MDBench: A Synthetic Multi-Document Reasoning Benchmark Generated with Knowledge Guidance

Joseph J. Peper[1] **Wenzhao Qiu**[1] **Ali Payani**[2] **Lu Wang**[1]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, MI
[2]Cisco Research, San Jose, CA
{jpeper,qwzhao,wangluxy}@umich.edu
apayani@cisco.com

Abstract

Natural language processing evaluation has made significant progress, largely driven by the proliferation of powerful large language models (LLMs). New evaluation benchmarks are of increasing priority as the reasoning capabilities of LLMs are expanding at a rapid pace. In particular, while *multi-document* (MD) reasoning is an area of extreme relevance given LLM capabilities in handling longer-context inputs, few benchmarks exist to rigorously examine model behavior in this setting. Moreover, the multi-document setting is historically challenging for benchmark creation due to the expensive cost of annotating long inputs.

In this work, we introduce **MDBench**, a new dataset for evaluating LLMs on the task of multi-document reasoning. Notably, MDBench is created through a novel synthetic generation process, allowing us to *controllably and efficiently generate challenging document sets* and the corresponding question-answer (QA) examples. Our novel technique operates on condensed structured seed knowledge, modifying it through LLM-assisted edits to induce MD-specific reasoning challenges. We then convert this structured knowledge into a natural text surface form, generating a document set and corresponding QA example. We analyze the behavior of frontier LLMs and prompting techniques, finding that MDBench poses significant challenges for all methods, even with relatively short document sets. We also see our knowledge-guided generation technique (1) allows us to readily perform targeted analysis of MD-specific reasoning capabilities and (2) can be adapted quickly to account for new challenges and future modeling improvements.

1 Introduction

The rapid advancements in natural language processing (NLP) have been largely driven by the development and deployment of large language models (LLMs). These models have showcased remarkable improvements in various tasks, including understanding, generating, and reasoning over text. However, despite these advancements, evaluation frameworks for NLP systems have struggled to keep pace (Chang et al., 2024), notably for tasks involving reasoning over multiple documents (Mavi et al., 2024).

Multi-document (MD) reasoning involves synthesizing and inferring information across multiple diverse texts (Caciularu et al., 2021), posing unique challenges not addressed by traditional single-document benchmarks. While LLMs are increasingly capable of handling longer-context multi-document inputs, there is a scarcity of benchmarks that rigorously examine the specific reasoning characteristics that are prominent in this setting. In addition, many existing benchmarks consist of static, hand-crafted datasets, which are labor-intensive to produce. These datasets are often susceptible to data contamination (Xu et al., 2024) over time, e.g., LLMs are exposed to public benchmarks during training. This can compromise the integrity of the evaluation.

In this work, we address these limitations with **MDBench**[1], a benchmark using a novel generation technique for multi-document reasoning evaluation. Our benchmark is generated through a synthetic process that leverages structured knowledge as seed information. This process uses a strong LLM (GPT-4o) to augment structured knowledge by injecting complexities that require advanced reasoning skills, then generates text documents from the augmented knowledge. Synthetic generation allows for efficient example creation and can be easily adapted to incorporate new reasoning skills and objectives.

Our benchmark generation pipeline (overviewed in Figure 1) begins with a structured knowledge

[1]Benchmark can be found at https://huggingface.co/launch/MDBench. Code and prompts are accessible at https://github.com/jpeper/MDBench.

Findings of the Association for Computational Linguistics: ACL 2025, pages 25592–25621
July 27 - August 1, 2025

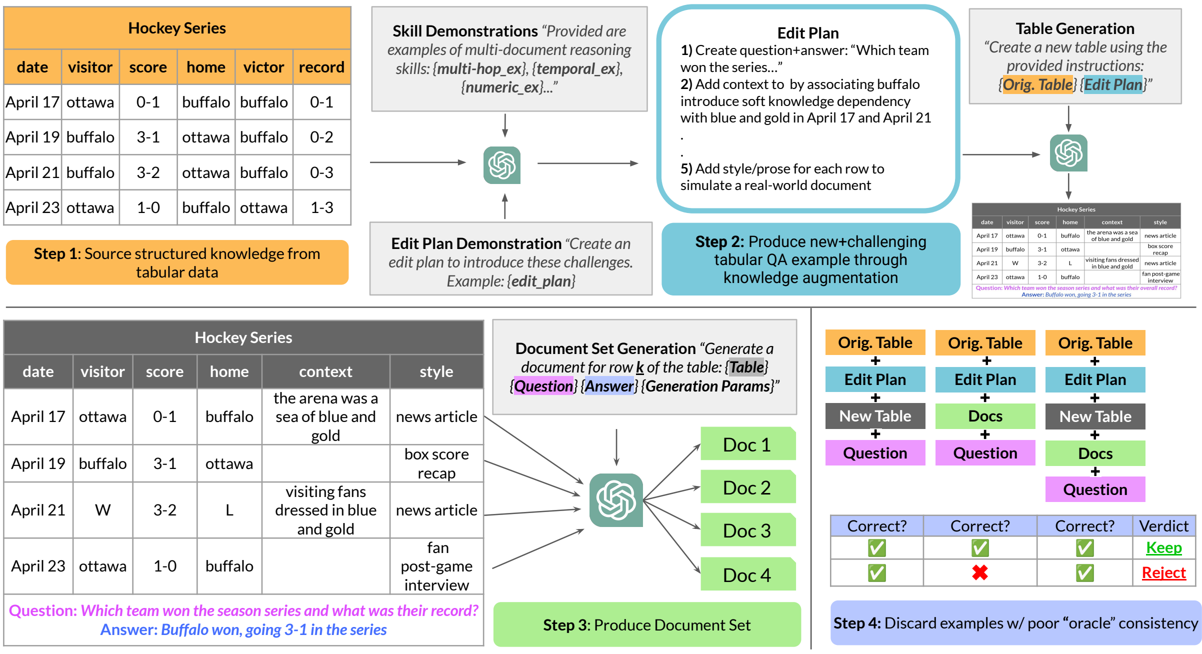

Figure 1: MDBENCH generation pipeline overview. We 1) source structured knowledge, then use in-context multi-document reasoning demonstrations to 2) intentionally modify the existing knowledge with challenging dependencies. We then 3) map this seed knowledge into document form to produce the multi-document QA example. Finally, we 4) employ an added validity check to filter inconsistent examples that yield erratic outputs in an oracle setting where example construction knowledge is provided.

source serving as the seed information. Each knowledge entry (i.e., a row of a table) encapsulates distinct knowledge that forms the basis of a document in the generated set. We use tabular data as these are widely utilized and inherently structured, with row entries often sharing topical connections, making them an ideal foundation for evaluating cross-document reasoning. We follow a multi-step augmentation process to source knowledge, augment knowledge, and generate high-quality document sets with multi-document reasoning challenges:

1. **Source Seed Knowledge (Fig. 1, Step 1)** We collect tabular data where each row contains information that will contribute to a generated document.

2. **Augment Knowledge (Fig. 1, Step 2)** Using GPT-4o, we edit the structured knowledge to inject challenging reasoning dependencies and enrich the context for document creation. By treating rows as proxies for documents, we model cross-document dependencies through cross-row knowledge interactions. In this step, we also generate question-answer pairs that utilize the introduced reasoning dependencies.

3. **Generate Natural Text (Fig. 1, Step 3)** We map the augmented knowledge into natural text by generating a corresponding multi-document set from the augmented table. This process allows us to systematically inject critical reasoning challenges while producing examples that are realistic and fluent.

4. **Automated Quality Validation (Fig. 1, Step 4)** To ensure high quality and consistency, we apply targeted consistency prompts and a novel oracle-setting self-consistency check that effectively filters out low-quality examples, enabling rapid benchmark scaling.

We produce 1,000 multi-document QA examples using this pipeline (300 human-validated, and 700 more automatically-validated for quality) and evaluate the performance of models from several prominent LLM families including GPT (OpenAI et al., 2024), Claude (Anthropic), Gemini (Team et al., 2024a), and LLaMA (Dubey et al., 2024). We find that:

- MDBENCH poses a strong challenge, even for state-of-the-art methods, with the best ones achieving ∼81% accuracy and ∼62% exact-match performance on this MD reasoning task.

- Frontier models such as GPT-4o, GPT-o1 and Gemini-2.5-Flash significantly outperform smaller LLMs across different prompting methods. This highlights the importance of model capacity and sophistication in handling complex multi-document reasoning tasks.

- When comparing performance on document reasoning versus tabular reasoning (i.e., structured format pre-document generation), we find that strong models are mostly performant in both settings, but all struggle more in the document setting. This suggests that *multi-document reasoning is influenced by both the fundamental reasoning complexity, and also from the nuances of the surface form.*

- Prompting techniques such as chain-of-thought (Wei et al., 2022) can improve performance across strong models. However, they are insufficient to significantly enhance the performance of weaker models like LLaMA-3-7B and even GPT-3.5. This indicates that while prompting strategies can aid reasoning, *underlying model capabilities remain a limiting factor for this task, which makes* MD-BENCH *suitable for future, advanced model evaluation.*

2 Related Work

Evaluating the capabilities of LLMs is a critical aspect of NLP research. As LLMs continue to improve rapidly, existing evaluation frameworks often lag behind, particularly in assessing complex reasoning abilities such as multi-document (MD) reasoning. As LLMs rapidly increase in reasoning capacity, there is a pressing need to develop evaluation methods that can capture these higher-order reasoning skills.

Multi-Document Reasoning MD reasoning involves synthesizing and inferring information across multiple texts. Existing work in this area includes datasets targeting specific phenomena such as temporal reasoning (Xiong et al., 2024; Wan, 2007), summarization (Xiao et al., 2021; Peper et al., 2023; Lior et al., 2024), multi-hop question answering with HotPotQA, MuSiQue and 2Wiki-MultiHop (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022) and ambiguous entity resolution (Lee et al., 2024). Notably, many of these MD datasets are publicly-sourced and often reliant on significant human effort to curate. For example, Zhu et al. (2024) introduce FanOutQA, a recent multi-hop, multi-document question answering dataset, which targeted decomposable QA examples sourced from public Wikipedia knowledge and relied on thousands of manual annotations. Our work seeks to use knowledge-controlled generation to offer a scalable alternative for producing nuanced and unseen multi-document reasoning examples.

Tabular Reasoning with LLMs LLMs have demonstrated strong performance in tasks involving structured knowledge, such as tabular data or knowledge bases (Lu et al., 2024; Li et al., 2023a). Recent studies have observed success in applying LLMs to table reasoning, manipulation, and augmentation (Lu et al., 2024; Li et al., 2023a). While there are limitations in LLM pre-training which can lead to formatting sensitivities and limitations with handling large tables, Nahid and Rafiei (2024) find improved performance by decomposing the tabular knowledge into a digestible size. Similarly, leveraging tabular knowledge within reasoning chains allows for compact and effective representation of complex problems, as explored in the Chain-of-Tables framework (Wang et al., 2024). These insights highlight the potential of using condensed knowledge as a foundation for generating challenging reasoning tasks. Although our work draws on structured data to simplify the creation of examples, our ultimate goal is to produce longer, more nuanced natural-language text document sets.

LLM-Supported Synthetic Benchmark Creation To address the need for more dynamic evaluation datasets, LLM-powered synthetic benchmark creation has gained significant traction (Long et al., 2024; Liu et al., 2024; Li et al., 2023b), particularly as there is growing concern of benchmark data contamination (Xu et al., 2024). Some work has been done in the multi-document setting, although automation is largely used for extending existing annotated multi-document benchmarks to more complex tasks (Schnitzler et al., 2024). While not directly modeling multi-document tasks, works like BoardGameQA (Kazemi et al., 2023) and MuSR (Sprague et al., 2023) explore synthetic generation in the related multi-step reasoning setting, with MuSR using a neurosymbolic generation algorithm which maps synthetic structure into natural text examples. Our method seeks to build off related work in synthetic generation to address efficient multi-document benchmark creation.

3 MDBENCH Generation Pipeline

In this section, we motivate and overview the generation process, and provide details on the components and steps taken to produce the MDBENCH evaluation benchmark.

Our design decisions for the MDBENCH benchmark are guided by several key objectives to ensure the resulting examples test a broad and relevant range of reasoning skills:

- **Contain Novel and Unseen Text**: We aim to produce examples that are not merely scraped from public datasets but rather contain newly-generated content. This ensures that models are tested on scenarios they have not encountered during training, avoiding overfitting to pre-existing benchmarks.
- **Contains Cross-Document Knowledge Dependencies**: A key focus is to produce examples that require reasoning across multiple documents. We design our benchmark to have intentional cross-document dependencies, making them particularly challenging to effectively test multi-document reasoning capabilities.
- **Grounded in Real-World Scenarios**: Even though the examples are synthetically generated, they should ideally remain grounded in real-world concepts and situations. This ensures that the reasoning challenges presented are realistic and relevant to practical NLP applications.
- **Counterfactual Alterations**: To further mitigate data contamination and leakage risks from public sources, we allow slight counterfactual or fictional twists on the real-world scenarios occurring in the seed dataset to necessitate that models leverage the extrinsic knowledge found in the input. This allows for a fresh take on familiar domains while maintaining the integrity of the benchmark.
- **Scalability and Control**: Our approach is designed to offer control during benchmark generation. We allow one to specify seed information such as domain and behavior types, and can control the complexity and nature of the reasoning tasks present in the benchmark.

	Min	Mean (std)	Max
# Rows per Table	5	8.31 (2.4)	17
# Table Columns	3	5.17 (1.2)	9
Token Length (Tab.)	121	256.0 (81.6)	554
Token Length (Doc.)	1048	2397.4 (738.4)	6493
Avg. Doc Length	177	268.1 (37.6)	388

Table 1: MDBENCH benchmark statistics. Each row in the tabular representation ultimately corresponds to a document within the multi-document example. We see a roughly nine-fold increase in surface form length when mapping the structured knowledge to natural text document format.

date	territory	showings
october 20, 2006	turkey	200
october 20, 2006	belgium	600

Baseline Question: Which country had the most showings and how many total?
Answer: Belgium had the most with 600 showings.
Answer rationale: Turkey had 200 showings and Belgium had 600. 600 > 200, therefore Turkey had the most showings.
Commentary: This is a simple reasoning process as it requires a simple comparison of two values with no additional reasoning required.

date	territory	showings
october 20, 2006	turkey	200
october 20, 2006	belgium	600
october 25, 2006	turkey	500

Harder Example: Which country had the most showings and how many total?
Answer: Turkey had the most with 700 showings.
Answer Rationale: Turkey had showings on two different days, so the total is 200+500=700 showings. 700 > Belgium's 600, therefore Turkey had the most.
Commentary: By adding a new row with complementary information, we necessitate an additional reasoning hop to correctly answer the question. Note that this table was edited specifically such that the answer (Turkey) is flipped from the original answer (Belgium) in the simple example.

Figure 2: Example skill description on "Multi-hop Reasoning". During knowledge augmentations, we demonstrate the multi-document skills relevant to the document sets.

3.1 Pipeline Overview

Our benchmark generation pipeline begins with structured knowledge sourced from tabular data, which serves as the seed for the augmentation process. This structured knowledge is systematically enriched and refined through a strong LLM to inject reasoning dependencies that challenge models to infer information across multiple documents. Figure 1 overviews the pipeline. Below, we provide details of the four major steps:

Step 1: Obtaining Seed Knowledge We start with an intuition that compressed structured knowledge provides an effective foundation for multi-document reasoning. Several valid sources of this exist, such as knowledge bases, tabular information, or even by performing information extraction to consolidate data from existing documents and text corpora. For the MDBENCH benchmark, we utilize the TabFact (Chen et al., 2020) dataset, which comprises 16,000 tables sourced from Wikipedia. Our motivation for exploring this dataset is threefold: (1) TabFact tables provide a reliable and curated

Reasoning Type	Description
Multi-hop Reasoning	Solving problems requiring multiple steps to arrive at the solution.
Numeric Reasoning	Handling numeric values and performing numerical operations.
Temporal Reasoning	Handle temporal information and temporal dependencies.
Knowledge Aggregation	Aligning, comparing and/or contrasting knowledge that may be present.
Soft Reasoning	Reasoning abductively and making informed decisions in cases where some uncertainty or fuzziness may be present, such as cross-document entity linking.

Table 2: Reasoning skills overview. For our benchmark, we focus on five goals which are especially relevant for the multi-document setting. We provide demonstrations of these reasoning types to inspire relevant knowledge edits during the generation process.

source of seed knowledge; (2) the data spans a wide range of domains, including news, sports, media, and technology; and (3) it has an emphasis on human-readability both in scale and content. This structured knowledge serves as the starting point for our knowledge augmentation process, which significantly transforms the raw data into more challenging and complex reasoning tasks. We heuristically filter the dataset to select tables which are rich in content yet manageable in size, choosing those with 5 to 17 rows and 3 to 9 columns.

Step 2: Knowledge Augmentation An important component of our technique is the knowledge augmentation step. This step modifies information, applying operations that inject complex knowledge dependencies and reasoning challenges. Figure 1 overviews our pipeline, while full detailed examples of the knowledge augmentation prompts are provided in Appendix A. Below, we describe two key techniques we employ:

- **Multi-document Reasoning Demonstrations** Prior to altering the existing information, we first demonstrate relevant skills for multi-document reasoning. Each skill is demonstrated in both 'simple' and 'challenging' forms (e.g., Figure 2). The demonstrations include examples, along with explanations and rationales for solving them. For the purpose of this benchmark, we define and focus on five reasoning components which are particularly relevant in the multi-document setting: (1) `Multi-hop`, (2) `Numeric`, (3) `Temporal`, (4) `Soft reasoning`, and (5) `Knowledge Aggregation`.

  We briefly describe these skills in Table 2, and provide full demonstrations of each in Appendix A.

- **Knowledge Augmentation Demonstrations** In addition to demonstrating relevant reasoning skills, we next provide *knowledge edit demonstrations*. These demonstrations illustrate plans for how simple tables can be enhanced to form nuanced QA examples. Each demonstration consists of an initial table, a series of edits, and a resultant augmented table and QA annotation. When performing knowledge augmentation, we sample one demonstration from a small set of high-quality curated examples. Figure 6 demonstrates a knowledge edit plan.

Through these two steps, we modify the tabular knowledge to form a more nuanced QA example with cross-row knowledge dependencies.

Step 3: Document Set Generation Once the tabular knowledge has been augmented, we map this information into natural language text; each row in the table is used to generate a document, with the augmented knowledge ensuring that reasoning across documents (rows) is required to solve the accompanying QA task. We independently generate each document, with the generation prompt parameterized by the following components: (1) the augmented table and title, (2) the column names, and (3) a specific row of content within the table indicated for generation. Iterating this process over all n rows in the table, we generate an n-document set. This approach of knowledge-grounded generation ensures the generated document set maintains logical coherence while presenting unique cross-document reasoning challenges. Figures 12 and 13 demonstrate a full generated MDBENCH multi-document example.

Step 4: Automated Quality Validation To ensure a high-quality and scalable benchmark, our pipeline employs a multi-step automated validation process. First, we apply targeted prompts (Appendix D) to verify that each component—the seed table, edit plans, intermediate edited table, and final document set—adheres to the intended instructions. Next, we perform an oracle self-consistency check based on the intuition that the procedural steps used to construct an example can also verify its consistency. Specifically, for each QA example, we prompt GPT-4o under three variations of *oracle knowledge* (original table and edit plan) and *context* (generated table and generated document set) as shown in Figure 1. An example is rejected if *any* of the three variations yield an answer which does not match the generated ground-truth. This comprehensive validation ensures internal consistency and *underpins scalable, automated quality control.*

3.2 MDBench Benchmark Generation Details

We utilize GPT-4o for both table augmentation and document generation. Table 1 summarizes the statistics of the resultant benchmark. We produce a benchmark comprising 1,000 examples.

Automated Validation Our pipeline integrates automated validation (including both localized validation checks as well as the oracle self-consistency check) to ensure quality and consistency throughout generation. During generation, we find the automated filtering retained approximately 32% of the generated examples.

Human Validation To assess the quality of the generated and auto-verified examples, we hire and train a native English-speaking graduate student with linguistics background for this task (compensated at $15 per hour for training and annotation). The annotator is instructed to validate the consistency of the generated multi-document QA examples (see Appendix 5). We find our automated validation yields an overall example validity of rate of 87%. Our final released benchmark comprises 300 human-validated examples and 700 additional machine-validated examples.

4 Experimental Setup

To assess the challenges of MDBench, we test the performance of many popular LLMs in combination with conventional prompting setups. Concretely, we test open-weights LLMs with Meta's **LLaMA-3** (Dubey et al., 2024), using the 8B-Instruct and 70B-Instruct variants. For API-based

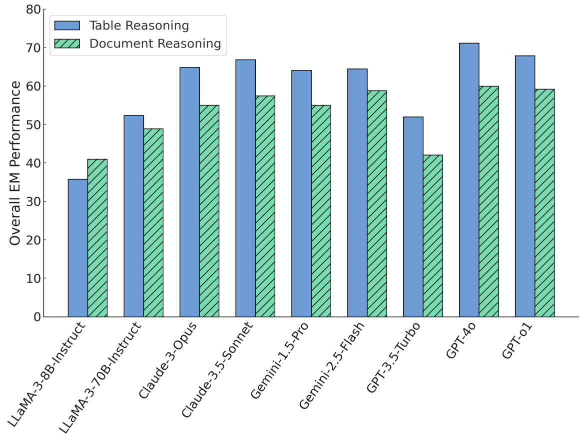

Figure 3: Overall exact-match performance of models on MDBench. Table Reasoning is when evaluated with the intermediate table QA examples. Document Reasoning refers to the performance on the *final task* of multi-document reasoning. Document reasoning poses a harder challenge, speaking to the importance of surface form in reasoning performance.

proprietary methods, we use models from the Anthropic **Claude**, OpenAI **GPT**, and Google **Gemini** families, which represent current state-of-the-art LLMs. For Claude we evaluate Claude-3-Opus `20240229` and Claude-3.5-Sonnet `20240620`[2]. For GPT we test GPT-3.5-Turbo-16k `0613`[3] (Ouyang et al., 2022), GPT-4o `2024-08-06`[4], and GPT-o1 `2024-12-17`[5]. For Gemini we include Gemini-1.5-Pro `0514` (Team et al., 2024b) and the recent Gemini-2.5-Flash-preview `05-20`[6].

We explore both *zero-shot* and *one-shot* QA prompting scenarios, noting that when prompting in the one-shot case we use a single representative demonstration across models for consistency. We use a conventional question-answering prompt, and also further instruct the models to 'think step by step' to additionally produce *chain-of-thought* (CoT) rationales. Examples of these prompt formats are provided in Appendix B. To evaluate on the QA task, we use GPT-4o as a reference-based scorer, first parsing the final answer from each output, then comparing the similarity of the predicted answer with the ground-truth answer (conditioned on the original question). We calculate

[2] https://www.anthropic.com/claude
[3] https://platform.openai.com/docs/models/gpt-3-5
[4] https://platform.openai.com/docs/models/gpt-4o
[5] https://platform.openai.com/docs/models/o1
[6] https://ai.google.dev/gemini-api/docs/models/2.5-flash

Model	Zero-shot	Zero-shot CoT	One-shot	One-shot CoT	Overall
LLaMA-3-8B-Instruct	42.7	43.1	39.3	38.9	41.0
LLaMA-3-70B-Instruct	51.2	50.2	45.0	49.3	48.9
Claude-3-Opus	55.5	54.5	56.4	53.6	55.0
Claude-3.5-Sonnet	59.2	56.4	58.3	55.9	57.5
Gemini-1.5-Pro	54.0	57.8	54.0	54.0	55.0
Gemini-2.5-Flash	57.8	58.3	58.8	**60.2**	58.8
GPT-3.5-Turbo	49.8	37.4	44.5	36.5	42.1
GPT-4o	**59.7**	**62.1**	**59.7**	58.3	**60.0**
GPT-o1	57.3	60.7	**59.7**	59.2	59.2

Model	Zero-shot	Zero-shot CoT	One-shot	One-shot CoT	Overall
LLaMA-3-8B-Instruct	65.7	66.8	60.7	63.5	64.2
LLaMA-3-70B-Instruct	76.1	75.6	67.4	71.6	72.7
Claude-3-Opus	79.3	74.2	78.8	74.2	76.6
Claude-3.5-Sonnet	78.0	77.3	80.3	77.9	78.4
Gemini-1.5-Pro	77.1	80.0	77.7	74.9	77.4
Gemini-2.5-Flash	77.9	79.2	79.7	79.7	79.1
GPT-3.5-Turbo	70.4	57.3	66.3	58.7	63.2
GPT-4o	79.1	79.8	80.8	78.3	79.5
GPT-o1	**81.1**	**83.3**	**82.2**	**82.4**	**82.2**

Table 3: Document Reasoning Overall Results. We report exact-match (top) and accuracy (bottom) results on the MDBENCH multi-document examples.

both an *exact match* score as well as an *accuracy* score, where the scorer can assign partial correctness credit (0-10 scale) for multi-part questions. Full post-processing and scoring prompts are provided in Appendix C.

5 Results and Analysis

Overall Findings Figure 3 overviews the overall task performance on the multi-document and intermediate tabular versions of the dataset, while Table 3 covers the detailed performance in the target multi-document setting. MDBENCH poses a strong challenge, with the best methods achieving roughly 60% exact-match performance. *GPT-4o leads with 60.0% overall EM* and 62.1% in the zero-shot CoT setting, with GPT-o1 close behind at 59.2% EM. Notably, we see *o1 attains the highest accuracy score across the board in the document setting* (82.2% overall). We generally observe that o1, with its deep reasoning capacities, performs better in the CoT and one-shot settings than in zero-shot. Gemini-1.5 Pro struggles relatively amongst proprietary models, however, the recent Gemini-2.5-Flash improves consistently over it in most settings, peaking at 60.2% EM under one-shot CoT. Overall, we see relatively strong performance from one or more models from each family, with Claude-3.5-Sonnet performing best among the Claude family. The use of CoT prompting yields performance gains in larger models like Gemini-Flash-2.5, GPT-o1 and GPT-4o, although the effect varies. We see open-weight models struggle overall, although LLaMA-3-70B bests GPT-3.5-Turbo, with the 8B variant comparable in performance to GPT-3.5.

Document vs. Tabular Reasoning To ascertain the impact of surface form on the reasoning task, we compare the performance of models on the full multi-document version of the benchmark versus the table version (i.e., stopping after step 2 in our pipeline). Table 4 overviews the table-reasoning results, and the comparison of overall results can be seen in Figure 3. We observe that performance is noticeably higher on the condensed tabular format than in the document setting, with GPT-4o achieving the highest table-reasoning performance at 71.2% overall exact-match, and GPT-o1 similarly performing strongly in the accuracy metric. A key exception to the tabular vs. document trend is with the LLaMA models, which struggle consider-

Model	Zero-shot	Zero-shot CoT	One-shot	One-shot CoT	Overall
LLaMA-3-8B-Instruct	38.4	45.0	19.9	39.8	35.8
LLaMA-3-70B-Instruct	57.3	64.9	30.8	56.4	52.4
Claude-3-Opus	65.4	62.1	65.9	66.4	64.9
Claude-3.5-Sonnet	69.2	67.3	67.3	64.0	66.9
Gemini-1.5-Pro	63.0	67.3	62.1	64.0	64.1
Gemini-2.5-Flash	64.0	65.4	63.5	64.9	64.5
GPT-3.5-Turbo	51.2	49.8	54.5	52.6	52.0
GPT-4o	**71.6**	**69.2**	**73.0**	**71.1**	**71.2**
GPT-o1	66.3	68.7	69.7	66.8	67.9

Model	Zero-shot	Zero-shot CoT	One-shot	One-shot CoT	Overall
LLaMA-3-8B-Instruct	59.0	65.1	28.0	54.1	51.5
LLaMA-3-70B-Instruct	75.3	79.4	39.7	73.5	66.9
Claude-3-Opus	79.8	80.4	80.1	80.3	80.1
Claude-3.5-Sonnet	84.1	83.3	82.8	79.7	82.5
Gemini-2.5-Flash	79.3	78.8	79.4	79.6	79.3
Gemini-1.5-Pro	80.8	81.8	78.3	81.0	80.5
GPT-3.5-Turbo	69.1	67.9	70.0	67.3	68.6
GPT-4o	**85.7**	81.8	**86.9**	83.6	**84.5**
GPT-o1	84.5	**84.9**	83.5	**84.1**	84.2

Table 4: Table Reasoning Overall Results. We report exact-match (top) and accuracy (bottom) when applying models to the augmented *tabular format* QA examples (as opposed to documents).

Ablation	GPT-4o	GPT-3.5
Original Doc. Set	59.7	49.8
- No Delimiter	55.5	45.0
- Shuffled	53.1	39.3
- No Delimiter + Shuff.	50.2	41.7

Table 5: We analyze the significance of document ordering and delimitation within our benchmark (in the zero-shot setting). We see that both shuffling the order of the original docs, and removing "Document $\langle num \rangle$" delimiters impact performance, implying the existence of temporal and cross-document dependencies within our reasoning problems.

ably more on the tabular reasoning task, performing significantly worse than GPT-3.5 in this setting.

Characteristic Breakdown We additionally evaluate the performance as a function of the example difficulty. To do this, we prompt GPT-4o to generate characteristic-level difficulty scores for each example. We use the same five characteristics as demonstrated in the generation process, and prompt the model with these definitions. Rather than generating absolute scores, we instead approximate difficulty by prompting GPT-4o to perform comparative ranking with two other randomly sampled examples for each characteristic. We aggregate these relative rankings over the entire dataset to form two difficulty bins per characteristic, as overviewed in Figure 4.

We see mostly consistent trends across characteristics, with temporal reasoning posing the steepest dropoff between the simple and hard bins. Interestingly, we see soft reasoning is impacted inversely, with performance increasing on the split of examples ranked to have harder soft-reasoning components. While some of this may be due to small sample size for for the hard bin (only 45 of 300 examples), we suspect there is an inverse relationship between soft reasoning and more 'explicit' characteristics such as numeric and temporal. For example, a table/example well-suited for temporal reasoning may naturally contain less 'soft' information requirements. Conversely, an example with significant soft reasoning requirements likely contains fewer hard reasoning requirements.

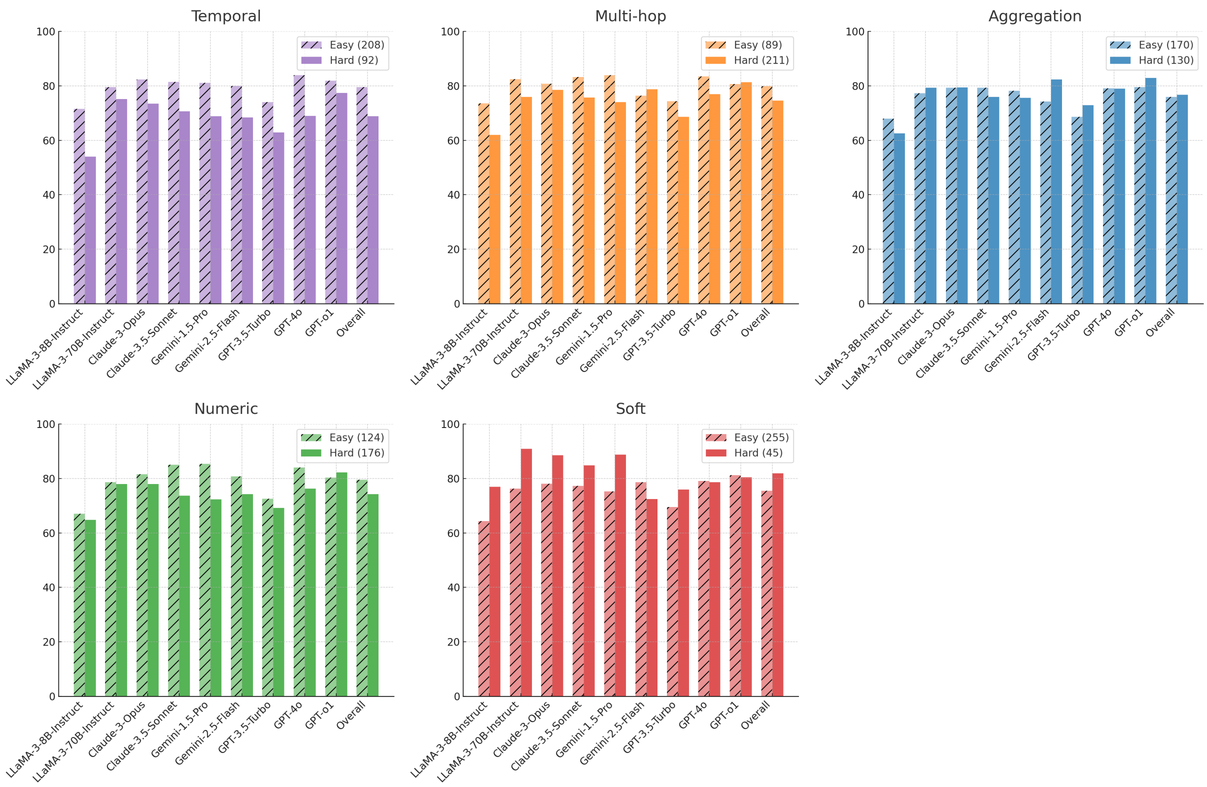

Figure 4: Characteristic-level performance breakdown. We partition the examples by difficulty for each characteristic (as assessed by an LLM judge) and report each model's overall accuracy on each of the bins (simple and hard).

Document Ordering Analysis Finally, we conduct a brief case study with GPT-4o and GPT-3.5 to assess the significance of document ordering. Table 5 shows that both explicit document delimiters and the canonical ordering of the documents are substantive cues for current LLMs. Removing the "Document $\langle num \rangle$" tags lowers GPT-4o accuracy by 4.2 points and GPT-3.5 by 4.8 points, indicating that models leverage these textual anchors to partition the multi-document contexts reliably. Shuffling the document order (while keeping delimiters) produces an even steeper decline (6.6 points for GPT-4o, 10.5 points for GPT-3.5), suggesting that the questions often rely on temporal or positional relationships established by the original document sequence. When both cues are ablated simultaneously, performance drops even further relative to the baseline. Taken together, the results confirm that MDBENCH examples are not solvable by treating each document in isolation. Rather, models must track cross-document dependencies that are sensitive to both ordering and explicit boundary markers.

6 Conclusion

We present MDBENCH, a novel benchmark for evaluating multi-document reasoning in LLMs. Leveraging structured seed knowledge and augmenting it with nuanced reasoning dependencies, MDBENCH enables the systematic creation of challenging QA examples while addressing data contamination and annotation costs. Our results reveal significant challenges for current models, and our approach facilitates scalable, targeted evaluation of multi-document reasoning capabilities, paving the way for more rigorous evaluation of models' abilities to handle real-world, multi-source information.

Acknowledgments

This work is supported by Cisco Research. We thank Greg Durrett, Grace Kim, Ramya Namuduri, Liyan Tang and Anisha Gunjal for their insights and valuable discussions. We also thank the reviewers of ICLR and ARR for their thoughtful comments.

Limitations

While MDBENCH presents a novel approach to evaluating multi-document reasoning in large language models, some directions warrant further exploration: First, we evaluate this method in an automation-focused setting, manually defining just a few foundational reasoning skills and demonstrations as prompts. Future work could explore the trade-offs in adopting more human-centric benchmark processes, such as identifying which pipeline components yield the greatest reduction in human effort while still allowing for dynamic human engagement during generation. Second, while our work has demonstrated the efficacy and potential of knowledge-grounded generation within the Wiki domain—spanning diverse topics such as sports, politics, and science—it is worthwhile to investigate the generalization and adaptability of this method to other, more niche domains, such as law or medicine. These specialized areas may introduce unique challenges due to their inherent complexity. Tabular knowledge may not be readily available or meaningful in such settings, however, it is likely there are other forms of structured knowledge representations that can be consumed in a similar manner. Understanding which prompt and pipeline design choices are best-suited for such targeted domains remains a promising avenue for future research.

References

Anthropic. Meet the claude 3 family. https://www.anthropic.com/news/claude-3-family. Accessed: 2024-09-06.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E Peters, Arie Cattan, and Ido Dagan. 2021. Cdlm: Cross-document language modeling. *arXiv preprint arXiv:2101.00406*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. *Preprint*, arXiv:1909.02164.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva

Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on 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. *Preprint*, arXiv:2306.07934.

Yoonsang Lee, Xi Ye, and Eunsol Choi. 2024. Ambigdocs: Reasoning across documents on different entities under the same name. *Preprint*, arXiv:2404.12447.

Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. 2023a. Sheetcopilot: Bringing software productivity to the next level through large language models. *Preprint*, arXiv:2305.19308.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Gili Lior, Avi Caciularu, Arie Cattan, Shahar Levy, Ori Shapira, and Gabriel Stanovsky. 2024. Seam:

A stochastic benchmark for multi-document tasks. *Preprint*, arXiv:2406.16086.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. *Preprint*, arXiv:2404.07503.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *Preprint*, arXiv:2406.15126.

Weizheng Lu, Jiaming Zhang, Jing Zhang, and Yueguo Chen. 2024. Large language model for table processing: A survey. *Preprint*, arXiv:2402.05121.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. Multi-hop question answering. *Preprint*, arXiv:2204.09140.

Md Mahadi Hasan Nahid and Davood Rafiei. 2024. Tabsqlify: Enhancing reasoning capabilities of llms through table decomposition. *Preprint*, arXiv:2404.10150.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Joseph J. Peper, Wenzhao Qiu, and Lu Wang. 2023. Pelms: Pre-training for effective lowshot multi-document summarization. *Preprint*, arXiv:2311.09836.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. Morehopqa: More than multi-hop reasoning. *Preprint*, arXiv:2406.13397.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2023. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko,

Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof

Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024a. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette,

Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip

Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, Xiang-Hai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi

Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024b. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Xiaojun Wan. 2007. Timedtextrank: adding the temporal dimension to multi-document summarization. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 867–868, New York, NY, USA. Association for Computing Machinery.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *Preprint*, arXiv:2401.04398.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *Preprint*, arXiv:2401.06853.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. *Preprint*, arXiv:2406.04244.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. *Preprint*, arXiv:2402.14116.

A Multi-document Reasoning Skills Demonstrations

Figures 7, 8, 9, 10, 11 overview the five reasoning skills we demonstrate during the creation of MDBENCH. Figure 6 demonstrates an edit plan provided to inspire the table augmentation.

B Model Evaluation Prompts

Simple QA Prompt

"You will be presented with a question and a context. You should answer the question based on the context. The last thing you generate should be ANSWER:[your answer here]"

Chain-of-thought QA Prompt

"You will be presented with a question and a context. You should answer the question based on the context. Explain your reasoning step by step before you answer. The last thing you generate should be ANSWER:[your answer here]"

C Example Post-processing and Scoring Prompts

These prompts are applied after example construction to (a) normalise the ground-truth format, (b) extract concise model predictions after inference, and (c) assign a scalar correctness score to the processed predictions.

QA Example Formatting Prompt

"Adjust the question and answer format such that the question includes output formatting instructions like "output the answer as a JSON with fields <fields>", and ensure the answer follows a corresponding JSON structure.
Question: {question}
Answer: {answer}"

Predicted Answer Extraction Prompt

"Take this response (which may include verbose reasoning) and concisely extract only the final answer to the provided question.
Question: {question}
Answer: {model_response}"

Scoring Prompt

"On a scale of 0–10, how consistent is the predicted answer with the ground-truth? Partial credit is acceptable if a portion of the response matches the ground-truth. RETURN ONLY AN INTEGER.
Question: {question}
Predicted Answer: {predicted_answer}
Ground-truth Answer: {ground_truth}"

D MDBENCH Pipeline Validity Prompts

We use the following prompts during the knowledge augmentation step to validate the edit plan execution and resultant QA example. Prompt 1 works through the generated problem (leveraging the full knowledge augmentation history) and attempts to rationalize the QA example. Then, prompt 2 evaluates whether this rationalization from Prompt 1 is valid and generates a 0-5 validity scalar, discarding values below 5.

Validity Prompt 1

Original Table Name: {table_title}
Original Table: {original_table}
Table Edits Applied: {edits_applied}
Resultant Table: {generated_table}
Resultant Question: {generated_question}
Resultant Answer: {generated_answer}

Prompt: I have provided an original table, and then an updated version (using the provided knowledge edits) which resulted in an augmented table with a corresponding new question and answer. Use this context and think step by step to come up with a solution rationale that provides a justification for the answer. Note that the original table + edits are provided mostly for added reference. Output the rationale as a string.

Validity Prompt 2

How consistent/valid is this reasoning in the following process for generating an example from a table? Score the validity and consistency of the resultant table+question+answer on a scale of 0-5. I want to be able to identify and ignore examples with low scores that I shouldn't include in my dataset. Output as a json with 'score' and 'explanation' fields. Here is the example: {prompt_1_output}

E Characteristic Breakdown

Table 6 overviews the overall model performance when binning examples by difficulty for each of the five considered characteristics.

Model	Aggregation E	Aggregation H	Multi-hop E	Multi-hop H	Numeric E	Numeric H	Soft E	Soft H	Temporal E	Temporal H
Support	170	130	89	211	124	176	255	45	208	92
LLaMA-3-8B-Instruct	68.0	62.6	73.6	62.0	67.1	64.8	64.4	77.0	71.5	54.0
LLaMA-3-70B-Instruct	77.3	79.3	82.5	75.9	78.6	78.0	76.3	90.9	79.5	75.2
Claude-3-Opus	79.3	79.4	80.8	78.5	81.6	78.0	78.1	88.5	82.4	73.5
Claude-3.5-Sonnet	79.3	76.0	83.3	75.7	85.1	73.7	77.3	84.8	81.5	70.6
Gemini-2.5-Flash	74.2	82.4	76.4	78.8	80.8	74.2	78.7	72.5	80.0	68.4
Gemini-1.5-Pro	78.2	75.6	83.9	74.0	85.4	72.3	75.3	88.8	81.1	68.9
GPT-3.5-Turbo	68.6	72.9	74.4	68.6	72.6	69.2	69.5	75.9	74.0	62.9
GPT-4o	79.1	79.0	83.5	77.0	84.0	76.3	79.1	78.7	83.9	69.0
GPT-o1	79.6	82.9	80.7	81.3	80.3	82.2	81.2	80.4	81.9	77.4
Overall	76.0	76.7	79.9	74.6	79.5	74.3	75.5	81.9	79.5	68.9

Table 6: Characteristic-level Performance Breakdown. We report overall accuracy.

Multi-document Dataset Annotation Guidelines

Your task is to evaluate the consistency of generated multi-document question answering examples.

You will be provided with the following for each example:

- A set of related documents (labeled Document 1 - Document N)
    - These contain the source information needed for the problem
- A question
    - This question will rely reasoning over the provided documents.
- An answer
    - You will solve the problem yourself, and compare your answer with the provided answer
- An example UUID
    - You will use this to refer to the example when recording your progress in the tracker spreadsheet

You will record your progress in the tracking spreadsheet. **Please contact <redacted> if you run into any issues or have questions.**

For each example, follow these steps:

1. **Read the Question:**
   Carefully read the question to understand what is being asked.
2. **Examine the Context:**
   Read through the entire provided document set. Understand the scenario, relevant details, and any nuances in the text. This may take a couple passes, and it may be helpful to highlight sections of text as you go, or feel free to take physical notes if this is helpful.
3. **Answer the Question:**
   Based solely on the context provided, determine what you believe the correct answer should be.
4. **Compare Answers:**
   Compare your answer with the provided ground-truth answer.
    - If your answer aligns with the provided answer, mark the example as **VALID** in the tracking sheet.
    - If your answer does not match, or if you find discrepancies indicating an inconsistency or error in the example, mark the example as **INVALID** in the sheet.
5. **Provide Comments:**
   If there are any unique observations or trends, please note them in the tracking sheet

Figure 5: MDBench human validation guidelines.

Original Table | Table Summary: Movie Sales by Country

date	territory	screens	rank	gross ($)
october 20 , 2006	turkey	378	1	146268
october 25 , 2006	belgium	6	19	38916
october 25 , 2006	germany	52	12	133228
october 26 , 2006	austria	4	13	41780
october 26 , 2006	netherlands	17	14	53749
october 27 , 2006	united kingdom	4	24	34704

Edit 1: Come up with a interesting question about this table. The question MUST have a concise verifiable answer. The question should go hand in hand with ensuring the augmentation introduces complex cross-row dependencies, as this will be used to create corresponding multi-document examples (one document per row). Make sure that the question + new table can only be answered if the model reasons correctly over documents.
Example: "Rank the movie's sales by country." -- requires reasoning/comparing over the different rows in the document. Note: we will edit the table further to make this even more challenging.

Edit 2: Remove extraneous columns to avoid overspecification in the resultant documents
Example: Remove the screens and rank columns since they're not relevant

Edit 3: Round some of the numeric values to eventually make the information more realistic in the articles
Example: Round the gross sales numbers to thousands

Edit 4: Add 'multi-hop' information, or additional rows that necessitate synthesizing information across documents
Example: Add an October 26th entry for Germany for $195k (now there are two rows for Germany) -- These need to be added into order to calculate the Germany sales.

Edit 5: Add secondary / peripheral / fictional information to contextualize/personalize the documents.
Example: Add a "context" column with some additional guidance to guide the document generation. This should include instructions on document length + writing style as well as superfluous content that might naturally occur in a document of this type. Also add a fictionalized film name (Nightmares of Glory).

Edit 6: Introduce cross-document dependencies by obfuscating some linked information. The dependencies must be utilized within the question answering process.
Example: The Germany Oct. 26th entry was modified. The country information was obfuscated, but the daily revenue was defined in terms of the prior day, allowing the model to refer back to the Oct. 25 row.

Resultant Augmented Table

date	country	daily revenue ($)	film	context
october 20 , 2006	turkey	146200	Nightmares of Glory	short article about total movie sales
october 25 , 2006	belgium	39000	Nightmares of Glory	article about total movie sales
october 25 , 2006	germany	135000	Nightmares of Glory	mid-length article about daily movie sales
october 26 , 2006	austria	42000	Nightmares of Glory	article about total movie sales
october 26 , 2006	netherlands	54000	Nightmares of Glory	report of national movie sales
october 27 , 2006	united kingdom	34700	Nightmares of Glory	article about total movie sales, and interviewing a fictional moviegoer
october 26 , 2006	[not explicitly stated]	195,000, 60,000 more than yesterday's sales	Nightmares of Glory	article about total movie sales

Augmented Table Question: Rank the movie's sales by country.
Augmented Table Answer: Germany (133000+195000), Turkey (146200), Netherlands (54000), Austria (42000), Belgium (39000), United Kingdom (34700)

Figure 6: Demonstration of table edit plan used during the knowledge augmentation component of the MDBENCH pipeline.

[Knowledge Aggregation] – The ability to align, compare and/or contrast knowledge that may be present. This includes non-numeric knowledge.

Baseline Example: Rank the teams by number of wins in the series.

race	pole position	winning team
May 7, 1992	nico valencia	ferrari
May 21, 1992	mark steedman	bmw
June 4, 1992	bonnie bobcat	mclaren
June 18, 1992	elio muchin	renault
July 2, 1992	tammy tiger	ford
July 16, 1992	tyrell eshar	ferrari
July 30, 1992	alain prost	ferrari
August 13, 1992	tigre trees	renault

Answer: Ferrari, Renault, and T-3 are BMW, McLaren and Ford.
Answer Rationale: Ferrari was listed as the winning team three times, Renault twice, and the others once each.
Commentary: This is a simple example that required calculating the number of appearances of each team in the 'winning team' column.

Harder Example: Identify the top two teams in this race series, and explain any correlation between their success and the weather.

race	pole position	winning team	notable conditions
May 7, 1992	nico valencia	ferrari	sunny + dry
May 21, 1992	mark steedman	bmw	rainy
June 4, 1992	bonnie bobcat	mclaren	heavy rain
June 18, 1992	elio muchin	renault	slick roads
July 2, 1992	tammy tiger	ford	cold and blustery
July 16, 1992	tyrell eshar	ferrari	sunny
July 30, 1992	alain prost	ferrari	overcast
August 13, 1992	tigre trees	renault	damp

Answer: Ferrari finished first and Renault finished second. Ferrari's wins were exclusively in conditions with dry pavement, whereas Renault won only in wet conditions.
Answer Rationale: Ferrari had three wins, and Renault had two wins. The rest of the teams had only one. Notably, Ferrari winning races were only in conditions where the roads were presumably dry (sunny+dry, sunny, and overcast), and Renault's wins were only on day where the conditions were wet (slick roads, and damp).
Commentary: This answer requires not only understanding the winning teams, but also realizing that there were patterns in the conditions for both teams. Namely, one had to ascertain that Ferrari performed well on dry days, whereas Renault did well on wet roads. This requires aggregating, comparing, and contrasting values across different rows and teams.

Figure 7: Knowledge Aggregation Skill Description

[Multi-hop Reasoning] – The ability to solve problems requiring multiple steps to arrive at the solution.

Baseline Example: Which country had the most showings and how many was this in total?

date	territory	showings
october 20, 2006	turkey	200
october 20, 2006	belgium	600

Answer: Belgium had the most with 600 showings.
Answer rationale: Turkey had 200 showings and Belgium had 600. 600 > 200, therefore Turkey had the most showings.
Commentary: This is a simple reasoning process as it requires a simple comparison of two values with no additional reasoning required.

Harder Example: Which country had the most showings and how many was this in total?

date	territory	showings
october 20, 2006	turkey	200
october 20, 2006	belgium	600
october 25, 2006	turkey	500

Answer: Turkey had the most with 700 showings.
Answer Rationale: Turkey had showings on two different days, so the total is 200+500=700 showings. 700 > Belgium's 600, therefore Turkey had the most.
Commentary: By adding a new row with complementary information, we necessitate an additional reasoning hop to correctly answer the question. Note that this table was edited specifically such that the answer (Turkey) is flipped from the original answer (Belgium) in the simple example. Edits like these ensure the reasoning cannot be shortcutted (e.g., by simply selecting the row with the highest showings).

Figure 8: Multi-hop Reasoning Skill Description

[Numeric Reasoning] – The ability to handle numeric values and perform numerical operations

Baseline Example: Rank each day by the total showings.

date	territory	showings
october 20, 2006	turkey	200
november 21, 2006	belgium	600
november 21, 2006	turkey	400
november 22, 2006	belgium	600

Answer: November 21st had the most showings with 1000, followed by November 22nd, then October 20th.
Answer Rationale: November 21st had 1000 totals showings – 600 in Belgium and 400 in Turkey. This was greater than the 600 on November 22nd and the 200 on October 20th.
Commentary: This is a simple case of performing numeric operations, having to sum values over different rows to identify the correct answer.

Harder Example: Rank each day by the total sales

date	territory	showings	Avg. sales per showing ($)
october 20, 2006	turkey	200	6000
november 21, 2006	belgium	600	1000
november 21, 2006	turkey	400	1000
november 22, 2006	belgium	600	500

Answer: October 20th had the highest sales, followed by November 21st, then November 22nd
Answer Rationale: October 20 had 200 showings * $6000 per showing = $1,200,000. November had 600*$1000 = $600,000 from Belgium and 400*$1000 = $400,000 from Turkey, totalling $1,000,000. November 22 had 600 * $500 = $300,000 in sales.
Commentary: This reasoning requires calculating values over two different columns, and then additionally summing values over associated rows (e.g. the november 21 entries).

Figure 9: Numeric Reasoning Skill Description

[Soft Reasoning] – The ability to reason abductively and make informed decision in cases where some uncertainty or fuzziness may be present.

Simple Example: Who had the most championships?

Year	Championship Winner
2008	Yusef
2009	Mattingly
2010	Tigre Trees
2011	Yusef “Skeeps” Mattingly
2012	Tigre
2013	John Smith
2014	John Smith
2015	Harrison Chevrolet

Answer: Yusef Mattingly, who had wins in 2008, 2009, and 2011
Answer Rationale: Although not clearly stated, some of the entries likely refer to the same person, just sometimes using only the first name, last name, or a nickname. We can reasonably assume ‘Yusef’, ‘Mattingly’, and ‘Yusef “Skeeps” Mattingly’ all refer to the same individual. Similarly, we see both a ‘Tigre Trees’ and ‘Tigre’ which likely refer to the same individual.
Commentary: This is an example abductive or ‘best guess’ soft reasoning where one could reasonably assume that some of the entries refere to the same canonical entity/person. Notably, this example is one where a wrong answer would be generated by using a simple exact match heuristic as ‘John Smith’ appears twice, which is less than Yusef Mattingly.

Harder Example: Rank the countries by total sales.

	Country	Sales ($)	Notes
October 20	Turkey	146200	
October 25	Belgium	39000	
October 25	Germany	134000	
October 26	Austria	42000	
October 26	Netherlands	54000	
October 27	United Kingdom	534700	
October 26	<one that was already mentioned>	195000, roughly 60k more than yesterday's sales.	A follow-up to a prior entry

Answer: United Kingdom, Germany, Turkey, Netherlands, Austria, Belgium
Answer Rationale: Most country sales are confined to just one row. However, the final row contains sales information that implicitly refers to a country. We see that this country is already mentioned and that this row is a follow-up to a previous entry with sales numbers. The sales value is $195,000 which is stated as 60k more than the prior day sales. We can use this to ascertain what the country is. Namely, we see that there are two entries for the prior day (October 25). Of these two, Germany’s sales were $134,000 which is approximately $60,000 less than $195,000. Belgium’s sales were much lower (over $150k less than $195,000). Therefore, we can reasonably conclude that the October 26 entry in mention refers to Germany. Combining the $134,000 from October 25 and $195,000 from October 26, we see Germany’s total sales are $329,000, which is less than the United Kingdom, but more than Turkey.
Commentary: This problem requires that one notices that the final row can be linked to a prior row. Once this is done, there is some soft reasoning that clearly leads to the proper solution. So, while there is some abduction reasoning required, it is very clear once you put the pieces together.

Figure 10: Soft Reasoning Skill Description

[Temporal Reasoning] – The ability to handle temporal information and dependencies.

Baseline Example: How many total showings were there in each month?

date	territory	showings
october 20, 2006	turkey	200
november 21, 2006	belgium	600
november 21, 2006	turkey	400
november 22, 2006	belgium	600

Answer: October 2006 had 200 showings, while November had 1,600
Answer Rationale: October had just one day with 200 showings. November had 3 showings total, summing to 600+400+600 showings total.
Commentary: This is fairly straightforward as we simply sum all rows sharing the same month.

Harder Example: How many total showings were there in each month?

date	territory	showings	notes
october 20, 2006	turkey	200	Opening day in Turkey
november 21, 2006	belgium	600	Opening day in Belgium
the week after opening day	turkey	400	
november 23, 2006	belgium	600	

Answer: October 2006 had 600 showings, while November had 1,200
Answer Rationale: In Turkey, the week after opening day fell in the month of October, therefore there were 200 (from opening day) + 400 (from the week after) = 600 showings in October. November had 600+600 = 1,200 showings, all from Belgium.
Commentary: We introduce a cross-row dependency here that requires temporal reasoning to solve. Namely, we need to intuit that, given opening day is on October 20th, the week immediately following it must fall within the month of October. Again, we intentionally edit the values in the table (and add a 'notes' column) to ensure that the answer (600 in October, 1200 in November) necessarily required resolving this cross-row dependency.

Figure 11: Temporal Reasoning Skill Description

Question: Each article refers to a different game in the running series. Which team won the series and which games did they win?
Answer: Buffalo won the series, winning games 1, 3, 6, 7. Ottawa won games 2, 4, an 5.

Document 1:

The chill of the arena was a stark contrast to the warmth that had begun to grace the city outside. Inside, the tension was palpable as the crowd's anticipation hung heavily in the air. The local team had been on a losing streak, but tonight there was a spark in their eyes that hadn't been seen in weeks. The arena was a sea of blue and gold, with a smattering of red among the stands, a small but loyal contingent that had made the trip to support their visiting team.Emma and her brother, Michael, had been looking forward to this game all season. They had grown up in a household divided between loyalty to Buffalo and affection for Ottawa, a dynamic that invigorated their sibling rivalry. As the clock ticked down the final minutes, Michael's grip tightened on the edge of his seat while Emma's cheers grew more fervent. The score had been stagnant since the second period when the home team had managed a slick goal that sent the crowd into a frenzy, a play that was now looking like it might secure the game. As the final buzzer sounded, the arena erupted, the home fans' relief and joy mingling in a deafening roar. Buffalo had managed to hold onto their slender lead, securing a much-needed victory. Emma's face was lit with a triumphant smile as she turned to her brother, whose disappointment was evident, though he managed a good-natured shrug. They filed out with the rest of the crowd, the crisp night air doing little to cool the elation of the home fans or the resigned contemplation of the visitors. For Emma, the walk home was going to be sweet, her team's victory a happy echo in her step, while Michael was already looking ahead, optimistic that the next game might tell a different story.

Document 2:

The city of Buffalo was alive with the buzz of anticipation as fans draped in blue and gold swarmed the downtown area, their spirits undampened by the brisk spring breeze. The local cafes and pubs were abuzz with predictions and friendly banter, while the streets echoed with the sounds of camaraderie. It was, after all, a night marked by the ritual of hockey, a sacred time when the community came together, united by their love for the Sabres. In one particular home, nestled in a cozy neighborhood where the scent of lilacs began to mingle with the crisp air, the Larson family gathered in their living room, adorned with memorabilia of games past. The youngest, Lily, wore her oversized jersey with pride, the number of her favorite player still crisp against the fabric. Her father, Michael, a man with a kind face and a voice worn from years cheering from the stands, was setting up the snacks on the coffee table with a quiet intensity. They had skipped the live game tonight, a rare occurrence, but Michael's work had demanded his presence later than usual. The game was to be a bonding moment, the television their window to the thrill. As the match commenced, the Larsons were perched at the edge of their seats, reacting to each play with a symphony of gasps, cheers, and the occasional groan. The evening waned and the tension within the four walls grew thick, mirroring the energy of the arena miles away. There was a magic to watching the game unfold from the comfort of home, a different kind of ritual observed by families like the Larsons, who lived and breathed by the puck's dance across the ice. When the final buzzer sounded, it was met with a mixture of emotions; the disappointment of a home defeat was palpable, yet there was a sense of pride in the hard-fought battle. They spoke of plays and strategy, of what might have been, Michael's arm draped over Lily's shoulder in a silent promise of next time. The game may have slipped away, but the spirit of it lingered in the hearts of the Buffalo faithful.

Document 3:

As the sun dipped below the horizon, painting the sky in hues of orange and purple, the city of Ottawa buzzed with an electric anticipation. It was more than just the changing seasons that had the locals in a state of excitement; it was the clash on the ice that had everyone talking. The arena was packed, the air thick with the scent of popcorn and the resonant echoes of cheering fans. In the heart of it all, Jeremy and his father, clad in their beloved team's colors, watched with bated breath as the players darted across the rink. It was a dance of power and precision, a tug of war for dominance that had them on the edge of their seats. The visitors, despite being on foreign ice, played with a relentless drive, their determination as palpable as the chill that hung in the air. The score was close, and each goal was met with a roar that reverberated off the walls, a testament to the passion that filled the space. The final minutes were a blur of motion and emotion. Jeremy's heart raced as he gripped his father's hand, both of them silently praying for a miracle. But as the clock dwindled down, it was clear that the visitors had edged out their opponents. The final buzzer sounded, a solemn toll that marked the end of the battle. Disappointment was heavy in the Ottawa crowd as they streamed out of the arena, a stark contrast to the jubilation of the few visiting fans who celebrated their hard-fought victory. On the ride home, Jeremy's father reminded him that this was the nature of the game–sometimes you win, sometimes you learn. And tonight, they had learned.

Figure 12: MDBENCH Document QA Example – Part 1

Document 4:

The city was abuzz with the excitement of the game as the evening approached. Downtown streets hummed with the chatter of fans clad in their team's colors, a sea of red and black flowing towards the arena. Inside the local pubs, patrons argued over potential plays and key players, their voices rising over the clinking of glasses and the low hum of pre-game broadcasts. Among them, Sarah and Jeff found a cozy corner booth, their eyes locked in a friendly but fierce debate about the night's outcomes. The arena itself was a cacophony of anticipation, with every seat filled and the air electric. The first period passed in a blur of motion, the crowd's roar ebbing and flowing with each close call and power play. Despite the visitors' valiant effort, their offense just couldn't break through the impenetrable defense they faced. Sarah, a hometown girl born and raised with a love for her city's team, clutched Jeff's hand tightly as the final minutes ticked down. Jeff, who had moved to the city years ago, had adopted her team as his own, their shared passion for the sport one of the many threads that had woven their lives together. As the final buzzer sounded, the arena erupted, the scoreboard confirming the hard-fought victory. Sarah leaped to her feet, her cheers mingling with the thousands around her, as she and Jeff celebrated the triumph. The players on the ice embraced, their faces a mixture of exhaustion and elation. It was a moment of pure joy for the home team, a narrow win that would be recounted in the history of the season. Outside, the night air was crisp, but the warmth of victory and the shared experience seemed to make the atmosphere glow. They walked home, their path lit by the city lights and the lingering excitement of a game that would be etched into the collective memory of the fans, a one-goal difference that had meant everything.

Document 5:

The early spring evening held a crisp chill that settled over the city of Buffalo, a reminder that winter's grip had only recently loosened. In the heart of downtown, a palpable excitement coursed through the crowd as they streamed towards the arena, their breaths visible in the twilight. The local fans, clad in blue and gold, were a sea of hopeful energy, ready to cheer on their beloved team. Inside the arena, the tension was almost tangible. The players on the home bench tapped their sticks anxiously, exchanging glances that were a mix of determination and nerves. The visiting team from Ottawa, undeterred by the raucous environment, seemed poised and focused during the warm-up, their red jerseys stark against the ice. The game commenced with the drop of the puck, and the home crowd's roar reached a fever pitch, echoing off the rafters. As the final buzzer sounded, a hush fell over the Buffalo fans, their earlier enthusiasm replaced by a respectful silence. The scoreboard, an unflinching digital adjudicator, displayed the outcome that had unfolded over the past few hours. The visitors had demonstrated a formidable defense and an offense that capitalized on every opportunity. It was clear to everyone leaving the arena that evening that the team from Ottawa had earned their win on foreign ice, besting their hosts with a strategy that was as effective as it was elegant. Later that night, as the Ottawa players boarded their bus, a sense of camaraderie enveloped them. This victory was a testament to their hard work and a crucial addition to their season's record. Meanwhile, the Buffalo team, though disheartened, knew that the season was far from over. There would be other games, other chances to prove themselves, and the unwavering support of their fans would be there every step of the way.

Document 6:

The air was thick with anticipation as the evening drew in over the capital city. The local arena, a coliseum of modern spectacle, was abuzz as fans donned in their team's colors made their pilgrimage to witness the clash on ice. Vendors hawked memorabilia, the scent of stadium food filled the air, and the collective heartbeat of the city seemed to pulse in time with the approaching faceoff. In the stands, there was a sea of red and black, punctuated by the occasional bold stripe of blue and gold. Families, friends, and die-hard enthusiasts shared in the camaraderie that only such a game could inspire. The home team had been struggling this season, but hope springs eternal, and there was a sense that tonight could mark a turning point. After all, it was a perfect evening for an underdog story. On the bench, the visiting team was a study in focus. Their strategy was clear, their execution clinical, and as the puck dropped, it became evident that they were not here to play into a narrative of redemption. The home team fought valiantly, their goalie making save after save, but as the final buzzer sounded, it was the silence of the home crowd that spoke volumes. The scoreboard, unyielding, displayed the tale of the tape, a testament to the visitor's triumph in the city of Ottawa that night.

Document 7:

As the evening sun dipped below the horizon, painting the sky in hues of lavender and orange, the city of Buffalo was alive with a palpable energy. It was a day marked in every local's calendar, a moment when hometown pride swelled in the hearts of the residents. The streets around the arena buzzed with anticipation, vendors calling out over the chatter, selling memorabilia and hot dogs to the sea of blue and gold. Inside the arena, you could cut the tension with a knife. The game was a nail-biter, with the crowd erupting into cheers and groans as the puck darted across the ice, a blur of potential victory or defeat. The players were relentless, sweat glistening on their brows as they fought for every inch, every opportunity to tip the scales. As the final buzzer sounded, the roar from the stands was deafening, a wave of elation that swept through every aisle and seat. The home team had snatched triumph from the jaws of a well-matched adversary, a testament to their resolve and spirit, much to the chagrin of the visitors who had hoped to silence the boisterous Buffalo fans. Later, as the night claimed the sky and the arena emptied, the streets of Buffalo remained abuzz. Families and friends recounted the game's most thrilling moments, reliving each goal with animated gestures and wide smiles. The city's iconic eateries filled with patrons, their conversations a mix of play-by-play analyses and plans for the next game. It was a victory that would be talked about for days to come, another storied chapter in the city's rich sports history, a testament to the heart and hustle that defined Buffalo not just on the ice, but in every aspect of its character."

Figure 13: MDBENCH Document QA Example – Part 2