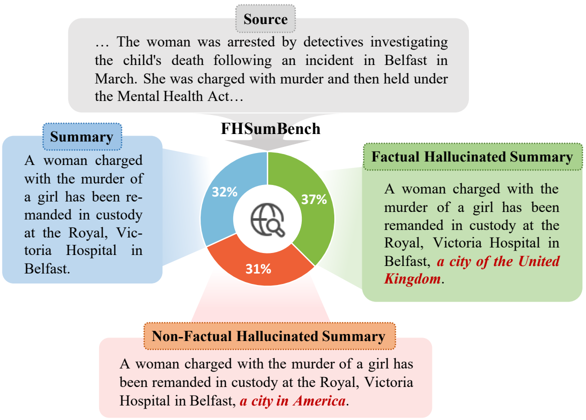
Evaluating LLMs' Assessment of Mixed-Context Hallucination Through the Lens of Summarization

Siya Qi[♣] **Rui Cao**[♠] **Yulan He**[♣,♢] **Zheng Yuan**[♡]
[♣]King's College London [♠]Cambridge University
[♢]The Alan Turing Institute [♡]The University of Sheffield
{siya.qi, yulan.he}@kcl.ac.uk
rc990@cam.ac.uk
zheng.yuan1@sheffield.ac.uk

Abstract

With the rapid development of large language models (LLMs), LLM-as-a-judge has emerged as a widely adopted approach for text quality evaluation, including hallucination evaluation. While previous studies have focused exclusively on single-context evaluation (e.g., discourse faithfulness or world factuality), real-world hallucinations typically involve mixed contexts, which remains inadequately evaluated. In this study, we use summarization as a representative task to comprehensively evaluate LLMs' capability in detecting mixed-context hallucinations, specifically distinguishing between factual and non-factual hallucinations. Through extensive experiments across direct generation and retrieval-based models of varying scales, our main observations are: (1) LLMs' intrinsic knowledge introduces inherent biases in hallucination evaluation; (2) these biases particularly impact the detection of factual hallucinations, yielding a significant performance bottleneck; and (3) the fundamental challenge lies in effective knowledge utilization, balancing between LLMs' intrinsic knowledge and external context for accurate mixed-context hallucination evaluation.[1]


Source
FHSumBench
Summary
32%
37%
31%
Factual Hallucinated Summary
Non-Factual Hallucinated Summary

… The woman was arrested by detectives investigating the child's death following an incident in Belfast in March. She was charged with murder and then held under the Mental Health Act…

A woman charged with the murder of a girl has been remanded in custody at the Royal, Victoria Hospital in Belfast.

A woman charged with the murder of a girl has been remanded in custody at the Royal, Victoria Hospital in Belfast, ***a city of the United Kingdom***.

A woman charged with the murder of a girl has been remanded in custody at the Royal, Victoria Hospital in Belfast, ***a city in America***.

Figure 1: Examples of our automated construction for mixed-context hallucination datasets, where "a city of the United Kingdom" is the correct description of "Belfast", constructed as factual hallucination. "A city in America" is an incorrect description of "Belfast", constructed as a non-factual hallucination.

1 Introduction

Large language models (LLMs) generate coherent text and follow instructions across diverse tasks (Zhao et al., 2024; Minaee et al., 2024). However, a critical challenge in scaling LLM applications is hallucination, where the generated content lacks factual grounding (Lin et al., 2022; Li et al., 2023) or deviates from the intended discourse context (Zhou et al., 2023). This issue makes hallucination evaluation essential for improving LLM reliability and developing more robust models (Huang et al., 2023; Zhang et al., 2023; Ji et al., 2023; Tonmoy et al., 2024). Recent studies have explored the potential of using LLMs themselves to assess hallucinations, reporting strong performance in hallucination detection with LLM-based evaluators (Gu et al., 2025; Min et al., 2023; Chern et al., 2023).

In practical applications, hallucinations typically derive from two primary sources: one originates from the discourse context, affecting faithfulness; while the other arises from external knowledge or the model's inherent knowledge, influencing factuality (Maynez et al., 2020). Previous LLM-based evaluators for hallucinations have focused on assessing either faithfulness (Manakul et al., 2023b; Wu et al., 2023; Zha et al., 2023) or factuality (Chern et al., 2023; Min et al., 2023; Dhuliawala et al., 2024; Hu et al., 2024) of the text under evaluation. However, hallucinations can emerge from a combination of discourse context and external information – a mixed-context setting that remains under-explored in LLM-based evaluation. Many real-world scenarios operate within this mixed-context setting. For instance, in summarization, a system might fabricate details while still pre-

[1] Code and data available at `https://github.com/cece00/FHSumBench`.

Findings of the Association for Computational Linguistics: ACL 2025, pages 16480–16503
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

senting factually correct information; in customer support chatbots, it may hallucinate from chat histories alongside accurate product information; and in image captioning, models might provide factually correct descriptions about non-existing objects in an image. Evaluating these mixed-context hallucinations requires simultaneous assessment of both faithfulness and factuality (Qi et al., 2024), which presents unique challenges for developing robust evaluation metrics and benchmarks. News summarization serves as an ideal testbed for this purpose, given its well-defined input references and the verifiability of factual content. In this work, we focus on the summarization task to examine the capabilities of LLMs under this evaluation setting.

With mixed-context considered, hallucinations in summarization can be categorized into two types:[2] **factual hallucinations**, where the generated content is factually correct but inconsistent with the source (e.g., the green block in Figure 1); and **non-factual hallucinations**, where the output contains outright factual inaccuracies (e.g., the red block in Figure 1). Existing datasets for mixed-context hallucinations in summarization (Maynez et al., 2020; Cao et al., 2022; Dong et al., 2022) are limited both in scale and exhibit unbalanced distributions of hallucination instances, making them inadequate to assess LLMs' capability in handling mixed-context hallucinations. To address this limitation, this study introduces the first automated pipeline for constructing mixed-context hallucination evaluation datasets, and aims to investigate the following research questions: **RQ1)** How do LLM-based hallucination evaluators perform in the mixed-context setting with different prompting strategies? Given the rapid advancement of retrieval-augmented generation (RAG) (Gao et al., 2024), we then examine **RQ2)** How can LLMs better leverage external knowledge to detect hallucinations? Among the hallucination categories, we ask **RQ3)** Which category benefits the most across evaluation methods and models? Finally, a fundamental question in LLM research, **RQ4)** Does scaling up model size lead to better hallucination assessment? These questions collectively address the challenges in mixed-context hallucination assessment, with our findings laying the groundwork for scalable LLM self-evaluation and self-evolution. The contributions and key observations of this work are as follows:

1. We comprehensively evaluate LLM-as-a-judge approaches for mixed-context hallucination detection through summarization, introducing an effective, well-balanced, and easily scalable benchmark, FHSumBench.

2. Our extensive experiments across direct generation and retrieval-based LLMs of varying scales reveal that model scaling does not guarantee better performance. However, the main performance bottleneck lies in the detection of factual hallucinations.

3. We identify that effective knowledge utilization remains the primary challenge in mixed-context hallucination evaluation. Prompt engineering yields greater improvements on smaller LLMs. While external knowledge augmentation enhances accuracy, it requires carefully designed retrieval strategies.

2 Related Work

Mixed-Context Hallucination Context plays a crucial role in hallucination evaluation. Hu et al. (2024) examine the hallucination detection capabilities of LLMs in the question-answer task under different contextual settings. However, their evaluation is limited to isolated contexts. For summarization, while the source document serves as a context input, the factuality of the summary is equally critical. Maynez et al. (2020) first systematically annotated the faithfulness and factuality separately on the XSum (Narayan et al., 2018) dataset, using summaries generated by small language models. Later, Cao et al. (2022) introduced another dataset with entity-level annotations, though it remained limited to XSum. The annotation process demands extensive effort from domain experts and heavily relies on annotator agreement, making it both time-consuming and resource-intensive.

Hallucination Evaluation Methods Most previous works only focus on one aspect of hallucination evaluation, which is either faithfulness or factuality. For faithfulness evaluation, a natural approach is to assess the entailment between the text and the context document (Wu et al., 2023; Zha et al., 2023; Goyal and Durrett, 2020) or to extract and compare answers by querying them (Manakul et al., 2023a; Durmus et al., 2020; Fabbri et al., 2022). For factuality evaluation, LLMs intrinsic

[2] We focus on the news summarization task, where source articles are assumed to be factually correct, so a hallucinated summary cannot be both faithful and non-factual.

knowledge (Chen et al., 2023; Huang et al., 2024) and external knowledge source (Min et al., 2023; Chern et al., 2023; Dhuliawala et al., 2024) are both used for detection. With both aspects considered, EntFA (Cao et al., 2022) utilizes entity original and context-conditioned features to detect factual and non-factual hallucinations.

3 FHSumBench

To evaluate models' ability to detect mixed-context hallucinations in summaries, an appropriate test set is essential. To the best of our knowledge, the XEnt dataset (Cao et al., 2022) is the only one providing factual hallucination annotations for summaries. However, among the 240 samples in the test set, only 29% of entities are annotated as factual or non-factual hallucinations, which is insufficient for assessing LLMs. Another related dataset is M-XSum[3] (Maynez et al., 2020), which is also built on the XSum dataset and includes annotations for separate labels: faithfulness and factuality. The faithfulness annotations classify hallucinations as either intrinsic or extrinsic, while the factuality annotations are based on world knowledge. After filtering out samples with inconsistent annotations, M-XSum remains an imbalanced dataset with 92% non-factual hallucination samples, as shown in Figure 2. Therefore, to obtain more diverse evaluation data, we propose an automated pipeline capable of accurately and effectively constructing both factual and non-factual hallucinations. Using this pipeline, we construct the **Factual Hallucination in Summarization Benchmark (FHSumBench)** to evaluate the performance of existing factuality and faithfulness evaluators, as well as LLM-as-a-judge approaches.

The core component of this pipeline is **fact injection**: injecting either factual or non-factual information into correct summaries. We use XEnt (Cao et al., 2022) and FactCollect (Feng et al., 2023) as our seed datasets, which provide annotated correct summaries, defined as being both faithful and factual.[4] The first step of the pipeline is to obtain correct summaries. For XEnt dataset, we remove spans annotated as "non-factual hallucinations," while for the FactCollect dataset, we use summaries labeled as "correct." The second step involves injecting factual information into the correct summaries. Here, we focus on entity knowledge as the basis for injection. Entities are extracted using the off-the-shelf NER tool (Honnibal and Montani, 2018), and a one-sentence description of each entity is used as the injected information. The injected facts are extracted from Wikidata (Vrandečić and Krötzsch, 2014). To create factual hallucinations, the correct description is appended to the corresponding entity. In contrast, non-factual hallucinations are built by appending a randomly selected, unrelated description to the entity, as in Figure 1. We also conducted a preliminary study over the data construction pipeline (see Appendix C.1).

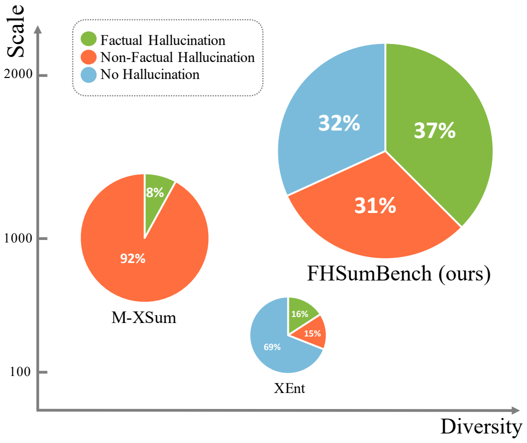


Figure 2: The size and distributions of FHSumBench, M-XSum, and XEnt datasets.

As a result, FHSumBench contains 1,336 samples, evenly distributed across three categories of summaries, providing a balanced dataset for comprehensive evaluation. This approach is more efficient than manual annotation, as it does not need experts or extensive time. Additionally, it is more controllable than data synthesis by LLMs, as it relies on precise descriptions for each entity. While preserving these advantages, this benchmark can effectively evaluate the ability of current LLMs and evaluators to assess factual hallucinations.

4 LLM-Based Evaluators on Mixed-Context Hallucination

In this section, we describe widely used LLM-based evaluators for hallucinations. Based on whether they retrieve knowledge from external sources, we categorize the evaluators into two types: **Direct Generation Evaluators** (§4.1) and **Retrieval-Based Evaluators** (§4.2). We will evaluate the performance of these approaches by analyz-

[3] Also referred to as *XSumFaith*. We use the name of *M-XSum* to prevent potential misunderstanding. We use both its faithfulness and factuality annotations.

[4] The relationship between FHSumBench and previous summarization datasets can be seen in Table 5, Appendix A.1.

ing their effectiveness in assessing mixed-context hallucination.

4.1 Direct Generation Evaluators

Vanilla Judge The vanilla judge evaluates based solely on the document-summary pair. The evaluation is designed to maximize self-contained judgment, ensuring that the LLM relies solely on its intrinsic knowledge, without additional modules. Specifically, the LLM assesses faithfulness by referencing the source document, while factuality is judged based on its intrinsic knowledge. For output Y and input X (comprising context C and query Q), where C includes the source document for vanilla judge, we have $P(Y \mid X) = P(Y \mid C, Q)$.

Prompting Strategies Following vanilla judge, we want to further investigate what capabilities are lacking in vanilla LLMs for objective and accurate judgment. Specifically, we seek to determine whether these challenges arise from constraints in reasoning ability or from difficulties in comprehending the task. To systematically analyze this, we explore two prompting strategies: in-context learning (+ICL) and chain-of-thought reasoning (+CoT). For **+ICL Judge**, we provide the LLM with three annotated examples, labeled for faithfulness and factuality, to facilitate in-context learning. For **+CoT Judge**, we incorporate the phrase "Think step by step" into the prompt to guide the LLM in generating a reasoning trajectory.

4.2 Retrieval-Based Evaluators

In this study, we explore evidence retrieval using previous evaluators and LLMs, comparing the following approaches to assess the faithfulness and factuality of summaries (see Figure 3).

Hybrid Score We select two evaluators from different evaluation aspects for each category and combined them as baselines. For faithfulness evaluation, we use WeCheck (Wu et al., 2023), which aggregates results from multiple NLI models, and AlignScore (Zha et al., 2023), which measures faithfulness through semantic alignment. For factuality evaluation, we employ FactScore (Min et al., 2023), which retrieves evidence from a knowledge base, and FacTool (Chern et al., 2023), which retrieves evidence through online searches. Faithfulness evaluation determines the presence of hallucinations, while factuality evaluation classifies them as factual or non-factual.

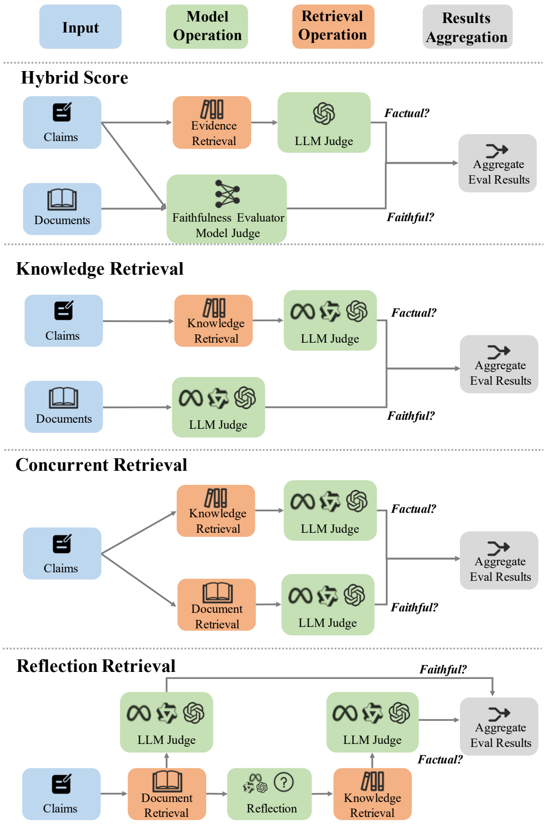


Figure 3: The pipelines of different retrieval-based evaluation methods.

Knowledge Retrieval This approach involves only retrieving summary-relevant knowledge and integrating it directly into the context alongside the document, offering a straightforward method for incorporating external information. Specifically, we break the summary S into individual claims $\{c_1, ..., c_l\}$ and retrieve knowledge evidence E^{kb} related to the entities in each claim c_i.

Concurrent Retrieval This approach involves simultaneously retrieving evidence from the source document and the knowledge base for each claim, referred to as Concurrent Retrieval. Specifically, LLMs generate a query q_i for each claim c_i. These queries are then used to retrieve the most relevant evidence from the evidence pool, which comprises the chunked source document E^d and entity-centric knowledge evidence E^{kb}, as Eq. 1 and Eq. 2.

$$e_i^d = \underset{e_j^d \in E^d}{\operatorname{argmax}} R(e_j^d \mid q_i), \forall q_i \in \{q_1, ..., q_l\} \quad (1)$$

$$e_i^{kb} = \underset{e_j^{kb} \in E^{kb}}{\operatorname{argmax}} R(e_j^{kb} \mid q_i), \forall q_i \in \{q_1, ..., q_l\} \quad (2)$$

Reflection Retrieval The summary hallucination existence is primarily judged based on the source

Methods		FHSumBench			M-XSum		
		P	R	F	P	R	F
Random (lower bound)		0.3185	0.3184	0.3185	0.3304	0.2120	0.2583
EntFA (Cao et al., 2022)		0.3295	0.3119	0.3205	0.3366	0.3489	0.3426
Vanilla Judge	Llama3-8B	0.3652	0.2977	0.3280	0.3477	0.1914	0.2469
	Qwen2.5-14B	0.4549	0.4652	0.4600	0.3710	0.4236	0.3955
	GPT-4o	0.4575	0.4650	0.4612	**0.3905**	0.4466	0.4166
+ICL	Llama3-8B	0.3770	0.3603	0.3685	0.3664	0.2994	0.3295
	Qwen2.5-14B	0.4333	0.4486	0.4408	0.3897	0.4347	0.4110
	GPT-4o	0.4519	0.4585	0.4552	0.3806	**0.4629**	**0.4178**
+CoT	Llama3-8B	0.3437	0.3044	0.3228	0.3575	0.3021	0.3275
	Qwen2.5-14B	**0.4717**	**0.4749**	**0.4733**	0.3735	0.3908	0.3819
	GPT-4o	0.4539	0.4739	0.4637	0.3662	0.3766	0.3713

Table 1: The results using direct generation LLM evaluators on FHSumBench and M-XSum datasets. We report the precision (P), recall (R) and F1-score (F) in this table. The best and second-best results are highlighted in **bold** and underlined, respectively. "+ICL" denotes in-context learning judge and "+CoT" denotes chain-of-thoughts judge, both of which are based on vanilla judge.

document. Therefore, this approach adopts a two-stage retrieval evaluation. In the first stage, we keep Eq. 1, where evidence e_i^d is retrieved from the source document to assess faithfulness. Based on this evaluation, the LLM reflects and identifies missing or incomplete information and generates a refined query q_i^r to retrieve additional evidence from the knowledge base, as Eq. 3.

$$q_i^r = P(y \mid [c_i, e_i^d], q_{reflect}) \tag{3}$$

This sequential process ensures that hallucination is evaluated with a more comprehensive set of evidence, which we call it Reflection Retrieval.

5 Experiment and Result

This section presents the experiment settings and offers a comprehensive analysis of LLM-based evaluators' performance in detecting mixed-context hallucinations. Through carefully designed experiments, we investigate our research questions while addressing critical challenges in hallucination evaluation using LLMs, while also proposing potential solutions.

5.1 Datasets

To evaluate the performance of LLMs on mixed-context hallucination evaluation, we use our newly introduced dataset, FHSumBench, and incorporate an additional dataset, M-XSum (Maynez et al., 2020), for comparable experiments. The difference between these two datasets is that FHSumBench is built based on knowledge fact injection, which can examine LLMs' ability to leverage internal parametric knowledge and external knowledge. Notably, the M-XSum dataset is constructed through multiple small masked language models generating summaries followed by human annotation, where non-factual hallucinations account for the vast majority (92%), making a supplement of this category.

5.2 Implementing Details

LLM Selection We evaluate open-source models, including Llama (Touvron et al., 2023) and Qwen (Yang et al., 2024) families, across parameter sizes from 0.5B to 72B, alongside the closed-source GPT-4o, to examine how model capacity influences mixed-context hallucination evaluation. For retrieval-based judgment, we use Llama-Index[5] for retrieval embedding and retrieving the highest-scoring evidence at each step. LLM inference is carried out using vLLM [6] on 4 NVIDIA A100$^{\text{80GB}}$ GPUs.

Evaluation Settings For our retrieval-based approaches, each retrieved piece of evidence is assessed for its relevance (whether it is related to the claim) and supportiveness (whether it supports or contradicts the claim). The system then assigns a faithfulness score and a factuality score to each claim while maintaining a complete trace of the retrieval process. The individual claim scores are aggregated to evaluate the overall summary. We report the overall macro precision, recall, and F1-score for all methods as well as the accuracy of each category. The full results can be seen in Appendix B, while a detailed analysis of the representative models is presented in §5.3.

[5] https://github.com/run-llama/llama_index

[6] https://github.com/vllm-project/vllm

Methods		FHSumBench			M-XSum		
		P	R	F	P	R	F
Hybrid	FS*WC	0.3914	0.3073	0.3443	0.3522	0.3458	0.3490
	FS*AS	0.3870	0.3366	0.3601	0.3470	0.2690	0.3030
	FT*WC	0.3508	0.3604	0.3555	0.4037	0.3715	0.3869
	FT*AS	0.3705	0.4034	0.3863	0.4069	0.2959	0.3426
KR	Llama3-8B	0.3548	0.2929	0.3209	0.3575	0.2298	0.2798
	Qwen2.5-14B	0.4628	0.4740	0.4683	0.4035	0.4716	0.4349
	GPT-4o	0.4669	0.4763	0.4715	0.3801	0.4009	0.3902
CR	Llama3-8B	0.4461	0.4420	0.4441	0.4794	0.4470	0.4626
	Qwen2.5-14B	**0.5395**	0.4167	0.4702	0.4579	0.4590	0.4584
	GPT-4o	0.5026	0.4878	0.4951	0.4214	0.4367	0.4289
RR	Llama3-8B	0.4664	0.4615	0.4640	**0.4869**	0.4752	**0.4810**
	Qwen2.5-14B	0.5325	0.4148	0.4663	0.4766	**0.4787**	0.4776
	GPT-4o	0.5135	**0.4891**	**0.5010**	0.4413	0.4613	0.4511

Table 2: The results using retrieval-based LLM evaluators on FHSumBench and M-XSum datasets. We report the precision (P), recall (R) and F1-score (F) in this table. The best and second-best results are highlighted in **bold** and underlined, respectively. Hybrid scores include the combination of FactScore (FS), FacTool (FT), WeCheck (WC) and AlignScore (AS). "KR" denotes knowledge retrieval, "CR" denotes concurrent retrieval, and "RR" denotes reflection retrieval.

5.3 Results and Analysis

RQ1: How do the LLMs perform in hallucination evaluation? Table 1 presents the evaluation results on the FHSumBench and M-XSum datasets, comparing baselines and LLM direct generation methods. The random score on both datasets presents a lower bound for the evaluation. The performance of vanilla judge varies significantly across models. GPT-4o consistently outperforms the other two LLMs. However, even in the best case, vanilla LLM judges exhibit moderate F1-scores, indicating room for improvement.

By incorporating different prompting strategies, most models achieve improvement over the vanilla judge baseline. Specifically, under the +ICL setting, Llama3-8B achieves an F1-score improvement of 12.3% and 33.5%, respectively, highlighting that few-shot examples offer greater advantages to smaller LLMs by serving as a crucial guide to task comprehension. CoT leads to noticeable gains, particularly for Qwen2.5-14B, which achieves the highest F1-score (0.4733) on FHSumBench. However, on M-XSum, CoT slightly underperforms for larger models due to knowledge bias and hallucinations in reasoning (see §6 for details).

RQ2: How can LLMs better leverage external knowledge? Overall, concurrent retrieval and reflective retrieval methods demonstrate superior performance compared to the other two approaches. However, knowledge retrieval, which simply integrates all entity knowledge in the summary, still shows improvement compared to direct generation evaluation. The advantage of retrieval-based methods likely comes from providing explicit external knowledge, thereby reducing the LLM's reliance on potentially incomplete or incorrect intrinsic knowledge. For mixed-context evidence retrieval, allowing the LLM to perform step-by-step retrieval by identifying missing information during the retrieval process proves to be more effective. FactScore utilizes Wikipedia database for retrieving knowledge evidence. However, its retrieval efficiency is relatively low, with only 55.2% of entities in FHSumBench and 45% in M-XSum successfully retrieving relevant knowledge. This significantly influences the accuracy of the evaluation. To address this, we employ GPT-4o to generate knowledge for entities and substantially improve evaluation performance (detailed analysis available in Appendix C.3). FacTool relies on online search to retrieve evidence, resulting in more noisy information. Consequently, the quality of evidence interpretation significantly impacts the final evaluation results.

Compared to direct generation evaluation, Llama3-8B has the highest improvement over all the methods on both datasets, which indicates that smaller models can benefit more from evidence retrieval. We also conduct an ablation study on the experimental setup of reflection retrieval (see Table 3). The results indicate that using Wikidata (Vrandečić and Krötzsch, 2014) as the knowledge base faces the same challenge as FactScore, where the eval-

		P	R	F
Reflection Retrieve		0.4664	0.4615	0.4640
KB	Wiki	0.4200	0.3611	0.3129
Reflection	loop	0.4532	0.4325	0.4156
Top-k	k=3	0.4631	0.4613	0.4599

Table 3: Ablation study of the RR method on FHSumBench. "Wiki" indicates using Wikidata to retrieve knowledge evidence. "Loop" in reflection refers to iterative evidence retrieval until the LLM can provide a definitive True or False judgment. "Top-k" specifies the retrieval of k evidence chunks from the evidence pool.

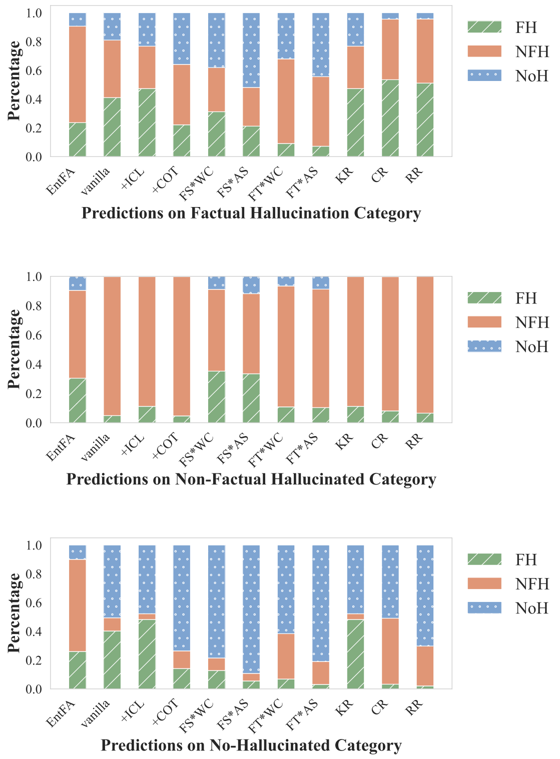


Figure 4: Percentages of the GPT-4o predictions on each category of FHSumBench.

uation results are heavily influenced by the entity matching rate. Additionally, we find that in this task, since both the document and knowledge are of moderate length, retrieving the top $k = 3$ results yields similar outcomes as retrieving only the top $k = 1$. Furthermore, we experiment with allowing the LLM to iteratively reflect until it retrieves sufficient evidence to draw a conclusion of either "support" or "contradict". However, we observe that during this process, the LLM exhibits a tendency to overcorrect its judgments.

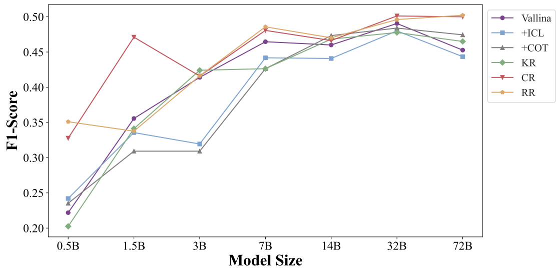


Figure 5: Model performance on F1-score with the increasing of model size in Qwen2.5 families.

RQ3: Which category dominates the results? We selected GPT-4o to analyze model predictions across three data categories (details of other models available in Appendix B). As shown in Figure 4, we report the prediction percentage for each category. Our analysis reveals a clear pattern in hallucination detection performance. For factual hallucinations, retrieval-based approaches demonstrate the highest accuracy, indicating that retrieving different contextual evidence helps detect mixed-context hallucinations. However, compared to a vanilla judge, CoT reasoning tends to classify more factual hallucinations as "no hallucination," suggesting that, in this case, LLMs are likely to deviate from predefined evaluation criteria during the reasoning process.

Additionally, most methods exhibit high sensitivity to non-factual hallucinations, indicating that unfaithful text is more likely to be classified as non-factual rather than factual. This behavior may be attributed to the fact that detecting clearly non-factual content is generally easier than identifying hallucinations that are factually accurate but unfaithful to the source.

While CoT reasoning significantly improves the accuracy of no-hallucination predictions, simpler approaches such as ICL show only marginal improvements over the vanilla judge. Furthermore, although the reflection retrieval judge demonstrates superior performance on non-hallucinated data compared to the concurrent retrieval judge, our comprehensive analysis indicates a notable trade-off: the improved accuracy in detecting factual hallucinations comes at the cost of decreased performance in the other two categories. The model size also plays a crucial role in evaluation stability, while larger LLMs show a more stable performance across the methods on each category (according to Figure 6,7,8 in Appendix B).

RQ4: Can scaling resolve the problem? Figure 5 demonstrates the performance of the best-performing reflection retrieve approach across the Qwen2.5 family models of varying sizes. Performance improves as the model size increases, with larger models (e.g., Qwen2.5-32B) achieving significantly higher overall accuracy, particularly in the factual hallucination category (see Table 7 in Appendix B), indicating that reflection with LLMs can help better identify and utilize the evidence in different context. However, the results from both Qwen2.5-72B and GPT-4o demonstrate that increased model size does not necessarily correlate with improved performance. We will further investigate the reason in §6.

6 Case Study

We choose the representative larger LLM, GPT-4o, as the case study model to investigate the reasons for the errors in hallucination detection (examples available in Appendix C.2).

Intrinsic Knowledge Bias of LLM Judge Since the CoT method exhibits the lowest score in detecting factual hallucinations using GPT-4o among direction generation LLM methods, we selected it to investigate the causes of error cases. We randomly selected and analyzed 30 error cases from the CoT outputs of GPT-4o. Among these cases, 63.3% were misclassified as "no hallucination" due to the LLM either relying on its intrinsic knowledge for reasoning or inherently assuming the injected information to be faithful. Additionally, 30% of the samples were mistakenly categorized as "non-factual hallucination" due to insufficient evidence being provided. The remaining 6.7% of errors resulted from issues in the reasoning process. These cases include instances where unfaithful content was correctly flagged as hallucination during the CoT process but was later misinterpreted as faithful in the final judgment. Conversely, some cases involved injected factual knowledge being accurately recognized during reasoning but ultimately misjudged as false in the factuality assessment. This indicates that, despite well-defined evaluation criteria, the model still struggles to some extent to distinguish between faithfulness and factuality.

The Influence of Retrieval Since the intrinsic knowledge of the LLM significantly influences its decision-making, retrieval-based experiments demonstrate that it can partially rectify erroneous classifications. To further investigate this, we examined the results of the CR and RR methods on GPT-4o using the aforementioned samples. Overall, evidence retrieval helps mitigate intrinsic knowledge bias. Among the analyzed error cases, the CR method successfully corrects 47.3% of samples misclassified as "no hallucination", while the RR method achieves a correction rate of 57.9%. However, the effectiveness of this approach still depends on the appropriateness of the knowledge base and the accuracy of retrieval. Although the RR method improves performance, there remains room for further enhancement.

7 Conclusion and Future Direction

In this paper, we systematically evaluate the ability of LLMs to assess hallucinations in mixed-context scenarios within summarization. The evaluation covers both open-source and closed-source models across various scales. Specifically, the distinction between factual and non-factual hallucinations is analyzed across different models and methods. Our findings reveal that LLMs still struggle to differentiate between faithfulness and factuality, highlighting the benefit of explicitly incorporating external knowledge to aid judgment. While larger models generally exhibit better performance, the trend suggests the existence of a performance plateau. Drawing insights from our experimental findings, several directions deserve further investigation:

Reducing Intrinsic Knowledge Bias Our experiments reveal that intrinsic knowledge bias significantly impacts LLM evaluation results, particularly for larger LLMs. While we have conducted preliminary investigations, deeper exploration is needed to understand fundamental aspects such as training data updates, internal information flow, and mechanisms for integrating internal and external knowledge.

Effective Knowledge Integration Mixed-context hallucination evaluation requires the fusion of multiple sources of information for decision-making. Our experiments demonstrate that specific information integration methods enable smaller LLMs to achieve performance comparable to larger LLMs. This highlights the urgent need for research into more effective information fusion approaches.

More Flexible Scenarios While this work primarily investigates mixed-context hallucination evaluation through the lens of summarization tasks, real-world scenarios often present greater complexity and flexibility. These include extended dialogues and more intricate task workflows that extend beyond knowledge-based hallucination. Further research is needed to address these diverse applications.

Limitation

Knowledge-Based Hallucination Construction While our approach of leveraging entity knowledge to construct factual and non-factual hallucinations is effective, it has certain limitations. This work does not address discourse-level or event-based hallucinations, leaving room for future exploration. As each instance is explicitly labeled, we rely on aggregate statistics rather than human evaluation, and do not analyze additional outputs from LLM evaluators beyond their predicted classifications. Moreover, given that news article contexts are factually accurate, we do not examine cases where summaries are faithful but not factual. Despite these constraints, our investigation approach remains practically applicable.

Benchmark Hacking Our benchmark primarily introduces hallucinations through appositive constructions, rather than simulating model-generated outputs. This approach may lead to relatively simple instances, due to the structural consistency of the hallucination format. While such simplicity might make the task appear easier, we observe that even under this controlled setting, current LLM-based evaluators exhibit notable limitations. An additional concern is that the consistent structure may encourage overfitting when models are trained specifically on these patterns. Nonetheless, this dataset is sufficient to evaluate the LLMs' capability to differentiate mixed-context hallucinations.

LLM-as-a-Judge Setting Although we extensively explored various common methods for employing LLMs as judges, our investigation was limited to text-based approaches. The potential of utilizing LLM prediction uncertainty for hallucination evaluation remains unexplored and presents a promising direction for future research.

Ethics Statement

Our study evaluates LLMs assessment of mixed-context hallucination using summarization datasets. We conduct experiments on both our constructed dataset and widely used datasets. Specifically, our dataset is derived from CNN/DailyMail and XSum, which contain news articles and summaries primarily in English. As such, our study focuses exclusively on English-language scenarios, which may limit the applicability to other languages and cultural contexts. Future research could incorporate more linguistically and culturally diverse datasets, and improve representation across different racial, ethnic, and cultural backgrounds, ultimately contributing to more equitable and globally relevant NLP models.

Acknowledgments

This work was supported by NetMind AI Gift Fund. SQ was funded by a PhD scholarship provided by CSC. YH was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2). The authors would also like to acknowledge King's Computational Research, Engineering and Technology Environment (CREATE) at King's College London.

References

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. pages 6325–6341, Singapore.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.10997 [cs].

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Matthew Honnibal and Ines Montani. 2018. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *Preprint*, arXiv:2405.14486.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv preprint*. ArXiv:2311.05232 [cs].

Zhaoheng Huang, Zhicheng Dou, Yutao Zhu, and Jirong Wen. 2024. Ufo: a unified and flexible framework for evaluating factuality of large language models. *arXiv preprint arXiv:2402.14690*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023a. MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. 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 39–53, Nusa Dua, Bali. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023b. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. pages 9004–9017, Singapore.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. pages 12076–12100, Singapore.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Siya Qi, Yulan He, and Zheng Yuan. 2024. Can we catch the elephant? a survey of the evolvement of hallucination evaluation on natural language generation. *Preprint*, arXiv:2404.12041.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

S. M. Towhidul Islam Tonmoy, S. M. Mehedi 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 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li, and Yajuan Lyu. 2023. WeCheck: Strong factual consistency checker via weakly supervised learning. pages 307–321, Toronto, Canada.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

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

A Supplementary Experiment Details

A.1 Dataset Details

Table 4 shows the data size, text length, and injected-fact length of each category in FHSumBench. Here we also present the meta-information of FHSumBench and previous related datasets on hallucinations of summarization evaluation. The relationship between different datasets is shown in Table 5.

	Factual Hallu	Non-Factual Hallu	No-Hallu	Overall
Data Size	501	410	425	1336
#Doc	433	448	471	449
#Sum	29	27	20	26
#Inject	6	4	-	5

Table 4: Meta information of FHSumBench, including the average word number per document (#Doc), summary (#Sum) and the injected fact (#Inject).

Constructed from XEnt and FactCollect by fact injection, FHSumBench comprises news articles from both XSum and CNN/DM, with summaries in CNN/DM being longer than those in XSum.

A.2 Previous Work Settings

Here we present the detailed descriptions and settings of the baselines in the experiments.

EntFA The EntFA (Cao et al., 2022) method aims to distinguish between factual hallucination and non-factual hallucination in abstractive summarization by evaluating both faithfulness and factuality as well. It performs evaluation at the entity level, utilizing both prior and posterior features of entities, which are extracted from a conditional masked language model. These two features are then used to train two separate KNN models to assess faithfulness and factuality individually. We label entities within the hallucination span as "hallucinated" and those outside the span as "non-hallucinated". The models then classify the faithfulness and factuality for each entity, which are used to determine the final categorization.

WeCheck WeCheck (Wu et al., 2023) is a consistency checker which verifies the alignment between generated summaries and their source documents. It employs weakly supervised learning to train a model that detects inconsistencies by leveraging large-scale noisy datasets and learning from multiple NLI models. The workflow includes extracting key facts from the source, comparing them with the summary, and scoring the consistency. The model is trained based on DeBERTaV3 (He et al.) to evaluate the faithfulness of the summaries.

AlignScore AlignScore (Zha et al., 2023) is a metric for evaluating the factual consistency of summaries by aligning information between the summary and the source document. It works by leveraging pretrained language models to compute semantic alignments and generate a consistency score that reflects how well the summary captures the source's key facts. The pretrained model used in AlignScore is RoBERTa (Liu, 2019) models (125M and 355M), which is fine-tuned on the task of faithfulness evaluation.

FactScore FactScore (Min et al., 2023) is an evaluation framework designed to assess the factual accuracy of text by assigning a score based on atomic fact veracity. It operates by breaking down content into claims, comparing them against knowledge bases. It uses a GPT-series model to extract claims and a Wikipedia knowledge base to retrieve evidence. We use GPT-4o to keep the consistency with our LLM-based experiments. Note that in FactScore, defining a topic for the input text is necessary. We identify entities in the summary as topics and then retrieve the corresponding evidence passages accordingly.

FacTool FacTool (Chern et al., 2023) is a tool designed for fact-checking to verify the factuality of statements by cross-referencing them with online search results. It supports multiple subtasks, among which we select the most relevant one, namely knowledge-based QA. FacTool operates by extracting claims, utilizing the Google Search API [7] to search the top pages and retrieve the most relevant search snippets, and then providing a confidence score for each validation. We use GPT-4o to extract claims and make judgments in our experiments to maintain the consistency of the setting.

A.3 LLM-Based Methods

Following FacTool and FactScore, claims for retrieval-based methods are generated by GPT-4o, with coreference resolution applied. The final classification is determined based on both the faithfulness score and the factuality score, where factual hallucination means unfaithful but factual, non-factual hallucination means unfaithful and unfactual, while no hallucination means both faithful

[7] https://serper.dev/

	Label Categories	Test Set Size	Data Source	Annotate Granularity
FHSumBench	factual hallucination, non-factual hallucination, no hallucination	1336	XEnt, FactCollect	summary
XEnt (Cao et al., 2022)	factual hallucination, non-factual hallucination, no hallucination	240	M-XSum	entity
FactCollect (Ribeiro et al., 2022)	hallucination, no hallucination	502	FactCC, M-XSum, QAGS, Frank	summary
M-XSum (Maynez et al., 2020)	intrinsic/extrinsic hallucination, factual/unfactual	912	XSum	span, summary
Frank (Pagnoni et al., 2021)	semantic frame errors, discourse errors, content verifiability errors	350	CNN/DM, XSum	span
QAGS (Wang et al., 2020)	hallucination, no hallucination	474	CNN/DM, XSum	summary
FactCC (Kryscinski et al., 2020)	hallucination, no hallucination	503	CNN/DM	claim

Table 5: The meta information and relationship between different datasets on faithfulness and factuality of summarization evaluation.

and factual. For direct generation, evaluators directly provide these scores. As mentioned in §5.2, retrieval-based methods aggregate results from individual claims. If any claim contains a non-factual hallucination, the summary is classified as a non-factual hallucination. Similarly, if any claim contains a factual hallucination, the summary is classified as a factual hallucination. Otherwise, if all claims remain faithful to the source document, it is classified as no hallucination.

B Full Results

Table 6 shows the full results for categories in FHSumBench and M-XSum, including factual hallucination accuracy (FH-Acc), non-factual hallucination accuracy (NFH-Acc), and no-hallucination accuracy (NoH-Acc). Figure 6,7,8 show the detailed prediction distribution on each category for Llama3-8B, Qwen2.5-14B and Qwen2.5-32B. Regarding factual hallucinations, all models demonstrate patterns consistent with GPT-4's results, indicating that CR and RR methods more accurately identify factual hallucinations. For Llama3-8B, all evaluation methods show decreased performance in detecting non-factual hallucinations. Furthermore, larger models like Qwen2.5-32B exhibit more consistent evaluation performance across all methods when assessing non-hallucinated content. Table 7 presents the full results for LLM-based methods on Qwen2.5 families models, where a darker background color indicates a higher score.

C Further Analysis

C.1 Dataset Preliminary Study

We conducted a preliminary study on FHSumBench to assess the feasibility of the data construction pipeline. Specifically, we manually examined 100 generated samples to assess the accuracy of entity-description matching. We found that some entities were linked to multiple conflicting descriptions, resulting in ambiguity. This issue arises when relying solely on entity matching to retrieve descriptions, as a single entity may correspond to multiple entries. To address this, we filtered out such ambiguous cases and retained only precise entity descriptions when constructing factual hallucinations. This design choice ensures that hallucinations are introduced in a controlled and interpretable manner, enabling reliable detection and elimination of problematic instances."

C.2 Case Study Examples

This section provides examples of the case study (§6), see Table 8, including the formatted generation of each method.

C.3 Checking on GPT-4o Generated Knowledge Base

In our attempt to integrate entity knowledge from the summaries, we first explored using Wikidata (Vrandečić and Krötzsch, 2014) as the knowledge base, leveraging entity aliases to expand the entries. However, the coverage proved insufficient, with only 35.9% of entities in the FHSubBench summaries matching Wikidata entries. This limitation is not unique to Wikidata, when using the Wikipedia database with FactScore, only 55.2% of entities in FHSumBench and 45% in M-XSum could be found with entity links in the knowledge graph. To overcome these coverage limitations, we developed an alternative approach using GPT-4o to generate a comprehensive knowledge base for

Methods		FHSumBench			M-XSum		
		FH-Acc	NFH-Acc	NoH-Acc	FH-Acc	NFH-Acc	NoH-Acc
Random		0.3167	0.3374	0.3012	0.3190	0.3171	-
EntFA		0.2362	0.6	0.0996	0.3214	0.7252	-
Vallina Judge	LLama-8B	0.1198	0.2756	0.7953	0.2154	0.3588	-
	Qwen-14b	0.4092	0.7902	0.6612	0.5429	0.7280	-
	GPT-4o	0.4052	0.9488	0.5059	0.5000	0.8397	-
+ICL	LLama-8B	0.2794	0.4463	0.7153	0.3000	0.5982	-
	Qwen-14b	0.4072	0.9000	0.4871	0.4714	0.8325	-
	GPT-4o	0.4711	0.8854	0.4776	**0.6857**	0.7031	-
+COT	LLama-8B	0.1956	0.3512	0.6706	0.4211	0.4852	-
	Qwen-14b	0.2595	0.8756	0.7647	0.3478	0.8245	-
	GPT-4o	0.2176	0.9439	0.7341	0.2571	0.8728	-
Hybrid	FS*WC	0.2535	0.3829	0.5929	0.2429	0.7945	-
	FS*AS	0.1756	0.3707	**0.8000**	0.2000	0.6069	-
	FT*WC	0.0878	0.7610	0.5929	0.2000	**0.9145**	-
	FT*AS	0.0699	0.7439	**0.8000**	0.1857	0.7019	-
KR	LLama-8B	0.0878	0.3122	0.7718	0.1667	0.5227	-
	Qwen-14b	0.3174	0.9244	0.6541	0.5857	0.8290	-
	GPT-4o	0.2495	**0.9732**	0.6824	0.3286	0.8741	-
CR	LLama-8B	**0.6128**	0.6756	0.2188	0.2273	0.6667	-
	Qwen-14b	0.4970	0.7927	0.3694	0.0588	0.8591	-
	GPT-4o	0.4770	0.8293	0.4776	0.0000	0.8733	-
RR	LLama-8B	0.5908	0.6878	0.1859	0.2069	0.7435	-
	Qwen-14b	0.4711	0.8146	0.3812	0.0789	0.8784	-
	GPT-4o	0.4393	0.8290	0.6538	0.0238	0.8988	-

Table 6: Full results for each category in FHSumBench and M-XSum, including factual hallucination accuracy (FH-Acc), non-factual hallucination accuracy (NFH-Acc), and no-hallucination accuracy (NoH-Acc).

the summary entities. Specifically, we employ the knowledge evidence generation prompt (detailed in Appendix D) to have GPT-4o create Wikipedia-like descriptions for all entities in the text to be evaluated, thereby constructing a knowledge evidence pool.

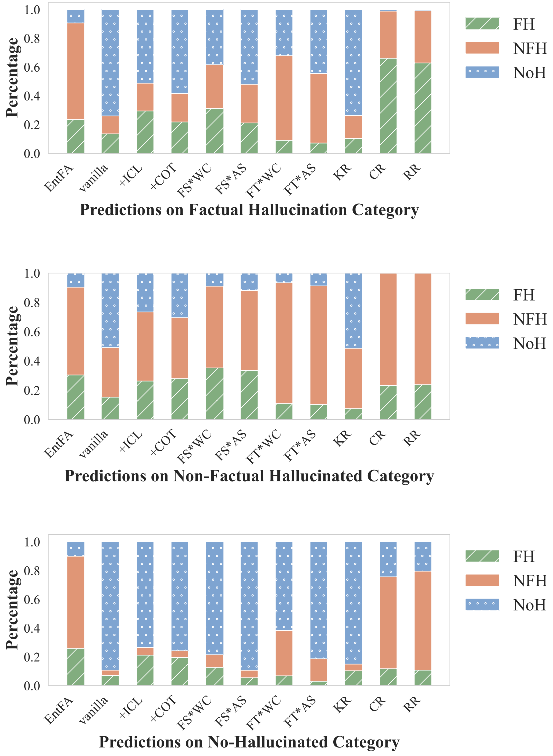
We have examined the performance of GPT-4o in generating Wikipedia-like knowledge for the entities found in the summaries. After examining 50 samples, we find that 94% of the generated knowledge is accurate. The remaining 6% contains errors due to the following reasons:

- Contextual ambiguity: The generated entity descriptions sometimes lack contextual information, because entities may have different meanings in different contexts.

- Hallucination: For certain entity types such as personal names, organization names, and business names, which may be out of vocabulary, the model occasionally fabricates information.

- Timeline issue: For past text content, the generated entity descriptions are based on the current situation. For example, in 2016, Donald Trump had not yet become president, but the generated knowledge states, "Donald Trump is the 45th President of the United States, serving from 2017 to 2021."

Despite the above issues, the knowledge generated by LLMs is still more informative than fixed databases like Wikipedia, making it more useful for evaluation.

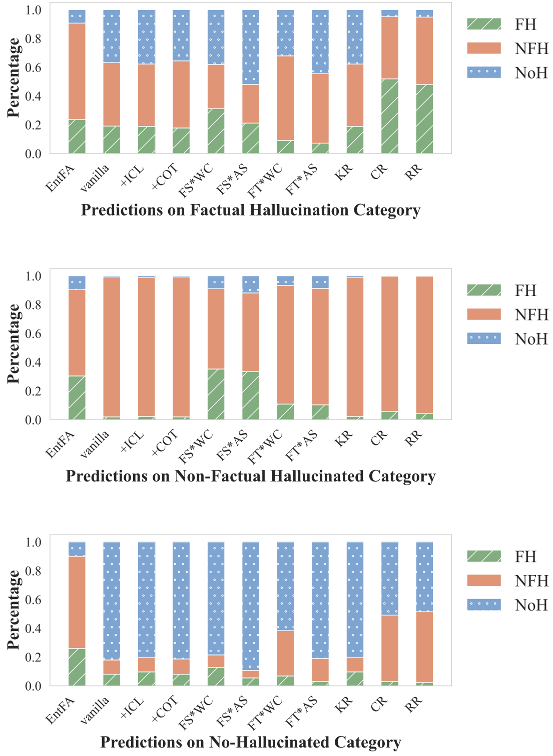


Predictions on Factual Hallucination Category

Predictions on Non-Factual Hallucinated Category

Predictions on No-Hallucinated Category

Figure 6: Predictions of Llama3-8B on categories of FHSumBench.

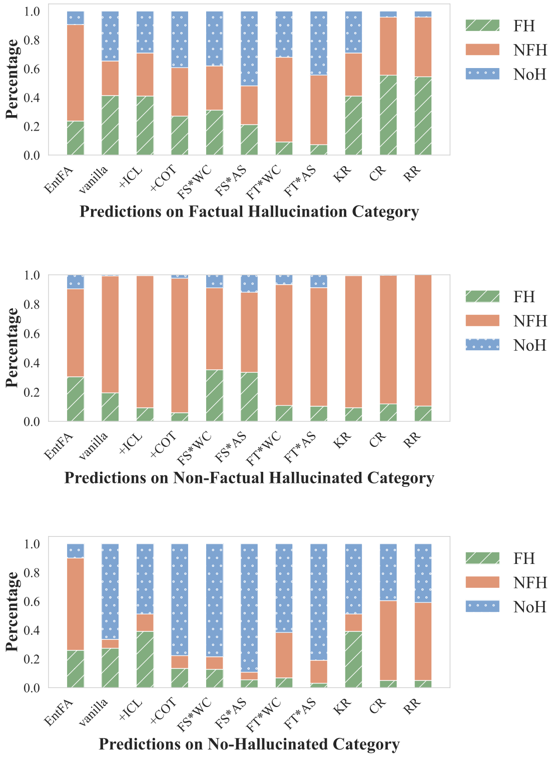


Predictions on Factual Hallucination Category

Predictions on Non-Factual Hallucinated Category

Predictions on No-Hallucinated Category

Figure 8: Predictions of Qwen2.5-32B on categories of FHSumBench.



Predictions on Factual Hallucination Category

Predictions on Non-Factual Hallucinated Category

Predictions on No-Hallucinated Category

Figure 7: Predictions of Qwen2.5-14B on categories of FHSumBench.

		FHSumBench					
Methods		**FH**	**NFH**	**NoH**		**Overall**	
		Acc	**Acc**	**Acc**	**P**	**R**	**F**
Vallina	Qwen2.5-0.5b	0.1437	0.1293	0.4941	0.2631	0.1918	0.2219
	Qwen2.5-1.5b	0.3214	0.5415	0.5388	0.3609	0.3504	0.3556
	Qwen2.5-3b	0.0419	1.0000	0.0118	0.5039	0.3512	0.4140
	Qwen2.5-7b	0.3373	0.9122	0.6329	0.4587	0.4706	0.4646
	Qwen2.5-14b	0.4092	0.7902	0.6612	0.4549	0.4652	0.4600
	Qwen2.5-32b	0.1896	0.9707	0.8165	0.4869	0.4942	0.4905
	Qwen2.5-72b	0.1836	0.9415	0.7600	0.4355	0.4713	0.4527
+ICL	Qwen2.5-0.5b	0.1517	0.2244	0.5035	0.2690	0.2199	0.2420
	Qwen2.5-1.5b	0.3613	0.5293	0.4071	0.3477	0.3244	0.3356
	Qwen2.5-3b	0.1597	0.9561	0.0918	0.3390	0.3019	0.3194
	Qwen2.5-7b	0.5329	0.7146	0.4894	0.4496	0.4342	0.4418
	Qwen2.5-14b	0.4072	0.9000	0.4871	0.4333	0.4486	0.4408
	Qwen2.5-32b	0.1856	0.9610	0.7976	0.4740	0.4861	0.4800
	Qwen2.5-72b	0.2874	0.8512	0.6871	0.4311	0.4564	0.4434
+COT	Qwen2.5-0.5b	0.2196	0.2098	0.4188	0.2643	0.2120	0.2353
	Qwen2.5-1.5b	0.1098	0.9049	0.0471	0.3704	0.2654	0.3092
	Qwen2.5-3b	0.1098	0.9049	0.0471	0.3704	0.2654	0.3092
	Qwen2.5-7b	0.3054	0.9073	0.4612	0.4345	0.4185	0.4263
	Qwen2.5-14b	0.2595	0.8756	0.7647	0.4717	0.4749	0.4733
	Qwen2.5-32b	0.1756	0.9634	0.8094	0.4808	0.4871	0.4840
	Qwen2.5-72b	0.1257	0.9098	0.8400	0.4801	0.4689	0.4744
KR	Qwen2.5-0.5b	0.1178	0.0902	0.4871	0.2427	0.1738	0.2025
	Qwen2.5-1.5b	0.2655	0.5098	0.5741	0.3452	0.3373	0.3412
	Qwen2.5-3b	0.0120	1.0000	0.0094	0.5622	0.3405	0.4241
	Qwen2.5-7b	0.2056	0.9244	0.6094	0.4182	0.4348	0.4263
	Qwen2.5-14b	0.3174	0.9244	0.6541	0.4628	0.4740	0.4683
	Qwen2.5-32b	0.1297	0.9878	0.8047	0.4750	0.4806	0.4777
	Qwen2.5-72b	0.1297	0.9659	0.8071	0.4549	0.4757	0.4650
CR	Qwen2.5-0.5b	0.3433	0.6927	0.0047	0.4425	0.2602	0.3277
	Qwen2.5-1.5b	0.0579	0.9854	0.0118	0.7134	0.3517	0.4711
	Qwen2.5-3b	0.4870	0.7439	0.1765	0.5085	0.3518	0.4159
	Qwen2.5-7b	0.5948	0.7122	0.4353	0.5362	0.4356	0.4807
	Qwen2.5-14b	0.4970	0.7927	0.3694	0.5325	0.4148	0.4663
	Qwen2.5-32b	0.4551	0.8683	0.4776	0.5650	0.4503	0.5012
	Qwen2.5-72b	0.5110	0.7976	0.4988	0.5596	0.4518	0.5000
RR	Qwen2.5-0.5b	0.4870	0.6415	0.0071	0.4598	0.2839	0.3510
	Qwen2.5-1.5b	0.0479	0.9780	0.0047	0.4895	0.2577	0.3376
	Qwen2.5-3b	0.4950	0.7366	0.2000	0.4970	0.3579	0.4161
	Qwen2.5-7b	0.5709	0.7341	0.4424	0.5470	0.4368	0.4858
	Qwen2.5-14b	0.4711	0.8146	0.3812	0.5395	0.4167	0.4702
	Qwen2.5-32b	0.4251	0.8902	0.4541	0.5648	0.4424	0.4961
	Qwen2.5-72b	0.4810	0.8073	0.5153	0.5666	0.4509	0.5022

Table 7: Full results for LLM-based methods on Qwen2.5 families models.

Text	**Content**
Document	Those who participated in the Aberdeen Children of the 1950s project, which saw all primary pupils aged seven to 12 surveyed by the Medical Research Council in 1962, have been contacted. They have been asked to take part in the Scottish Family Health Study, which is recruiting 50,000 people. It aims to investigate why diseases such as cancer can run in families. Those recruited will have their health tracked, with the intention of creating a Scottish b̈io-bankc̈ontaining genetic, medical and family history and lifestyle information. This will allow researchers to investigate the contribution of genetic and environmental factors to common conditions. Blair Smith, a GP and professor of primary care medicine at Aberdeen University, will run the project. He said: Ẅe know that many diseases such as cancer, heart disease, mental illness and diabetes run in families and that tells us that genes are important. B̈ut we don't know yet which genes are important or how they relate to other things such as lifestyle. Ẅe can get more information about genes by studying DNA which carries the genetic blueprint handed down through the generations. T̈hat is why we are collecting and analysing blood samples and health information from thousands of families across Scotland.Ḧe said the data gathered would help future research into the prevention, treatment and diagnosis of illnesses.
Summary	Aberdeen children from the 1950s have been invited to take part in a major study of family health in Scotland, home nation of the United Kingdom, as part of a new project.
Label	Factual hallucination ("home nation of the United Kingdom" is a correct description of "Scotland").
Method	**Response**

Vanilla Judge	{
	"Faithfulness": false,
	"Factuality": false
	}

+ICL	{
	"Faithfulness": true,
	"Factuality": true
	}

Continued on next page...

Continued from previous page...

Method	Response
+CoT	To evaluate the faithfulness and factuality of the summary in relation to the document, let's first examine the key points from each: 1. **Document Overview**: - Participants from the Aberdeen Children of the 1950s project are invited to the Scottish Family Health Study. - The study aims to recruit 50,000 people to investigate hereditary diseases. - The study will create a "bio-bank "including genetic, medical, history, and lifestyle information. - Led by GP and professor Blair Smith, the project will analyze DNA to understand genetic factors in diseases. - The broader aim is to aid future research in disease prevention, treatment, and diagnosis. 2. **Summary Overview**: - States that Aberdeen children from the 1950s are invited to join a family health study in Scotland. 3. **Analysis**: - **Faithfulness**: - The summary accurately conveys that individuals from the Aberdeen Children of the 1950s project are being contacted for a new study. - It specifically highlights the location (Scotland) and timing (from the 1950s project). - It does not mention the scope, goals, leader, or the extensive nature of the Scottish Family Health Study, which indicates missing information but not necessarily deviation or misrepresentation from what's present. - **Factuality**: - The statement regarding Aberdeen children from the 1950s and their involvement in a new project aligns with the document. - The phrase "home nation of the United Kingdom " is factual but irrelevant to the core details of the study, as the invitation pertains to a Scottish study regardless of broader geopolitical context. **Conclusion**: The summary does not contain any statements that deviate from or misrepresent the document (it is faithful to what it covers, but lacks completeness). It also does not present any factual inaccuracies; it simply omits details present in the document. Therefore, based on what's provided, the summary is both faithful and factual. { "Faithfulness": true, "Factuality": true } "

Knowledge
Retrieve
	{
	"Faithfulness": true,
	"Factuality": true
	}

Continued on next page...

Continued from previous page...

Method	**Response**
Concurrent Retrieve	queries: 1. "1950s Aberdeen children invited major study 2023", 2. "family health study Scotland site:.gov OR site:.ac.uk", 3. "Scotland home nation United Kingdom official status", 4."new project study evidence 2023 site:.edu” judgement: 1. claim: Aberdeen children from the 1950s have been invited to a major study., judge: faithful 2. claim: The study is about family health in Scotland., judge: faituful 3. claim: Scotland is a home nation of the United Kingdom., judge: factual hallucination 4. claim: The study is part of a new project., judge: unknown
Reflection Retrieve	queries: 1. "1950s Aberdeen children invited major study 2023", 2. "study on family health outcomes in Scotland 2023", 3. "Scotland status as home nation of United Kingdom evidence", (reflective query: "Scotlands political status within the United Kingdom") 4. "study is part of a new project " evidence 2023 academic research publication" judgement: 1. claim: Aberdeen children from the 1950s have been invited to a major study., judge: faithful 2. claim: The study is about family health in Scotland., judge: faithful 3. claim: Scotland is a home nation of the United Kingdom., judge: factual hallucination 4. claim: The study is part of a new project., judge: faithful

Table 8: Case Study.

D Prompts

The following are the prompts used in the experiments.

Vanilla Judge Prompt

Task Description: You are tasked with assessing the hallucination of a summary for a given document. Your role is to evaluate whether the summary:

1. Contains any information that deviates from or misrepresents the content of the document (Faithfulness).
2. Includes factual inaccuracies based on common knowledge or common sense (Factuality).

You will have access to the document to assist in your judgment. Provide your judgment in the following JSON format:

{
"Faithfulness": true/false,
"Factuality": true/false
}

Definitions:
1. Faithfulness: Determines if the information in the summary is directly inferred or logically entailed from the content of the document.
2. Factuality: Assesses whether the information in the summary aligns with commonly accepted facts, general knowledge, or common sense.

You should try your best to determine if the summary contains non-factual or hallucinated information according to the above hallucination types. Always provide answers in JSON format, start with "{", adhering strictly to the schema provided above. Do not include any explanations or extra content.

Document:{doc}
Summary:{summary}
Your Judgement:

ICL Judge Prompt (3-shot)

(... previous task description omitted ...)

Example 1:
Document: Two leading groups, Jaysh al-Islam and Ahrar al-Sham, which formed a pact last year, both say the plane was shot down. Syrian state media said the crash was caused by a technical fault......
Your Judgement:
{
"Faithfulness": false,
"Factuality": true
}

Example 2:
Document: These are external links and will open in a new window. The units in North Tyneside and Northumberland have been shut between midnight and 08:00 since December. Overnight emergencies have been diverted to the recently-opened Northumbria Hospital in Cramlington......
Your Judgement:
{
"Faithfulness": true,
"Factuality": true
}

Example 3:
Document: In a damning new report, the group also called for an ïndependent and impartialïnquiry into cases of abuse. The law, AFSPA, was introduced in the region in 1990 as a response to violence by insurgent groups......
Your Judgement:
{
"Faithfulness": false,
"Factuality": false
}

(... following task description omitted ...)

CoT Judge Prompt

(... previous task description omitted ...)

You should try your best to determine if the summary contains non-factual or hallucinated information according to the above hallucination types. Think step by step, and provide the trajectory before your judgement within 200 tokens. Always provide your final judgement in JSON format, start with "{", adhering strictly to the schema provided above.

Below are the prompts used in retrieval-based methods.

Claim Extraction Prompt

Task description: Break the given text into several independent claims, resolve the coreference. A claim is a statement that represents a clear and self-contained fact, opinion, or assertion, which can be verified by humans. Your task is to accurately identify and extract every claim stated in the provided text. Then, resolve any coreference (pronouns or other referring expressions) in the claim for clarity. Each claim should be concise (less than 15 words)Each claim should represent a clear and self-contained fact, opinion, or assertion. Split the claims with "\n". Start with the claims. Do not omit any information in the sentence. DO NOT RESPOND WITH ANYTHING ELSE.

Example 1:
Sentence: "This company offers high-quality products, and its customer service is highly regarded in the industry."
Claims:
This company offers high-quality products.
This company has customer service.
The customer service is highly regarded in the industry.

Example 2:
Sentence: "Former Wales captain Martyn Williams says Dan Biggar's decision to sign a new contract with Ospreys will benefit the region."
Claims:
Martyn Williams is a former Wales captain.
Martyn Williams says about Dan Biggar.
Dan Biggar has decided to sign a new contract with Ospreys
Martyn Williams believes this decision will benefit the Ospreys region.

Example 3:
Sentence: "This city not only has a rich cultural heritage but is also an important economic center that attracts a lot of investment."
Claims:
This city has a rich cultural heritage.
This city is an important economic center.
This important economic center attracts a lot of investment.

Now, please break down the following sentence:
Sentence: {summary}
Claims:

Knowledge Evidence Generation Prompt

Describe the following entity in a concise and informative manner. Include key characteristics, functions, and any relevant context that helps explain its significance or role. Aim for a clear and engaging description within a few sentences. Please only reply with the description, DO NOT include any extra content.

Entity: {entity}
Description:

Reflection on Evidence Prompt

What key information is missing from our current evidence to make a judgement about this claim? Please reply the missing information in the format of a sentence within 20 tokens. If there are multiple missing information, reply the most important one. DO NOT reply any other information.

Claim: {claim}
Current Evidence: {retrieved evidence list}
Missing Information:

Judge on Evidence Prompt

Analyze how the evidence relates to the claim. DO NOT make any assumption, reasoning or use previously owned knowledge EXCEPT the evidence. Start with the answer, DO NOT reply any other information. Answer in the following format, split with "\n":
Rationale: Give the rationale for the answer, within 30 tokens. Relevant means whether the evidence provides connection to the claim or contains the information of the claim. If Relevant is Yes, then make a judgement on relation.
Relevant: Yes / No
Relation: Support / Contradict

Example1:
Claim: New York is an eastern state in the United States of America.
Evidence: New York is known for its iconic landmarks like the Statue of Liberty and bustling New York City.
Answer:
Rationale: Even though the evidence is about New York, it does not directly support the claim that "New York is an eastern state in the United States of America." Instead, it mentions iconic landmarks and features of New York, such as the Statue of Liberty and New York City, without addressing its geographical location or status as an eastern state. Therefore, the evidence is not relevant to the claim.
Relevant: No
Relation: None

Example2:
Claim: Regular exercise improves mental health.
Evidence: Studies show that individuals who engage in physical activity report lower levels of stress and anxiety.
Answer:
Rationale: The evidence indicates a positive relationship between physical activity and reduced mental health issues, aligning with the claim.
Relevant: Yes
Relation: Support

Example3:
Claim: the pilot of the crashed jet killed himself.
Evidence: Jaysh al-Islam, the larger of the two groups, posted footage online which it claimed showed the pilot being held after ejecting from the jet. The video, bearing Jaysh al-Islam's logo, showed an object engulfed in flames followed by an interview with the supposed pilot.
Answer:
Rationale: The evidence addresses the topic of the pilot living status. The evidence suggests the pilot survived and was captured, which is inconsistent with the claim of suicide.
Relevant: Yes
Relation: Contradict

Now, please answer based on the claim and evidence:
Claim: {claim}
Evidence: {candidate}
Answer: