FAITH: A Framework for Assessing Intrinsic Tabular Hallucinations in Finance

Mengao Zhang[†]
mengaoz@nus.edu.sg
Asian Institute of Digital Finance,
National University of Singapore
Singapore

Jiayu Fu[*]
jiayu.fu@u.nus.edu
Asian Institute of Digital Finance,
National University of Singapore
Singapore

Tanya Warrier[*]
tanya.warrier@u.nus.edu
Asian Institute of Digital Finance,
National University of Singapore
Singapore

Yuwen Wang
wangyuwen@u.nus.edu
Asian Institute of Digital Finance,
National University of Singapore
Singapore

Tianhui Tan
tant@nus.edu.sg
Asian Institute of Digital Finance,
National University of Singapore
Singapore

Ke-Wei Huang
huangkw@comp.nus.edu.sg
Asian Institute of Digital Finance,
National University of Singapore
Singapore

ABSTRACT

Hallucination remains a critical challenge for deploying Large Language Models (LLMs) in finance. Accurate extraction and precise calculation from tabular data are essential for reliable financial analysis, since even minor numerical errors can undermine decision-making and regulatory compliance. Financial applications have unique requirements, often relying on context-dependent, numerical, and proprietary tabular data that existing hallucination benchmarks rarely capture. In this study, we develop a rigorous and scalable framework for evaluating intrinsic hallucinations in financial LLMs, conceptualized as a context-aware masked span prediction task over real-world financial documents. Our main contributions are: (1) a novel, automated dataset creation paradigm using a masking strategy; (2) a new hallucination evaluation dataset derived from S&P 500 annual reports;[1] and (3) a comprehensive evaluation of intrinsic hallucination patterns in state-of-the-art LLMs on financial tabular data. Our work provides a robust methodology for in-house LLM evaluation and serves as a critical step toward building more trustworthy and reliable financial Generative AI systems.

KEYWORDS

Large Language Models, Hallucination, Benchmarking, Financial NLP, Table Reasoning

1 INTRODUCTION

While Large Language Models (LLMs) are rapidly transforming financial services through capabilities such as automated information extraction [4] and client-facing chatbots [16], their deployment also introduces substantial risks, among which hallucination poses serious threats to decision-making and stakeholder trust [8, 19–22]. In response, regulators such as the Monetary Authority of Singapore (MAS) have underscored the need for robust model risk management frameworks tailored to advanced AI applications [15]. Despite growing regulatory scrutiny, systematic methods for evaluating and mitigating hallucinations in financial contexts remain largely undeveloped. As a foundational step, this study proposes a scalable, comprehensive framework for evaluating and analyzing LLM hallucinations in finance, quantifies how increasing information-extraction complexity exacerbates hallucination errors, and offers actionable guidance for researchers and practitioners.

This paper focuses on intrinsic hallucinations—a form of LLM error where the generated output is inconsistent with the provided input context [3]. Many financial tasks require precise extraction, summarization, calculation, interpretation, and reasoning based on structured inputs such as financial tables. Errors in these processes can significantly undermine decision-making. Intrinsic hallucinations are already concerning in such cases; when LLMs rely instead on web search results or internal knowledge rather than the actual financial input, the risk of hallucination becomes even greater. Undetected hallucinations can propagate through complex, automated systems, leading to misleading analyses, flawed investment strategies, or regulatory breaches.

Yet, evaluating hallucinations in a systemic and robust approach is a non-trivial challenge, mainly due to the lack of finance-specific hallucination evaluation datasets. Most existing datasets are derived from general-domain texts such as Wikipedia [3, 6], whereas financial applications often involve complex, numerically grounded data - including tables, charts, and document inputs - and require specialized reasoning and contextual understanding. More importantly, there is a clear need for an automated and scalable approach to creating evaluation datasets for hallucination detection in real-world financial settings. Manual annotation is resource-intensive and can hardly keep up with evolving LLM behaviors or the diversity of proprietary financial data. Thus, inspired by recent advancements [7, 12, 26], our study addresses this gap by proposing a rigorous and scalable methodology that uses actual financial annual reports—where ground truth values are explicitly known—to automatically construct benchmarking tasks for evaluating intrinsic hallucinations in financial LLM applications.

In this study, we address a core challenge in financial LLM applications: accurately identifying, calculating, and presenting numerical values based on financial statements (input context). We introduce a novel context-aware masked span prediction task, where numerical values from real annual reports are masked and LLMs are prompted to recover them. To ensure reliable evaluation, we

[*] Work done during internship at the Asian Institute of Digital Finance, National University of Singapore.
[†] Correspondence.
[1] Dataset available upon request.

develop a novel filtering method that selects only those values with a unique, consistent, and answerable ground truth. A central contribution is our taxonomy of four mutually exclusive and exhaustive financial reasoning types: *Direct Lookup*, *Comparative Calculation* (The value is a function of the same variable in different periods, e.g., year-over-year growth), *Bivariate Calculation* (the value is a function of two variables, e.g., gross margin), and *Multivariate Calculation* (more complex dependencies). This classification enables structured evaluation across reasoning complexity. Using our approach, we build a large-scale benchmark dataset from S&P 500 annual reports and evaluate leading LLMs on their intrinsic hallucination rates. Results show that even top models frequently hallucinate, especially on complex tasks.

Our work provides an essential foundation for evaluating hallucinations in finance, a critical first step before effective mitigation. The main contributions are as follows:

- We propose a novel and scalable paradigm for dataset creation based on a masking strategy, enabling automatic construction of evaluation datasets for both public and proprietary financial documents.
- We release a new dataset for evaluating intrinsic hallucinations in finance, with varying reasoning complexities.
- We conduct a comprehensive analysis of intrinsic hallucination patterns in state-of-the-art LLMs on financial tabular data, including a detailed breakdown by reasoning type.

2 RELATED WORKS

2.1 Hallucination Evaluation Datasets

LLM hallucinations can be categorized as extrinsic or intrinsic, based on the reference source [3]. The evaluation of extrinsic hallucinations assesses whether a model's output, generated from its internal knowledge alone, aligns with established world facts. Benchmarks in this category [10, 11, 13] generally evaluate models against static, open-domain corpora like Wikipedia. Financial LLMs, however, operate in a domain where information is highly time-sensitive, context-intensive, and often proprietary, requiring evaluation methods that prioritize contextual accuracy over pretrained, parametric knowledge [8, 19–22].

Therefore, intrinsic hallucination evaluation is more relevant and critical for financial applications, as it directly measures a model's faithfulness to a specific, provided context. While intrinsic evaluation datasets have emerged, especially for Retrieval-Augmented Generation (RAG) applications [6, 14, 17], those based on general-purpose text, mostly are limited to simple look-up tasks rather than complex reasoning [2, 17].

Our work addresses this critical gap, focusing on intrinsic hallucinations within complex financial tabular data—a unique and high-stakes domain largely overlooked by current evaluation methodologies.

2.2 Tabular Reasoning and Evaluation

Our focus on intrinsic hallucinations with tables connects to the broader research area of tabular reasoning. Foundational benchmarks such as `RealHiTBench` [24] and `TableBench` [25] have been pivotal in assessing core LLM skills like numerical reasoning. In the more intricate financial domain, specialized benchmarks have emerged to address higher complexity. Datasets like `FinQA` [5] and `TAT-QA` [30] test joint reasoning over tables and text, while others focus on navigating complex hierarchical structures [9, 23, 27].

While these benchmarks have significantly advanced LLM reasoning capabilities, their evaluation scope is primarily focused on the correctness of the final answer. This focus on accuracy creates a critical blind spot: as task complexity increases, so does the risk that a model generates a correct-seeming output through unfaithful reasoning or hallucination. Addressing this gap requires moving beyond traditional question-answering formats to develop scalable methodologies that can directly probe for such intrinsic errors. Our study proposes a novel framework to rigorously and automatically evaluate intrinsic hallucination in financial tabular reasoning.

2.3 Automatic Dataset Construction

The high cost and limited scalability of manual annotation have driven the development of automated evaluation dataset construction methods. Some approaches achieve fine-grained analysis by decomposing generated content into atomic facts for verification [1, 7, 13], while others leverage the inherent structure of relational data to automatically generate complex, verifiably question-answer pairs [18]. Adversarial test cases can be dynamically created through data perturbation [26]. Although using LLMs as annotators may introduce biases [28], recent work shows that carefully designed automated pipelines can yield high-quality annotations, sometimes rivaling human performance [12]. We build on these advancements by adapting and extending automated benchmarking paradigms to the unique challenges of financial tabular analysis.

3 TASK DEFINITION

In this section, we focus on the core requirement of retrieving information from context and providing accurate answers in complex financial scenarios, which is a common need for financial LLM applications. We formulate the task as a context-aware masked span prediction over tabular financial data, for the goal of intrinsic hallucination evaluation. Figure 1 demonstrates key flow.

3.1 Problem Formulation

Let a financial document D be composed of a set of structured tables $\mathcal{T}$ and a body of text. This text is partitioned into two distinct, non-overlapping sets:

- A set of explanatory pre-texts $\mathcal{P}$, where each pre-text $P_T \in \mathcal{P}$ introduces the table $T \in \mathcal{T}$.
- A sequence of general sentences $S = (s_1, s_2, ..., s_N)$, which constitutes the remaining text of the document.

Within a given sentence $s_i \in S$, we identify a set of non-overlapping spans of interest, $\mathcal{M}_i = \{m_{i,1}, m_{i,2}, ..., m_{i,k}\}$, based on specific criteria. A task instance is constructed by selecting a single span $m_{i,j} \in \mathcal{M}_i$ and replacing its content with a `[MASK]` token, producing a corrupted sentence $\tilde{s_i}$.

The objective of the model f is to recover the original content of the masked span $m_{i,j}$. The model is conditioned on the corrupted sentence and a context set C_i:

$$\hat{m}_{i,j} = f(\tilde{s_i}, C_i) \tag{1}$$

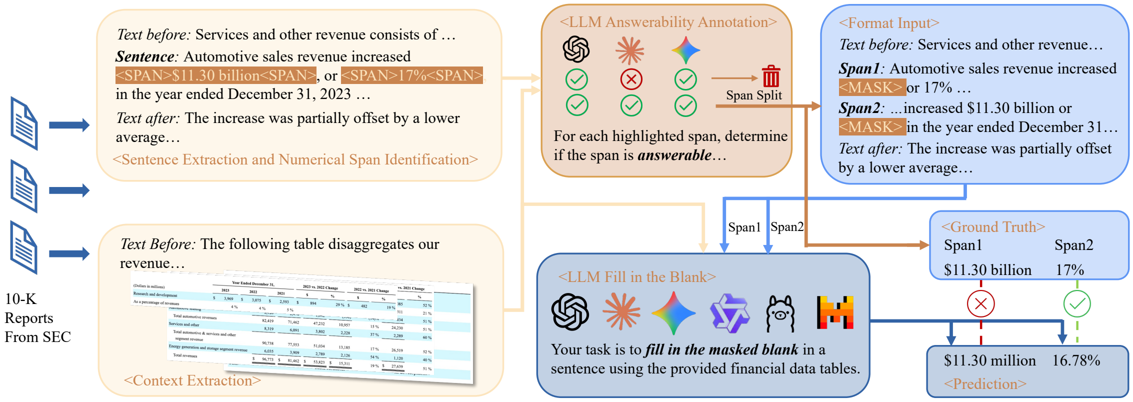
Text before: Services and other revenue consists of …
Sentence: Automotive sales revenue increased <SPAN>$11.30 billion<SPAN>, or <SPAN>17%<SPAN> in the year ended December 31, 2023 …
Text after: The increase was partially offset by a lower average…
<Sentence Extraction and Numerical Span Identification>
10-K Reports From SEC
Text Before: The following table disaggregates our revenue…
<Context Extraction>
<LLM Answerability Annotation>
Span Split
For each highlighted span, determine if the span is answerable…
<Format Input>
Text before: Services and other revenue…
Span1: Automotive sales revenue increased <MASK> or 17% …
Span2: ...increased $11.30 billion or <MASK> in the year ended December 31…
Text after: The increase was partially offset by a lower average…
Span1
Span2
<LLM Fill in the Blank>
Your task is to fill in the masked blank in a sentence using the provided financial data tables.
<Ground Truth>
Span1 Span2
$11.30 billion 17%
$11.30 million 16.78%
<Prediction>

Figure 1: Illustration of the task definition and data processing.

The context $C_i = \{\mathcal{T}, \mathcal{P}, s_{i-1}, s_{i+1}\}$ includes the tables, their pretext, and the immediately preceding and succeeding sentences. The goal is for the predicted span $\hat{m}_{i,j}$ to be matched with the original span $m_{i,j}$

3.2 Masking Criteria

To ensure the task setup yields meaningful and unbiased hallucination evaluation, three key assumptions must hold for each masked span. These assumptions guide our masking strategy and directly impact the reliability of evaluation outcomes:

- **Uniqueness:** The masked span must have a unique correct answer, preventing multiple plausible completions.
  - *Example:* A sentence like "The company's [MASK] has improved" would be excluded, as several financial metrics such as "revenue," "profit margin," or "cash flow" could all be reasonable answers.
- **Consistency:** The ground-truth span must be consistent with the context, avoiding internal misalignments within the source document.
  - *Example:* If the table reports operating income as *$500 million*, the masked sentence should not mistakenly state "Operating income was *$500 thousand*" as the to-be-masked truth. Evaluating based on the incorrect masked content would unfairly penalize a model that correctly aligns with the table.
- **Answerability:** The masked span must be inferable from the provided context, ensuring the task is solvable by an LLM.
  - *Example:* If the sentence states "The company's revenue increased by [MASK] in 2024 compared to 2023," both revenue figures for 2024 and 2023 should be present in the tables or described in the surrounding text; otherwise, the span is not answerable and should be excluded from evaluation.

To enforce these assumptions and ensure reliable evaluation, we adopt the following design:

(1) **Numeric Span Selection:** We restrict masking to numeric spans that include units or verbal scales (e.g., "million," "USD"), ensuring that the masked content is both specific and uniquely recoverable. We also implement normalization strategies for evaluation to account for equivalent numeric representations.

Algorithm 1 Precision-Relaxed Matching with Unit Groups

```
Require: Ground truth span $m$, predicted span $\hat{m}$, unit groups $\mathcal{U}$
Ensure: $is\_numeric\_match$ and $is\_unit\_match$.
Let $\mathcal{U}_{\text{scales}} \subset \mathcal{U}$ be the subset of scale units.
$(v_m, p_m) \leftarrow \text{NormalizeNumber}(m, \mathcal{U}_{\text{scales}})$
$(v_{\hat{m}}, p_{\hat{m}}) \leftarrow \text{NormalizeNumber}(\hat{m}, \mathcal{U}_{\text{scales}})$
                                   ▷ Normalize to base value and determine precision
$p_{coarsest} \leftarrow \max(p_m, p_{\hat{m}})$
$is\_numeric\_match \leftarrow (\lfloor v_m / p_{\min} \rceil = \lfloor v_{\hat{m}} / p_{\min} \rceil)$
                                   ▷ Compare numbers after rounding
$U_m \leftarrow \text{ExtractUnits}(m, \mathcal{U} \setminus \mathcal{U}_{\text{scales}})$
$U_{\hat{m}} \leftarrow \text{ExtractUnits}(\hat{m}, \mathcal{U} \setminus \mathcal{U}_{\text{scales}})$
                                   ▷ Extract non-scale unit groups from each span
$is\_unit\_match \leftarrow (U_m \subseteq U_{\hat{m}})$
                                   ▷ Check for set equality of unit groups
return $is\_numeric\_match, is\_unit\_match$
```

(2) **Reliable Source Documents:** We use company annual reports (i.e., 10-K reports) as the data source, minimizing the likelihood of contradictions due to their regulatory rigor and editorial consistency.

(3) **Answerability Annotation:** We utilize LLMs to annotate the answerability of the spans and conducted a comprehensive pilot study by comparing human annotations with LLM-generated annotations, demonstrating that LLMs can reliably support the answerability labeling process.

3.3 Robust Evaluation

Our evaluation protocol is designed to robustly handle the nuances of numerical text, ensuring that valid predictions are not penalized due to formatting or semantic variations. We specifically account for two potential sources of ambiguity where a simple string comparison would prove insufficient:

(1) **Compromised Uniqueness:** In certain contexts, a masked span could be correctly expressed in multiple values, violating the Uniqueness assumption. For instance, in *"the company's revenue increased by [MASK],"* both a percentage change (*"10%"*)

and an absolute value (*"$5 million"*) could be factually correct based on the source table.

(2) **Formatting Variations:** A single numeric value can be written in many equivalent string formats (e.g., *$1,230 million* vs. *USD 1.23 billion*). A simple string match would incorrectly penalize valid predictions.

To address the first challenge, we guide the model by incorporating a **hint** into the prompt that specifies the expected value type (e.g., percentage or absolute value), thereby restoring uniqueness. To solve the second, we introduce a **precision-relaxed evaluation protocol** that normalizes and compares the numeric predictions. This protocol is detailed in Algorithm 1 and consists of two main components:

Numeric Matching. This process first normalizes the ground-truth span m and the predicted span $\hat{m}$ into base values. It parses numbers and scale indicators (e.g., "million," "billion") to derive a base number v (e.g., "1.23 billion" becomes 1.23×10^9). It then determines the numerical **precision** from the least significant digit. For instance, the number 1,230,000 has a precision p of 10^4, as its last non-zero digit is in the ten thousands place. A numeric match is declared if the ground-truth and predicted values are equal after being rounded to the coarser of their two precisions.

Unit Matching. We define a set of **unit groups** $\mathcal{U}$, where each group contains aliases for a single unit (e.g., {$, USD, dollars}, {mil, million, M}). We extract the set of unit groups present in each span by greedily matching the longest aliases first. For non-scale units, matching is successful only if all groups found in m are present in $\hat{m}$

3.4 Financial Reasoning Scenarios

To facilitate a fine-grained analysis of model performance, we categorize each masked span based on the complexity of the financial reasoning required to restore its content. We define four distinct scenarios:

A. **Direct Lookup:** The answer can be found by directly extracting a single cell in a table.
B. **Comparative Calculation:** The answer requires a simple calculation on a single metric across different time periods or categories (e.g., computing a year-over-year change).
C. **Bivariate Calculation:** The answer involves a simple calculation between two distinct metrics explicitly present in the table (e.g., computing a financial ratio).
D. **Multivariate Calculation:** The answer requires multi-step reasoning over three or more metrics or a sequence of arithmetic operations.

Recognizing that a masked span can sometimes be derived through multiple valid reasoning paths, we do not let LLM determine a fixed scenario label during answerability annotation step. Instead, during evaluation time, we instruct each model to classify its own reasoning process into one of the four scenarios. To establish a reliable scenario classification for our analysis, we aggregate the scenario classifications from all models that correctly predict the span's content. The final scenario for that span is the class with highest frequency, ensuring our analysis is grounded in successful and verifiable reasoning steps from LLM.

Table 1: Agreement among LLMs and human annotation for Answerability Annotation. "Yes" and "No" represent "answerable" and "unanswerable" respectively. The LLMs used are GPT-4.1, Claude-sonnet-4, and Gemini-2.5-pro.

	Human		
LLM Agreement Pattern	**Yes**	**No**	**Total**
All 3 Agree (Unanimous)			
3 LLMs - No	1	626	627
3 LLMs - Yes	276	11	287
2-1 Split (Disagree)			
2 LLMs - Yes, 1 LLM - No	12	23	35
1 LLM - Yes, 2 LLMs - No	11	164	175
Total	300	824	1124

3.5 Pilot Study

To validate our evaluation dataset construction methodology, we conduct a pilot study to assess the feasibility of using LLMs for annotation.

Manual Annotation. We begin by creating a ground-truth dataset. We sample 1,124 text spans from the *"Item 7. Management's Discussion and Analysis"* section of 10-K reports from nine companies across diverse industries (e.g., healthcare, finance, technology). Each span is labeled for answerability by at least two financial experts, with a senior reviewer adjudicating disagreements. This process yields a high inter-annotator agreement (Fleiss' Kappa = 0.905) and results in 300 answerable and 824 unanswerable spans. The annotators also classify the 300 answerable spans into our four financial reasoning scenarios (A: Direct Lookup, B: Comparative, C: Bivariate, D: Multivariate), resulting in a distribution of 212, 58, 22, and 8 respectively, which reflects the natural rarity of more complex reasoning tasks.

LLM Annotation. We then prompt three leading LLMs (GPT-4.1, Claude-Sonnet-4, Gemini-2.5-Pro) with the answerability annotation task. The results, summarized in Table 1, show that **unanimous LLM consensus is an exceptionally strong indicator of answerability correctness**. When all three models agreed, they correctly identified 626 of 627 unanswerable spans (99.8% accuracy) and 276 of 287 answerable spans (96.2% accuracy).

This pilot study validates that leveraging unanimous LLM consensus is a highly reliable and scalable strategy for annotating answerability. It is worth noting that we only mask existing text rather than generating new content; thus, using LLMs as annotators does not introduce fabrications. This ensures the soundness of our approach for creating intrinsic hallucination evaluation benchmarks.

4 EVALUATION DATASET

4.1 Data Collection and Processing

Our dataset is built upon publicly available 10-K reports from S&P 500 companies filed in 2024, ensuring the data reflects real-world financial reporting practices and covers a wide spectrum of industries. We source these documents in XBRL format from the SEC's

Table 2: Comparative statistics of the Pilot (human-annotated) and Main (LLM-annotated) dataset splits.

Metric	Pilot Split	Main Split
Number of companies	9	453
Avg. context length (chars)	14,148	12,843
Avg. number of tables	14.9	19.2
Number of sentences	164	1,122
Number of answerable spans	300	2,406

EDGAR database[1]. The collected filings then undergo several processing steps to extract high-quality numerical claims and their surrounding context.

Item 7 Extraction. From each 10-K filing, we extract "Item 7: Management's Discussion and Analysis of Financial Condition and Results of Operations" (MD&A). This section is selected for its rich narrative analysis and manageable length, providing a dense source of contextualized financial data. We apply a keyword-based search to locate the MD&A section, supplementing this automated process with manual curation to ensure data quality.

Context and Sentence Extraction. To avoid overlaps between context and sentence, we first extract tables and their immediate preceding sentence as the explanatory pre-texts. Then we parse the rest of the content into plain text. After which a sentence split is performed using spaCy to get a list of candidate sentences which incorporated custom rules to merge fragments that typically arose as artifacts from the document conversion process.

Numerical Span Identification. From this sentence corpus, we isolate claims containing numerical data. This is a two-stage process:

- **Initial Detection:** We employ spaCy's Named Entity Recognition (NER) to detect entities such as MONEY, PERCENT, CARDINAL, and QUANTITY.
- **Span Expansion:** To ensure the extracted spans are semantically complete, we expand the initial NER outputs with rules to include relevant currency symbols (e.g., $) and a comprehensive vocabulary of financial units and scales (e.g., "million," "billion," "per share").

LLM-based Answerability Annotation. To create a representable sample, we randomly select 10 sentences from each 10-K report. Following the same setting in pilot study, three popular LLMs are employed for annotating whether each span is answerable given the provided context. A span is retained in the final dataset only if all three models unanimously classified it as answerable.

4.2 Dataset Statistics

Table 2 presents a statistical overview of our dataset, which is composed of a human-annotated Pilot Split and a larger, LLM-annotated Main Split covering 453 S&P 500 companies after processing and filtering.

A key characteristic of our dataset is its realistic complexity. With an average context length exceeding 12,800 characters and an average of 19.2 tables per document, the dataset is designed to mirror the information-dense nature of real-world financial reports. This substantial context poses a significant challenge for LLMs, providing a robust testbed for evaluating their reasoning and retrieval capabilities under practical, real-world conditions.

5 EXPERIMENTS

5.1 Experimental Setup

We evaluate a range of state-of-the-art open-source and proprietary LLMs on our dataset's pilot and main splits. During evaluation, we prompt the models to generate a step-by-step rationale before predicting the final masked value and to self-classify their reasoning process into one of the four financial scenarios (A-D). The detailed prompt is available in Appendix A.1.

Accuracy is used as the primary evaluation metric, computed for overall predictions and separately for the numeric value and unit components. In addition, we report the accuracy stratified by financial scenario. For the pilot split, scenario labels are assigned by human annotators. For the main split, we use the strategy stated in section 3.4. This stratification allows for a detailed analysis of error types.

All models are run with a temperature of 0 to ensure reproducibility. Proprietary models are accessed via their official APIs, while open-source models are run on 4 H200 GPUs using the LLaMA-Factory framework [29]. We implement a retry mechanism (up to 3 attempts) to handle instances where a model failed to produce a valid JSON output.

5.2 Experiment Results

This section presents a detailed analysis of the performance of various LLMs on our financial hallucination benchmark. The comprehensive results, detailed in Table 3, facilitate an objective assessment of current model capabilities and limitations in finance. Our findings reveal a clear stratification of model performance, underscore the direct relationship between reasoning complexity and hallucination risk, and identify dominant error patterns that highlight specific model weaknesses.

5.2.1 A Quantifiable Hierarchy of Model Reliability. The results demonstrate a distinct performance hierarchy among the evaluated models.

A top tier of proprietary models, led by **Claude-Sonnet-4** (95.6% on Main Split) and **Gemini-2.5-Pro** (91.9% on Main Split), exhibits high overall accuracy. Their strong performance, which remains largely consistent between the Pilot and Main splits, validates the robustness of our benchmark. However, the 4-8% error rate, while low relative to other models, remains a significant consideration for financial applications where precision is non-negotiable.

Below this top tier, we observe a second group of capable models, including **GPT-4.1** (89.2%) and **GPT-4.1-mini** (88.2%), which deliver respectable but demonstrably lower accuracy. A further sharp decline is evident in the majority of other models, particularly smaller open-source variants. Models such as **Llama-3.1-8B** (47.5%) and **Qwen-3-8B** (30.6%) exhibit high error frequencies, rendering them unsuitable for tasks requiring financial fidelity.

[1] https://www.sec.gov

Table 3: Model Accuracy (%) on Pilot and Main Splits, including scenario performance. Scenario abbreviations: (A) Direct Lookup, (B) Comparative Calculation, (C) Bivariate Calculation, and (D) Multivariate Calculation.

Model	Pilot Split							Main Split						
	Overall	Value	Unit	A	B	C	D	Overall	Value	Unit	A	B	C	D
				(n=212)	(n=58)	(n=22)	(n=8)				(n=1606)	(n=635)	(n=135)	(n=10)
Gemini-2.5-pro	**95.0**	**95.7**	**97.7**	**96.7**	91.4	**95.5**	**75.0**	91.9	**97.8**	93.1	91.8	**94.0**	**96.3**	**90.0**
Gemini-2.5-flash	91.0	92.3	96.7	94.8	82.8	90.9	50.0	88.7	91.1	93.7	90.2	88.0	87.4	70.0
Gemini-2.5-flash-lite	55.7	59.0	83.0	67.0	27.6	27.3	37.5	50.2	57.9	57.8	45.8	59.4	70.4	20.0
Claude-sonnet-4	93.0	93.7	97.3	96.2	**93.1**	81.8	37.5	**95.6**	95.8	**98.7**	**97.0**	82.6	94.1	80.0
Claude-haiku-3	64.7	66.0	92.7	71.7	50.0	59.1	0.0	81.3	81.8	93.9	84.1	80.6	66.7	30.0
GPT-4.1	89.7	90.3	95.3	93.9	87.9	72.7	37.5	89.2	89.5	92.6	91.1	90.4	77.8	30.0
GPT-4.1-mini	78.0	79.3	91.7	85.4	70.6	40.9	37.5	88.2	89.4	94.2	89.7	89.0	80.7	70.0
GPT-4.1-nano	31.3	43.3	62.3	36.8	19.0	22.7	0.0	70.0	71.9	92.1	71.8	72.3	53.3	10.0
Qwen-3-8B	20.3	32.0	44.3	24.5	10.3	13.6	0.0	30.6	35.6	36.1	27.1	35.7	54.1	0.0
Qwen-3-32B	68.0	70.0	96.0	80.7	34.5	45.5	37.5	73.9	76.7	83.7	73.5	76.1	81.5	40.0
Ministral-8B	22.3	23.0	75.7	29.7	1.7	9.1	12.5	40.8	41.8	74.7	45.3	33.2	31.9	0.0
Mistral-small-24B	63.7	63.7	**97.7**	77.8	29.3	31.8	25.0	45.6	64.2	53.4	45.3	46.0	57.0	0.0
Llama-3.1-8B	27.0	29.0	68.3	35.3	6.9	9.1	0.0	47.5	47.8	85.5	47.8	49.9	43.7	10.0
Llama-3.3-70B	56.0	57.0	95.3	66.0	36.2	31.8	0.0	37.0	40.9	42.5	34.7	40.9	53.3	10.0
Gemma-3-12B	32.0	34.3	83.7	42.5	6.9	9.1	0.0	15.2	19.7	37.7	16.3	12.8	17.0	0.0
Gemma-3-27B	37.3	39.0	91.0	45.7	20.7	13.6	0.0	33.8	37.3	51.0	31.7	38.6	43.7	0.0

5.2.2 *Reasoning Complexity as the Primary Performance Differentiator.*

The analysis of performance across the four reasoning scenarios (A-D) reveals that task complexity is the most significant factor influencing model reliability.

Degradation in Multi-step Scenarios. While most models perform adequately on **Direct Lookup (A)**, accuracy systematically decreases as tasks require calculation and logical inference. The most pronounced performance drop occurs in the **Multivariate Calculation (D)** scenario, which serves as a stress test for multi-step reasoning. On this task, a significant number of models, including several with relatively higer parameter counts (e.g., Llama-3.3-70B, Mistral-small-24B), score at or near 0.0%. This indicates a fundamental breakdown in their ability to reason under complex context, leading to the fabrication of outputs.

Resilience in Top-Tier Models. In contrast, the top-performing models demonstrate notable, albeit imperfect, resilience on complex tasks. **Gemini-2.5-Pro** is particularly robust in the **Multivariate Calculation (D)** scenario, achieving 90.0% accuracy on the Main Split. **Claude-Sonnet-4** also performs strongly at 80.0%. While these results are promising and highlight advanced reasoning capabilities, a 10-20% failure rate on the most complex calculations represents a critical barrier. This difficulty in maintaining factual consistency through multi-step logic remains a primary vector for intrinsic hallucinations.

5.2.3 *Case Study.*

Beyond aggregate metrics, we perform a qualitative analysis to diagnose the models' specific failure modes. Our analysis reveals several recurring patterns, with one of the most significant being scale error. This occurs when a model correctly identifies a numerical value but fails to associate it with the proper magnitude (e.g., reporting "$150" instead of "$150 million"). Correcting for this single error type in Llama-3.3-70B's outputs, for instance, would boost its value accuracy from 37.0% to 57.7%, highlighting this as a critical vulnerability in contextual numerical grounding.

Table 4: Key financial attributes for Mohawk Commons used in the masked value inference task. *(Table headers and rows have been truncated for clarity.)*

Investment	Ownership %	Debt ($M)	Rate	Maturity
Gotham	49%	$8.5	8.36%	Mar 2024
Mohawk Commons	18.1%	$7.2	5.80%	Mar 2028
Total		**$188.8**		

However, more profound failures stem from a deficient understanding of financial concepts that require multi-step reasoning across different data sources. To illustrate these deeper challenges, we now present a representative detailed case study on a sample that required latent variable inference from both tabular and textual context. The target span was the equity investment value ($20.2 million) in the following scenario:

> *On January 27, 2023, Fund V acquired a 90% interest in an unconsolidated venture for* ***$20.2 million****, which purchased a shopping center, Mohawk Commons, located in Schenectady, New York, for $62.1 million, inclusive of transaction costs.*

Table 4 contains the relevant details for Mohawk Commons, including the 18.1% ownership and the pro-rata share of mortgage debt ($7.2M).

Model Reasoning Patterns: Only Gemini-2.5-Pro accurately inferred the masked value. Its reasoning chain involved several inferential steps: The model first identified the relevant table entry (Mohawk Commons) that matched the acquisition described in the text. The model then inferred the total mortgage debt on the asset by dividing the pro-rata debt ($7.2 M) by the ownership percentage

(18.1%), yielding \$39.7 M. Subtracting this value from the purchase price (\$62.1 M) gave the total equity (\$22.4 M), to which the model applied the Fund's 90% interest, ultimately arriving at the correct equity investment value of \$20.2 M.

In contrast, GPT-4.1 and Claude-Sonnet-4 failed to utilize the debt information from the table. Instead, both models based their prediction solely on the information in the sentence, by simply calculating 90% of the purchase price (62.1 M × 0.90 = 55.9 M). This neglects the mortgage debt and indicates insufficient integration of tabular data.

Implications for Multimodal Reasoning: This example reveals a key limitation in current LLM capabilities: the inference of latent variables that require synthesizing information across modalities and reasoning over implicit relationships. Only Gemini 2.5 Pro reconstructed the correct financial logic chain, whereas other models defaulted to predicting the masked span without grounding answers in the provided numerical data.

6 CONCLUSION AND FUTURE WORK

In this paper, we introduced a new dataset and a dynamic framework for evaluating hallucinations of LLMs in the financial domain, providing a nuanced view of their current capabilities. Our findings indicate that even state-of-the-art models struggle with the nuances of financial tabular data. While leading models are approaching the accuracy required for less critical applications, the fundamental challenges of ensuring factual integrity, particularly regarding numerical scale in complex reasoning, remain the primary barrier to their safe deployment in accuracy-critical financial workflows.

For future work, we plan to expand the benchmarking datasets to include more document types and more complex reasoning scenarios. Another promising direction is to study how factors like table size and the number of tables in the context affect hallucination rates. In summary, we believe that robust, domain-specific evaluation is a crucial step towards building more reliable and trustworthy LLMs for financial applications.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of National University of Singapore IT for providing access to its high-performance computing resources (Grant: NUSREC-HPC-00001).

REFERENCES

[1] Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica Salinas, Victor Alvarez, and Erwin Cornejo. 2024. HalluMeasure: Fine-grained Hallucination Measurement Using Chain-of-Thought Reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*. Association for Computational Linguistics, 15020–15037. https://doi.org/10.18653/V1/2024.EMNLP-MAIN.837

[2] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 967–976. https://doi.org/10.18653/v1/2023.findings-emnlp.68

[3] Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM Hallucination Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 24128–24156. https://doi.org/10.48550/ARXIV.2504.17550

[4] Yupeng Cao, Zhi Chen, Qingyun Pei, Nathan Lee, K. P. Subbalakshmi, and Papa Momar Ndiaye. 2024. ECC Analyzer: Extracting Trading Signal from Earnings Conference Calls using Large Language Model for Stock Volatility Prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance, ICAIF 2024*. ACM, 257–265. https://doi.org/10.1145/3677052.3698689

[5] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3697–3711. https://doi.org/10.18653/v1/2021.emnlp-main.300

[6] Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH-v2: Scaling Analytical Hallucination Annotation of Large Language Models. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2407.04693

[7] Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Knowledge-Centric Hallucination Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6953–6975. https://doi.org/10.18653/v1/2024.emnlp-main.395

[8] Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. *CoRR* (2023). https://doi.org/10.48550/ARXIV.2311.15548

[9] Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. BizBench: A Quantitative Reasoning Benchmark for Business and Finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 8309–8332. https://doi.org/10.18653/v1/2024.acl-long.452

[10] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6449–6464. https://doi.org/10.18653/v1/2023.emnlp-main.397

[11] 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)*. Association for Computational Linguistics, Dublin, Ireland, 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

[12] Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. HalluDial: A Large-Scale Benchmark for Automatic Dialogue-Level Hallucination Evaluation. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2406.07070

[13] 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*. Association for Computational Linguistics, 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

[14] Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows". *CoRR* (2024). https://doi.org/10.48550/ARXIV.2410.03727

[15] Monetary Authority of Singapore. 2024. Artificial Intelligence (AI) Model Risk Management. (dec 2024). Available at: https://www.mas.gov.sg/-/media/mas-media-library/publications/monographs-or-information-paper/imd/2024/information-paper-on-ai-risk-management-final.pdf.

[16] Syed Shariyar Murtaza, Yifan Nie, Elias Avan, Utkarsh Soni, Wanyu Liao, Adam Carnegie, Cyril John Mathias, Junlin Jiang, and Eugene Wen. 2025. Implementing Retrieval Augmented Generation Technique on Unstructured and Structured Data Sources in a Call Center of a Large Financial Institution. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 3: Industry Track, Albuquerque, New Mexico, USA, April 30, 2025*. Association for Computational Linguistics, 598–606. https://doi.org/10.18653/V1/2025.NAACL-INDUSTRY.48

[17] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 10862–10878. https://doi.org/10.18653/v1/2024.acl-long.585

[18] Jio Oh, Soyeon Kim, Junseok Seo, Jindong Wang, Ruochen Xu, Xing Xie, and Steven Euijong Whang. 2024. ERBench: An Entity-Relationship based Automatically Verifiable Hallucination Benchmark for Large Language Models. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2403.05266

[19] Sohini Roychowdhury. 2024. Journey of Hallucination-minimized Generative AI Solutions for Financial Decision Makers. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*. Association for Computing Machinery, 1180–1181. https://doi.org/10.1145/3616855.3635737

[20] Sohini Roychowdhury, Andres Alvarez, Brian Moore, Marko Krema, Maria Paz Gelpi, Punit Agrawal, Federico Martin Rodriguez, Angel Rodriguez, Jose Ramon

Cabrejas, Pablo Martinez Serrano, and Arijit Mukherjee. 2023. Hallucination-minimized Data-to-answer Framework for Financial Decision-makers . In *2023 IEEE International Conference on Big Data (BigData)*. IEEE Computer Society, 4693–4702. https://doi.org/10.1109/BigData59044.2023.10386232

[21] Bhaskarjit Sarmah, Dhagash Mehta, Stefano Pasquali, and Tianjie Zhu. 2024. Towards reducing hallucination in extracting information from financial reports using Large Language Models. In *Proceedings of the Third International Conference on AI-ML Systems (AIMLSystems '23)*. Association for Computing Machinery, Article 39. https://doi.org/10.1145/3639856.3639895

[22] Agam Shah, Liqin Ye, Sebastian Jaskowski, Wei Xu, and Sudheer Chava. 2025. Beyond the Reported Cutoff: Where Large Language Models Fall Short on Financial Knowledge. *CoRR* (2025). https://doi.org/10.48550/ARXIV.2504.00042

[23] Yan Wang, Yang Ren, Lingfei Qian, Xueqing Peng, Keyi Wang, Yi Han, Dongji Feng, Xiao-Yang Liu, Jimin Huang, and Qianqian Xie. 2025. FinTagging: An LLM-ready Benchmark for Extracting and Structuring Financial Information. *CoRR* (2025). https://doi.org/10.48550/ARXIV.2505.20650

[24] Pengzuo Wu, Yuhang Yang, Guangcheng Zhu, Chao Ye, Hong Gu, Xu Lu, Ruixuan Xiao, Bowen Bao, Yijing He, Liangyu Zha, Wentao Ye, Junbo Zhao, and Haobo Wang. 2025. RealHiTBench: A Comprehensive Realistic Hierarchical Table Benchmark for Evaluating LLM-Based Table Analysis. *CoRR* (2025). https://doi.org/10.48550/ARXIV.2506.13405

[25] Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Tongliang Li, Zhoujun Li, and Guanglin Niu. 2025. TableBench: A Comprehensive and Complex Benchmark for Table Question Answering. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*. AAAI Press, 25497–25506. https://doi.org/10.1609/AAAI.V39I24.34739

[26] Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. 2024. ReEval: Automatic Hallucination Evaluation for Retrieval-Augmented Large Language Models via Transferable Adversarial Attacks. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*. Association for Computational Linguistics, 1333–1351. https://doi.org/10.18653/V1/2024.FINDINGS-NAACL.85

[27] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 6588–6600. https://doi.org/10.18653/v1/2022.acl-long.454

[28] 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. *CoRR* (2023). https://doi.org/10.48550/ARXIV.2306.05685

[29] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics, 400–410. https://doi.org/10.18653/v1/2024.acl-demos.38

[30] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 3277–3287. https://doi.org/10.18653/v1/2021.acl-long.254

A PROMPT TEMPLATES

A.1 Financial Span Answerability Annotation

You are given the following tables from a 10-K annual report: `{tables_with_pretext}`
Filing date: {`filing_date`}
Sentence: {`pre_sentence`} {`sentence`} {`post_sentence`}
Your task:
For each highlighted span substring (shown as `<SPAN>{span}</SPAN>`) in the sentence, determine if the span is:

- unanswerable: Spans that
  - do not represent some type of numeric financial data (e.g., phone numbers, pincodes, or any other noise), or
  - the span cannot be derived from or supported by the table by any method.
- answerable: Spans that
  - can be directly found in the table, or
  - can be derived through some calculations (such as addition, subtraction, multiplication, or division), or
  - can be inferred via deeper reasoning involving multiple table entries

Instructions:

(1) Carefully analyze every given table and the given sentence.
(2) For each `<SPAN>{span}</SPAN>`, in the sentence: - Label the span as answerable or unanswerable
(3) Provide a concise explanation for your reasoning.

Output Format:

~~~
```json
{
 "reasoning": <A detailed explanation for each highlighted
  span in the sentence - if it is  answerable: how it can
  be matched, derived, or inferred from the table,
  or else if it  is unanswerable: a brief reason why.>
  "spans": {
    "<span1>" : "answerable" | "unanswerable",
    ...
  }
}
```
~~~

A.2 Financial Metric Prediction

You are a financial analyst.
Your task is to fill in the masked blank in a sentence using the provided financial data tables.
Hint: The masked content is a single positive value that fits the context `{unit_description}`.
Instructions:

(1) **Analyze the Request:** Carefully read the sentence and examine the provided tables to understand what information is needed to fill in the blank.
(2) **Reason Step-by-Step:** Before providing the final answer, write out your reasoning process. Explain how you will find the value, including any calculations.
   - If you are extracting a value, mention which table and cell you are getting it from.
   - If you are performing a calculation, show the formula and the values you are using.
(3) **Categorize Your Reasoning:** Based on your reasoning, classify it into one of the following scenarios:
   - **A. Direct Lookup:** The value is directly extracted from a single cell in a table.
   - **B. Simple Calculation (Single Metric):** The result is calculated from a single metric across different time periods, categories, or rows (e.g., calculating a year-over-year change).
   - **C. Simple Calculation (Two Metrics):** The result is calculated using two different metrics (e.g., calculating a ratio).
   - **D. Complex Calculation:** The reasoning involves more than two metrics or multiple complex steps.
(4) **Format the Final Answer:** After your reasoning, provide the final answer in a JSON block with the following structure.

JSON Output Format:

~~~
```json
{
"results": {
    "answer": "<The calculated or extracted value>",
    "scenario": "<A, B, C, or D>",
    "necessary_metrics": ["<metric_name_1>",
                          "<metric_name_2>", ...],
    "reference": ["<table_identifier_1>",
                  "<table_identifier_2>", ...]
   }
}
```
~~~

Field Explanations:

- ‘answer‘: The value to fill in the masked blank. Format it professionally (i.e., with rounding and units, etc.).
- ‘scenario‘: One of "A", "B", "C", or "D" based on your reasoning.

```
        - ‘necessary_metrics‘: A list of all metric names from the tables
          required to derive the answer.
        - ‘reference‘: A list of all table identifiers for the tables used.
**Inputs:**
**Tables:** {tables_with_pretext}
**Sentence:** {pre_sentence} {sentence} {post_sentence}
```