega^k) \geq \xi\}]$. In this way, the LLM is guided to generate reliable responses, even in the case that the context contains a large amount of conflicting information.

4 Experiments

In this section, we verify our theoretical analysis and demonstrate that Swin-VIB can reliably guide LLMs to adapt to conflicting external context and generate the correct responses.

4.1 Experimental Settings

Datasets We use three popular datasets to evaluate the performance of Swin-VIB. Note that these datasets contain supplementary contexts and conflicting contexts, but don't have any mixed contexts.

- ConflictQA (Xie et al. 2023) (2,839 samples): Each q contains internal knowledge "parametric memory" and conflicting external context "counter memory", mapping them to r_s and r_c, respectively.
- DRUID (Hagström et al. 2024) (1,003 samples): Each q contains a two-option item with mutually exclusive claim A/B (each paired with supporting evidence); the verified claim becomes r_s, the refuted one r_c.
- TruthfulQA (Lin, Hilton, and Evans 2021) (817 samples): Each q has one best-answer candidate which serves as the ground truth, multiple other correct responses, and a collection of incorrect alternatives. We use LLMs to select among the above options as r_s, unselected options are r_c.

Our construction only requires including queries paired with contexts from different sources. Specifically, given a query q, the LLM first answers, the chosen option is treated as supplementary to its internal memory, and the unchosen option as conflicting. For LLMs not covered by the original dataset, we follow the two-step procedure: (i) obtain a closed-book answer and rationale as Internal Memory; (ii) prompt ChatGPT to draft a coherent passage that contradicts those facts as External Conflicting Context. To provide the mixed contexts,

Prompt of Multiple-Choice

[INST] <<SYS>> According to the given information and your knowledge, choose the best choice from the following options. Follow the format: Option: your option. <</SYS>>
• Question: <Query> • Information: <External Context>
Options: A: <option a> B: <option b> [INST]

Prompt of Open-ended QA

[INST] <<SYS>> You are an assistant for question answering tasks. Use the following pieces of retrieved context and your own knowledge to answer the question. Among the multiple contexts retrieved , some are correct while others are incorrect. Use two sentences maximum and keep the answer concise. <</SYS>>
• Question: <Query> • Information: <External Context>
Answer: [INST]

Figure 5: Prompt templates in conflicting scenarios.

a supplementary context r_s and a conflicting context r_c are sampled from the dataset and interleaved with each other by every four tokens, and form a mixed context r_m.

Models Our evaluations are performed on five open-source LLMs, including Llama 2-7B, Llama 2-13B, Llama 2-70B (Touvron et al. 2023), Qwen3-8B (Team 2025), and DeepSeek-LLM-7B-base (DeepSeek-AI 2024).

Tasks and Metrics We include two primary tasks in our experiments to evaluate the response generation of Swin-VIB when LLM retrieves conflicting contexts, and the corresponding RAG performance:

1) *Multiple-choice*: This task can not only quantify the response generation preferences of LLMs but also can be used to evaluate the ability to exist methods to accommodate knowledge conflicts. Specifically, we provide two options for each query, one of which is obtained from the internal memory, and the other is from the external context. Only one option is correct, which constrains the generation space and yields confirmed answers.

- Memorization Preference Rate (MPR $= \frac{f_m}{S}$);
- Context Preference Rate (CPR $= \frac{f_c}{S}$);
- Uncertain Answer Rate (UAR $= \frac{S - f_m - f_c}{S}$),
- Accuracy Rate(ACC $= \frac{C_r}{S}$)
- Correction Rate (CR $= \frac{C_{crt}}{S-L}$)
- Resistance Rate (RR $= \frac{C_{def}}{L}$);
- Instance-level Uncertainty (Mean-ψ): Equal to the mean of the token-level negative log-likelihood values, its detailed computation procedure is provided in Appendix B.5;
- Total Response Entropy (TRE): Equal to $-(ACC \log_2(ACC + \epsilon) + (1 - ACC - UAR) \log_2[(1 - ACC - UAR) + \epsilon] + UAR \log_2(UAR + \epsilon))$;

where S is the total number of dataset instances; f_m is the number of instances for which the LLM relies on internal memory, and f_c is the number of instances using external contexts. L is the number of instances for which the LLM can generate correct responses without any context. C_r is the number of instances for which the LLM can generate correct responses with conflicting contexts. C_{crt} counts instances that are answered incorrectly without external contexts but

LLM	Method	ConflictQA						DRUID					
		ACC↑	CR↑	RR↑	Mean-ψ ↓	UAR↓	TRE↓	ACC↑	CR↑	RR↑	Mean-ψ ↓	UAR↓	TRE↓
Llama2-7B	Closed-book	20.61	-	-	0.31	-	-	50.58	-	-	0.34	-	-
	In-context	79.68	96.49	16.92	0.34	2.96	0.69	50.08	90.11	12.74	0.38	8.89	1.34
	CD^2	77.32	88.05	29.16	0.29	2.01	0.87	58.14	92.30	22.05	0.30	4.32	1.18
	Rowen-CL	80.10	92.59	28.55	0.28	1.30	0.79	52.35	94.76	17.30	0.29	5.92	1.26
	CK-PLUG	78.17	95.78	25.29	0.28	1.87	0.85	58.88	94.24	18.00	0.29	5.39	1.21
	Swin-VIB	**84.04**	**98.18**	**29.40**	**0.24**	**0.21**	**0.64**	**63.24**	**94.87**	**25.63**	**0.26**	**2.85**	**1.09**
Llama2-13B	Closed-book	21.37	-	-	0.30	-	-	50.93	-	-	0.33	-	-
	In-context	79.83	97.33	16.02	0.33	1.98	0.82	51.73	92.21	11.19	0.37	7.16	1.29
	CD^2	80.20	93.52	27.23	0.28	4.85	0.88	60.72	93.12	18.55	0.29	4.04	1.15
	Rowen-CL	80.36	95.42	28.55	0.27	1.10	0.76	52.61	94.85	16.93	0.28	4.36	1.21
	CK-PLUG	81.44	97.27	27.38	0.27	1.78	0.78	61.07	**95.18**	18.83	0.28	5.47	1.19
	Swin-VIB	**85.68**	**99.59**	**30.40**	**0.23**	**0.18**	**0.61**	**63.43**	95.16	**25.48**	**0.25**	**1.96**	**1.06**
Llama2-70B	Closed-book	20.99	-	-	0.29	-	-	49.59	-	-	0.32	-	-
	In-context	80.29	98.21	20.91	0.32	1.12	0.78	52.81	90.65	19.77	0.36	3.66	1.18
	CD^2	76.20	96.25	30.00	0.27	2.10	0.89	54.86	94.32	24.55	0.28	1.19	1.07
	Rowen-CL	81.43	97.04	29.46	0.26	0.95	0.74	54.44	95.60	23.89	0.26	0.97	1.06
	CK-PLUG	83.41	99.08	28.13	0.27	0.43	0.68	56.70	**97.16**	24.42	0.27	1.36	1.07
	Swin-VIB	**87.68**	**99.48**	**31.06**	**0.22**	**0.03**	**0.54**	**62.94**	96.86	**27.37**	**0.24**	**0.16**	**0.97**
DeepSeek-7B	Closed-book	16.42	-	-	0.30	-	-	37.93	-	-	0.33	-	-
	In-context	77.49	97.04	9.23	0.33	3.36	0.91	46.69	96.59	10.51	0.37	4.45	1.22
	CD^2	79.02	90.21	29.80	0.28	2.48	0.85	51.64	95.95	11.73	0.30	4.28	1.21
	Rowen-CL	81.83	96.29	50.20	0.28	2.32	0.78	43.49	94.21	9.87	0.28	4.77	1.22
	CK-PLUG	78.82	96.20	29.82	0.27	1.97	0.84	53.19	97.57	12.24	0.27	5.52	1.24
	Swin-VIB	**84.13**	**98.83**	**58.65**	**0.24**	**0.79**	**0.68**	**56.39**	**98.75**	**16.55**	**0.27**	**1.97**	**1.10**
Qwen3-8B	Closed-book	23.53	-	-	0.31	-	-	53.51	-	-	0.34	-	-
	In-context	78.36	97.25	15.89	0.34	2.11	0.85	56.88	96.12	12.24	0.38	3.44	1.16
	CD^2	74.42	95.18	26.44	0.30	3.06	0.96	57.60	98.58	12.49	0.32	3.97	1.17
	Rowen-CL	80.38	97.72	21.42	0.31	1.88	0.80	54.54	97.87	14.33	0.33	4.50	1.21
	CK-PLUG	78.61	96.11	28.15	0.29	1.89	0.84	57.86	98.52	**16.78**	0.29	3.50	1.17
	Swin-VIB	**82.81**	**98.68**	**32.47**	**0.25**	**0.64**	**0.82**	**58.69**	**99.09**	15.62	**0.28**	**1.83**	**1.08**

Table 2: Evaluation results of multiple-choice.

corrected after LLM retrieved external conflicting contexts. C_{def} counts instances that are answered correctly without external contexts and remain correctly answered after LLM retrieved conflicting contexts. ϵ is a small constant used to avoid numerical issues in logarithmic calculations.

2) *Open-ended QA*: We make the LLM generate responses, and evaluate response generation quality of RAG as well as its robustness in real-world conflicting scenarios. For each query, to evaluate an open-ended QA with retrieved context, the top five correct or incorrect answers for a query are retrieved to fill the <external context> slot in the prompt template. The template is depicted in Figure 5. According to the experimental configuration of open-ended QA tasks (Yu, Zhang, and Feng 2024), we use EM and Faithfulness (Es et al. 2023) to evaluate whether the model can provide correct answers, as well as METEOR (Denkowski and Lavie 2014) to assess the quality of the response generation.

Baselines and Implementation Details We compare Swin-VIB with five SOTA baselines (Closed-book, In-context, CD^2 (Jin et al. 2024a), CK-PLUG (Bi et al. 2025), Rowen-CL (Ding et al. 2024)) on the multiple-choice task, while integrating Swin-VIB and these baselines into three advanced RAG methods (Naive RAG, Self-RAG (Asai et al. 2023), Astute RAG (Wang et al. 2024)) for the open-ended QA task (see their implementation details in Appendix B.4). We implement Swin-VIB with fixed-length sliding windows (7 tokens) and set (ξ, β) to $(0.68, 10^{-5})$ on all datasets. Data preprocessing, hardware/software environment, and model configurations are described in Appendix B.1–B.3. During the training stage, the bottleneck loss converges within $\approx$ 200 epochs, and deeper decoder layers provide more robustness (see Appendix C.2 for convergence analysis). Swin-VIB achieves advanced generation quality to β in the range 10^{-5}–10^{-3}, and remains stable when the threshold ξ lies between 0.60 and 0.8. A window length of 7 tokens can trade off accuracy and inference latency. The windows with fewer tokens degrade the precision to accept these windows, whereas the windows with more than 7 tokens also slightly degrade acceptance precision. The parameters β and ξ, and the window length are analyzed in Appendix C.3.

4.2 Verification of Our Theoretical Analysis

To verify our theoretical analysis 2.2, we explore the impact of information difference on LLMs' preference and generation uncertainty on the multiple-choice task. To achieve that, the internal-memory facts (supplementary information) are collected and combined into the external context with controlled proportions. As depicted in Figure 2, we get two observations:

- The preferences of the LLM are influenced by the proportion of external conflicting and supplementary information. As the information difference grows, LLM becomes more focused on a certain preference, and the curve in Figure 2 mirrors this trajectory. (see Appendix C.1 for analysis of answer preference distributions)
- When the proportions are set at 2:2, the UAR values reach their maximum. This indicates that when the information difference is slight, the uncertainty of response generation is the greatest. The empirical "rise–then–fall" trend of the UAR in Figure 6 is consistent with the curve in Figure 2.

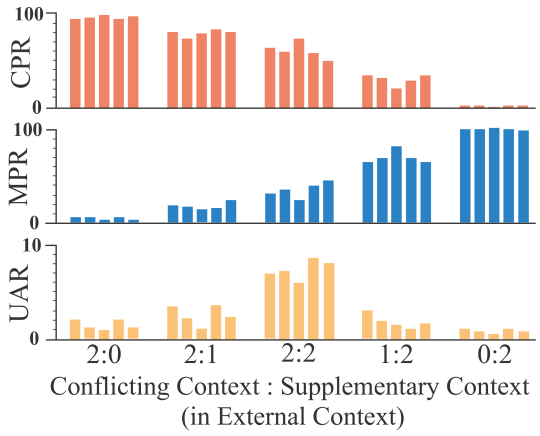

Figure 6: The uncertainty answer ratio of LLMs under varying proportions of external conflicting information from ConflictQA, e.g., 1:2, means that the external context includes one conflicting context and two supplementary contexts. For each bar group, the results for different LLMs are listed according to the order of Llama2-7B, Llama2-13B, Llama2-70B, Deepseek-7B and Qwen3-8B.

4.3 Comparisons with State of the Arts

Through the comparison of Swin-VIB and baselines, we can observe the following in Table 2: Swin-VIB significantly outperforms the state-of-the-art baselines on all settings and outperforms the strongest baseline by up to 6.24%. This demonstrates its robustness of response generation in the conflicting scenario. Swin-VIB attains the best trade-off between CR and RR, reflecting both strong capability for error recovery and resilience to misleading context. This enables Swin-VIB to generate more reliable responses. On the other hand, it considerably reduces UAR, so the LLMs have less probability of skipping answering the queries. To specifically measure the uncertainty of response generation, following the literature (Zhu et al. 2024), we take the output distribution of the LLM to measure the macro-level uncertainty of LLMs with the metric, total response entropy (TRE). The corresponding results prove that the adaptation principle of Swin-VIB is quite effective in accommodating knowledge conflicts and improving response reliability. Similarly, the decrease in Mean-ψ confirms that Swin-VIB maintains smaller micro-level uncertainty. Actually, the Pearson correlation between $Mean - \psi$ and TRE is 0.81 (see Appendix C.5), validating the high robustness of Swin-VIB. Finally, to verify whether Swin-VIB has selected the windows with larger information difference, we conducted a small-scale human annotation study. We find that 76 % of the windows discarded by Swin-VIB were judged "undecidable" (no clear conflicting or supplementary information) and the result confirms that our method indeed amplifies the information difference in a rational way (see Appendix C.6).

4.4 Evaluation of RAG with Swin-VIB

This section examines how integrating Swin-VIB into advanced RAG frameworks, enhances practical applicability under real-world settings, including evaluating generation quality and inference latency.

Methods	EM↑	Meteor↑	Faithfulness ↓
Naive RAG	46.50	42.85	73.03
Naive RAG + Swin-VIB	60.14	50.95	66.58
Self-RAG	39.39	43.38	70.10
Self-RAG + Swin-VIB	58.32	55.29	64.32
Astute RAG	53.02	44.69	65.23
Astute RAG + Swin-VIB	64.16	53.70	64.09

Table 3: Evaluation of Open-ended QA tasks on TruthfulQA

LLM	Naive RAG	Our cost per window↓	Our cost↓
Llama2-7B	0.4912s	0.08ms	0.3913s
Llama2-13B	0.8791s	0.14ms	0.5221s
Llama2-70B	4.6013s	0.55ms	2.7830s
Qwen3-8B	0.3862s	0.12ms	0.2740s
DeepSeek-7B	0.3770s	0.11ms	0.4293s

Table 4: Efficiency Evaluation of Open-ended QA tasks on TruthfulQA

As shown in Table 3, Astute-RAG employs an iterative fusion step in a data augmented stage, and thus raises decision variance and generation uncertainty. From the beginning, Swin-VIB drops the windows with high uncertainty context, alleviates Astute-RAG's uncertainty, and achieves an improvement of EM by 11.14%. Additionally, we observe a significant reduction in faithfulness by 6.45%, indicating that the LLM no longer copies the retrieved context totally. This is because the sliding-window strategy of Swin-VIB trades some cross-segment coherence for higher intra-segment consistency. On the other hand, the naive RAG and Self-RAG lack conflicting accommodation resolution, often copy the retrieved context, and fall behind on both EM and METEOR. The high EM scores of Swin-VIB show that it can adapt the LLM's preference on external contexts and select the most relevant external information, even when the top-5 retrieved contexts contain conflicting information. These results demonstrate that Swin-VIB can improve response generation quality and reliability for mitigating knowledge conflicts in RAG.

Table 4 evaluates the overhead of Swin-VIB in response generation delay. For Llama2-7B, Swin-VIB involves an additional delay of 0.3913 seconds. It can also be found that after the Swin-VIB has been integrated into the RAG systems, the increased number of model parameters does not significantly increase the latency. The Rag systems remain sufficiently lightweight, since Swin-VIB trades off efficiency and accuracy in response generation. (see Appendix C.4 for computational complexity analysis).

5 Related Work

Efforts to mitigate risks associated with knowledge conflicts in response generation can be categorized into three groups: a) Internal knowledge-driven methods (Shi et al. 2023b; Jin et al. 2024b; Zhang, Yu, and Feng 2024): Fine-tuning or editing the model so that the retrieved evidences can override

conflicts; b) External-validation methods incorporate verification module to ensure the reliability of retrieved contexts (Yu, Zhang, and Feng 2024; Kortukov et al. 2024; Bi et al. 2025); c) Adaptive methods at the decoding stage, take LLM parameters frozen and detect token-level conflicts through adjusting parametric and context logits (Shi et al. 2023a; Yuan et al. 2024; Jin et al. 2024a; Bi et al. 2025; Huang et al. 2025). However, these existing methods have limitations in elucidating LLMs' knowledge preference due to: a) Internal catastrophic forgetting of LLMs caused by fine-tuning (Xu et al. 2024), b) Overconfidence in external sources caused by external validation (Xie et al. 2023); c) Adaptive single-token adjustments struggling with long text analysis, thus limiting their practical applicability in real-world RAG deployments. The recent empirical works (Jin et al. 2024b; Zhao et al. 2024; Yuan et al. 2025) focus on heuristic interventions on response generation of LLMs.

In contrast, our theoretical study demonstrates that the information difference between external context and internal representations is the true force to drive preference shifts and information selection. According to this insight, we propose a novel response generation method (Li et al. 2024), Swin-VIB. Its bottleneck-based adapter is involved to regulate the compression ratio of the retrieved information, without retraining the pre-trained LLMs. In this way, Swin-VIB accommodates conflicting knowledge, avoiding merely relying on internal or external information. This information-theoretic insight furnishes a unified, model-agnostic yardstick that combines theoretical breadth with practical simplicity.

6 Conclusion

This work proposes a novel theoretical framework to analyze and address the issue of information conflicts encountered by LLMs in RAG systems. Leveraging analysis on the knowledge conflicts and preferences of LLMs from an information theory perspective, we find that the uncertainty of LLMs can be mitigated by adapting the external information difference of LLMs. Insight from this, we propose Swin-VIB to optimize how RAG handles external contexts to achieve reliable response generation. Experimental results demonstrate that Swin-VIB significantly accommodates conflicts, reduces uncertainty in LLM outputs, and generates more accurate, consistent, and context-aware responses. Moreover, Swin-VIB enhances retrieval system performance, facilitating its real-world applications. Future work will explore extending this approach to more types of response generation tasks to validate further and to improve its effectiveness.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (62272248 and 61962045), the Natural Science Foundation of Tianjin (25JJJJC0016 and 25JCZDSN00040), in part by the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region under Grant NJYT23104, and the Basic Scientific Research Expenses Program of Universities directly under Inner Mongolia Autonomous Region under Grant JY20220273 and Grant JY20240002.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Bi, B.; Liu, S.; Wang, Y.; Xu, Y.; Fang, J.; Mei, L.; and Cheng, X. 2025. Parameters vs. Context: Fine-Grained Control of Knowledge Reliance in Language Models. *arXiv preprint arXiv:2503.15888*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

DeepSeek-AI. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Ding, H.; Pang, L.; Wei, Z.; Shen, H.; and Cheng, X. 2024. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *arXiv preprint arXiv:2402.10612*.

Es, S.; James, J.; Espinosa-Anke, L.; and Schockaert, S. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2: 1.

Hagström, L.; Marjanović, S. V.; Yu, H.; Arora, A.; Lioma, C.; Maistro, M.; Atanasova, P.; and Augenstein, I. 2024. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. *arXiv preprint arXiv:2412.17031*.

Huang, P.; Liu, Z.; Yan, Y.; Yi, X.; Chen, H.; Liu, Z.; Sun, M.; Xiao, T.; Yu, G.; and Xiong, C. 2025. Pip-kag: Mitigating knowledge conflicts in knowledge-augmented generation via parametric pruning. *arXiv preprint arXiv:2502.15543*.

Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; Jiang, X.; Xu, J.; Li, Q.; and Zhao, J. 2024a. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409*.

Jin, Z.; Cao, P.; Yuan, H.; Chen, Y.; Xu, J.; Li, H.; Jiang, X.; Liu, K.; and Zhao, J. 2024b. Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models. *arXiv preprint arXiv:2402.18154*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kortukov, E.; Rubinstein, A.; Nguyen, E.; and Oh, S. J. 2024. Studying large language model behaviors under realistic knowledge conflicts. *arXiv e-prints*, arXiv–2404.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, X.; Jin, J.; Zhou, Y.; Zhang, Y.; Zhang, P.; Zhu, Y.; and Dou, Z. 2024. From matching to generation: A survey on generative information retrieval. *arXiv preprint arXiv:2404.14851*.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Robertson, S. E.; and Jones, K. S. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3): 129–146.

Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; and Yih, S. W.-t. 2023a. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Shi, W.; Min, S.; Lomeli, M.; Zhou, C.; Li, M.; Lin, V.; Smith, N. A.; Zettlemoyer, L.; Yih, S.; and Lewis, M. 2023b. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, F.; Wan, X.; Sun, R.; Chen, J.; and Arık, S. Ö. 2024. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint arXiv:2410.07176*.

Wu, K.; Wu, E.; and Zou, J. 2024. How faithful are RAG models? Quantifying the tug-of-war between RAG and LLMs' internal prior. *arXiv preprint arXiv:2404.10198*.

Xie, J.; Zhang, K.; Chen, J.; Lou, R.; and Su, Y. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Xu, R.; Qi, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024. Knowledge Conflicts for LLMs: A Survey. *arXiv preprint arXiv:2403.08319*.

Yu, T.; Zhang, S.; and Feng, Y. 2024. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. *arXiv preprint arXiv:2403.07556*.

Yuan, X.; Yang, Z.; Huang, Z.; Wang, Y.; Fan, S.; Ju, Y.; Zhao, J.; and Liu, K. 2025. Exploiting Contextual Knowledge in LLMs through V-usable Information based Layer Enhancement. *arXiv preprint arXiv:2504.15630*.

Yuan, X.; Yang, Z.; Wang, Y.; Liu, S.; Zhao, J.; and Liu, K. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. *arXiv preprint arXiv:2402.11893*.

Zhang, S.; Yu, T.; and Feng, Y. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

Zhao, Y.; Devoto, A.; Hong, G.; Du, X.; Gema, A. P.; Wang, H.; He, X.; Wong, K.-F.; and Minervini, P. 2024. Steering knowledge selection behaviours in LLMs via sae-based representation engineering. *arXiv preprint arXiv:2410.15999*.

Zhu, T.; Liu, Q.; Wang, F.; Tu, Z.; and Chen, M. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.