Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.086343

ARTICLE

Do LLMs Know When Evidence is Insufficient? An Evidence Sufficiency Benchmark for Answer-Abstention Calibration in Retrieval-Augmented Generation

Hantian Zhang[1] **and Wentai Wu**[2,*]

[1]Department of Mathematics, Jinan University, Guangzhou, China
[2]Department of Computer Science, Jinan University, Guangzhou, China

*Corresponding Author: Wentai Wu. Email: wentaiwu@jnu.edu.cn

Received: 28 May 2026; Accepted: 29 June 2026; Published: 13 August 2026

ABSTRACT: Large language models (LLMs) are increasingly used in retrieval-augmented generation (RAG) systems, where they are expected to answer questions based on retrieved evidence. In many cases, however, the right behavior is not to answer. A model should abstain when the evidence is insufficient, irrelevant, or contradictory. Existing evaluations mainly focus on final-answer accuracy, and they often pay less attention to whether models can recognize evidence quality before responding. To study this problem, we propose the Evidence Sufficiency Benchmark, a five-level benchmark for evaluating answer-abstention calibration. The benchmark covers evidence conditions from L1 Full Support to L5 Conflicting Evidence, including fully supportive, partially supportive, irrelevant, absent, and conflicting evidence. We evaluate seven LLMs from five families on the full L1–L5 gradient under three prompting strategies. The results show that current LLMs still have clear limitations in evidence-based abstention. Under L5 conflicting evidence, all evaluated models show high over-answer rates, ranging from 65% to 91%. The evidence sufficiency curves show that models reduce their answer rates as evidence quality decreases, but their abstention behavior remains unreliable. Chain-of-thought prompting improves abstention for some models, although the effect is not consistent across model families. Human validation on 200 samples further supports the reliability of the automatic evaluation. Overall, our findings suggest that current LLMs still struggle to recognize when evidence is insufficient in RAG settings.

KEYWORDS: Large language models; retrieval-augmented generation; evidence sufficiency; abstention calibration; over-answering; benchmark evaluation

1 Introduction

Retrieval-augmented generation (RAG) has become a standard paradigm for improving the factuality and traceability of large language model (LLM) outputs by conditioning generation on retrieved evidence [1–3]. Existing RAG evaluation frameworks mainly measure whether the retrieved context is relevant and whether the generated answer is faithful or correct [4,5]. However, real retrieval results are often imperfect: evidence may be incomplete, irrelevant, missing, or mutually contradictory [6,7]. In such cases, a reliable model should not merely produce a plausible answer; it should recognize when the evidence is insufficient and abstain.

This paper studies this problem as *evidence sufficiency calibration*: whether an LLM can calibrate its answer-abstention behavior according to the quality of externally provided evidence. This differs from standard factuality evaluation, which focuses on final-answer correctness, and from binary unanswerable

QA, which treats answerability as a two-way distinction [8,9]. Although LLMs encode substantial parametric knowledge [10,11], such knowledge cannot substitute for faithfully reasoning about external evidence. We instead ask how model behavior changes across a controlled gradient of evidence quality.

To this end, we propose the *Evidence Sufficiency Benchmark*, a five-level benchmark for evaluating answer-abstention behavior in RAG-style settings. The five evidence levels are: L1 Full Support, where the evidence fully supports the answer; L2 Partial Support, where the evidence provides weaker but still useful support; L3 Irrelevant Evidence, where the context is unrelated; L4 No Context, where no external evidence is provided; and L5 Conflicting Evidence, where the evidence contains contradictions or supports a conflicting answer. Fig. 1 gives an overview of our benchmark construction, prompting, and behavioral evaluation pipeline.

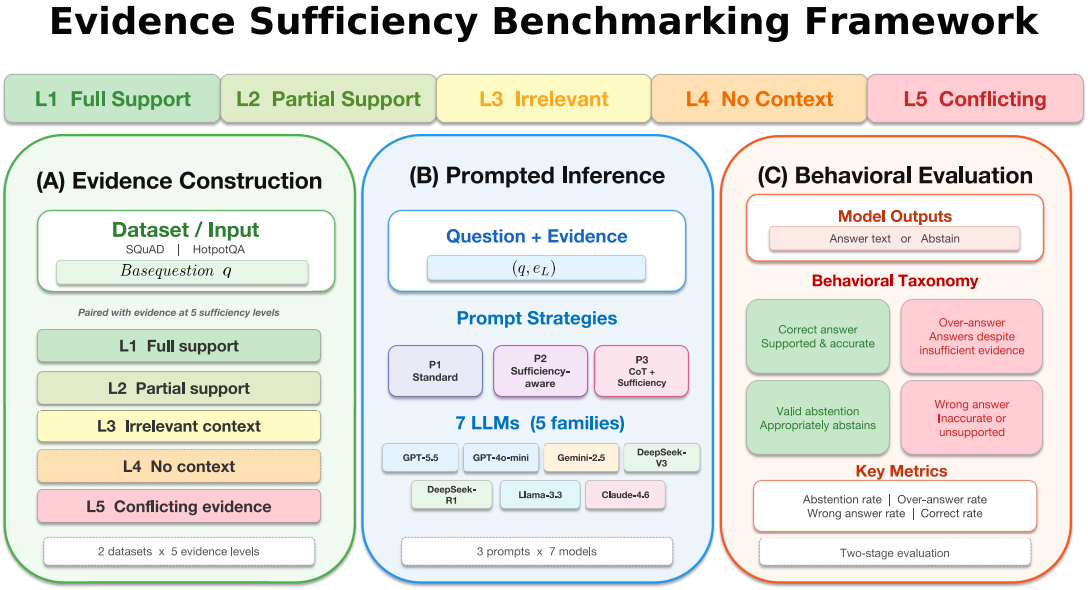

Figure 1: Overview of the evidence sufficiency benchmark. We construct five levels of evidence conditions from L1 full support to L5 conflicting evidence, evaluate LLMs under three prompting strategies, and measure their answer/abstention behaviors through a behavioral taxonomy and key metrics.

We evaluate seven models from five families on the full L1–L5 evidence sufficiency gradient under three prompting strategies. Our results show that current models are often poorly calibrated to evidence sufficiency. Under L5 Conflicting Evidence, all evaluated models exhibit high over-answer rates, ranging from 65% to 91%. Evidence sufficiency curves further reveal systematic failures near the boundary between sufficient and insufficient evidence. Although sufficiency-aware and chain-of-thought prompts improve abstention for some models, their effects are inconsistent across model families. Human validation on 200 samples supports the reliability of our automatic evaluation.

Our contributions are summarized as follows:

- We introduce the Evidence Sufficiency Benchmark, a controlled five-level benchmark for evaluating answer-abstention calibration in RAG-style settings.
- We propose an evaluation protocol that measures accuracy, abstention rate, over-answer rate, wrong-answer rate, and evidence sufficiency curves.

- We conduct a multi-model empirical study showing that, under our benchmark setting, evaluated LLMs tend to over-answer under insufficient and conflicting evidence, especially under L5 Conflicting Evidence.

2 Related Work

2.1 RAG and Search-Augmented Generation

Retrieval-augmented generation combines parametric language models with non-parametric retrieval to improve factual accuracy in knowledge-intensive tasks [1,2], with recent surveys offering broader overviews [3,6,7]. CRAG proposes corrective retrieval that refines retrieved documents before generation [12], Self-RAG enables models to reflect on their retrieval and generation quality [13], and Xie et al. study LLM behavior under parametric–contextual knowledge conflict [14]. Evaluation frameworks such as RAGAS and ARES measure context relevance, faithfulness, and correctness [4,5], while other work studies LLM robustness under noisy retrieval [15,16]. These studies target answer quality and retrieval refinement; they do not systematically test whether models recognize when evidence is *insufficient* and abstain accordingly—the gap our work addresses.

2.2 Faithfulness, Attribution, and Hallucination Evaluation

A growing body of work asks whether outputs are grounded in source evidence. Hallucination has been broadly surveyed [17,18]. Rashkin et al. formalize attribution [19]; Liu et al. evaluate verifiability of generative search engines [20]; Adlakha et al. jointly evaluate correctness and faithfulness in QA [21]. TruthfulQA targets imitation of misconceptions [9], and FActScore decomposes long-form claims for fine-grained verification [22]. These works check whether an answer is faithful *after* it is produced; we instead ask whether the model should have answered at all given the available evidence.

2.3 Abstention, Uncertainty, and Unanswerable QA

Abstention and selective prediction reject uncertain inputs to reduce error [23]. Wen et al. survey abstention in LLMs [24], and Madhusudhan et al. find that LLMs often fail to recognize their own knowledge boundaries [25]. SQuAD 2.0 introduces unanswerable questions with a binary answerable/unanswerable distinction [8], and several studies examine calibrated uncertainty in LLMs [26,27]. Our work differs by targeting the RAG setting with externally controlled evidence quality and a five-level gradient from sufficient to conflicting.

2.4 Context Sensitivity and Position Bias

LLMs struggle to use information in the middle of long contexts, exhibiting position bias [28]. HotpotQA evaluates multi-hop reasoning over supporting facts [29], and FEVER studies claim verification [30]. These findings motivate our L5 analysis (Section 7), where we randomize the order of gold and conflict passages to test whether models robustly detect contradictions regardless of presentation order.

2.5 What Makes ESB Different

Several lines of prior work are related but distinct. Selective QA and abstention research [23,24] asks whether models know the limits of their parametric knowledge. Confidence calibration [26,27] measures whether expressed confidence matches accuracy. Unanswerable QA [8] tests detection that a passage contains no answer. Self-RAG [13] and CRAG [12] add verification modules that filter low-quality retrieval before generation. Our Evidence Sufficiency Benchmark (ESB) differs in evaluating whether models recognize the *quality of externally provided evidence* and adjust behavior accordingly: unlike confidence calibration (which

targets internal uncertainty), ESB tests responses to *external* evidence conditions; unlike unanswerable QA (binary answerable/unanswerable), ESB covers a full gradient from sufficient to conflicting; unlike Self-RAG and CRAG (architectural solutions), ESB is a diagnostic benchmark that measures the problem itself.

3 Task Definition

We formulate evidence sufficiency evaluation as an *answer-abstention decision problem*: given a question q, an evidence context e_l at evidence level $l \in \{L1, \ldots, L5\}$, and a prompt strategy p, a model M produces $y = M(q, e_l, p)$. The goal is to evaluate whether the output behavior matches the sufficiency of the provided evidence.

3.1 Evidence Sufficiency Levels

The five evidence levels are defined in Table 1. L1 and L2 are answerable conditions; L3–L5 are insufficient-evidence conditions requiring abstention. The key *evidence sufficiency boundary* lies between L2 and L3.

Table 1: Evidence sufficiency levels. L1–L2 are answerable conditions, while L3–L5 require abstention because the provided evidence is irrelevant, missing, or conflicting.

Level	Name	Evidence Condition	Construction	Expected
L1	Full Support	Evidence fully contains the information needed to answer.	Original gold passage	Answer
L2	Partial Support	Partial or weaker support; answer still inferable.	Weaken supporting context	Answer
L3	Irrelevant Evidence	Unrelated to the question; does not support any valid answer.	Replace with unrelated passage	Abstain
L4	No Context	No external evidence provided.	Empty or no-context input	Abstain
L5	Conflicting Evidence	Contains information that contradicts the gold answer.	Pair gold with contradictory passage	Abstain

3.1.1 Why Five Levels

The taxonomy is motivated by common retrieval failure modes in real-world RAG systems: L1 ideal retrieval; L2 partial or noisy retrieval (tangentially relevant documents); L3 off-topic retrieval; L4 retrieval failure (no context); L5 contradictory retrieval (sources disagree). Together they cover the practical spectrum from fully supported to fully contradictory evidence. We do not claim this is the only possible taxonomy; rather, it provides a systematic and reproducible framework for evaluating evidence-aware abstention behavior.

3.1.2 Evidence Sufficiency Calibration

We use *calibration* in a behavioral sense rather than the probabilistic sense of confidence calibration [26]: *evidence sufficiency calibration* asks whether a model's answer-abstention decision is aligned with the quality of external evidence. A well-calibrated model answers when evidence is sufficient (L1–L2) and abstains when it is insufficient (L3–L5); the evidence sufficiency curve (abstention rate across L1–L5) visualizes this calibration. To complement the curve with a scalar metric, we define the *Behavioral Expected Calibration*

Error:

$$\text{B-ECE} = \frac{1}{N} \sum_{l \in \{L1,\ldots,L5\}} n_l \cdot |\bar{a}_l - a_l^*|, \tag{1}$$

where n_l is the number of instances at level l, $N = \sum_l n_l$, $\bar{a}_l$ is the observed abstention rate, and a_l^* is the ideal behavior (0 for L1–L2, 1 for L3–L5). A perfectly calibrated model achieves 0%; larger values indicate stronger miscalibration. Unlike standard ECE, which requires probabilistic confidence outputs, B-ECE operates on the model's decisions and applies to instruction-following LLMs that do not expose calibrated probabilities.

3.2 Behavioral Taxonomy and Metrics

Each output y is mapped to a behavioral category $c(y) \in \{\text{answer}, \text{abstain}\}$: *answer* if it provides a concrete answer, *abstain* if it explicitly states that the evidence is insufficient. The ideal behavior is $c^*(y) =$ answer for $l \in \{L1, L2\}$ and $c^*(y) =$ abstain for $l \in \{L3, L4, L5\}$. *Over-answering* is producing an answer under insufficient evidence: $\text{OverAnswer}(y, l) = \mathbb{I}[c(y) = \text{answer} \wedge l \in \{L3, L4, L5\}]$.

We report three primary metrics. The *abstention rate* is the fraction of outputs classified as abstain; the *over-answer rate* is computed only over $l \in \{L3, L4, L5\}$ instances:

$$\text{Over-answer Rate} = \frac{\sum_i \mathbb{I}[c(y_i) = \text{answer} \wedge l_i \in \{L3, L4, L5\}]}{\sum_i \mathbb{I}[l_i \in \{L3, L4, L5\}]}; \tag{2}$$

accuracy is computed only over answerable instances:

$$\text{Accuracy} = \frac{\sum_i \mathbb{I}[\text{Correct}(y_i) \wedge l_i \in \{L1, L2\}]}{\sum_i \mathbb{I}[l_i \in \{L1, L2\}]}. \tag{3}$$

Plotting these across the five evidence levels yields the *Evidence Sufficiency Curve*.

4 Benchmark Construction

We build the Evidence Sufficiency Benchmark to examine whether LLMs adjust their answering behavior to the quality of the given evidence. The pipeline has three stages: (i) construct five controlled evidence levels, (ii) query each model with three prompt strategies, and (iii) aggregate the behavioral metrics. Algorithm 1 presents the full procedure.

Algorithm 1: Evidence sufficiency benchmark construction and evaluation

Require: QA dataset $\mathcal{D} = \{(q_i, a_i, c_i)\}_{i=1}^N$, levels $\mathcal{L} = \{L_1, \ldots, L_5\}$, prompts $\mathcal{P} = \{P_1, P_2, P_3\}$, models $\mathcal{M}$
Ensure: Evaluation metrics under each evidence condition

Stage I: Evidence Construction

1: **for** each $(q_i, a_i, c_i) \in \mathcal{D}$ **do**
2: $e_i^{L_1} \leftarrow$ **Full Support**: gold context with sufficient evidence for a_i
3: $e_i^{L_2} \leftarrow$ **Partial Support**: weakened evidence; a_i still inferable
4: $e_i^{L_3} \leftarrow$ **Irrelevant Evidence**: unrelated text not containing a_i
5: $e_i^{L_4} \leftarrow$ **No Context**: empty evidence field
6: $e_i^{L_5} \leftarrow$ **Conflicting Evidence**: contradicts a_i or supports alternative
7: **end for**

(Continued)

Algorithm 1 (continued)

Stage II: Model Querying

for each model $m \in \mathcal{M}$, prompt $p \in \mathcal{P}$, level $L \in \mathcal{L}$, question q_i **do**

  Generate $r_{i,m,p,L} = m(\Pi_p(q_i, e_i^L))$

end for

Stage III: Classification and Metrics

for each response $r_{i,m,p,L}$ **do**

  Classify as CORRECT, WRONG, ABSTAIN, or INVALID

end for

Compute accuracy on L_1–L_2 (answerable)

Compute abstention rate, over-answer rate on L_3–L_5 (insufficient)

Aggregate by dataset, model, prompt, and evidence level

return Metrics and diagnostic cases

4.1 Evidence Level Construction

We construct five evidence levels moving from sufficient support to conflicting evidence (Table 1). L1–L2 are answerable; L3–L5 are insufficient-evidence cases requiring abstention. The evidence sufficiency boundary lies between L2 and L3.

Fig. 2 provides a concrete example of how the five evidence sufficiency levels are constructed from a single question.

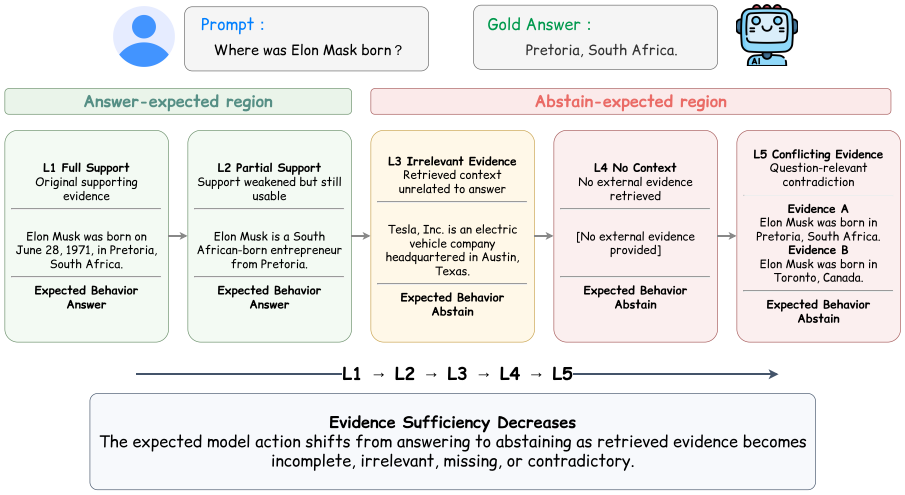

Figure 2: Construction of the five evidence sufficiency levels. The five levels correspond to common retrieval conditions encountered in real-world RAG systems, including evidence degradation, irrelevant retrieval, missing retrieval, and conflicting retrieval. Together they form a gradual transition from answer-expected to abstention-expected conditions.

4.2 Data Sources

We use two QA datasets: SQuAD for single-passage extractive QA and HotpotQA for multi-hop reasoning [8,29]. Both provide gold-annotated answer spans and supporting passages, which enables controlled evidence manipulation. We sample 100 questions per dataset and construct five evidence levels each, giving $2 \times 100 \times 5 = 1000$ evidence-level instances and, with three prompt strategies, 3000 prompt-level instances per model (approximately 29,400 model calls in total across the seven models). Although

SQuAD and HotpotQA enable precise evidence-level control, evidence-sufficiency calibration is also critical in high-stakes specialized domains; we therefore identify medical (e.g., PubMedQA [31], BioASQ [32]), legal (CUAD [33]), financial (FinQA [34]), and educational QA as priority directions for extending ESB, where retrieval failures carry greater downstream risk.

4.3 Prompt Strategies

We use three prompt strategies with increasing sufficiency guidance (Table 2): P1 Standard answers based on the provided context; P2 Sufficiency-aware adds an explicit instruction to abstain when evidence is insufficient; P3 CoT + Sufficiency requires the model to first assess sufficiency before deciding.

Table 2: Prompt strategies used in the benchmark. P1 provides no abstention guidance, P2 explicitly permits abstention when evidence is insufficient, and P3 requires the model to assess evidence sufficiency before answering.

ID	Name	Key Instruction
P1	Standard	Answer based on the provided context.
P2	Sufficiency-aware	Answer only if evidence is sufficient; otherwise abstain.
P3	CoT + Sufficiency	Assess evidence sufficiency first, then answer or abstain.

4.4 Quality Assurance

We verify that L3/L4 evidence does not contain the gold answer string, and that L5 evidence contains question-relevant information conflicting with the gold answer. The authors manually reviewed 200 stratified samples to confirm evidence-level labels and automatic response classifications.

5 Experimental Setup

5.1 Models

We evaluate seven LLMs from five families on the full L1–L5 gradient: GPT-5.5 and GPT-4o-mini (OpenAI), Claude Sonnet 4.6 (Anthropic), Gemini 2.5 Flash (Google), DeepSeek Chat (V3) and DeepSeek Reasoner (R1) (DeepSeek), and Llama-3.3-70B (Meta, open-weight). L5 is additionally evaluated with order randomization.

5.2 Evaluation Procedure and Output Classification

Each model is evaluated on $2 \times 100 \times 5 \times 3 = 3000$ prompt-level instances across L1–L5, plus L5 order-randomized evidence pairs ($2 \times 200 \times 3 = 1200$), totaling approximately 29,400 model calls across the seven models. Each response is classified as Answer (concrete answer), Abstain (explicitly states evidence is insufficient), or Invalid (empty/malformed); for Answer outputs we check gold-answer match via case-insensitive string matching, and under L5 we additionally flag known-wrong answers from the conflicting passage.

5.3 Evaluation Metrics

We report four metrics: abstention rate, accuracy, and over-answer rate (defined in Section 3), plus the *wrong-answer rate* = #Known Wrong Answers/#Total Responses which measures adoption of incorrect answers from conflicting evidence. Together, accuracy captures helpfulness under sufficient evidence, while the other three capture safety under insufficient or conflicting evidence.

5.4 Statistical Analysis

We use 2000 bootstrap resamples for confidence intervals on all primary metrics, and proportion-based significance tests for prompt/model comparisons. Effect sizes are reported as percentage-point changes.

6 Results

6.1 Evidence Sufficiency Curve

Fig. 3 and Table 3 show the evidence sufficiency curve for all seven models. Under L1–L2, abstention is low (<10%), confirming correct answering under sufficient evidence. Abstention rises under L3 (27%–67%) and L4 (34%–84%), but drops sharply under L5 to 9.5%–34.8%—the *conflict trap*: models detect evidence absence (L4) more readily than evidence conflict (L5).

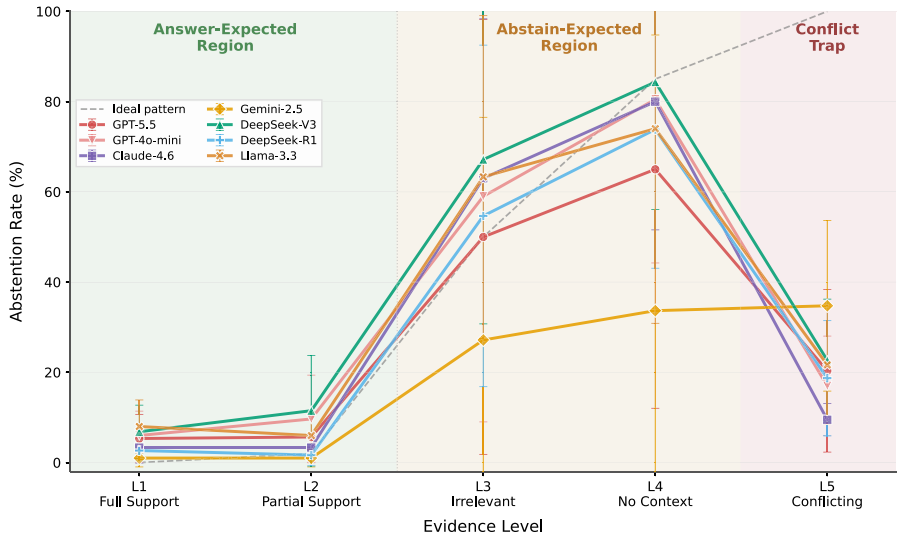

Figure 3: Evidence sufficiency curve across L1–L5 for all seven models. Models rarely abstain under L1–L2, abstain more under L3–L4, but abstention drops sharply under L5 conflicting evidence, revealing the conflict trap.

Table 3: Main results across L1–L5 for all seven models. L1–L2: accuracy (%); L3–L5: abstention rate (%); B-ECE: Behavioral expected calibration error (%, lower is better).

Model	L1 Acc.	L2 Acc.	L3 Abst.	L4 Abst.	L5 Abst.	B-ECE ↓
GPT-5.5	84.3	83.0	50.0	65.0	20.3	49.95
GPT-4o-mini	84.7	80.5	59.0	80.5	17.0	40.36
Claude Sonnet 4.6	86.0	88.7	63.0	80.0	9.5	53.21
Gemini 2.5 Flash	47.0	46.0	27.2	33.7	34.8	45.28
DeepSeek Chat	84.3	79.5	67.2	84.3	22.7	36.92
DeepSeek Reasoner	77.2	76.2	54.7	73.8	18.7	39.60
Llama-3.3-70B	84.3	85.3	63.3	73.9	21.7	48.77

6.2 Cross-family Results under Conflicting Evidence

Under L5 (Fig. 4), all seven models over-answer above 65%. Gemini 2.5 Flash attains the highest abstention (34.8%) but still over-answers 65.2%; Claude Sonnet 4.6 has the lowest abstention (9.5%, 90.5%

over-answering); GPT-5.5 has the highest wrong-answer rate (39.2%), frequently adopting the contradictory answer rather than recognizing the conflict. The conflict trap affects all model families: stronger answer generation does not imply stronger conflict recognition.

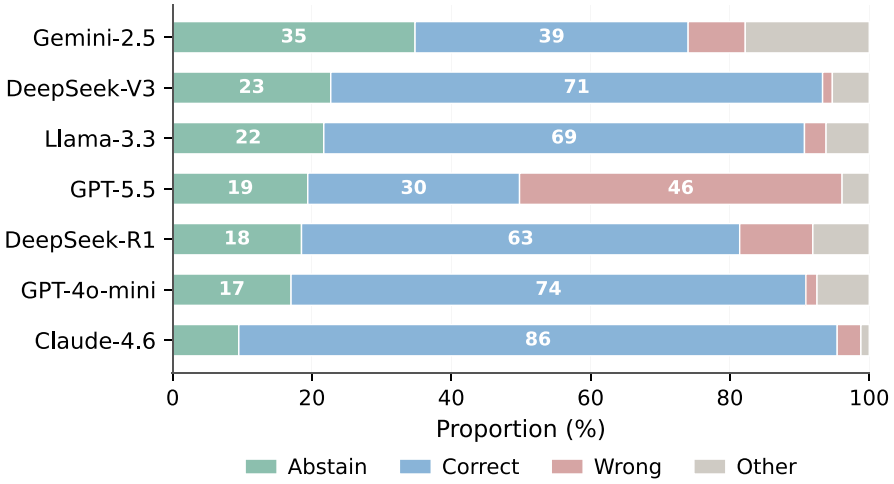

Figure 4: Response distribution under L5 conflicting evidence across seven models, showing proportions of abstain, correct, wrong, and other outputs.

To examine model behavior more closely, we categorize each L5 response into six fine-grained types (Table 4). The dominant failure mode across all models is *Evidence Picking*—selecting one of the two conflicting passages. GPT-5.5 is uniquely vulnerable to picking the wrong answer (39.2%); Claude Sonnet 4.6 shows a distinctive Conflict-Aware Answering pattern (32.9%) where it acknowledges the conflict but still answers; only Gemini exceeds 30% abstention.

Table 4: Fine-grained L5 response categories (%). CA-Abst: Conflict-aware abstention; Gen-Abst: Generic abstention; CA-Ans: Conflict-aware answering; EP-Cor/Wrong: Evidence picking (correct/wrong); Unsup: Unsupported answer.

Model	CA-Abst	Gen-Abst	CA-Ans	EP-Cor	EP-Wrong	Unsup
GPT-5.5	7.3	7.8	3.1	24.1	39.2	3.2
GPT-4o-mini	0.9	16.2	0.5	73.2	1.6	7.5
Claude Sonnet 4.6	1.2	1.3	32.9	61.7	2.0	0.8
Gemini 2.5 Flash	5.2	28.7	2.1	38.0	8.2	17.8
DeepSeek Chat	7.0	15.2	2.8	68.2	1.4	5.3
DeepSeek Reasoner	6.9	11.5	2.2	60.0	9.9	9.4
Llama-3.3-70B	2.2	19.7	1.3	67.8	2.8	6.2

6.3 Prompt Sensitivity

Prompting affects abstention but unevenly (Fig. 5). Under P1, abstention is below 8% for almost all models. From P1 to P3, Gemini improves from 5.2% to 55.2% and GPT-5.5 from 0% to 47.9%, but Claude Sonnet 4.6 only from 7.2% to 11.2%—some models remain answer-biased even with explicit sufficiency guidance.

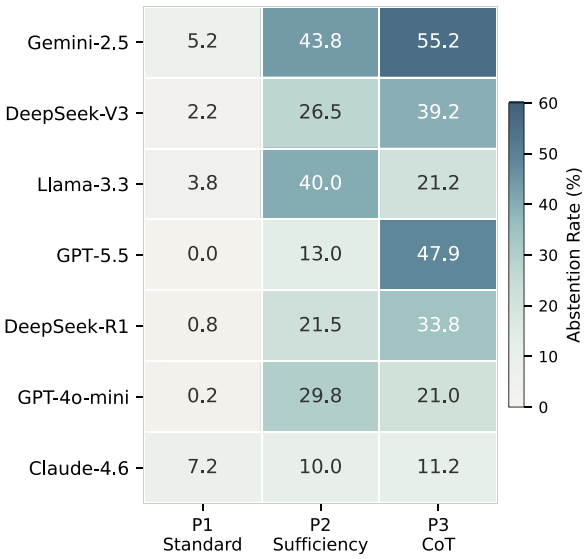

Figure 5: Prompt sensitivity heatmap under L5. Abstention rates (%) under P1, P2, and P3 for each model.

6.4 Position Bias under Conflicting Evidence

Randomizing the order of gold and conflict passages in L5 reveals significant position bias in six of seven models (Fig. 6), with the largest absolute difference 20.5%. Claude Sonnet 4.6 shows an opposite bias (−6.7%), indicating model families integrate conflicting evidence differently.

6.5 Helpfulness–Safety Trade-off

Fig. 7 plots accuracy (L1–L2) against over-answer rate (L3–L5). Claude Sonnet 4.6 attains the highest accuracy (87.4%) but a high over-answer rate (69.8%); Gemini's low accuracy (46.5%) reflects excessive abstention under some prompts; DeepSeek Chat strikes the best balance (81.9% accuracy, 50.8% over-answer). Answer-only evaluation is therefore insufficient: helpfulness and safety must be assessed jointly.

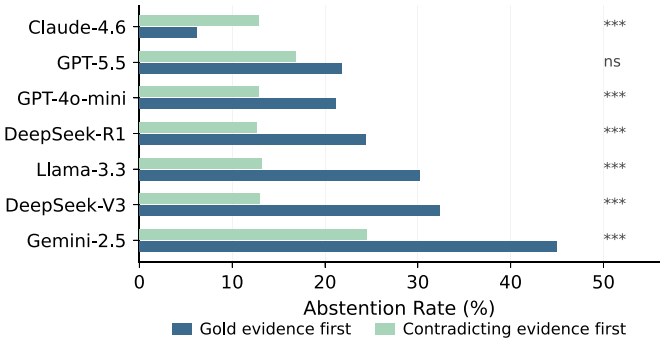

Figure 6: Position bias under L5. Significance: $^{***}p < 0.001$, ns = not significant.

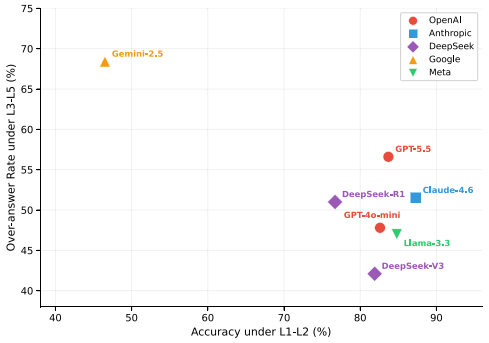

Figure 7: Helpfulness–safety trade-off under L5. Upper-right models answer correctly more often but also over-answer more.

7 Analysis

7.1 Why is Conflicting Evidence Harder than No Context?

Models abstain far more under L4 (64%–80%) than L5 (9.5%–21.7%). In L4, the absence of evidence makes insufficiency explicit; in L5 the context still contains plausible answer-bearing information, so models select one side rather than detect the contradiction. The failure is not in finding an answer but in recognizing internal inconsistency in retrieved evidence.

7.2 Model-Specific Behavior under Conflict

Claude Sonnet 4.6 abstains well under L3/L4 but only 9.5% under L5, indicating it detects absence better than conflict. GPT-5.5 shows moderate abstention (19.4%) coupled with an anomalously high wrong-answer rate (46.3%), meaning it frequently adopts the incorrect alternative. These patterns confirm that evaluating only correctness hides important safety failures.

7.3 Prompting Helps but Does Not Solve Over-Answering

From P1 to P3, Gemini improves from 5.2% to 55.2% while Claude improves only from 7.2% to 11.2%. P3 is not always better than P2—for some models, chain-of-thought reasoning focuses on answerable clues rather than detecting conflict. Prompt engineering partially mitigates over-answering but is not a stable solution.

7.4 Dataset Comparison and Position Bias

Six of seven models show significant L5 position bias (Fig. 6), confirming models do not robustly detect conflict regardless of passage order. HotpotQA (multi-hop) shows different abstention patterns than SQuAD (Fig. 8), indicating that evidence complexity interacts with sufficiency calibration.

7.5 Error Diagnostics

Fig. 9 reveals four recurring failure modes: (1) treating conflicting evidence as ordinary support; (2) relying on parametric knowledge under L3/L4 instead of the provided context; (3) behavior flipping across prompt templates; (4) false abstentions under L1/L2.

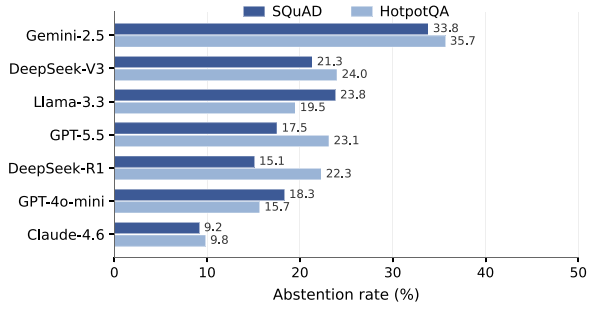

Figure 8: Dataset comparison: SQuAD (single-passage) vs. HotpotQA (multi-hop).

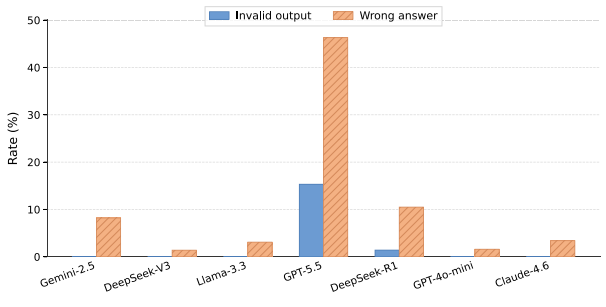

Figure 9: Error diagnostics across evidence levels and models.

7.6 Case Study: Example Outputs

Table 5 shows two representative GPT-5.5 failures. Under L4 the model ignores irrelevant context and answers from parametric memory—factually correct here, but unsafe because the provided evidence does not support any answer. Under L5 the model fails to detect the contradiction and adopts the fabricated wrong answer.

Table 5: Case study (GPT-5.5). L4: over-answering from parametric memory under irrelevant context. L5: adopting the wrong answer from conflicting evidence.

Lvl	Content
L4	**Q:** In what country is Normandy located? **Ctx:** "Boston is a town. . .in Lincolnshire. . ." (irrelevant) **Out:** "France." × over-answering from parametric memory
L5	**Q:** What country did the Normans invade in 1169? **Ctx:** gold says "Ireland"; conflict says "Chief of Protocol" **Out:** "Chief of Protocol" × adopted wrong answer

7.7 Human Validation

The authors manually reviewed 200 stratified samples (40 per level) as an internal consistency check, examining both evidence-level validity and automatic-label agreement (Table 6). All evidence items were judged valid; overall agreement was 89.0%—high under L1–L4 and reasonably strong (75.0%) under L5. The lower L5 agreement reflects intrinsic difficulty: models often acknowledge the conflict yet still select one side, motivating the fine-grained L5 categories in Section 6.

Table 6: Human validation on 200 stratified samples (40 per level). All evidence items were judged valid; overall automatic–manual agreement was 89.0%, with L5 the most challenging due to mixed conflict-aware responses.

Level	Valid%	Agree%
L1 (Full Support)	100.0	100.0
L2 (Partial Support)	100.0	100.0
L3 (Irrelevant)	100.0	80.0
L4 (No Context)	100.0	90.0
L5 (Conflict)	100.0	75.0
Overall	100.0	89.0

7.8 Why Models Differ in Abstention Behavior

Abstention rates vary substantially across models [11] (e.g., Gemini 34.8% vs. Claude 9.5% under L5). Plausible contributing factors include *safety alignment* (stronger safety-oriented RLHF rewards caution under uncertainty), *instruction-following fidelity* (higher adherence yields stronger responses to abstention instructions), *reasoning depth* (reasoning models may detect contradictions yet still resolve to an answer), and *training-data composition* (heavier QA-style training fosters an answer-producing prior). These factors interact, and disentangling them is beyond the scope of this benchmark study; the evidence sufficiency curve nonetheless serves as a useful diagnostic for comparing models along this dimension.

7.9 Practical Implications for RAG Systems

Our findings have direct implications for RAG design. The evidence sufficiency curve serves as a diagnostic for retrieval failures: a flat curve across L1–L5 signals an LLM that cannot distinguish evidence quality. Abstention thresholds in production should incorporate the LLM's own evidence-sensitivity profile rather than rely solely on retrieval confidence. Verification frameworks such as Self-RAG [13] and CRAG [12] can themselves be evaluated on ESB to test whether they actually prevent over-answering, and for high-stakes deployments (medical, legal, financial) ESB identifies which models are most prone to confidently producing wrong answers under conflicting evidence.

Why current verification architectures do not eliminate the L5 conflict trap. Self-RAG emits reflection tokens for per-passage relevance and supportedness; CRAG uses an external evaluator that triggers corrective retrieval when documents look unreliable. Both leave the L5 trap largely unaddressed because (i) both rely on the base LLM (or a same-family scorer) to judge passage quality—so if the base model fails to detect contradictions (as our results show across all seven), the verification module inherits the blind spot; (ii) neither explicitly models *pairwise contradictions* among multiple "relevant" passages—Self-RAG's tokens are per-passage, and CRAG scores retrieval quality holistically.

A contradiction-aware retrieval verifier. A natural next step is to insert a dedicated contradiction-detection module between retrieval and generation: for each pair of top-k retrieved passages, run a lightweight NLI- or LLM-judge-based pairwise check; if cross-passage contradictions are detected above a threshold, route the query to an abstention or clarification branch rather than standard generation. ESB provides a controlled testbed for such a verifier because L5 instances directly probe the failure mode it targets. We leave full implementation and evaluation to future work.

8 Limitations

Benchmark scale and domain coverage. ESB is built from 200 seed QA pairs (100 SQuAD + 100 HotpotQA). Five-level evidence manipulation and three prompts yield 3000 instances per model, but the underlying question diversity is limited and the seed datasets are general open-domain QA. Because evidence-sufficiency calibration is especially consequential in high-stakes specialized domains, we plan to extend ESB to medical (PubMedQA [31], BioASQ [32], MedQA), legal (CUAD [33]), financial (FinQA [34]), and education-oriented QA, as well as multilingual settings.

Rule-based abstention detection. Our classifier matches explicit refusal phrases for reproducibility; this may miss implicit refusals or mixed responses, most notably under L5 (75% automatic–manual agreement). Future work should integrate LLM-as-judge evaluation and richer response taxonomies.

Annotation methodology and inference-time scope. Human validation was an internal consistency check by the authors rather than a formal study with multiple external annotators; future work should recruit independent annotators and report inter-rater reliability (e.g., Cohen's or Fleiss' Kappa). We also evaluate models only at inference time without fine-tuning or specialized modules; supervised abstention training, preference optimization, and contradiction-aware retrieval verifiers (Section 7) are complementary mitigations to investigate.

9 Conclusion

We introduced the Evidence Sufficiency Benchmark (ESB), a controlled benchmark evaluating whether LLMs calibrate answer-abstention behavior to the quality of external evidence in RAG—asking not only whether the final answer is correct but whether the model should answer at all. Across five evidence levels from L1 Full Support to L5 Conflicting Evidence, current LLMs are often poorly calibrated: models perform

well under L1/L2 but continue answering under irrelevant, missing, or conflicting evidence; the most severe failure occurs under L5, where all seven models over-answer above 65%. Prompting partially improves abstention but its effect is highly model-dependent, so high answer accuracy does not imply reliable evidence use or safe RAG behavior.

These findings indicate that RAG systems need explicit evidence-quality checks and dedicated conflict-detection mechanisms. Future work includes extending ESB to multilingual and domain-specific datasets (medical, legal, financial, educational), incorporating LLM-as-judge response classifiers, and developing contradiction-aware retrieval verifiers that improve evidence sufficiency calibration in deployment.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Guangdong Provincial Undergraduate Innovation and Entrepreneurship Training Program (Project No. S202610559029) at Jinan University, a nationwide initiative administered by the Ministry of Education. We gratefully acknowledge the funding and project supervision provided through Jinan University during the full course of this research.

Author Contributions: The authors confirm their contributions to the paper as follows: conceptualization, Hantian Zhang and Wentai Wu; methodology, Hantian Zhang and Wentai Wu; software, Hantian Zhang; validation, Hantian Zhang; formal analysis, Hantian Zhang; investigation, Hantian Zhang; data curation, Hantian Zhang; writing—original draft preparation, Hantian Zhang; writing—review and editing, Hantian Zhang and Wentai Wu; visualization, Hantian Zhang; supervision, Wentai Wu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The benchmark data and evaluation code will be made publicly available upon acceptance.

Ethics Approval: This study did not involve human participants, human-subject experiments, or animal experiments. The manual validation was conducted by the authors as an internal check. No external participants were recruited, and no participant data were collected.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2020. p. 9459–74.
2. Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. In: III HD, Singh A, editors. Proceedings of the 37th International Conference on Machine Learning. Vol. 119 of Proceedings of Machine Learning Research. Cambridge, MA, USA: PMLR; 2020. p. 3929–38.
3. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. arXiv:2312.10997. 2024.
4. Es S, James J, Espinosa Anke L, Schockaert S. RAGAs: automated evaluation of retrieval augmented generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. Stroudsburg, PA, USA: ACL; 2024. p. 150–8.
5. Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: an automated evaluation framework for retrieval-augmented generation systems. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL; 2024. p. 338–54.
6. Wang W, Bao J, Zhou W, Ren N, Rao M, Qi P. A survey on RAG meeting LLMs: towards retrieval-augmented large language models. arXiv:2405.06211. 2024.
7. Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, et al. Retrieval-augmented generation for AI-generated content: a survey. arXiv:2402.19473. 2024.

8. Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2018. p. 784–9.
9. Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2022. p. 3214–52.
10. Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y, et al. Language models as knowledge bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2019. p. 2463–73.
11. OpenAI. GPT-4 technical report. arXiv:2303.08774. 2024.
12. Yan SQ, Gu JC, Zhu Y, Ling ZH. Corrective retrieval augmented generation. arXiv:2401.15884. 2024.
13. Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: learning to retrieve, generate, and critique through self-reflection. arXiv:2310.11511. 2023.
14. Xie J, Zhang K, Chen J, Lou R, Su Y. Adaptive chameleon or stubborn sloth: revealing the behavior of large language models in knowledge conflicts. arXiv:2305.13300. 2024.
15. Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. Proc AAAI Conf Artif Intell. 2024;38(16):17754–62. doi:10.1609/aaai.v38i16.29728.
16. Ren R, Wang Y, Qu Y, Zhao WX, Liu J, Tian H, et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation. arXiv:2307.11019. 2023.
17. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38. doi:10.1145/3571730.
18. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv:2311.05232. 2023.
19. Rashkin H, Nikolaev V, Lamm M, Aroyo L, Collins M, Das D, et al. Measuring attribution in natural language generation models. Comput Linguist. 2023;49(4):777–840. doi:10.1162/coli_a_00486.
20. Liu NF, Zhang T, Liang P. Evaluating verifiability in generative search engines. In: Findings of the association for computational linguistics: EMNLP 2023. Stroudsburg, PA, USA: ACL; 2023. p. 7001–25.
21. Adlakha V, BehnamGhader P, Lu XH, Meade N, Reddy S. Evaluating correctness and faithfulness of instruction-following models for question answering. Trans Assoc Comput Linguist. 2024;12:681–99. doi:10.1162/tacl_a_00667.
22. Min S, Krishna K, Lyu X, Lewis M, Wt Y, Koh P, et al. 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. Stroudsburg, PA, USA: ACL; 2023. p. 12076–100.
23. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 4878–87.
24. Wen B, Yao J, Feng S, Xu C, Tsvetkov Y, Howe B, et al. Know your limits: a survey of abstention in large language models. Trans Assoc Comput Linguist. 2025;13:529–56.
25. Madhusudhan N, Madhusudhan ST, Yadav V, Hashemi M. Do LLMs know when to NOT answer? Investigating abstention abilities of large language models. arXiv:2407.16221. 2024.
26. Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, et al. Language models mostly know what they know. arXiv:2207.05221. 2022.
27. Lin S, Hilton J, Evans O. Teaching models to express their uncertainty in words. arXiv:2205.14334. 2022.
28. Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, et al. Lost in the middle: how language models use long contexts. Trans Assoc Comput Linguist. 2024;12:157–73.
29. Yang Z, Qi P, Zhang S, Bengio Y, Cohen W, Salakhutdinov R, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2018. p. 2369–80.
30. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a large-scale dataset for fact extraction and VERification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2018. p. 809–19.

31. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research 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). Stroudsburg, PA, USA: ACL; 2019. p. 2567–77.
32. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform. 2015;16(1):1–28. doi:10.1186/s12859-015-0564-6.
33. Hendrycks D, Burns C, Chen A, Ball S. CUAD: an expert-annotated NLP dataset for legal contract review. arXiv:2103.06268. 2021.
34. Chen Z, Chen W, Smiley C, Shah S, Borova I, Langdon D, et al. FinQA: a dataset of numerical reasoning over financial data. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: ACL; 2021. p. 3697–711.