# Bridging External and Parametric Knowledge: Mitigating Hallucination of LLMs with Shared-Private Semantic Synergy in Dual-Stream Knowledge

**Yi Sui[1], Chaozhuo Li[2], Chen Zhang[3], Dawei Song[1,4] *, Qiuchi Li[1]**
[1]Beijing Institute of Technology, China
[2]Beijing University of Posts and Telecommunications, China
[3]Meituan, China [4]The Open University, UK
**Correspondence:** dawei.song2010@gmail.com

## Abstract

Retrieval-augmented generation (RAG) aims to mitigate the hallucination of Large Language Models (LLMs) by retrieving and incorporating relevant external knowledge into the generation process. However, the external knowledge may contain noise and conflict with the parametric knowledge of LLMs, leading to degraded performance. Current LLMs lack inherent mechanisms for resolving such conflicts. To fill this gap, we propose a Dual-Stream Knowledge-Augmented Framework for Shared-Private Semantic Synergy (DSSP-RAG). Central to it is the refinement of the traditional self-attention into a mixed-attention that distinguishes shared and private semantics for a controlled knowledge integration. An unsupervised hallucination detection method that captures the LLMs' intrinsic cognitive uncertainty ensures that external knowledge is introduced only when necessary. To reduce noise in external knowledge, an Energy Quotient (EQ), defined by attention difference matrices between task-aligned and task-misaligned layers, is proposed. Extensive experiments show that DSSP-RAG achieves a superior performance over strong baselines.

## 1 Introduction

Large Language Models (LLMs) excel in natural language processing (NLP) tasks but face challenges in adapting to the rapid evolution of knowledge (Meng et al., 2022; Fan et al., 2025). The discrepancy between static pre-trained knowledge and continuously evolving external information leads to outdated, inconsistent, and insufficient knowledge within LLMs, contributing to hallucination and ultimately limiting model performance (Huang et al., 2024; Tonmoy et al., 2024).

Retrieval-Augmented Generation (RAG) offers a cost-effective solution by incorporating retrieved external knowledge into the generation process of LLMs to alleviate hallucinations (Ram et al., 2023; Qian et al., 2024; Yu et al., 2025; Li et al., 2025). However, as shown in Figure 1, the classical RAG methods face two core challenges. Firstly, there is a lack of an effective mechanism for real-time hallucination detection, resulting in blind injection of external knowledge into LLM (Du et al., 2022; Pan et al., 2023), but neglecting the need for proper integration of knowledge. Studies have shown that LLMs often exhibit a tendency to over-rely on external evidence (Wu et al., 2024), in which noise information may exist and inadvertently degrade the overall performance (inter-context conflict). Secondly, the external knowledge may conflict with the parametric knowledge acquired during LLM pre-training (context-memory conflict) (Xu et al., 2024). The lack of dedicated mechanisms to resolve such knowledge conflicts, coupled with the implicit knowledge encoding in current LLMs, would affect generation stability.

To ensure more effective RAG, hallucination detection and knowledge filtering are essential for identifying and mitigating irrelevant or misleading information prior to knowledge integration. Existing approaches of hallucination detection (Xiong et al., 2024; Mahaut et al., 2024) and knowledge filtering (Yu et al., 2023; Wang et al., 2024b) often rely on supervised training that inherently limits the generalization to out-of-domain scenarios. Moreover, some denoising strategies use the confidence scores of the original LLM, rendering them susceptible to the model's internal biases (Li et al., 2022; Zhou et al., 2024). Therefore, it is necessary to incorporate accurate and unsupervised detection modules.

More recent approaches to knowledge augmentation employ knowledge editing (Zhang et al., 2024a; Hase et al., 2024; Wang et al., 2025b) or iterative reasoning (Wang et al., 2024a), but remain constrained by delayed integration and the absence of explicit semantic disentanglement (Ju

*Corresponding author.



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10846–10870
November 4-9, 2025 

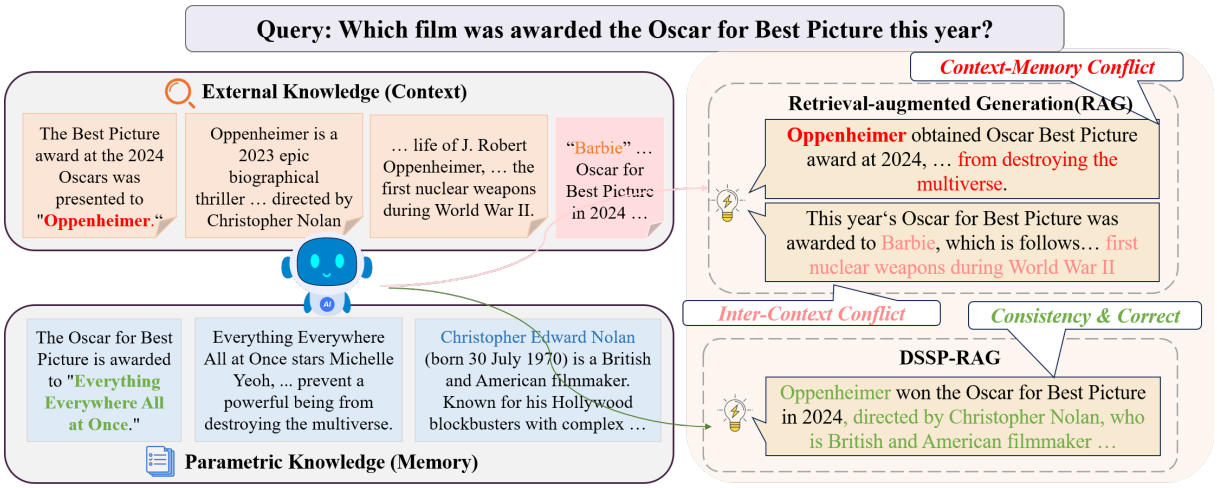



Figure 1: Comparison of Inference Results Using External vs. Parameterized Knowledge Across Methods

et al., 2024). The core limitation of current LLM architectures lies in their reliance on implicit fusion of internal and external knowledge, which not only lacks interpretability but also incurs substantial computational overhead due to extended context lengths (Jin et al., 2024) and quadratic cost of the self-attention mechanism (Yang et al., 2024).

To address the aforementioned challenges, we propose a Dual-Stream Knowledge-Augmented Framework for Shared-Private Semantic Synergy (DSSP-RAG). To tackle the inter-context conflicts, DSSP-RAG preemptively incorporates a two-stage control mechanism into RAG, consisting of an unsupervised hallucination detection module to determine when the external knowledge is necessary, and a knowledge filtering module to maintain the accuracy of the external knowledge by filtering the noisy or redundant information. The former is grounded in intrinsic cognitive uncertainty of LLMs and detects hallucination by analyzing subspace variations in response to semantically equivalent prompts. For the latter, we identify key and offset layer that differentially capture the task-aligned versus misaligned information, and propose an Energy Quotient (EQ) that is derived from the attention difference matrices between the pair of the key and offset layer to amplify key knowledge fragments while suppressing noise.

To mitigate the context-memory conflicts and optimize knowledge augmentation by maximizing the synergy between the model's intrinsic capabilities and external resources, we propose a mixed attention mechanism to decompose dual-stream knowledge into shared and private semantics in a pluggable way. Shared semantics captures the consistency between internal-external knowledge, aiming to improve the reliability of generation. On the other hand, private semantics preserves their unique contributions that are dynamically adjusted through a weighted fusion layer based on contextual relevance, thus mitigating the conflicts and improving generation accuracy.

Our contributions are summarized as follows: (1) We propose an integrated framework that combines unsupervised hallucination detection, knowledge filtering and augmentation to precisely regulate the integration of external knowledge and mitigate hallucinations. (2) DSSP-RAG employs a mixed attention mechanism to resolve context-memory conflicts and enhance dual-stream knowledge utilization, improving generation reliability and accuracy. (3) Extensive experiments over four benchmarks show that our approach consistently outperforms the state-of-the-art baselines.

## 2 Related Work

**Hallucination Detection.** LLMs acquire knowledge during pretraining, but the limitations of static corpora may result in factual inaccuracies, resulting in outputs based on unreliable or fabricated information that undermine response credibility (Li et al., 2023; Wang et al., 2023a; He et al., 2024). Researchers have proposed various techniques to detect hallucination. Current approaches primarily include training multi-layered probes to extract factual confidence scores from hidden states (Azaria and Mitchell, 2023; Kadavath et al., 2022; Burns et al., 2023), leveraging the averaged log probabilities assigned to a sequence of output tokens to estimate factual confidence (Xiong et al., 2024; Yin et al., 2023; Lin et al., 2022), prompting the model to generate responses accompanied by con-

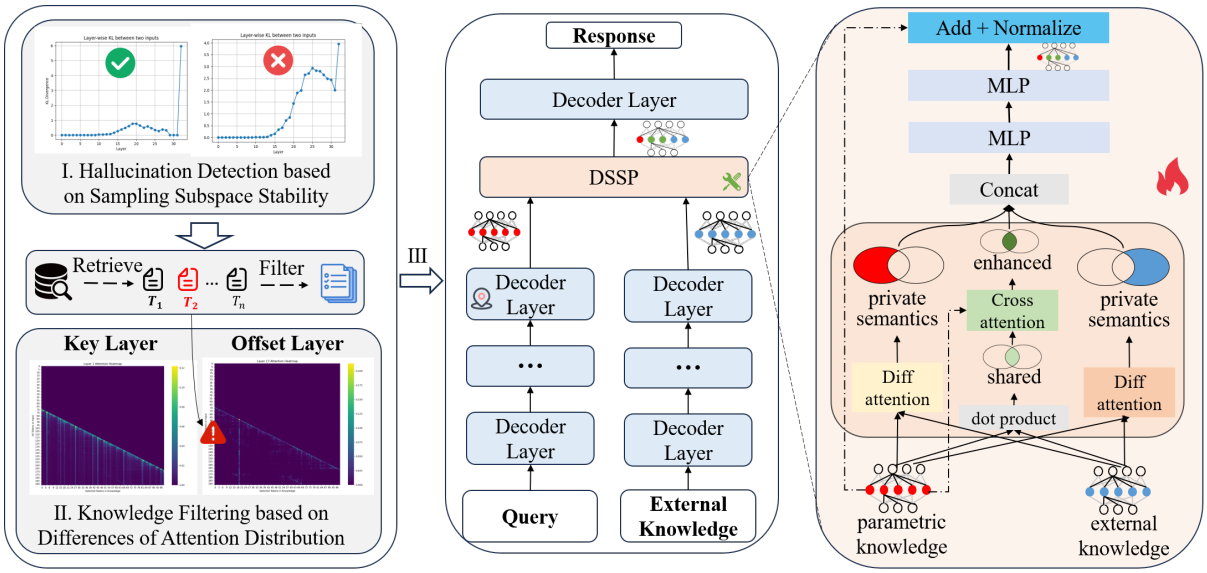



Figure 2: Dual-Stream Knowledge-Augmented Framework for Shared-Private Semantic Synergy (DSSP-RAG)

fidence levels (Verbalization) (Tian et al., 2023), using prompts to guide the model in outputting specific tokens to assess factual accuracy, and evaluating consistency based on multiple sampling results (Wang et al., 2023b; Manakul et al., 2023; Ding et al., 2024). Recent studies indicate that training probes based on the model's internal states represents one of the most reliable methods for hallucination detection (Mahaut et al., 2024). Hallucination detection serves as the foundation for mitigating hallucination issues, and our proposed method stands out by eliminating the need for supervised training, ensuring greater generalizability.

**RAG for Hallucination.** A straightforward approach to mitigate hallucination is to employ RAG methods to retrieve accurate and comprehensive external facts as a complementary resource to guide the generation process of LLMs (Huang et al., 2023; Gao et al., 2025; Wang et al., 2025a). However, in the face of complex or multi-hop reasoning tasks, the knowledge obtained through single-round retrieval often proves insufficient to provide adequate information for subsequent inference steps. Thus, ITER-RETGEN uses an iterative approach to coordinate retrieval and generation, leveraging LLM responses at each iteration to retrieve more relevant knowledge, thereby improving inference accuracy in subsequent rounds (Shao et al., 2023; Lee et al., 2025). Moreover, Adaptive-RAG dynamically adjusts retrieval strategies based on task complexity, allowing flexible switching between single retrieval, iterative retrieval and no retrieval, to meet various task demands (Jeong et al., 2024; Guan et al., 2025). Integrating external knowledge through in-context learning can conflict with the model's internal knowledge and introduce noise, degrading LLM performance. We address this by filtering noise, resolving conflicts, and enabling collaborative utilization of dual-stream knowledge to mitigate hallucinations.

## 3 Methodology

### 3.1 Overview

Figure 2 presents an overview of DSSP-RAG, which comprises three core modules designed to integrate external knowledge with internal representations while mitigating hallucinations. First, DSSP-RAG identifies hallucination-prone instances through subspace stability analysis, enabling adaptive retrieval timing. Second, an attention difference matrix is constructed to filter noise from retrieved knowledge, ensuring the relevance of the incorporated information. Finally, the DSSP module proposes a mixed attention mechanism to disentangle external and internal knowledge into shared and private semantics, resolving potential conflicts and enhancing the reliability of the output.

### 3.2 Hallucination Detection Based on Stability of Sampling Subspaces

Hallucination detection provides a guiding signal for the necessity and timing of integrating external knowledge. To overcome domain-specific limitations and the reliance on labeled data in previous work (Azaria and Mitchell, 2023; Kuhn et al., 2023), we propose an unsupervised detection method based on the so-called cognitive uncertainty (Yadkori et al., 2024), which manifests

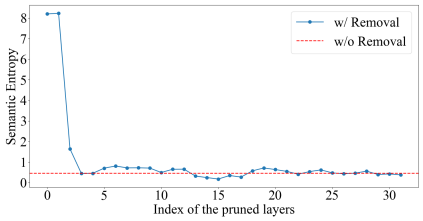



Figure 3: Performance variation with layer-wise pruning

as inconsistent outputs in response to semantically equivalent but syntactically varied prompts. However, the outputs can be influenced by input noise and model biases. LLMs reason within high-dimensional sampling (probabilistic) subspaces shaped by parametric knowledge from pre-training. Thus, we employ the stability of the model's sampling subspaces (Ji et al., 2024; Zhang et al., 2024b) as a metric to detect hallucinations, enabling an improved generalization across tasks and a more reliable assessment of the reasoning process.

To this end, we prompt the LLM to automatically generate syntactically distinct but semantically similar alternative expressions for each query. To assess the stability of sampling subspaces, we employ Jensen-Shannon Divergence (JSD) to quantify the differences between the subspaces of paired queries $(x, \hat{x})$ across model layers.

$$d(s_l(\cdot|x), \hat{s}_l(\cdot|\hat{x})) = \text{JSD}(s_l(\cdot|x)||\hat{s}_l(\cdot|\hat{x})). \quad (1)$$

Here, $s_l$ denotes the sampling subspace activated by the LLM in layer $l$ for a given input. The JSD distance $d$ below a predefined threshold $\delta$ suggests an effective internalization of correct knowledge from the pretraining corpus of the LLM. Cognitive uncertainty may cause a divergence in response, leading to an increase in JSD in deeper layers, as illustrated in the top-left subfigure of Figure 2. A $d$ exceeding the threshold would indicate an instability in the sampling subspace, leading to an increased risk of hallucination. JSD is preferred over alternatives like KL-divergence due to its symmetric and bounded properties, see Appendix C for a detailed analysis.

### 3.3 Filtering of External Knowledge

The core of the RAG is to enhance the LLM performance by retrieving external factual knowledge to support generation process. Current knowledge filtering approaches exhibit poor domain generalization and heightened sensitivity to model-induced biases (Fan et al., 2024). Therefore, inspired by the work of Zhang et al. (2023), we perform a layer-wise pruning study to assess the influence of external knowledge on LLM's decision-making.

By systematically removing each layer of the LLM (denoted as w/ Removal) and comparing the performance differences with the original model (denoted as w/o Removal), the variations are quantitatively evaluated using semantic entropy, which measures the uncertainty of the model's output:

$$H(Y|Q) = -\sum_{y} p(y|Q) \log p(y|Q). \quad (2)$$

Here, $p(y|Q)$ denotes the model's predicted probability for a given output token $y$, conditioned on input query $Q$. The higher entropy value indicates a greater uncertainty in the model prediction.

The results (Figure 3) reveal performance fluctuations caused by layer ablation. Specifically, removing certain layers leads to an increased semantic entropy, underscoring their role in encoding task-relevant information. These layers often exhibit broadly distributed attention to contextual semantics. In contrast, eliminating other layers reduces entropy, suggesting that excessively narrow attention may encode noise or irrelevant features. Thus, we propose a classification of layers:

- Key Layer whose removal causes significant performance degradation, indicating its role in extracting task-relevant semantics and constructing global representations aligned with inference objectives;
- Offset Layer whose removal leads to the greatest performance improvement, suggesting it amplifies noise or irrelevant features due to attention misalignment with task goals.

Furthermore, from an information-theoretic perspective, we construct the Energy Quotient (EQ) as a filtering matrix by analyzing the difference in attention distribution, $\Delta A = A_\alpha - A_\beta$, between an offset layer ($\alpha$) and a key layer ($\beta$), thereby quantifying the significance of external context:

$$EQ_i = \frac{\exp(-\lambda \Delta A_i)}{\sum_j \exp(-\lambda \Delta A_j)}. \quad (3)$$

Here, $\lambda$ denotes the temperature coefficient. The higher $\Delta A_i$, the poorer the stability of the representation space. Thus, a smaller $EQ_i$ indicates that the feature $i$ increases uncertainty in the reasoning process and is more likely to be identified as noise. In contrast, a larger $EQ_i$ signifies a higher informational relevance and a greater contribution to the reasoning process.

Considering that the degree of noise varies in different tasks, we further introduce a dynamic weighting coefficient $\varepsilon$, formulated as follows, to adjust the filtering effect of EQ on the external context.

$$\varepsilon = \begin{cases} \log\left(1 - \frac{\Delta\hbar}{\hbar_{\text{orig}}}\right), & \Delta\hbar < -0.1 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Here, $\triangle\hbar = \hbar_\alpha - \hbar_{orig}$, where $\hbar_{orig}$ represents the semantic entropy of the original model response with external context, and $\hbar_\alpha$ corresponds to the semantic entropy after removing the offset layers. A negative and larger absolute value of $\triangle\hbar$ indicates that the information attended to by the offset layers has a significant negative impact, representing increased noise. Conversely, if the difference is close to zero or positive, it indicates low noise in external information, and knowledge filtering is not required. Finally, the encoded external information $\hat{D}$ is updated based on EQ-based filtering and coefficient weighting, as follows:

$$\hat{D} = \begin{cases} \varepsilon \cdot EQ \cdot D, & \Delta\hbar < -0.1 \\ D, & \text{otherwise} \end{cases} \quad (5)$$

### 3.4 Dual-Stream Knowledge Augmentation

A primary limitation for LLMs in resolving context-memory conflicts stems from the absence of explicit semantic decomposition in the conventional attention, hindering the distinction between private (source-specific) and shared (overlapping) semantics (Xie et al., 2024; Li et al., 2024; Dong et al., 2025). The inherent uniformity bias in attention weights further exacerbates this issue by amplifying the redundancy in shared semantics while neglecting the conflicts in private semantics, leading to ambiguous output. Furthermore, the extended context length introduced by external knowledge increases computational overhead and diminishes the model's focus on salient information (Jin et al., 2024; Liu et al., 2024). In response to these challenges, a mixed attention of the DSSP module is proposed to decompose dual-stream knowledge into shared and private semantics at the target layer, identified via the maximal JSD between semantically identical prompts. DSSP dynamically balances the integration of parametric knowledge and external sources while mitigating the context-memory conflicts. A mathematical derivation of the mixed attention is detailed in Appendix E.

Shared semantics refers to the overlapping portion of knowledge between internal knowledge $I$ and external knowledge $\hat{D}$, which exhibits high consistency between different sources of knowledge. To capture this shared information, we compute the similarity matrix between dual-stream knowledge using the dot product similarity:

$$sim(I, \hat{D}) = \varphi\left(\frac{(W_{share}I)(W_{share}\hat{D})}{\sqrt{d_k}}\right), \quad (6)$$

where $\varphi$ denotes the softmax function. Based on the similarity matrix, the top-$T$ tokens from external knowledge most similar to internal knowledge are selected as shared semantics, denoted as $U_{\text{share}}$. Then a cross-attention mechanism is applied to derive $U_{\text{enhance}}$, which integrates $U_{\text{share}}$ into the internal knowledge stream, amplifying the consistency between external and internal representations.

Private semantics refers to the unique information within internal $I$ or external knowledge $\hat{D}$, which helps fill the gaps in the other source. This provides additional useful information for the model's predictive behavior. To extract these private features, we introduce a differential attention mechanism that subtracts the shared components from each knowledge stream. Self-attention within each source models token-level dependencies and contextual interactions.

$$\tau(X, X) = \varphi\left(\frac{(W_Q^s X)(W_K^s X)}{\sqrt{d_k}}\right)(W_V^s X). \quad (7)$$

Meanwhile, the cross-attention models the mutual influence and semantic relationships between the dual knowledge, capturing the shared information:

$$\eta(X, Y) = \varphi\left(\frac{(W_Q^c X)(W_K^c Y)}{\sqrt{d_k}}\right)(W_V^c Y). \quad (8)$$

Thus, we define the differential attention mechanism as the difference between self-attention $\tau(\cdot)$ of single knowledge source and cross-attention $\eta(\cdot)$ of dual-stream knowledge:

$$U_{\text{private}} = \mathcal{D}_{\text{attn}}(X, Y) = \tau(X, X) - \eta(X, Y). \quad (9)$$

When $X$ and $Y$ represent internal knowledge and external knowledge respectively, the private semantics within the internal knowledge can be extracted through differential attention, and similarly, the private semantics within the external knowledge can be obtained. The explicit decoupling mechanism enables DSSP-RAG to adaptively reweight knowledge sources according to task demands:

$$U = \text{concat}(U_{enhance}, U^I_{private}, U^{\hat{D}}_{private}), \quad (10)$$

$$\hat{U} = \text{LN}(W_o\,(\text{ReLU}\,(W_f U + b_f)) + b_o). \quad (11)$$

To ensure the stability of model training, we apply a residual connection to combine the original representation $I$ and the aggregated knowledge representation $\hat{U}$. This combined representation is then passed into the subsequent layers of the model encoding and further-depth knowledge fusion.

### 3.5 Loss Function

LLMs are typically trained with cross-entropy loss, but their inherent confidence bias may cause overconfidence in incorrect predictions or undue uncertainty in correct ones. Moreover, models without knowledge adaptation struggle to effectively attribute the influence of external knowledge , resulting in abrupt probability distribution shifts that may undermine the predictive accuracy.

To address this issue, we introduce a conditional entropy $H(P(\hat{U}|I)) = -\sum_{\hat{y}} P(\hat{y}|I) log P(\hat{y}|\hat{U})$ to regulate the predictive uncertainty, allowing the model to dynamically adjust confidence based on the reliability of external knowledge. Accurate external knowledge reduces uncertainty, while noisy knowledge increases it, providing a basis for filtering external information. However, relying solely on conditional entropy may compromise training stability, as it lacks constraints on distribution shifts across different inputs, potentially leading to overfitting or bias toward erroneous knowledge. To enhance stability, we incorporate the KL divergence $D_{KL}(P(\hat{y}|\hat{U})||P(\hat{y}|I))$ to regularize the distributional shift of external knowledge, preventing excessive deviations. This regularization ensures that the model maintains prediction stability while avoiding overreliance on noisy information. The final loss function is then formulated as:

$$L = L_{CE} + \mu H(P(\hat{U}|I)) + \nu D_{KL}, \quad (12)$$

where $L_{CE}$ represents the cross-entropy loss, $\mu$ and $\nu$ respectively controls the regularization strength of conditional entropy and KL divergence.

## 4 Experiments

### 4.1 Tasks and datasets

DSSP-RAG is evaluated on five datasets, including two single-hop QA datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019) which is based on real user queries, and TriviaQA (Joshi et al., 2017) that consists of trivia questions; and two multi-hop QA datasets: Adversarial HotpotQA (Yang et al., 2018), and 2WikiMultihopQA (Ho et al., 2020), which assess cross-document reasoning and information integration using Wikipedia data. In addition, PubHealth (Kotonya and Toni, 2020) is a closed-set generative dataset that comprises medical claims on various biomedical topics. More details are included in the Appendix A.

### 4.2 Experimental setups and baselines

All experiments are conducted on NVIDIA RTX A6000 GPU with 49 GB of memory. The training data consists of question-answer pairs from HotpotQA and 2WikiMultihopQA, leveraging the supporting and non-supporting facts provided within the datasets. The effectiveness of DSSP-RAG is validated in different model architectures, including **LLaMA2-7B-Chat** (Touvron et al., 2023), **LLaMA3-8B-Instruct** (Meta, 2024) and **Qwen2.5-3B-Instruct** (Qwen et al., 2025). In our main experiment, we employ the BM25 (Robertson et al., 2009) as retriever, utilizing the English version of Wikipedia dumped on December 20, 2018 as the retrieval corpus and retrieving five documents.

To evaluate the effectiveness of DSSP-RAG, we compare it with two representative baselines of RAG. (1) **Simple RAG**(SR-RAG): which directly prepends the external knowledge obtained from a single retrieval to the LLM's prompt. (2)**Adaptive RAG**: an active RAG framework that decides when and what to retrieve during generation, including **Self-RAG** (Asai et al., 2023), **FLARE** (Jiang et al., 2023), **DRAGIN** (Su et al., 2024), **SEAKR** (Yao et al., 2024). Following SEAKR, we re-implement FLARE with IRCoT strategy (Trivedi et al., 2022) to support evaluation on complex QA. IRCoT interweaves CoT reasoning with retrieval-augmented generation strategy. Meanwhile, we implement DSSP-RAG with simple and adaptive retrieval, yielding two variants: **DSSP_SR** and **DSSP_AR**.

### 4.3 Results and analysis

#### 4.3.1 Main results

The experimental results in Table 1 demonstrate that DSSP-RAG consistently achieves optimal performance in different retrieval strategies over baseline methods. For complex queries, the static and broad nature of simple retrieval strategies often fails to provide precise matches, thus impairing model performance. However, DSSP-RAG effectively leverages both internal and external knowledge while mitigating noise interference, leading to significant performance gains.

Although adaptive retrieval expands the knowledge scope, it also introduces more noise and redundancy. DSSP-RAG mitigates this by integrat-

| Model | LLaMA2-7B-Chat | | | | | LLaMA3-8B-Instruct | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **NQ** | **TriviaQA** | **HotpotQA** | **2Wiki** | **PubH** | **NQ** | **TriviaQA** | **HotpotQA** | **2Wiki** | **PubH** |
| | **EM** | **EM** | **F1** | **F1** | **ACC** | **EM** | **EM** | **F1** | **F1** | **ACC** |
| *Simple RAG* | | | | | | | | | | |
| No Retrival | 13.8 | 30.5 | 27.5 | 22.3 | 33.4 | 22.6 | 55.7 | 28.4 | 33.9 | 50.2 |
| SR-RAG | 20.7 | 42.5 | 25.0 | 25.5 | 30.7 | 30.1 | 58.3 | 35.8 | 29.6 | 51.3 |
| **DSSP_SR** | 19.5 | 45.3 | 29.3 | 26.7 | 34.2 | 28.6 | 58.9 | 37.6 | 36.8 | 52.7 |
| *Adaptive RAG* | | | | | | | | | | |
| Self-RAG | **32.3** | 57.0 | 17.5 | 19.6 | - | **36.4** | 38.2 | 29.6 | 25.1 | - |
| FLARE | 25.3 | 50.7 | 22.1 | 24.3 | 37.1 | 22.5 | 55.8 | 28.8 | 33.9 | 44.5 |
| DRAGIN | 23.2 | 54.0 | 29.2 | 30.0 | 52.4 | - | - | 44.6 | 37.8 | 66.6 |
| SEAKR | 25.6 | 54.1 | **38.1** | 36.0 | 59.3 | 31.0 | 59.4 | 47.7 | 48.1 | 70.2 |
| **DSSP_AR** | 25.1 | **55.4**$^{\dagger}$ | 37.7 | **37.8**$^{\dagger}$ | **62.1**$^{\dagger}$ | 30.8 | **60.2**$^{\dagger}$ | **48.6**$^{\dagger}$ | **49.8**$^{\dagger}$ | **73.5**$^{\dagger}$ |
| **Δ (%)** | 1.95$^{\downarrow}$ | **2.40**$^{\uparrow}$ | 1.05$^{\downarrow}$ | **5.00**$^{\uparrow}$ | **4.72**$^{\uparrow}$ | 0.65$^{\downarrow}$ | **1.35**$^{\uparrow}$ | **1.89**$^{\uparrow}$ | **3.54**$^{\uparrow}$ | **4.55**$^{\uparrow}$ |

Table 1: Performance comparison across methods and datasets on Llama family. † refers to a statistically significant improvement achieved by DSSP-RAG over SEAKR (the best performing baseline) according to paired t-test ($p < 0.05$). Self-RAG is fine-tuned from LLaMA2-7B-chat using NQ-style data.

| **Method** | **NQ** | **TriviaQA** | **HotpotQA** | **2Wiki** |
|---|---|---|---|---|
| | **EM** | **EM** | **F1** | **F1** |
| *Simple RAG* | | | | |
| No Retrival | 29.3 | 57.6 | 28.7 | 30.2 |
| SR-RAG | 38.5 | 60.8 | 37.1 | 27.4 |
| DSSP-SR | 39.7 | 61.3 | 40.1 | 32.3 |
| *Adaptive RAG* | | | | |
| FLARE | 30.1 | 58.8 | 27.5 | 31.7 |
| DRAGIN | - | - | 42.6 | 38.7 |
| SEAKR | 42.1 | 63.7 | 48.1 | 40.1 |
| **DSSP_AR** | **43.6** | **65.3** | **48.4** | **44.2** |
| **Δ (%)** | **3.56**$^{\uparrow}$ | **2.51**$^{\uparrow}$ | **0.62**$^{\uparrow}$ | **10.22**$^{\uparrow}$ |

Table 2: Performance comparison across different methods and datasets on Qwen2.5-3B-Instruct.

ing a hallucination detection module to assess the necessity of external knowledge and leveraging attention difference matrices to filter irrelevant content. In contrast, baseline methods, except SEAKR, lack such mechanisms, resulting in interference with internal knowledge by noise and reduced output reliability. DSSP-RAG slightly underperforms SEAKR on HotpotQA with LLaMA2-7B-Chat, as SEAKR's self-aware signals better capture uncertainties in critical intermediate steps in multi-hop reasoning. SEAKR utilizes LLM-based uncertainty to guide retrieval but fails to resolve conflicts between external and internal knowledge. In contrast, DSSP-RAG's mixed attention mechanism exploits shared semantics for reliable information activation and adaptively balances private semantics to mitigate conflicts and external interference. This adaptability ensures stable performance of DSSP_AR in cross-domain tasks, whereas baseline models suffer from fluctuations due to domain shifts.

DSSP-RAG consistently outperforms baselines across tasks, with improvements varying by model scale and dataset characteristics. Gains are most significant in multi-hop QA, which requires extensive knowledge integration, while single-hop QA tasks like NQ show limited improvement. As illustrated in Table 2, DSSP-RAG in Qwen2.5-3B-Instruct achieves higher performance gains than baselines compared to LLaMA3-8B-Instruct. These results across architectures and scales demonstrate the strong robustness of DSSP-RAG. In addition, we further evaluate the generalizability of DSSP-RAG on the ASQA (long-form QA) dataset, see Appendix B for detailed results.

#### 4.3.2 Evaluation of DSSP under Knowledge Conflict Scenarios

To evaluate DSSP-RAG's effectiveness in resolving knowledge conflicts, we compare it with two recent state-of-the-art baselines: EditCoT (Wang et al., 2024a) and CK-PLUG (Bi et al., 2025). For a fair comparison, hallucination detection and knowledge filtering are removed from DSSP-RAG, leaving only the DSSP module that modulates reliance between parametric and retrieved knowledge. As shown in Table 4, DSSP consistently outperforms both baselines across multiple backbone models.

EditCoT updates knowledge at the chain-of-thought level, constraining activation of parametric knowledge, while CK-PLUG regulates reliance on parametric versus retrieved knowledge through entropy-based confidence gain, which is prone to output bias. In contrast, DSSP integrates knowledge at the representation level via mixed-attention, enabling fine-grained coordination between inter-

nal and external knowledge. These results highlight DSSP's superior robustness and adaptability in resolving knowledge conflicts.

| Method | Method | TriviaQA | HotpotQA |
|---|---|---|---|
| | | EM | F1 |
| **LLaMA3 -8B- Instruct** | EditCoT | 54.7 | 43.1 |
| | CK-PLUG | 57.5 | 45.8 |
| | w/ DSSP | **58.3** | **46.4** |
| **Qwen2.5 -3B- Instruct** | EditCoT | 59.2 | 44.6 |
| | CK-PLUG | 61.8 | 46.1 |
| | w/ DSSP | **63.2** | **47.1** |

Table 3: Performance comparison of knowledge conflict mitigation technologies.

To evaluate the effectiveness of DSSP module in balancing internal and external knowledge under conflicts, we follow CK-PLUG by constructing a synthetic RAG setup with two deliberately incorrect statements in the retrieved context, enabling stress-testing of LLM conflict resolution. Table 4 reports three metrics: ConR (alignment with the retrieved context), ParR (alignment with model parameters), and EM.

| Model | Method | ConR | ParR | EM |
|---|---|---|---|---|
| **LLaMA2 -7B- Chat** | w/o RAG | - | - | 27.5 |
| | w/o DSSP | 65.7 | 30.1 | 23.6 |
| | w/ DSSP | 51.6 | 32.3 | 25.9 |
| **LLaMA3 -8B- Instruct** | w/o RAG | - | - | 28.4 |
| | w/o DSSP | 56.8 | 25.4 | 29.7 |
| | w/ DSSP | 44.7 | 31.6 | 31.5 |
| **Qwen2.5 -3B- Instruct** | w/o RAG | - | - | 28.7 |
| | w/o DSSP | 46.1 | 21.7 | 32.1 |
| | w/ DSSP | 37.4 | 28.3 | 33.6 |

Table 4: Performance (%) of DSSP Module in controlling knowledge reliance.

Across all model configurations, replacing standard self-attention with our mixed-attention mechanism consistently reduces reliance on misleading context (lower ConR) while strengthening alignment with parametric knowledge (higher ParR). This shift is crucial when external knowledge introduces noise or contradictions. Furthermore, DSSP consistently improves EM over baselines, underscoring the effectiveness of mixed-attention in synergistically leveraging internal and external knowledge for enhanced reasoning and conflict resolution. See Appendix G for Representative examples.

#### 4.3.3 Ablation study

We conduct an ablation study on Llama-7b-chat to assess the components of DSSP-RAG (Table 5). The DSSP module is critical, as its removal (w/o DSSP) causes a substantial drop in EM and F1 across datasets. By converting self-attention into mixed attention, DSSP enables effective integration of external knowledge while mitigating conflicts and improving the reliability of the output.

Knowledge filtering (K.F.) has a greater impact than hallucination detection (H.D.). Removing K.F. (w/o K.F.) leads to a more significant decline, especially in F1, underscoring its role in removing irrelevant or noisy knowledge and emphasizing the importance of external knowledge accuracy for response quality. Although removing hallucination detection (w/o H.D.) results in a smaller performance decrease, it still noticeably degrades generation quality, indicating that unregulated incorporation of external knowledge can introduce inconsistencies or biases. The H.D. module mitigates this by identifying hallucination-prone cases.

| Method | HotpotQA | | 2Wiki | | TriviaQA |
|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM |
| DSSP-RAG | 28.5 | 37.7 | 31.6 | 37.8 | 55.4 |
| *Ablating hallucination detection module* | | | | | |
| Trained Probes | 28.6 | 37.1 | 32.1 | 37.7 | 54.5 |
| Verbalization | 27.1 | 36.6 | 29.8 | 36.5 | 52.7 |
| Hyb-Cons | 27.8 | 36.8 | 31.4 | 37.1 | 54.3 |
| *Ablating knowledge filtering module* | | | | | |
| Entropy-based | 26.7 | 36.9 | 29.1 | 35.6 | 53.3 |
| Attention-based | 27.4 | 37.1 | 30.7 | 36.5 | 54.1 |
| w/o H.D. | 27.6 | 36.8 | 30.1 | 36.0 | 52.9 |
| w/o K.F. | 27.0 | 37.0 | 29.5 | 35.6 | 53.1 |
| w/o DSSP | 25.3 | 34.8 | 26.9 | 33.7 | 51.7 |

Table 5: Ablation study.

Furthermore, we replace the hallucination detection module in DSSP-RAG with two existing methods: Trained Probes, Verbalization and Hybrid Consistency (Hyb-Cons). Verbalization suffers from LLMs' overconfidence (Ni et al., 2024), underperforming even the baseline w/o H.D., whereas our unsupervised detection based on sampling subspace stability matches the supervised Trained Probes in effectiveness. The Hyb-Cons (Ding et al., 2024) integrates cross-language and cross-model scores to assess the consistency of perturbed responses for identical queries. DSSP-RAG framework mitigates hallucination more robustly than consistency-based baselines by analyzing sampling subspace stability rather than surface-level output. Its lightweight, unsupervised design integrates efficiently into RAG pipelines without auxiliary models or complex decoding, demonstrating both effectiveness and generalizability.

To better assess the effectiveness of the EQ-based knowledge filtering module, we replaced

the module in DSSP-RAG with two representative approaches: Entropy-based Filtering (Qiu et al., 2025), which assigns preference weights to documents based on token distribution entropy, and Attention-based Filtering (Peng et al., 2025), which scores relevance via cross-layer attention with a fixed bias. The results demonstrate that EQ-based method achieves superior factual accuracy and more robust performance across datasets. DSSP-RAG departs from raw attention-based importance measures by adopting an information-theoretic view, modeling cross-layer attention variation to capture semantic stability. A temperature-controlled Energy Quotient enhances sensitivity to representational instability, while an entropy-aware adaptive filter dynamically adjusts external knowledge contributions, thereby reducing irrelevant interference and improving robustness.

#### 4.3.4 Hyperparameter and efficiency analysis

We evaluate DSSP-RAG and two baselines on 2WikiMultihopQA using LLaMA-7B-Chat with varying numbers of retrieved documents (Table 6). SR-RAG initially benefits from knowledge augmentation, but deteriorates with more than five documents due to increased noise. SEAKR, despite incorporating filtering and ranking, struggles with relevance in extended contexts. In contrast, DSSP-RAG consistently outperforms baselines by leveraging mixed attention to disentangle shared and private semantics, enabling effective knowledge integration and mitigation of hallucinations.

Furthermore, inference efficiency is measured with each method retrieving five external documents per query. The results show that DSSP-RAG effectively balances inference efficiency and resource consumption. Due to the design of DSSP module, its autoregressive decoding time is 0.72s substantially faster than SR-RAG (1.6s) and SEAKR (1.25s), which encodes internal and external knowledge in parallel and delays fusion via hybrid attention until the target layer. Despite introducing moderate overhead, DSSP-RAG maintains a low peak memory usage (17.34), outperforming SEAKR (24.03) and remaining comparable to SR-RAG (16.08). SEAKR's higher resource demand stems from its reliance on repeated LLM sampling.

As illustrated in Figure 4, we validate the effectiveness of DSSP-RAG's dynamic plug-in strategy, which selects the DSSP insertion layer based on the maximal distance between sampling subspaces of semantically equivalent but syntactically

| Method | 3 | 5 | 10 | 20 | Inf. | Mem. |
|---|---|---|---|---|---|---|
| SR-RAG | 15.5<br>% | 16.9<br>**9.0**$^{\uparrow}$ | 8.3<br>50.9$^{\downarrow}$ | 5.5<br>33.7$^{\downarrow}$ | 1.6 | 16.08 |
| SEAKR | 30.2<br>% | 30.5<br>**1.0**$^{\uparrow}$ | 25.7<br>15.7$^{\downarrow}$ | 21.4<br>16.7$^{\downarrow}$ | 1.25 | 24.03 |
| **DSSP-RAG** | 28.4<br>% | 31.6<br>**11.3**$^{\uparrow}$ | 32.1<br>**1.6**$^{\uparrow}$ | 31.3<br>2.5$^{\downarrow}$ | 0.72 | 17.34 |

Table 6: Analysis of inference efficiency and document count of knowledge-enhanced methods. Inf. (s) measures the time for autoregressive decoding only; Mem. (GB) indicates peak GPU memory usage.

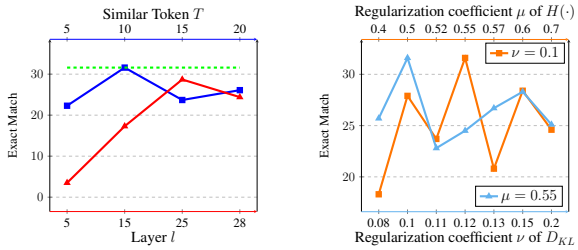



Figure 4: The analysis of hyperparameters.

different prompts. Compared with fixed plug-in positions at early, middle, and late layers, the dynamic strategy consistently outperforms all fixed settings, including the best static configuration at the 25th layer, highlighting the layer-wise variability in knowledge encoding. Dynamically selecting the most activated layer enables more effective interaction with external knowledge and improves information extraction. Moreover, we observe that when the number of similar tokens $T$ between internal and external knowledge is 10, the collaborative effect of shared and private semantics reaches its peak. Through grid search, we determine that DSSP-RAG achieves optimal performance with regularization coefficients of 0.55 for conditional entropy and 0.1 for KL divergence.

## 5 Conclusions

We have proposed DSSP-RAG to address the inter-context and context-memory knowledge conflicts in RAG. A mixed attention mechanism is developed to disentangle the internal and external knowledge into shared and private semantics, enabling fine-grained control over knowledge utilization. An unsupervised hallucination detection based on sampling subspace divergence guides the integration of external knowledge, and an Energy Quotient is introduced to filter the noise via attention difference matrices. Extensive experiments demonstrate that DSSP-RAG significantly outperforms a wide range of the state-of-the-art baselines.

## 6 Limitations

To mitigate the hallucination issues, DSSP-RAG incorporates hallucination detection, knowledge filtering, and a mixed-attention module for shared-private semantic synergy, enhancing the effectiveness of external knowledge augmentation in LLMs. Our hallucination detection method operates without annotated data or supervised training, identifying cognitive uncertainty in LLMs by assessing the stability of sampling subspaces constructed from internal states. This enables a precise determination of when external knowledge should be introduced. However, LLMs may occasionally generate consistent responses to syntactically different but semantically identical prompts due to coincidence, or they may produce stable but incorrect answers that deviate from the intended response. To address these limitations, we will propose further advancements in hallucination detection methods based on internal model states and a more in-depth investigation of LLMs' internal representations.

Additionally, we evaluated the robustness of DSSP-RAG and the baseline models under varying numbers of external documents. Although DSSP-RAG shows a general performance improvement with an increasing number of retrieved documents, a slight decline was observed when the document count reached 20. This suggests potential interference from redundant or noisy information, leading to diminishing returns. In future research, we aim to refine external knowledge selection mechanisms by integrating more advanced noise filtering strategies and context-aware integration techniques to enhance the stability and generalization capabilities of DSSP-RAG under larger-scale external knowledge conditions.

## Acknowledgments

This work is funded in part by Natural Science Foundation of China (grant number: 62376027).

## References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Findings '23*, volume 1, page 967–976.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

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, ICLR '23*.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *arXiv preprint arXiv:2402.10612*.

Qian Dong, Qingyao Ai, Hongning Wang, Yiding Liu, Haitao Li, Weihang Su, Yiqun Liu, Tat-Seng Chua, and Shaoping Ma. 2025. Decoupling knowledge and context: An efficient and effective retrieval augmented generation framework via cross attention. In *Proceedings of the ACM on Web Conference 2025*, page 4386–4395, New York, NY, USA. Association for Computing Machinery.

Yibing Du, Antoine Bosselut, and Christopher D. Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. In *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, page 10581–10589.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Yuxin Fan, Yuxiang Wang, Lipeng Liu, Xirui Tang, Na Sun, and Zidong Yu. 2025. Research on the online update method for retrieval-augmented generation (rag) model with incremental learning. *arXiv preprint arXiv:2501.07063*.

Yunfan Gao, Yun Xiong, Yijie Zhong, Yuxi Bi, Ming Xue, and Haofen Wang. 2025. Synergizing rag and reasoning: A systematic review. *Preprint*, arXiv:2504.15909.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. 2025. Deeprag: Thinking to retrieval step by step for large language models. *Preprint*, arXiv:2502.01142.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2024. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36.

Qiyuan He, Yizhong Wang, and Wenya Wang. 2024. Can language models act as knowledge bases at scale? *arXiv preprint arXiv:2402.14273*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arik. 2024. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *Preprint*, arXiv:2410.05983.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, volume 1, page 1601–1611.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. *Preprint*, arXiv:2402.16061.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *Preprint*, arXiv:2010.09926.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR '23*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Zhicheng Lee, Shulin Cao, Jinxin Liu, Jiajie Zhang, Weichuan Liu, Xiaoyin Che, Lei Hou, and Juanzi Li. 2025. Rearag: Knowledge-guided reasoning enhances factuality of large reasoning models with iterative retrieval augmented generation. *Preprint*, arXiv:2503.21729.

Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yuanbo Xu, Feiran Huang, Xi Zhang, and Philip S Yu. 2025. Loki's dance of illusions: A comprehensive survey of hallucination in large language models. *arXiv preprint arXiv:2507.02870*.

Rui Li, Xu Chen, Chaozhuo Li, Yanming Shen, Jianan Zhao, Yujing Wang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang, et al. 2023. To copy rather than memorize: A vertical learning paradigm for knowledge graph completion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6335–6347.

Rui Li, Chaozhuo Li, Yanming Shen, Zeyu Zhang, and Xu Chen. 2024. Generalizing knowledge graph embedding with universal orthogonal parameterization. *arXiv preprint arXiv:2405.08540*.

Rui Li, Jianan Zhao, Chaozhuo Li, Di He, Yiqi Wang, Yuming Liu, Hao Sun, Senzhang Wang, Weiwei Deng, Yanming Shen, et al. 2022. House: Knowledge graph embedding with householder parameterization. In *International conference on machine learning*, pages 13209–13224. PMLR.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: on reliability and robustness of current estimators. *arXiv preprint arXiv:2406.13415*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP '23*, page 9004–9017.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date, 2024. *URL https://ai. meta. com/blog/meta-llama-3/. Accessed on April*, 26.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When do llms need retrieval augmentation? mitigating llms' overconfidence helps retrieval augmentation. *arXiv preprint arXiv:2402.11457*.

Liangming Pan, Wenhu Chen, Min Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 525–539, Nusa Dua, Bali. Association for Computational Linguistics.

Han Peng, Jinhao Jiang, Zican Dong, Wayne Xin Zhao, and Lei Fang. 2025. Cafe: Retrieval head-based coarse-to-fine information seeking to enhance multi-document qa capability. *arXiv preprint arXiv:2505.10063*.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2025. Entropy-based decoding for retrieval-augmented large language models. 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 4616–4627.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP '23*, page 967–976.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2024a. Knowledge editing through chain-of-thought. *arXiv preprint arXiv:2412.17727*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, Chen Luo, Xianfeng Tang, Monica Cheng, Tuo Zhao, and Jing Gao. 2024b. Blendfilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. *arXiv preprint arXiv:2402.11129*.

Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025a. Chain-of-retrieval augmented generation. *Preprint*, arXiv:2501.14342.

Pengbo Wang, Chaozhuo Li, Chenxu Wang, Liwen Zheng, Litian Zhang, and Xi Zhang. 2025b. Two birds with one stone: Improving factuality and faithfulness of llms via dynamic interactive subspace editing. *arXiv preprint arXiv:2506.11088*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR '23*.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. *Preprint*, arXiv:2305.13300.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR '24*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. 2024. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2024. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. *arXiv preprint arXiv:2406.19215*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Findings '23*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184.

Kaiyan Zhang, Ning Ding, Biqing Qi, Xuekai Zhu, Xinwei Long, and Bowen Zhou. 2023. Crash: Clustering, removing, and sharing enhance fine-tuning without full large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9637, Singapore. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024a. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024b. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

## A Datasets and Settings

The fine-tuning dataset for DSSP-RAG consists of 30,000 samples from the train set of HotpotQA and 10,000 samples from the train set of 2WikiMultihopQA. For hyperparameter tuning, we utilize a subset of 3,000 HotpotQA samples and 2,000 2WikiMultihopQA samples. The details of the inference dataset statistics are shown in Table 7. And, the hyperparameters of DSSP-RAG on are listed in Table 8.

The tuning process of the regularization coefficients follows a systematic grid search strategy to determine the optimal parameter combination:

| | milti-hop QA | | single-hop QA | | closed-set of public health |
|---|---|---|---|---|---|
| | **HotpotQA** | **2WikiMultihopQA** | **NQ** | **TriviaQA** | **PubHealth** |
| #Examples | 1000 | 1000 | 3,610 | 11,313 | 11,800 |

Table 7: The statistics of inference dataset.

- The KL divergence regularization term $\nu$ is introduced to constrain distributional shifts when integrating external knowledge, preventing the model from becoming overly reliant on external sources, which could lead to prediction biases. Given that a smaller initial value is typically preferable to avoid excessive constraints, $\nu$ is initialized at 0.12. The search space for $\nu$ is set within the range $[0.05, 0.15]$ with a step size of 0.01.
- The conditional entropy regularization term $\mu$ is designed to mitigate uncertainty when leveraging external knowledge, thereby improving knowledge utilization efficiency. A relatively larger initial value helps the model reduce uncertainty more effectively, facilitating improved accuracy in knowledge integration. Accordingly, $\mu$ is initialized at 0.5, with a coarse search space defined as $[0.4, 0.7]$ using a step size of 0.1. For finer adjustments, the step size is refined to 0.01 within the range $\mu \in [0.5, 0.6]$.

To ensure optimal parameter selection, cross-validation is performed in the neighborhood of the most promising parameter $(0.1 \pm 0.02, 0.55 \pm 0.03)$. The results confirm that the globally optimal combination is $\nu = 0.1$ and $\mu = 0.55$.

| **Hyperparameters** | **Value** |
|---|---|
| JSD threshold $\delta$ | 1.0 |
| Similar token $T$ | 10 |
| Regularization coefficient $\mu$ | 0.55 |
| Regularization coefficient $\nu$ | 0.1 |
| Learning rate | 4e-5 |
| Epoch | 7 |
| Warm-up ratio | 0.1 |

Table 8: Hyperparameters of DSSP-RAG.

## B Evaluation on the ASQA Dataset

ASQA (Stelmakh et al., 2022) transforms QA into a summarization-style generation task, thereby allowing us to evaluate DSSP-RAG's capabilities in scenarios that demand both factual grounding and discourse-level generation quality. The results shown in Table 9 further demonstrate the extensibility of DSSP-RAG beyond traditional QA and reaffirm the effectiveness of our proposed modules under more generative settings.

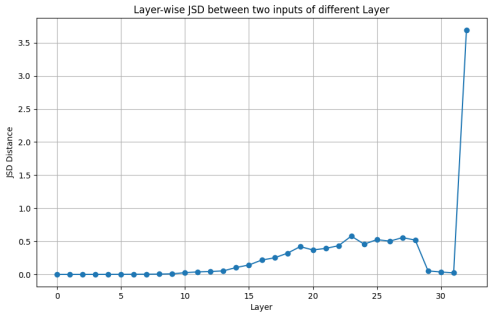
| **Model** | **LLaMA2 -7B-Chat** | | **LLaMA3 -8B-Instruct** | |
|---|---|---|---|---|
| **Method** | **str-em** | **R-L** | **str-em** | **R-L** |
| *Simple RAG* | | | | |
| No Retrieval | 16.7 | 15.5 | 29.7 | 33.4 |
| SR-RAG | 24.2 | 29.6 | 35.0 | 36.3 |
| **DSSP_SR** | **25.7** | **30.2** | **35.8** | **37.9** |
| *Adaptive RAG* | | | | |
| Self-RAG | 30.0 | 35.7 | 36.3 | 38.4 |
| FLARE | 28.5 | 33.6 | 34.3 | 35.1 |
| SEAKR | 30.8 | 36.4 | 40.1 | 40.6 |
| **DSSP_AR** | **32.2** | **38.1** | **42.9** | **43.3** |
| $\Delta$ (%) | ↑**4.54** | ↑**4.67** | ↑**6.98** | ↑**6.65** |

Table 9: Performance comparison on ASQA dataset.

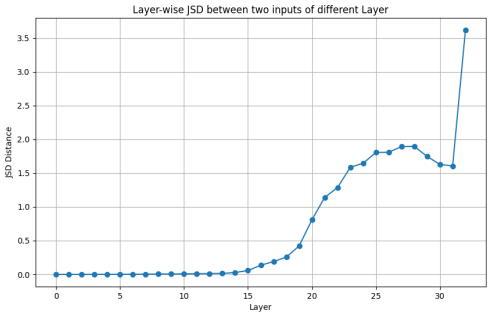
## C The details of hallucination detection

Figure 5 presents the Jensen-Shannon Divergence (JSD) between sampling subspaces across different layers of the LLM when prompted with semantically equivalent but syntactically different queries. The accompanying Table 10 details the model's responses to each prompt. When the factual answer is known, the model provides consistent responses across prompts, resulting in a small JSD between subspaces. However, due to the cognitive uncertainty, variations in the prompts can lead to divergent responses, reflected in an increasing JSD with deeper layers. If the JSD exceeds a predefined threshold($\delta = 1.0$), it indicates hallucination in the LLM's generation.

In the early stage of exploration and experimentation, we compare two metrics-KL Divergence (KL-Divergence) and Jensen-Shannon Divergence (JSD)-in the hallucination detection module during the testing phase of our work. The choice of JSD is supported by both empirical observations and

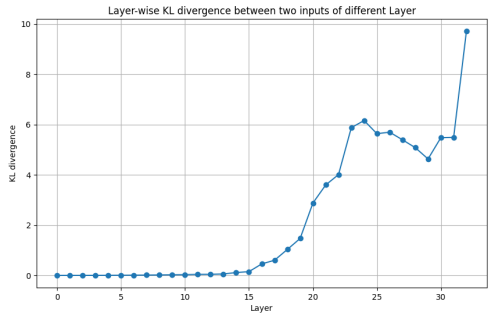



(a) Non-Hallucination

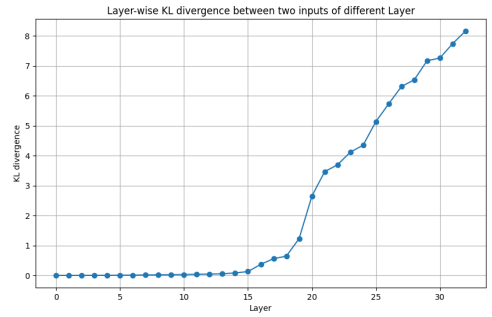



(b) Hallucination

Figure 5: JSD Distances Between Sampling Subspaces of different prompts Under Non-Hallucination and Hallucination Scenarios in LLMs.




(a) Non-Hallucination




(b) Hallucination

Figure 6: KL-Divergence Between Sampling Subspaces of different prompts Under Non-Hallucination and Hallucination Scenarios in LLMs.

| Prompt | Label | Model Answer |
|---|---|---|
| Where was the place of death of Yazid III's father? | Damascus | Damascus |
| father of Yazid III, where was place of death of him? | | Damascus, Syria |
| Where did the director of film Akcja Pod Arsenałem graduate from? | National Film School in Łódź | University of Warsaw |
| film Akcja Pod Arsenałem's director, where did him graduate from? | | Poland |

Table 10: Comparison of Model Responses to Different Prompts for Two Questions

mathematical analysis, as detailed below.

**Empirical Observations.** In our hallucination detection experiments, we measured the divergence of internal representations across different syntactic formulations of the same query using both JSD and KL-Divergence. The results reveal a key distinction: When hallucinations occur, JSD exhibits a sharp increase in intermediate layers (e.g., layers 15–25) and remains above a certain threshold, while it stabilizes below the threshold in non-hallucinatory scenarios (as shown is Figure 5). However, KL-Divergence shows elevated values in intermediate layers regardless of hallucination status, failing to differentiate hallucinatory outputs from confident predictions (as shown is Figure 6).

This divergence highlights fundamental differences in their mathematical properties and their applicability to uncertainty quantification.

**Mathematical Analysis of JSD and KL-Divergence.** Let $P$ and $Q$ denote the probability distributions of model responses to two syntactically distinct but semantically equivalent prompts. The formal definition of KL-Divergence is as follows:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{13}$$

1. Asymmetry: $D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$
2. Unbounded Range:$D_{\mathrm{KL}} \in [0, +\infty)$, which is sensitive to outliers when $Q(i) \to 0$.
3. Interpretation: Quantifies the information loss when approximating $P$ with $Q$.

The formal definition of JSD is as follows:

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M), \tag{14}$$

where $M = \frac{P+Q}{2}$.

1. Symmetry: $\mathrm{JSD}(P \parallel Q) = \mathrm{JSD}(Q \parallel P)$
2. Bounded Range: $\mathrm{JSD} \in [0, 1]$, which is robust to distribution tails.
3. Interpretation: Measures the average divergence of $P$ and $Q$ from their midpoint distribution $M$.

The suitability of JSD for hallucination detection arises from its symmetric and bounded nature, which aligns well with the requirements for detecting hallucinations caused by cognitive uncertainty: First, hallucinations manifest as inconsistencies in reasoning paths under varying syntactic contexts. JSD symmetrically captures the mutual disagreement between $P$ and $Q$, reflecting intrinsic uncertainty in the model's knowledge representation. In contrast, KL-Divergence's asymmetry introduces bias: $D_{\mathrm{KL}}(P \parallel Q)$ penalizes $Q$'s deviations from $P$ but not vice versa, amplifying noise in intermediate layers where syntactic variations naturally induce transient shifts in representation (e.g., rephrasing-triggered attention redistribution), which can lead to false positives. Second, intermediate layers often encode syntax-sensitive features (e.g., grammatical structures), which vary significantly across paraphrased queries even in non-hallucinatory scenarios. KL-Divergence's unbounded sensitivity to such variations causes spurious spikes, whereas JSD's boundedness filters out syntax-driven noise while retaining semantic-level inconsistencies.

The symmetry, boundedness, and robustness of JSD to syntactic noise make it particularly well-suited for detecting hallucinations driven by cognitive uncertainty. In contrast, the sensitivity of KL-Divergence to transient syntactic variations and directional bias limits its effectiveness in this context. Therefore, we have selected JSD as the metric

for our unsupervised hallucination detection module.

**Hyperparameter analysis** To evaluate the robustness of our hallucination detection threshold $\delta$, we conduct a sensitivity analysis by varying $\delta$ from 0.6 to 1.5 in increments of 0.1, and measuring the resulting EM scores on three datasets: HotpotQA, 2Wiki, and TriviaQA. As shown in Table 11, when the hallucination detection threshold is set too low, external knowledge is introduced without sufficient control, which may interfere with the LLM's reasoning process. As the threshold $\delta$ increases from 0.6 to 1.0, performance improves steadily, reaching its optimal at $\delta = 1.0$. Within the range $\delta \in [0.8, 1.2]$, EM scores remain stable with minimal fluctuation. However, when $\delta$ exceeds 1.2, the performance begins to degrade, likely due to the over-restriction of external knowledge retrieval.

| $\delta$ | HotpotQA | 2Wiki | TriviaQA |
|---|---|---|---|
| | EM | EM | EM |
| 0.6 | 27.6 | 30.1 | 52.9 |
| 0.7 | 27.9 | 30.4 | 53.3 |
| 0.8 | 28.1 | 30.9 | 53.9 |
| 0.9 | 28.4 | 31.4 | 54.3 |
| **1.0** | **28.5** | **31.6** | **54.5** |
| 1.1 | 28.3 | 31.5 | 54.4 |
| 1.2 | 27.9 | 31.0 | 54.0 |
| 1.3 | 25.3 | 28.5 | 50.1 |
| 1.4 | 23.1 | 25.3 | 47.6 |
| 1.5 | 20.4 | 19.5 | 44.3 |

Table 11: Sensitivity analysis of detection threshold $\delta$ on EM scores.

In the future work, we will introduce a more principled and adaptive mechanism for threshold selection to enhance the generality and flexibility of the framework. For example, potential future directions include: (1) dynamically adjusting $\delta$ based on statistical properties (e.g., the distribution of JSD across layers or samples), and (2) clustering-based threshold calibration. In the latter, we can apply unsupervised clustering algorithms (e.g., K-means or Gaussian Mixture Models) to the distribution of JSD values obtained from semantically perturbed prompts. By identifying natural groupings in the JSD space, for example, stable vs. unstable subspaces, the boundary between clusters can be used to automatically determine a context-dependent threshold $\delta$, eliminating the need for manual tuning and improving adaptability across models and domains.

## D The details of knowledge filtering

We evaluate LLMs' attention distribution and response behavior when external knowledge is incorporated into the input query using the following prompts, where the red-highlighted segments indicate noise. As described in the methodology, the key layers in LLMs are responsible for capturing abstract global semantics, focusing on task-specific feature representation, and playing a central role in reasoning, as illustrated in Figure 7. In contrast, offset layers tend to amplify noise, leading to deviations in attention distribution and adversely affecting prediction performance, as shown in Figure 8. Excessive reliance on external knowledge, coupled with noise interference, results in erroneous model outputs. Using attention distribution differences between the key and offset layers, we can effectively identify noise and mitigate its negative impact.

**Prompt of knowledge filtering**

*User:*
Please answer the given question with format referring to the knowledge provided below.
**Question:** Where was the performer of song Get A Life – Get Alive born?
**Knowledge:** The song is sung by Eric Papilaya, and was written by Greg Usek (music) and Austin Howard (lyrics). Bernard Bonvoisin, known as Bernie Bonvoisin( born 9 July 1956 in Nanterre, Hauts- de- Seine), is best known for having been the singer of Trust. Eric Papilaya (born 9 June 1978 in Vöcklabruck, Upper Austria) is a singer from Austria who represented his home country in the Eurovision Song Contest 2007 in Helsinki, Finland.

*Assistant:*
**A: The answer is**

*Output:* 9 July 1956 in Nanterre, Hauts- de- Seine

Figure 10: Prompt of knowledge filtering

Figure 9 presents the performance variation curves of the original and fine-tuned models under layer-wise pruning on the same test cases. The results indicate that semantic entropy, when computed based on the probability distribution of the original model, fails to serve as a reliable metric for evaluating key and offset layers. However, after fine-tuning with the proposed DSSP components and loss function, semantic entropy effectively captures the performance differences between the key and offset layer compared to the LLM without integrated knowledge.

Our approach leverages conditional entropy to mit-

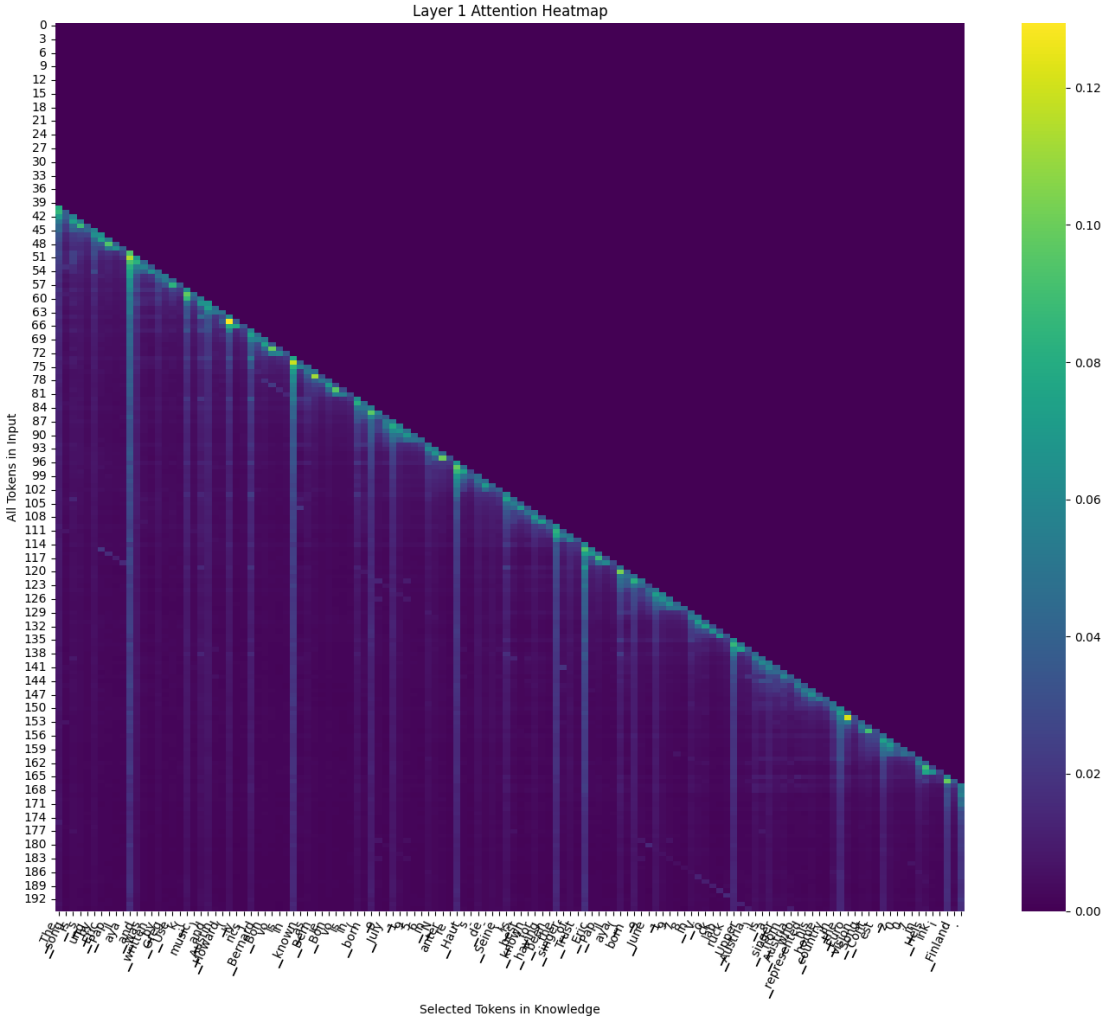



Figure 7: Attention heatmap of key layer

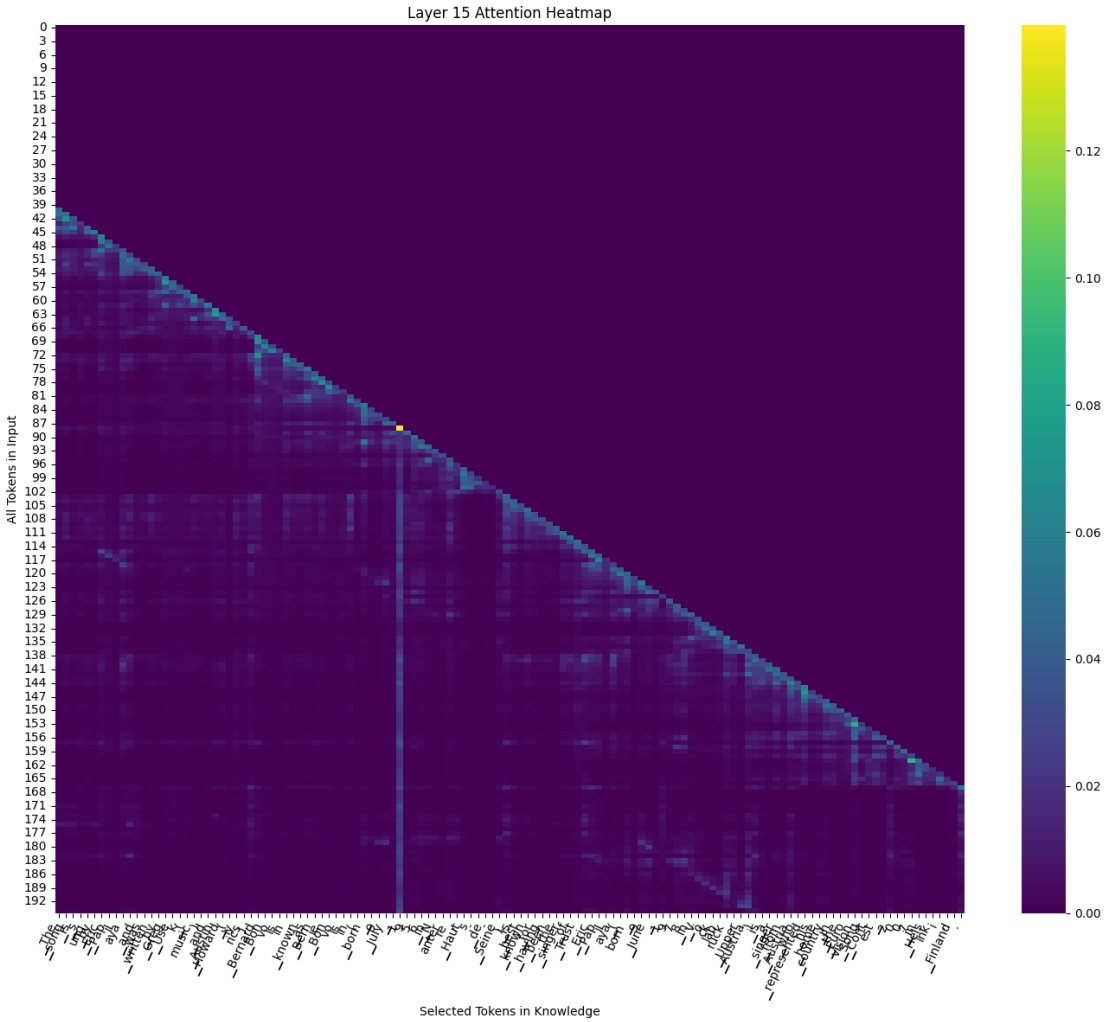
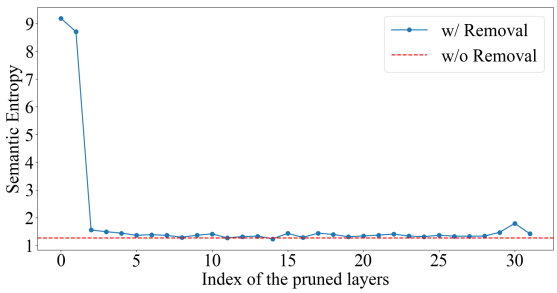
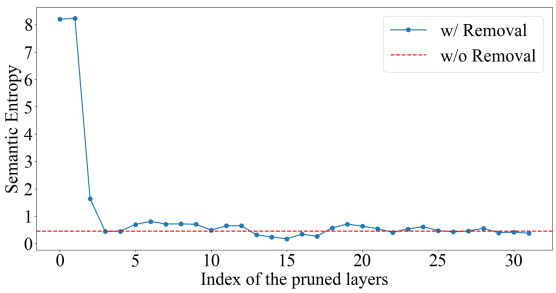



Figure 8: Attention heatmap of offset layer




(a) Performance variation curves of the original models




(b) Performance variation curves of the fine-tuned models

Figure 9: Performance variation curves of the original and fine-tuned models under layer-wise pruning.

igate uncertainty while using KL divergence to enhance distribution stability, thereby improving external knowledge filtering during the reasoning process. The fine-tuned model alleviates the inherent confidence bias of the original model, which may lead to excessive confidence in incorrect predictions or undue uncertainty in correct predictions.

## E Mathematical derivation of the mixed attention for shared-private semantic decomposition

In our work, to enhance generation process of LLM with the filtered external knowledge, DSSP module employs mixed attention to decompose dual-stream knowledge into shared and private semantics at the target layer. The semantic relevance between internal knowledge $X$ and external knowledge $Y$ is estimated based on dot product similarity, where the top $T$ tokens in the external knowledge that exhibit the highest similarity to the internal knowledge are selected as shared semantics. Furthermore, we propose a differential attention mechanism to extract private semantics between internal and external knowledge. In the following, the detailed explanation and analysis of the underlying principles of differential attention are provided.

The differential attention mechanism is defined as the difference between self-attention $\tau(\cdot)$ of a single knowledge source and cross-attention $\eta(\cdot)$ of dual-stream knowledge:

$$\mathcal{D}_{\text{attn}}(X, Y) = \tau(X, X) - \eta(X, Y) \tag{15}$$

The mathematical explanation via explicit operator decomposition is as follows.

Let the latent representations of internal knowledge $X$ and external knowledge $Y$ be decomposed as follows:

$$X = S + P_X + N_X, \tag{16}$$

$$Y = S + P_Y + N_Y, \tag{17}$$

where $S$ represents the shared semantics (common component), $P_X, P_Y$ denote the private semantics of internal and external knowledge, respectively, and $N_X, N_Y$ correspond to noise components.

Traditional cross-attention $\eta(X, Y)$ primarily models the associations between $X$ and $Y$, and its output can be approximated as:

$$\begin{aligned}\eta(X, Y) &\propto \text{softmax}(XW_Q(YW_K)^T)YW_V \\ &\approx \alpha S + \beta(N_X + N_Y),\end{aligned} \tag{18}$$

where $\alpha$ and $\beta$ are weighting coefficients. This formulation indicates that cross-attention predominantly captures a mixture of the shared semantics $S$ and noise components $N_X, N_Y$, while failing to effectively distinguish private semantics.

Conversely, self-attention $\tau(X, X)$ models the internal dependencies within $X$, and its output can be approximated as:

$$\begin{aligned}\tau(X, X) &\propto \text{softmax}(XW_Q(XW_K)^T)XW_V \\ &\approx \gamma S + \delta P_X + \epsilon N_X,\end{aligned} \tag{19}$$

which captures a combination of shared semantics $S$, private semantics $P_X$, and noise $N_X$.

By computing the difference between self-attention and cross-attention, $U_{\text{private}} \approx \tau(X, X) - \eta(X, Y)$, the contribution of the shared semantics $S$ is effectively removed:

$$\begin{aligned}U_{\text{private}} &\approx (\gamma S + \delta P_X + \epsilon N_X) \\ &\quad - (\alpha S + \beta(N_X + N_Y)) \\ &= (\gamma - \alpha)S + \delta P_X + (\epsilon - \beta)N_X - \beta N_Y,\end{aligned} \tag{20}$$

By optimizing parameters such that $\gamma \approx \alpha$ and $\epsilon \approx \beta$, the formulation is simplified as follows:

$$U_{\text{private}} \approx \delta P_X - \beta N_Y. \tag{21}$$

The above derivation process demonstrates that the differential attention mechanism suppresses both shared semantics $S$ and internal noise $N_X$, while preserving private semantics $P_X$ and attenuating the impact of external noise $N_Y$.

## F Prompt of query variation

The prompt designed to enable the LLM to automatically generate alternative expressions, which are semantically similar but syntactically distinct from the original queries, is as follows:

**Prompt of query variation**

*User:*
Please provide another expression with the same meaning as the given sentence following the given example, and ensuring that both expressions have the same tokenization length using AutoTokenizer.
**Sentence:** Who is the paternal grandfather of Tancred, King Of Sicily?
**Another Sentence:** Tancred, King Of Sicily's paternal grandfather is who?
**Sentence:** What is cause of death of performer of song How Do U Want It?
**Another Sentence:** performer of song How Do U Want It, what is cause him die?
**Sentence:** Where was the place of death of Yazid III's father?
**Another Sentence:** father of Yazid III Where was place of death of him?
**Sentence:** [input]
*Assistant:*
**Another Sentence:**

Figure 11: Prompt of query variation

## G Case study

To illustrate the practical implications of DSSP-RAG's Shared-Private Semantic Synergy in real-world tasks, we present representative examples in Table 12. The first case demonstrates how the DSSP module effectively prevents the LLM from over-relying on noisy information introduced by the retrieved context during reasoning. The second case illustrates that, under partial guidance from external knowledge, the DSSP module enables the LLM to successfully leverage its parametric knowledge to produce correct reasoning, thereby achieving a synergistic integration of internal and external knowledge.

## H In context learning examples

For DSSP_SR, the retrieved external knowledge and query are provided separately as inputs to the LLM. At the target layer, the hybrid attention mechanism of the DSSP module integrates external knowledge with the parametric knowledge of the model to generate responses, as illustrated in Figure 2. For DSSP_adaptive, to maximally activate the LLM's parametric knowledge and reasoning capabilities, we follow the SEAKR approach by providing identical in-context learning examples for single-hop QA datasets, as shown in Figure 12. For multi-hop QA datasets, we leverage IRCoT to supply dataset-specific in-context learning examples: HotpotQA (Figure 13) and 2WikiMultiHopQA (Figure 14). The retrieved knowledge remains a separate input to the LLM, where it is first encoded and subsequently integrated with the parametric knowledge of the model through the hybrid attention mechanism of the DSSP module, enhancing both the generation process and the accuracy of the reasoning.

| | | |
|---|---|---|
| **Question** | Where was the performer of song *Get A Life – Get Alive* born? | |
| **Ground Truth** | Vöcklabruck | |
| **Context** | Bernard Bonvoisin, known as Bernie Bonvoisin (born 9 July 1956 in Nanterre, de-Seine), is best known for having been the singer of Trust **[Noise]**. The song is sung by Eric Papilaya, and was written by Greg Usek (music) and Austin Howard (lyrics). Eric Papilaya (born 9 June 1978 in Vöcklabruck, Upper Austria) is a singer from Austria who represented his home country in the Eurovision Song Contest 2007 in Helsinki, Finland. | |
| **Output** | w/o DSSP | Nanterre, de-Seine |
| | w/ DSSP | **Vöcklabruck** |
| **Question** | Who is the paternal grandfather of Tancred, King Of Sicily? | |
| **Ground Truth** | Roger II of Sicily | |
| **Context** | Roger's first public act took place at Melfi in 1129, where, though still a child, he accepted the fealty of some rebellious barons along with his father and his younger brother Tancred **[Noise]**. Tancred was King of Sicily from 1189 to 1194. He was born in Lecce, an illegitimate son of Roger III, Duke of Apulia (the eldest son of King Roger II) by his mistress Emma, a daughter of Achard II, Count of Lecce. | |
| **Output** | w/o DSSP | Roger I of Sicily |
| | w/ DSSP | **Roger II of Sicily** |

Table 12: Case study: DSSP mitigates hallucinations via synergizing parametric and external knowledge under conflicting contexts.

single-hop QA

**Question:** Nobody Loves You was written by John Lennon and released on what album that was issued by Apple Records, and was written, recorded, and released during his 18 month separation from Yoko Ono?
**Answer:** The album issued by Apple Records, and written, recorded, and released during John Lennon's 18 month separation from Yoko Ono is Walls and Bridges. Nobody Loves You was written by John Lennon on Walls and Bridges album. So the answer is Walls and Bridges.
**Question:** What is known as the Kingdom and has National Route 13 stretching towards its border?
**Answer:** Cambodia is officially known as the Kingdom of Cambodia. National Route 13 stretches towards the border to Cambodia. So the answer is Cambodia.
**Question:** Jeremy Theobald and Christopher Nolan share what profession?
**Answer:** Jeremy Theobald is an actor and producer. Christopher Nolan is a director, producer, and screenwriter. Therefore, they both share the profession of being a producer. So the answer is producer.
**Question:** What film directed by Brian Patrick Butler was inspired by a film directed by F.W. Murnau?
**Answer:** Brian Patrick Butler directed the film The Phantom Hour. The Phantom Hour was inspired by the films such as Nosferatu and The Cabinet of Dr. Caligari. Of these, Nosferatu was directed by F.W. Murnau. So the answer is The Phantom Hour.
**Question:** Vertical Limit stars which actor who also played astronaut Alan Shepard in 'The Right Stuff'?
**Answer:** The actor who played astronaut Alan Shepard in 'The Right Stuff' is Scott Glenn. The movie Vertical Limit also starred Scott Glenn. So the answer is Scott Glenn.
**Question:** Which car, produced by Ferrari from 1962 to 1964 for homologation into the FIA's Group 3 Grand Touring Car category inspired the Vandenbrink GTO?
**Answer:** The car produced by Ferrari from 1962 to 1964 for homologation into the FIA's Group 3 Grand Touring Car category is the Ferrari 250 GTO. The Ferrari 250 GTO also inspired the Vandenbrink GTO's styling. So the answer is Ferrari 250 GTO.

Following the examples above, answer the question by reasoning step-by-step.
**Question:** [Question]

Figure 12: Examples for single-hop QA datasets

HotpotQA

**Question:** Jeremy Theobald and Christopher Nolan share what profession?
**Answer:** Jeremy Theobald is an actor and producer. Christopher Nolan is a director, producer, and screenwriter. Therefore, they both share the profession of being a producer. So the answer is producer.
**Question:** What film directed by Brian Patrick Butler was inspired by a film directed by F.W. Murnau?
**Answer:** Brian Patrick Butler directed the film The Phantom Hour. The Phantom Hour was inspired by the films such as Nosferatu and The Cabinet of Dr. Caligari. Of these, Nosferatu was directed by F.W. Murnau. So the answer is The Phantom Hour.
**Question:** How many episodes were in the South Korean television series in which Ryu Hye-young played Bo-ra?
**Answer:** The South Korean television series in which Ryu Hye-young played Bo-ra is Reply 1988. The number of episodes Reply 1988 has is 20. So the answer is 20.
**Question:** Were Lonny and Allure both founded in the 1990s?
**Answer:** Lonny (magazine) was founded in 2009. Allure (magazine) was founded in 1991. Thus, of the two, only Allure was founded in the 1990s. So the answer is no.
**Question:** Vertical Limit stars which actor who also played astronaut Alan Shepard in The Right Stuff ?
**Answer:** The actor who played astronaut Alan Shepard in The Right Stuff is Scott Glenn. The movie Vertical Limit also starred Scott Glenn. So the answer is Scott Glenn.
**Question:** What was the 2014 population of the city where Lake Wales Medical Center is located?
**Answer:** Lake Wales Medical Center is located in the city of Lake Wales, Polk County, Florida. The population of Lake Wales in 2014 was 15,140. So the answer is 15,140.
**Question:** Who was born first? Jan de Bont or Raoul Walsh?
**Answer:** Jan de Bont was born on 22 October 1943. Raoul Walsh was born on March 11, 1887. Thus, Raoul Walsh was born first. So the answer is Raoul Walsh.
**Question:** In what country was Lost Gravity manufactured?
**Answer:** The Lost Gravity (roller coaster) was manufactured by Mack Rides. Mack Rides is a German company. So the answer is Germany.

Following the examples above, answer the question by reasoning step-by-step.
**Question:** [Question]

Figure 13: Examples for HotpotQA

2WikiMultihopQA

**Question:** Who was born first out of Martin Hodge and Ivania Martinich?
**Answer:** Martin Hodge was born on 4 February 1959. Ivania Martinich was born on 25 July 1995. Thus, 4 February 1959 is earlier than 25 July 1995 and Martin Hodge was born first. So the answer is Martin Hodge.
**Question:** When did the director of film Hypocrite (Film) die?
**Answer:** The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So the answer is 19 June 2013.
**Question:** Are both Kurram Garhi and Trojkrsti located in the same country?
**Answer:** Kurram Garhi is located in the country of Pakistan. Trojkrsti is located in the country of Republic of Macedonia. Thus, they are not in the same country. So the answer is no.
**Question:** Do the director of film Coolie No. 1 (1995 Film) and the director of film The Sensational Trial have the same nationality?
**Answer:** Coolie No. 1 (1995 film) was directed by David Dhawan.
The Sensational Trial was directed by Karl Freund. David Dhawan's nationality is Indian. Karl Freund's nationality is German. Thus, they do not have the same nationality. So the answer is no.
**Question:** Who is Boraqchin (Wife Of Ögedei)'s father-in-law?
**Answer:** Boraqchin is married to Ögedei Khan. Ögedei Khan's father is Genghis Khan. Thus, Boraqchin's father-in-law is Genghis Khan. So the answer is Genghis Khan.
**Question:** When did the director of film Laughter In Hell die?
**Answer:**The film Laughter In Hell was directed by Edward L. Cahn. Edward L. Cahn died on August 25, 1963. So the answer is August 25, 1963.

Following the examples above, answer the question by reasoning step-by-step.
**Question:** [Question]

Figure 14: Examples for 2WikiMultihopQA