

Characterizing Ambiguity in Financial Document Question Answering

Kelsey Rook¹, Clare Arrington¹, Tomek Strzalkowski¹, Razvan Bunescu²

¹Rensselaer Polytechnic Institute

²University of North Carolina at Charlotte

{rookk, arrinj, tomek}@rpi.edu, rbunescu@charlotte.edu

Abstract

Large Language Models (LLMs) have demonstrated strong performance in question answering (QA) over general knowledge due to the wealth of information used in training. However, naturally occurring ambiguity in text poses significant challenges. Users may omit important contextual information from their queries, available context may be insufficient to answer a question without making assumptions, while frontier models tend to give confident answers rather than ask clarifying questions. Research on ambiguity in QA has focused primarily on factoid and general-knowledge settings that rely on LLM-internal knowledge, leaving a gap in research on domain-specific and document QA.

We examine how natural ambiguity arises in financial document QA, focusing on hybrid documents containing both text and tables. To this aim, we annotate 200 questions sampled from the MultiHiertt dataset and find that 77% of them exhibit some degree of ambiguity, with 52% of that ambiguity at least partially due to the data itself. We propose an extended taxonomy that accounts for ambiguity source (document vs. query) and degree. Upon evaluating frontier LLMs on ambiguity detection, we observe sharp drops in accuracy across the ambiguity spectrum and high false positive rates. Additionally, inter-model agreement is low, suggesting that models lack a consistent notion of ambiguity.

1 Introduction

Natural ambiguity in human dialogue arises from various linguistic and contextual factors, and is often a feature rather than a limitation. In many cases, information is intentionally underspecified to enhance the flow of conversation, or to allow conversation partners to take more initiative in directing the conversation topic. Ambiguity lets speakers say more with fewer words, and rely on

context to fill in the rest (Piantadosi et al., 2012). In other instances, ambiguity stems from a fault in assumed shared knowledge or accidental disfluencies; however, these are typically resolved interactively through follow-up questioning, leading to successful communication.

Ambiguities pose significant challenges for natural language QA systems. Standard evaluation paradigms expect precise outputs matching a single gold answer, leaving no room for multiple valid interpretations. While LLMs have increased resilience to surface-level linguistic variance, fundamental issues remain: LLMs frequently fail to distinguish among unambiguous, ambiguous, and unanswerable queries, and tend to prioritize giving an immediate answer over seeking clarification (Kuhn et al., 2022). This is compounded by traditional dataset annotation pipelines: when multi-annotator disagreement reflects genuine ambiguity, disputed questions are typically discarded or forced into a singular gold label, eliminating the distribution of possible answers (Jayaweera and Dorr, 2025).

Research on ambiguity in QA has sought to automate the detection and resolution of ambiguous queries (Keyvan and Huang, 2022; Liu et al., 2025b; Tanjim et al., 2025), primarily focusing on trivia and common-sense settings. Domain-specific document QA introduces additional complexity that these approaches do not address. Without full knowledge of a reference document, users may struggle to articulate precise information needs, making it difficult for agents to qualify intent based on a single query (Keyvan and Huang, 2022). In the financial domain, analysts must extract and synthesize information from dense documents containing text and tables that must be cross-referenced. The document structure can introduce ambiguity, and a query unambiguous in isolation may admit multiple defensible answers once attached to the context of a specific document. Currently, no taxonomy

distinguishes this type of ambiguity.

Financial data is a frequent choice for studying complex information extraction, logical reasoning, and mathematical reasoning due to the requirement of public companies to disclose extensive financial reports. The ecosystem of existing benchmarks synthesized from these reports makes finance a natural target for studying what existing evaluation practices miss (Chen et al., 2021; Zhu et al., 2021; Zhao et al., 2022). Existing benchmarks uniformly assume a single gold answer, ignoring nuances in user intent and data representation.

To address these gaps, we investigate the nature of ambiguity in financial document QA. This work makes the following contributions:

- We establish a performance baseline for financial document QA by evaluating three frontier LLMs (GPT-5.2, Haiku 4.5, Gemini 2.5 Pro) on the MultiHiertt (Zhao et al., 2022) dataset, demonstrating a substantial gap between model and human expert performance (Section 3).
- We sample and manually annotate 200 questions from MultiHiertt for ambiguity degree, type, and source, finding that a majority (77.5%) exhibit some degree of ambiguity and that a significant portion of that ambiguity (52.5%) originates due to the nature of the reference document rather than the query alone. We release a dataset including the original 200 samples alongside additional queries labeled as unambiguous and unanswerable ($N = 223$).
- We demonstrate that existing ambiguity taxonomies, developed for open-domain QA, fail to capture document-sourced ambiguities that arise in financial document QA, and we propose an extended two-dimensional taxonomy that addresses this gap.
- We evaluate the ability of frontier LLMs to detect and characterize ambiguities in this setting, and analyze the relationship between ambiguity and model error.

2 Related Work

2.1 Ambiguity in Question Answering

Ambiguity in QA is broadly defined as non-uniqueness, where an ambiguous question insufficiently expresses its determined meaning resulting in multiple plausible interpretations with different answers. The task of AmbigQA (Min et al., 2020), defined alongside the release of the AmbigNQ dataset, formalizes this distinction by asking sys-

tems to pair each plausible answer with a rewritten, disambiguated question. Ambiguous questions are distinguished from unanswerable questions, which lack any plausible answer within a given context. While some work specifically considers entity reference resolution as an ambiguity source (Xu et al., 2019), more recent work considers any word, phrase, or query with multiple interpretations (Tanjim et al., 2025). A further divide occurs between research that considers ambiguity as a property of the query alone or of its interaction with contexts. While some studies consider ambiguity as a query shortfall alone (Nakano et al., 2022; Xu et al., 2019), others show that questions can become ambiguous due to context including time, place, user profile, and document content (Liu and Roth, 2025; Zhang and Choi, 2021).

The majority of studies on ambiguity in QA operate within the single-turn QA task, in which the answering system is required to resolve ambiguities and return a precise answer without follow-up questioning. Recently, studies on ambiguity have focused on how Large Language Models (LLMs) perceive and manage uncertainty. The CLAMBER benchmark (Zhang et al., 2024) introduces a taxonomy that categorizes query uncertainty according to the following categories: *epistemic misalignment*: inherent knowledge within the LLM conflicts with understanding about the query, *linguistic ambiguity*: a statement can be interpreted in multiple ways due to unclear meaning (lexical or semantic ambiguity), *aleatoric output*: the query is well-formed but missing essential elements (who, where, when, what).

ConfuseBench (Liu et al., 2025a) distinguishes the following categories of uncertainty: *document scarcity*: the model lacks essential factual information to answer a question, *limited capacity*: the query is too complex for the model to resolve effectively, *query ambiguity*: the query itself is unclear. The AT-CoT prompting paradigm (Tang et al., 2025) proposes a three-category taxonomy of ambiguity: *semantic* (multiple valid interpretations), *generalize* (scope too narrow), and *specify* (scope too broad), designed to guide LLM actions in response to specific ambiguity types. Other studies examine the ability of LLMs to handle questions for which low-confidence answers are expected (Amayuelas et al., 2024), and assess their own ambiguity (Kim et al., 2024). Significant research on ambiguity has also occurred within the context of the text-to-SQL task (Wang et al., 2023;

Ding et al., 2025)

While existing taxonomies successfully capture several dimensions of query ambiguity, they are limited by their focus on the general knowledge and trivia domains. These taxonomies lack consideration for ambiguity source, which must be noted in tasks where the QA system must consider the user query and intent, its own internal knowledge, and additional knowledge sources (financial report or some other document). ConfuseBench’s taxonomy captures missing documents, not one that has multiple applicable elements; AmbiSQL covers structured schemas but not interleaved text and tables. Additionally, none of these taxonomies consider whether domain specificity introduces new types of ambiguity, limiting their broad applicability. The taxonomy proposed in this work aims to fill these gaps.

2.2 Financial Question Answering

Financial QA demands numerical and common-sense reasoning over heterogeneous documents with text, footnotes, and tables that frequently require cross-reference. FinQA (Chen et al., 2021) provides expert-annotated QA pairs over single tables and surrounding textual context from S&P earnings reports. ConvFinQA (Chen et al., 2022) expands FinQA with conversational chain-of-thought, and TAT-QA (Zhu et al., 2021) grounds questions in multiple tables and specifically selected paragraphs. The MultiHiertt (Zhao et al., 2022) benchmark emulates the most realistic scenario: questions span entire documents with multiple tables, requiring reasoning over multiple sections without prior knowledge of where information resides. Critically, all of these benchmarks assume a single gold answer.

Srivastava et al. (2024) evaluate a range of models including GPT-4 and MAMMoTH 13B on a range of financial benchmarks, achieving performance of up to 70% accuracy with GPT-4 on MultiHiertt using a specialized prompt. This is still well below the 83.1% human expert ceiling, a gap we hypothesize is partly attributable to unaccounted for ambiguity.

3 LLMs for Financial Document QA

We evaluate GPT-5.2, Haiku 4.5, and Gemini-2.5-Pro on MultiHiertt (Zhao et al., 2022) under direct prompting. Our definition of financial document QA is based on the problem

formulation established in FinQA (Chen et al., 2021), with minor changes such as the inclusion of a full financial report with multiple in-place tables, as follows:

Presented with a financial report R , consisting of text and multiple tables, and given a question Q , the task for an LLM is to auto-regressively generate an answer A . Results from multiple models can be found in Table 1, compared against human financial expert performance. MultiHiertt includes evaluation programs that use two metrics: exact match (EM), which measures the percentage of predictions that match the ground truth answer exactly, and a numeracy-focused F_1 score, which measures overlap between a bag-of-words representation of the predicted and gold answers. All models underperform on questions requiring arithmetic relative to span extraction.

Notably, applying a tolerance of ± 0.005 to arithmetic answers produces a large gain in performance; the highest increase occurs for GPT 5.2, achieving $F_1 = 0.48$ with a tolerance of 0.005. Full F_1 scores with various degrees of numeric tolerance can be seen in Table 6. This suggests that many errors are floating-point imprecisions rather than reasoning failures, a direct consequence of internalizing multi-step arithmetic within the forward pass rather than delegating to an external calculator. However, the best model (Haiku 4.5, $EM = 0.54$) still falls 29 points below human expert performance. We examine whether ambiguity contributes to this gap.

While MultiHiertt presents the most realistic scenario (QA over full documents with multiple tables), the dataset itself contains errors and ambiguities. This is partially substantiated by the human performance and inter-annotator agreement found in the original paper. Analysis of LLM performance and manual human-evaluation of the dataset reveals a varying degree of ambiguities in many questions. This poses issues as the MultiHiertt dataset is designed as a textual and numerical reasoning dataset in which queries result in a single numeric or concise textual answer (yes/no, single entity name, etc), that leaves no room for ambiguities to be acknowledged.

4 Ambiguity Taxonomy

Ambiguity is a common feature in QA datasets constructed from naturally generated sources. Early research on web search behavior indicates that ap-

	Arithmetic $EM = F_1$	Span $EM \quad F_1$		Combined $EM \quad F_1$	
Human Expert	–	–	–	83.1	87.0
MT2Net (RoBERTa-large)	–	–	–	36.2	38.4
Gemini 2.5 Pro	38.4	59.6	64.2	41.8	42.6
Haiku 4.5	51.4	66.8	73.2	54.2	55.6
GPT 5.2	27.9	59.4	65.5	34.0	35.2

Table 1: Performance comparison of baseline and frontier models on MultiHiertt using direct prompting. EM = exact match; F_1 = numeracy-aware F1.

proximately 16% of queries were naturally ambiguous (Song et al., 2007). More recently, Min et al. (2020) found that the average human inter-annotator agreement rate was 49.2% in the Natural Questions (NQ-OPEN) dataset, constructed from real Google search queries, largely due to ambiguous questions. Beyond open-domain text, Gomm et al. (2025) recently identified high rates of ambiguity across 15 tabular QA, text-to-SQL, and data analysis benchmarks in various domains.

MultiHiertt is no exception to the ambiguity trends noticed above. Manual inspection reveals a substantial proportion of questions with multiple valid interpretations (Section 5.1). However, while previous ambiguity taxonomies focus on ambiguity types, we find two crucial dimensions along which ambiguity can be classified in the document QA setting: ambiguity source (query or document) and ambiguity severity. We propose an extension to existing taxonomies for document QA based on these two new dimensions.

4.1 Ambiguity Source

Min et al. (2020) argue that ambiguity is a function of the interaction between question and evidence corpus, not of the question alone. In document QA, this is particularly consequential. A question that sounds unambiguous on its own may have multiple defensible answers following consultation of the reference document. We identify three categories based on source: query, data, or both.

Query ambiguity: The query has multiple interpretations; all of which can be answered for any given document. For example, in *"What is the total of Granted Exercised - Canceled and Expired in 2006 for Options for Shares?"*, the dash could be read as subtraction or as a list separator, yielding different results regardless of the accompanying document’s contents.

Data ambiguity: The query has multiple interpretations, and the document’s structure or content determines which interpretations are answerable. There exists some document for which the question would have a unique answer; it is the specific document that creates the ambiguity. For example, the query *"What is the sum of CET1 capital, Tier 1 capital, and total capital in 2017?"* may appear ambiguous on initial inspection to someone with enough financial knowledge to recognize that gross vs. net values were not specified. The question could be disambiguated by a document containing net values alone; however, the source document contains both gross and net values, resulting in an underspecified scope.

Both (Query and Data): The query has multiple interpretations. Some interpretations are answerable for any given document, and for at least one interpretation, whether it is answerable depends on the specific document. For example, the query *"What was the average of the Total operating revenues in the year where other investment income is positive? (in million)?"* allows interpretations where:

- "year" is a typo for "years"; the intent is the average across total operating revenues for years meeting the condition. However, data for 2011, a year which meets the condition, is missing, making this interpretation unanswerable.
- "average" is meant as a mean across investment sub-categories within a single year, but the document lists several overlapping subcategories, leaving the scope of the average ambiguous and in need of additional clarification.

Query-sourced ambiguity can in principle be resolved by rewriting the question, while data-sourced ambiguity requires knowledge of the specific document and cannot be anticipated from the query alone. This distinction influences the ap-

propriate resolution strategy, and has no analog in existing taxonomies.

4.2 Ambiguity Levels

The severity of ambiguity partially depends on the evaluation setting. In a conversational setting, many ambiguities are benign, as a model can present multiple candidate answers with brief explanations, allowing the user to obtain more context and select the one that aligns with their intent. In a single-turn setting against a fixed gold label, the same questions have a high chance to cause failures as the model is penalized for a valid interpretation with no mechanism to expose the ambiguity. We distinguish among three levels of ambiguity.

1. **Unambiguous:** There are no conditions unstated in the query that, when changed, also change the correct answer or add additional interpretations. The evaluation setting is largely irrelevant.

2. **Inferable:** Multiple interpretations exist but the ambiguity does not require follow-up dialogue to handle in practice. We present two sub-cases:

(2.a) **Representational:** Multiple interpretations align with the user’s broad intent, and would be acceptable to most human interlocutors. A common example are queries asking for a rate of change, in which the intended meaning is unambiguous, but the final answer representation may differ. In a conversational setting, these cases are benign. For example, consider the question *"What is the mathematical range for average train speed (mph) for 2008-2010?"* The answer could be expressed as "23.5-27.3 mph", the upper and lower bound for that time period, or "3.8 mph", the range obtained by subtracting the lower from the upper bound. In single-turn evaluation settings, proper normalization can typically resolve the issue of undue penalties; however, consideration must be taken in the creation of datasets and scoring programs to account for multiple possible representations of an answer, when both are correct and helpful.

(2.b) **Pragmatic:** One or more technically correct answers coexist with the clearly intended one, and domain knowledge or communicative context strongly favors the intended read-

ing. For example, consider the question *"In the year with the largest amount of 'Office Net Charge-offs', what's the increasing rate of 'Office Net Charge-offs'?"* The associated document reveals that the rate of 'Office Net Charge-offs' is actually decreasing in the year that meets the condition. While flagging the question as unanswerable due to a contradiction is technically correct, determining that the user overspecified and instead providing the decreasing rate is ultimately more useful.

3. **Ambiguous:** There are multiple interpretations of user intent, each which differing answers. In a conversational setting, the appropriate response is a follow-up question to resolve unspecified conditions of the original question. In single-turn evaluation, the model must select an interpretation based on the most likely intent, though this retains a chance of failure. Additional ambiguity occurs for questions with multiple interpretations due to differences of opinions, expertise, or knowledge. This includes questions about "known unknowns", as addressed by [Amayuelas et al. \(2024\)](#) and discussed in Section 2.1.

4.3 Relation to Existing Taxonomies

In their framework of QA ambiguity for tabular data analysis, [Gomm et al. \(2025\)](#) identify the need to categorize query types more finely for precise evaluation of a system’s ability to handle ambiguity. A system must distinguish between cooperative and uncooperative ambiguity, where cooperative ambiguity can be resolved through reasonable inference. However, scenarios with reference information make the cooperation category less clear. The cooperativity of a query will vary based on the information available, but when the user does not know the precise nature of the information in the reference document, their intent stays the same. AT-CoT’s categories (semantic, generalize, specify) all belong to the category of query-sourced ambiguity, but their taxonomy does not have any analog for cases of data-sourced ambiguity. While some works acknowledge the issue of data-sourced ambiguity, it is currently not integrated into any studies that have formalized taxonomies. Finally, while some works distinguish between ambiguous and unanswerable questions, others include unanswerable questions in their taxonomy ([Zhang et al., 2024](#)).

5 Experimental Setup

5.1 Data Annotation

Based on the LLMs from Table 1 and their generated answers with explanations, we conducted an initial pass through MultiHiertt by recording model agreement and how many models answered correctly (out of 3) for each QA pair. We sampled ~ 200 pairs, annotating according to our ambiguity dimensions and providing a text description of the precise source of ambiguity. Additionally, we assigned ambiguity types from the AT-CoT and CLAMBER taxonomies in order to assess their coverage. These taxonomies were chosen due to their broad applicability and complementary goals. While AT-CoT prioritizes mutual exclusivity and the actionability of each category, CLAMBER captures LLM-specific failure modes such as unfamiliar entities and contradictions.

Mapping inferable and fully ambiguous questions to singular categories in the AT-CoT or CLAMBER taxonomy proved difficult. Most often, the ambiguity originated from the document rather than the query itself, a distinction neither taxonomy encodes. This gap reflects a difference in design goals. The majority of existing taxonomies are action-oriented based on the ambiguity type, or to serve as tools for studying QA system performance on the ambiguity types most frequently observed in a particular dataset, rather than to be mutually exclusive and full-coverage.

Following this initial annotation pass, we observed that 77% of the randomly sampled questions contained some level of ambiguity. In order to provide an even sampling across data categories, we added additional unambiguous samples. The resulting dataset serves as the basis for the remainder of our analysis, and is available at <https://github.com/shmewtep/MultihierTTAmbig>.

5.2 Ambiguity Detection Task

Given a financial report R , consisting of text and multiple tables, and a question Q , a model is tasked with identifying all plausible interpretations of Q and producing, for each interpretation I_j , a candidate answer along with an estimated probability $P(I_j|Q)$ that I_j reflects the question’s intended meaning, where $\sum_{j=1}^m P(I_j|Q) = 1$. The model’s output is thus a set of m interpretation-answer-probability triples $\{(I_j, A_j, P(I_j|Q))\}_{j=1}^m$, and the overall answer distribution is given by:

Degree	N	GPT	Gemini	Haiku
Unambiguous	73	52.1	74.0	60.3
Inferable (Repr.)	31	38.7	60.0	71.0
Inferable (Prag.)	42	28.6	32.5	45.2
Ambiguous	71	12.7	11.3	11.3

Table 2: QA accuracy by annotated degree of ambiguity for GPT 5.2, Gemini 2.5 Pro, and Haiku 4.5.

Source	N	GPT	Gemini	Haiku
Data	63	30.2	26.2	36.5
Query	39	21.1	45.9	42.1
Both	42	14.6	14.6	22.0

Table 3: QA accuracy by annotated source of ambiguity for GPT 5.2, Gemini 2.5 Pro, and Haiku 4.5.

$$P(A | R, Q) = \sum_{j=1}^m P(A | R, Q, I_j) P(I_j | Q) \quad (1)$$

where I_j is a plain-text description of the j -th plausible interpretation of Q , and $P(A|R, Q, I_j)$ is the model’s answer distribution conditioned on interpretation I_j and report R . By observing the set of interpretation-answer-probability triples, we can estimate the ambiguity of a QA pair according to the model used.

6 Results

6.1 Ambiguity and Model Error

We evaluate QA accuracy against our annotated data based on two sets of ambiguity labels: the degree (or level) and the source.

Ambiguity Level Table 2 shows QA accuracy stratified by annotated degree of ambiguity. Across all models, accuracy degrades monotonically from unambiguous to fully ambiguous questions. The contrast between the two inferable sub-cases is particularly informative: two models (Haiku and Gemini) performed well on questions with *representational* ambiguity, comparable to or exceeding unambiguous performance. Questions with *pragmatic* ambiguity, which require the model to recognize that a literal reading is unhelpful, are substantially harder. This divergence confirms that the two sub-cases impose different demands and should not be conflated in evaluation.

Ambiguity Source For two out of three models, Gemini and Haiku, accuracy when data-sourced ambiguities are present is generally lower than

query-sourced. GPT 5.2, however, has greater success in resolving data-sourced ambiguities compared to query-sourced ambiguities. Instances with ambiguity from both sources produces the lowest accuracy across all models.

6.2 Ambiguity Detection

To measure how much uncertainty a model expresses about its own interpretations, we compute the Shannon entropy of its predicted probability distribution over candidate interpretations for each question. Intuitively, entropy is low when a model concentrates its probability mass on a single interpretation, and high when that mass is spread more evenly across several. Formally, for a query Q with interpretation set I , we compute:

$$H(I | Q) = - \sum_{j=1}^m P(I_j | Q) \log P(I_j | Q)$$

Confidence entropy is a statistically significant predictor of error across all models: incorrect predictions exhibit substantially higher entropy than correct ones (Haiku: 1.026 vs. 0.630 bits; GPT: 1.228 vs. 1.001 bits; Gemini: 0.555 vs. 0.302 bits). Gemini’s consistently lower absolute entropy values reflect its tendency toward higher top-answer confidence overall, but the direction and significance of the relationship holds across all three models.

Figure 1 shows the rank in which the correct answer appears in the resulting answer-probability distribution. For each model, we show the proportion of questions in which the gold answer appears as the top-ranked candidate interpretation (Rank 1), a lower-ranked candidate (Rank 2 or Rank 3+), or does not appear among the model’s predicted candidates at all (Incorrect). Across all three models, the gold answer is absent from the candidate

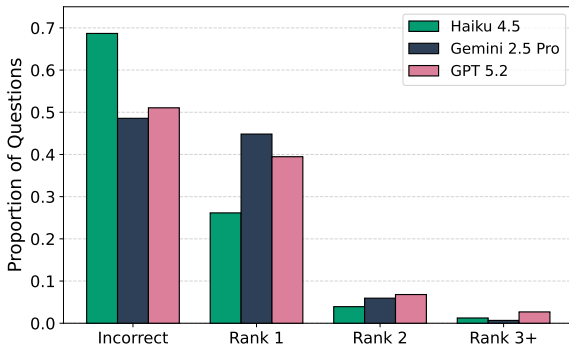


Figure 1: Distribution of gold answer ranking within ambiguity detection results by model.

FP Definition	GPT	Gemini	Haiku
Multi-Answer	100.0	57.5	83.6
Low Confidence	76.7	6.8	47.9

Table 4: False positive rates (%) on unambiguous questions ($N = 73$) under two definitions of over-predicted ambiguity: (1) multi-answer prediction and (2) top-candidate confidence < 0.80 .

set entirely for the largest share of questions, underscoring that the dominant failure mode is not misranking a correct interpretation but failing to generate it at all. Where a model does produce the correct interpretation, it is likely to rank it first.

False positive rates. We find that LLMs consistently over-predict on unambiguous questions, though the degree varies substantially by model, as shown in Table 4. GPT is the most aggressive, generating multiple candidate answers for 100% of unambiguous questions. Haiku similarly presents multiple possible answers for the majority of questions, but indicates less ambiguities in its spread of probabilities, suggesting that it accounts for more edge cases but recognizes that the probability of their being correct or useful is low. Gemini performs notably better under the confidence-based definition: only 6.8% of unambiguous questions fall under the 0.80 top-confidence threshold, reflecting higher-confidence outputs. More comprehensive results including calibration analysis, as well as precision and recall for individual ambiguity categories and across categories can be found in Appendix C.

Inter-model agreement. Pairwise agreement on ambiguity detection is low across all model pairs, which holds regardless of which ambiguity metric is used. Under multiple-answer output, κ ranges from 0.002–0.324 across model pairs; under confidence-based detection, κ ranges from 0.019 to 0.333. These ranges overlap substantially, and the ordering of pairs is consistent across both definitions—agreement is highest between Gemini and Haiku, and lowest between GPT and Gemini. These results confirm that models do not share a stable notion of what constitutes an ambiguous question, and that any single model’s flag for ambiguity should not be considered reliable. More detailed inter-model agreement metrics can be found in Appendix C.

Performance across AT-CoT and CLAMBER categories. Under both taxonomies, performance on ambiguous questions is low and inconsistent across models and categories (0–36.4% for individual AT-CoT categories; 0–66.7% for CLAMBER, with the high-end figures reflecting tiny sample sizes $N \leq 3$). No consistent accuracy ordering emerges across categories, and within-category variance is high. This is consistent with our argument that AT-CoT and CLAMBER categories, developed for open-domain QA, do not cleanly discriminate the ambiguity types present in financial document QA, and that classification under these schemes does not reliably predict model performance.

7 Conclusion

We introduced two dimensions that extend existing taxonomies with a document-sourced ambiguity category absent from prior work, and demonstrated that the split between full ambiguity and the types of inferable ambiguity predicts meaningfully different accuracy levels. We show that LLM prediction of ambiguity is poor, due to high false positive rates in detecting ambiguity, and inter-model agreement is low to moderate, demonstrating a lack of consistency in what is flagged as ambiguity in frontier models. However, models articulate the source of ambiguity most accurately in the cases they fail to resolve, suggesting that recognizing and resolving ambiguity are substantially different tasks, particularly under single-turn constraints.

We observe that single-gold-answer evaluation causes ambiguity to manifest as a silent failure: a model is penalized for choosing an interpretation the annotator did not anticipate (Min et al., 2020). Our annotation of MultiHiertt shows this is not a rare edge case. 77.5% of questions exhibit some ambiguity, and over half of that ambiguity is at least partially data-sourced, originating in the document rather than the query and therefore undiscoverable without consulting the specific reference material. Model accuracy drops significantly from unambiguous to ambiguous questions, and confidence entropy (κ) is a statistically significant predictor or error across all models.

These findings motivate the treatment of ambiguity as a prioritized annotation in future benchmarks, as it is clearly a standing issue in existing datasets. Annotations of ambiguity degree enable more comprehensive evaluation and support the development

of clarification-seeking systems. Critically, they make inferable ambiguities distinguishable from genuinely ambiguous questions. Format-variant and interpretation-variant questions are resolvable through post-processing or through contextualization in a conversational setting, and should be distinguished from questions that require additional turns of clarification dialogue to clarify the user’s assumptions. Conflating all three as incorrect with no further nuance misrepresents model capability and obscures actionable directions for resolution.

8 Limitations

Due to the scope of this project, we relied primarily on secondary financial data and established literature rather than direct consultation with industry financial experts. This is a primary limitation particularly in our labeling of MultiHiertt; future work should include re-annotation from multiple financial experts and reporting of inter-annotator agreement. While we develop and evaluate this taxonomy on financial document QA, we hypothesize that the source dimension (query, data, or both) generalizes to document-grounded QA more broadly since the underlying mechanism of a query’s interpretability depending on a specific retrieved document’s structure is not specific to financial reports. We are more cautious about the severity dimension: the Pragmatic Ambiguity category’s reliance on domain expertise to disambiguate a technically valid reading from the intended one suggests it may extend most naturally to other expert or technical domains (e.g., legal or medical document QA) rather than to open-domain QA generally, where this kind of domain-expertise gap is less consistently present. We leave empirical validation of this generalization to future work. A more thorough validation of this taxonomy also requires a more operationalized version of the taxonomy, which might include different concrete criteria for the Pragmatic Ambiguity classification depending on the domain.

Additionally, this work is limited in its evaluation since we only consider questions from the MultiHiertt dataset. A larger scale analysis on the extent and coverage of these ambiguity types would provide a richer understanding on how LLMs handle ambiguity in financial QA settings. However, existing financial document QA datasets may not reflect questions that would be asked in real scenarios, since questions in FinQA, MultiHiertt, etc. are designed from sample documents rather than

arising naturally. As such, evaluation would benefit from the inclusion of data from a natural source like earnings calls, which are conducted after public companies release earnings reports. Calls typically consist of a scripted presentation followed by a Q&A session, and transcripts are made publicly available afterwards. This dataset could be an interesting complement to datasets with single gold answers such as MultiHieertt, as the questions are largely more open-ended than single-number answers, and responses may include clarifications, discussion, and repairs.

Acknowledgments

This paper is based upon work supported in part by the United States Air Force under Contracts No. FA8750-21-C-0075 and FA8750-23-F-0080. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of USAF or the U.S. Government.

References

- Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. [Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432, Bangkok, Thailand. Association for Computational Linguistics.
- 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*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. [ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zhongjun Ding, Yin Lin, Tianjing Zeng, Rong Zhu, Bolin Ding, and Jingren Zhou. 2025. Ambisql: Interactive ambiguity detection and resolution for text-to-sql. *arXiv preprint arXiv:2508.15276*.
- Daniel Gomm, Cornelius Wolff, and Madelon Hulsebos. 2025. Are we asking the right questions? on ambiguity in natural language queries for tabular data analysis. *arXiv preprint arXiv:2511.04584*.
- Chathuri Jayaweera and Bonnie Dorr. 2025. From disagreement to understanding: The case for ambiguity detection in nli. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 37–46.
- Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. *ACM Computing Surveys*, 55(6):1–40.
- Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. [Aligning language models to explicitly handle ambiguity](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1989–2007, Miami, Florida, USA. Association for Computational Linguistics.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.
- Jingyu Liu, JingquanPeng JingquanPeng, Xiaopeng Wu, Xubin Li, Tiezheng Ge, Bo Zheng, and Yong Liu. 2025a. [Do not abstain! identify and solve the uncertainty](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17177–17197, Vienna, Austria. Association for Computational Linguistics.
- Siyi Liu and Dan Roth. 2025. [Conflicts in texts: Data, implications and challenges](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10073–10091, Suzhou, China. Association for Computational Linguistics.
- Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. 2025b. A survey of text-to-sql in the era of llms: Where are we, and where are we going? *IEEE Transactions on Knowledge and Data Engineering*.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Yuya Nakano, Seiya Kawano, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. 2022. [Pseudo ambiguous and clarifying questions based on sentence structures toward clarifying question answering system](#). In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 31–40, Dublin, Ireland. Association for Computational Linguistics.

- Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.
- Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2007. Identifying ambiguous queries in web search. In *Proceedings of the 16th international conference on World Wide Web*, pages 1169–1170.
- Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, and Dan Roth. 2024. [Evaluating LLMs’ mathematical reasoning in financial document question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3853–3878, Bangkok, Thailand. Association for Computational Linguistics.
- Anfu Tang, Laure Soulier, and Vincent Guigue. 2025. Clarifying ambiguities: on the role of ambiguity types in prompting methods for clarification generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 20–30.
- Mehrab Tanjim, Yeonjun In, Xiang Chen, Victor Burszty, Ryan A. Rossi, Sungchul Kim, Guang-Jie Ren, Vaishnavi Muppala, Shun Jiang, Yongsung Kim, and Chanyoung Park. 2025. [Disambiguation in conversational question answering in the era of LLMs and agents: A survey](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9537–9550, Suzhou, China. Association for Computational Linguistics.
- Bing Wang, Yan Gao, Zhoujun Li, and Jian-Guang Lou. 2023. [Know what I don’t know: Handling ambiguous and unknown questions for text-to-SQL](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5701–5714, Toronto, Canada. Association for Computational Linguistics.
- Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. [Asking clarification questions in knowledge-based question answering](#). 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 1618–1629, Hong Kong, China. Association for Computational Linguistics.
- Michael Zhang and Eunsol Choi. 2021. [SituatQA: Incorporating extra-linguistic contexts into QA](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. [CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand. Association for Computational Linguistics.
- 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)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.
- 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)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Ambiguity Annotations

Ambiguity Category	Example Question	Annotator Note	AT-CoT	CLAMBER
Unambiguous	For December 31, 2009, what was the total value of segregated collateral for the benefit of brokerage customers, in millions?	—	N/A	N/A
Inferable (Representational)	What was the change in advertising costs from 2001 to 2002?	"Change" could refer to the difference between values, or change in percentage.	Specify	Linguistic ambiguity (lexical)
Inferable (Pragmatic)	What is the proportion of Pensions' Service cost of net periodic benefit cost in U.S. Plans to the total in 2006?	It is unclear whether the "2006" condition also applies to the first clause of the query.	Specify	Linguistic ambiguity (semantic)
Fully Ambiguous	What is the sum of the Consumers in the section where Consumers is greater than 1 million?	"Section" is ambiguous.	Generalize	Aleatoric output (what)
Fully Unanswerable	As of February 8, 2019, what was the number of shares outstanding?	Asks for number of outstanding shares but the only available data is about shareholders.	Specify	Aleatoric output (what)

Table 5: Representative example questions from MultiHiertt spanning the four ambiguity degrees and additional unanswerable case, with justification for the presence of ambiguity and the corresponding AT-CoT and CLAMBER taxonomy labels.

B Question Answering Task

Model	0.000	0.001	0.005	0.010	0.050	0.100
Haiku 4.5	51.43%	51.66%	58.79%	59.74%	61.88%	63.54%
Gemini 2.5 Pro	38.44%	40.33%	54.52%	55.65%	57.41%	58.92%
GPT 5.2	27.91%	31.00%	47.98%	48.81%	50.48%	52.49%

Table 6: F1 score sensitivity to numeric-match tolerance on arithmetic questions.

C Ambiguity Detection Task

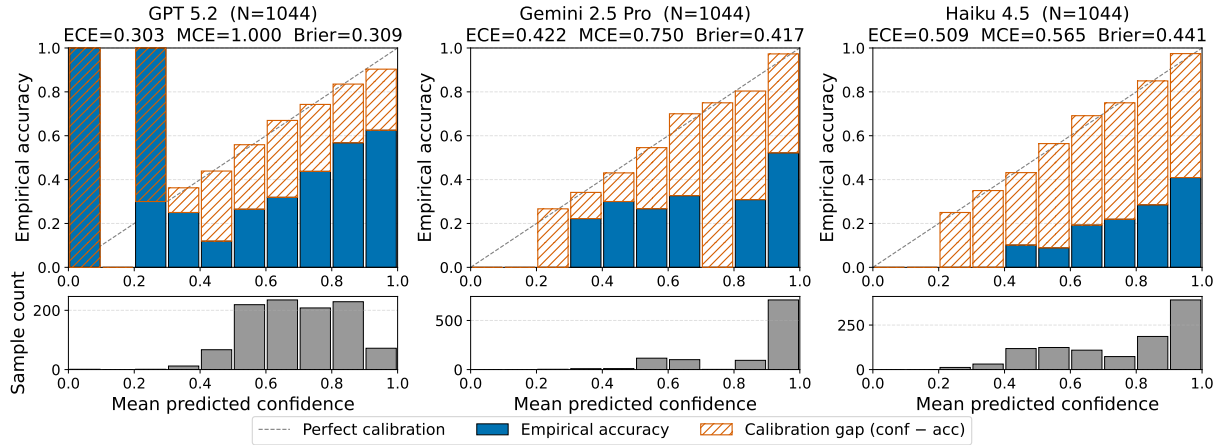


Figure 2: Answer-confidence calibration of the ambiguity-prompting outputs. For each question, the model’s highest-confidence candidate answer is scored against the MultiHiertt gold answer and binned into confidence intervals. All three models are substantially overconfident (points well below the diagonal).

Definition	Model Pair	Agreement (%)	κ
Multiple Candidate Answers	GPT 5.2, Gemini 2.5 Pro	53.6	0.0022
	GPT 5.2, Haiku 4.5	75.7	0.0059
	Gemini 2.5 Pro, Haiku 4.5	67.4	0.3242
	All 3 Models	48.4	—
Top-Candidate Confidence < 0.80	GPT 5.2, Gemini 2.5 Pro	50.4	0.1918
	GPT 5.2, Haiku 4.5	65.1	0.3325
	Gemini 2.5 Pro, Haiku 4.5	68.4	0.3298
	All 3 Models	42.0	—

Table 7: Pairwise inter-model agreement on ambiguity detection under two metric definitions.

Model	Definition	Precision (strict)	Recall (strict)	Recall (lenient)	F1 (strict)
GPT 5.2	Multi-Answer	0.493	1.000	1.000	0.660
	Conf. < 0.90	0.497	1.000	0.993	0.664
	Conf. < 0.80	0.541	0.930	0.900	0.684
Gemini 2.5	Multi-Answer	0.600	0.887	0.800	0.716
	Conf. < 0.90	0.800	0.789	0.613	0.794
	Conf. < 0.80	0.894	0.592	0.440	0.712
Haiku 4.5	Multi-Answer	0.516	0.915	0.920	0.660
	Conf. < 0.90	0.566	0.901	0.867	0.696
	Conf. < 0.80	0.615	0.789	0.687	0.691

Table 8: Ambiguity detection by definition and model. Precision and F1 use the strict ground truth (ambiguous vs. unambiguous); strict recall counts only fully-ambiguous questions, lenient recall also credits detection of inferably-ambiguous and unanswerable questions.

Model	Definition	Unambiguous	Represent.	Pragmatic	Ambiguous
GPT 5.2	Multi-Answer	100.0	100.0	100.0	100.0
	Conf. <.90	98.6	96.8	100.0	100.0
	Conf. <.80	76.7	80.6	92.9	93.0
Gemini 2.5 Pro	Multi-Answer	57.5	64.5	83.3	88.7
	Conf. <.90	19.2	41.9	50.0	78.9
	Conf. <.80	6.8	25.8	33.3	59.2
Haiku 4.5	Multi-Answer	83.6	87.1	97.6	91.5
	Conf. <.90	67.1	77.4	92.9	90.1
	Conf. <.80	47.9	41.9	73.8	78.9

Table 9: Detection rate by annotated ambiguity category (%). For *Unambiguous*, this is the false-positive rate; the remainder are recall. Category sizes: Unambiguous $N=73$, Representational $N=31$, Pragmatic $N=42$, Ambiguous $N=71$.