Are We on the Right Way to Assess Document Retrieval-Augmented Generation?

Wenxuan Shen[1]*, Mingjia Wang[3]*, Yaochen Wang[3], Dongping Chen[3], Junjie Yang[1], Yao Wan[3], Weiwei Lin[1,2]†

[1]South China University of Technology,
[2]Pengcheng Laboratory, Shenzhen 518066, China,
[3]Huazhong University of Science and Technology,
linww@scut.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) systems using Multimodal Large Language Models (MLLMs) show great promise for complex document understanding, yet their development is critically hampered by inadequate evaluation. Current benchmarks often focus on specific part of document RAG system and use synthetic data with incomplete ground truth and evidence labels, therefore failing to reflect real-world bottlenecks and challenges. To overcome these limitations, we introduce DOUBLE-BENCH: a new large-scale, multilingual, and multimodal evaluation system that is able to produce fine-grained assessment to each component within document RAG systems. It comprises 3,276 documents (72,880 pages) and 5,168 single- and multi-hop queries across 6 languages and 4 document types with streamlined dynamic update support for potential data contamination issues. Queries are grounded in exhaustively scanned evidence pages and verified by human experts to ensure maximum quality and completeness. Our comprehensive experiments across 9 *state-of-the-art* embedding models, 4 MLLMs and 4 *end-to-end* document RAG frameworks demonstrate the gap between text and visual embedding models is narrowing, highlighting the need in building stronger document retrieval models. Our findings also reveal the over-confidence dilemma within current document RAG frameworks that tend to provide answer even without evidence support. We hope our fully open-source DOUBLE-BENCH provide a rigorous foundation for future research in advanced document RAG systems. Our benchmark and code are open source at https://double-bench.github.io/.

Introduction

Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) has emerged as a transformative technique in textual information retrieval, enhancing Large Language Models (LLMs) by retrieving the most relevant content from knowledge bases in response to queries. This approach has driven significant advances in context engineering, particularly for knowledge-intensive NLP tasks. Besides text-only scenarios, vision documents—including scanned files (Breci, Guarnera, and Battiato 2024), charts (Masry et al. 2022), and slides (Tanaka et al. 2023)—serve as rich information sources that have traditionally required substantial manual effort to examine. These documents are now being efficiently enhanced and handled through multimodal document RAG systems, enabling advanced document understanding (Faysse et al. 2024; Wang et al. 2025; Cho et al. 2024).

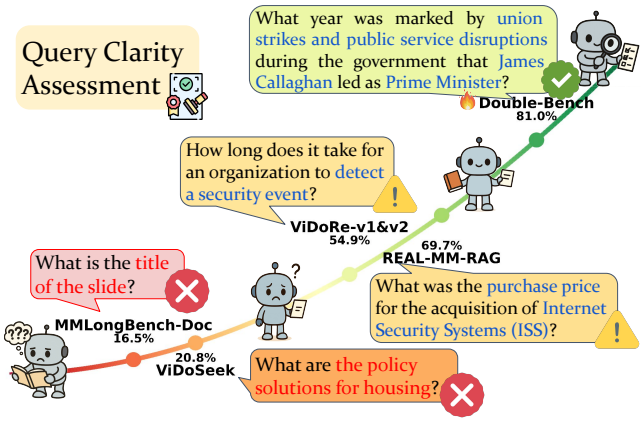

Figure 1: Existing document RAG benchmarks suffer from ambiguous queries that are insufficiently specified to retrieve relevant results, failing to authentically evaluate current document retrieval models and systems.

Despite the growing importance of document RAG systems (Mortaheb et al. 2025a; Yu et al. 2024; Mortaheb et al. 2025b), effectively evaluating them in detail presents significant challenges. Existing document RAG evaluation benchmarks (Friel, Belyi, and Sanyal 2024) suffer from four critical shortcomings as we identified through pilot experiments shown in Figure 1: **(1) Limited evaluation scope**: Current benchmarks only focus on specific parts such as embedding model or VQA model, failing to reveal the bottlenecks of current RAG system in a holistic and comprehensive way. **(2) Unrealistic prior knowledge assumptions:** Many VQA-style benchmarks (Li et al. 2024; Wu et al. 2025) presuppose that the target page or document is already known, making queries inappropriate for evaluating real-world global retrieval scenarios. **(3) Ambiguous or non-unique evidence:** Synthetic queries are often crafted from a single page and assume a one-to-one mapping between query and evidence, neglecting cases where multiple pages could be relevant. **(4)**

*Equal contribution.
†Corresponding author.

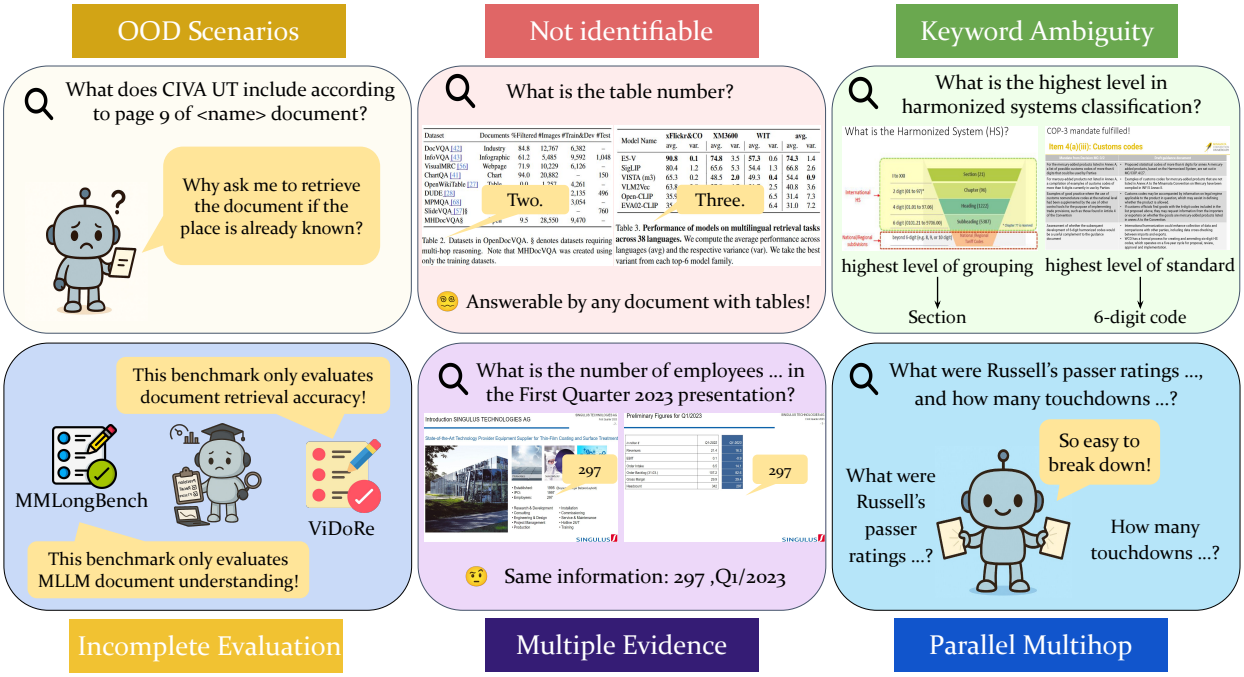

Figure 2: Our pilot study reveals critical limitations in existing document RAG system benchmarks, which make them unable to reliably assess current system in a fine-grained and realistic manner.

Unlinked multi-hop compositions: Sythesized multi-hop queries are frequently constructed from loosely connected single-hops, failing to evaluate retrieval models' ability on multi-hop reasoning across documents and modalities.

To overcome these limitations, we introduce DOUBLE-BENCH (**DO**cument **U**nified **B**road-coverage **L**ogical hops **E**valuation Benchmark), consisting of 5,168 human-validated single- and multi-hop queries and 3,276 documents in 6 languages and 4 types of document data. First, we assemble and preprocess a large diverse document corpus spanning PDFs, scanned documents, slides and web pages by a two-stage filtering and modality decomposition. Next, we synthesize and rigorously label both single- and multi-hop queries (Zhang, Feng, and Zhang 2025; Tang and Yang 2024), using a combination of iterative clarity-oriented refinement, knowledge-graph-guided composition and exhaustive evidence search. Finally, expert annotators review (Chiang et al. 2024; Chen et al. 2024a) and correct machine-assigned evidence to ensure high-precision ground truth for large-scale realistic multimodal retrieval evaluation. As shown in Figure 2, DOUBLE-BENCH contains high quality queries with low ambiguity. To avoid potential data contamination issues, DOUBLE-BENCH is also designed to support streamlined dynamic updates with minimal human intervention.

Based on DOUBLE-BENCH, we conduct extensive experiments across 9 *state-of-the-art* textual, visual and multimodal embedding models, 5 MLLMs and 4 advanced document RAG frameworks. Our findings demonstrate that text embedding models are narrowing the gap with visual embedding models, and `Colqwen2.5-3b` achieve a strong performance with an 0.795 averaged hit@5 score. Regarding different languages, retrieval models perform generally better on high-resource language than low-resource like Arabic and French. Regarding different document types, clean and structured documents, such as PDFs and HTML pages, are generally easier for models to inspect. Moreover, MLLMs' low accuracy across both single- and multi-hop queries demonstrates the inherent challenges in multimodal long document understanding. Multi-hop queries prove particularly challenging for current document RAG frameworks, achieving only 0.655 accuracy even when ground truth pages are directly provided.

In summary, our contributions are three-fold:

- We diagnose several major limitations in existing document RAG evaluation, including incomplete scope, unrealistic prior knowledge assumptions, ambiguous or non-unique evidence, and non-grounded multi-hop query design.
- We introduce DOUBLE-BENCH, the first-of-its-kind live evaluation system for multilingual and multimodal document RAG system, featuring a diverse document corpus, fine-grained page decomposition, and high-quality single- and multi-hop QA pairs with manually labeled evidence.
- Our experiments across 9 *state-of-the-art* embedding models, 4 MLLMs, and 4 document RAG system uncover critical limitations in current RAG frameworks, providing insights and findings for the research community.

Limitations of Existing Document RAG System Evaluation: A Pilot Study

Task Formulation

Let C be a large corpora consisting of documents $\{d_1, d_2, ..., d_n\}$. Each document d_i is stored by page images

Benchmarks	Size			Queries			Labels		Evaluation Target			Document		
	Doc	Avg. #Pages	Query	Clarity	i.i.d.	M.H.	GT	M.H. Chain	Embed Model	MLLMs	RAG System	Lang.	Dyna.	Type
DocVQA	6,071	1.0	50,000	✗	✗	✗	✗	-	✗	✓	✗	1	✗	
MMLongbench-Doc	135	47.5	1,082	✗	✓	✓	✗	✗	✗	✓	✗	1	✗	
MMDocIR	6,818	65.1	73,843	✗	✗	✓	✗	✗	✓	✗	✗	1	✗	
UDA-QA	2,965	46.3	29,590	✗	✓	✓	✗	✗	✓	✗	✗	1	✗	
ViDoRe v1	5,000	1.0	500	✓	✓	✗	✓	-	✓	✗	✗	2	✗	
ViDoRe v2	65	48.6	913	✓	✓	✗	✓	-	✓	✗	✗	2	✗	
ViDoSeek	1,142	18.4	1,142	✓	✗	✓	✗	✗	✓	✗	✓	1	✗	
REAL-MM-RAG	163	49.1	4,553	✓	✓	✗	✓	-	✓	✗	✗	1	✗	
DOUBLE-BENCH	3,276	22.3	5,168	✓	✓	✓	✓	✓	✓	✓	✓	**6**	✓	

Table 1: Comparison between existing multimodal document related benchmarks and the proposed DOUBLE-BENCH, where each symbol represents: PDFs; Scanned documents; Slides; HTML pages. Half-tick denotes dependent on the specific benchmark component or insufficient evaluation. ***GT***: Ground Truth evidence labels. ***M.H.***: Multi-Hop. ***Lang.***: Supported number of languages. ***Dyna.***: Opensource automated pipeline for dynamic benchmark updates.

$\{p_1^i, p_2^i, ..., p_m^i\}$. Given a query Q, the objective is to retrieve top-k possible evidence pages E_r from the entire corpora to formulate the answer A. For single-hop queries Q_s, answer A can be found if one or more evidence pages from the evidence set E_q is successfully retrieved. For multi-hop queries Q_m, the requirement extends to having one or more evidence page for every evidence set $E_{q,j}$ of each query hop j.

Four Major Overlooked Issues for Document RAG Evaluation

Though practical document RAG scenarios typically involve queries that clearly state an informational need, the core utility of RAG is demonstrated when users do not possess specific prior knowledge about individual documents, such as their titles, filenames, or the precise location of some document component, even if users have general familiarity with the topics within the collection.

Therefore, we start by investigating whether existing benchmarks are fully appropriate for evaluating real-world document RAG scenarios. By screening existing benchmarks with concrete rules, we illustrate these limitations in Figure 2.

Current benchmarks fail to comprehensively evaluate document RAG systems with fine-grained breakdown. As shown in Table 1, current document-related benchmarks usually focus on embedding-based retrieval models (Macé, Loison, and Faysse 2025) or response models (Ma et al. 2024), which are only components within document RAG systems, failing to provide a comprehensive assessment. This fragmented evaluation obscures critical interaction effects between retrieval and generation components that often determine real-world system performance.

Benchmark design issues with prior knowledge assumptions. VQA benchmarks, such as MMLongbench-Doc (Ma et al. 2024), DocVQA (Mathew, Karatzas, and Jawahar 2021), and MMDocIR (Dong et al. 2025), are inherently designed with a given page or document as prior knowledge, rendering their queries ambiguous for identifying the ground-truth page within a global document corpus. Manually inspected benchmarks designated for RAG, such as ViDoSeek (Wang et al. 2025) and MRAMG-Bench (Yu et al. 2025), have significant gains in query information. However, such benchmarks tend to insert the exact name or page of the ground document, failing to align with intended applications where users do not have any specific prior knowledge over individual documents. Such queries create gaps between evaluation and real use.

Queries with multiple interpretations and scattered evidence. Most benchmarks construct queries by selecting a ground truth page beforehand, and assume the evidence used is unique (Chen et al. 2024b; Tang and Yang 2024; Saad-Falcon et al. 2023). This generally holds true when the corpus is small, such as individual benchmarks in ViDoRe (Faysse et al. 2024), but the problem becomes uncontrolled when the corpus scales up. Some queries may also have unexpected multiple interpretations given different content in the same document, which further undermines the unique assumption.

The linearity in multi-hop query synthesis is overlooked. The inclusion of trivial multi-hop queries in evaluations may overstate the reasoning capabilities of RAG frameworks, inflating perceived performance without accurately assessing genuine multi-step reasoning. These queries are essentially simple linkings of independent parts, do not necessitate complex reasoning to deconstruct and can be processed in parallel (Hui, Lu, and Zhang 2024).

DOUBLE-BENCH: The Benchmark

To address existing limitations, we introduce DOUBLE-BENCH, a benchmark with manually verified multi-modal, multi-lingual content and an automatic benchmark construction suite via a three-stage pipeline shown in Figure 3. Detailed benchmark statistics can be found in Figure 4 and Appendix. We also provide extensive metadata, such as queried modality, language, evidence chains/lists and parsed page chunks to advance document RAG research community.

Metadata Collection and Preprocessing

This section details the preprocessing steps applied to the raw document corpus: ***(1)*** Large Corpus Collection, ***(2)*** Two-stage Filtering, and ***(3)*** Modality Split.

Metadata Collection. To ensure comprehensive evaluation, we collect a diverse range of document types and languages. The initial database comprises four popular types of documents collected from various sources. As shown in Figure

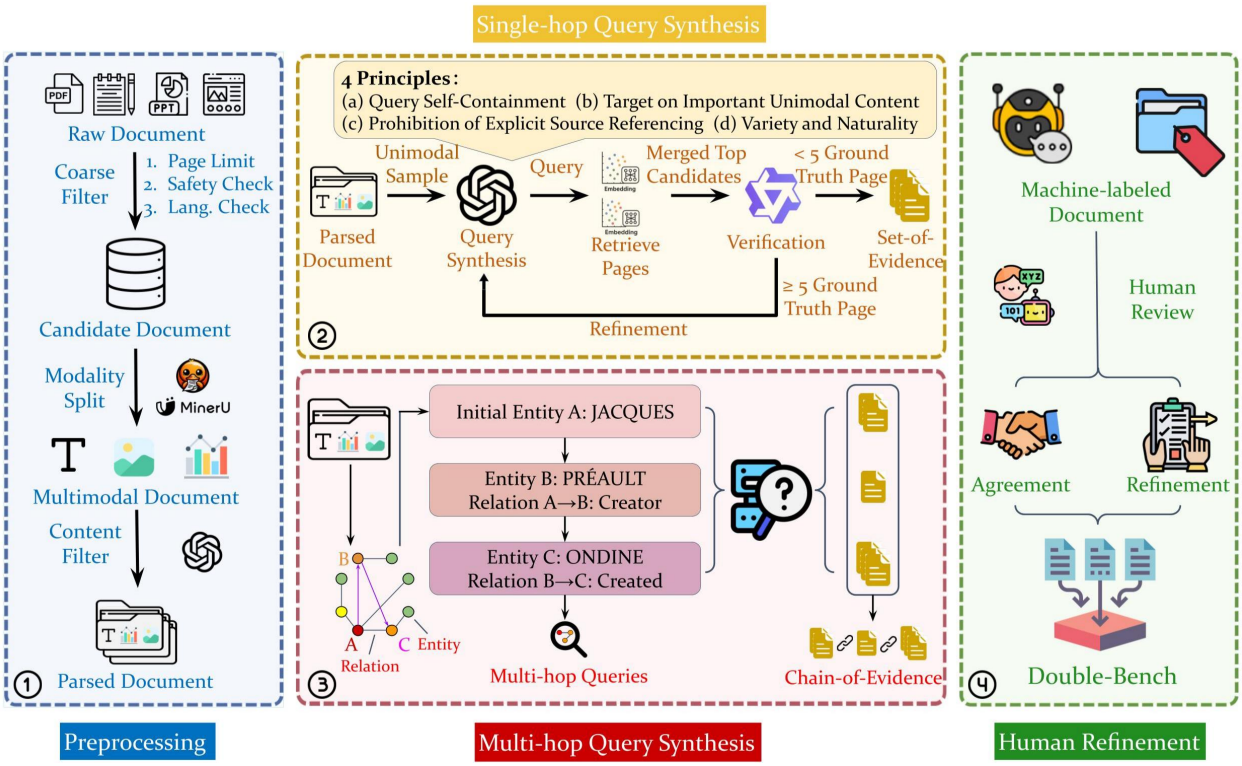

Figure 3: Overview of the DOUBLE-BENCH construction pipeline. The preprocessing stage filters the collected corpora and splits the content by modality. To alleviate identifed problems explained in Figure 2, an iterative clarity-oriented refinement pipeline is introduced for single-hop query generation, while knowledge graphs are additionally constructed to assist multi-hop query generation. All document pages are thoroughly checked by annotators to produce list-of-evidence and set-of-evidence labels.

4, we include high-quality PDF files, scanned documents, slides, and HTML pages to ensure the diversity coverage of our raw data. See appendix for more details.

Filtering and Preprocessing. Our preprocessing pipeline begins with a coarse-grained, rule-based filter (Pu et al. 2025) to select documents that meet basic structural and language requirements. GPT-4o (OpenAI 2024) reads the first three pages of every document to determine the primary language. Only documents with 10 to 50 pages and a primary language listed in Figure 4 are retained. Following this initial selection, each document page undergoes a modality split, where it is parsed and decomposed into its constituent text, table, and figure components using tools like Docling (Livathinos, Auer et al. 2025) and MinerU (Wang, Xu, and Zhao 2024). Finally, we apply a fine-grained content filter where parsed chunks are reviewed with their adjacent context to ensure semantic coherence, filtering out any content that is irrelevant or lacks meaningful information for query generation.

Single-hop Query Synthesis

Principles. Single-hop VQA queries often lack enough detail for precise document retrieval. We enhance them by adding supportive descriptions, making queries self-contained, focused on key unimodal information, and diverse in type including factual and analytical. This produces robust queries for evaluating single-hop retrieval.

Synthesis Process. The Single-hop query synthesis process goes through an iterative refinement process to ensure clarity. Provided a parsed page component obtained by preprocessing, we leverage `GPT-4o` to formulate an initial query based on the following four principles: (1) Query self-containment; (2) Target on most significant unimodal content; (3) Prohibition of explicit source eeferencing; (4) Variety and naturality in queries. This initial query is then validated against the corpus. Two high-performance embedding models of different modalities, `colqwen` and `qwen3-embedding`, are used to retrieve the top-10 candidate pages each. Subsequently, `Qwen2.5-VL-32B` processes the merged candidates seperately to identify all ground truth pages containing a direct answer. If more than five ground truth pages are found, the model is prompted to refine the query by incorporating a distinguishing detail extracted from one of the identified ground truth pages. This validation and refinement loop continues until the query yields five or fewer ground truth pages, ensuring the overall difficulty.

Multi-hop Query Synthesis

Principles. While multi-hop queries benefit from information across hops, which mitigates the lack of information problem, their direct generation by LLMs is challenging, even with techniques like Chain-of-Thought or inference scaling. Problems in current synthetic multi-hop benchmarks,

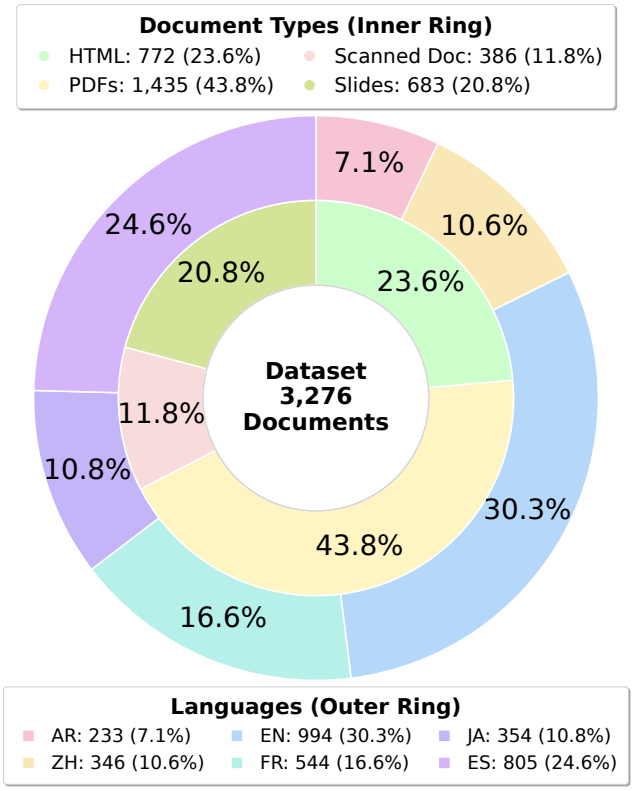

Figure 4: Statistics of the DOUBLE-BENCH dataset.

such as trivial connections and non-realistic intents, have undermined accurate evaluations of RAG frameworks. Our multi-hop query synthesis pipeline addresses this by using knowledge graphs and intent-driven walks. This approach simplifies sub-query linking by replacing key entities with new sub-queries, forming linearly combined queries.

Synthesis Process. Our multi-hop query synthesis process uses a knowledge graph-based approach to generate complex reasoning chains. We construct knowledge graphs for each document using LightRAG (Guo et al. 2024), ensuring extracted relationships uniquely identify target entities to prevent sub-query ambiguity. An LLM agent selects an initial node from high-degree entities, infers query intent based on the node and document summary, then performs a guided graph walk by iteratively selecting neighbors that best align with the inferred intent. The final multi-hop question is built iteratively by cumulating sub-queries along this path while rearranged for natural flow. This transforms a simple entity-relation path into a grammatically natural, complex question requiring sequential reasoning to answer.

Post-Processing

Query Quality Inspection. The query drafting module strives to generate high-quality queries, but a quality inspection is required to ensure all criteria are met. For single-hop queries, the checklist reviews all generation requirements. For multi-hop queries, we create a separate checklist, assessing: (1) Final question quality (clarity, specificity, no explicit final answer); (2) Logical necessity and correctness of intermediate reasoning steps; (3) Uniqueness of step answers and rigor of relations; and (4) Significance and relevance of the overall query. Queries failing any criteria are discarded.

Evidence Labeling. For each query, we locate all ground truth by thoroughly searching each page within the document. Pages are marked as evidence only if they directly provide or lead to the answer. We provide set-of-evidence labels for single-hop queries—a set of all evidence pages. Chain-of-evidence labels, which distinguishes the set-of-evidence labels for every hop, is provided for multi-hop queries.

Human Refinement. To improve benchmark quality and ensure accuracy, we further conduct human refinement. Although automated evidence labeling is sufficiently accurate, human annotators reviewed and adjusted 8% of labels with discrepancies. With 92% agreement, this step ensures precise ground truth data, enhancing DOUBLE-BENCH 's reliability.

Experiments

Evaluation Protocol. Following task formulation setting, we define the hit rate for retrieval accuracy evaluation of single- and multi-hop queries. The accuracy of the final answer is evaluated using LLM-as-a-judge (Zheng et al. 2023). `GPT-4o` rates the correctness the generated answer compared to the ground truth answer on a scale of 0 to 10. Answers with a score not lower than 7 count as correct, not higher than 3 count as incorrect, others count as partially correct.

Evaluated Models and Frameworks. We evaluate 4 competitive text embedding models, namely `bge-m3` (Chen et al. 2024c), `gte-Qwen2` (Li et al. 2023), `NV-Embed-v2` (Lee et al. 2024), `Qwen3-Embedding` (Zhang et al. 2025), 5 competitive open-source document page embedding models, namely `colpali` (Faysse et al. 2024), `colqwen` (Faysse et al. 2024), `gme` (Zhang et al. 2024), `vdr-2b` (LlamaIndex 2025), `jina-embeddings-v4` (Günther et al. 2025), and 3 advanced document RAG frameworks, namely `M3DocRAG` (Cho et al. 2024), `MDocAgent` (Han et al. 2025), `VidoRAG` (Wang et al. 2025). To understand how each part of RAG framework affects answer accuracy, we add a reference framework, `Colqwen-gen`, by directly pairing the strongest embedding model `Colqwen` with `GPT-4o`. Baseline results are reported with `GPT-4o` directly answering the queries without RAG. The Oracle setting estimates the upper-bound performance of RAG frameworks. In this setting, we provide the parsed contents of all the ground truth pages to the MLLM, together with a prompt that instructs the MLLM to first extract relevant information and think before providing the final answer. See Appendix for models and framework details.

Main Results

DOUBLE-BENCH is high-quality and low contaminated so that MLLMs still need to retrieve details for fully correct answers. As shown in Table 3, *state-of-the-art* MLLMs like `GPT-4o`, `Gemini`, and `Qwen` are able to make general responses without context, with 50% to 70% of responses being partially correct. Providing evidence pages to MLLMs

Model	Average			Single Hop			2-Hop			3-Hop		
	hit@1	hit@3	hit@5	hit@1	hit@3	hit@5	hit@1	hit@3	hit@5	hit@1	hit@3	hit@5
Text Embedding Models												
Qwen3-Embedding-4B	**0.489**	**0.699**	**0.776**	**0.726**	**0.852**	**0.886**	0.314	0.598	0.663	0.235	**0.531**	**0.668**
NV-Embed-v2	0.443	0.650	0.724	0.626	0.756	0.796	**0.333**	**0.604**	**0.689**	**0.240**	0.526	0.641
gte-Qwen2-7B-instruct	0.404	0.611	0.697	0.585	0.749	0.804	0.288	0.503	0.603	0.205	0.466	0.588
bge-m3	0.355	0.525	0.591	0.527	0.648	0.695	0.180	0.366	0.428	0.182	0.412	0.502
Visual & Multimodal Embedding Models												
colqwen2.5-3b-multilingual	**0.533**	**0.727**	**0.795**	**0.778**	**0.865**	**0.895**	**0.326**	**0.622**	**0.693**	**0.277**	**0.579**	**0.696**
vdr-2b-multi	0.463	0.648	0.725	0.688	0.813	0.847	0.283	0.491	0.589	0.225	0.482	0.606
jina-embeddings-v4	0.451	0.641	0.720	0.671	0.804	0.844	0.264	0.468	0.570	0.222	0.479	0.603
gme-Qwen2-VL-7B-Instruct	0.428	0.614	0.697	0.638	0.775	0.822	0.249	0.472	0.579	0.208	0.449	0.570
colpali-v1.3	0.403	0.571	0.646	0.584	0.679	0.717	0.230	0.440	0.525	0.220	0.469	0.588

Table 2: Retrieval accuracy of *state-of-the-art* text and multimodal embedding models across query types, showing performance degradation as reasoning complexity increases.

Model & Setting	Single Hop			Multi Hop		
	✓	⚔	✗	✓	⚔	✗
Models w.o. RAG						
Qwen3-32B text-only	0.242	0.488	0.271	0.193	0.293	0.515
Qwen2.5-VL-7B w.o. RAG	0.053	0.557	0.390	0.127	0.168	0.705
GPT-4o w.o. RAG	0.109	0.748	0.144	0.197	0.332	0.472
Qwen2.5-VL-32B w.o. RAG	0.200	0.621	0.179	0.159	0.319	0.521
Llama 4 Maverick w.o. RAG	**0.245**	0.480	0.275	**0.215**	0.193	0.592
Models Oracle						
Qwen2.5-VL-7B Oracle	0.406	0.490	0.104	0.456	0.241	0.303
GPT-4o Oracle	0.678	0.141	0.181	0.538	0.271	0.191
Llama 4 Maverick Oracle	0.601	0.350	0.049	0.524	0.192	0.284
Qwen2.5-VL-32B Oracle	**0.874**	0.061	0.066	**0.643**	0.312	0.045

Table 3: Evaluation of MLLMs' long document understanding capability. *Oracle*: directly providing evidence page.

substantially boosts accuracy, with 3x to 5x responses being completely correct compared to *w.o.* RAG setting. This indicates that our benchmark is well-suited for evaluating the retrieval and synthesis components of RAG systems, as it clearly distinguishes context-grounded reasoning from a model's inherent knowledge. Notably, the robust performance of `Qwen2.5-VL` observed in the upper bound setting, which closely mirrors our benchmark curation pipeline, further suggesting the robustness and effectiveness of our pipeline in identifying correct evidence pages of queries.

Document-specified embedding model outperform general ones, and gap between text and image embedding models is narrowing. DOUBLE-BENCH provides a clear divergence in the retrieval performance of various embedding models, as detailed in Table 2. The model rankings within DOUBLE-BENCH align well with popular text embedding leaderboards MTEB (Muennighoff et al. 2022) and document retrieval benchmark ViDoRe v2 (Macé, Loison, and Faysse 2025), demonstrating the robustness of our benchmark.

`ColQwen2.5-3B` significantly outperforms general multimodal embedding models like `jina-embeddings-v4` and `GME`, achieving a 9% higher average hit rate and demonstrating strong potential in document retrieval. Other multimodal models struggle, even trailing the text-only `Qwen3-Embedding`. We attribute this to advanced text training techniques—such as hard negative sampling and data synthesis—which are difficult to transfer to visual models due to cost and data constraints. Although visual embedding models have inherent advantages for visual content retrieval, the semantic complexity of document RAG tasks negates this advantage. The critical influence of both visual observation and textual understanding abilities incentive combined strategies such as interleaved embedding models and advanced multimodal understanding pipelines.

Document RAG frameworks bottleneck still lies on retrieval accuracy, where designing advanced strategies may help. Most frameworks strive to design complex information mining pipelines to extract maximum value from retrieved pages, yet tend to pay little attention to the retrieval stage itself. However, our experiments demonstrate strong correlations between retrieval accuracy and answer accuracy, as shown in Table 4. Equipped with a single MLLM pass, `Colqwen-gen` even partially outperforms `MDocAgent` on multi-hop queries, despite the latter seamlessly integrating multiple agents to provide final answers. This underscores the critical importance of optimizing the retrieval stage, potentially through finer-grained document preprocessing, exploiting the hierarchical and semantic structure of documents and developing more powerful or integrated embedding models.

In-Depth Analysis

The overconfidence dilemma: trading trustworthiness for answers. To investigate the bottleneck in existing RAG frameworks, we breakdown each reponse of `M3DocRAG` and `MDocAgent` to analyze whether the error comes from retrieval or answering, and look into the trade-off between answering accuracy and the ability to identify insufficient information (also known as honesty).

Figure 5 reveals notable divergence in agent behavior. Simpler agents like `M3DocRAG` adopt a cautious strategy, answering a lower proportion of queries with successfully retrieved context but reliably identifying retrieval failures and refusing to respond. In contrast, more complex agents like `MDocAgent` and `ViDoRAG` exhibit significant overconfidence. While they achieve higher accuracy on retrieval hits, they indiscrimi-

Framework	Average				Single Hop				2-Hop				3-Hop			
	Retrieval	Answer			Retrieval	Answer			Retrieval	Answer			Retrieval	Answer		
	hit@5	✓	✗	✗	hit@5	✓	✗	✗	hit@5	✓	✗	✗	hit@5	✓	✗	✗
MDocAgent (Han et al. 2025)	0.688	**0.645**	0.126	0.229	0.830	**0.757**	0.132	0.111	0.572	**0.567**	0.065	0.367	0.549	0.532	0.135	0.332
ViDoRAG (Wang et al. 2025)	0.682	0.536	0.138	0.326	0.822	0.623	0.144	0.233	0.539	0.457	0.112	0.431	0.544	0.447	0.137	0.416
M3DOCRAG (Cho et al. 2024)	0.608	0.451	0.121	0.428	0.709	0.538	0.138	0.324	0.490	0.330	0.088	0.582	0.519	0.382	0.110	0.508
Colqwen-gen (Faysse et al. 2024)	**0.795**	0.604	0.135	0.261	**0.895**	0.676	0.160	0.164	**0.693**	0.462	0.143	0.395	**0.696**	**0.554**	0.100	0.346

Table 4: Performance of RAG Frameworks. `Colqwen-gen` achieves comparable performance with `MDocAgent`, the best among evaluated frameworks. This observation highlights the need for more advanced retrieval stage frameworks.

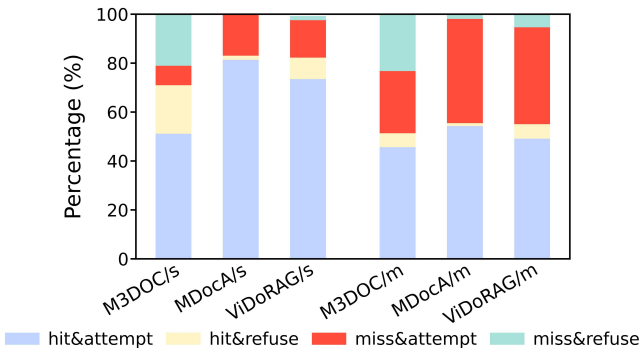

Figure 5: Breakdown of retrieval and response model performance of frameworks under single (s) and multi-hop (m) queries. Our analysis reveals that performance drops on multi-hop questions are mainly due to retrieval failures that cause models to abstain from answering.

nately attempt to answer nearly every query, regardless of whether sufficient information was retrieved.

This observation indicates that recent document RAG development has over-emphasized maximizing answer generation at the expense of *"epistemic humility"*, *i.e.*, the crucial skill of knowing what it doesn't know and admitting when an answer cannot be found. Consequently, we argue that future research should pursue more trustworthy RAG frameworks where identifying informational gaps is as valued as accuracy.

Time efficiency of frameworks. Agent efficiency is as important a metric as effectiveness. Since API completion times may vary across models, Appendix reports the normalized time efficiency of the evaluated frameworks. Both `MDocAgent` and `ViDoRAG` employ a sequential agent coordination pattern, which significantly increases their inference time. Note that `ViDoRAG` dynamically controls the generation process, so we report the lower and upper bound theoretical time efficiency of `ViDoRAG` estimated by API call times.

Inference patterns of MLLMs as response model. We also observe different answering strategy in MLLMs in Table 4 and Appendix. When directly provided with a multi-hop query, response model tend not to process them hop-by-hop. On the contrary, they first collect signature information—the most distinguishing or identifiable pieces—from the various hops. Following this, models tend to perform a direct inclusion based elimination to arrive the final answer. This mechanism differentiates significantly from our expectation of how models might sequentially solve multi-hop queries. This provides a compelling point of view: merely increasing the number of hops may not increase its difficulty.

Related Work

Multimodal Document Retrieval. Different from traditional text retrieval (Zhao et al. 2024), documents (Masry et al. 2022; Tanaka et al. 2023) often contain multimodal information, which may be time consuming and would cause information loss when directly parsed to text. Therefore, recent works have dedicated great effort to improve the accuracy and efficiency of document retrieval with MLLMs. One line of work adopts high quality synthetic data (Zhang et al. 2024; Chen et al. 2025), hardness aware training (Lan et al. 2025; Lee et al. 2024) and retrieval-optimized network architectures (Faysse et al. 2024) for more precise embedding models. Another line of work leverages LLM/MLLM agentic flows to process different modalities in parallel (Han et al. 2025) and perform iterative inference steps for more grounded and informative answers (Wang et al. 2025).

Document RAG Benchmarks. The increasing attention on Document RAG (Ma et al. 2024) and VQA (Mathew, Karatzas, and Jawahar 2021) necessitates comprehensive multimodal retrieval benchmarks. Common practices often use VQA dataset queries (Friel, Belyi, and Sanyal 2024; Faysse et al. 2024; Cho et al. 2024), but these are document-specific and lack information for global retrieval. Other benchmarks (Wang et al. 2025; Dong et al. 2025) craft informative queries from single pages, yet often only mark that single page as relevant, ignoring other potential matches and risking evaluation inaccuracies. Some recent benchmarks (Macé, Loison, and Faysse 2025; Wasserman et al. 2025) have identified contextual gaps between artificial and realistic queries, and strive to provide evaluations that fully reflect real use scenarios.

Conclusion

We introduce DOUBLE-BENCH, a large-scale, multimodal, multilingual benchmark designed to reflect realistic retrieval-augmented generation scenarios, overcoming limitations of prior work with validated chain-of-evidence and comprehensive assessment. Evaluations of leading embedding models and RAG frameworks reveal several crucial bottlenecks. We hope our fully open-sourced code, framework and dataset establish a strong foundation for document RAG system.

Acknowledgements

This work is supported by Guangdong Provincial Natural Science Foundation Project (2025A1515010113), Shandong Provincial Natural Science Foundation Project (ZR2024LZH012), Guangxi Key Research and Development Project(2024AB02018) and the Major Key Project of PCL, China under Grant PCL2025AS11.

References

Breci, E.; Guarnera, L.; and Battiato, S. 2024. A novel dataset for non-destructive inspection of handwritten documents. *arXiv preprint arXiv:2401.04448*.

Chen, D.; Chen, R.; Zhang, S.; Wang, Y.; Liu, Y.; Zhou, H.; Zhang, Q.; Wan, Y.; Zhou, P.; and Sun, L. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Chen, H.; Wang, L.; Yang, N.; Zhu, Y.; Zhao, Z.; Wei, F.; and Dou, Z. 2025. mmE5: Improving Multimodal Multilingual Embeddings via High-quality Synthetic Data. *arXiv preprint arXiv:2502.08468*.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17754–17762.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024c. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M.; Gonzalez, J. E.; et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Cho, J.; Mahata, D.; Irsoy, O.; He, Y.; and Bansal, M. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Dong, K.; Chang, Y.; Goh, X. D.; Li, D.; Tang, R.; and Liu, Y. 2025. MMDocIR: Benchmarking Multi-Modal Retrieval for Long Documents. *arXiv preprint arXiv:2501.08828*.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; and Colombo, P. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Friel, R.; Belyi, M.; and Sanyal, A. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Günther, M.; Sturua, S.; Akram, M. K.; Mohr, I.; Ungureanu, A.; Eslami, S.; Martens, S.; Wang, B.; Wang, N.; and Xiao, H. 2025. jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval. *arXiv preprint arXiv:2506.18902*.

Guo, Z.; Xia, L.; Yu, Y.; Ao, T.; and Huang, C. 2024. LightRAG: Simple and Fast Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.05779*.

Han, S.; Xia, P.; Zhang, R.; Sun, T.; Li, Y.; Zhu, H.; and Yao, H. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

Hui, Y.; Lu, Y.; and Zhang, H. 2024. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *arXiv preprint arXiv:2406.15187*.

Lan, Z.; Niu, L.; Meng, F.; Zhou, J.; and Su, J. 2025. LLaVE: Large Language and Vision Embedding Models with Hardness-Weighted Contrastive Learning. *arXiv preprint arXiv:2503.04812*.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, Y.; Li, Y.; Wang, X.; Jiang, Y.; Zhang, Z.; Zheng, X.; Wang, H.; Zheng, H.-T.; Huang, F.; Zhou, J.; et al. 2024. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; and Zhang, M. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Livathinos, N.; Auer, C.; et al. 2025. Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion. *arXiv preprint arXiv:2501.17887*.

LlamaIndex. 2025. vdr-2b-multi-v1. Hugging Face Model Card.

Ma, Y.; Zang, Y.; Chen, L.; Chen, M.; Jiao, Y.; Li, X.; Lu, X.; Liu, Z.; Ma, Y.; Dong, X.; et al. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*.

Macé, Q.; Loison, A.; and Faysse, M. 2025. ViDoRe Benchmark V2: Raising the Bar for Visual Retrieval. *arXiv preprint arXiv:2505.17166*.

Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Mathew, M.; Karatzas, D.; and Jawahar, C. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2200–2209.

Mortaheb, M.; Khojastepour, M. A. A.; Chakradhar, S. T.; and Ulukus, S. 2025a. Rag-check: Evaluating multimodal retrieval augmented generation performance. *arXiv preprint arXiv:2501.03995*.

Mortaheb, M.; Khojastepour, M. A. A.; Chakradhar, S. T.; and Ulukus, S. 2025b. Re-ranking the Context for Multimodal Retrieval Augmented Generation. *arXiv preprint arXiv:2501.04695*.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

OpenAI. 2024. GPT-4o. https://openai.com/gpt-4o. Accessed: 2025-05-01.

Pu, S.; Wang, Y.; Chen, D.; Chen, Y.; Wang, G.; Qin, Q.; Zhang, Z.; Zhang, Z.; Zhou, Z.; Gong, S.; et al. 2025. Judge Anything: MLLM as a Judge Across Any Modality. *arXiv preprint arXiv:2503.17489*.

Saad-Falcon, J.; Khattab, O.; Potts, C.; and Zaharia, M. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*.

Tanaka, R.; Nishida, K.; Nishida, K.; Hasegawa, T.; Saito, I.; and Saito, K. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13636–13645.

Tang, Y.; and Yang, Y. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Wang, B.; Xu, C.; and Zhao, X. 2024. MinerU: An Open-Source Solution for Precise Document Content Extraction. *arXiv preprint arXiv:2409.18839*.

Wang, Q.; Ding, R.; Chen, Z.; Wu, W.; Wang, S.; Xie, P.; and Zhao, F. 2025. ViDoRAG: Visual Document Retrieval-Augmented Generation via Dynamic Iterative Reasoning Agents. *arXiv preprint arXiv:2502.18017*.

Wasserman, N.; Pony, R.; Naparstek, O.; Goldfarb, A. R.; Schwartz, E.; Barzelay, U.; and Karlinsky, L. 2025. REAL-MM-RAG: A Real-World Multi-Modal Retrieval Benchmark. *arXiv preprint arXiv:2502.12342*.

Wu, Y.; Long, Q.; Li, J.; Yu, J.; and Wang, W. 2025. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *arXiv preprint arXiv:2502.16636*.

Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; and Liu, Z. 2024. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, 102–120. Springer.

Yu, Q.; Xiao, Z.; Li, B.; Wang, Z.; Chen, C.; and Zhang, W. 2025. MRAMG-Bench: A BeyondText Benchmark for Multimodal Retrieval-Augmented Multimodal Generation. *arXiv e-prints*, arXiv–2502.

Zhang, X.; Zhang, Y.; Xie, W.; Li, M.; Dai, Z.; Long, D.; Xie, P.; Zhang, M.; Li, W.; and Zhang, M. 2024. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv preprint arXiv:2412.16855*.

Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; et al. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.

Zhang, Z.; Feng, Y.; and Zhang, M. 2025. LevelRAG: Enhancing Retrieval-Augmented Generation with Multi-hop Logic Planning over Rewriting Augmented Searchers. *arXiv preprint arXiv:2502.18139*.

Zhao, W. X.; Liu, J.; Ren, R.; and Wen, J.-R. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4): 1–60.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.