WikiMixQA: A Multimodal Benchmark for Question Answering over Tables and Charts

Negar Foroutan[1], Angelika Romanou[1], Matin Ansaripour[1]
Julian Martin Eisenschlos[23], Karl Aberer[1], Rémi Lebret[1]
[1]EPFL, [2]Google DeepMind, [3] Universidad Nacional de Córdoba
Correspondence: {negar.foroutan}@epfl.ch

Abstract

Documents are fundamental to preserving and disseminating information, often incorporating complex layouts, tables, and charts that pose significant challenges for automatic document understanding (DU). While vision-language large models (VLLMs) have demonstrated improvements across various tasks, their effectiveness in processing long-context vision inputs remains unclear. This paper introduces WikiMixQA, a benchmark comprising 1,000 multiple-choice questions (MCQs) designed to evaluate cross-modal reasoning over tables and charts extracted from 4,000 Wikipedia pages spanning seven distinct topics. Unlike existing benchmarks, WikiMixQA emphasizes complex reasoning by requiring models to synthesize information from multiple modalities. We evaluate 12 state-of-the-art vision-language models, revealing that while proprietary models achieve ∼70% accuracy when provided with direct context, their performance deteriorates significantly when retrieval from long documents is required. Among these, GPT-4-o is the only model exceeding 50% accuracy in this setting, whereas open-source models perform considerably worse, with a maximum accuracy of 27%. These findings underscore the challenges of long-context, multi-modal reasoning and establish WikiMixQA as a crucial benchmark for advancing document understanding research.[1]

1 Introduction

Documents serve as a fundamental medium for preserving and exchanging information, with millions being generated daily across various domains. Beyond plain text, they often include complex layouts, tables, and images, making automatic document understanding (DU) a crucial challenge. NLP-driven DU enables efficient extraction, organization, and interpretation of this information, supporting real-world information retrieval and decision-making.

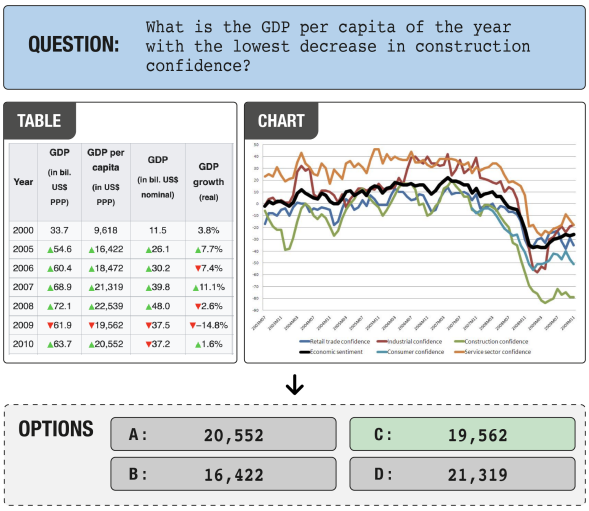


Figure 1: An example from WikiMixQA illustrating a question whose answer relies on the information presented in the accompanying table and chart.

A notable challenge in DU arises from the prevalence of documents containing large tables and charts, which can be difficult for humans to process and analyze. A question-answering (QA) system would help humans get insights from documents with such interjections easily. Over the past few years, several Visual Question Answering (VQA) benchmarks have been developed to assess the DU capabilities of vision-language large models (VLLMs) across various aspects, including handling tables, charts, and document layouts. However, most existing benchmarks primarily focus on single-page documents.

Another key challenge in automatic DU is the ability to answer complex questions that require integrating information across multiple sections of a document and reasoning over different modalities. Current benchmarks largely lack multi-hop questions where models must synthesize information from multiple modalities—such as text, tables, and charts—to derive correct answers (Ma et al., 2024; Van Landeghem et al., 2023). Furthermore,

[1]Code and dataset is released here.

Findings of the Association for Computational Linguistics: ACL 2025, pages 24941–24958

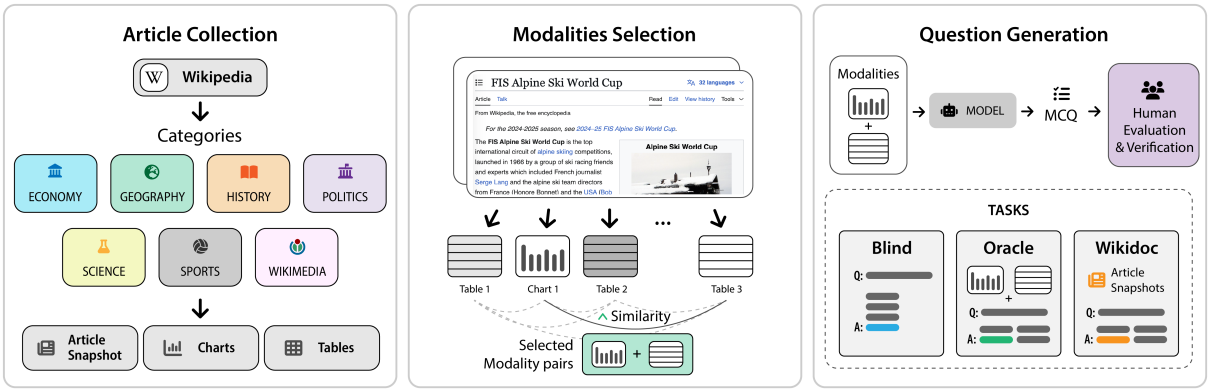


Figure 2: WikiMixQA Creation Pipeline: (1) We collect Wikipedia articles that contain tables and charts. (2) For each article, we identify table-chart pairs that exhibit semantic similarity. (3) We employ GPT-4-turbo to generate multiple-choice questions (MCQs) based on each table-chart pair. (4) Human annotators assess and validate the quality of the generated questions to ensure accuracy and relevance.

existing datasets lack controlled evaluation settings that isolate the specific modalities and types of information required to answer a question, making it difficult to conduct fine-grained analyses of VLLMs' limitations in DU tasks.

To bridge this gap, we introduce WikiMixQA, a benchmark dataset constructed using Wikipedia as the primary source of documents. WikiMixQA focuses on long, digital-only documents that contain multiple large tables and charts. The questions in our dataset are specifically designed to require multi-modal reasoning over these structured elements, covering a broad spectrum of topics. The dataset comprises 1,000 multiple-choice questions (MCQs) derived from ∼4k Wikipedia pages, with each document averaging 24.18 pages and 1815.01 ± 2825.16 textual tokens. The questions span diverse domains, including *Economy, Geography, History, Politics, Science, Sport*, and *Wikimedia*. To ensure the need for multi-modal reasoning, questions are structured to require information from either two tables (table-table), two charts (chart-chart),[2] or a combination of one table and one chart (table-chart).

The dataset construction follows a systematic pipeline: (1) We collect approximately 7,200 Wikipedia pages containing at least three tables and one chart; (2) We identify highly similar table-table, chart-chart, or table-chart pairs to ensure meaningful cross-modal reasoning; (3) We employ GPT-4-turbo to generate MCQs based on predefined criteria; and (4) A rigorous human curation process is conducted to refine and validate 1,000 fully curated questions.

We conduct an extensive evaluation on WikiMixQA using four open-source and eight closed-source state-of-the-art VLLMs across three different experimental settings, where we vary the level of contextual information provided to the models. The results, summarized in Table 1, reveal that while closed-source models perform relatively well (∼70%) when provided with the exact relevant information, they struggle significantly when required to retrieve relevant context from long documents before answering the questions. Notably, GPT-4-o is the only model to exceed 50% accuracy in such a setting, while other closed-source models exhibit near-random performance. Open-source models perform even worse, with the highest accuracy reaching only 27% when exact information is provided as input. These findings underscore the persistent challenge of long-context multi-modal document understanding for VLLMs.

2 Dataset Construction

This section outlines the pipeline employed for collecting Wikipedia documents, extracting and selecting their associated modalities (*i.e.*, tables and charts), and generating the multiple-choice question (MCQ) samples that constitute WikiMixQA. An overview of the WikiMixQA creation pipeline is illustrated in 2, with a detailed explanation provided in the Appendix B.

[2]We define a chart as a graphical representation of data, including: (a) data charts such as diagrams or graphs that organize and display numerical or qualitative information; (b) maps enhanced with additional data; and (c) other domain-specific constructs, such as chord charts or record charts.

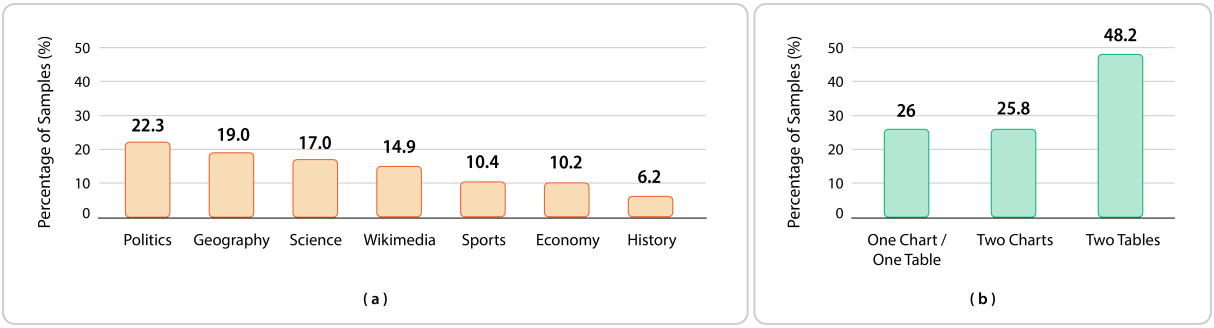


Figure 3: (a) Distribution of question-answer pairs across seven topics. (b) Distribution of question-answer pairs by modality type.

2.1 Document collection

To construct the dataset, we used the WTabHTML[3] project's preprocessed English Wikipedia dumps from March 2022, initially comprising over 4 million entries. We filtered out articles with fewer than three tables to eliminate small, less relevant tables, narrowing the set to 392,223 entries.

To ensure multimodality, we downloaded over a million images from these articles and filtered out non-chart images using a fine-tuned Vision Transformer (ViT) model,[4] retaining relevant formats like PNG and JPEG. Articles with at least one valid chart were further filtered, reducing the set to 15,164 entries. To promote diversity, we categorized each document using Wikipedia's *"Instance of"* property and grouped similar categories into a custom taxonomy of seven main categories: *Economy, Geography, History, Politics, Science, Sport,* and *Wikimedia*. This final step reduced the dataset to 7,258 documents, ensuring broad coverage across various topics. Table 4 shows the document statistics for each category.

2.2 Modalities selection

Wikipedia documents often contain multiple tables and charts. If questions are generated from randomly selected tables and charts, the resulting question set may lack diversity. Furthermore, if the selected tables and/or charts are not semantically relevant, generating meaningful and challenging questions becomes infeasible. To address these issues, we focus on generating questions that involve structured modality pairs: two tables (*table-table*), one table and one chart (*table-chart*), or two charts (*chart-chart*). Figures 1 and 5 show examples for each of these question types.

[3]Available at https://github.com/phucty/wtabhtml

[4]https://huggingface.co/facebook/dinov2-base-imagenet1k-1-layer

To avoid irrelevant pairings, we selected pairs based on the textual similarity of the descriptions of each pair. Since most tables lacked captions, we used the `Llama-3-8B-Instruct`[5] language model to generate descriptions from their raw HTML. For images identified as potential charts, we employed the vision-language model `GPT-4-turbo` to confirm whether they were charts and extract key information in fewer than 200 words. This approach ensured meaningful modality descriptions, with full prompt details provided in Appendix E.

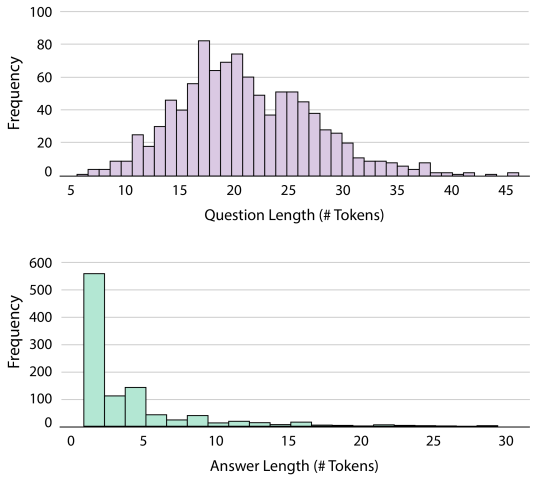


Figure 4: Distribution of questions and answers lengths.

Once we generated textual descriptions for both HTML tables and images, we calculated similarity scores for each possible modality pair (*table-table*, *table-chart*, and *chart-chart*) within a document. We used the *cross-encoder* model `BAAI/bge-reranker-v2-m3` (Chen et al., 2024), available on HuggingFace, which directly outputs a similarity score for two inputs instead of generating embeddings. Although slower than embed-

[5]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

ding models, this reranker model provides more accurate results. Since each document contains a limited number of modalities, this approach balances speed and accuracy effectively. Only images identified as charts by GPT-4-turbo were considered for this process. Appendix C provides more information on what we consider as charts.

2.3 Question generation

To ensure meaningful and challenging questions, we filtered modality pairs (*table-table*, *table-chart*, and *chart-chart*) based on similarity scores. We calculated the macro mean similarity score for each pair type across topics and retained pairs with scores between the macro mean and 0.9. Tables with fewer than 512 characters were also excluded for their limited information content. We aimed for a balanced distribution of modality types and topics by selecting an equal number of pairs per type and topic, capping the number of pairs from the same document. Three types of multiple-choice questions were generated: two using individual modalities and one combining both. Each question had four options, one correct answer, and an explanation. Ultimately, 3,528 question-answer pairs were generated using GPT-4-turbo.

Quality Control Through manual inspection of the generated question-answer pairs, we observed that some of the generated questions were invalid. Despite efforts to select related modalities, there were cases where the information between the two modalities did not overlap (*e.g.*, tables with differing date ranges or charts representing unrelated regions). To reduce the number of invalid questions reaching the annotation phase, we utilized a state-of-the-art vision-language model, `OpenGVLab/InternVL2-Llama3-76B`, available on HuggingFace[6].

The model was provided with both modalities, and we prompted it to determine whether sufficient information was present in the charts and/or HTML tables to answer the given question (see the full prompt in Appendix G). If the model's response was "yes," we followed up with the question: Is this answer correct: `gpt4_full_answer`? Answer with "yes" or "no." Here, `gpt4_full_answer` refers to the answer originally suggested by GPT-4 during the generation process. Only pairs passing this two-step evaluation were retained for further processing.

2.4 Human curation

Out of the 3,528 generated candidates, we selected 2,001 question-answer pairs for annotation. This included 938 pairs positively evaluated by the InternVL2 model and 1,063 pairs randomly sampled from the remaining dataset. Three Master's students in Computer Science annotated all the selected pairs. The annotation process consisted of two steps:

1. *Validity Check*: Annotators first determined if a question could only be answered by integrating information from both provided modalities. This ensured there was no informational overlap between modalities and that both were essential context for the question.

2. *Answer Assessment*: For question-answer pairs deemed valid, annotators assessed the correctness of the provided answers, labeling each pair as "Correct", "Wrong", or "Small Edit". The "Small Edit" label was used for cases where the question could be retained after minor revisions. During this step, annotators also verified that incorrect answer options were plausible and contextually grounded.

Using majority voting, 595 questions were labeled as Correct. Invalid pairs were revised for issues like overly detailed questions or multiple correct answers. Pairs labeled "Small Edit" were refined, resulting in 405 additional corrected pairs being added to the final dataset. Figure 8 shows the interface of our annotation tool.

2.5 The `WikiMixQA` benchmark

Combining questions labeled as *Correct* with the revised questions, the final dataset comprises 1000 question-answer pairs derived from 526 unique Wikipedia documents. The distribution of question-answer pairs across topics is relatively balanced, reflecting the natural topical distribution of Wikipedia, as shown in Figure 3.

Approximately 515 of the 1000 pairs were validated as *Correct* by our AI evaluator (`InternVL2-Llama3-76B` model), underscoring the value of sampling from initially rejected questions to enhance dataset diversity. The distribution of question-answer pairs by modality type is illustrated in Figure 3. Notably, nearly half of the pairs

[6]Available at https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B

involve reasoning across two tables. Figure 4 indicates the distribution of questions' and answers' lengths.

3 Evaluation

Evaluation Setup To rigorously assess the performance of state-of-the-art Vision-Language Learning Models (VLLMs), we design three distinct evaluation setups that vary the amount of contextual information supplied to the model. These setups are defined as follows: (1) `blind`: In this scenario, no contextual information is provided to the model. The model is tasked with answering the question based solely on its internal knowledge or reasoning capabilities. (2) `oracle`: Here, the model is supplied with the necessary visual or tabular data, such as charts or tables, that are essential for answering the question. This setup isolates the model's ability to interpret and reason with the provided structured data (3) `wikidoc`: In this case, the model is given snapshots of the Wikipedia page from which the question was derived. This setup evaluates the model's capacity to process and utilize textual information from a comprehensive and unstructured source. Due to thecomputational-heavy nature of this setup, we only use it for closed-source models.

Models We conduct an evaluation of the benchmark using state-of-the-art vision large language models (VLLMs), encompassing both open-source and closed-source models across various scales. For open-source models, we include the Qwen family of models (Yang et al., 2024) (*Qwen2-VL-7B-Instruct* and *Qwen2-VL-72B-Instruct*), OpenGVLab's InternVL2 (Chen et al., 2023) series (*InternVL2.5-1B*, *InternVL2.5-5B*, *InternVL2.5-26B*, *InternVL2.5-78B*), and Meta's Llama (Dubey et al., 2024) series (*Llama-3.2-11B-Vision-Instruct*). For closed-source models, the benchmark evaluation is conducted on *GPT-4o* (Achiam et al., 2023), *Gemini-1.5-Flash*, *Gemini-1.5-Pro* (Team et al., 2024), and *Claude3.5-Sonnet* (2024). To evaluate open-source models, we use the vLLM library,[7] on a machine with 8 NVIDIA A100 GPUs (40GB memory).

Evaluation Metrics For each evaluation setting, we provide the model with a set of questions and their corresponding contextual information, prompting it to generate the correct answer choice (*i.e.*, A, B, C, or D). We then assess model performance by measuring *accuracy* and comparing results across different models. Appendix H provides details regarding the prompt design and usage.

[7] https://docs.vllm.ai/en/latest/

Model	Blind	Oracle	Wikidoc
GPT-4o	33.46	71.42	55.24
Gemini-2.0-pro	22.67	69.53	23.47
Gemini-2.0-flash	23.27	67.52	24.47
Claude3.5-Sonnet	11.28	70.82	35.56
InternVL2.5-1B-MPO	19.88	23.17	-
InternVL2.5-5B-MPO	22.17	27.87	-
InternVL2.5-26B-MPO	11.48	26.37	-
InternVL2.5-78B-MPO	03.09	27.37	-
InternVL2.5-78B	03.29	27.67	-
Qwen2.5-VL-7B-Instruct	12.68	22.87	-
Qwen2.5-VL-72B-Instruct	0.39	23.17	-
Llama-3.2-11B-Vision-Instruct	10.68	14.08	-
Human Experts	-	87.50	-

Table 1: Models' performance (**accuracy %**) under three different evaluation settings. Random baseline is 25%.

4 Analysis

Table 1 presents the performance of the evaluated models on `WikiMixQA`. In the `blind` setting, where models lack access to contextual information, all models perform below random chance, with the exception of GPT-4o, which achieves a slightly higher accuracy of 33%. This result is expected, as questions in `WikiMixQA` require contextual information to be answered correctly, and without such context, models are unable to make informed predictions.

In the `oracle` setting, where models receive direct access to relevant context, proprietary models perform significantly better. GPT-4o achieves the highest accuracy, with Claude and Gemini models closely following. In contrast, open-source models exhibit poor performance, performing at or near random levels.

For the `wikidoc` setting, we exclude closed-source models due to their limited context length, which prevents them from processing the full Wikipedia snapshots required for answering questions. Among open-source models, only GPT-4o surpasses 50% accuracy. This observation suggests that while these models perform well when provided with explicitly relevant information, they struggle when required to process long-context inputs. The challenge arises from their need to first locate relevant information within extensive text

Model	History	Politics	Geography	Sports	Science	Economy	Wikimedia
GPT-4o	74.12	68.61	**76.32**	**75.96**	**72.94**	58.25	72.48
Gemini-2.0-flash	74.19	65.02	72.11	70.19	65.88	56.31	70.47
Gemini-2.0-pro	**75.81**	71.75	72.11	69.23	69.41	52.43	72.48
Claude3.5-Sonnet	67.74	**76.68**	69.47	71.15	68.24	**61.17**	**74.50**
InternVL2.5-1B-MPO	30.65	22.87	18.95	25.00	25.29	15.53	27.52
InternVL2.5-5B-MPO	32.26	30.49	28.42	21.15	27.65	25.24	28.19
InternVL2.5-26B-MPO	32.26	26.01	25.79	24.04	24.12	23.30	31.54
InternVL2.5-78B-MPO	33.87	22.42	27.89	27.88	24.71	32.04	30.87
InternVL2.5-78B	35.48	24.66	25.79	31.73	24.12	33.01	28.86
Qwen2.5-VL-7B-Instruct	27.42	22.87	17.37	23.08	24.12	21.36	27.52
Qwen2.5-VL-72B-Instruct	30.65	22.87	18.95	25.00	25.29	15.53	27.52
Llama-3.2-11B-Vision-Instruct	06.45	14.35	16.84	14.42	17.06	11.65	11.41

Table 2: Models' performance (**accuracy %**) across various topics in the `oracle` evaluation setting.

Model	2 Charts	2 Tables	1 Chart/1 Table
GPT-4o	71.31	71.63	71.15
Gemini-2.0-flash	53.48	74.12	69.23
Gemini-2.0-pro	54.65	77.43	69.61
Claude3.5-Sonnet	66.66	73.29	70.38
InternVL2.5-1B-MPO	20.93	24.43	23.07
InternVL2.5-8B-MPO	25.19	31.26	24.23
InternVL2.5-26B-MPO	24.03	29.19	23.46
InternVL2.5-78B-MPO	24.41	29.39	26.53
InternVL2.5-78B	24.41	30.02	26.53
Qwen2.5-VL-7B-Instruct	18.99	24.63	23.46
Qwen2.5-VL-72B-Instruct	22.48	24.22	21.92
Llama-3.2-11B-Vision-Instruct	28.29	02.27	21.92

Table 3: Models' performance (**accuracy %**) across three question types in the `oracle` setting.

before formulating an answer, highlighting a key limitation in handling extended vision-context scenarios.

In our human evaluation, annotators achieved an accuracy of 87% in the `oracle` setting, revealing a substantial performance gap of approximately 17% between current VLLMs and human performance. This discrepancy underscores the challenges that LVLMs face in document understanding and highlights the necessity of our benchmark for advancing research in this area.

Question Type Analysis Table 3 presents a breakdown of model performance across three question types: (i) questions involving two charts, (ii) questions involving two tables, and (iii) questions involving a combination of one chart and one table.

For questions that require interpreting two charts, GPT-4o demonstrates the highest performance, outperforming Claude and Gemini models by 5% and 17%, respectively. This result suggests that GPT-4o is particularly effective at processing and reasoning over chart-based data compared to other models.

In contrast, for questions involving two tables, GPT-4o exhibits a relatively consistent performance across all question types. However, Gemini-2.0-pro achieves the highest accuracy in this category, indicating its superior capability in handling tabular data.

Finally, for questions that involve both a chart and a table, all proprietary models achieve similar performance levels. Given that closed-source models perform at a level indistinguishable from random chance, further fine-grained analysis is not meaningful in this context.

Topic Analysis Table 2 presents a breakdown of models performance across different question types in the `oracle` setting. The results indicate that models perform relatively consistently across various topics, with the exception of the Economy topic, where all models exhibit slightly lower performance. A potential explanation for this discrepancy is that Economy-related questions frequently involve bar and line charts, necessitating both chart interpretation and comparative analysis. These tasks may pose greater challenges for the models, leading to a decline in performance on these questions.

5 Related Work

Visual Question Answering (VQA) is a crucial subtask of document understanding (DU), where the objective is to generate natural language answers to questions based on a given visual document. Previous research has explored the DU capabilities of vision large language models (VLLMs) by introducing new datasets and benchmarks. Many

Topic	Subtopic	Docs	Images	Charts	Tables	Table Rows	Tokens
Economy	Budget	3	1.33 ± 0.58	0.67 ± 0.58	3.00 ± 0.00	21.78 ± 12.47	1626.33 ± 936.56
	GDP	60	3.13 ± 2.71	2.27 ± 2.50	5.37 ± 4.19	15.63 ± 13.12	6465.23 ± 3028.17
	Reform	1	4.00	0.00	5.00	7.00 ± 1.67	2901.00
	Stock market	15	1.20 ± 0.41	1.00 ± 0.53	3.67 ± 1.07	29.29 ± 35.65	994.40 ± 907.67
	Tax	1	4.00	4.00	5.00	5.20 ± 2.71	11179.00
	Total / Avg	80	2.73 ± 2.47	1.96 ± 2.25	4.95 ± 3.73	17.43 ± 18.61	5272.36 ± 3545.17
Geography	City	509	1.28 ± 0.61	0.35 ± 0.58	4.52 ± 3.42	10.21 ± 10.36	4307.26 ± 4144.97
	Country	87	2.89 ± 1.98	1.36 ± 1.44	6.36 ± 5.86	13.72 ± 21.72	8364.64 ± 5031.71
	Region	460	2.04 ± 2.56	0.92 ± 2.12	5.39 ± 3.13	11.17 ± 12.56	4384.26 ± 4183.70
	Transport	242	1.16 ± 0.46	0.36 ± 0.50	3.94 ± 1.51	10.55 ± 13.41	2534.54 ± 2539.62
	Total / Avg	1298	1.63 ± 1.74	0.62 ± 1.42	4.84 ± 3.35	10.95 ± 13.09	4275.99 ± 4186.91
History	Battle	19	1.89 ± 1.37	0.58 ± 1.12	3.84 ± 2.30	11.67 ± 10.22	8128.32 ± 4793.27
	Dynasty	8	2.12 ± 1.46	0.62 ± 0.74	6.00 ± 3.81	11.17 ± 15.79	2532.62 ± 2023.55
	Other	38	2.29 ± 1.64	0.89 ± 1.06	6.32 ± 5.28	12.05 ± 13.59	6164.74 ± 3820.45
	Total / Avg	65	2.15 ± 1.52	0.77 ± 1.03	5.55 ± 4.57	11.86 ± 13.31	6291.68 ± 4299.70
Politics	Composition of parliament, government	1026	1.13 ± 0.63	0.24 ± 0.48	14.77 ± 14.80	8.35 ± 10.87	594.59 ± 1109.90
	Election results	1382	1.45 ± 1.64	0.72 ± 0.90	14.58 ± 14.81	9.56 ± 14.14	1233.96 ± 1900.37
	Foreign relations	9	1.22 ± 0.67	0.67 ± 0.71	5.89 ± 1.37	26.62 ± 15.67	2969.89 ± 1596.50
	Total / Avg	2420	1.32 ± 1.31	0.52 ± 0.79	14.63 ± 14.79	9.07 ± 12.88	969.34 ± 1647.15
Science	Astronomy	340	3.28 ± 2.41	0.25 ± 0.63	3.71 ± 0.97	8.31 ± 4.87	513.51 ± 457.57
	Biology	73	2.64 ± 1.64	1.11 ± 1.06	4.00 ± 1.57	10.03 ± 7.85	988.05 ± 539.53
	Chemistry	82	1.50 ± 0.95	0.67 ± 0.82	3.91 ± 1.06	6.51 ± 4.48	5215.02 ± 2306.80
	Demography	248	2.63 ± 2.51	1.77 ± 1.59	9.35 ± 7.36	16.37 ± 23.56	2695.69 ± 2716.78
	Total / Avg	743	2.81 ± 2.33	0.89 ± 1.29	5.64 ± 5.07	12.75 ± 18.34	1807.38 ± 2356.32
Sport	Events	319	1.11 ± 0.35	0.51 ± 0.57	10.07 ± 11.64	11.08 ± 17.66	1616.35 ± 1844.19
	Results	52	1.21 ± 0.67	0.62 ± 0.60	10.23 ± 9.49	10.62 ± 22.21	1364.02 ± 2596.41
	Teams	1264	1.11 ± 0.36	0.24 ± 0.44	10.51 ± 7.60	10.50 ± 12.10	1217.02 ± 1760.18
	Total / Avg	1637	1.11 ± 0.37	0.31 ± 0.49	10.42 ± 8.60	10.61 ± 13.73	1299.26 ± 1815.95
Wikimedia	Article	973	3.32 ± 8.74	0.87 ± 2.47	12.61 ± 12.46	16.17 ± 32.29	920.20 ± 1885.78
	Information	9	1.22 ± 0.44	0.33 ± 0.50	8.89 ± 7.23	18.49 ± 21.94	571.78 ± 582.64
	Person	4	2.00 ± 1.41	0.50 ± 0.58	16.50 ± 12.52	15.06 ± 35.96	508.25 ± 276.44
	Timeline	15	3.07 ± 4.15	0.67 ± 0.62	20.13 ± 18.03	19.60 ± 28.09	1278.53 ± 2995.44
	Overview	14	6.57 ± 8.89	5.14 ± 7.87	9.86 ± 14.63	12.16 ± 16.87	3935.00 ± 3021.31
	Total / Avg	1015	3.34 ± 8.64	0.92 ± 2.63	12.66 ± 12.60	16.22 ± 32.05	962.36 ± 1948.86
Total / Avg	-	7258	1.79 ± 3.59	0.60 ± 1.37	10.55 ± 11.48	11.02 ± 18.25	1815.01 ± 2825.16

Table 4: Document Statistics by Topic and Subtopic.

of these datasets are designed to evaluate specific components, such as tables (Herzig et al., 2021; Chen et al., 2020) or charts (Chaudhry et al., 2020; Methani et al., 2020; Masry et al., 2022; Kantharaj et al., 2022; Tanaka et al., 2023a), and are often constrained to single-page documents.

While recent benchmarks have attempted to extend document VQA beyond single-page settings, they still face limitations in terms of cross-page reasoning, domain diversity, and question complexity. For example, MP-DocVQA (Tito et al., 2023), an extension of DocVQA (Mathew et al., 2021), does not include cross-page questions. DUDE (Van Landeghem et al., 2023) introduces a small proportion of cross-page questions but is limited by its reliance on crowd-sourced annotations, which often result in less challenging and rigorous questions, many of which focus on document layout rather than deeper content understanding. Similarly, SlideVQA (Tanaka et al., 2023b) incorporates cross-page questions but is tailored to slide decks, which typically contain lower information density compared to other document types. Doc2SoarGraph (Zhu et al., 2024) includes a few multi-page queries but remains limited to PDFs as its data format.

FinanceBench (Islam et al., 2023) addresses some of these challenges by including long-context documents and practical cross-page questions.

Benchmark	Sources	Origin	# Docs.	# Questions	Cross-page	# Tokens	Answer Type	Evidence Source
MMLongBench-Doc	Multi	Digital + scan	135	1k	✓	21214.1	Abst + Ext.	T, L, F, Ch, M
DUDE	Multi	Digital + scan	5k	41k	✓	1,831.53	Abst. + Ext.	T, L, F, Ch, M
MP-DocVQA	Industry docs	Mostly scans	6k	50k	✗	2026.6	Ext.	T, L, F, Ch
VisualMRC	Web pages	Digital	10k	30k	✗	154.19	Abst.	T, L, F, Ch
InfographicsVQA	Infographics	Digital	5.4k	30k	✗	287.98	Abst + Ext.	T, L, F, Ch, M
TAT-DQA	Finance reports	Digital	2.7k	16k	✗	576.99	Abst. + Ext.	T, L
WikiMixQA	Wikipedia	Digital	$\sim 4k$	1k	✓	1815.01	Abst. + Ext.	T, L, F, Ch, M

Table 5: Comparison between WikiMixQA and existing VQA benchmarks. Evidence Sources are abbreviated as (T)able, (L)ist, (F)igure, (Ch)art, and M(ap). Answer types are Extractive (Ext.) and Abstractive (Abst.)

However, its exclusive focus on financial reports and open-ended answer formats requires expert-level manual evaluation, restricting its applicability to broader domains. Similarly, CRAB (Romanou et al., 2023) focuses on question-answering based on events spanning multiple documents, however, its scope is limited to the domain of causal understanding. More recently, MMLongBench-Doc (Ma et al., 2024) was introduced, incorporating questions from diverse sources, with approximately one-third being cross-page questions. Despite this, the dataset is derived from only 130 documents, limiting its domain coverage and diversity.

A more detailed comparison of existing datasets is presented in Table 5, highlighting the unique contributions of WikiMixQA in advancing research on multi-modal reasoning and document-based question answering.

6 Conclusion

In this paper, we introduce WikiMixQA, a multi-modal visual question-answering (VQA) benchmark designed to assess the long-context document understanding (DU) capabilities of vision-language large models (VLLMs). Our benchmark consists of 1,000 multiple-choice questions (MCQs) that necessitate complex, multi-hop reasoning over visual data, including charts and tables.

We conducted an extensive evaluation of both closed-source and open-source models. Our results indicate that while closed-source models perform well when provided with precisely relevant information, their performance degrades significantly in settings where they must process and extract relevant details from long-context visual data. In contrast, state-of-the-art open-source models exhibit performance close to random, suggesting fundamental challenges in their reasoning capabilities.

These findings highlight that current VLLMs struggle with VQA tasks requiring the extraction and integration of dispersed information from extended contexts. We hope that WikiMixQA serves as a valuable resource for the research community in identifying and addressing the limitations of VLLMs in reasoning and long-context visual understanding.

7 Limitations

WikiMixQA is constructed using Wikipedia as the primary source of documents. Consequently, the generated questions are constrained to Wikipedia-style content, and the benchmark currently covers only seven topics.

A key characteristic of this benchmark is that it includes questions requiring information from multiple modalities. However, this introduces a limitation, as the dataset does not yet support more complex multi-hop reasoning. To address this, we plan to release the full dataset, including the filtered Wikipedia pages, along with the extracted charts and tables. This will enable future research to extend the dataset with more sophisticated question formulations.

Another limitation of this study is the long-context evaluation methodology. Specifically, our evaluation is conducted using image-based inputs (snapshots of Wikipedia pages), without incorporating the textual representations of these pages. Future work could enhance the evaluation by integrating text-based inputs to improve model performance and robustness.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments and constructive feedback. We are also grateful to the members of the LSIR and NLP laboratories at EPFL for their support and helpful input. Additionally, we thank Google for funding and supporting this project.

References

AI Anthropic. 2024. Claude 3.5 sonnet.

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

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. *arXiv preprint arXiv:2103.12011*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022. OpenCQA: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

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

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. Crab: Assessing the strength of causal relationships between real-world events. *arXiv preprint arXiv:2311.04284*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023a. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023b. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan

Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. 2024. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Fengbin Zhu, Chao Wang, Fuli Feng, Zifeng Ren, Moxin Li, and Tat-Seng Chua. 2024. Doc2SoarGraph: Discrete reasoning over visually-rich table-text documents via semantic-oriented hierarchical graphs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5119–5131, Torino, Italia. ELRA and ICCL.

A Custom Wikipedia Category Taxonomy

Category	Subcategory	Instance of (P31)
Politics	Election results	*election*, legislative election, referendum
	Composition of parliament, government	legislative district of *, electoral unit, constituency of the *, United States congressional district, federal electoral district of Canada, parliamentary constituency of *, political party, provincial electoral district of *, ward or electoral division of the United Kingdom, local government areas of *
	Foreign relations	foreign policy, foreign relations
Sport	Results	sports competition, Rugby World Cup qualification, Olympic medal table, championships, UEFA European Championship qualifying
	Teams	association football club, association football team, college sports team, women's national association football team, association football team season, Olympic delegation, * team
	Events	sport season, *Grand Prix, 24 Hours of Le Mans, Tour de France, Summer Olympic Games, Winter Olympic Games, Olympic sporting event, derby, nation at sport competition, multi-sport event, qualification event, qualification for the FIFA World Cup
History	Battle	battle, war, world war
	Dynasty	noble family, * dynasty
	Other	aspect of history
Science	Demography	demographics of country or region, ethnic group
	Astronomy	* eclipse, asteroid, potentially hazardous asteroid, near-Earth object
	Biology	gene, protein, group or class of transmembrane transport proteins, protein family associated with domain, protein family
	Chemistry	chemical element, synthetic element
Geography	City	city or town, big city, municipality of *, independent city, million city, city in the United States, human settlement, largest city, megacity, highly urbanized city
	Country	country population, historical country, country
	Region	geography of geographic location, region of *, regions of *, U.S. region, U.S. state, province of *, county, county of *, township of *, aspect in a geographic region, geographic region
	Transport	rapid transit railway line, commercial traffic aerodrome, international airport, airport, railway station, railway line, airline, rapid transit, railway company, transport company
Economy	GDP	national economy, regional economy
	Stock market	stock market index
	Budget	budget, military budget
	Reform	economic reform
	Tax	tax system
Wikimedia	Article	Wikimedia list article
	Information	Wikimedia information list

Category	Subcategory	Instance of (P31)
	Person	Wikimedia list of persons
	Timeline	Wikimedia timeline
	Overview	Wikipedia overview article

Table 6: Categories and Subcategories with Examples

B WikiMixQAcreation

B.1 Wikipedia article collection

To collect Wikipedia documents containing tables and charts, we leveraged the dataset curated by the WTabHTML project[8]. We downloaded preprocessed English-language dumps extracted from the 2022-03-01 Wikipedia dump. This dataset included 4,291,914 entries across 1,480,422 articles.

Many articles contained very small tables (e.g., one-row sports results, as shown in Figure 6), which were less useful for our purposes. To address this, we filtered out articles with fewer than three tables, resulting in a subset of 392,223 entries.

Next, we focused on articles containing charts to ensure a multimodal dataset. Images were downloaded from the 392,223 documents using the `pyWikiCommons` Python package, resulting in 1,041,062 images. Since most images in Wikipedia are natural images, flags, icons, or buttons rather than charts, we filtered them using a Vision Transformer (ViT) model trained with the DINOv2 method (Oquab et al., 2023) and fine-tuned on ImageNet-1k categories. The model, available on HuggingFace as `facebook/dinov2-base-imagenet1k-1-layer`, classified images into relevant categories. We restricted image formats to JPEG, JPG, and PNG, converting all SVG files to PNG beforehand. Images classified under `web site`, `website`, `internet site`, `site` (∼90%) or `oscilloscope`, `scope`, `cathode-ray oscilloscope`, `CRO` (∼10%, primarily line charts) were retained. Additional rules excluded irrelevant categories like flags and sports-related images based on filenames. Articles with at least one remaining chart were preserved, reducing the dataset to 15,164 entries.

B.2 Promoting diversity

To promote diversity in question generation, we sampled documents by category, aiming for a varied set of question-answer pairs. Each document was labeled with a category using the Wikipedia property *Instance of (P31)*[9], retrieved via the `pywikibot` Python package. Frequently occurring categories included politics and sports, covering documents like election results, parliamentary compositions, and sports team rosters.

To unify categories, we created a custom taxonomy grouping similar *instance of* classes into subcategories (details in Appendix A). For example, documents related to election results were grouped under the *Election results* subcategory of *Politics* using regular expressions (`*election*`). The final taxonomy included seven main categories: *Economy*, *Geography*, *History*, *Politics*, *Science*, *Sport*, and *Wikimedia*. This step reduced the dataset to 7,258 documents, as shown in Table 4.

Balancing Categories The initial classification of charts using the ImageNet model, while efficient, resulted in a number of false positives. To improve chart identification, we utilized GPT-3.5-turbo to analyze image filenames and distinguish likely charts from non-charts (see Appendix D.1). In the most populated subtopics—such as Geography/City, Politics/Composition of Parliament or Government, Politics/Election Results, Science/Astronomy, and Sport/Teams—we excluded only documents that lacked any identified charts, as inferred by GPT-3.5. This refinement step reduced the dataset to 4,292 documents.

Downloading the Final Documents The final set of Wikipedia documents was downloaded in HTML format using the official API[10] in March and April 2024. Each HTML page was converted into a JPG image using the `imgkit` Python package[11]. Due to large heights, images were split into segments of 768 pixels with a 32-pixel overlap using `pillow`[12]. Note that some very long documents were truncated by `imgkit`, resulting in missing content. HTML tables were extracted using the WTabHTML extractor[13]. Additionally, textual data was extracted for reference using the `wikipediaapi` Python package.

B.3 Selection of modality pairs

Wikipedia documents often contain multiple tables and charts. If questions are generated from randomly selected tables and charts, the resulting question set may lack diversity. Furthermore, if the selected tables and/or charts are not semantically relevant, generating meaningful and challenging questions becomes infeasible. To address these issues, we focus on generating questions that involve structured modality pairs: two tables (*table-table*),

[8] Available at https://github.com/phucty/wtabhtml

[9] See documentation at https://www.wikidata.org/wiki/Property:P31

[10] https://en.wikipedia.org/w/api.php

[11] https://github.com/jarrekk/imgkit

[12] https://github.com/python-pillow/Pillow

[13] https://github.com/phucty/wtabhtml

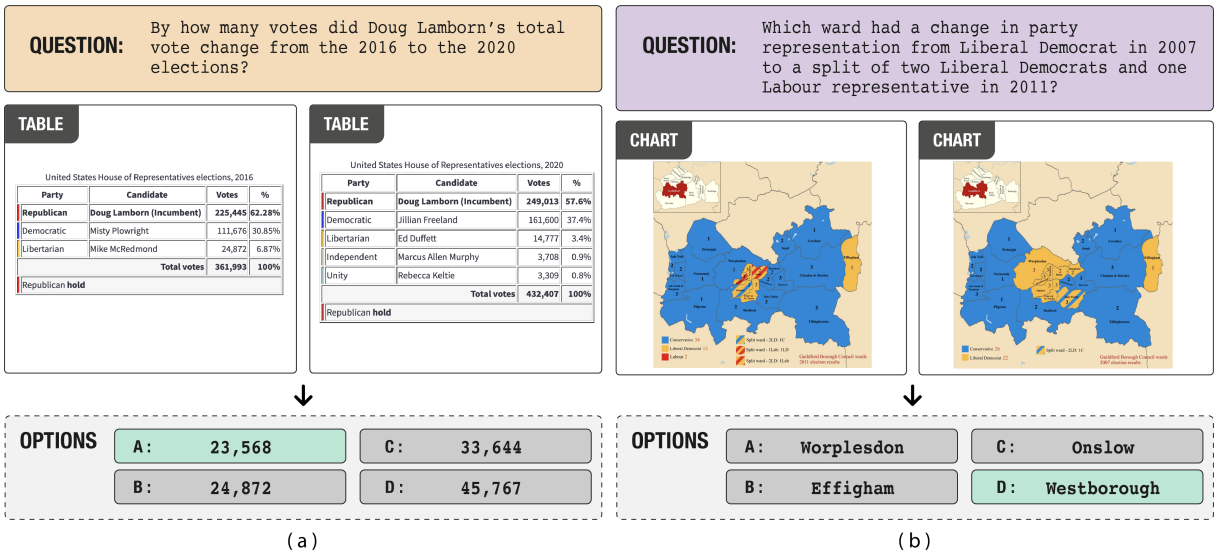


Figure 5: Examples from `WikiMixQA` illustrating (a) a question whose answer relies on the information presented in two tables and (b) a question whose answer relies on the information presented in two charts.

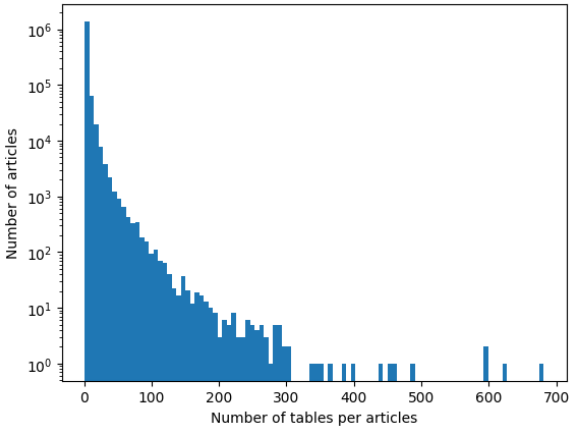


Figure 6: Distribution of tables in selected Wikipedia pages.

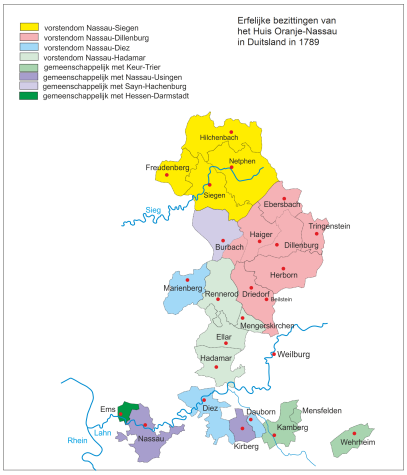
one table and one chart (*table-chart*), or two charts (*chart-chart*). To mitigate the selection of irrelevant pairs, we rely on the similarities in the textual descriptions between the modalities.

Table description While some HTML tables include captions, the majority do not. To address this, we used an open-source large language model to generate descriptions for all tables, using the raw HTML of the tables as input. We chose a medium-sized model, `Llama-3-8B-Instruct`, to minimize costs, as the prompt for some tables can be quite large. The details of the prompt can be found in Appendix E.

Image description To generate descriptions for images identified as potential charts, we used a vision-language model, `GPT-4-turbo`. These images were previously identified as potential charts by the ImageNet classifier (see Appendix B.1). We prompted the model with the image content, asking it to determine whether the image is a chart, and if so, to specify the type of chart (data chart or map, according to Wikipedia's definitions). Finally, the model was asked to extract the most relevant information from the chart in fewer than 200 words, which serves as the image description. The full prompt can be found in Appendix D.2.

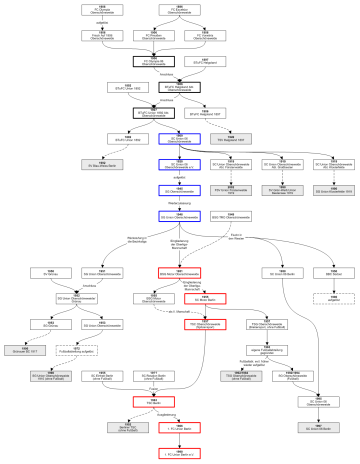
C Chart Examples

We define a chart as a graphical representation of data, including: (a) data charts such as diagrams or graphs that organize and display numerical or qualitative information; (b) maps enhanced with additional data; and (c) other domain-specific constructs, such as chord charts or record charts. Figure 7 shows some chart examples from different topics and subtopics.

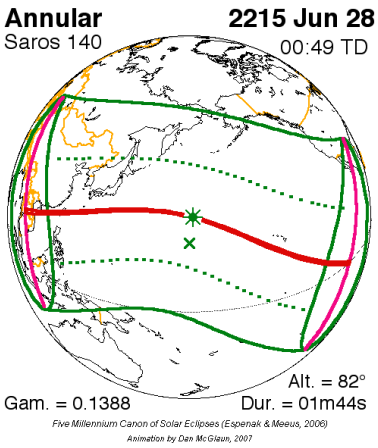


(a) History (Dynasty)

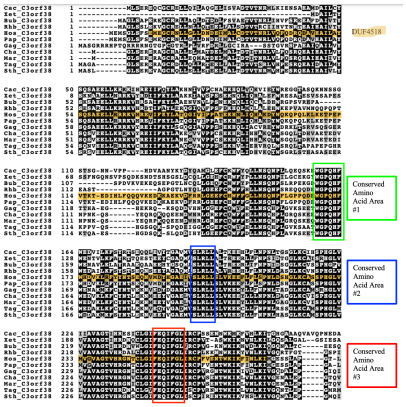
(b) Geography (Region)



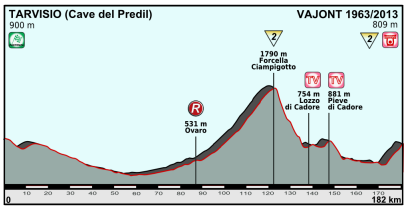
(c) Sport (Teams)



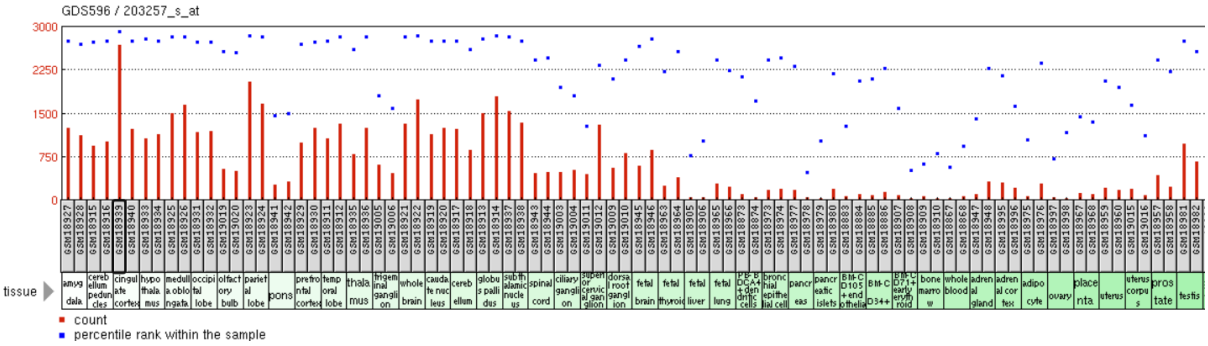
(d) Science (Astronomy)



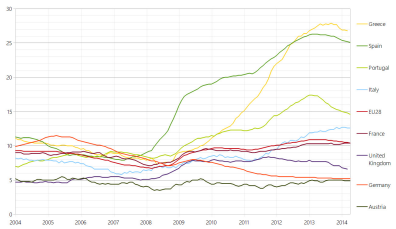
(e) Science (Biology)



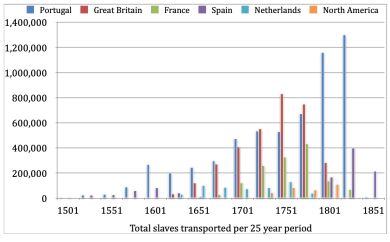
(f) Wikimedia (Article)



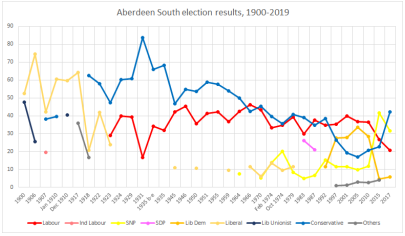
(g) Science (Biology)



(h) Economy (GDP)



(i) History (Other)



(j) Politics (Composition of parliament, government)

Figure 7: Chart examples from different topics and subtopics.

D GPT prompts for chart selection

D.1 GPT-3.5 Prompt

We used the GPT-3.5-turbo to predict whether an image is likely to represent a chart based on the filename.

```
You are an image filename reader and you
    should guess whether the filename
    describes an image representing a
    chart or not. You should answer by
    yes or no in the JSON format.
```

D.2 GPT-4-Turbo Prompt

We prompted GPT-4 by providing the model the definition from Wikipedia article about "Chart": https://en.wikipedia.org/wiki/Chart which identifies three types of charts (data chart, maps and other).

```
Is the image a chart?
A chart (sometimes known as a graph) is
    a graphical representation for data
    visualization, in which "the data is
    represented by symbols, such as
    bars in a bar chart, lines in a line
    chart, or slices in a pie chart". A
    chart can represent tabular numeric
    data, functions or some kinds of
    quality structure and provides
    different info.
If yes, please specify the type of chart
    between data chart, maps or other.
A data chart is a type of diagram or
    graph, that organizes and represents
    a set of numerical or qualitative
    data.
Maps that are adorned with extra
    information (map surround) for a
    specific purpose are often known as
    charts, such as a nautical chart or
    aeronautical chart, typically spread
    over several map sheets.
Then extract the most relevant
    information from the chart in less
    than 200 words.
You are communicating with an API, not a
    user. Begin all AI responses with
    the character '{' to produce valid
    JSON. Here is an example:
{
   "chart": "yes",
   "type": "data chart",
   "data": "information extracted from
       the image"
}
```

E Llama3 prompt for table selection

We used `Llama3-8B-Instruct` with `torchtune`[14] to get a textual description of each HTML table. The following prompt has been used:

```
You are a table-to-text assistant. I'll
    give you a table in HTML format,
    please write a textual description
    of the table content."
```

[14] Available at https://github.com/pytorch/torchtune.

F GPT-4-turbo prompt for question generation

Prompt used with `GPT-4-turbo` model for generating the questions with one chart and one HTML table as modalities:

```
You are given one chart and one table.
    First design a multiple-choice
    question based on the given chart (
    Q1). Then design a multiple-choice
    question based on the given table (
    Q2). Finally, combine the
    information from both chart and
    table to create a multiple-choice
    question that can be answered ONLY
    by combining the information from
    all the given charts and tables (Q3)
    .
Each question should have 4 options, one
     of which is the correct answer.
    Please explain how to reach the
    correct answer from the given
    context.
You are communicating with an API, not a
     user. Begin all AI responses with
    the character '{' to produce valid
    JSON. Here is an example:
{
    "Q1":{
    "Question": "<question>",
     "A": "<option1> ",
     "B": "<option2>",
     "C": "<option3>",
     "D": "<option4>",
     "Answer": "<correct_option>",
     "Explanation": "<explanation>"
     },
     "Q2":{
    "Question": "<question>",
     "A": "<option1> ",
     "B": "<option2>",
     "C": "<option3>",
     "D": "<option4>",
     "Answer": "<correct_option>",
     "Explanation": "<explanation>"
     },
     "Q3":{
     "Question": "<question>",
     "A": "<option1> ",
     "B": "<option2>",
     "C": "<option3>",
     "D": "<option4>",
     "Answer": "<correct_option>",
     "Explanation": "<explanation>"
     }
 }
Tables:
{html_table}
{table description if it is available}

<chart image>
```

We modified the first paragraph of the prompt and the modalities attached in the end accordingly when two charts or two tables were provided.

G InternVL2 prompt for question-answer pair evaluation

Prompt used with `InternVL2` model for evaluating the generated question with one chart and one HTML table as modalities:

```
Image:
<image>

HTML table:
{html_table}

Given the information provided in the
    image and HTML table, can you answer
     to the following question: {
    generated_question}
Answer only by yes or no.
```

We modified the start of the prompt accordingly when two charts or two tables were provided.

H Answer Generation Prompts

Oracle **Setting:** Prompt used in answer-generation step, where we task models to answer a MCQ given provided chart and table (*table-chart*):

```
Given the following chart and table (in
    HTML format), which of the following
     answer is correct? A, B, C or D.
    Please answer in the format A, B, C
    or D.

Chart:
<image>

HTML table:
{html_table}
```

The prompt used in answer-generation step, where we task models to answer a MCQ given two provided tables (*table-table*):

```
Given the following chart and table (in
    HTML format), which of the following
     answer is correct? A, B, C or D.
    Please answer in the format A, B, C
    or D.

HTML table_1:
{html_table}

HTML table_2:
{html_table}
```

Prompt used in answer-generation step, where we task models to answer a MCQ given two provided charts (*chart-chart*):

```
Given the following chart and table (in
    HTML format), which of the following
     answer is correct? A, B, C or D.
    Please answer in the format A, B, C
    or D.

Chart_1:
<image>

Chart_2:
<image>
```

Wikidoc **Setting:** The prompt used in answer-generation step, where we task models to answer a MCQ snapshots of the relavant Wikipedia page:

```
Given the following document, which
    includes tables and charts, which of
     the following answer is correct? A,
     B, C or D.
Extract and analyze the relevant data
    from the document to select the
    correct answer. If the document does
     not contain enough information,
    infer the most plausible answer or
    state 'Unable to determine'.
Please answer in the format A, B, C, D,
    or 'Unable to determine'.

Image:
<image>
.
.
.
Image:
<image>
```

Blind **Setting:** The prompt used in answer-generation step, where we task models to answer a MCQ given NO ccontextual information:

```
Which of the following answer is correct
    ? A, B, C or D. Please answer in the
     format A, B, C or D.
            If the chart or table is
                unavailable, please
                infer the most plausible
                 answer based on general
                 reasoning or state '
                Unable to determine' if
                no inference is possible
                .
            Please answer in the format
                A, B, C, D, or 'Unable
                to determine'.
```

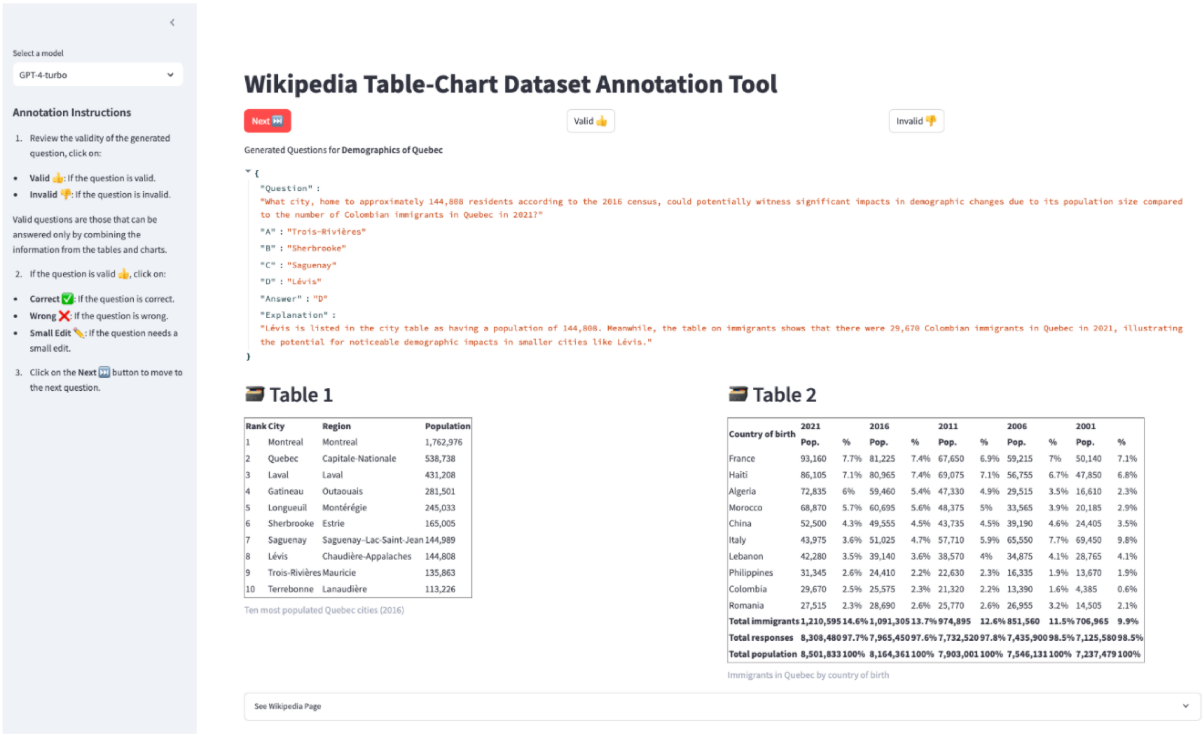


Figure 8: Interface of the annotation tool used for human curation.