Benchmarking LLM's Capability in Reasoning over Conflicting Web References

Yizhen Yuan[1], Rui Kong[2], Dongze Li[3], Yuanchun Li[1,†], Yunxin Liu[1],
[1]Institute for AI Industry Research (AIR), Tsinghua University
[2]Shanghai Jiao Tong University [3]Independent

yuanyz21@mails.tsinghua.edu.cn, kongrui@sjtu.edu.cn, c1r9n0ai@gmail.com,
{liyuanchun, liuyunxin}@air.tsinghua.edu.cn

Abstract

Large language models (LLMs) integrated with retrieval-augmented generation (RAG) have become a dominant framework for building intelligent assistants. In real-world applications such as ChatGPT with web search, the retrieved document often comes from diverse, potentially unreliable sources and may contain inconsistent claims. Unlike traditional search engines that rely on users to manually compare information, LLM-based systems typically feed *all* retrieved content into the model's context, requiring LLMs to autonomously identify, differentiate, and reason over conflicting viewpoints. Unlike mainstream LLM evaluation tasks like math and code generation that are primarily focused on reasoning with factual context, question-answering with multi-source references requires fundamentally different capabilities to identify and reason over knowledge contradictions. In this paper, we introduce CONFRAG, a benchmark for evaluating LLMs' reasoning capability over real-world conflicting documents retrieved from the web. It consists of 1,814 real-world questions, each paired with an average of 9.58 retrieved paragraphs from heterogeneous online sources. A total of 57.2% of the questions exhibit explicit contradictions. We further propose three structured evaluation tasks, *answer clustering*, *answer coverage*, and *reason coverage*, to quantify a model's ability to organize and explain contradictory content. Experiments with state-of-the-art models such as GPT-4.1 and Claude-3-7-Sonnet reveal substantial performance gaps, highlighting the need for more targeted research in contradiction-aware question answering. To the best of our knowledge, CONFRAG is the first benchmark specifically designed to evaluate contradiction-aware reasoning on real-world long web documents.

1 Introduction

Large language models (LLMs) have achieved remarkable progress in recent years. One especially promising application is natural language interaction enhanced with retrieval-augmented generation (RAG) (Fan et al., 2024; Gao et al., 2023), where documents retrieved from the external databases are used to support LLM reasoning and factual grounding. This functionality is offered by most chatbot services of popular model providers, such as ChatGPT (Achiam et al., 2023), Claude (Cla, 2024), Grok (Thompson, 2025), etc., with the promise to become the next major entry of information retrieval for end-users.

However, this new paradigm also brings new concerns of answer biases and inaccuracies. In many real-world deployment scenarios of RAG-based LLM systems, the sources of retrieved documents may be diverse and lack control. These may include forums, news sites, academic pages, and personal blogs that may contain contradictory information. Some content may even be purposely optimized by third parties to increase visibility or strengthen certain viewpoints (Aggarwal et al., 2024). Unlike traditional search engines where users manually compare sources, LLM-based chat systems typically absorb all retrieved content into the context window and autonomously synthesize a response. This mode of interaction imposes fundamentally new demands on the underlying models: instead of merely improving the accuracy of question-answering under clean-context settings, the models must deal with the potential inconsistencies and conflicts in the retrieved documents. This motivates the need for benchmarks that explicitly evaluate contradiction reasoning.

We believe that reasoning over conflicting information is an under-explored while important capability for LLMs. A model may give the correct answer while ignoring meaningful disagree-

†Corresponding author.

ments across references, which is something human users rarely tolerate. Moreover, this setting challenges the architecture of transformer-based LLMs: since all tokens are treated equally in attention layers, there is no built-in mechanism for source differentiation.

Despite its practical significance, the benchmarks explicitly designed to evaluate this capability are limited. Most existing benchmarks for LLMs and RAG systems assume a benign and consistent context (Xu et al., 2024; Lin et al., 2022). Although a few QA datasets have been proposed for contradiction retrieval, they are generally limited in scale, realism, and source diversity. For instance, WikiContradict (Hou et al., 2024) contains only 253 examples, each consisting of exactly two viewpoints, with each viewpoint represented by a single sentence. BoardGameQA (Kazemi et al., 2023) includes a larger number of examples, but all samples are synthetically constructed based on predefined rules, rather than collected from real-world sources, allowing arbitrary instance generation but sacrificing realism. Thus, existing resources either lack scale or realism.

To address this gap, we introduce the CONFRAG[1]. We begin by collecting 1,814 open-domain questions from diverse sources, including Natural Questions (Kwiatkowski et al., 2019), ELI5 (Fan et al., 2019), Yahoo Answers (Zhang et al., 2015), and manually curated examples. For each question, we retrieve an average of 9.58 real-world web documents using keyword-based search. After filtering and cleaning, we use LLMs to extract implied answers and rationales from each document. These are then clustered into coherent viewpoints via human-in-the-loop annotation. Rather than assuming a single gold answer, each question is annotated with multiple answer clusters, each supported by a subset of references and accompanied by structured rationales. Among these, 57.2% of samples feature strong contradictions between answers, reflecting the diversity of perspectives found in real-world discourse.

In addition to the dataset, we propose three structured evaluation tasks *answer clustering*, *answer coverage*, and *reason coverage* to assess a model's ability to organize and explain disagreement. The first task evaluates whether the model correctly partitions the input documents into semantically distinct answer clusters. The second measures whether all gold answer types are recovered, and the third examines whether supporting reasons are faithfully captured within each cluster. All three tasks are scored on a $[0, 1]$ scale. For answer clustering, we use Normalized Mutual Information (NMI) (Estévez et al., 2009), where 1 indicates a perfect match between predicted and gold clusters and 0 reflects no meaningful agreement. For answer and reason coverage, we use bipartite maximum matching over keyword-based substring similarity, where 1 means all gold perspectives and their supporting reasoning are faithfully recovered.

We benchmark a diverse set of LLMs, including GPT-4.1, GPT-4.1-mini (Achiam et al., 2023), Claude-3-7-Sonnet (Cla, 2024), Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-72B-Instruct, and Qwen2.5-7B-Instruct (Yang et al., 2024). The results have shown a limited performance of all tested LLMs on this task. Claude has the highest structural compliance, while GPT-4.1 achieves top performance with an NMI score of 0.47, answer coverage of 0.45, and reason coverage of 0.15. The performance gap between 7B and 72B Qwen models confirms that larger models do better at capturing multi-viewpoint semantics.

Other models, including Claude and Qwen variants, score significantly lower on answer coverage and reason coverage, with answer coverage value under 0.37 and reason coverage typically below 0.14. Many smaller models also struggle to produce structurally valid outputs. These results highlight a pressing need for models that are not only factual, but also contradiction-aware and source-sensitive.

To summarize, our main contributions are:

- We introduce CONFRAG, a realistic benchmark for contradiction-aware reasoning over retrieved web content, featuring paragraph-level inputs, multi-answer annotation, and structured rationales.
- We define three complementary evaluation tasks including answer clustering, answer coverage, and reason coverage, which jointly assess a model's ability to separate, recover, and justify divergent perspectives.
- Beyond the dataset itself, we present a generalizable pipeline for contradiction benchmark

[1]The code and dataset are available at https://github.com/XaiverYuan/ConfRAG.

construction: given any user-defined question, our method retrieves real-world web content, identifies conflicting viewpoints, and structures them into answer clusters and rationales, enabling future expansion or domain-specific evaluation.

2 Related Work

Analyzing Contradictions in Input. Existing studies on contradictions in input primarily focus on pairs of short texts. For instance, natural language inference (NLI) tasks classify the relation between a premise and a hypothesis as entailment, contradiction, or neutral. Benchmarks such as MultiNLI (Williams et al., 2018), SNLI (Camburu et al., 2018), and ANLI (Nie et al., 2020) are constructed on short, sentence-level pairs, which restricts their applicability to concise and pre-aligned target sentences. WikiContradict (Hou et al., 2024) extends this idea to factual inconsistencies mined from Wikipedia revisions but still contains only 253 examples, each consisting of two contradictory sentences. BoardGameQA (Kazemi et al., 2023) introduces contradictions by defining artificial rules and generating synthetic reasoning tasks. However, its content is not derived from real web data, and therefore cannot evaluate model behavior under realistic, naturally occurring conflicts.

All the benchmarks mentioned above are limited to binary contradictions between two viewpoints and involve short texts. They also predefine the exact scope of the conflicting content, which is unrealistic in open-domain web retrieval. In contrast, most real-world web documents are long, noisy, and diverse. We argue that language models should be evaluated on their ability to identify and reason about contradictions within such lengthy, complex, and multi-source information.

Evaluating Contradictions in Output. Beyond input contradictions, several studies analyze contradictions that emerge in model outputs, often as manifestations of hallucination or bias (Manakul et al., 2023; Niu et al., 2024). TruthfulQA (Lin et al., 2022) examines whether models reproduce human misconceptions or socially biased beliefs. FalseQA (Hu et al., 2023) investigates models ability to refute false presuppositions in questions, while CREPE (Yu et al., 2023) identifies and rewrites questions containing erroneous assumptions. These datasets primarily focus on detecting, rejecting, or correcting false or inconsistent answers to ensure factual coherence. However, when a user's question itself is reasonable but associated with inherently conflicting viewpoints, such cases are not covered by existing benchmarks, leaving models untested on how to properly handle genuine contradictions in input sources.

Table 1: Comparison of contradiction-related QA benchmarks.

Dataset	Contradict Labels	Long Doc	With Reason
MultiNLI (Williams et al., 2018)	✓	✗	✗
SNLI (Camburu et al., 2018)	✓	✗	✓
ANLI (Nie et al., 2020)	✓	✗	✗
WikiCon (Hou et al., 2024)	✓	✗	✗
TruthfulQA (Lin et al., 2022)	✗	✗	✗
NQ (Kwiatkowski et al., 2019)	✗	✓	✗
ELI5 (Fan et al., 2019)	✗	✗	✗
FalseQA (Hu et al., 2023)	✗	✗	✓
CREPE (Yu et al., 2023)	✗	✓	✗
BoardGame (Kazemi et al., 2023)	✓	✗	✓
Ours	✓	✓	✓

We summarize the differences between our benchmark and prior datasets in Table 1. To our knowledge, our work is the only benchmark that systematically examines contradictions among more than two clusters of viewpoints. Further analysis of multi-cluster reasoning and error patterns can be found in Section 4.4.

3 Dataset and Task Design

3.1 Task Definition

We formulate a structured benchmark task aimed at evaluating whether language models can identify, organize, and explain multiple conflicting answers retrieved from the web. Each sample consists of a single natural language question and a set of associated web documents (9.58 websites per question on average), each represented as a paragraph of AI-friendly (retrieved by jina.ai (Sturua et al., 2024; Wang et al., 2025)) text each with a unique index. The documents are retrieved via open-domain web search and reflect realistic online content including Wikipedia articles, forums, blogs, and news outlets. An example of input, output, and corresponding answer is shown in Figure 1.

Here, each item in the list (as shown in blue blocks in Figure 1) represents a distinct *answer cluster*, summarizing one coherent viewpoint derived from a subset of the documents. The field

`index` denotes which documents support this cluster, and `reason` provides justifications for the answer cluster. Both `answer` and `reason` fields are evaluated via keyword-based substring matching.

Evaluation. We define three evaluation tasks:

- **Answer Clustering**: Whether the model groups documents into semantically coherent clusters matching the ground truth. Unlike other benchmarks with one gold answer, we emphasize the ability to identify which documents share the same viewpoint. In terms of model output, it is represented as the index clustering.
- **Answer Coverage**: Whether the model recovers all major gold clusters. After the model has demonstrated its ability to understand which answer is aligned with which answer, it is also important to check whether the model understands what each group of answers tries to express.
- **Reason Coverage**: Whether the model includes correct reasoning aspects associated with each answer. Beyond understanding the main point of each cluster, we further require the model to attend on each separate document too instead of only focusing on the clustered viewpoint, bringing the model's attention back to each document with respect to each clustered viewpoint.

3.2 Dataset Construction Pipeline

To construct our benchmark, we design an automated pipeline that begins with a pool of controversial questions and produces structured examples with web-based evidence, clustered answers, and supporting rationales as shown in Figure 2. The pipeline is composed of five major components: (1) question collection, (2) web retrieval, (3) paragraph selection, (4) implied answer and reason extraction, and (5) answer clustering. All prompts and LLM interactions used in the following steps are executed with GPT-4o.

Controversial Question Selection. We begin by generating our own set of questions across six topical domains: health, trivia, lifestyle habits, disease prevention, science and technology, and nutrition/food. This generation process was guided by prompts listed in Appendix A.1. Each generated question was manually reviewed to ensure it was non-offensive and appropriate for public release. To further enhance topic diversity, we additionally incorporate questions from three publicly available datasets: Natural Questions (Kwiatkowski et al., 2019), ELI5 (Fan et al., 2019), and Yahoo! Questions (Zhang et al., 2015). We then apply a filtering process to all four sources to identify controversial questions that may elicit divergent opinions among internet users. The filtering prompts used for this step are provided in Appendix A.3.

Web Retrieval and Paragraph Filtering. For each question, we first extract a set of keywords for better search performance. Then we perform web retrieval via Google Search using SerpAPI, based on keywords extracted from the question. We search up to 40 websites for each question, sorted from the most relevant to the least relevant website. For each retrieved page, we employ Jina's document reader pipeline (Sturua et al., 2024; Wang et al., 2025) to extract an AI-friendly plain-text representation of the page content. Due to problems in crawling or other issues, some websites may fail to be crawled. When 10 valid webpages have been collected, we stop processing. On average, each question is associated with 9.58 documents, with each document containing approximately 2,025 words.

Implied Answer and Reason Extraction. From each retained document, we use the prompt in Appendix A.4 to extract an *implied answer* and a list of short *reasons* (as strings).

Answer Clustering. We cluster similar answers into coherent viewpoints using LLM-based similarity judgment . The prompt is shown in Appendix A.5. Each cluster is assigned an answer string, a set of supporting document indices, and a list of key reasons. We define a question as *contradictory* if its clusters contain mutually exclusive claims supported by different sources. Among our dataset, 1,037 (57.2% of total) questions are labeled as contradictory. An example of clustered answer can be found in Figure 1 and Supplementary materials. Documents that appeared to support multiple conflicting clusters simultaneously were excluded during this stage to ensure each retained document has a clear, singular viewpoint.

3.3 Human-in-the-loop Annotation Protocol

To ensure the quality of automatically generated clusters, we established a structured human-in-the-loop protocol. Each example was independently reviewed by two annotators, who received

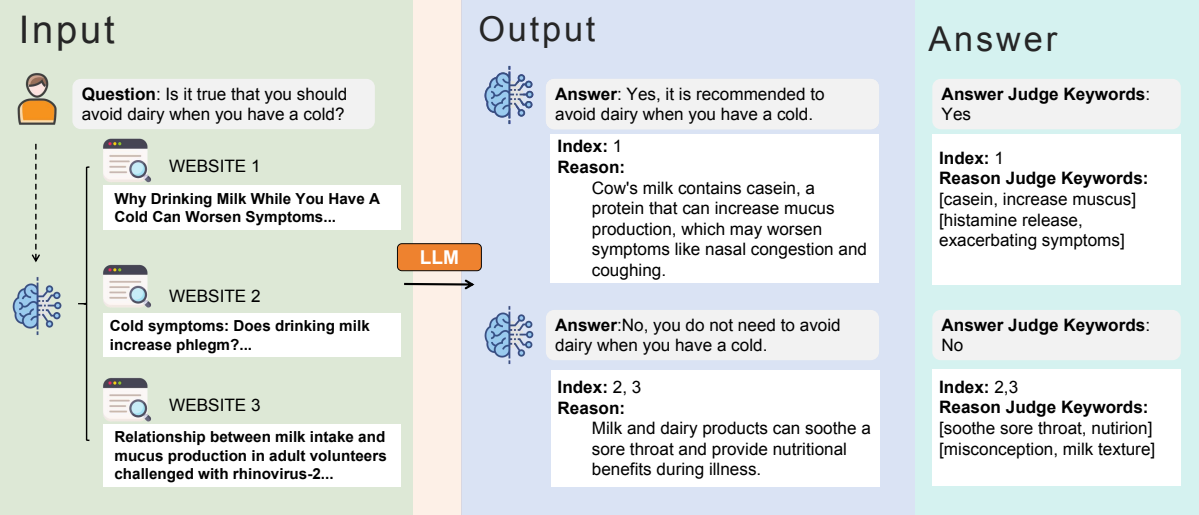

Figure 1: An example of model input, expected output and the correct answer

the question, the set of LLM-extracted answers, reasons, and their preliminary cluster assignments. Annotators were asked to (1) verify whether documents within each cluster expressed a consistent viewpoint, (2) check the correctness and completeness of the summarized answers and supporting reasons, and (3) mark whether contradictions existed between clusters. When the two annotators disagreed, a third adjudicator reviewed their judgments and facilitated a joint discussion to reach consensus, occasionally revising cluster boundaries or rationales. Five trained annotators with at least undergraduate-level education participated, following a written manual defining cluster formation, contradiction criteria, and acceptable reasons. Across 1,814 examples, we achieved an inter-annotator agreement (IAA) (Artstein and Poesio, 2008) of 89.2% and a Cohen's κ (Cohen, 1960) of 0.642, indicating moderate consistency.

3.4 Evaluation Protocol

Each example in our benchmark consists of a question and a set of retrieved indexed documents. Given this input, a model is expected to output a structured response in JSON format containing a set of answer clusters as shown in Figure 1. All evaluations are fully automated and use a standardized JSON interface.

Task 1: Answer Clustering. This task measures how well the model groups supporting documents into semantically coherent clusters. We compare the predicted document-to-cluster assignments with gold annotations using Normalized Mutual Information (NMI) (Estévez et al., 2009), which is a widely used metric for clustering quality. The formulas are listed in Appendix D.1. For example, the instance in Figure 1 results in a grade of 1 (two sets were identical). For any valid partition, (every index shown once and only once in the answer), they will get a non-zero score. A score of 0 occurs when the partition is invalid, we further discuss this problem in Section 4.4.

Task 2: Answer Coverage. Each gold answer is associated with a list of `answer judge keywords`. We use substring matching to match the answer. A gold answer is considered matched if any of its judge keywords appears as a substring in the predicted answer. These judge keywords are generated by GPT-4o and verified by human annotators to ensure accuracy and coverage. Substring matching is chosen over semantic similarity to keep evaluation cost-effective and fully reproducible without additional model calls; as a trade-off, minor lexical variations (e.g., abbreviations or paraphrases) may occasionally cause a valid answer to go unmatched, leading to a slight underestimation of true coverage. An answer can be matched to at most one list of answer judge keywords while a list of answer judge keyword also can only be matched to at most one answer. The score is computed by formula in Appendix D.2. For example, the instance in Figure 1 results in a grade of 1 (One answer contains 'Yes' while another answer contains 'No')

Task 3: Reason Coverage. This task uses the same evaluation principle as Task 2. Each reason is associated with an answer. If an answer is matched, then we can start to match its reasons.

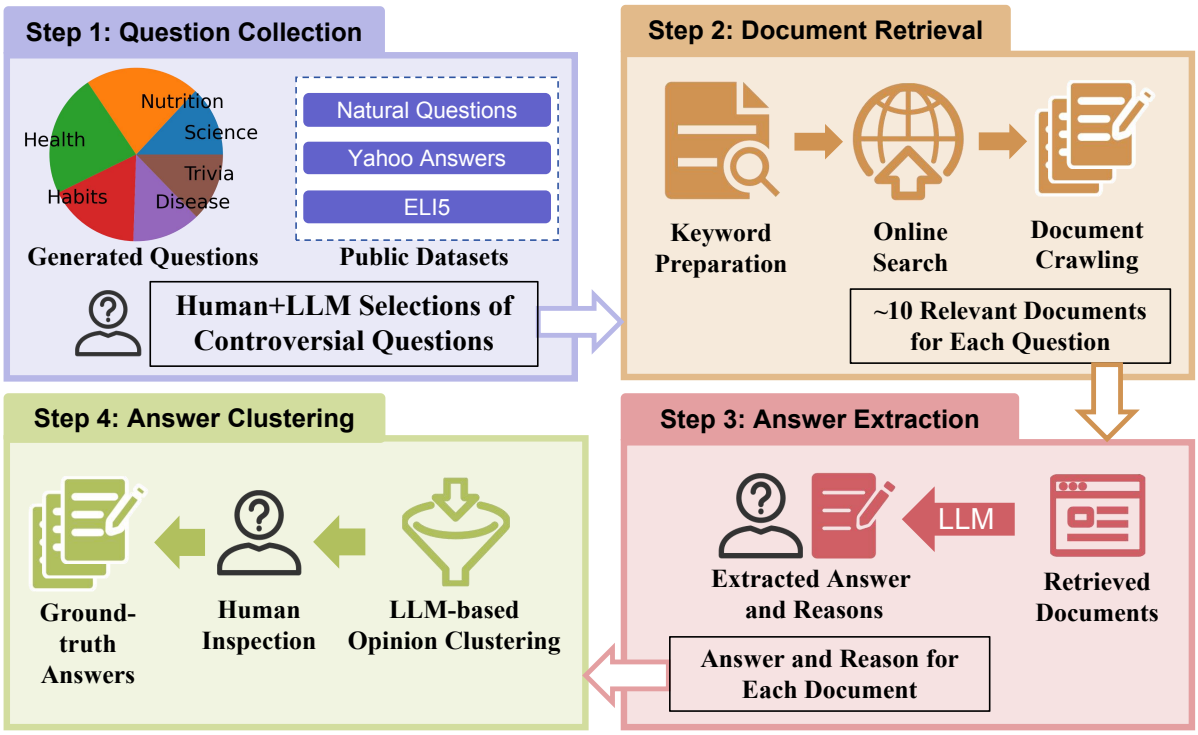

Figure 2: The Workflow of Dataset Construction

Table 2: Results on three benchmark tasks. All models use the same prompt and JSON format; values are mean(std) on valid outputs.

Model	Answer Clustering	Answer Coverage	Reason Coverage	Valid Partition Rate
GPT-4.1-mini	0.370 (0.383)	0.452 (0.299)	0.151 (0.161)	0.648
GPT-4.1	0.466 (0.406)	0.448 (0.315)	0.157 (0.181)	0.763
Qwen2.5-7B-Instruct	0.147 (0.283)	0.312 (0.290)	0.083 (0.134)	0.308
Qwen2.5-72B-Instruct	0.373 (0.376)	0.368 (0.308)	0.133 (0.169)	0.706
Claude-3-7-Sonnet	0.459 (0.356)	0.362 (0.296)	0.114 (0.142)	0.849
Llama-3.1-8B-Instruct	0.161 (0.260)	0.276 (0.271)	0.071 (0.120)	0.428
Llama-3.1-70B-Instruct	0.178 (0.286)	0.331 (0.321)	0.105 (0.160)	0.489
HumanEval	0.691 (0.301)	0.510 (0.357)	0.275 (0.164)	1.000

Each gold reason has an associated `reason judge keyword` list. We also use substring matching to match the reason to the reason judge keyword. A reason can be matched to at most one list of reason judge keywords within the matched answer's reason judge keywords list, while one list of reason judge keywords also can only be matched to at most one reason. The score is computed by formula in Appendix D.3. For example, the instance in Figure 1 results in a grade of 0.5 (reasons in each output cluster are correct, but did not cover all reasons in the answer).

This three-part evaluation captures complementary aspects of reasoning under contradiction: the ability to group documents, understanding the main point of each cluster, and explain them concisely. Together, these metrics quantify a model's capacity to structure conflicting web-based content for end-users.

4 Experiment

4.1 Experimental Setup

To evaluate the difficulty of our benchmark and establish baseline performance, we conduct experiments using a diverse set of open-source and proprietary language models. The models evaluated include: GPT-4.1-mini, GPT-4.1, Qwen2.5-7B-Instruct, Qwen2.5-72B-Instruct, Claude-3-7-Sonnet, Llama-3.1-8B-Instruct, Llama-3.1-70B-

Instruct.

For each model, we randomly sample five documents from the approximately ten retrieved web documents available per question (We assume that the documents have already been retrieved, with further discussion of the retriever component provided in the Limitations section). These documents are paired with their corresponding integer indices and passed into the model along with the target number of answer clusters (i.e., the known ground-truth cluster count for the selected subset of webpages). This setting is used to reduce computational costs while maintaining evaluation fidelity. The same random sampling protocol applies to the human evaluation subset, ensuring methodological consistency between human and LLM evaluations.

We use a standardized prompt template that clearly defines the task, constraints, and required output format. The model is instructed to: (1) Partition the provided set of documents into exactly k non-overlapping clusters (2) For each cluster, produce a distinct core answer, a list of supporting indices, and a list of concise reasons grounded in the corresponding documents (3) Output a valid JSON object following a strict schema

To enforce output validity, the prompt includes format examples, attention warnings, and postconditions (e.g., index sets must be disjoint and complete). An example prompt is provided in Appendix F.

All experiments are conducted in a zero-shot setting without finetuning. No ground-truth answers, reasoning annotations, or keyword metadata are exposed to the model. Models are evaluated using the automatic metrics described in Section 3.4.

SS

Violation of Output Constraints. Despite explicit instructions in the prompt, many models fail to satisfy key structural constraints. A common violation is that the same document index appears in multiple clusters, even though the prompt explicitly enforces that clusters must be disjoint and collectively exhaustive. Other failure cases include missing indices, malformed JSON outputs, or redundant answer strings across clusters.

These violations highlight a broader challenge in using LLMs for structured reasoning tasks: even with clear schema and format instructions, models may ignore constraints unless fine-tuned or externally validated[2]. Our benchmark provides a setting to systematically test such behaviors.

4.2 Main Results

Across all models, we observe that Answer Clustering scores (NMI) range from 0.14 to 0.47, indicating that even strong language models struggle to consistently group conflicting evidence. The best-performing model (GPT-4.1) reaches an NMI of 0.466, while Claude is close behind with 0.459, which are both significantly lower than human-level (0.69).

Answer Coverage scores are also modest, with only two models exceeding 0.40. This suggests that models often fail to identify all distinct plausible answers, frequently collapsing multiple perspectives into a single dominant cluster. Importantly, Answer Coverage is computed independently of clustering correctness: it relies solely on substring matching between predicted and gold answers. However, poor clustering can indirectly affect this metric. When two gold clusters are incorrectly merged, one may dominate the output while the other is ignored or replaced with unrelated content.

Reason Coverage scores are uniformly low (all below 0.16), reflecting the compounded difficulty of generating fine-grained explanations after successful answer matching. Since reason evaluation is gated by matched answer clusters, these scores further highlight the limitations of current models in both semantic reasoning and evidence-grounded justification.

We additionally report the *Valid Partition Rate*, which measures structural compliance with the output format. This does not refer to JSON syntax validity, but to logical partitioning constraints. Specifically, whether the predicted index sets are disjoint, complete, and drawn only from the input. Top models such as Claude-3-7-Sonnet and GPT-4.1 produce structurally valid outputs in over 75% of cases. In contrast, smaller models frequently violate partition rules by omitting, duplicating, or hallucinating document indices. Importantly, the high Valid Partition Rate of top models (Claude-3-7-Sonnet: 84.9%, GPT-4.1: 76.3%) demonstrates that structural compliance is not the primary bottleneck. Human annotators, using the same JSON format and instructions, achieved a 100% Valid

[2]Constrained decoding (Poesia et al., 2022) could enforce strict schema compliance and eliminate json structural violations entirely; we leave its integration to future work.

Partition Rate while substantially outperforming all models on reasoning metrics. This confirms that the observed performance gap reflects inherent difficulty in semantic reasoning over conflicting evidence, rather than unfamiliarity with the output format.

Together, these results demonstrate that current models are not yet capable of robustly reasoning over contradictory web-based content, particularly when multi-cluster organization and explanation are required.

4.3 Human Evaluation

To better contextualize model performance, we conducted a small-scale human evaluation to establish an upper bound on the benchmark tasks. The prompt can be found in Appendix G. A total of 150 randomly selected examples were annotated manually by 5 volunteers with at least undergraduate-level education. Each volunteer was asked to group the input web documents into semantically distinct answer clusters, and provide a representative answer and list of concise supporting reasons for each cluster. On average, annotating one website took 5-10 minutes, reflecting the cognitive complexity of the task.

Table 2 compares the performance of seven LLMs with the human annotations using the same automatic metrics. The human annotators achieved an NMI of 0.691, answer coverage of 0.510, and reason coverage of 0.275 — all higher than the best-performing model (GPT-4.1), which scored 0.466, 0.448, 0.157 respectively. In addition, the human outputs were fully compliant with all structural constraints, yielding a 100% Valid Partition Rate. Since the reason score is based on the answer score, i.e., if an answer is not matched, none of its associated reasons will receive credit either, as we discussed in Section 3.4. Since most models perform poorly in Answer Coverage, their Reason Coverage score is also low. These results confirm that while the benchmark is solvable by humans, it remains challenging even for top-tier language models.

4.4 Ablation Study and Error Analysis

Contradictory vs. Non-Contradictory Samples. Contrary to intuition, models perform *better* on questions labeled as *contradictory* than on non-contradictory ones across all three metrics, as shown in Figure 5 (in Appendix). For instance, GPT-4.1 achieves an NMI of 0.51 on contradictory samples versus only 0.40 on non-contradictory ones. A similar trend holds for Answer Coverage and Reason Coverage.

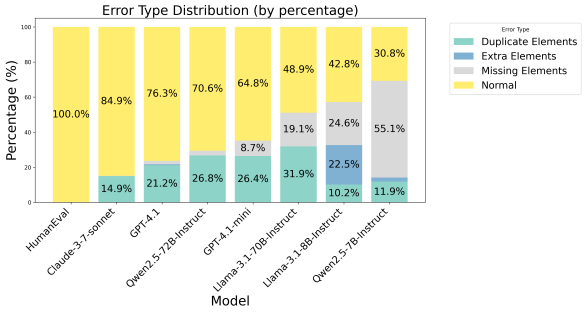

Figure 3: Distribution of structural error types across models. Claude and GPT-4.1 maintain the highest structural validity, while smaller models often duplicate, omit, or hallucinate indices.

First, contradictory questions typically present more distinct, separable viewpoints, making the clustering task easier compared to ambiguous or redundant non-contradictory inputs. Second, models may pay more attention to distinguishing differences when input references clearly diverge, whereas subtle variation in non-contradictory data may lead to under-clustering or oversimplification.

This result suggests that the difficulty of contradiction lies less in its presence and more in how well-defined the underlying perspectives are. It also highlights the importance of designing benchmarks with both types of examples, as they expose different reasoning failures.

Impact of Gold Cluster Count. We further investigate how the number of gold answer clusters affects model performance. Table 3 (in Appendix) report NMI, answer coverage, and reason coverage scores for each model, conditioned on the gold cluster count.

Across most models, performance degrades as the gold cluster number increases for answer and reason coverage. This trend suggests that current models struggle to identify and recover diverse viewpoints when the space of valid answers is more fine-grained. For instance, GPT-4.1's answer score drops from 0.48 to 0.32 when going from 2-cluster to 4-cluster samples, while its reason coverage falls by nearly half. A possible reason is that as the number of clusters increases, it becomes more difficult to separate answers into the correct groups, increasing the likelihood of merging distinct clusters or misallocating irrelevant content into isolated clusters.

An interesting observation is the human NMI score increases while the number of clusters increases, in contrast to models' NMI score trends. We further dive into this phenomenon, and find that the invalid partition rate increases while cluster number is increasing, which causes the final NMI score to decrease. Since our volunteers check their partition before they submit, they do not make those formatting errors, resulting in a monotonic increase in NMI score.

Overall, the results highlight a fundamental limitation of current models: while they can often resolve binary controversies, they fail to generalize when multiple distinct yet valid perspectives coexist. This limitation further underscores the importance of our benchmark. All previous contradictory datasets are limited to binary clustering.

Error Type Distribution Across Models.

To analyze structural robustness, we categorize partition errors into three types: (1) *Duplicate Elements*: same index appears in multiple clusters, (2) *Missing Elements* :not all input indices are assigned, and (3) *Extra Elements* :model includes non-existent indices. Figure 3 shows the error type distribution for each model.

Claude and GPT-4.1 maintain high structural validity, while smaller models often omit, duplicate, or hallucinate indices.

5 Conclusion

We present CONFRAG, a benchmark for evaluating how LLMs reason over conflicting web content. It pairs each question with retrieved documents and clusters answers by implied conclusions. Three evaluation tasks measure how models organize and explain competing viewpoints. Results show that even strong LLMs struggle with contradiction-aware reasoning, highlighting the need for better handling of source disagreement.

Limitations

First, our benchmark currently focuses on English-language open-domain questions and retrieved web documents. Extending the dataset to cover multilingual or domain-specific queries (e.g., medical or legal) may introduce new reasoning behaviors not captured in our current setup.

Second, our benchmark is limited to text-only documents, without incorporating multimodal inputs such as images or videos. Extending the dataset to multimodal settings may reveal richer patterns of conflict and reasoning.

Third, our benchmark isolates the generator component of retrieval-augmented generation (RAG) while assuming that retrieval has already been performed. RAG systems consist of two indispensable components: the retriever, which determines the scope and quality of the input evidence, and the generator, which performs reasoning over the retrieved content. Both components significantly influence end-to-end performance, but comparing them simultaneously would confound whether observed differences arise from retrieval or generation, especially since retrievers and generators are not orthogonal, a particular retriever may pair unusually well or poorly with a specific generator. While retriever design has been studied prior to the recent rise of large language models, it remains a crucial yet evolving component of RAG systems. Due to the limited scope of this work, we focus on the generator side and defer a systematic study of retrievers (particularly their ability to capture and represent conflicting evidence) to future research.

Fourth, our benchmark provides models with the ground-truth number of answer clusters k as part of the input. Without k, a model has no way to determine the appropriate granularity of clustering: it could trivially achieve a valid partition by placing all documents into a single cluster or assigning each document its own cluster, both of which are structurally correct yet semantically meaningless. Providing k eliminates this degeneracy and ensures that evaluation measures genuine reasoning ability rather than granularity preference. That said, this setting does not reflect real-world deployments where k is unknown. Evaluating models in a k-free setting where the model must also infer the number of distinct perspectives is a natural extension left for future work.

Fifth, our experimental setting evaluates LLMs end-to-end on the full contradiction-aware reasoning task. A promising direction for future agentic systems is a *classify-then-reason* pipeline: a model first detects whether the retrieved documents are contradictory, then selects a tailored strategy (e.g., a single-answer prompt for non-contradictory inputs and a multi-cluster prompt for contradictory ones). Such a two-stage design may reduce unnecessary complexity for benign inputs while allowing richer reasoning when conflicts are detected. We leave this direction to future work.

5.1 Broader Impact

CONFRAG is designed to improve the robustness and transparency of language models in retrieval-augmented settings by modeling real-world contradictions. However, the ability to cluster viewpoints could be misused to selectively promote certain narratives while suppressing others, or to simulate balanced debate in contexts with established scientific consensus, thereby undermining public trust. We recommend that CONFRAG be used primarily as a diagnostic tool for evaluating model robustness in the presence of conflicting evidence, and not as a training set for belief modeling. When used, it should be accompanied by source trust metadata and filtering strategies to mitigate risks of misuse.

Ethical Considerations

All data used in CONFRAG are collected from publicly available datasets or open-access web pages retrieved through search engines.

Human annotators involved in the quality assurance process participated voluntarily with informed consent, and all annotation tasks were limited to publicly accessible information. No sensitive demographic, personal, or identifying data were recorded or stored.

While CONFRAG is designed as a diagnostic benchmark to evaluate model robustness under conflicting web evidence, it may also pose potential risks if misused. For example, the ability to cluster and summarize divergent viewpoints could be exploited to selectively emphasize or suppress particular narratives, especially in politically or scientifically sensitive contexts. In a RAG system, filtering harmful or misleading content is primarily the responsibility of the retrieval component; the generator's role is to faithfully reason over the provided context. CONFRAG focuses on evaluating this generator-side reasoning capability and is not intended as a tool for content moderation or misinformation detection. We therefore recommend that the dataset be used strictly for research and evaluation purposes rather than for model training or content moderation.

Overall, this work aims to promote transparent, reproducible, and accountable evaluation of language models, rather than to influence public discourse or real-world decision-making.

Licenses and Terms of Use All datasets and models used in this work comply with their respective open licenses or terms of service. Specifically, ELI5 is released under CC BY-NC-SA 3.0, Natural Questions under Apache 2.0, and Yahoo Questions does not specify an explicit license but is cited in accordance with academic fair use. For language models, GPT-4.1 and Claude-3-7-Sonnet were accessed under their providers Terms of Use (OpenAI and Anthropic, respectively). The open-source models follow permissive or community licenses: Qwen2.5 is released under Apache 2.0, and Llama-3.1 under the Llama 3.1 Community License Agreement.

Our benchmark dataset, CONFRAG, will be released under the CC BY 4.0 license to encourage open research use, redistribution, and adaptation with proper attribution.

Crowdsourcing and Annotator Demographics All human annotators participated voluntarily with informed consent. They were fluent English speakers aged between 20 and 30, all based in China, and had at least undergraduate-level education. No personal, demographic, or identifying information beyond these general characteristics was collected or stored.

Use of AI Assistants We used ChatGPT to assist with dataset quality verification, coding, and writing refinement, while all methodological and analytical decisions were made by the authors.

Experiment Cost For closed-source models, running experiment consumed approximately 2.5×10^{10} tokens (each output token is counted twice to account for its generation cost). For models larger than or equal to 70B, inference typically required around 24 hours on a 4 $\times$ A800 GPU setup. For models smaller than 10B, it completed within approximately 12 hours.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No.62272261), Tsinghua University (AIR)–AsiaInfo Technologies (China) Inc. Joint Research Center, and Wuxi Research Institute of Applied Technologies, Tsinghua University under Grant 20242001120.

References

2024. The claude 3 model family: Opus, sonnet, haiku. Technical report by Anthropic.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2024. Geo: Generative engine optimization. In *KDD 2024 - Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 5–16. Association for Computing Machinery. Publisher Copyright: © 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.; 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024 ; Conference date: 25-08-2024 Through 29-08-2024.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. 2009. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, 20(2):189–201.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *Advances in Neural Information Processing Systems*, 37:109701–109747.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643, Toronto, Canada. Association for Computational Linguistics.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems*, 36:39052–39074.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models. In *International Conference on Learning Representations*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and 1 others. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Alan D. Thompson. 2025. What's in Grok? (Independent Grok-3 Paper). Subscription required for access.

Feng Wang, Zesheng Shi, Bo Wang, Nan Wang, and Han Xiao. 2025. Readerlm-v2: Small language model for html to markdown and json. *arXiv preprint arXiv:2503.01151*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Prompt used in dataset construction

A.1 Prompt used to generate questions

You are assisting in constructing a dataset of realistic, English-language, controversial or commonly misunderstood **questions** that people might ask online.

Your task is to generate up to 30 **realistic, fact-seeking, controversial questions** based on the topic: "keyword". These should:

- Reflect common misunderstandings, misinformation, or conflicting viewpoints found online
- Be suitable for retrieval-based question answering
- Be clearly phrased, answerable in principle, and free of subjective opinion
- Be questions that could plausibly appear on public forums like Quora or Reddit
- Avoid vague prompts like "What do you think about..."
- Each question should be distinct and cover different aspects of the topic

Note: If you feel like you can't generate enough different questions, just return as many as you can.

Return your answer as a **JSON array of plain English strings**, with **no duplicates**, and **no explanations**.

A.2 Prompt used to generate more questions

You are assisting in constructing a dataset of realistic, English-language, controversial or commonly misunderstood **questions** that people might ask online.

These questions will be used to evaluate hallucination defenses in retrieval-augmented generation (RAG) systems.

Topic keywords: `keywords`

Currently included questions: `exists_formatted`

Your task: Generate as many new, diverse, fact-seeking questions as you reasonably can under the above topic (ideally 20–25, but fewer is acceptable if the topic is narrow). These should:

- Reflect real-world misunderstandings, debates, or misinformation

- Be the kind of question where different sources may give conflicting answers
- Be phrased in a way that invites nuanced answers (Avoid overly generic or binary yes/no questions if possible)
- Not repeat or slightly rephrase any of the currently included questions

Output format:

```
{"more related questions": ["question1", "
    ↪ question2", ..., "questionN"]}
```

Do not include any explanation. Output only the JSON object.

A.3 Prompt used to filter controversial questions

You are a controversy detector for real-world questions.

Below is a question a user might ask on the internet. Determine if the question is likely to be controversial or misunderstood, meaning it may receive conflicting or polarized answers online.

Question: `question`

Answer yes or no, and provide a one-line explanation why or why not.

Output format:

controversial: true/false reason: <one-line explanation>

A.4 Prompt used to get the answer and reasons from one website

You are given a factual question and the full content of a webpage that may contain an answer to the question. Your task is to:

1. **Extract the core answer**: Identify the core point or conclusion from the webpage that directly answers the question. Make sure the answer is clear, concise, and directly addresses the question.
2. **Extract the supporting reasons**: Identify the reasons or explanations given in the webpage that support the core answer. Each reason should be a distinct, understandable point that explains why the answer is valid.
3. **Avoid hallucinations**: Ensure that each reason is directly or indirectly supported by the content of the webpage. Do not add information that is not present in the source content. Double-check that the extracted content is genuinely from the text and not inferred or assumed.
4. **Evaluate the trustworthiness**: Assign a **trust_score** from 0 to 10, where 10 indicates highly trustworthy content (e.g., supported by facts or reputable sources) and 0 indicates unreliable or non-relevant content.
5. **If there is an issue** (e.g., the content is completely meaningless, irrelevant, or cannot be extracted), use the **additional field** to explain the issue. For example, if the webpage content is entirely an advertisement or unrelated to the question, mark it as "meaningless content."

Output format:

```
{
  "answer": "The core answer extracted from the
      ↪ webpage that directly
  answers the question.",
  "reason": [
    "Reason 1 explaining the answer.",
    "Reason 2 explaining the answer.",
    "Reason 3 explaining the answer."
  ],
  "additional": "", // Leave empty unless there
      ↪ is an issue with the content, e.g., "
      ↪ meaningless content"
  "trust_score": 8 // Score based on the
      ↪ trustworthiness of the
  content
}
```

A.5 Prompt used to summarize to clusters

You are given a set of structured answer-reason-index triples that have been extracted from multiple webpages in response to a specific factual question. Each entry includes:

- An `index` indicating the source webpage number
- A proposed `answer` that summarizes the main viewpoint expressed on that page
- A list of `reasons` that support that answer

Your task is to:

1. **Cluster the answers into semantically consistent groups**, where each group represents a distinct, coherent viewpoint.
    - If multiple answers express the same meaning using different wording, group them into one cluster.

- Each webpage index can appear in **only one answer cluster**. Do not assign the same index to multiple answers.

2. **Aggregate the reasons** for each answer cluster:
   - Combine reasons with similar meaning from different indexes into a unified reason entry.
   - Each reason should include the list of indexes where it was found.
   - An index may support multiple reasons, but only within the same answer group.
3. **Discard meaningless or unrelated content**:
   - If a given index corresponds to irrelevant, incoherent, or meaningless content, do not include it in any answer cluster.
   - In that case, include its index in the `additional` field with a brief explanation.
4. **Detect contradictions** between answer clusters:
   - If two or more answer clusters express mutually exclusive claims, set `contradicts: true`; otherwise, use `false`.
5. **Avoid hallucinations**:
   - You must not add any content that is not present in the input.
   - Each answer and each reason must directly reflect the inputs they came from.

Output format (JSON):

```
{
  "answers": [
    {
      "answer": "Clustered answer representing a
          ↪ shared viewpoint.",
      "answer judge keyword": ["keyword1", "
          ↪ keyword2"],
      "index": [0, 3, 5],
      "reason": [
        {
          "explain": "A unified reason that
              ↪ supports this answer.",
          "reason judge keyword": ["
              ↪ reason_keyword1", "
              ↪ reason_keyword2"],
          "index": [0, 3]
        },
        {
          "explain": "Another reason with
              ↪ different emphasis.",
          "reason judge keyword": ["
              ↪ reason_keyword3"],
          "index": [5]
        }
      ]
    },
    {
      "answer": "A different answer expressing a
          ↪ contradictory or
      alternative viewpoint.",
      "answer judge keyword": ["keyword3", "
          ↪ keyword4"],
      "index": [1, 2],
      "reason": [
        {
          "explain": "A reason explaining this
              ↪ opposing viewpoint.",
          "reason judge keyword": ["
              ↪ reason_keyword5"],
          "index": [1]
        }
      ]
    }
  ],
  "contradicts": true,
  "additional": "Indexes 8 and 17 were discarded
      ↪  due to irrelevant or
  meaningless content."
}
```

Notes for the model:

- Be strict: Do not hesitate to discard data. If an index includes content that is off-topic, spammy, or incoherent, you must explain it in the additional field and leave it out of all answer clusters.
- Be grounded: Do not invent keywords, reasons, or relationships. Everything you include must come from the inputs provided.
- `answer judge keyword` should be a **minimal but sufficient** list of phrases or words that, when **any are matched**, indicate that a model has provided this answer.
- Avoid using overly generic words like "stages", "highest-ranking", or single years like "1857" unless they are paired with context. Use phrases instead (e.g., "control by 1818", "direct British rule").
- Do not use judge keywords that overlap significantly with those from other answer clusters.

B Dataset Statistics

Our final benchmark comprises 1,814 questions paired with 17,372 web documents, averaging

9.58 documents per question. These documents cover a wide range of domains (including government portals, scientific sources, news outlets, forums, and blogs) and have an average length of 2,025 words.

Among the 1,814 questions, 1,037 (57.2% of total) are labeled as containing strong contradictions between answer clusters. The remaining questions also exhibit variations in reasoning and emphasis, although their claims may not be directly incompatible.

The number of answer clusters per question ranges from 2 to 8, with the majority containing 2 to 3 clusters. Each answer cluster is supported by 2 to 3 documents on average and contains up to 3 short textual reasons. In total, the dataset includes 5,323 gold answer clusters and 35,868 unique rationale statements.

Figure 4 summarizes key statistics about document length, answer cluster complexity, support density, and dataset composition.

C Example Data

C.1 Example model answer


```
{
    "answer": {
        "answers": [
            {
                "answer": "There is no scientific
                    ↪ evidence that you need to
                    ↪ avoid dairy when you have
                    ↪ a cold; dairy does not
                    ↪ increase mucus production
                    ↪ .",
                "index": [
                    1,
                    4,
                    9,
                    11
                ],
                "reason": [
                    "Mayo Clinic (1) states that
                        ↪ drinking milk does not
                        ↪ cause the body to make
                        ↪ more phlegm and studies
                        ↪  have found no link
                        ↪ between dairy and
                        ↪ increased mucus.",
                    "Kids Plus Pediatrics (4)
                        ↪ explains that while
                        ↪ milk may coat the
                        ↪ throat and make you
                        ↪ feel more mucousy,
                        ↪ there's no evidence it
                        ↪ actually increases
                        ↪ mucus production;
                        ↪ drinking milk can help
                        ↪ stay hydrated.",
                    "Reddit r/askscience (9) cites
                        ↪  studies showing no
                        ↪ association between
                        ↪ dairy intake and
                        ↪ increased mucus or
                        ↪ respiratory symptoms
                        ↪ with colds; sensations
                        ↪ of thicker mucus are
                        ↪ subjective and not
                        ↪ confirmed by
                        ↪ measurements.",
                    "BabyCenter (11) reviews
                        ↪ multiple studies and
                        ↪ finds no conclusive
                        ↪ evidence that dairy
                        ↪ increases congestion;
                        ↪ most research shows no
                        ↪ connection, noting milk
                        ↪  may make mucus feel
                        ↪ thicker but does not
                        ↪ actually produce more.
                        ↪ Experts recommend not
                        ↪ withholding dairy from
                        ↪ sick children."
                ]
            },
            {
                "answer": "Dairy may worsen
                    ↪ congestion and cold
                    ↪ symptoms in some people,
                    ↪ and avoiding it could
                    ↪ alleviate these symptoms,
                    ↪ especially for those with
                    ↪ dairy sensitivities or
                    ↪ allergies.",
                "index": [
                    10
                ],
                "reason": [
                    "Switch4Good (10) claims that
                        ↪ cow's milk contains
                        ↪ casein, which
                        ↪ researchers hypothesize
                        ↪  can increase mucus
                        ↪ production. Consuming
                        ↪ milk may worsen a runny
                        ↪  nose, congestion, and
                        ↪ the sensation of mucus,
                        ↪  especially for those
                        ↪ with dairy allergies.
                        ↪ They recommend removing
                        ↪  dairy to alleviate
                        ↪ congestion during a
                        ↪ cold."
                ]
            }
        ]
    }
}
```


As mentioned in Section 4.1, we selected 5 websites to be tested.

C.2 Example Data


```
{
    "id": 11,
    "question": "Are children more susceptible to
        ↪  radiation from electronic devices?",
    "websites": [
        {
            "content": "Title: Children and Cell
                ↪ Phones: Is Phone Radiation
```

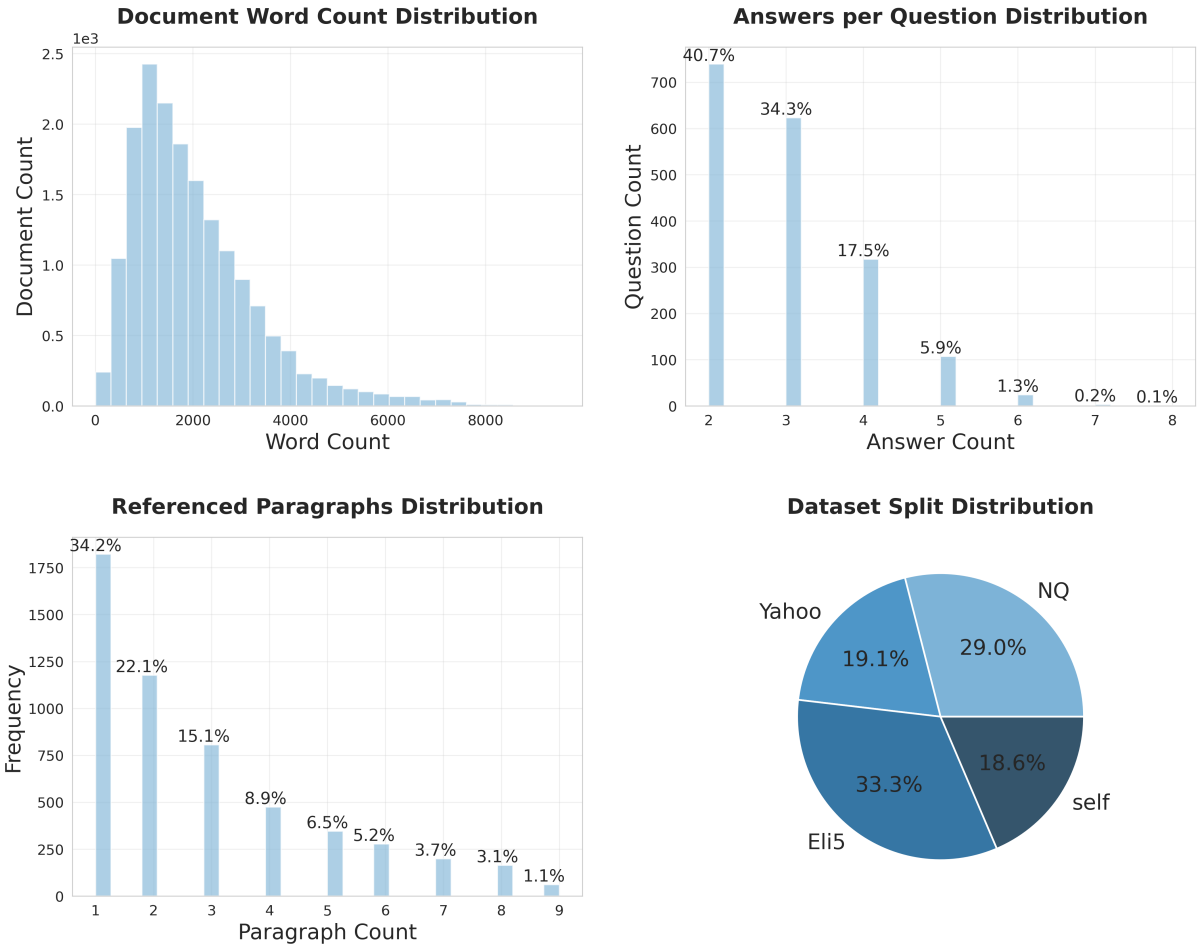

Figure 4: Dataset statistics. Top-left: Distribution of word counts per document. Top-right: Number of answer clusters distribution. Bottom-left: Number of supporting documents per answer cluster. Bottom-right: Source dataset distribution. (NQ stands for Natural Questions dataset)


```
        ↪ Risky for Kids? - National
        ↪ Center for Health Research\n\
        ↪ nURL Source: https://www.
        ↪ center4research.org/children-
        ↪ cell-phones-phone-radiation-
        ↪ risky-kids/\n",
    // total 1638 words.
    "answer": "Yes, children are more
        ↪ susceptible to radiation from
        ↪ electronic devices.",
    "reason": [
        "Children's brains absorb 2-3
            ↪ times more radiation than
            ↪ adults' brains due to
            ↪ their smaller heads and
            ↪ brains, which receive the
            ↪ same levels of cell phone
            ↪ radiation as adults.",
        "The American Academy of
            ↪ Pediatrics states that RF
            ↪ energy deposition is
            ↪ higher in children, with
            ↪ two times higher in the
            ↪ brain and ten times higher
            ↪  in the bone marrow of the
            ↪  skull compared to adults
            ↪ .",
        "Research suggests that starting
            ↪ cell phone use before age
            ↪ 20 increases the risk of
            ↪ developing brain tumors
            ↪ due to thinner tissues and
            ↪  bones in children, which
            ↪ allow closer proximity of
            ↪ radiation to the brain.",
        "Cell phone exposure has been
            ↪ linked to behavioral
            ↪ problems in children, such
            ↪  as ADHD, especially when
            ↪ combined with high levels
            ↪ of lead in the blood."
    ],
    "trust_score": 8,
    "index": 2,
    "website": "https://www.
        ↪ center4research.org/children-
        ↪ cell-phones-phone-radiation-
        ↪ risky-kids/"
}
// 9 more websites are omitted due to the
    ↪  length of the paper
],
"from": "Self generated",
"contradicts": true,
```

```
"answers": [
    {
        "answer": "Yes, children are more
            ↪ susceptible to radiation from
            ↪ electronic devices.",
        "answer judge keyword": [
            "children",
            "radiation",
            "electronic devices"
        ],
        "index": [
            1,4,7,9,11,12,16,19
        ],
        "reason": [
            {
                "explain": "Children's
                    ↪ developing bodies and
                    ↪ brains absorb more
                    ↪ radiation due to
                    ↪ anatomical differences
                    ↪ such as thinner skulls
                    ↪ and higher water
                    ↪ content in tissues.",
                "reason judge keyword": [
                    "developing bodies",
                    "thinner skulls",
                    "absorb more radiation"
                ]
            },
            {
                "explain": "Research and
                    ↪ expert recommendations
                    ↪ indicate higher
                    ↪ susceptibility and
                    ↪ potential health risks,
                    ↪  including behavioral
                    ↪ problems and cancer.",
                "reason judge keyword": [
                    "higher susceptibility",
                    "health risks",
                    "behavioral problems",
                    "cancer"
                ]
            },
            {
                "explain": "Current standards
                    ↪ may not adequately
                    ↪ protect children, and
                    ↪ recommendations suggest
                    ↪  reducing exposure.",
                "reason judge keyword": [
                    "current standards",
                    "reduce exposure",
                    "not adequately protect"
                ]
            }
        ]
    },
    {
        "answer": "Children are not more
            ↪ susceptible to radiation from
            ↪ electronic devices,
            ↪ specifically non-ionizing
            ↪ electromagnetic fields (EMFs),
            ↪  as there is no consistent
            ↪ evidence linking these
            ↪ exposures to cancer in
            ↪ children.",
        "answer judge keyword": [
            "non-ionizing EMFs",
            "no consistent evidence",
            "cancer"
        ],
        "index": [
            10
        ],
        "reason": [
            {
                "explain": "Non-ionizing EMFs
                    ↪ are not known to damage
                    ↪  DNA or cells directly,
                    ↪  and studies have not
                    ↪ found consistent
                    ↪ evidence of cancer risk
                    ↪  in children.",
                "reason judge keyword": [
                    "non-ionizing EMFs",
                    "not damage DNA",
                    "no consistent evidence",
                    "cancer risk"
                ]
            }
        ]
    }
]
}
```


D Formulas for Evaluation Protocol

D.1 Answer Clustering

NMI is defined as:

$$\mathrm{NMI}(U,V) = \frac{2 \cdot I(U;V)}{H(U) + H(V)} \tag{1}$$

Let $U = \{U_1, ..., U_K\}$ and $V = \{V_1, ..., V_L\}$ be the sets of gold and predicted clusters. Then:

$$I(U;V) = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{|U_i \cap V_j|}{N} \log \left(\frac{N \cdot |U_i \cap V_j|}{|U_i| \cdot |V_j|} \right) \tag{2}$$

$$H(A) = -\sum_{i=1}^{K} \frac{|A_i|}{N} \log \left(\frac{|A_i|}{N} \right), for A = U, V \tag{3}$$

where N is the total number of documents.

D.2 Answer Coverage

Let $G = (A, B, E)$ be a bipartite graph where A is the set of gold answers (or reasons), B is the set of predicted ones, and $(a_i, b_j) \in E$ if a_i is a case-insensitive substring of b_j (case-free). We define M as the size of the maximum matching on G:

$$M = \max_{M' \subseteq E} |M'|$$
$$\text{s.t.} \quad \forall (a_i, b_j), (a_{i'}, b_{j'}) \in M',\ i \neq i',\ j \neq j' \tag{4}$$

Then the final score is:

$$\mathrm{Score}_{\text{answer}} = \frac{M}{\sqrt{N_{\text{pred}} \cdot N_{\text{gold}}}} \in [0, 1] \tag{5}$$

where N_{pred} and N_{gold} are the number of predicted and gold answers respectively. In ideal case, $N_{\text{pred}} = N_{\text{gold}}$ since we have told the LLM how many clusters we are expecting.

D.3 Reason Coverage

Let $A = \{a_1, ..., a_m\}$ be the set of gold answers and $B = \{b_1, ..., b_n\}$ the set of predicted answers. Each a_i and b_j is associated with a set of reasons $R(a_i)$ and $R(b_j)$.

We first compute a maximum matching $M \subseteq A \times B$ such that $(a_i, b_j) \in M$ iff a_i and b_j match via keyword overlap (as in Task 2).

For each matched pair $(a_i, b_j) \in M$, define $M_r^{i,j}$ as the size of the maximum matching between reasons in $R(a_i)$ and $R(b_j)$ (via keyword overlap), and compute:

$$\text{Score}_{i,j}^{\text{reason}} = \frac{M_r^{i,j}}{\sqrt{|R(a_i)| \cdot |R(b_j)|}} \in [0, 1] \quad (6)$$

The final reason score is the average over all matched answer pairs:

$$\text{Score}_{\text{reason}} = \frac{\sum_{(a_i,b_j)\in M} \text{Score}_{i,j}^{\text{reason}}}{\sqrt{N_{pred} * N_{gold}}} \in [0, 1] \quad (7)$$

E Figure and Table for Ablation Study

See Figure 5 for Model performance comparison on contradictory vs. non-contradictory samples. See Table 3 for Model performance across questions with different gold cluster counts.

F Prompt for doing experiment

You are given: - A question - A set of webpage contents, each with a corresponding index - A target number of clusters (clustering number)

Task: Your task is to process the webpage contents and organize them into the required number of clusters based on their answers to the question.

Rule: Each cluster should contain:

1. A single core "answer" that represents the main viewpoint of the group.

2. A list of "index" values representing the webpages that support this answer.

3. A list of "reason" entries explaining why this answer is supported. Each reason should be a concise, specific point.

Instructions: - You must create exactly [CLUSTERING NUMBER] clusters, no more, no fewer.

- You must include all index in your clusters, do not miss any.

- Each answer should be distinct and represent a unique viewpoint.

- Supporting Reasons must be grounded in the content of the webpages. Do not invent information that is not present in the original content.

- Be concise and factual. Avoid redundant or overly generic answers and reasons.

- If a webpage content supports multiple reasons under the same answer, include all relevant reasons under the corresponding answer.

- Output must be a valid JSON object exactly as specified. Do not include any additional text, comments, or explanations outside the JSON.

- Clusters must form a non-overlapping partition of all indexes: each index must appear in one and only one cluster.

Example: Output strictly in the following JSON format:

```
{
  "answers": [
    {
      "answer": "The core answer representing one
          ↪  cluster.",
      "index": [1, 3],
      "reason": ["Supporting reason 1", "
          ↪ Supporting reason 2"]
    },
    {
      "answer": "The core answer representing
          ↪ another cluster.",
      "index": [2],
      "reason": ["Supporting reason 3", "
          ↪ Supporting reason 4"]
    }
    // Continue until [CLUSTERING NUMBER]
        ↪ clusters are completed
  ]
}
```

Attention: - The numeric contents of the index arrays in different answers must not overlap. Each index should appear in exactly one index list.(such as "index": 1, 3 in [1, 3] and 2 in index [2] are mutually exclusive)

- Each reason must be clearly supported by content from the corresponding webpage(s). Do not fabricate or generalize beyond what is explicitly stated.

- The entire output must be a valid JSON object that follows the given format strictly, no additional comments, metadata, or text.

- Ensure that the total set of indexes across all clusters forms a complete partition: every original index is included and none are duplicated.

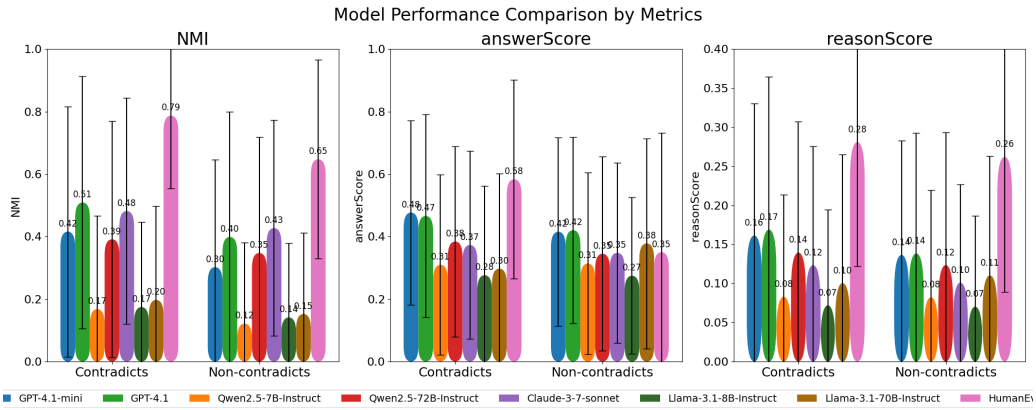

Figure 5: Model performance comparison on contradictory vs. non-contradictory samples. Left: Answer Clustering (NMI). Middle: Answer Coverage. Right: Reason Coverage. Surprisingly, models achieve higher scores on contradictory samples, likely due to the clearer separation of viewpoints and more explicit conflict in the source content. The human eval reason Score is especially high since the reason score is based on answerScore. If one answer is not matched, its all reason is effectively useless. The error bar shows a standard deviation

the '[CLUSTERING NUMBER]' will be replaced by real clustering number when doing experiment. A json contains websites content and index will be added at last. Due to the length of website content (on average 2,025 words), we do not show it here.

G Human Eval Prompt

The prompt for each question is the same as the prompt given to LLM which is shown in Appendix F. The only difference is we replace the website content added in prompt of LLM to the url. If user found the url is unavailable, we will give them another url to make sure they get 5 websites every time.

Table 3: Model performance across questions with different gold cluster counts. Each row shows the model's performance on samples with exactly 2, 3, or 4 answer clusters. As the number of gold clusters increases, performance generally declines, indicating higher difficulty for multi-cluster separation and justification. The data are reported as mean (standard deviation)

model	NMI			Invalid Partition Rate			Answer Score			Reason Score		
cluster	2	3	4	2	3	4	2	3	4	2	3	4
GPT-4.1-mini	0.39 (0.39)	0.33 (0.36)	0.37 (0.43)	0.22	0.51	0.56	0.48 (0.31)	0.44 (0.28)	0.34 (0.27)	0.16 (0.17)	0.15 (0.15)	0.10 (0.13)
GPT-4.1	0.47 (0.40)	0.47 (0.41)	0.44 (0.43)	0.16	0.39	0.48	0.48 (0.33)	0.44 (0.30)	0.32 (0.26)	0.17 (0.19)	0.16 (0.18)	0.09 (0.13)
Qwen2.5-7B-Instruct	0.13 (0.26)	0.18 (0.30)	0.17 (0.33)	0.66	0.72	0.80	0.33 (0.31)	0.30 (0.26)	0.23 (0.21)	0.09 (0.14)	0.08 (0.13)	0.07 (0.09)
Qwen2.5-72B-Instruct	0.40 (0.37)	0.35 (0.36)	0.27 (0.42)	0.12	0.47	0.70	0.41 (0.33)	0.33 (0.28)	0.27 (0.24)	0.15 (0.17)	0.12 (0.17)	0.07 (0.10)
Claude-3-7-Sonnet	0.42 (0.34)	0.52 (0.34)	0.46 (0.44)	0.05	0.25	0.47	0.40 (0.31)	0.32 (0.27)	0.26 (0.24)	0.13 (0.15)	0.10 (0.13)	0.08 (0.11)
Llama-3.1-8B-Instruct	0.16 (0.23)	0.19 (0.30)	0.11 (0.27)	0.45	0.69	0.85	0.31 (0.30)	0.23 (0.23)	0.25 (0.22)	0.08 (0.13)	0.06 (0.10)	0.06 (0.10)
Llama-3.1-70B-Instruct	0.18 (0.28)	0.18 (0.28)	0.16 (0.32)	0.51	0.69	0.80	0.36 (0.35)	0.30 (0.27)	0.27 (0.26)	0.11 (0.17)	0.11 (0.14)	0.07 (0.12)
HumanEval	0.64 (0.34)	0.77 (0.14)	0.87 (0.13)	0.00	0.00	0.00	0.53 (0.35)	0.48 (0.39)	0.44 (0.34)	0.30 (0.17)	0.22 (0.14)	0.25 (0.11)