# From Single to Multi: How LLMs Hallucinate in Multi-Document Summarization

**Catarina G. Belem*** 
University of California Irvine
cbelem@uci.edu

**Pouya Pezeshkpour**
Megagon Labs
pouya@megagon.ai

**Hayate Iso**
Megagon Labs
hayate@megagon.ai

**Seiji Maekawa**
Megagon Labs
seiji@megagon.ai

**Nikita Bhutani**
Megagon Labs
nikita@megagon.ai

**Estevam Hruschka**
Megagon Labs
estevam@megagon.ai

## Abstract

Although many studies have investigated and reduced hallucinations in large language models (LLMs) for single-document tasks, research on hallucination in multi-document summarization (MDS) tasks remains largely unexplored. Specifically, it is unclear how the challenges arising from handling multiple documents (*e.g.*, repetition and diversity of information) affect models outputs. In this work, we investigate how hallucinations manifest in LLMs when summarizing topic-specific information from a set of documents. Since no benchmarks exist for investigating hallucinations in MDS, we leverage existing news and conversation datasets, annotated with topic-specific insights, to create two novel multi-document benchmarks. When evaluating 5 LLMs on our benchmarks, we observe that on average, up to 75% of the content in LLM-generated summary is hallucinated, with hallucinations more likely to occur towards the end of the summaries. Moreover, when summarizing non-existent topic-related information, `gpt-3.5-turbo` and `GPT-4o` still generate summaries about 79.35% and 44% of the time, raising concerns about their tendency to fabricate content. To better understand the characteristics of these hallucinations, we conduct a human evaluation of 700+ insights and discover that most errors stem from either failing to follow instructions or producing overly generic insights. Motivated by these observations, we investigate the efficacy of simple *post-hoc* baselines in mitigating hallucinations but find them only moderately effective. Our results underscore the need for more effective approaches to systematically mitigate hallucinations in MDS.

## 1 Introduction

Multi-document summarization (MDS) has numerous real-world applications, including planning treatments and diagnosing patients based on their medical history (*i.e.,* doctor notes, lab reports) (Tang et al., 2023b), forming legal arguments by linking precedents (Rodgers et al., 2023; Wu et al., 2023), or screening resumes to match candidates to job descriptions (Wang et al., 2024; Du et al., 2024). These tasks often involve linking information across lengthy documents, making the summarization process time- and labor-intensive (Van Veen et al., 2024). Recently, large language models (LLMs) have been proposed as more efficient approaches to reduce the human effort and increase scalability (Katz et al., 2023; Van Veen et al., 2024; Liu et al., 2024).

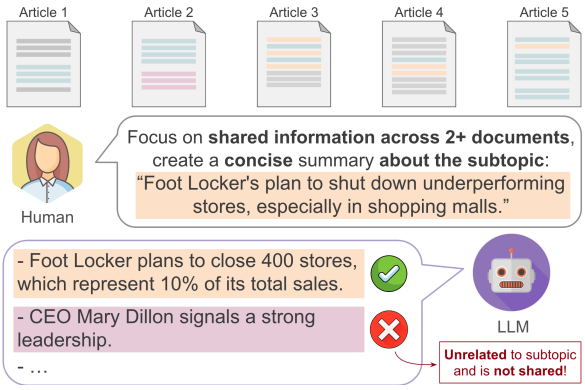



Figure 1: **An illustrative example of a summary generation from news articles**. Concerned about the credibility of the information, a human instructs the model to focus on shared, subtopic-related information. However, the LLM summarizes unrelated information that is not shared, raising concerns about the trustworthiness of LLMs in MDS.

However, despite the remarkable advances of LLMs in various tasks (Bubeck et al., 2023; Zhao et al., 2023), the frequent generation of ungrounded yet plausible-sounding text, referred to as "hallucinations", undermines trust in their output (Maynez et al., 2020; Uluoglakci and Temizel, 2024; Kalai and Vempala, 2024). While LLM *hallucinations* have been extensively studied in single-document tasks (Ji et al., 2023; Huang et al., 2025), resulting in the development of various evaluation bench-

* Work done while interning at Megagon Labs.



*Findings of the Association for Computational Linguistics:*
*NAACL 2025*, pages 5291–5324

marks (Lin et al., 2022; Wang et al., 2022; Yang et al., 2023) and taxonomies (Rawte et al., 2023; Zhang et al., 2023; Mishra et al., 2024; Dahl et al., 2024), little is known about how processing multiple documents affects the hallucinatory behavior of LLMs in MDS. Specifically, by focusing on the aggregated quality metrics of LLM outputs, including coverage (Fabbri et al., 2019; Lu et al., 2020; Laban et al., 2024) and faithfulness (Pu et al., 2023; Huang et al., 2024), existing work offers limited insight into how the inherent challenges of MDS (*e.g.,* repeated and diversity of information) correlate with LLM hallucinations. Consider the motivating example in Figure 1: when summarizing the repeated information from multiple documents, the LLM fails to link overlapping information across documents and satisfy the subtopic condition (highlighted text in the figure).

In this work, we comprehensively investigate the hallucinatory behavior of 5 popular LLMs in MDS tasks. Focusing on news articles and dialogues summarization, we examine the properties of hallucinations produced by LLMs when summarizing topic-specific information from multiple documents. Specifically, we aim to understand the patterns of hallucinations, their frequency, and their relationship with the number of input documents, as well as how these patterns change with the number of input documents and task focus (*e.g.,* summarizing undiscussed topics, or focusing the summary on repeated information). To this end, we design an evaluation protocol based on fine-grained annotations concerning relevant *insights*—units of information—within each document. We use the insight-level annotations from the SummHay datasets (Laban et al., 2024), designed for evaluating LLMs in long-context summarization, and create benchmarks with combinations of up to 10 documents in both the conversational and news domains. Our goal here is to leverage the insight-level annotations to automatically assess the correctness of the LLM-generated summaries.[1]

Our empirical evaluation of 5 prominent LLMs using the proposed evaluation protocol reveals that, up to 45% and 75% of the content in LLM-generated summaries is hallucinated in the news and conversation domain, respectively. We also observe that increasing the volume of input documents affects LLMs differently: for instance, as the document count increases from 2 to 10, most models experience only marginal changes in hallucinated content ($\pm$ 5%) whereas `gemini-1.5-flash` shows up to a 10% increase. Examining the composition of LLM-generated summaries, we find that regardless of the number of input documents and summary focus, hallucinations are more likely to appear in the later sections of the summaries. Furthermore, we find that both `GPT-4o` and `GPT-3.5-Turbo` show a strong predisposition (about 44% and 79.35%, respectively) to generate summaries even when there is no topic-specific insight in the input, raising concerns about their tendency to fabricate content.

To better understand the characteristics of these hallucinations, we manually evaluate 700+ insights spanning all 5 evaluated models and discover that most errors stem from either failing to follow instructions (*e.g.,* topic-unrelated information and/or redundant information) or producing overly generic insights (*e.g.,* paraphrases of the topic). Based on these observations, we investigate the effectiveness of five simple post-processing approaches, including both rule-based and LLM-based approaches, in mitigating hallucinations. We find these methods result in marginal improvements, on average: hallucinated content is decreased by up to 7% points but at the cost of excluding relevant information by up to 6%. Together, these results suggest that reducing hallucinations without compromising model performance remains challenging, underscoring the need for further research to better understand and systematically prevent LLM hallucinations in MDS.

## 2 Investigating Hallucination in MDS

Our goal is to shed light on the hallucination patterns in LLMs when addressing multi-document summarization (MDS). We begin by carefully defining the problem and its key assumptions. We then introduce two benchmarks, as well as the necessary evaluation metrics to discern between hallucinated and non-hallucinated LLM outputs.

### 2.1 Problem Formulation

We frame the MDS task as follows: given $N$ documents ($d_1, \ldots, d_N$), and $K$ conditions ($c_1, ..., c_K$), an LLM must generate a summary $\hat{\mathbf{y}}$ such that it satisfies all $K$ conditions and is grounded in the documents. Examples of conditions include matching specific topics, adhering to length constraints, or following particular writing styles. For simplic-

[1]The code and data have been made publicly available: https://github.com/megagonlabs/Hallucination_MDS.

ity, we assume that the $N$ documents in our formulation are of *the same type* (*e.g.,* all documents are news articles), rather than from diverse sources (*e.g.,* job descriptions and resumes, doctor notes and prescriptions). The summarization of multiple documents may involve capturing information that is either diverse (Huang et al., 2024), contradictory, or common across documents. The latter is especially important for combating misinformation, as it emphasizes "trustworthy" information supported by multiple sources. To analyze model behavior in both contexts, we require *fine-grained annotations* of the *information units* contained in each document.

### 2.2 Dataset Creation

Recently, Laban et al. (2024) introduced two MDS datasets with *insight-level annotations*.[2] The datasets cover two distinct domains—*news* and *conversation*—each containing 500 documents, evenly distributed across 5 topics, with documents averaging around 750 words. In particular, the news dataset (SummHay-News) is entity-centric and quantitative, featuring brands (*e.g.,* Tesla), banks (*e.g.,* JP Morgan), celebrities (*e.g.,* Elon Musk), along with numbers and dates. In contrast, the conversation dataset (SummHay-Conv) involves everyday scenarios with 2 to 4 participants, such as medical appointments and debates. Originally developed to evaluate LLMs in long-context tasks (*e.g.,* retriever-augmented generation, multi-document summarization), the documents were designed to ensure insights are repeated across at least 6 per topic and categorized into various subtopics, with no contradictory insights present.

By selecting SummHay-News and SummHay-Conv as testbeds, we can exploit their insight-level annotations to systematically assess model behavior across different settings, such as number of input documents and task focus. To control for input length (Li et al., 2024) while evaluating model behavior, we organize documents in the SummHay dataset into sets of $N$ documents, referred to as *combinations of* $N$ or $N$*-documents* (see Figure 7 in Appendix for an illustration of our benchmark creation process). Specifically, given a corpus with insight-level annotations and a subtopic $q$, we create our benchmarks by grouping documents into combinations of $N$ where multiple subtopic-related insights co-occur in two or more documents. Because we have the insight-level annotations, we can automatically identify the ground truth insights related to subtopic $q$ (called *reference insights*). To fully benefit from these fine-grained annotations, we instruct LLMs to succinctly[3] summarize the information as individual insights, presented in the form of bullet-point lists. We coin this setting "subtopic". To further investigate how the model handles common information, we introduce a "subtopic+trustworthy" setting by refining the prompt and restricting the reference insights to only those that are shared across documents.

With the intent of investigating the impact of document volume in the model's tendency to hallucinate, we employ the previous methodology to generate combinations of N for both SummHay-News and SummHay-Conv across five different combination sizes $N = \{2, 3, 4, 5, 10\}$. Since analyzing model behavior across all existing combinations-subtopic pairs is prohibitively expensive, we conduct our analysis on 500 randomly selected combinations for each $N$. Refer to Appendix A for additional information on the resulting benchmarks.

Note that while the proposed methodology applies to any MDS dataset with high-quality insight-level annotations, such information is rarely available. Therefore, we limit our evaluation to the SummHay dataset and leave the evaluation of other datasets, including non-English datasets and those with contradictory insights, for future work.

### 2.3 Automatic Evaluation

With reference insights for each combination, we can automatically evaluate the *correctness* of LLM outputs, distinguishing between hallucinatory and non-hallucinatory content. The correctness metric should assess whether predicted insights capture all essential information from the reference insights while avoiding the inclusion of irrelevant details (Huang et al., 2024). This ensures that predicted insights do not contain unsupported or potentially fabricated information.

Previous work uses an *LLM-as-a-judge* approach (Zheng et al., 2023) to assess whether reference insights are fully, partially, or not covered by the predicted insights, showing strong correlation with human evaluations (Laban et al., 2024).

[2]The authors define *insights* as "units of information".

[3]Unlike Laban et al., we do not instruct LLMs to limit their generation to a fixed number of bullet-points. Instead, we study their behavior "in the wild" when the exact number of insights is unknown. For further details related to prompts, refer to Appendix B.

However, this metric only measures reference coverage and ignores the validity of additional details in predictions. To enhance this, we assess how well predicted insights align with reference insights by applying the original metric twice with swapped inputs—once for reference insight and once for predicted insight——yielding two coverage labels per pair. We then combine these into a single correctness label using a conservative approach based on the order: `NO` $\prec$ `PARTIAL` $\prec$ `FULL`. By selecting the worst label (i.e., select `NO` over `PARTIAL` and select `PARTIAL` over `FULL`), we ensure the final assessment captures discrepancies related to missing relevant information or any potentially unfaithful content that is generated by the LLM.

### 2.4 Metrics

A key question in MDS tasks is whether LLM-generated summaries include all the correct information. To estimate average LLM correctness, we calculate *macro-recall* by determining the fraction of reference insights covered in each generated summary and averaging these scores across all summaries to obtain a single score. We also report the average fraction of hallucinated content in generated summaries by computing the average proportion of predicted insights that do not correspond to any reference insight across all summaries. This metric is referred to as the false discovery rate (*macro-FDR*). Both metrics range from 0 to 1, with a perfect LLM scoring 1 for macro-recall and 0 for macro-FDR. For simplicity, we refer to *macro-recall* as **recall** and *macro-FDR* as **hallucination rate**.

## 3 Experimental Settings

Before delving into the details of our investigation into LLM behavior in MDS, we briefly discuss key aspects of the conducted evaluation:

**Models.** We examine the capabilities of 5 popular LLMs, that span both open-source and closed-source models. As part of the open-source models, we evaluate the instruction-tuned version of `Llama 3.1 (70B)` (Meta, 2024) and `Qwen 2 (72B)` (Yang et al., 2024) models due to their competitive performance and instruction-following capabilities (Chiang et al., 2024; Fourrier et al., 2024; Galileo, 2024). As for the closed-source models, we assess OpenAI's `gpt-3.5-turbo-0125` and `gpt-4o-2024-05-13` models (OpenAI, 2024b), as well as Google's `gemini-1.5-flash` (Anil et al., 2024). More details are provided in Appendix C.

**Automatic Evaluation.** As previously mentioned, we adopt a few-shot *LLM-as-a-judge* approach. However, instead of using `gpt-4o` like Laban et al. (2024), we resort to a competitive yet more budget-friendly option—`gpt-4o-mini-2024-07-18`. The authors manually annotated 100 insight pairs and found strong alignment between both models, confirming the suitability of `gpt-4o-mini-2024-07-18` as an evaluator. Refer to Appendix D for more details.

**Metrics.** There is a mismatch between metrics: recall and hallucination use binary labels (correct/incorrect), but the LLM-based metric uses three coverage labels (`NO`, `PARTIAL`, `FULL`). To report recall and hallucination, we must map these three labels into correctness labels. We find the exact mapping has minimal impact (<5%) in the news domain but a larger effect ($\sim$30%) in the conversation domain. In the following sections, we opt for an optimistic approach, treating both partially and fully covered insights as correct (more details in Appendix E.1).

## 4 Experimental Results

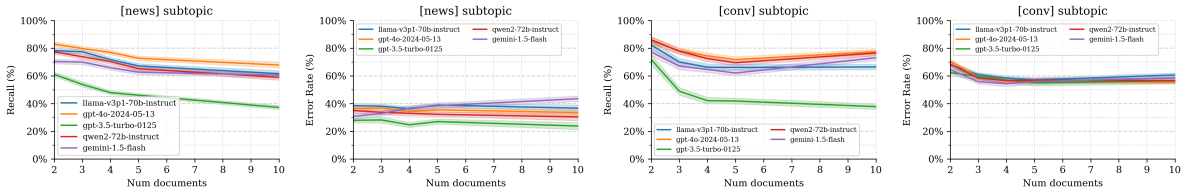
In this section, we first investigate the extent of LLMs hallucinations in MDS under the proposed "subtopic" and "subtopic+trustworthy" settings. Then, we analyze LLMs' hallucinatory behavior when synthesizing information on non-existent topics and investigate the link between hallucinated content and characteristics of both the input and output (we use the prompts listed in Appendix B).

### 4.1 LLMs hallucinate in MDS

To investigate LLMs hallucinatory behavior in MDS, we instruct each LLM using the "subtopic" prompt to generate a summary for 500 examples of the datasets created in Section 2.4. We then evaluate the recall and hallucination rate associated with each summary. Our main findings are as follows:[4]

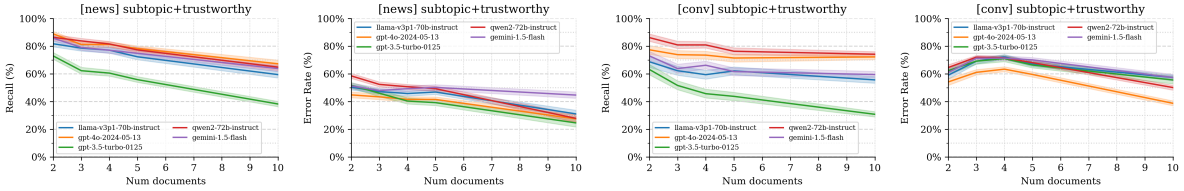
***LLMs exhibit substantial hallucination rates*** (see Figure 2). Across models and combination size ($N$), we observe an average hallucination rate greater or equal than 20% and 52% (and up to 45% and 75%) for the news and conversation domains, indicating that a non-negligible portion of LLM-generated content is hallucinated. Moreover, we observe that, ***hallucination rate changes only***

[4]Refer to Appendix E for additional results.




(a) Recall (news) (b) Error Rate (news) (c) Recall (conv) (d) Error Rate (conv)

Figure 2: **Performance metrics as a function of input documents counts in the "subtopic" setting**. Each line represents the mean value, with shaded areas indicating the 95% confidence intervals. Generally, recall drops significantly as document count increases, while average error rate changes only slightly across models and domains.




(a) Recall (news) (b) Error Rate (news) (c) Recall (conv) (d) Error Rate (conv)

Figure 3: **Performance metrics as a function of the number of input documents in the "subtopic+trustworthy" setting**. Each line represents the mean value, with shaded areas indicating the 95% confidence intervals.

***marginally with input size.*** Intuitively, increasing the number of documents introduces distracting information, which can affect the LLM-generated summaries—potentially, resulting in lower recall and higher error rates. Figure 2 shows that, although there is an overall downward trend in recall, the error rate remains almost constant ($\pm$5%), increasing only for `Gemini (Flash)` ($<$10%). Since we withhold the expected number of insights from LLMs, we hypothesize that the observed patterns stem from a mismatch between the number of predicted insights and reference insights. Upon further analysis (see Appendix E.2), we validate this pattern between observed recall drops and the ratio of predicted-to-reference insights.

***Models make more mistakes when summarizing conversations than news articles.*** Overall, we find a 20-30% hallucination rate disparity between the two domains. One possible explanation may be rooted in the dataset properties. In particular, the SummHay-News is entity-centric, discussing various concerns about common celebrities and companies (*e.g.,* Twitter), whereas the SummHay-Conv is more contextual discussing multi-turn everyday interactions between different participants. We indeed verify this is the case in practice, after manually inspecting 25 insights from each domain.

## 4.2 Shared Insights, Greater Errors

Compared to the previous "subtopic" setting, analyzing model behavior in "subtopic+trustworthy" requires adjusting the evaluation process and summarization prompt to focus only on shared insights. As before, we prompt models to summarize all 500 examples per benchmark and measure recall and hallucination rates. Our findings are as follows:

***LLMs generate shorter summaries but hallucinate more in general.*** To understand how responsive models are to the *shared instruction*, we compare summary lengths between "subtopic" and "subtopic+trustworthy" settings (see Appendix E.2). Summaries in "subtopic+trustworthy" are shorter, suggesting models respond to the *shared* instruction. However, shorter summaries do not guarantee quality: hallucination rates are, on average, higher in "subtopic+trustworthy" than "subtopic" (news: +10.47%, conv: +4.20%), indicating models may struggle to identify shared insights. We also observe that, ***LLMs struggle to identify subtopic-related shared insights.*** Compared to the "subtopic" scenario, we observe an average increase in summarization performance in news domain (+6.93%) but a slight decrease in the conversation (-2.91%). Overall recall trends remain the same: larger combination sizes lead to up to a 33% drop, with the sharpest declines in `GPT-3.5-Turbo`. Notably, `Qwen 2 (72B)` is mostly on par with, sometimes superior to, `GPT-4o` (58.9-86.5% *vs* 67.3-86.9%). However, unlike `GPT-4o`, `Qwen 2 (72B)` generates longer summaries and has a higher hallucination rate (see Appendix E.2).

## 4.3 Summarizing the Unsummarizable

Thus far, we have examined model behavior in well-defined scenarios where, by design, models summarize information that is known to exist in the input

documents. However, this assumption may not hold true in practice. For instance, when summarizing opinions (Angelidis et al., 2021a; Amplayo et al., 2021; Hosking et al., 2023), it is possible that models are instructed to synthesize information along specific aspects (*e.g.,* "product reliability" or "location of the resort") that were not mentioned.

Ideally, a reliable LLM would avoid generating summaries in such *adversarial* cases, but it remains uncertain whether the evaluated LLMs will be able to do so, especially given their sycophantic tendencies (Sharma et al., 2024; Rrv et al., 2024). To assess LLMs' ability to abstain from generating incorrect information, we modify 250 samples from the proposed benchmarks and pair them with a subtopic $q$ that, while still related to the theme, it is not explicitly discussed in the input documents. We instruct LLMs to output "No insights found" when relevant insights are absent and measure how often they *abstain* from summary generation.

***Models mistakenly generate summaries even if no relevant information is provided.*** Figure 4 shows that models actually generate summaries for non-existent subtopics: LLMs only abstain in 20.65% (`GPT-3.5-Turbo`, conversation) and up to 71.08% (`Llama 3.1 (70B)`, news). Moreover, ***as the document count increases, LLMs are more prone to mistakenly generate summaries.*** In general, our findings show a sharp decline in LLMs' capability to output "No insights found" as the number of input documents increases—this decline is especially pronounced in the conversation setting. Finally, we observe that open-source models tend to make *fewer mistakes* compared to proprietary models. Notably, `Llama 3.1 (70B)` outperforms other models (71.08% on average) and is least affected by input size (<8% drop). In contrast, OpenAI models show lower performance and greater sensitivity to the number of documents, incorrectly generating summaries in up to 44% and 79.35%.

### 4.4 Which documents do errors stem from?

So far, we have presented empirical evidence that models consistently hallucinate, irrespective of the number of input documents or task focus. In this section, we investigate the source of model hallucinations by examining how they correlate with the input documents, *i.e.,* we ask the question *where do hallucinations come from*?[5] We tackle this question by first determining which input document (if any) the hallucinated insight is copied from. In particular, we modify the string matching approach used to determine data contamination in the GPT-4 paper (OpenAI, 2024b), leaving the exploration of other techniques (Xu et al., 2024) for future work. Specifically, we assume that an hallucinated insight $i$ originates from a document $d$ if any 50-character string from $i$ matches a string in $d$.

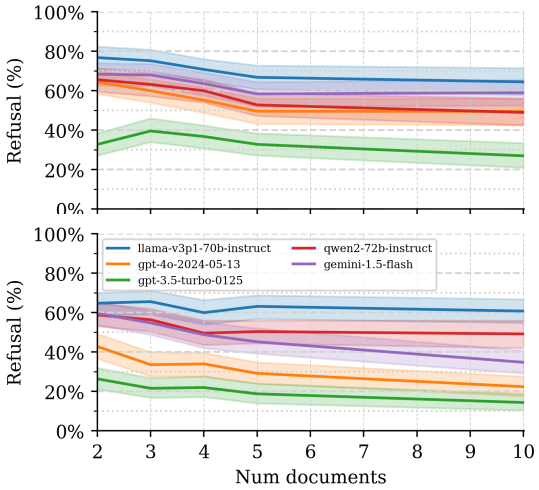



Figure 4: **Mean and 95% confidence intervals of summary refusal rate (%) for the news (top) and conversation (bottom) domains**. Notably, OpenAI models perform the worst, while `Llama 3.1 (70B)` consistently abstains from generating summaries over 60+% of the time, regardless of document count.

We find that ***3 out of 5 models exhibit a slight recency bias towards the last documents***. As observed in Figure 5, this recency bias is particularly prominent in `GPT-3.5-Turbo` and `Gemini (Flash)` in the news domain, but also present in `GPT-3.5-Turbo` and `Llama 3.1 (70B)` in the conversation domain. We hypothesize that such biases may emerge from the proximity to the list of instructions or due to the models' positional encoding, whose exploration we leave for future work.

### 4.5 Output order: A Proxy for Correctness?

Having explored how errors relate to the input, we now turn our attention to how hallucinations manifest in LLMs outputs. Given that models are instructed to produce summaries in bullet-point format, we can gauge the likelihood of the i-th bullet-point (*i.e.,* predicted insight) being hallucinated. To be more precise, we assess the relationship between the insights' position and their accuracy rate.

We observe that ***insights positioned earlier in the summary are more likely to be accurate than those located later*** (see Figure 6), consistent with

[5]Prior work has studied this question with respect the faithfulness score (Huang et al., 2024).

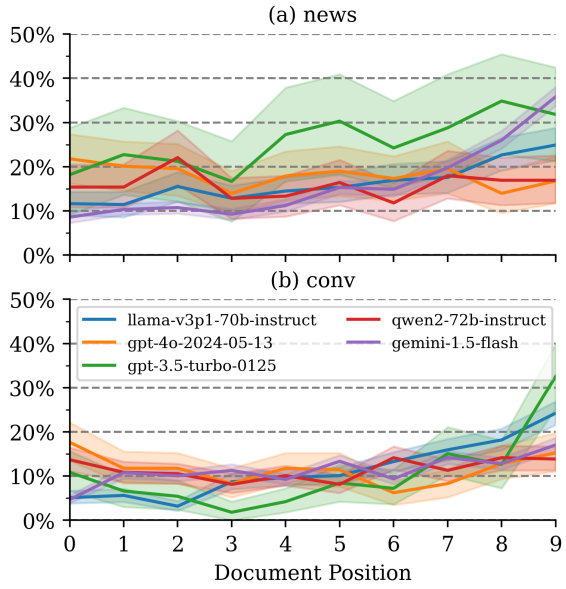



Figure 5: **Likelihood of an insight coming from a document (y-axis) based on its position in the input (when summarizing 10 documents)**. `GPT-3.5-Turbo`, `Llama 3.1 (70B)` and `Gemini (Flash)` mistakes seem to be more likely to originate from later documents on average than from earlier ones.

previous findings (Min et al., 2023; Chen et al., 2023a). In Appendix E.5, we show that this pattern holds across all summarization models, regardless of input document count. Interestingly, the opposite pattern appears in the top three summary positions. Manual inspection reveals that lower accuracy at the start is due to the generation of broader scope insights (*e.g.,* opening remarks or preliminary statements). We also find that `GPT-4o` and `Gemini (Flash)` tend to generate 1 or 2 "takeaway" or "concluding" insights at the end of the summary to emphasize the main point or lesson to remember. We hypothesize this structure may stem from discourse coherence (Jurafsky and Martin, 2024; Zhu et al., 2024) or the models' ability to mimic input document patterns. Future research could explore these possibilities further.

## 5 What type of mistakes do LLMs make?

Understanding the nature and frequency of LLM hallucinations is crucial for evaluating the reliability of these models in practical applications. While prior work acknowledges failures in LLM outputs (Laban et al., 2024), these failures are not well-studied in MDS. To bridge this gap, we collect human annotations for 150+ LLM-generated summaries and proposed a taxonomy based on the recurring mistakes observed. Specifically, we manu-

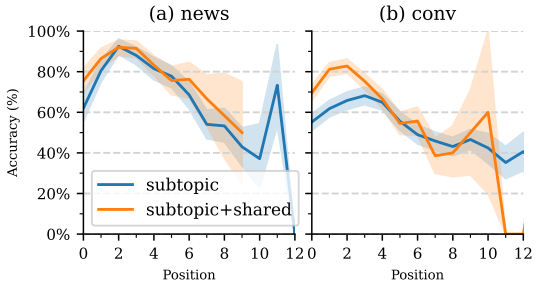



Figure 6: **Accuracy rate of `GPT-4o` generated insights by position (when summarizing 10 input documents)**. Each solid line shows the mean, with shaded areas representing 95% confidence intervals. Overall, accuracy rate declines as insight position increases.

ally inspect over 700 predicted insights across models, domains, and task focus. To assess whether predicted insights are *faithful*, we also include the input documents in the annotation process. However, as previously observed (Chang et al., 2023), analyzing long documents is quite challenging. Consequently, instead of analyzing multiple combination sizes, we limit our analysis to N=2, leaving the analysis of 2+ documents for future work. The main analysis is conducted by two authors of this paper (see Appendix F for further details on the annotation protocol). Finally, to ensure the robustness of the proposed error categories, we collect the category annotations of two additional authors across 50 examples spanning both domains for each category (Chang et al., 2023). We discover that, on average, 80.71% of the annotated insights are assigned the same label by 3 or more annotators.

We identify three common LLM errors, detailed in Table 1: (1) focusing on high-level details (*pedantic*), (2) failing to follow instructions (*instruction inconsistency*), and (3) misrepresenting specific details (*context inconsistency*). Occasionally, though less frequently, we also observe insights that deviate from the original input (*fabrication*). During the annotation process, we also encounter examples that fall into multiple categories, such as an overly generic insight (pedantic) related to a different subtopic (instruction inconsistency), suggesting the complexity associated with the errors observed in MDS settings.

Table 2 shows each error's frequency in the news domain. Overall, we find all LLMs to be fairly faithful to input documents, as emphasized by the low rates of context inconsistency (9%-37%) and fabrication errors (0%-9%). The majority of the errors seems to be related to the generation of uninfor-

Table 1: **Definition of the hallucination types found in LLM-generated summaries across both domains**.

| Type | Definition |
|---|---|
| Pedantic | Insights that, while correct, add no new information. These include paraphrasing the subtopic, contextual statements (*e.g.,* opening remarks, takeaway), as well as insights that are overly generic (and, thus, non-informative), or overly specific (*e.g.,* explanations, particular background information). |
| Instruction Inconsistency | Insights that are unrelated to the subtopic, redundant, or not shared, violating the conditions specified in the prompt: (1) focus on a subtopic, (2) create concise summaries, and, in the "subtopic+trustworthy" setting, (3) focus on shared information. |
| Context Inconsistency | Insights that are subtly misrepresented, either by oversimplifying or exaggerating details. Examples include: *overgeneralization*, where individual opinions are turned into broader claims (*e.g.,* individual's opinion reported in the predicted insight as a generic claim), or *oversimplification*, where the model narrows details to limited contexts inappropriately (*e.g.,* when a document mentions both banks and government agencies, but the predicted insight only focuses on a specific bank). |
| Fabrication | Insights that contradict or are not supported by the information presented in the provided documents. In practice, fabrications are hard to identify and often manifest as slight changes in wording. |

Table 2: **Observed hallucinations (and their frequency) when summarizing news articles for different task focus ("subtopic" and "subtopic+trustworthy")**. Reported values represent the fraction of error mistakes that are categorized as either **P**edantic, **I**nstruction Inconsistency, **C**ontext Inconsistency, and **F**abrication. We also report the **T**otal number of errors for each LLM.

| | Model | P | I | C | F | T |
|---|---|---|---|---|---|---|
| subtopic | GPT-3.5 | 51.51 | 38.71 | 19.35 | 0.00 | 31 |
| | GPT-4o | 61.88 | 52.73 | 21.81 | 0.00 | 55 |
| | Gemini | 78.72 | 23.40 | 10.64 | 8.51 | 47 |
| | Llama 3.1 | 60.00 | 72.00 | 16.00 | 4.00 | 25 |
| | Qwen 2 | 53.33 | 86.67 | 36.37 | 3.33 | 60 |
| shared | GPT-3.5 | 38.46 | 75.00 | 12.82 | 0.00 | 32 |
| | GPT-4o | 28.15 | 79.49 | 9.30 | 0.00 | 39 |
| | Gemini | 52.73 | 70.91 | 9.09 | 0.00 | 55 |
| | Llama 3.1 | 29.27 | 70.73 | 17.07 | 2.4 | 41 |
| | Qwen 2 | 15.51 | 86.20 | 17.24 | 1.7 | 58 |

mative insights (pedantic) and unrelated subtopics (instruction inconsistency). For instance, GPT-4o and Qwen 2 (72B) often include redundant and off-topic insights, while also adding coherence-enhancing insights such as takeaway statements. In "subtopic+trustworthy" results, the rate of instruction inconsistency errors exceeds 70% of all hallucinations. Of these, 80% to 95.45% (58.54% to 67.24% of all hallucinations) stem from insights not shared across documents, as required by the prompt.

## 6 Mitigating Hallucination

With a clearer understanding of LLMs hallucinations, we now explore whether we can mitigate them through simple post-processing heuristics. Focusing on LLMs exhibiting higher rate of instruction inconsistent and pedantic errors, we investigate whether simple heuristics suffice to reduce incorrect insights (lower error rate) while preserving correct ones (maintaining recall). In particular, we explore 4 mitigation methods that include output truncation (`top-k`), removal of redundant insights (`redundant`), subtopic paraphrases (`st-paraphrase`), and subtopic-unrelated (`st-unrelated`) (for technical details, see Appendix G). We apply each mitigation method to previously generated summaries in the 2-document "subtopic" setting and report the absolute change in average F1-score.

Across mitigation methods, F1-score variation is minimal ($\pm 3\%$), with a simple *top-k* showing the most improvement (see Table 3). Additionally, `st-unrelated` and `redundant` have less impact than expected, despite targeting common errors in generated summaries. This may be due to poor performance of adopted LLM-based classifiers in the tested domains. Even with a narrow focus ("subtopic", N=2), our findings highlight the complexity of reducing hallucinated errors in LLM-generated summaries within a MDS setting, calling for further research to address these challenges. To minimize hallucination in MDS, future work could explore hierarchical approaches, where each document is summarized individually before combining the results into a final summary (Chang et al., 2023; Tang et al., 2024).

## 7 Related Work

***Multi-document summarization*** is a broad and versatile multi-document NLP task that consists of generating a summary from multiple source documents, including opinion/reviews (Angelidis et al., 2021b; Iso et al., 2022), scientific articles (Lu et al.,

Table 3: **Absolute difference in average F1-score after applying four simple mitigation methods to summaries generated from two input documents (N=2)**. All methods show minimal impact on average F1-score (±3%), with truncating summaries to the top 5 bullet-points being the most effective.

| | Strategy | GPT-4o | Llama 3.1 | Qwen 2 |
|---|---|---|---|---|
| news | top-5 | 2.51% | 1.69% | 0.42% |
| | st-unrelated | -2.61% | -1.49% | -1.95% |
| | st-paraphrase | -1.19% | -1.19% | -0.86% |
| | redundant | -0.49% | -0.28% | -0.98% |
| conv | top-5 | 2.28% | 1.52% | 1.52% |
| | st-unrelated | 0.85% | 0.85% | 0.43% |
| | st-paraphrase | -0.11% | -0.46% | -0.29% |
| | redundant | 0.46% | -0.08% | 0.18% |

2020; DeYoung et al., 2021; Yang et al., 2023), or news articles (Fabbri et al., 2019). Assessments of LLMs capabilities can be carried using generic summarization benchmarks (Fabbri et al. (2019); Lu et al. (2020); DeYoung et al. (2021); Wolhandler et al. (2022); Huang et al. (2024), *inter alia*) or using focus-specific benchmarks, that resort to aspects (Angelidis et al. (2021b); Hayashi et al. (2021); Amar et al. (2023); Yang et al. (2023), *inter alia*), queries (Kulkarni et al. (2020); Bolotova-Baranova et al. (2023); Huang et al. (2024); Chen et al. (2024); Laban et al. (2024), *inter alia*), or perspectives (Naik et al., 2024). However, these benchmarks mainly focus on overall performance metrics and rarely conduct in-depth error analysis, offering limited insight into LLMs behavior and its relation to the input.

***Hallucination evaluation benchmarks*** have been proposed for single-document NLP tasks with the purpose of evaluating models' propensity to generate hallucinated content (Huang et al., 2025). These benchmarks are often designed to probe model errors, for instance, by testing knowledge of false beliefs and misconceptions (Lin et al., 2022; Cheng et al., 2023), current events (Vu et al., 2024; Kasai et al., 2023), similar but incorrect statements (Muhlgay et al., 2024), or even a models' ability to recover from nonsensical questions (Pal et al., 2023). While single-document hallucination benchmarks have been crucial in identifying and mitigating non-factual outputs in current LLMs (Touvron et al., 2023; Yang et al., 2024; Achiam et al., 2024; Wei et al., 2022), hallucinatory behavior in multi-document tasks remains under-explored, especially how it varies with factors like document count or repeated or contradictory information.

***Automatic evaluation***, while less reliable than human evaluation (Kryscinski et al., 2019; Fabbri et al., 2021), is often sought for its scalability and efficiency in large-scale experiments, where human evaluation is impractical (Krishna et al., 2023; Chang et al., 2023; Huang et al., 2024). To evaluate the grounding of LLM-generated text, specialized entailment-based (Laban et al., 2022; Tang et al., 2024; Goyal and Durrett, 2021) and question-answering-based (Fabbri et al., 2022) metrics have been proposed. Recent work has found that, despite their higher cost, zero-shot prompting general-purpose LLMs performs as well as or better than specialized metrics when used for fact-checking while requiring no fine-tuning(Manakul et al., 2023; Tang et al., 2024).[6] Building on prior successes of LLM-as-a-judge approaches, we use an LLM-based metric with three demonstration examples (Chang et al., 2023; Laban et al., 2024) to determine the information coverage between reference and predicted insights. While a variant of this metric has been rigorously validated (Laban et al., 2024), we further confirm its reliability through human evaluations (see Sections 3 and 5), finding strong alignment with LLM judgments.

## 8 Conclusion

In this paper, we carry the first in-depth investigation of hallucinatory behavior of popular instruction-tuned LLMs in a multi-document summarization task. Through controlled experiments where we vary the number of documents and task focus, we find that up to 75% of LLM-generated content is hallucinated. More surprisingly, our results show that when prompting models to summarize information related to non-existing subtopics, models still generate plausible-sounding summaries in more than 20% of the examples, with `GPT-3.5-Turbo` incorrectly generating summaries in 79.35% of the samples. Subsequent manual analysis of 700+ insights reveals that, albeit *faithful* to the input, a large fraction of insights is overly generic or inconsistent with the instructions. To mitigate such errors, we experiment with simple *post-processing* mitigation strategies but observe a trade-off with models' performances. Together, our results underscore the need for more effective approaches that systematically mitigate hallucinations in MDS.

[6] Similarly, prompting LLMs (with task specific prompts) has been shown to correlate better with human judgments than specialized models (Liu et al., 2023a; Farquhar et al., 2024).

## Limitations

While we design our evaluation protocol to ensure the reliability of our analysis, we acknowledge some limitations with our work.

*Grounding evaluation in reference insights.* To automate evaluation while keeping costs manageable, we compare LLM-predicted insights with the reference insights instead of using the documents. This assumes that the reference insights cover all relevant document content. This assumption is supported by the fact that SummHay documents were artificially generated from the reference insights and rigorously validated for completeness (Laban et al., 2024). However, this simplification could result in misclassifying a predicted insight as hallucinated, even if it is covered in the document. Empirically, across 700+ insights, we find that most "false hallucination" results from models merging two high-level ideas, rather than omitting important information. Alternatively, future work could explore document-level (Yin et al., 2021), sentence-level (Huang et al., 2025), or fact-level (Tang et al., 2024) approaches to directly cross-check predicted insights with the documents themselves, rather than with reference insights.

*Insights may not represent a single unit of information.* Underlying our evaluation protocol was the assumption that insights—both predicted and reference—represent a unique unit of information and "are expected to mention a number, a date, or an entity" and "are independent of each other" (Laban et al., 2024). However, we observe that the assumption does not always hold in practice (see examples in Appendix A), which has repercussions for the evaluation metric, which is produces a 1-to-1 coverage mapping between lists of insights. To address this, one solution could be to modify the evaluation prompt (listed in Figure 15) to output a list of "coverage" judgments (as opposed to a single JSON object). However doing so may introduce additional challenges for evaluation, such as anchoring effects for multi-attribute judgments (Stureborg et al., 2024). Alternatively, we could decompose the reference and predicted insights into smaller units of information (Min et al., 2023; Wanner et al., 2024) and apply the metric on decomposed facts.

*The use of information coverage as a proxy for measuring similarity of meaning.* We acknowledge the pragmatic and semantic nuances involved in determining whether two insights have the same meaning. Linguistic phenomena, such as entailment, implicature, and presupposition (**?**Jeretic et al., 2020; Jiang and de Marneffe, 2021, 2019) could prove useful in determining the semantic relationships between two sentences (Goyal and Durrett, 2020). Alternatively, researchers have also explored other approaches (Tang et al., 2023a), based on iterative question generation and response (Wang et al., 2020; Scialom et al., 2021) or even based on prompting LLM using a few demonstrations (Chang et al., 2023; Laban et al., 2024). In this work, similarly to previous work (Huang et al., 2024; Laban et al., 2024), we use the more abstract notion of *coverage* (as opposed to other linguistic phenomena) due to its previous reports of high agreement with both experts and crowdsourcing annotators for abstractive summarization tasks.

*Hallucination Mitigation Trade-off.* In this paper, we explore the efficacy of simple heuristics in the mitigation of hallucinations across LLMs, which were designed to tackle specific issues with the LLM-generated summaries, including redundancy (Xiao and Carenini, 2020). However, we observe limited effect in LLMs' hallucination rate in practice, raising important questions about the efficacy of these methods. One potential explanation is that the LLM-based classifiers, used to detect these errors, perform poorly in the specific domains tested in this work. Moreover, we observe that most insights (both reference and predicted) are rarely atomic, containing a mixture of redundant and non-redundant pieces of information, making it difficult to detect using binary judgments. Future work could explore the use of more fine-grained atomic evaluation methodologies (Min et al., 2023; Chen et al., 2023b; Tang et al., 2024).

## Acknowledgments

We thank the members of Megagon Labs and UCI NLP for their valuable feedback and fruitful discussions.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, et al. 2024. GPT4 Technical Report. *Preprint*, arXiv:2303.08774.

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. OpenAsp: A benchmark for multi-document open aspect-based summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*,

pages 1967–1991, Singapore. Association for Computational Linguistics.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021a. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021b. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Anthropic. 2024. Introducing the next generation of claude.

Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023. WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5314, Toronto, Canada. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. 2023a. Understanding retrieval augmentation for long-form question answering. *Preprint*, arXiv:2310.12150.

Peter Baile Chen, Yi Zhang, Chunwei Liu, Sejal Gupta, Yoon Kim, and Michael Cafarella. 2024. Mdcr: A dataset for multi-document conditional reasoning.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023b. FELM: benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yingpeng Du, Di Luo, Rui Yan, Xiaopei Wang, Hongzhi Liu, Hengshu Zhu, Yang Song, and Jie Zhang. 2024. Enhancing job recommendation through llm-based generative adversarial networks. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 8363–8371. AAAI Press.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. 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.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Galileo. 2024. Llm hallucination index: Rag special.

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.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Xiaobo Guo and Soroush Vosoughi. 2023. Length does matter: Summary length can bias summarization metrics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15869–15879, Singapore. Association for Computational Linguistics.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

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 Comput. Surv.*, 55(12).

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213, Florence, Italy. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2021. He thinks he knows better than the doctors: BERT for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 20, 2024.

Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, page 160–171, New York, NY, USA. Association for Computing Machinery.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: what's the answer right now? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. 2023. Natural language processing in the legal domain. *SSRN Electronic Journal*.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yong Liu, Shenggen Ju, and Junfeng Wang. 2024. Exploring the potential of chatgpt in medical dialogue summarization: a study on consistency with human preferences. *BMC Medical Informatics and Decision Making*, 24(1).

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark 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*, pages 9004–9017, Singapore. Association for Computational Linguistics.

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.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem. Med. (Zagreb)*, 22(3):276–282.

Meta. 2024. Introducing llama 3.1: Our most capable models to date.

Andrianos Michail, Simon Clematide, and Juri Opitz. 2024. Paraphrasus : A comprehensive benchmark for evaluating paraphrase detection models.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta. Association for Computational Linguistics.

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! perspective-aware healthcare answer summarization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15919–15932,

Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Hello gpt-4o.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *ArXiv preprint*, abs/2309.09558.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Ian Rodgers, John Armour, and Mari Sako. 2023. How technology is (or is not) transforming law firms. *Annual Review of Law and Social Science*, 19(1):299–317.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12717–12733, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023a. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023b. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Cem Uluoglakci and Tugba Temizel. 2024. HypoTermQA: Hypothetical terms dataset for benchmarking hallucination tendency of LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 95–136, St. Julian's, Malta. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny

Zhou, Quoc Le, and Thang Luong. 2024. Fresh-LLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13697–13720, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin Jr. au2, and Maria Perez-Ortiz. 2024. Jobfair: A framework for benchmarking gender hiring bias in large language models.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How "multi" is multi-document summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5761–5769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 516–528, Suzhou, China. Association for Computational Linguistics.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey.

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.

Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023. OASum: Large-scale open domain aspect-based summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4381–4401, Toronto, Canada. Association for Computational Linguistics.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

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. 2023. A survey of large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*

*36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Dawei Zhu, Wenhao Wu, Yifan Song, Fangwei Zhu, Ziqiang Cao, and Sujian Li. 2024. CoUDA: Coherence evaluation via unified data augmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 967–978, Mexico City, Mexico. Association for Computational Linguistics.

## A Dataset Details

This section describes the dataset card and the access to the the datasets used. We also report the statistics of the benchmarks we created and that we base our analysis on.

### A.1 Original Datasets

To the best of our knowledge, there is still no MDS benchmark focused on the investigation of hallucinations. To this end, we leverage an existing dataset—SummHay(Laban et al., 2024)—that provides fine-grained level annotations about the relevant information presented in each insight. Originally proposed to evaluated LLMs' performance in long-context summarization, SummHay puts forward two synthetic datasets spanning different domains: news and conversation, which we refer to as SummHay-News and SummHay-Conv, respectively. Each dataset comprises 500 documents, uniformly distributed across 5 different topics (see list of topics in Table 4). Each document is carefully crafted using `GPT-4o` to ensure that facts (or *insights*) are repeated across 6+ documents within the same topic. To this end, the authors begin by prompting `GPT-4o` to generate a list of candidate subtopics that are unique and expandable into more than 3 distinct insights. Each subtopic is then automatically validated to ensure that no subtopics overlap thematically and to verify that there are at least 3 insights that are specific to that subtopic and not other. For additional details on the dataset creation, refer to the original paper.

#### A.1.1 What is an insight?

Next, we enumerate a few examples of insight-level annotations found in the original dataset.

- **News domain**: The insights in this domain are generally more quantitative, richer in details, and descriptive of the events in the documents. However, we notice that certain subtopics tend to be more abstract and subjective (*e.g.,* insights related to historical context or NHRA regulations). Moreover, despite the insight quality checks conducted in the original dataset, we find evidence that insights are not always independent and sometimes exhibit large lexical and even semantic overlap (as emphasized by the reference insights related to *Foot Locker*). Specifically, based on our manual analysis, we notice that in examples containing partially overlapping insights, LLMs tended to generate a single insight (instead of multiple) and therefore naturally exhibited lower recall for that summary.

  1. "Dodge's switch to E85 for the Demon 170 marks a significant shift from their historical reliance on gasoline, driven by the high performance demands of the vehicle."
  2. "The journey from the original 426 Hemi in 1964 to the 2023 Demon 170 highlights significant engineering milestones, such as the introduction of the supercharged Hellcat in 2015, producing 707 horsepower."
  3. "Many tech startups suddenly faced cash flow issues as they scrambled to find new banking partners after SVB failed. Some had difficulty understanding new banking systems and integrating them with their accounting software, causing delays in operations."
  4. "Foot Locker reported Q4 2022 sales of $2.33 billion, exceeding consensus expectations of $2.15 billion, despite a 0.3% year-over-year decline."
  5. "Dillon's leadership is set to position Foot Locker for growth in 2024 and beyond, focusing on expanding wallet share and broadening customer reach."
  6. "Despite Tesla's reputation, the expanding EV market has allowed competitors to gain traction, diluting Tesla's early advantage."
  7. "During the 2021 forest fires in Turkiye, the '#HelpTurkey' hashtag on Twitter generated a nationalistic reaction as it was perceived to imply that the country could not handle the crisis, highlighting the power of hashtags in influencing public perception."
  8. "Banks are expected to adopt more diversified financial models to mitigate future risks. For example, banks might integrate asset-backed lending and communal investment pools to spread out risk effectively."
  9. "Foot Locker plans to ensure that approximately 40% of its revenue will come from non-Nike brands by 2026 to reduce dependency on any sin-

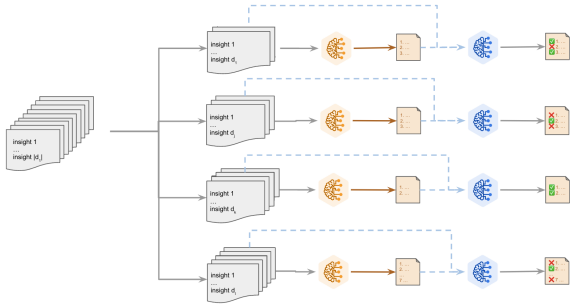


Figure 7: **Proposed evaluation protocol to investigate hallucinations in LLMs as a function of number of documents and task focus**. Given a corpus of documents with insight-level annotations, we craft combinations of $N$ documents such that part of the information about the queried subtopic is present in 2+ documents.

gle brand."

10. "Foot Locker is focusing on higher-margin non-Nike products to improve profitability while reducing reliance on any single footwear brand."
11. "Foot Locker expects to double its revenue from non-Nike brands by focusing on sneakers that cater to different occasions and activities, thereby appealing to a wider audience."
12. "The company noted a broadening consumer base that requires a wider range of brands, including non-Nike offerings, to meet various footwear needs."
13. "The company plans to broaden its assortment of lifestyle and performance shoes from various brands during the holiday seasons to cater to diverse consumer preferences."

- **Conversation domain**: Unlike the news domain, the conversation (conv) dataset tend to describe multi-turn interactions between participants in ordinary situations (*e.g.,* doctor appointments, company meetings, student debates). Notably, while the reference insights remain detail-specific, aiming to accurately capture the participants' interactions and responses (*e.g.,* by specifying time, duration, cost, or other specific details related to the interactions), they are participant-agnostic. That is, they are not tailored to a participant with a specific name. Additionally, note how some insights simultaneously convey multiple pieces of information, rather than focusing on a single unit of information.

1. "The doctor questions whether the symptoms have been getting worse, better, or staying the same, and the patient indicates that they have been gradually worsening over the last few days."
2. "The doctor asks the patient to rate the severity of their symptoms on a scale from 1 to 10, and the patient rates their symptoms as a 7."
3. "The project manager notes that the user manual and API documentation are 50% complete, with a target to finish by the end of the month."
4. "The project manager allocates additional resources to a critical task that requires more attention, assigning 2 extra developers to ensure timely completion."
5. "They explore the necessity of using task management tools like Asana or Trello to assign tasks, track progress, and hold team members accountable for their deliverables."
6. "The sales rep asks about the primary sources of customer data for the business, and the customer lists their website forms, social media platforms, and email marketing campaigns as the main sources."
7. "The sales representative and customer agree to schedule a live demonstration for the following Tuesday at 2 PM."
8. "The sales rep inquires about the tools the customer is currently using to manage their business processes, and the customer responds that they are using a mix of Excel sheets and a simple CRM software from a lesser-known vendor."
9. "One student offers to scan their handwritten notes and convert them into PDFs, which they believe might be useful for others who prefer reading handwritten material."
10. "One student finds that their best study time is at night, typically starting their sessions at 10 PM and studying until 2 AM."

### A.2 Evaluated Datasets.

One of our driving research questions was to understand how models' predisposition to hallucinate changed with increasing number of documents. To this end, we repurpose the SummHay dataset by manipulating the insight-level annotations associated with each document. In particular, for every topic (each containing 100 documents), we create combinations of $N$ documents s.t. when

| SummHay-News | SummHay-Conv |
|---|---|
| Comprehensive Analysis of Dodge V8 Muscle Cars: Evolution, Performance, and Future Trends | A 25-minute meeting between a doctor and a patient discussing the patient's recent health concerns and treatment options |
| Economic and Regulatory Implications of the Silicon Valley Bank Collapse | A 20-minute meeting between a project manager, two team members, and a stakeholder discussing the progress of a software development project at IBM |
| Strategic Growth and Operational Changes at Foot Locker | A 50-minute debate between three colleagues that work at a big company discussing the pros and cons of remote work |
| Recent Developments and Challenges in Twitter's Algorithms and Services | A 30-minute phone conversation between a Salesforce sales representative and either a current, prospective, or former customer |
| Current Financial Market Dynamics and Institutional Responses | A 25-minute study group session where three students discuss their strategies and insights for an upcoming exam |

Table 4: **List of topics included in the SummHay dataset (Laban et al., 2024), highlighting the thematic scope and domain coverage**. The topics proposed in the news domain are entity-centric, grounded on well-known entities (*e.g.,* Dodge V8 Muscle cars, Silicon Valley Bank, Twitter's, Foot Locker), but also more likely to be quantitative. Conversely, the topics in the conversation domain include 2-4 participants and are more broad.

paired with a subtopic $q$, the resulting combination is guaranteed to have at least 2 subtopic-related insights that are shared across 2 or more documents. We perform this for various combination sizes $N = \{2, 3, 4, 5, 10\}$ for all 10 topics, resulting in two benchmarks that we use in our evaluations. The statistics of the benchmarks are summarized in Tables 5 and 6, including the input length[7] (`total length`), number of shared subtopics (`# shared subtopics`), and shared insights (`# shared insights`) From these benchmarks, we sample 500 combinations for each $N$, totalling 2.5k examples per domain.

## B Prompt Selection

This paper outlines the reasoning and proposed modifications to the prompts used in the SummHay paper. The prompts used to report the results in the main paper are:

- Figure 11, used in the news, "subtopic" setting;
- Figure 13, used in the news, "subtopic+trustworthy" setting;
- Figure 12, used in the conv, "subtopic" setting;
- Figure 14, used in the conv, "subtopic+trustworthy" setting;
- Figure 15, used to parameterize the LLM-evaluator and measure information coverage.

Early on in our experiments, we explored a modified version of the news summarization prompt proposed in the SummHay paper (Laban et al., 2024). Since our goal was to investigate hallucinations in MDS (and not in model's capability to cite different sources), we remove the instructions related to the citation task, therefore avoiding potential anchoring effects that could be associated with performing multiple tasks in a single shot (Stureborg et al., 2024). Additionally, because we are interested in analyzing model behavior in scenarios where the number of relevant insights in the input documents is unknown, we do not instruct the model on the ideal summary length. Instead, in our earlier experiments, to avoid biasing evaluation based on the length of the summaries (Liu et al., 2023b; Guo and Vosoughi, 2023), we explore instructing the

[7]The length of the combinations is estimated using `tiktoken`'s encoder `cl100k_base` to encode all documents in a combination.

Table 5: **Different statistics for varying number of document combinations (N) in the proposed evaluation benchmark for the news domain.** In addition to the total number of combinations (`# combinations`), we report the average (and standard deviation) over the number of combinations for the total document length in tokens (`total length`), as well as for the number of insights and subtopics included in each combination.

| | N = 2 | N = 3 | N = 4 | N = 5 | N = 10 |
|---|---|---|---|---|---|
| # combinations | 1.5k | 1.5k | 1.5k | 1.5k | 1.2k |
| total length | $1766.2_{\pm 243.2}$ | $2626.9_{\pm 328.0}$ | $3522.8_{\pm 397.3}$ | $4363.6_{\pm 464.6}$ | $8643.7_{\pm 895.0}$ |
| # subtopics | $2.9_{\pm 0.4}$ | $3.6_{\pm 0.5}$ | $4.3_{\pm 0.7}$ | $4.8_{\pm 0.7}$ | $6.8_{\pm 0.9}$ |
| # shared subtopics | $1.1_{\pm 0.4}$ | $1.4_{\pm 0.5}$ | $1.7_{\pm 0.6}$ | $2.0_{\pm 0.6}$ | $3.9_{\pm 0.7}$ |
| # insights | $11.4_{\pm 1.8}$ | $14.8_{\pm 2.2}$ | $18.4_{\pm 2.6}$ | $21.6_{\pm 3.0}$ | $34.5_{\pm 3.5}$ |
| # subtopic insights | $5.2_{\pm 0.8}$ | $5.6_{\pm 0.8}$ | $6.6_{\pm 1.0}$ | $7.3_{\pm 1.0}$ | $9.2_{\pm 0.8}$ |
| # shared insights | $2.8_{\pm 1.0}$ | $3.9_{\pm 1.4}$ | $4.9_{\pm 1.6}$ | $6.3_{\pm 1.8}$ | $14.4_{\pm 2.4}$ |
| # subtopic shared insights | $2.6_{\pm 0.6}$ | $2.4_{\pm 0.6}$ | $2.8_{\pm 0.9}$ | $3.5_{\pm 1.0}$ | $7.0_{\pm 1.0}$ |

Table 6: **Different statistics for varying number of document combinations (N) in the proposed evaluation benchmark for the conversation domain.** In addition to the total number of combinations (`# combinations`), we report the average (and standard deviation) over the number of combinations for the total document length in tokens (`total length`), as well as for the number of insights and subtopics included in each combination.

| | N = 2 | N = 3 | N = 4 | N = 5 | N = 10 |
|---|---|---|---|---|---|
| # combinations | 341 | 1.5k | 1.5k | 1.5k | 1.5k |
| total length | $2054.6_{\pm 235.2}$ | $3017.5_{\pm 306.5}$ | $3992.3_{\pm 354.3}$ | $4947.4_{\pm 405.8}$ | $9870.2_{\pm 622.2}$ |
| # subtopics | $2.9_{\pm 0.3}$ | $3.7_{\pm 0.5}$ | $4.4_{\pm 0.6}$ | $5.0_{\pm 0.7}$ | $7.3_{\pm 0.9}$ |
| # shared subtopics | $1.1_{\pm 0.3}$ | $1.3_{\pm 0.5}$ | $1.6_{\pm 0.6}$ | $1.9_{\pm 0.6}$ | $3.8_{\pm 0.8}$ |
| # insights | $5.2_{\pm 0.7}$ | $8.5_{\pm 0.9}$ | $11.2_{\pm 1.1}$ | $13.4_{\pm 1.2}$ | $21.7_{\pm 1.6}$ |
| # subtopic insights | $2.0_{\pm 0.0}$ | $3.6_{\pm 0.5}$ | $4.5_{\pm 0.7}$ | $5.0_{\pm 0.7}$ | $5.9_{\pm 0.3}$ |
| # shared insights | $2.0_{\pm 0.2}$ | $2.2_{\pm 0.4}$ | $2.6_{\pm 0.7}$ | $3.3_{\pm 0.9}$ | $7.3_{\pm 1.3}$ |
| # subtopic shared insights | $2.0_{\pm 0.0}$ | $2.0_{\pm 0.2}$ | $2.3_{\pm 0.5}$ | $2.8_{\pm 0.7}$ | $5.3_{\pm 0.7}$ |

model to restrict the number of words used in the summaries. To determine this number, we analyse the ground truth summaries (concatenation of reference insights for each dataset) and find 300 words to be a reasonable upper bound over the reference insights (see Figure 8), therefore allowing for some semantic and lexical variations in LLM summaries.

**Ablation: Instructing LLMs to adhere to a fixed number of words.** Using the modified prompt, we instruct 4 different LLMs to generate 250 summaries for different number of document combinations and investigate the properties of the generated summaries, including their sentence length. To this end, we decompose the summaries into their words and punctuation using the `word_tokenize` from the `nltk` Python library[8] and report the number of generated words. Lengthwise, with the exception of `gpt-3.5-turbo-0125 (Any)`, we observe a large discrepancy (up to 200 words difference) between LLM-generated summaries and ground truth summaries, especially in cases involving fewer input documents (see Figure 9). One potential reason for the observed discrepancy is that models attempt to satisfy the length condition (of 300 words) and, in doing so, generate overly verbose summaries. Such verbosity may negatively impact models' ability to digest information and increase the models' hallucination rate.

**Ablation: Instructing LLMs to be succinct.** In many practical applications, it is difficult to exactly antecipate how many words (or insights) would be enough to summarize the relevant information, since it may depend on various factors, including the amount of information in the input (*e.g.,* number, length, and diversity of documents), as well as the degree of specificity of the summary (*e.g.,* depth vs more general insights). Consequently, we opt for instructing models with a more nuanced length restriction that emphasizes the clarity and conciseness of the generated summary: *Your summary should be concise and clear.*. Empirically, we found that replacing the fixed-length instruction led to significant reductions (by 50 to 150 words) in

[8] https://www.nltk.org/api/nltk.tokenize.html#nltk-tokenize-package

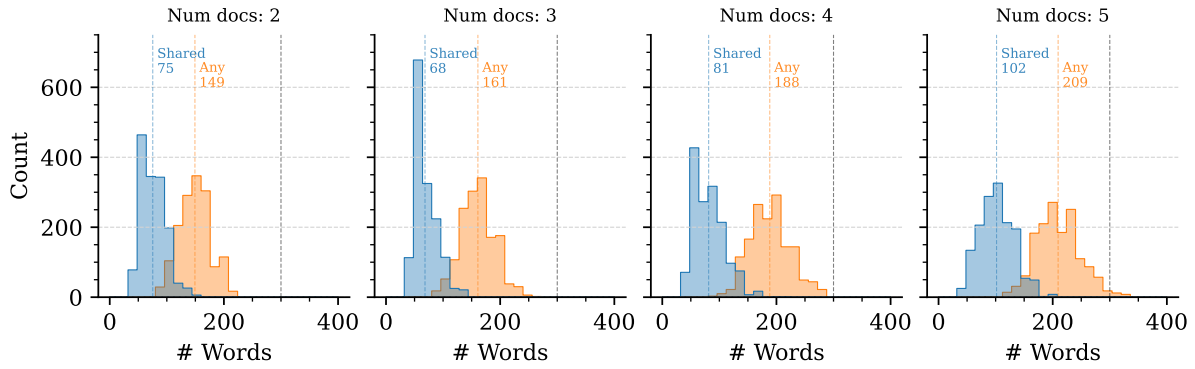



Figure 8: **Distribution of the number of words of the reference summaries for the proposed evaluation benchmark in the news domain**. The length of reference summaries is computed by concatenating ground truth reference insights for the different examples in the evaluation benchmarks and counting the number of words separated by whitespaces. `Any`, and `Shared` refer to the "subtopic" and "subtopic+trustworthy" settings, respectively. We observe that, across the different document combinations, the number of words (including punctuation) is fewer than the limit of 300 words specified in the prompt proposed by Laban et al. (2024). Thus suggesting that 300 words is more than enough to faithfully summarize information within the provided documents.

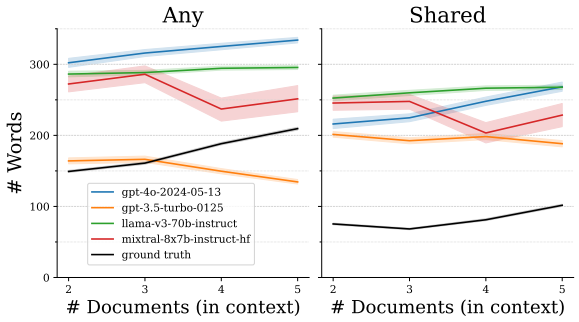



Figure 9: **Number of words in summaries as a function of the number of input documents (news domain)**. On the left, we observe the length of LLM-generated summaries when instructed to summarize insights in the "subtopic" setting. On the right, we observe the length of LLM-generated summaries when instructed to summarize shared subtopic-related insights ("subtopic+trustworthy" setting). In general, we observe a large gap between the total length of ground truth summaries (black solid line) and the LLM-generated ones, suggesting that models attempt to generate 300 words-long summaries even when the relevant information can be condensed in using fewer words.

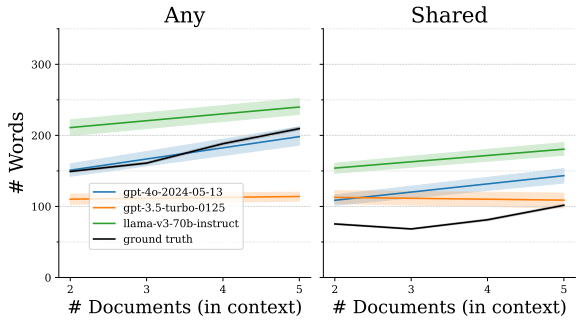



Figure 10: **Number of words in summaries as a function of the number of input documents (news domain)**. The summaries are generated by instructing the models on the generation of *clear and concise* summaries instead of *no longer than 300 words in total*. We observe smaller discrepancies between ground truth summaries and LLM-generated summaries. Compared to Figure 9, LLM-generated summaries are shorter by up to 150 words, especially for fewer documents. These results suggest that specifying the length of the summaries may bias the model to produce unnecessarily verbose summaries.

the length of LLM-generated summaries, leading to summaries that length-wise are more aligned with the reference summaries (see Figure 10). Based on these findings, we replace the fixed-length condition in the prompt with this more nuanced instruction.

**Addressing mispecifications in the input.** Finally, to provide some format to the summaries and limit potential confounding aspects related to the bullet-point or the inability to identify subtopic-related insights in the prompt, we include two additional instructions in the prompt: *Represent each bullet-point using "-".* and *If you do not find related insights, write "No insights found" and nothing else.*.

## C Model Access Details

This section describes the model card, access endpoints, and hyperparameter configurations used for each model.

**News Summarization - "subtopic" Prompt**

```
You are given {{n_articles}} news articles about the main subject "{{topic}}".

```
Article 1:
{{article_1}}

Article 2:
{{article_2}}
{{remaining_articles}}
```

Your objective is to clearly and concisely summarize all insights regarding
the topic of "{{subtopic}}". If you don't find insights regarding the topic of
"{{subtopic}}", return "No insights found".

Careful:
- [Format] You should format your summary as a bullet point list, where each
bullet point is a different insight consisting of a single sentence. Represent
each bullet point using "-". If you don't find related insights, write "No
insights found" and nothing else.
- [Length] Your summary should be concise and clear.
```

Figure 11: **Prompt used to summarize news articles**. The "subtopic" prompt instructs the model to summarize *any* insight in the documents that relates to the specified `{{subtopic}}`. `{{placeholders}}` will be replaced with corresponding values.

**Model Access.** We use the official OpenAI API[9] to access three LLMs, namely `gpt-3.5-turbo-0125`, `gpt-4o-2024-05-13`, and `gpt-4o-mini-2024-07-18`. To access `gemini-1.5-flash` we use Google's Generative AI API[10]. The remaining models, which include `llama-v3p1-70b-instruct` and `qwen2-72b-instruct`, are accessed through the Fireworks API[11]. Early on we also used the Fireworks API to evaluate LLMs, including Llama-3 (70B) (`llama-v3-70b-instruct`), and Mixtral (`mixtral-8x7b-instruct-hf`) but found numerous artifacts in their generations and, for that reason, excluded them from our analysis. For instance, `Llama-3 (70B)` generates preambles before generating list items (*e.g., Here is a summary ...*, *Here are the insights regarding ...*) but also postambles in 43% of its generations. In the postambles `Llama-3 (70B)` either re-iterate the instructions (*e.g., Note: These insights are summarized from both Article 1 and Article 2, These insights ... and are summarized in a concise and clear manner.*) or summarizes the listed bullet-points (*e.g., These insights highlight the issues with Twitter's complex tweet recommendation algorithm ..., which is a critical step in improving the platform's user experience and rebuilding trust with its users.*). Conversely, `mixtral-8x7b-instruct-hf`'s generations either end abruptly, leading to incomplete summaries (*e.g., Foot Locker aims to ensure that approximately 40% of its revenue*, *Twitter, under Elon Musk's leadership, has announced plans to*), or conclude with the total word count (*e.g., (Word count: 241)*, *(Note: The word count is exactly 200 words without the headings.)*). Throughout our experiments, we explicitly focus on the evaluation of larger models due to their superior instruction following capability. All experiments were carried in 2024, with the main experiments (in Section 4) being conducted between July 1st to September 5th, and the mitigation experiments (described in Section 6) being conducted between September

[9] https://platform.openai.com/docs/quickstart

[10] https://ai.google.dev/gemini-api/docs/quickstart?authuser=2&lang=python

[11] https://fireworks.ai/models

**Conv Summarization - "subtopic" Prompt**

```
You are given {{n_conversations}} conversations about the following scenario:
"{{topic}}".
The conversations involve the following participants: {{participants}}. In
each conversation, the participants might be different, with different names,
but all the conversations fall into the same scenario.

```
Conversation 1:
{{conversation_1}}

Conversation 2:
{{conversation_2}}
{{remaining_conversations}}
```

Your objective is to clearly and concisely summarize all insights in the
document regarding the topic of "{{subtopic}}".
If you don't find insights regarding the topic of "{{subtopic}}", write "No
insights found".

Careful:
- [Format] You should format your summary as a bullet point list, where each
bullet point is a different insight consisting of a single sentence. Represent
each bullet point using "-". If you don't find related insights, write "No
insights found" and nothing else.
- [Length] Your summary should be concise and clear.
```

Figure 12: **Prompt used to summarize multiple conversations**. The "subtopic" prompt instructs the model to summarize *any* insight in the documents that relates to the specified `{{subtopic}}`. `{{placeholders}}` will be replaced with corresponding values.

10th and October 12th.

**Summary generation.** The summaries generated in this paper were all generated using the following configurations: `temperature=1`, `top_p=0.9`, and `max_tokens=800`.

**LLM-based evaluation.** Automatic evaluation is conducted using `gpt-4o-mini-2024-07-18` with parameters `n=1`, `temperature=0`, `top_p=1`, and `response_format: { "type": "json_object"}`. In Appendix D, we validate that the metric correlates strongly with human judgements of correctness.

## D Automatic Metric Validation

In the main paper, we adopt an LLM-as-a-judge approach (Zheng et al., 2023) to determine whether the predicted insights *faithfully* cover the information in the input. Our evaluation approach is based on the metric proposed in the SummHay's paper (Laban et al., 2024), which comprehensively compared the accuracy and human-alignment of different LLMs as evaluators, including `gpt-4o` (OpenAI, 2024b), `Gemini-1.5-pro` (Anil et al., 2024), and `Claude` (Anthropic, 2024). They find that among models they evaluate, `Gemini-1.5-pro` and `gpt-4o` were found to be positively correlated with human-level annotations. In this paper, however, we adopt `gpt-4o-mini-2024-07-18` as a more cost-efficient but still competitive evaluator.[12] To ensure that replacing the evaluator model does not negatively impact evaluation quality (La-

[12] `gpt-4o-mini-2024-07-18` has been shown to be competitive with `gpt-4o` across various math, coding, and instruction-following benchmarks alike (OpenAI, 2024a; Chiang et al., 2024).

**News Summarization - "subtopic+trustworthy" Prompt**

```
You are given {{n_articles}} news articles about the main subject "{{topic}}".

```
Article 1:
{{article_1}}

Article 2:
{{article_2}}
{{remaining_articles}}
```

Your objective is to clearly and concisely summarize all insights regarding
the topic of "{{subtopic}}" that are mentioned in at least two articles. In
other words, your summary should include any insight that is related to the
topic and is  mentioned in two or more articles.
If you don't find insights regarding the topic of "{{subtopic}}", return "No
insights found".

Careful:
- [Format] You should format your summary as a bullet point list, where each
bullet point is a different insight consisting of a single sentence. Represent
each bullet point using "-". If you don't find related insights, write "No
insights found" and nothing else.
- [Length] Your summary should be concise and clear.
```

Figure 13: **Prompt used to summarize commonalities across news articles**. The "subtopic+trustworthy" prompt instructs the model to summarize insights that are *shared* across two or more documents and that relate to the specified `{{subtopic}}`. `{{placeholders}}` will be replaced with corresponding values.

ban et al., 2024), we investigate the agreement between the two LLMs across 45.3k reference insights, spanning 4 summarization models and 2 distinct prompts. Given the complex nature of the evaluation process, our analysis is decomposed in terms of *linking agreement*—focused on whether models agree on which predicted insight (if any) covers the information of the reference insight—and *coverage agreement*—measuring to which extent the models agree on the coverage label (`NO`, `PARTIAL`, `FULL`).

**Model Coverage Agreement.** Among the matching examples, `gpt-4o` and `gpt-4o-mini` exhibit strong inter-annotator agreement, achieving a Cohen's Kappa of 0.84 and a Kendall Tau correlation coefficient of 0.94 (McHugh, 2012; Kendall, 1938). In practice, the two models agree on the coverage labels for about 91% of these examples (see Table 7).

Table 7: **Breakdown of information coverage labels for the 43.2k examples where `gpt-4o` and `gpt-4o-mini` agree on the candidate insight.** The two models output the same coverage label in 91% of the examples, with `gpt-4o-mini` being more optimistic than `gpt-4o` in the cases where they disagree.

| gpt-4o | gpt-4o-mini | Counts | Frequency |
|---|---|---|---|
| full | full | 13084 | 30.24 |
| full | partial | 18 | 0.04 |
| partial | full | 3877 | 8.96 |
| partial | partial | 2255 | 5.21 |
| no | no | 24029 | 55.55 |

**Model Linking Agreement.** Our results show that the models' linking predictions match in about 95.5% of the examples, suggesting strong linking agreement between the two models. Out of the 2049 mismatching examples, we find that `gpt-4o-2024-05-13` and `gpt-4o-mini-2024-07-18` assign the same cov-

**Conv Summarization - "subtopic+trustworthy" Prompt**

```
You are given {{n_articles}} conversations about the following scenario:
"{{topic}}".
The conversations involve the following participants: {{participants}}. In
each conversation, the participants might be different, with different names,
but all the conversations fall into the same scenario.

```
Conversation 1:
{{conversation_1}}

Conversation 2:
{{conversation_2}}
{{remaining_conversations}}
```

Your objective is to clearly and concisely summarize all insights regarding
the topic of "{{subtopic}}" that are mentioned in at least two conversations.
In other words, your summary should include any insight that is related to the
topic and is mentioned in two or more conversations.
If you don't find insights regarding the topic of "{{subtopic}}", write "No
insights found".

Careful:
- [Format] You should format your summary as a bullet point list, where each
bullet point is a different insight consisting of a single sentence. Represent
each bullet point using "-". If you don't find related insights, write "No
insights found" and nothing else.
- [Length] Your summary should be concise and clear.
```

Figure 14: **Prompt used to summarize commonalities across conversations**. The "subtopic+trustworthy" prompt instructs the model to summarize insights that are *shared* across two or more documents and that relate to the specified `{{subtopic}}`. `{{placeholders}}` will be replaced with corresponding values.

erage label to 24.35% of them (see Table 8). Through manual inspection of 50 such examples, we find that the candidate insights exhibit high semantic overlap or complementarity, which is not well captured by the current evaluation process—only 1 candidate insight can be attributed to the reference insight. When comparing the win rate of the two models, we find that humans agree with both models 55% times and disagree with both models about 10% of the times. We also note that while `gpt-4o-2024-05-13` favors predicted insights with higher lexical overlap with the reference insight, humans prefer the answers from `gpt-4o-mini-2024-07-18` (over `gpt-4o-2024-05-13`) 30% of the times.

Next, we consider the cases where `gpt-4o-2024-05-13` suggests a greater information coverage of the reference insight than `gpt-4o-mini-2024-07-18`—representing 39.09% of the examples with linking disagreement. Again, we manually inspect 50 such examples and find that humans prefer `gpt-4o-mini-2024-07-18` over `gpt-4o-2024-05-13` in 60% examples. One potential reason for the high disagreement is due to the fact that the predicted insights only match the information content of reference insights at a high-level, missing important details or adding irrelevant information, therefore compromising the usefulness of predicted insights in practice. As a result, humans may prioritize insights that cover part of the details within the reference insight (as opposed to insights focusing on the

**Evaluation Prompt**

```
You are given a list of bullet points (each with a unique number), and a specific reference insight. Your
objective is to determine whether the reference insight is covered in any of the bullet points. You must
further determine if the insight is partially covered ("PARTIAL_COVERAGE") or fully covered
("FULL_COVERAGE") by the bullet points. If the insight is not covered at all, you must return
"NO_COVERAGE". See examples below:

Example Reference Insight 1: "The doctor asks the patient about their medical history".

Example Bullet Points 1:
{"bullets": [
    {"bullet_id": 1, "text": "The patient often mention that they are worried about medication
    side-effect."},
    {"bullet_id": 2, "text": "The doctor and patient spend time going over symptoms, particularly the
    initial symptoms and the progression in the last few months."},
    {"bullet_id": 3, "text": "The doctor and patient discuss medical history within the patient's family,
    with the patient often unaware that some of the conditions are hereditary."}
]}

Example Output 1:
{"coverage": "FULL_COVERAGE", "bullet_id": 3}

Example Reference Insight 2: "The doctor asks the patient about their medical history".

Example Bullet Points 2:
{"bullets": [
    {"bullet_id": 1, "text": "The patient often mention that they are worried about medication
    side-effect."},
    {"bullet_id": 2, "text": "The doctor and patient spend time going over symptoms, particularly the
    initial symptoms and the progression in the last few months."}
]}

Example Output 2:
{"coverage": "NO_COVERAGE", "bullet_id": "NA"}

Example Reference Insight 3: "The doctor asks the patient about their medical history".

Example Bullet Points 3:
{"bullets": [
    {"bullet_id": 1, "text": "The patient often mention that they are worried about medication
    side-effect."},
    {"bullet_id": 2, "text": "The doctor and patient catch up after a long time, with the patient
    mentioning feeling unwell for a while, and knowing of other family member's similar experiences."},
    {"bullet_id": 3, "text": "The doctor and patient spend time going over symptoms, particularly the
    initial symptoms and the progression in the last few months."}
]}

Example Output 3:
{"coverage": "PARTIAL_COVERAGE", "bullet_id": 2}

Now complete the task for the following insight and bullet points:

Reference Insight:
{{reference_insight}}

Bullet Points:
{{candidate_insights}}


Requirements:
- Do not hallucinate that the insight is covered by the bullet points if it is not.
- Your response should only be the JSON output in the format above, such that it can directly parsed by
Python's json module. DO NOT OUTPUT ANY EXPLANATION OR ANYTHING THAT IS NOT THE JSON RESPONSE.
```

Figure 15: **Few-shot prompt to automatically evaluate information coverage**. Proposed by Laban et al., the prompt determines whether the information conveyed by the `{{reference_insight}}` is fully, partially, or not covered by any of the `{{candidate_insights}}`.

Table 8: **Breakdown of information coverage labels for the 2049 non-matching examples**. In 24.35% of the mismatches, the models predict the same coverage label (*full*, *partial*, *no*) for different insights. `gpt-4o-mini` is more pessimistic than `gpt-4o` in 39.09% of the examples and more optimistic in the remaining 36.56%.

| gpt-4o | gpt-4o-mini | Counts | Relative Freq |
|---|---|---|---|
| full | full | 344 | 16.79 |
| full | partial | 1 | 0.05 |
| partial | full | 336 | 16.40 |
| partial | partial | 155 | 7.56 |
| partial | no | 800 | 39.04 |
| no | full | 54 | 2.64 |
| no | partial | 359 | 17.52 |

higher-level information). Similar human trends are observed when analyzing cases for which `gpt-4o-mini-2024-07-18` attributes higher information coverage than `gpt-4o-2024-05-13`. Across 50 manually annotated examples, humans prefer `gpt-4o-mini`'s outputs about 46% of the times *versus* 28% of examples where they prefer `gpt-4o` outputs, having 18% of examples where both are equally preferred.

# E Additional Results

This section constitutes additional results that are complementary to the main paper.

- Appendix E.1 discusses the impact of coverage labels in the reported metrics;
- Appendix E.2 provides insights about the length of the generated summaries;

## E.1 Impact of Coverage Labels

In Section 2.3, we described the LLM-based metric that we use to automatically determine the *faithfulness* of the generated summary. By design, our metric outputs three coverage labels (`NO`, `PARTIAL`, `FULL`). But to be able to compute the recall and hallucination rates, we must covert these into a single correctness label (*correct/incorrect*). In the original paper (Laban et al., 2024), the authors sidestep this conversion by computing a *coverage score* (that is similar to recall) but is computed as follows:

> **Coverage Score**: For each insight, the summary receives a score of 100 for full coverage, 50 for partial coverage, and 0 otherwise. The final coverage score of a summary is the average coverage on all the insights of the subtopic, such that it ranges from 0 to 100.

Instead, we adopt an optimistic approach and report the best-case scenario. In other words, by assuming that both `PARTIAL` and `FULL` insights are correct, we get a ceiling on the model's performance in the best case scenario, *e.g.,* if a model achieves 20% recall or 60% error rate, we immediately know that in a more conservative scenario, models will exhibit lower performance. Note, however, that in practice this difference between `PARTIAL` and `FULL` is only meaningful if a `PARTIAL` represent a large fraction of the coverage labels assigned by our metric. Figure 16 shows that indeed the gap between the two is minimal for the news domain. Conversely, Figure 17 illustrates an up to 30% points difference when considering `PARTIAL` correct (fc+pc) or incorrect (fc).

## E.2 Analysis of the Number of Insights

As previously mentioned (in Appendix B), we do not to condition models to generate summaries with a fixed number of insights. As a consequence, and because of the different training dynamics and fine-tuning procedures associated with different LLMs, we expect models to generate summaries with different lengths. Figure 18 shows the average number of insights for the evaluated benchmarks. We observe that, with the exception of `gpt-3.5-turbo-0125`, models tend to generate longer summaries.

Figure 19 shows the ratio of predicted-to-reference insights for the summaries generated for the two evaluated domains—news and conversation—as the number of input documents increases. By observing this ratio we can examine each models' propensity to over- or under-generate insights. The first observation is that model behavior varies considerably depending on the prompt setting ("subtopic" or "subtopic+trustworthy") and the domain. Secondly, we observe that all models tend to over-estimate the number of relevant insights in the input documents when processing lower number of documents (N<3). In general, we find that

## E.3 F1-score Results

Figure 22 reports the F1-scores obtained for both "subtopic" and "subtopic+trustworthy" settings across the news (Figure 22a) and conversation domain (Figure 22b).

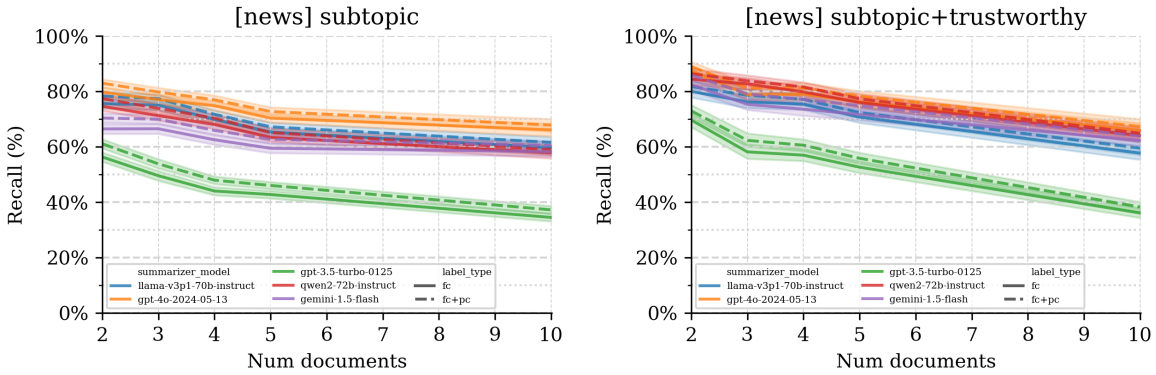



Figure 16: **Impact of coverage labels in the macro-recall metric for the SummHay-News across both prompt settings**. Overall, we observe minimal differences ($< 5\%$) across evaluated LLMs, number of documents, and pr

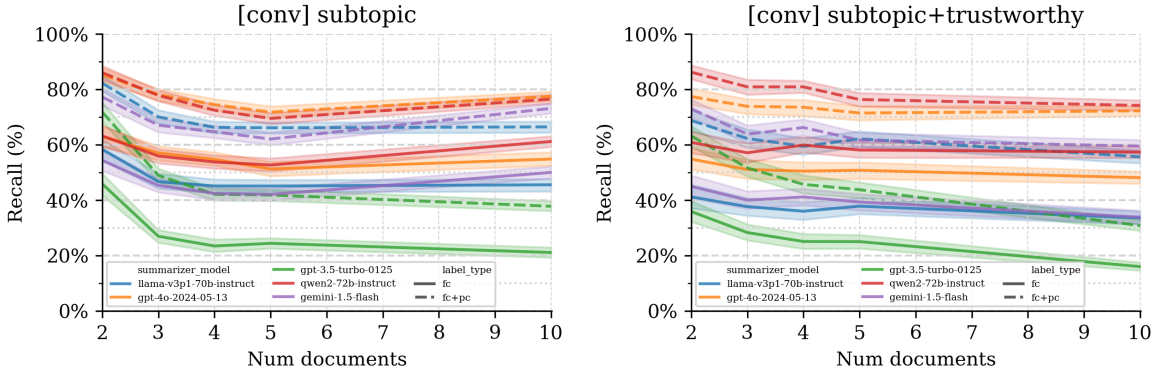



Figure 17: **Impact of coverage labels on recall measures in the SummHay-Conv**. Unlike in the news domain, we observe large disparities of up to 30% percentage points depending on whether we consider partial coverage (fc+pc) or not (fc). We hypothesize that the higher fraction of PARTIAL labels is associated with the fact that insights in the conversation domain are more contextual and are less entity-centric than the news domain and, as a consequence, make it more difficult to assess but also to summarize.

### E.4 Analysis Breakdown

In this section, we report additional results about the types of errors. Tables 9 summarize the percentage of correctness when using fc+pc label type.

### E.5 Correctness vs Outputs

In the main paper, we report the results with respect to a single model and when processing combination of 10 documents. In this section, we show evidence of the same behavior across all evaluated models (see Figure 23), as well as for combination size N=2 (see Figure 24).

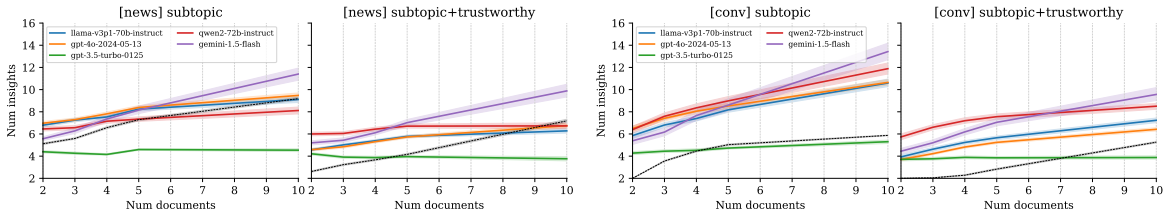
## F Hallucination Taxonomy's Annotation Details

To create the hallucination taxonomy, we collected human annotations for over 150 LLM-generated summaries and developed a taxonomy based on recurring mistakes. Due to the challenge of analyzing long documents, as noted in prior work (Chang et al., 2023), we limited our analysis to N=2, leaving the exploration of larger combinations for future research.

Two authors independently conducted the main analysis, manually inspecting over 700 predicted insights, which were nearly uniformly distributed across models, task focus, and domains. Each annotation sample included two input documents, the query containing the target subtopic, all reference insights, and the predicted insight. Annotators first determined whether the predicted coverage label (full, partial, or not covered) of the insight was correct and, then, provided a description for the type of error in the predicted insight. After gathering error descriptions, the annotators together discussed the observed patterns, defined the hallucination taxonomy based on those patterns, and categorized the hallucinated insights under each type.

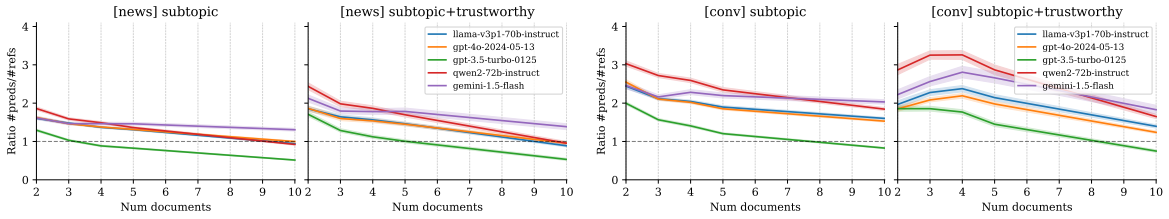
Table 9: **Analysis of summarizer predictions in the conversation domain for varying number of in-context documents, when using the subtopic prompt**.

| Summarizer | % Shared | % Subtopic | % Shared-Sub | % Context | % Not Context |
|---|---|---|---|---|---|
| **N = 2** | | | | | |
| gpt-3.5-turbo-0125 | $37.85_{\pm 23.66}$ | $37.61_{\pm 23.56}$ | $37.61_{\pm 23.56}$ | $45.48_{\pm 23.91}$ | $54.52_{\pm 23.91}$ |
| gpt-4o-2024-05-13 | $29.43_{\pm 14.37}$ | $29.34_{\pm 14.29}$ | $29.34_{\pm 14.29}$ | $35.9_{\pm 17.18}$ | $64.1_{\pm 17.18}$ |
| llama-v3p1-70b | $32.01_{\pm 17.17}$ | $31.93_{\pm 17.13}$ | $31.93_{\pm 17.13}$ | $38.03_{\pm 18.53}$ | $61.97_{\pm 18.53}$ |
| **N = 3** | | | | | |
| gpt-3.5-turbo-0125 | $28.48_{\pm 18.87}$ | $40.42_{\pm 25.18}$ | $28.0_{\pm 18.75}$ | $49.74_{\pm 26.02}$ | $50.26_{\pm 26.02}$ |
| gpt-4o-2024-05-13 | $25.37_{\pm 12.84}$ | $40.79_{\pm 17.86}$ | $25.17_{\pm 12.86}$ | $49.33_{\pm 20.53}$ | $50.67_{\pm 20.53}$ |
| llama-v3p1-70b | $25.24_{\pm 13.22}$ | $39.18_{\pm 18.64}$ | $24.92_{\pm 13.01}$ | $46.43_{\pm 20.78}$ | $53.57_{\pm 20.78}$ |
| **N = 4** | | | | | |
| gpt-3.5-turbo-0125 | $29.76_{\pm 21.23}$ | $43.04_{\pm 26.12}$ | $28.77_{\pm 20.81}$ | $51.72_{\pm 26.44}$ | $48.28_{\pm 26.44}$ |
| gpt-4o-2024-05-13 | $26.31_{\pm 13.0}$ | $43.67_{\pm 17.23}$ | $25.65_{\pm 12.7}$ | $52.93_{\pm 18.78}$ | $47.07_{\pm 18.78}$ |
| llama-v3p1-70b | $25.56_{\pm 13.56}$ | $41.6_{\pm 18.27}$ | $24.85_{\pm 13.13}$ | $48.68_{\pm 19.65}$ | $51.32_{\pm 19.65}$ |
| **N = 5** | | | | | |
| gpt-3.5-turbo-0125 | $32.75_{\pm 22.17}$ | $44.95_{\pm 25.24}$ | $31.84_{\pm 21.71}$ | $53.55_{\pm 26.63}$ | $46.45_{\pm 26.63}$ |
| gpt-4o-2024-05-13 | $28.41_{\pm 12.9}$ | $44.13_{\pm 16.78}$ | $27.9_{\pm 12.88}$ | $52.26_{\pm 18.91}$ | $47.74_{\pm 18.91}$ |
| llama-v3p1-70b | $28.15_{\pm 14.87}$ | $42.92_{\pm 17.82}$ | $27.63_{\pm 14.56}$ | $49.21_{\pm 19.09}$ | $50.79_{\pm 19.09}$ |
| **N = 10** | | | | | |
| gpt-3.5-turbo-0125 | $44.06_{\pm 25.68}$ | $43.51_{\pm 25.37}$ | $41.55_{\pm 24.96}$ | $55.15_{\pm 27.01}$ | $44.85_{\pm 27.01}$ |
| gpt-4o-2024-05-13 | $43.33_{\pm 14.44}$ | $44.23_{\pm 13.8}$ | $40.92_{\pm 13.59}$ | $54.88_{\pm 16.33}$ | $45.12_{\pm 16.33}$ |
| llama-v3p1-70b | $38.49_{\pm 15.73}$ | $39.33_{\pm 16.04}$ | $36.82_{\pm 15.33}$ | $47.43_{\pm 17.39}$ | $52.57_{\pm 17.39}$ |




(a) news (b) conversation

Figure 18: **Number of predicted insights reported for the the proposed evaluation benchmarks**. Each line represents the average number of insights, with shaded region representing the 95% confidence interval. Despite being less perceptible in the news domain, we observe systematic differences between the two prompt settings, suggesting that models exhibit different behavior when instructed to focus on the shared insights.




(a) news (b) conv

Figure 19: **Ratio of predicted insights to reference insights reported for two datasets**. Each line represents the average number of insights, with shaded region representing the 95% confidence interval. A ratio greater than 1 indicates the model predicted more insights than those referenced and, thus, indicates larger propensity to make mistakes (or possibly more redundant information). A ratio less than 1 signifies fewer predicted insights, indicating the model fails to identify some of the relevant information in the input documents.

## G Hallucination Mitigation

Based on the observed patterns during our manual annotation, we seek out to investigate the effectiveness of simple *post-hoc* heuristics. In particular, we make two observations: (1) with the exception of `GPT-3.5-Turbo`, LLMs tend to over-generate insights (more than reference insights—See Figure 19) and (2) a large fraction of LLM mistakes is due to the generation of overly generic and/or repeated information. To assess the extent to which simple heuristics —both LLM-based and rule-based—may help reducing such errors, we conduct a small-scale experiment and measure the impact in the measured recall and hallucination rate. Ideally, one would like mitigation approaches to reduce the hallucination rate with minimal impact on recall.

### G.1 Methods

As previously mentioned, we propose two categories of methods: rule-based and model-based. The **rule-based methods** are faster to run and do not depend on existing LLMs. The proposed rule-based methods are listed below:

- **Truncate summaries (Top-K)**: this method is based on our findings that LLMs' insights generated earlier in the summary are more likely to be accurate than those generated later (see 4.5). Given a summary composed of $I$ insights (expressed as bullet-points), this method truncates the summary, keeping the first $K$ insights.[13]

In addition to rule-based methods, we also explore three different **model-based methods**. As the name indicates, this class of methods rely on LLMs and, hence, may be considered more time-consuming and costly than rule-based methods. However, these methods are also more versatile and nuanced facilitating the manipulation of semantic relationships between two texts. We explore these capabilities to mitigate some of the patterns observed in the LLMs generations, such as generating insights that are unrelated to the subtopic, redundant, and/or paraphrases of the subtopic:

- **Unrelated Subtopic (`st-unrelated`)**: given the queried subtopic $q$ and a list of predicted insights (*i.e.,* a summary in bullet-point format), this method filters out the insights that are not related to the subtopic $q$. Particularly,

[13] Early on, we also experimented with the removal of the first $L$ insights but found it to cause significant drops in recall and, therefore, chose not to include those results in the experiments.

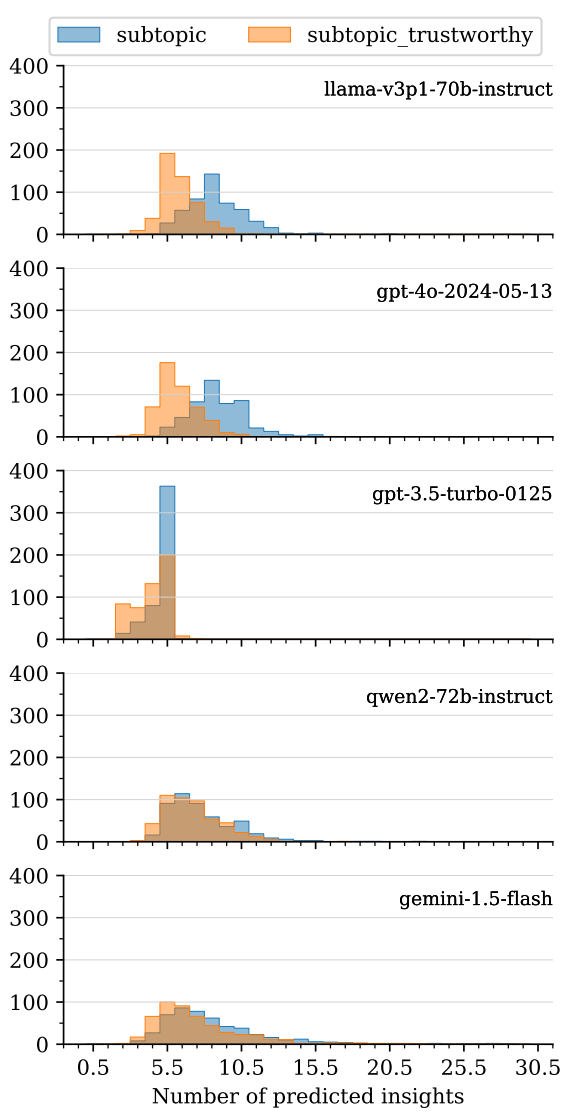



Figure 20: **Distribution of the number of predicted insights per model in the news domain for combinations of 5 documents (N=5).**

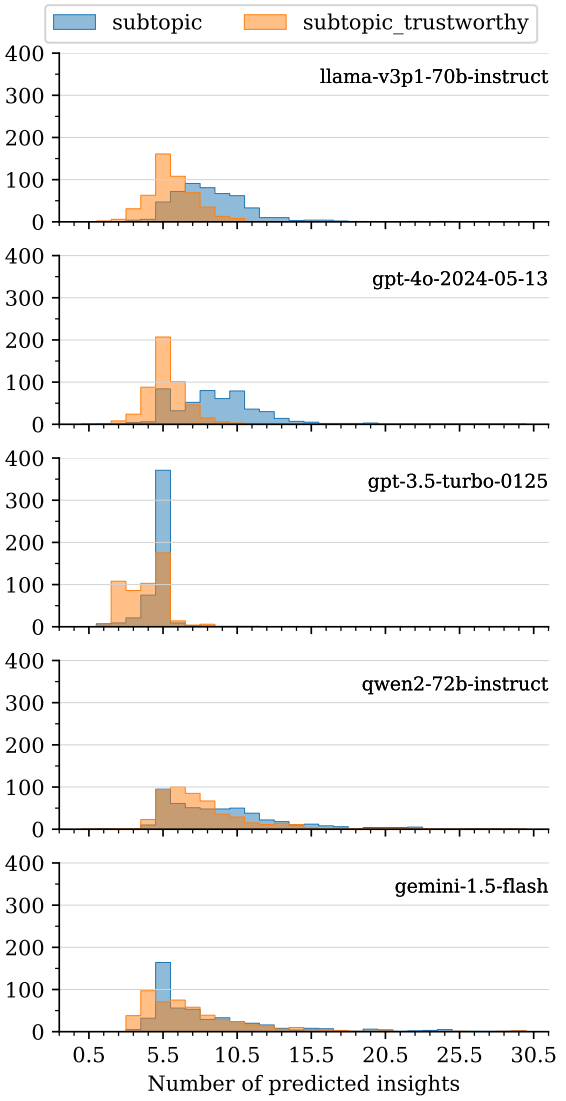
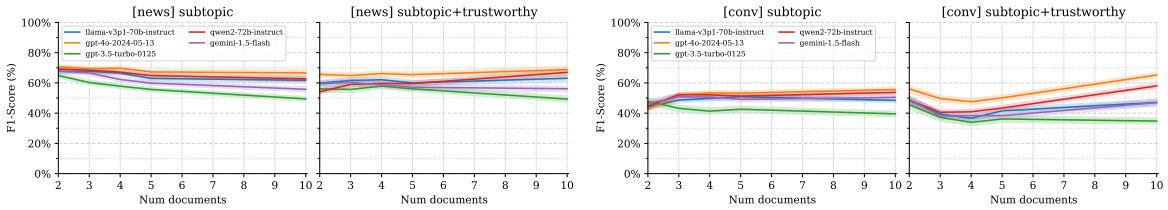



Figure 21: **Distribution of the number of predicted insights per model in the conv domain for combinations of 5 documents (N=5).**




(a) news (b) conv

Figure 22: **F1-score as a function of the number of input documents** . Each line represents the mean value, with shaded areas indicating the 95% confidence intervals. Overall, `gpt-3.5-turbo-0125` and `gemini-1.5-flash` are among the worst performers in both domains, due to their tendencies to generate overly short and long summaries, respectively. Surprisingly, `qwen2-72b-instruct` reveals to be on par with `gpt-4o-2024-05-13`, with the former exhibiting slightly superior performance overall (<5% points).

for each predicted insight $i$, we zero-shot ask `gpt-4o-mini-2024-07-18` whether $i$ is related to the subtopic $q$ and use the greedy decoded answer (yes/no) to determine whether to keep the insight or not. Alternatively, we also experiment using different confidence thresholds $\alpha_u \in [0, 1]$, where we filter out the predicted insight if the likelihood assigned to the answer "yes" is below a threshold $\alpha_u$. We denote this method using the notation `st-unrelated-`$\alpha_u$.

- **Paraphrases of Subtopic (`st-paraphrase`)**: the goal of this method is to identify and filter out the insights that have a similar meaning to subtopic $q$ or are superficial in nature. Following previous work (Michail et al., 2024; Farquhar et al., 2024), we explore the use of two different approaches to detect meaning similarity: (1) a zero-shot LLM-based approach (again instantiated using `gpt-4o-mini-2024-07-18`) and (2) an entailment approach (instantiated using `microsoft/deberta-v2-xlarge-mnli` (He et al., 2021)), which we denote `st-paraphrase` and `st-paraphrase-nli`, respectively.[14] By default, we consider the greedy answer or label when determining whether an insight $i$ is a paraphrase of subtopic $q$. We also experiment using a the model's confidence $\alpha_p \in [0, 1]$ to regulate this prediction, which we denote `st-paraphrase-`$\alpha_p$ or `st-paraphrase-nli-`$\alpha_p$ depending on whether we use the zero-shot or the entailment approach.

- **Redundant Bullet-points (`redundant`)**: To detect redundancy in the summaries, we reuse the previous zero-shot and entailment approaches but instead of using it to compare a subtopic with an insight, we use it to compare every pair of predicted insights in the summary. Whenever we find a redundant insight, we drop one of them. Like before, depending on whether the zero-shot LLM approach is being used or the nli approach, we refer to these methods as `redundant` and `redundant-nli`, appending a suffix -$\alpha_r$ to each method depending on the confidence threshold used.

[14] Results reported in the main paper refer to the LLM approach.

## G.2 Results

In this section, we present the full results corresponding to the hallucination mitigation. Table 10 shows the average F1-score improvements after applying the various methods, whereas Tables 11 and 12 summarize the average improvements in terms of recall and hallucination rate, respectively.

Table 10: **Absolute difference (in percentage points) in average F1-score after applying four simple mitigation methods to summaries generated from two input documents (N=2)**. "Top-k" is the only rule-based method and all other methods are model-based. By default, model-based methods use greedily decoded with no confidence threshold (unless explicitly identified through "-$\alpha$" for $\alpha \in [0, 1]$. Models with "-nli" suffix are implemented using bidirectional entailment model (`microsoft/deberta-v2-xlarge-mnli`) opposed to general purpose LLM (`gpt-4o-mini-2024-07-18`). All mitigation methods have little to no impact in terms of average F1-score ($\pm$3%), highlighting the need for further research to address hallucinations in multi-document scenarios more systematically.

| | Strategy | Gemini (Flash) | GPT-3.5-Turbo | GPT-4o | Llama 3.1 (70B) | Qwen 2 (72B) |
|---|---|---|---|---|---|---|
| news | top-5 | 0.61% | 0.09% | 2.51% | 0.42% | 1.69% |
| | st-unrelated | 0.15% | -0.37% | -0.64% | -0.27% | -0.02% |
| | st-paraphrase | -1.69% | -2.00% | -1.19% | -0.86% | -1.19% |
| | redundant | -0.46% | -0.02% | -0.09% | -0.41% | 0.16% |
| | st-unrelated-0.8 | -1.88% | -0.38% | -2.61% | -1.95% | -1.49% |
| | st-paraphrase-nli-0.6 | 0.03% | -0.01% | 0.05% | 0.09% | 0.01% |
| | redundant-nli-0.6 | -0.80% | -1.23% | -0.49% | -0.98% | -0.28% |
| conv | top-5 | 1.37% | 0.19% | 2.28% | 1.52% | 1.52% |
| | st-unrelated | 0.24% | 0.50% | 0.76% | 0.36% | 0.63% |
| | st-paraphrase | -0.23% | -0.77% | -0.11% | -0.29% | -0.46% |
| | redundant | 0.33% | 0.11% | 0.48% | 0.40% | 0.22% |
| | st-unrelated-0.8 | 0.09% | 0.64% | 0.85% | 0.43% | 0.85% |
| | st-paraphrase-nli-0.6 | 0.00% | -0.04% | 0.01% | 0.00% | 0.00% |
| | redundant-nli-0.6 | 0.36% | 0.53% | 0.46% | 0.18% | -0.08% |

Table 11: **Absolute difference (in percentage points) in average Recall after applying four simple mitigation methods to summaries generated from two input documents (N=2)**. "Top-k" is the only rule-based method and all other methods are model-based. By default, model-based methods use greedily decoded with no confidence threshold (unless explicitly identified through "-$\alpha$" for $\alpha \in [0, 1]$. Models with "-nli" suffix are implemented using bidirectional entailment model (`microsoft/deberta-v2-xlarge-mnli`) opposed to general purpose LLM (`gpt-4o-mini-2024-07-18`). Overall, we find slight drops in average recall (<10%), we observe lower improvements (< 5%) in average hallucination rate (in Table 12), which highlights the complexity of mitigating these hallucination errors and the need for further exploration.

| | Strategy | Gemini (Flash) | GPT-3.5-Turbo | GPT-4o | Llama 3.1 (70B) | Qwen 2 (72B) |
|---|---|---|---|---|---|---|
| news | top-5 | -2.72% | -0.07% | -4.66% | -5.67% | -3.58% |
| | st-unrelated | -2.27% | -2.82% | -3.90% | -2.58% | -3.20% |
| | st-paraphrase | -3.30% | -3.34% | -3.14% | -3.18% | -3.10% |
| | redundant | -1.28% | -0.12% | -0.49% | -1.96% | -0.39% |
| | st-unrelated-0.8 | -5.11% | -3.02% | -8.55% | -5.08% | -7.40% |
| | st-paraphrase-nli-0.6 | -0.05% | -0.07% | -0.03% | 0.00% | 0.00% |
| | redundant-nli-0.6 | -2.89% | -1.94% | -2.21% | -4.12% | -1.96% |
| conv | top-5 | -3.08% | -0.15% | -4.55% | -2.64% | -5.57% |
| | st-unrelated | -0.15% | -0.44% | -0.29% | -0.29% | -0.15% |
| | st-paraphrase | -2.05% | -1.61% | -0.44% | -2.35% | -1.47% |
| | redundant | -1.32% | 0.00% | -0.88% | -0.88% | -1.03% |
| | st-unrelated-0.8 | -0.59% | -0.44% | -0.44% | -0.44% | -0.15% |
| | st-paraphrase-nli-0.6 | 0.00% | -0.15% | 0.00% | 0.00% | 0.00% |
| | redundant-nli-0.6 | -3.96% | -1.61% | -3.23% | -4.69% | -4.25% |

Table 12: **Absolute difference (in percentage points) in average hallucination rate after applying four simple mitigation methods to summaries generated from two input documents (N=2)**. "Top-k" is the only rule-based method and all other methods are model-based. By default, model-based methods use greedily decoded with no confidence threshold (unless explicitly identified through "-$\alpha$" for $\alpha \in [0, 1]$. Models with "-nli" suffix are implemented using bidirectional entailment model (`microsoft/deberta-v2-xlarge-mnli`) opposed to general purpose LLM (`gpt-4o-mini-2024-07-18`). Overall, we find slight drops in average recall (< 10%) (in Table 11), we observe lower improvements (< 5%) in average hallucination rate, which highlights the complexity of mitigating these hallucination errors and the need for further exploration.

| | Strategy | Gemini (Flash) | GPT-3.5-Turbo | GPT-4o | Llama 3.1 (70B) | Qwen 2 (72B) |
|---|---|---|---|---|---|---|
| news | top-5 | 2.29% | 0.18% | 6.09% | 3.58% | 4.28% |
| | st-unrelated | 0.61% | 0.93% | 0.59% | 0.43% | 0.98% |
| | st-paraphrase | 0.67% | 0.87% | 0.28% | 0.92% | 0.44% |
| | redundant | 0.14% | 0.19% | 0.17% | 0.66% | 0.53% |
| | st-unrelated-0.8 | 0.70% | 1.17% | 1.87% | 0.49% | 2.98% |
| | st-paraphrase-nli-0.6 | 0.14% | 0.14% | 0.11% | 0.17% | 0.01% |
| | redundant-nli | 1.10% | 0.23% | 0.57% | 1.26% | 0.92% |
| conv | top-5 | 1.26% | 0.17% | 2.21% | 1.44% | 1.61% |
| | st-unrelated | 0.22% | 0.66% | 0.80% | 0.42% | 0.61% |
| | st-paraphrase | 0.80% | -0.61% | -0.04% | 0.16% | -0.24% |
| | redundant | 0.70% | 0.12% | 0.65% | 0.53% | 0.40% |
| | st-unrelated-0.8 | 0.25% | 0.84% | 0.89% | 0.52% | 0.83% |
| | st-paraphrase-nli-0.6 | -0.00% | -0.00% | 0.01% | -0.00% | -0.00% |
| | redundant-nli | 0.95% | 1.12% | 0.83% | 0.86% | 0.52% |

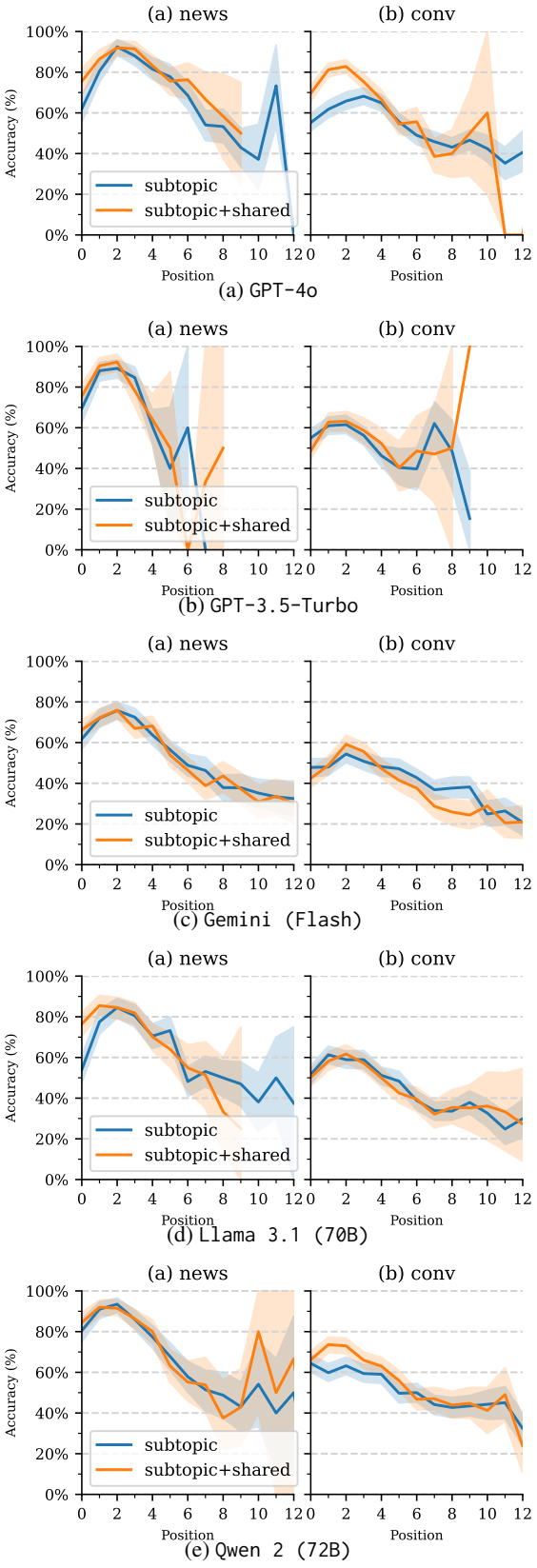



Figure 23: **Accuracy rate of LLM-generated insights by position (when summarizing 10 input documents)**. Each solid line represents the mean value, with shaded areas indicating the 95% confidence intervals. Overall, we observe the same pattern across all models: accuracy decreases as LLMs generate more insights.

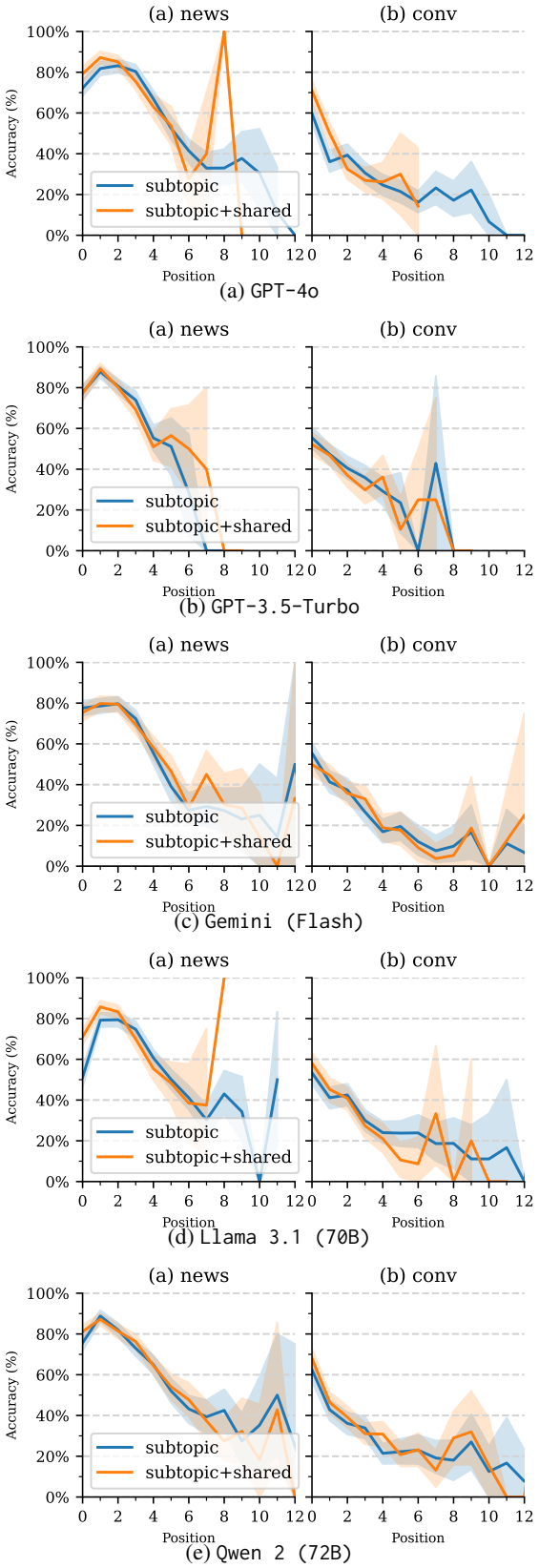



Figure 24: **Accuracy rate of LLM-generated insights by position (when summarizing 2 input documents)**. Overall, we observe the same pattern across all models: accuracy decreases as LLMs generate more insights.