Chart-MRAG: Benchmarking Multimodal Retrieval Augmented Generation on Chart-based Documents

Yuming Yang[1], Jiang Zhong[1*], Li Jin[2*], Xiao Sun[1], Jingwang Huang[1], Jingpeng Gao[1], Qing Liu[2], Yang Bai[2], Jingyuan Zhang[3], Rui Jiang[1], Qin Lei[4], Kaiwen Wei[1*]

[1]College of Computer Science, Chongqing University

[2]Aerospace Information Research Institute, Chinese Academy of Sciences, China

[3]Kuaishou Technology, Beijing, China

[4]The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

ymyang@cqu.edu.cn, zhongjiang@cqu.edu.cn, weikaiwen@cqu.edu.cn

Abstract

Multimodal Retrieval-Augmented Generation (MRAG) enhances reasoning capabilities by integrating external knowledge. However, existing benchmarks primarily focus on simple image-text interactions, overlooking complex visual formats like charts that are prevalent in real-world applications. In this work, we introduce a novel task, **Chart-based MRAG**, to address this limitation. To generate high-quality evaluation samples, we propose **CHARGE** (**CHAR**t-based document question-answering **GE**neration), a semi-automatic framework for generating evaluation samples through multimodal keypoint extraction, knowledge graph construction, and qa pair synthesis. By combining CHARGE with expert validation, we construct **Chart-MRAG Bench**, a comprehensive benchmark for chart-based MRAG evaluation, featuring 4,738 question-answering pairs across 8 domains from real-world documents. Our experiments reveal three critical limitations in current approaches: (1) unified multimodal embedding retrieval methods struggles in chart-based scenarios, (2) even with ground-truth retrieval, state-of-the-art Multimodal Large Language Models (MLLMs) achieve only 71.15% Correctness and 80.74% Coverage scores, and (3) Widely-used MLLMs demonstrate consistent text-over-visual modality bias. These findings highlight great challenges in processing information-dense visual formats. The dataset and code are available at Chart-MRAG.

1 Introduction

Multimodal retrieval-augmented generation (MRAG) (Zhao et al., 2023) enhances multimodal reasoning by retrieving relevant external knowledge, and leveraging multimodal large language models (MLLMs) for informed response generation (OpenAI, 2023; Zhang et al., 2024a). This approach substantially mitigates hallucinations and improves factual grounding (Gao et al., 2023).

*Corresponding authors.

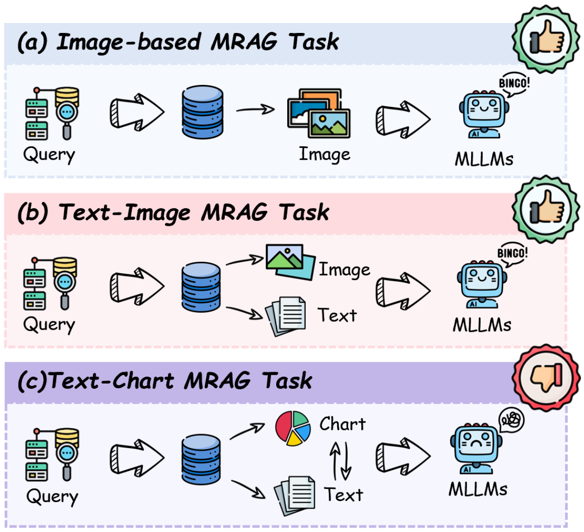


Figure 1: Comparison of two common MRAG scenarios, image-based and text-image, and the proposed text-chart task. In the text-chart MRAG scenario, models need to capture intricate chart details and retrieve both chart and text information to generate correct answers.

Effectively evaluating MRAG systems requires high-quality benchmarks that assess both retrieval and generation. Existing benchmarks such as MRAG-Bench (Hu et al., 2024) and Dyn-VQA (Li et al., 2024b) have made strides in assessing MRAG capabilities through manually curated question-answering (QA) pairs. However, as illustrated in Fig.1(a) and (b), these benchmarks primarily focus on scenarios involving images or simple combinations of images and text. Such settings fail to capture the complex interactions between visual details and corresponding text, particularly when dealing dense and structured information like charts, which are widely used in real-world applications (Masry et al., 2022). This leaves a critical gap in MRAG evaluation (detailed discussion refer to Appendix 6).

To bridge this gap, we propose a new task: **Chart-based MRAG**. For a given text query, this

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25392–25445

task involves three RAG sub-tasks: (1) *Text-Chart MRAG*, as illustrated in Fig. 1(c), both textual and chart data must be jointly retrieved to generate correct answers. In addition, to allow for the separate evaluation of each modality's contributions, it also provides (2) *Text-only RAG*, where answers can only be found in textual information; and (3) *Chart-only MRAG*, where answers depend exclusively on chart data. To comprehensively evaluate these tasks, a major challenge is how to semi-automatically generate high-quality QA pairs that accurately capture text-chart interactions.

To overcome this challenge, we propose **CHAR**t-based document question-answering **GE**neration (CHARGE), a framework for automatically generating QA pairs from real-world chart-document data. CHARGE follows a three-stage pipeline: it begins with multi-modal keypoint extraction from both text and charts, then constructs a keypoints knowledge graph, and finally generates question-answer pairs to model complex multimodal interactions. Moreover, to further challenge the chart-based MRAG task, MLLMs are employed to generate QA pairs that require multi-hop reasoning based on intra-document or inter-document retrieval.

Building on CHARGE, we introduce **Chart-MRAG Bench**, a high-quality, human-checked benchmark tailored for Chart-based MRAG. With CHARGE, 5,866 qualified QA pairs were initially generated, after that, 4,738 (nearly 80%) were meticulously selected through expert evaluation based on clarity, accuracy, multimodal coherence, and ethical considerations. As shown in Table 1, Chart-MRAG Bench comprises 267 documents spanning 8 domains, 8 types of questions, 1,283 paragraphs, and 627 charts, capturing complex cross-modal interactions in realistic scenarios.

We conducted a systematic evaluation of mainstream retrieval methods and MLLMs on Chart-MRAG Bench. In our evaluation, keypoint-based Correctness and Coverage metrics were introduced to rigorously assess accuracy and comprehensiveness. The results reveal that unified multimodal embedding retrieval methods, which rely on a single vector store, perform poorly in high-density chart scenarios. Furthermore, even with ground-truth retrieval, the best-performing Claude-4.5-Sonnet (Team et al., 2024) only achieved 71.15 Correctness and 80.74 Coverage metrics, highlighting persistent challenges in text-chart multimodal reasoning. In summary, the contributions of this paper are:

1) We present Chart-based MRAG, the first extension of MRAG to chart scenarios that introduces a new dimension for evaluating cross-modal reasoning in information-dense visual contexts.

Table 1: Comparison between existing MRAG benchmarks and the proposed Chart-MRAG Bench.

Benchmark	MRAG Type	QA Pairs	Includes Chart	Expert Annot.
M²RAG (Ma et al., 2024c)	Image	0.8k	✗	✗
MultiTableQA (Zou et al., 2025)	Table	2.6k	✗	✗
Dyn-VQA (Li et al., 2024b)	Image	1.5k	✗	✓
MMSearch (Jiang et al., 2025)	Image	0.3k	✗	✓
ViDoSeek (Wang et al., 2025)	Image	1.1k	✓	✓
MRAG-Bench (Hu et al., 2024)	Image	1.4k	✗	✓
SSMQG (Wu et al., 2024)	Image	1.0k	✗	✗
Chart-MRAG *(Ours)*	Chart	4.7k	✓	✓

2) We propose CHARGE, an automated framework for generating QA pairs in real-world scenarios through a structured pipeline.

3) We establish Chart-MRAG Bench based on CHARGE. It is a human-verified benchmark for chart-based MRAG, covering 8 scenarios, 8 question types, and 4,738 QA pairs, with a subset designed for multi-hop reasoning.

4) We introduce two robust evaluation metrics to assess MRAG quality. Extensive experiments highlight the limitations of existing retrieval and generation methods in chart-centric tasks.

2 Related Work

Multimodal RAG Methods. Recent advances in Retrieval-Augmented Generation (RAG) (Izacard et al., 2022; Zhang et al., 2024b; Jia et al., 2025; Zhu et al., 2025; Wei et al., 2025a, 2026, 2025b) have successfully extended to multimodal domains (Chen et al., 2022; Zhao et al., 2023, 2024; Li et al., 2026b,a; Zhang et al., 2026), enabling cross-modal tasks through MLLMs (Yao et al., 2024; Team, 2024). While researchers have proposed various approaches (Ma et al., 2024a; Faysse et al., 2024; Yu et al., 2024; Zou et al., 2025; Wang et al., 2025) for cross-modal retrieval, current evaluation methodologies predominantly rely on Visual Question Answering (VQA) datasets (Marino et al., 2019; Talmor et al., 2021; Schwenk et al., 2022; Masry et al., 2022). These evaluations fall short in addressing retrieval-specific challenges.

Multimodal RAG Benchmarks. The effectiveness of MRAG systems necessitates comprehensive evaluation benchmarks. While several benchmarks (Hu et al., 2024; Li et al., 2024b; Zhou et al., 2024) explore vision-based retrieval for question

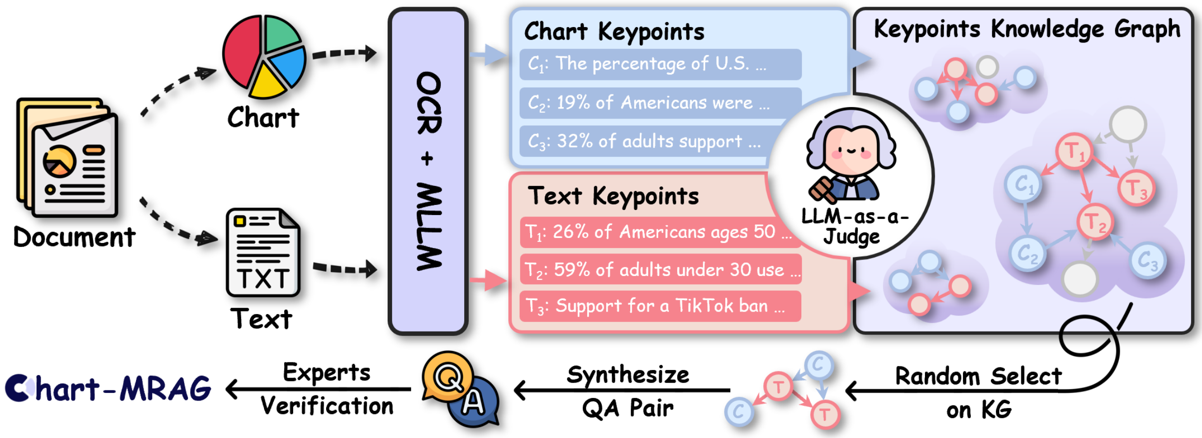


Figure 2: The proposed CHARGE framework for creating QA pairs from document-chart data, consisting of three steps: (1) Extracting multimodal keypoints from both textual and visual content, (2)Constructing a knowledge graph from the extracted keypoints, and (3) Generating diverse QA pairs by performing random selection on the graph.

answering through manual annotation, they neglect the critical dimension of cross-modal collaborative generation. Some studies (Dong et al., 2025; Ma et al., 2024b; Ding et al., 2024) consider hybrid modality retrieval, yet they primarily rely on manual question-answering. Furthermore, although some studies (Es et al., 2023; Mathew et al., 2021; Li et al., 2024a; Wu et al., 2024) have investigated automated processes for generating cross-modal QA pairs, their scope focus on simplistic natural images with singular subjects, the chart-based scenarios largely unexplored. To bridge this gap, this paper introduce Chart-MRAG Bench. Table 1 illustrates the differences between existing MRAG benchmarks and Chart-MRAG Bench.

3 CHARGE Framework

We present CHARGE, a framework for generating multimodal multi-hop QA pairs from chart-based documents. CHARGE operates in three stages: (1) extracting multimodal keypoints from both textual and visual content, (2) constructing a knowledge graph from the extracted keypoints, and (3) generating diverse QA pairs by performing random selects on the graph to enable multi-hop reasoning.

3.1 Multimodal Keypoint Extraction

CHARGE initially processes textual content and charts into keypoints. These keypoints consist of information units that capture factual statements, logical inferences, or conclusive summaries (e.g., "33% of U.S. adults say they use TikTok").

The processing module combines OCR and MLLM to capture both textual information and chart content. For textual information, we utilize GPT-4o to extract keypoints $T = \{T_1, ..., T_m\}$ from paragraphs, as shown in Fig 34. For chart content, we employ a two-step approach: first extracting numerical values using *ChartOCR* (Luo et al., 2021), then applying GPT-4o to structure the extracted values into keypoints $C = \{C_1, ..., C_n\}$, ensuring both contextual comprehension and numerical precision, as detailed in Fig 35. The complete workflow is presented in Appendix .12.

3.2 Knowledge Graph Construction

Following keypoint extraction, CHARGE constructs a knowledge graph to deeply model information relationships across documents and modalities. Specifically, we transform individual keypoints into structured representations by extracting entities, relationships, and attributes from each keypoint.

During graph construction, we verify keypoints whose semantic similarity exceeds a threshold τ. Concretely, if two keypoints have a semantic similarity above τ, *GPT-4o* is invoked to determine whether they are semantically redundant. If both keypoints convey identical information, they are discarded and excluded from the candidate set for QA generation (as illustrated by the gray nodes in Fig.2). This verification mechanism ensures the uniqueness of keypoint sources in the graph, thereby maintaining answer consistency in subsequent retrieval-augmented generation.

After building the keypoint knowledge graph, we further partition the keypoints into communities to organize information by topic. Identifying semantically related keypoint communities enhances the topical relevance and logical coherence of the

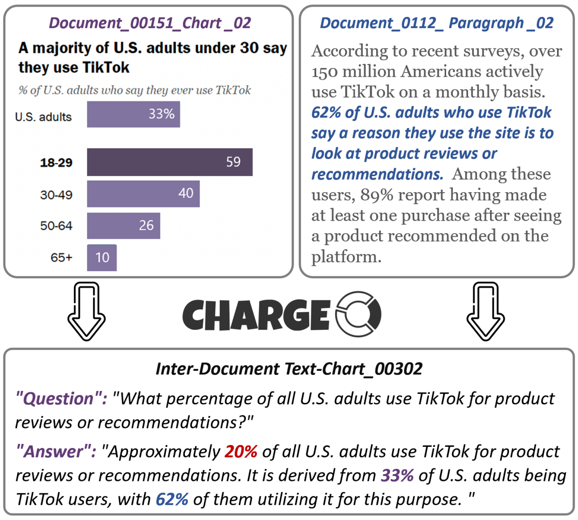


Figure 3: An inter-document multi-hop QA example from Chart-MRAG Bench, generated by CHARGE.

generated high-quality question-answer pairs.

3.3 Question-Answer Pair Generation

CHARGE employs a random select approach on the knowledge graph for generating question-answer pairs. The process begins by randomly selecting a community to determine the topical theme, followed by a four-step generation procedure:

Starting Point Selection: A node K_{Start} (e.g., "TikTok") is randomly chosen from the selected community as the starting point.

Related Node Retrieval: The top-n nodes semantically related to K_{Start} are retrieved to form a candidate pool $\{K_1, K_2, K_3, ..., K_n\}$, maintaining thematic and entity consistency to ensure informational coherence in QA generation.

Node Pair Matching: The candidate nodes are processed by an LLM, which analyzes their relationships and selects the most appropriate pair $\{K_i, K_j\}$ for joint QA generation.

High-quality QA Generation: A specialized LLM agent generates precise question-answer pairs QA_{ij} that strictly based on $\{K_i, K_j\}$.

To ensure QA_{ij} exclusively depends on both $\{K_i, K_j\}$, we implement a rigorous verification protocol using an LLM-as-a-Judge. The QA pair is retained only if it satisfies the following criteria: correct when provided with both $\{K_i, K_j\}$, but incorrect under direct answering or when given only K_i or K_j individually. The complete procedure for QA generation is outlined in Algorithm 1.

We categorize the generated QA pairs by **Document Source** and **Modality**. For **Document Source**, categories include *Intra-Document* (keypoints $\{K_i, K_j\}$ from the same document) and *Inter-Document* (keypoints from different documents). For **Modality**, categories include *Text-only* (keypoints $\{T_i, T_j\}$ from text), *Chart-only* (keypoints $\{C_i, C_j\}$ from charts), and *Text-Chart* (keypoints $\{C_i, T_j\}$ from mixed modalities).

For example, as illustrated in Fig 3, CHARGE generates an Inter-Document Text-Chart QA_{ij}: "What percentage of all U.S. adults use TikTok for product reviews or recommendations?", which requires integrating C_i (from chart): "33% of U.S. adults say they use TikTok" with T_j (from text): "62% of U.S. adults who use TikTok say a reason they use the site is to look at product reviews or recommendations".

Furthermore, to cover varying difficulty levels, CHARGE also supports single-keypoint QA generation, denoted as QA_i derived from *Text-only* keypoint T_i or *Chart-only* keypoint C_i. These QA pairs are constructed based on individual keypoints and undergo the same LLM-as-a-Judge verification process. We categorize these as *Single-Point Text-only* and *Single-Point Chart-only* QA pairs.

Algorithm 1: QA Generation via CHARGE

Input : Knowledge graph $G = (V, E)$; Community set $\mathcal{C} = \{C_1, C_2, ..., C_m\}$
Output : Question-Answer pair (q, a)
`// Step 1: Random Starting Point Selection`
$C \leftarrow$ Randomly select one community from $\mathcal{C}$
$K_{start} \leftarrow$ Randomly select one keypoint from C
`// Step 2: Related Node Retrieval`
$S \leftarrow$ Retrieve(K_{start}, V, n)
$K_{candidates} \leftarrow \{K_1, K_2, ..., K_n\}$ from S
`// Step 3: Node Pair Matching`
$\{K_i, K_j\} \leftarrow \mathrm{LLM}_{\mathrm{pair}}(K_{candidates})$
`// Step 4: High-quality QA Generation`
$(q, a) \leftarrow \mathrm{LLM}_{\mathrm{qa}}(K_i, K_j)$
return (q, a)

4 Chart-MRAG Bench

By utilizing the CHARGE framework, we generated an initial pool of question-answer pairs. These pairs underwent rigorous expert evaluation to ensure high quality, culminating in the Chart-MRAG Bench. This process was guided by 4 principles:

Authenticity and Diversity. The benchmark is based on real-world data collected from the official website[1], a trusted source of high-quality social research. We collected data from September 2023 to September 2024, encompassing 267

[1]www.pewresearch.org

Statistics	Reasoning Step	Number
- Single-Point Text-only	1-hop	499 (10.53%)
- Single-Point Chart-only	1-hop	763 (16.10%)
- Intra-Document Text-only	2-hop	666 (14.06%)
- Intra-Document Chart-only	2-hop	587 (12.39%)
- Intra-Document Text-Chart	2-hop	746 (15.74%)
- Inter-Document Text-only	2-hop	547 (11.54%)
- Inter-Document Chart-only	2-hop	472 (9.96%)
- Inter-Document Text-Chart	2-hop	458 (9.67%)

Table 2: Chart-MRAG question types, spanning complexities and multi-modalities, are designed by selecting keypoints from diverse sources and modalities.

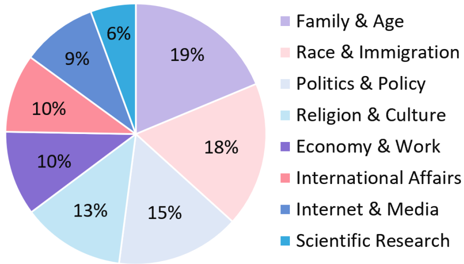


Figure 4: Distribution of Chart-MRAG Bench across 8 domains by controlling the community topics in the knowledge graph, representing key areas of real-world.

documents containing 1,283 text passages and 627 charts. As illustrated in Table 2 and Fig 4, Chart-MRAG Bench encompasses 8 distinct domains, integrating over 10 chart types and 8 QA types.

Annotation Reliability. We engaged 12 expert annotators with Master's degrees. All annotators were proficient in English, with an average TOEFL score of 92 or equivalent language proficiency. The annotation process took 34 working days to complete. Our annotation protocol involved three independent reviewers evaluating each sample, achieving a Fleiss's kappa (Fleiss and Cohen, 1973) of 0.82, indicating substantial inter-annotator agreement.

Rigorous Quality Control. Through meticulous manual review, we refined the dataset from 9,600 initial candidates to 5,866 validated pairs by systematically eliminating 2,631 samples with OCR errors and 1,103 redundant samples. A consensus-based sampling strategy required validation from at least two reviewers, resulting in 4,738 high-quality samples (nearly 80% of the validated pairs).

High Information Complexity. Statistical analysis reveals the benchmark's sophistication: approximately 70% of charts contain more than 8 critical information points (mean: 13.87), and over 73% of text passages include more than 6 keypoints (mean: 8.31). This information-rich environment rigorously evaluates models' capacity to process intricate and dense data representations.

For illustrative examples of Chart-MRAG Bench question-answer pairs across different domains and reasoning types, please refer to Appendix 6.

5 Experiments

5.1 Baselines and Evaluation Metrics

We conduct comprehensive evaluations using 3 distinct retrieval methods and 8 diverse MLLMs. Including **Multimodal Retrievers**: CLIP (Radford et al., 2021), JINA (Koukounas et al., 2024), SigLIP (Zhai et al., 2023), BGE-M3-base/large (Chen et al., 2024) and E5-base/large (Wang et al., 2022). And **Backbone MLLMs**: GPT-4o (version 2024-11-20) (Radford et al., 2021), GPT-5 (version 2025-08-07)(OpenAI, 2025), Gemini-2.5-Pro (version preview-06-05)(Comanici et al., 2025), Claude-4.5-Sonnet (version 2025-05-14) (Anthropic, 2024), SAIL-VL-2B (Team, 2024), Qwen2-VL-7B-instruct (Wang et al., 2024), MiniCPM-V-2.6 (8B) (Yao et al., 2024), and Llama-3.2-90B-Vision (Dubey et al., 2024).

Following (Wu et al., 2024), we evaluate multimodal retrieval models using Recall@5 (R@5) and Recall@10 (R@10). Please refer to Appendix .9 and Appendix .10 for details of the retrieval setup and metrics. Moreover, since chart-based MRAG is a newly proposed task, existing evaluation metrics are inadequate. Therefore, we introduce Correctness and Coverage metrics to assess the quality of responses. The rationale behind these new metrics and their fairness evaluation are thoroughly discussed in Appendix .11 and Appendix .12.

Correctness. It measures the exact match between response and ground truth keypoints. Given a question-answer pair $\{Q, A, K^{gt}\}$ with ground truth keypoints $K^{gt} = \{k_1^{gt}, ..., k_n^{gt}\}$, we extract keypoints $K^r = \{k_1^r, ..., k_m^r\}$ from the model's response using an LLM. The score is defined as:

$$\text{Correctness}(K^r, K^{gt}) = \mathbb{1}[K^r \equiv K^{gt}], \quad (1)$$

where $K^r \equiv K^{gt}$ implies complete keypoint matching and equal cardinality. This binary metric requires perfect accuracy, with zero tolerance for missing information or errors.

Coverage. It quantifies the proportion of correctly captured ground truth keypoints:

$$\text{Coverage}(K^r, K^{gt}) = \frac{|K^m|}{|K^{gt}|}, \quad (2)$$

Model	Overall		Single-Point				Intra-Document						Inter-Document					
			Text-only		Chart-only		Text-only		Chart-only		Text-Chart		Text-only		Chart-only		Text-Chart	
	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10
Method 1: Unified Multimodal Embedding and Single Vector Store																		
SigLIP	11.69	15.95	50.00	57.07	0.00	0.00	19.63	30.74	0.00	0.00	0.00	0.00	16.20	27.31	0.00	0.00	0.00	0.00
CLIP	13.26	19.07	56.06	65.15	0.00	0.00	24.44	42.22	0.00	0.00	0.00	0.00	16.20	28.70	0.00	0.00	0.00	0.00
JINA	23.14	29.02	77.78	85.35	0.00	0.00	47.04	63.70	0.00	0.00	0.00	0.00	41.20	56.94	0.00	0.00	0.00	0.00
Method 2: Multimodal Embeddings and Combined Vector Stores (Caption generated by GPT-4-Vision)																		
BGE-M3-base	22.89	31.21	39.90	47.47	52.24	62.09	9.63	18.52	17.01	31.29	9.09	15.51	9.72	15.28	13.43	19.40	4.46	11.61
BM25	27.02	36.46	51.01	54.55	52.24	63.88	16.67	26.67	12.93	25.17	7.49	19.79	23.15	31.94	10.45	16.42	12.50	21.43
BGE-M3-large	27.64	39.52	64.65	70.71	43.58	59.70	29.26	42.96	10.20	17.69	8.02	18.72	18.98	32.41	5.97	14.93	8.93	22.32
E5-base	35.27	47.59	67.17	73.74	66.27	80.90	21.48	34.81	23.13	47.62	15.51	25.13	20.37	27.78	20.90	35.07	14.29	23.21
E5-large	41.53	59.54	72.73	79.80	64.78	79.40	38.89	60.74	23.13	48.30	18.18	41.71	35.65	53.24	20.90	41.04	22.32	40.18
Method 3: Multimodal Embeddings and Separate Vector Stores																		
JINA + BM25	23.83	33.90	48.48	53.03	45.67	59.10	11.85	20.37	14.97	28.57	9.63	18.72	15.74	26.39	7.46	19.40	14.29	21.43
CLIP + BGE-M3-base	25.64	36.09	34.34	41.92	66.57	77.61	5.93	12.96	24.49	51.70	10.70	19.25	6.48	12.04	14.93	35.07	11.61	12.50
CLIP + BGE-M3-large	33.40	46.97	57.58	68.18	66.57	77.61	20.74	34.07	24.49	51.70	19.79	32.09	13.89	24.07	14.93	35.07	16.07	25.89
SigLIP + E5-base	37.96	52.47	64.65	69.70	84.18	93.73	15.56	29.63	39.46	74.15	17.11	28.34	13.89	24.54	14.18	44.78	14.29	28.57
SigLIP + E5-large	42.53	61.10	68.69	75.76	84.18	93.73	25.19	47.41	39.46	74.15	24.06	41.71	21.76	43.06	14.18	44.78	22.32	40.18

Table 3: Performance Comparison of Different Multimodal Retrieval Models (%) on Chart-MRAG benchmark, evaluating three strategies: Unified Multimodal Embedding with Single Vector Store, Multimodal Embeddings with Combined Vector Stores, and Multimodal Embeddings with Separate Vector Stores (best scores highlighted in blue).

where K^m represents matched GT keypoints. This metric in [0,1] enables granular evaluation.

To compute the Correctness and Coverage scores in a robust and automated manner, we employ a jury of multiple advanced LLMs (GPT-4.1, Qwen2.5-Max, Grok-3, and Claude-4.5-Sonnet), where each model independently scores every response and the final score is derived by averaging their individual ratings to enhance metric stability.

Table 3 reveals significant challenges in multimodal retrieval. While existing retrievers exhibit strong single-modal performance (JINA-CLIP achieves 77.78% Recall@5 in text-only questions and SigLIP + E5 reaches 84.18% Recall@5 in chart-only tasks), Inter-Document Text-Chart questions yielded only 22.32% retrieval accuracy. The key findings demonstrate that storing and retrieving charts and text separately in the database substantially improves performance, achieving recall rates of 42.53% and 61.10% at k=5 and k=10.

Unified multimodal embeddings fail in knowledge-intensive scenarios. While Method 1 outperforms all other approaches in pure text-only QA, it achieves zero recall (0.00%) in chart-only QA and Text-Chart QA tasks. This phenomenon reveals a critical limitation: current unified multimodal embedding models excel at representing knowledge-sparse content (e.g., identifying a dog in an image) but struggle with knowledge-intensive scenarios (e.g., retrieving specific numerical values from charts in a multimodal repository).

5.2 Retrieval Performance Comparison

Chart captioning enables simple yet effective multimodal retrieval. Methods 2 and 3 achieve comparable performance (Recall@5: 41.53% vs 42.53%), with differences primarily in chart re-

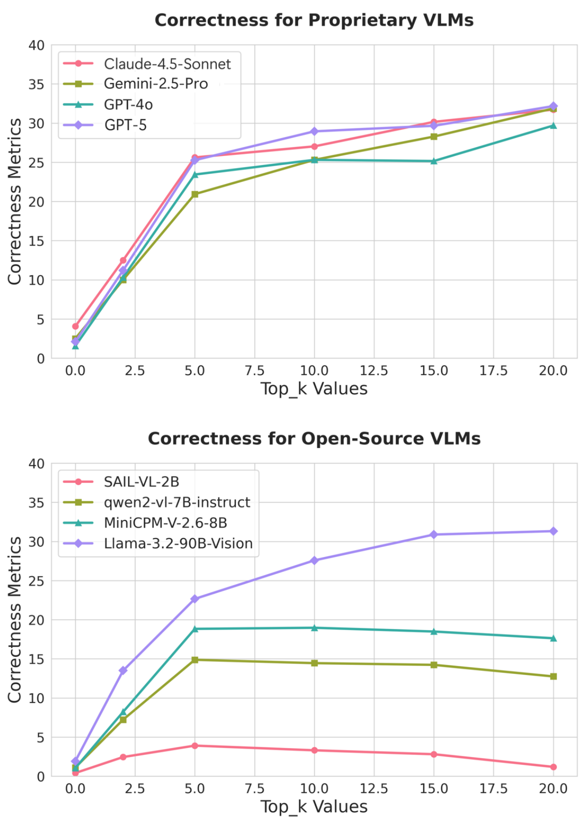


Figure 5: Trade-off analysis between retrieval coverage and answer accuracy across different k settings, illustrating how larger retrieval windows increase recall while compromising answer correctness.

Model	Overall		Single-Point		Intra-Document						Inter-Document					
	Corr.	Cov.	Text-only	Chart-only	Text-only		Chart-only		Text-Chart		Text-only		Chart-only		Text-Chart	
			Corr.	Corr.	Corr.	Cov.	Corr.	Cov.	Corr.	Cov.	Corr.	Cov.	Corr.	Cov.	Corr.	Cov.
Open-Source MLLMs																
SAIL-VL-2B	0.38	1.58	1.52	0.30	0.74	2.59	0.00	0.00	0.00	1.60	0.00	3.47	0.00	0.37	0.00	1.79
+ RAG *(k=5)*	3.88	8.51	19.19	4.18	1.48	14.44	0.00	2.04	0.00	1.34	2.78	13.89	0.00	0.75	0.00	1.79
+ RAG *(k=10)*	3.19	7.71	14.65	4.48	2.22	13.89	0.00	1.36	0.00	2.14	0.46	11.11	0.00	0.75	0.00	3.12
+ RAG *(GT)*	19.82	29.44	63.64	9.85	31.30	57.22	2.72	6.80	0.53	5.61	29.86	54.17	0.00	3.36	3.57	12.05
qwen2-VL-7B-instruct	1.16	4.45	4.55	2.09	0.56	7.22	0.00	1.19	0.00	3.48	0.46	7.41	0.00	1.49	0.00	5.36
+ RAG *(k=5)*	13.51	23.55	50.51	4.78	22.41	46.73	1.36	3.40	1.07	8.82	15.51	41.67	1.49	2.99	0.00	9.82
+ RAG *(k=10)*	14.45	23.57	51.01	3.28	26.11	49.81	0.68	2.38	1.60	7.75	20.14	43.75	0.75	1.49	0.00	8.48
+ RAG *(GT)*	33.15	42.46	78.28	11.04	64.26	81.30	2.04	9.52	5.88	20.86	62.73	80.56	2.99	6.72	9.82	27.23
MiniCPM-V-2.6-8B	0.88	4.05	2.02	2.69	0.37	6.67	0.00	1.70	0.00	3.74	0.00	6.71	0.00	2.24	0.00	4.46
+ RAG *(k=5)*	17.32	31.32	47.98	25.67	14.81	42.59	3.40	13.61	4.01	20.59	15.97	43.52	1.49	11.07	6.25	22.77
+ RAG *(k=10)*	17.60	31.51	48.48	19.70	21.11	50.43	2.72	12.24	4.81	19.79	19.68	46.30	1.49	11.07	4.46	23.21
+ RAG *(GT)*	46.94	59.41	79.29	48.66	65.37	80.99	12.93	29.93	22.46	46.79	69.68	83.10	14.18	32.34	20.98	47.32
Llama-3.2-90B-Vision	1.22	4.36	5.56	2.09	0.37	8.15	0.00	1.36	0.00	1.87	0.23	5.79	0.00	3.36	0.00	4.46
+ RAG *(k=5)*	20.42	34.68	50.51	30.15	21.85	49.20	5.44	16.21	4.81	20.05	17.82	45.83	1.49	12.69	8.04	28.57
+ RAG *(k=10)*	23.11	37.31	53.54	31.94	26.67	53.15	4.76	16.67	5.88	24.33	23.38	49.54	4.48	15.67	8.93	30.36
+ RAG *(GT)*	50.16	64.03	79.80	58.81	63.33	81.36	21.09	40.95	21.66	48.13	59.49	78.24	32.09	46.27	29.46	57.59
Proprietary MLLMs																
GPT-4o	2.05	7.37	8.59	5.97	0.37	13.64	0.00	2.38	0.00	5.08	0.46	12.27	0.75	2.61	0.00	8.48
+ RAG *(k=5)*	22.88	33.43	47.98	45.97	24.44	43.89	11.90	20.75	8.29	22.99	19.44	37.96	6.72	13.06	13.39	30.36
+ RAG *(k=10)*	26.11	37.93	47.98	47.76	28.52	51.85	13.27	23.47	10.96	26.74	25.93	43.98	14.93	23.88	15.62	34.38
+ RAG *(GT)*	60.50	68.86	90.20	63.88	94.07	97.22	24.83	38.10	32.19	55.19	92.96	95.46	36.57	51.12	51.52	60.95
Gemini-2.5-Pro	1.75	7.52	7.23	4.81	0.51	12.11	0.00	5.61	0.00	2.78	1.26	17.75	0.00	4.09	0.00	7.18
+ RAG *(k=5)*	23.21	32.47	47.61	46.79	23.57	44.70	14.25	25.21	7.10	23.68	23.51	39.32	7.84	15.62	10.01	28.12
+ RAG *(k=10)*	27.67	40.37	50.98	49.93	33.12	53.91	13.54	25.07	12.56	27.91	25.98	51.03	14.66	25.23	15.48	36.01
+ RAG *(GT)*	63.61	72.34	95.52	69.14	92.03	94.28	29.01	44.08	39.51	61.22	93.51	99.05	42.13	55.62	48.43	60.23
GPT-5	2.30	8.30	7.90	4.01	2.16	17.13	0.00	2.99	0.00	5.32	2.02	15.71	2.01	6.06	0.89	8.21
+ RAG *(k=5)*	25.79	42.93	53.32	47.03	26.33	55.06	14.91	33.67	10.01	34.91	23.72	47.91	11.32	29.31	16.67	40.24
+ RAG *(k=10)*	32.69	51.75	58.32	57.42	41.22	64.31	18.01	40.09	11.21	41.02	26.83	55.37	18.81	41.86	26.34	55.23
+ RAG *(GT)*	68.81	78.48	97.91	69.55	97.74	99.07	38.79	59.43	42.34	71.62	96.30	99.34	55.52	67.34	56.32	64.44
Claude-4.5-Sonnet	4.17	10.34	7.88	6.77	2.13	16.05	0.00	7.18	5.33	11.37	5.66	15.33	2.58	7.33	2.05	10.21
+ RAG *(k=5)*	26.25	50.25	47.24	56.79	24.67	57.02	16.55	38.27	7.35	40.05	19.67	51.82	12.42	35.61	19.67	46.43
+ RAG *(k=10)*	30.70	51.73	46.96	62.15	30.55	62.78	21.87	47.01	8.67	42.05	27.68	56.04	22.50	46.03	20.06	46.07
+ RAG *(GT)*	71.15	80.74	97.97	72.75	96.81	99.19	42.75	62.88	48.32	76.68	96.98	99.88	58.15	69.91	58.16	66.34

Table 4: Performance Comparison of Different MLLMs (%) on Chart-MRAG benchmark. The optimal retrieval configuration (*SigLIP + E5-large*) is employed across all experiments to ensure controlled comparison (best scores for open-source and proprietary models highlighted in blue and red, respectively).

trieval due to the inherent limitations of text-based chart representations. However, considering the maintenance overhead of separate modal stores, caption-based retrieval provides a practical approach that preserves effectiveness while significantly reducing system complexity.

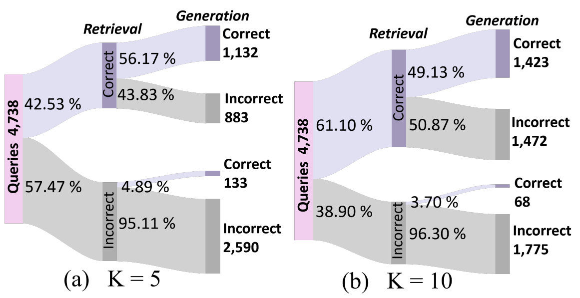


Figure 6: Impact of retrieval size k across different parameter scales, demonstrating that larger models consistently benefit from increased retrieval context while smaller models show performance degradation.

5.3 Generative Performance Comparison

Table 4 presents the comprehensive experimental results of mainstream MLLMs, with retrieval method 3 consistently applied across all evaluations to ensure controlled comparison. The results reveal that state-of-the-art MLLMs achieve only modest performance metrics (Correctness = 4.17 and Coverage = 10.34) without multimodal RAG knowledge, highlighting Chart-MRAG Bench's exceptional challenging nature that surpasses existing benchmarks in knowledge leakage control.

Claude-4.5-Sonnet demonstrates superior overall performance. The experimental results validate our keypoint-based evaluation methodology. With ground truth retrieval, Claude-4.5-Sonnet achieves Correctness of 71.15% and Coverage of 80.74%, outperforming mainstream MLLMs across various retrieval scenarios. It only falls behind GPT-5 in

Intra-Document Text-only QA.

Model performance generally scales with parameter count. Among open-source MLLMs, Llama-3.2-90B-Vision consistently outperforms models with smaller parameters across various retrieval settings. Similarly, in proprietary MLLMs, GPT-5, with its presumably larger model size, demonstrates marginally better than GPT-4o.

Architectural optimizations can mitigate MLLMs' parameter constraints. By incorporating SigLip-400M and optimizing multi-image understanding, MiniCPM-V-2.6-8B achieves a Correctness of 46.94 and Coverage that surpasses its base model qwen2-VL-7B-instruct by 13.79 and 16.95 respectively. Most notably, despite using only 7B parameters, it approaches the performance of Llama-3.2-90B-Vision, with gaps of 3.22 in Correctness and 4.62 in Coverage, demonstrating that thoughtful architecture design can largely compensate for parameter constraints. Detailed case studies are provided in Appendix 6.

5.4 Further Analysis

In this study, we examine the influence of retrieval rate (k) and modality bias of MLLMs in multimodal question answering. Our analysis shows:

Model performance in multimodal retrieval significantly correlates with parameter scale. Empirical analysis reveals a strong correlation between model scale and multimodal retrieval performance. We evaluated eight models of varying parameter sizes under different retrieval settings (k = 2, 5, 10, 15, 20), where retrieved items were balanced between images and text (split equally for even k, with text receiving one additional item for odd k). For each model, we selected 40 question-answer pairs per category, totaling 320 pairs for comprehensive evaluation, as shown in Fig 5. The results demonstrate that larger models consistently achieve superior performance across all retrieval settings. In contrast, smaller models show no significant improvement (even exhibit declining) in performance as the number of retrieved items increases.

Larger retrieval windows lead to a non-trivial trade-off between retrieval coverage and answer quality. To systematically investigate the impact of Top_k on response generation, we conducted extended experiments as visualized in Fig 6. With k=5, the system achieves a R@5 = 42.53 and 56.17 correctness. When increasing k=10, although the 61.10 Recall, the answer get 49.13 correctness. Notably, while this adjustment results in an increase in absolute correct answers from 1,132 to 1,423, the improvement sacrifices precision.

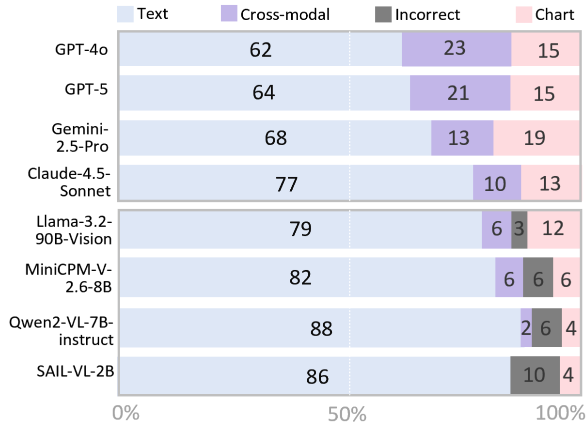


Figure 7: Analysis of modality preference in MLLMs when presented with redundant information across text and charts, revealing systematic modality bias.

MLLMs demonstrate consistent text-over-visual modality bias. To systematically investigate modality bias in MLLMs, we manually curated 100 specialized question-answer pairs where answers could be derived from both textual and visual information simultaneously, but with varying levels of granularity (e.g., "one third" in text versus "35.2%" in charts). All model responses were then carefully evaluated by human experts to ensure accurate assessment of modality preferences. As shown in Fig. 7, our human-verified analysis reveals a consistent preference across models for text-only responses, even when charts contain more precise information. Notably, larger MLLMs demonstrate superior ability in detecting information redundancy and actively acknowledge this in their responses. For instance, GPT-4o proactively identified information redundancy in 23% of its responses. In contrast, smaller models show limited sensitivity to such information redundancy. Detailed examples can be found in Appendix .12.

6 Conclusion

This paper introduces Chart-based MRAG, a novel task to bridge the evaluation gap for chart formats in MRAG systems. To support this, we propose CHARGE, an automated framework for generating Cross-modal evaluation samples with keypoint-based metrics. Combining CHARGE with expert validation, we construct Chart-MRAG Bench, comprising 4,738 high-quality QA pairs across 8 domains. Experiments expose key limitations in current MRAG approaches; specialized architectures are needed for high-density visual interactions.

Limitations

While our work presents promising results, we acknowledge several limitations that warrant consideration in future research.

First, although we ensured the accuracy of chart information in Chart-MRAG Bench through manual verification, the CHARGE framework would benefit from more advanced OCR techniques to further enhance the accuracy of question generation, especially in handling complex chart layouts and diverse visual elements.

Second, due to computational constraints, our evaluation was confined to a select set of MRAG methods and MLLMs. A more comprehensive evaluation across diverse model architectures and frameworks would likely yield additional insights into the generalizability of our findings and potentially reveal new directions for improvement.

Ethical Considerations

This research was conducted under the approval of our institution's ethics review board. All procedures were designed to ensure participant welfare and data privacy throughout the study.

Participant Recruitment and Compensation. We recruited expert annotators through Amazon, a professional data annotation platform. Annotators were compensated at a rate of $28.5 per hour. This rate was determined by:

- Conducting pilot studies with 5 annotators to establish an average task completion time of 45 minutes
- Accounting for additional training time (30 minutes) and regular breaks
- Considering local living wage standards across different regions
- Adding a 20% premium for specialized expertise required

For a typical 8-hour workday including training, we ensure fair payment while maintaining data quality. Regular feedback from annotators confirmed the compensation was considered fair for the required expertise and effort.

Informed Consent and Instructions. All annotators received comprehensive instructions detailing the task requirements, data usage policies, and potential content exposure. The instruction package included:

- Task objectives and annotation guidelines
- Examples of expected annotations
- Data privacy and usage policies
- Right to withdraw from participation

Annotators provided explicit consent for their contributions to be used in academic research and public datasets.

Annotator Demographics. Our annotation team consisted of 12 professional annotators with backgrounds in data science and visualization. The annotators represented diverse geographical locations (3 North America, 3 Europe, 6 Asia) and possessed relevant domain expertise. All demographic information was self-reported during the recruitment process.

Data Collection and Privacy. The datasets used in this study, including those for generating multimodal question-answer pairs, were collected and processed in compliance with GDPR and relevant data privacy regulations. We ensured that:

- No personally identifiable information was collected
- All chart data was anonymized before annotation
- Participants were informed about data usage and sharing plans

Bias Mitigation. We implemented several measures to minimize potential biases in our dataset and evaluation metrics:

- Diverse annotator selection to ensure varied perspectives
- Regular quality checks for systematic biases in annotations
- Balanced representation of different chart types and domains

The resulting benchmark will be made publicly available for academic research purposes, accompanied by detailed documentation of the collection process and annotator guidelines. All materials will be released through established academic repositories to ensure transparency and reproducibility.

Acknowledgements

We gratefully acknowledge the support of the National Key R&D Program of China (No.

2024YFC2707805), the National Natural Science Foundation of China (No. 62176029, No. 62506050), the Open Competition Program of Chongqing Municipal Commission of Economy and Information Technology (No. YJX-202500100200X), the China Postdoctoral Science Foundation Funded Project (No. 2024M763867), and the Chongqing Higher Education Teaching Reform Research Project (No. 242009).

The experimental and computational work in this research runs on the Huawei Cloud AI Compute Service. We appreciate the stable compute supply from this platform. We sincerely thank the anonymous reviewers for their insightful comments and constructive suggestions.

References

Anthropic. 2024. Claude 4: Constitutional ai with harmlessness from ai feedback. Technical report, Anthropic.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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. *arXiv preprint arXiv:2402.03216*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025. Mmdocir: Benchmarking multi-modal retrieval for long documents. *arXiv preprint arXiv:2501.08828*.

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*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

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

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2024. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 1(2):4.

Caijun Jia, Nan Xu, Jingxuan Wei, Qingli Wang, Lei Wang, Bihui Yu, and Junnan Zhu. 2025. Chartreasoner: Code-driven modality bridging for long-chain reasoning in chart question answering. *arXiv preprint arXiv:2506.10116*.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, Yu Liu, et al. 2025. Mmsearch: Unveiling the potential of large models as multi-modal search engines. In *The Thirteenth International Conference on Learning Representations*.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, et al. 2024. Jina clip: Your clip model is also your text retriever. *arXiv preprint arXiv:2405.20204*.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. 2026a. Modeling uncertainty trends for timely retrieval in dynamic rag. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31527–31535.

Bo Li, Mingda Wang, Gexiang Fang, Shikun Zhang, and Wei Ye. 2026b. Retrieval as generation: A unified framework with self-triggered information planning. *Preprint*, arXiv:2604.11407.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024a. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Yangning Li, Yinghui Li, Xingyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S Yu, Fei Huang, et al. 2024b. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1917–1925.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. 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. 2024b. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*.

Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Heyan Huang, and Xian-Ling Mao. 2024c. Multi-modal retrieval augmented multi-modal generation: A benchmark, evaluate metrics and strong baselines. *arXiv preprint arXiv:2411.16365*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

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.

OpenAI. 2025. GPT-5: Artificial general intelligence is near. Technical report, OpenAI.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.

Bytedance Douyin Content Team. 2024. Sail-vl: Scalable vision language model training with high quality data curation.

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*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Jingxuan Wei, Nan Xu, Junnan Zhu, Gaowei Wu, Qi Chen, Bihui Yu, Lei Wang, et al. 2025a. Chartmind: A comprehensive benchmark for complex real-world multimodal chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4555–4569.

Kaiwen Wei, Xiao Liu, Jie Zhang, Zijian Wang, Ruida Liu, Yuming Yang, Xin Xiao, Xiao Sun, Haoyang Zeng, Changzai Pan, et al. 2025b. Cfvbench: A comprehensive video benchmark for fine-grained multimodal retrieval-augmented generation. *arXiv preprint arXiv:2510.09266*.

Kaiwen Wei, Rui Shan, Dongsheng Zou, Jianzhong Yang, Bi Zhao, Junnan Zhu, and Jiang Zhong. 2026. Mirage: Scaling test-time inference with parallel graph-retrieval-augmented reasoning chains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33818–33826.

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Pakazad, Tongshuang Wu, and Graham Neubig. 2024. Synthetic multimodal question generation. *arXiv preprint arXiv:2407.02233*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024b. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.

Junnan Zhu, Jingyi Wang, Bohan Yu, Xiaoyu Wu, Junbo Li, Lei Wang, and Nan Xu. 2025. Tableeval: A real-world benchmark for complex, multilingual, and multi-structured table question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7137–7157.

Jiaru Zou, Dongqi Fu, Sirui Chen, Xinrui He, Zihao Li, Yada Zhu, Jiawei Han, and Jingrui He. 2025. Gtr: Graph-table-rag for cross-table question answering. *arXiv preprint arXiv:2504.01346*.

Comparison of MRAG Tasks

While existing multimodal large language models have made significant progress on image-based MRAG tasks and text-image MRAG tasks, they still exhibit substantial limitations when handling chart-based tasks. To systematically analyze the capabilities of MLLMs in handling chart-text interactions, we conducted comprehensive evaluations across 8 distinct interaction patterns. Our experiments with Claude-4.5-Sonnet-20241022 reveal a clear performance hierarchy across different task categories, as shown in Table 5.

To further contextualize the unique challenges of chart-text interactions, we conducted comparative analyses with text-image tasks. Table 6 presents the retrieval performance across different modalities, while Table 7 shows the MLLM performance when provided with ground truth retrieved context.

The results demonstrate a significant performance gap between text-chart and text-image tasks. Notably, both retrieval systems and MLLMs show substantially lower performance on text-chart interactions compared to traditional text-image tasks, highlighting the unique challenges posed by chart understanding and reasoning. This performance disparity is particularly evident in the retrieval phase, where CLIP's effectiveness drops by approximately 47 percentage points when handling charts instead of natural images.

Chart-MRAG Bench Cases

To illustrate the diverse chart categories in Chart-MRAG Bench, we present representative examples as shown in Figure 8.

We categorize the question-answering pairs in Chart-MRAG into eight distinct categories, as summarized in Table 8, encompassing various combinations of single-point, intra-document, and inter-document scenarios across text-only, chart-only, and text-chart contexts. These categories are illustrated through representative examples: Single-Point Text-Only QA (Fig. 26), Single-Point Chart-Only QA (Fig. 27), Intra-Document Text-Only QA (Fig. 28), Intra-Document Chart-Only QA (Fig. 29), Intra-Document Text-Chart QA (Fig. 30), Inter-Document Text-Only QA (Fig. 31), Inter-Document Chart-Only QA (Fig. 32), and Inter-Document Text-Chart QA (Fig. 33).

Chart-MRAG Bench Case Study

.1 Case Study: Single-Point Text-only Question

Parameter size strongly correlates with retrieval accuracy, with smaller models exhibiting critical failures despite having ground truth context. We selected a representative Single-Point Text-only example, as shown in Figure 17, which details the question-answer pair, ground truth retrieval passage, and responses from eight different MLLMs when provided with the ground truth context. In this task, proprietary large models (Claude-4.5-Sonnet, Gemini-2.5-Pro, GPT-5, GPT-4o) demonstrated perfect or near-perfect accuracy, capturing all key relationships in the ground truth. Medium-sized models (Llama-3.2-90B-Vision, MiniCPM-V-2.6) maintained factual correctness with minor stylistic variations. However, Qwen2-VL-7B-instruct completely missed the mark, producing irrelevant information about news consumption rather than surname preferences, despite having access to the correct context. SAIL-VL-2B maintained accuracy but padded its response with unnecessary prefatory statements. This case reveals that even with ground truth retrieval handed to them on a silver platter, smaller models (<10B parameters) still struggle with correctly identifying and extracting relevant information, highlighting the persistent challenge of developing parameter-efficient MLLMs for effective MRAG systems.

.2 Case Study: Single-Point Chart-only Question

Chart interpretation reveals significant accuracy challenges across all model scales, with most models failing to extract the correct percentage despite ground truth availability. We examined a Single-Point Chart-only example (Figure 18) investigating Republicans' perceptions of economic fairness. Here, model performance showed alarming inconsistency. Only Claude-4.5-Sonnet correctly identified that 32% of Republicans believe the economic system is generally fair. Gemini-2.5-Pro provided the correct percentage but delivered a minimal, contextually incomplete response. Surprisingly, both GPT-5 and Qwen2-VL-7B-instruct claimed inability to answer despite having the ground truth chart available. Even more concerning, other models confidently provided incorrect figures: GPT-4o (39%), Llama-3.2-90B-Vision (36%), MiniCPM-V-2.6 (47%), and SAIL-

Task Type	Sub-task	Correctness (%)	Coverage (%)
Text-only tasks	Single-Point Text-only	83.33	83.33
	Intra-Document Text-only	69.81	85.19
	Inter-Document Text-only	68.98	85.88
Chart-only tasks	Single-Point Chart-only	70.75	70.75
	Intra-Document Chart-only	32.65	60.32
	Inter-Document Chart-only	48.51	67.91
Text-Chart tasks	Intra-Document Text-Chart	20.32	56.68
	Inter-Document Text-Chart	35.71	63.84

Table 5: Performance breakdown across different MRAG task categories

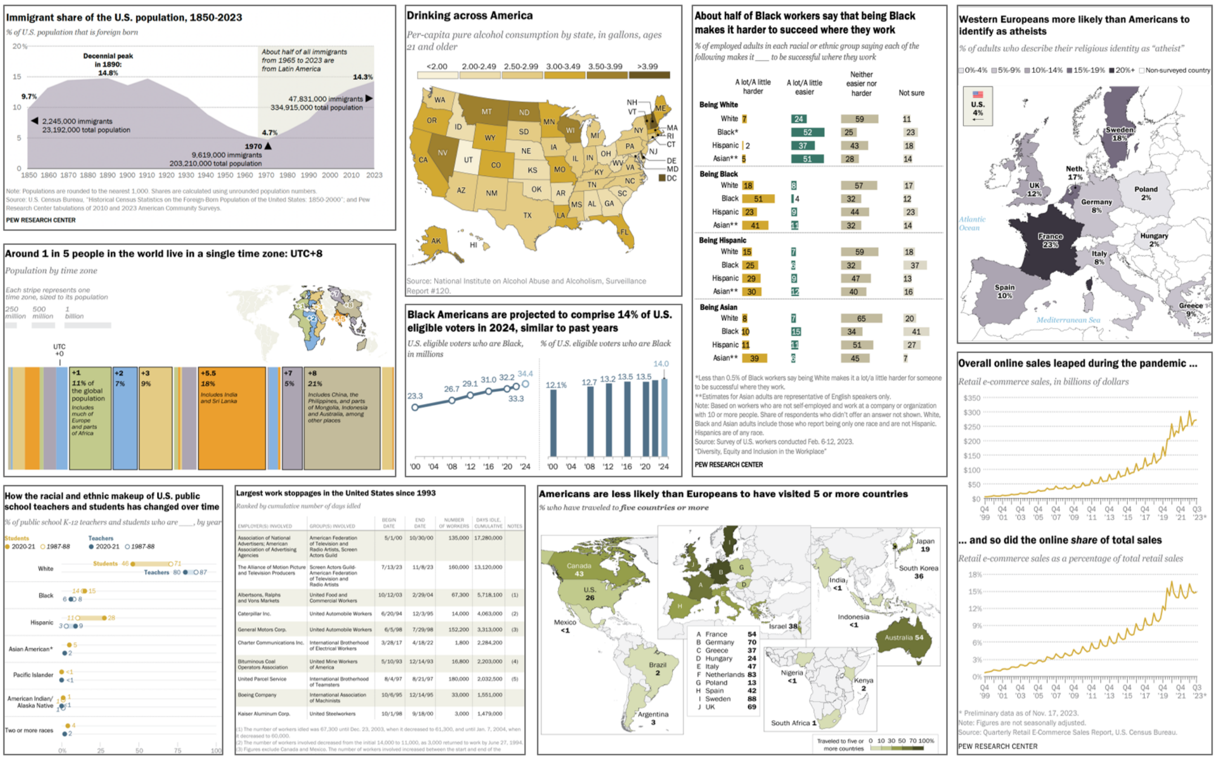


Figure 8: Representative visualization categories from Chart-MRAG Bench, showcasing temporal trend analysis (line charts), geospatial data visualization (choropleth maps), categorical comparisons (bar charts), compositional analysis (stacked bars), and integrated text-chart representations. The diversity of these examples demonstrates the comprehensive scope of Chart-MRAG Bench in representing complex statistical information across multiple domains and visualization paradigms.

```
# Single-Point_Text-only_00316

## Query
In 2024, what percentage of U.S. adults will support prioritizing renewable
energy over fossil fuel expansion?

## Answer
67% of American adults support prioritizing renewable energy over expanding oil,
coal and natural gas production.

## Source

### Source 1
"id": "paragraph_00189_07",
"text": "Democrats are much more likely than Republicans to say protecting the
environment (63% vs. 23%) and dealing with climate change (59% vs. 12%) should
be top policy priorities for 2024. In fact, addressing climate change ranks last
on Republicans' list of priorities this year. Views of the Biden
administration's current climate policies also differ sharply by party. Eight-
in-ten Democrats say the federal government is doing too little to reduce the
effects of climate change, compared with 29% of Republicans, according to a
Center survey from spring 2023. Overall, a majority of U.S. adults (67%) support
prioritizing the development of renewable energy, such as wind and solar, over
expanding the production of oil, coal and natural gas. But Democrats are far
more likely than Republicans to prefer this (90% vs. 42%). Still, the public
overall is hesitant about a full energy transition: Just 31% say the U.S. should
phase out fossil fuels completely.“

## keypoints
"paragraph_00189_07_04": "Overall, a majority of U.S. adults (67%) support
prioritizing renewable energy sources, such as wind and solar, over expanding
oil, coal and natural gas production."
```

Figure 9: A sample case of single-point text-only question answering.

Single-Point_Chart-only_ 00317

Query

In the public policy debate about climate change, what percentage of Democrats or those who lean Democratic think climate scientists have the appropriate amount of influence?

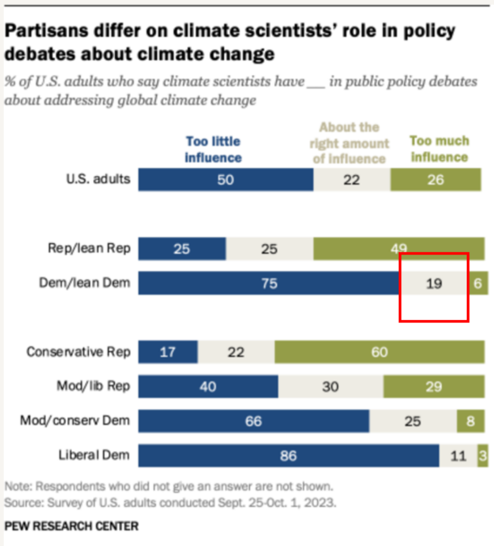
Answer

Nineteen percent of Democrats or those who lean Democratic believe climate scientists have the appropriate influence in the public policy debate over addressing global climate change.

Source

Source 1

"id": "chart_00077_04"



keypoints

"chart_00077_04": "19% of Dem/lean Dem who say climate scientists have About the right amount of influence in public policy debates about addressing global climate change"

Figure 10: A sample case of single-point chart-only question answering.

```
# Intra-Document_Text-only_00571

## Query
In Japan, what percentage of adults say their family owns a cemetery? Among
these adults, how many of those with no religious affiliation have made
offerings to their ancestors in the past 12 months?

## Answer
85% of Japanese adults claim that their family owns a cemetery. Among these
adults, 59% of those with no religious affiliation have prepared food, water, or
beverages for their ancestors in the past 12 months.

## Source

### Source 1
"id": "paragraph_00128_01"
"text": "People in Japan are preparing to celebrate Obon – a festival devoted to
celebrating ancestors that features lighting lanterns and maintaining family
gravesites. In Japan, 85% of adults say their family has such a gravesite, and
79% say they have looked after this gravesite by sweeping or cleaning it in the
past year, according to a recent Pew Research Center survey. Obon …"

### Source 2
"id": "paragraph_00128_02"
"text":" We also asked survey respondents … For instance, 59% of Japan's
religiously unaffiliated adults say they have offered food, water or drinks in
the past 12 months to care for their ancestors. Christians generally are less
likely to engage in these sorts of activities. However, many Vietnamese
Christians have burned incense, offered flowers or lit candles to care for
ancestors in the last year. "

## keypoints
"paragraph_00128_01_02": "85% of adults in Japan say their family has a
gravesite",

"paragraph_00128_02_07": "59% of Japan's religiously unaffiliated adults have
offered food, water, or drinks in the past 12 months to care for their
ancestors"
```

Figure 11: A sample case of intra-document text-only question answering.

Intra-Document_Chart-only_00257

Query

Among Hispanic women, what percentage strongly prefers a Spanish-speaking physician for routine care? Among these women, what percentage of those with a bachelor's degree or higher have a primary care provider?

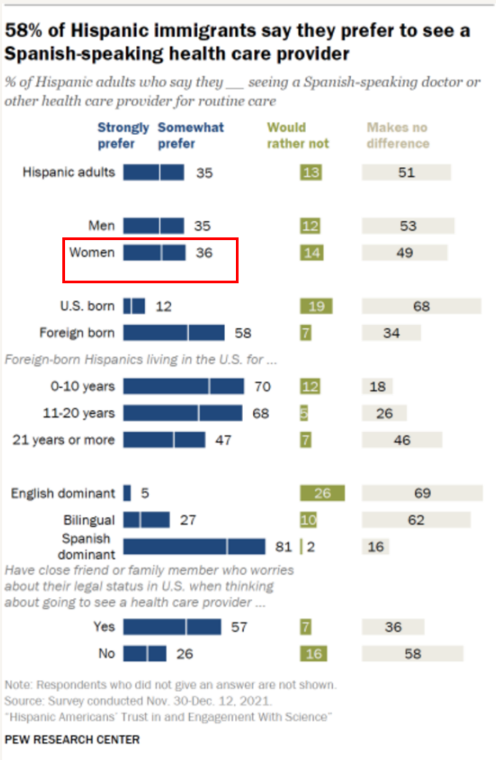
Answer

The data showed that 36% of Hispanic women strongly preferred to have a Spanish-speaking physician for their routine care. Of these women, 73% of college graduates had a primary care provider.

Source

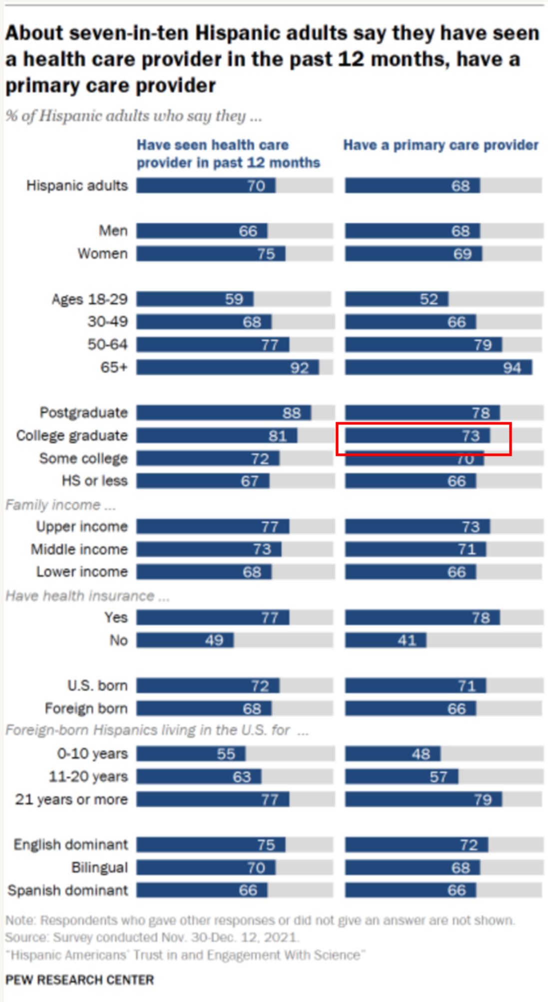
Source 1

chart_00080_05



Source 2

chart_00080_02



keypoints

"chart_00080_05": "36% of Hispanic adults who say they Strongly prefer seeing a Spanish-speaking doctor or other health care provider for routine care among Women",
"chart_00080_02": "73% of Hispanic adults who are College graduate have a primary care provider"

Figure 12: A sample case of intra-document chart-only question answering.

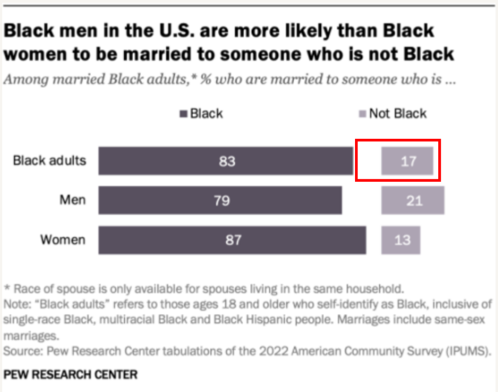
Intra-Document_Text-Chart_00225

Query
Among married black adults, what percentage of them have a non-black spouse? How does this compare to the percentage of black women who are divorced?

Answer
Among married black adults, 17% have a non-black spouse. In comparison, 14% of black women are divorced. This indicates that the rate of interracial marriage among black adults is slightly higher than the divorce rate among black women.

Source

Source 1
"id": "chart_00058_09"



Source 2
"id": "paragraph_00058_08"
"text": "About a third of Black adults (32%) are currently married. That compares with 53% of adults who are not Black. Among Black adults, 36% of men are married, compared with 29% of women. Black women, in turn, are slightly more likely than Black men to be divorced (14% vs. 10%) or widowed (8% vs. 2%). "

keypoints
"chart_00058_09": "17% of married Black adults who are Black adults married to Not Black",

"paragraph_00058_08_02": "14% of Black women are divorced"

Figure 13: A sample case of intra-document text-chart question answering.

```
# Inter_Document_Text-only_00406

## Query
How important is spending time with family in the lives of American adults? In
comparison, how many adults say having children or being married is equally
important to life satisfaction?

## Answer
Nine-in-ten Americans say spending time with family is very or one of the most
important things to them personally, compared to just 26% who say having
children and 23% who say being married is extremely or very important to life
satisfaction.

## Source

### Source 1
"id": "paragraph_00090_02_03"
"text": "According to a Center survey from this spring, a large majority of U.S.
adults (73%) say family time is one of the most important aspects of their life
- and none of the other priorities we asked about comes close. Overall, nine-in-
ten Americans say spending time with family is either very important or one of
the most important things to them personally, regardless of how much time they
actually devote to it. What's more, spending time with family is the top
priority for Americans regardless of political affiliation. About nine-in-ten
Republicans and Democrats (including those who lean to each party) say family
time is either very important or one of the most important aspects of their life.
"

### Source 2
"id": "paragraph_00103_02_04"
“text”: “According to … A Center survey conducted in April found that relatively
few Americans see marriage as essential for people to live a fulfilling life
compared with factors like job satisfaction and friendship. While majorities say
that having a job or career they enjoy (71%) and having close friends (61%) are
extremely or very important for living a fulfilling life, far fewer say this
about having children (26%) or being married (23%). "

## keypoints
"paragraph_00090_02_03": "Nine-in-ten Americans say spending time with family is
either very important or one of the most important things to them personally",

"paragraph_00103_02_04": "Far fewer Americans say having children (26%) or being
married (23%) are extremely or very important for living a fulfilling life"
```

Figure 14: A sample case of inter-document text-only question answering.

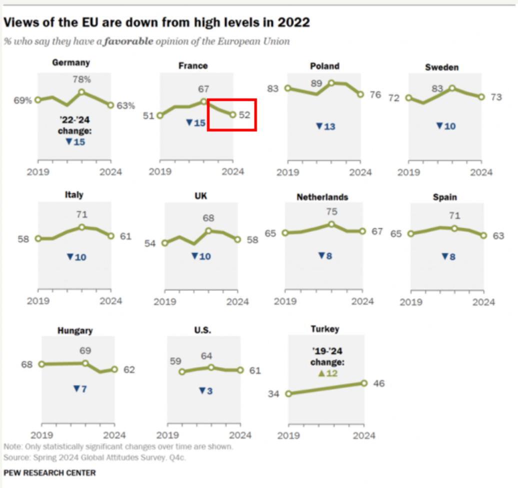
Inter_Document_Chart-only_00513

Query
What percentage of people in France will have a positive view of the EU in 2024? In comparison, what percentage did this hold in Germany in 2019?

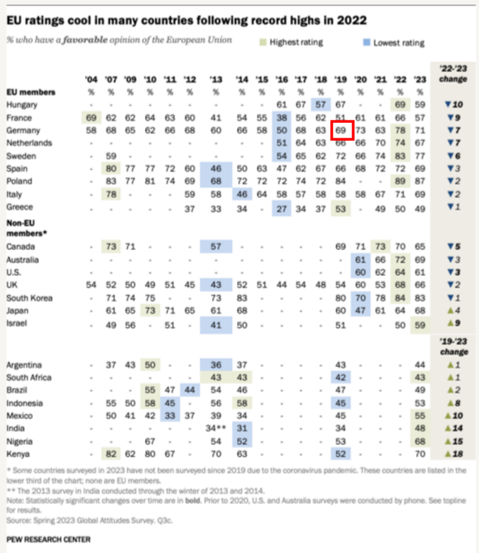
Answer
In 2024, 52% of people in France will have a positive view of the EU. In 2019, this figure was 69% in Germany.

Source

Source 1
"id": "chart_00112_02"



Source 2
"id": "chart_00076_02",



keypoints
"chart_00112_02": "52% of people in France who say they have a favorable opinion of the European Union in 2024",

"chart_00076_02": "69% who have a favorable opinion of the European Union among Germany in 2019"

Figure 15: A sample case of inter-document chart-only question answering.

Inter_Document_Text-Chart_01016

Query

Among people aged 65 and older, what percentage are not too concerned or not at all concerned about the impact of AI in the 2024 election? And among these same people, what percentage are mostly concerned about the growing use of AI in daily life?

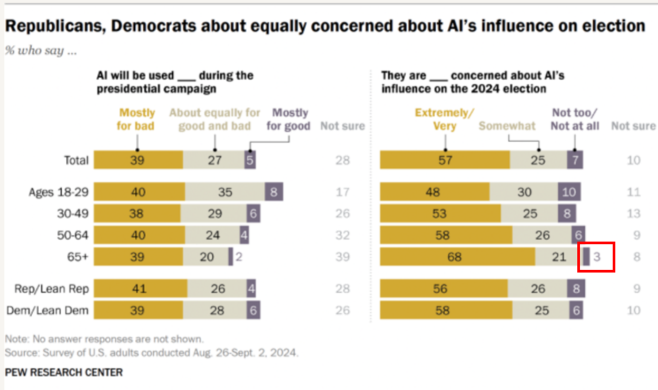
Answer

Among those 65 and older, 3% are not too concerned or not at all concerned about the impact of AI in the 2024 election. Meanwhile, about 61% of this age group are mostly concerned about the growing use of AI in daily life.

Source

Source 1

"id": "chart_00257_04"



Source 2

"id": "paragraph_00025_01_06"
"text": " A growing share of Americans express concern about the role artificial intelligence (AI) is playing in daily life, according to a new Pew Research Center survey. Overall, 52% of Americans say they feel more concerned than excited about the increased use of artificial intelligence. Just 10% say they are more excited than concerned, while 36% say they feel an equal mix of these emotions. The share of Americans who are mostly concerned about AI in daily life is up 14 percentage points since December 2022, when 38% expressed this view. Concern about AI outweighs excitement across all major demographic groups. Still, there are some notable differences, particularly by age. About six-in-ten adults ages 65 and older (61%) are mostly concerned about the growing use of AI in daily life, while 4% are mostly excited. "

keypoints

"chart_00257_04": "3% of 65+ who are Not too/Not at all concerned about AI's influence on the 2024 election",

"paragraph_00025_01_06": "About 61% of adults ages 65 and older are mostly concerned about the growing use of AI in daily life"

Figure 16: A sample case of inter-document text-chart question answering.

Retriever	Text-Chart		Text-Image(Wu et al., 2024)	
	R@5	R@10	R@5	R@10
CLIP	13.26	19.07	61.1	67.1
BM25	27.02	36.46	53.2	56.6

Table 6: Retrieval performance comparison between text-chart and text-image tasks

Model	Text-Chart	Text-Image(Mathew et al., 2021)
Claude-4.5-Sonnet	71.15	83.36
GPT-4o	60.50	88.40
Qwen2-VL-7B	33.15	62.60

Table 7: MLLMs performance comparison between text-chart and text-image tasks

VL-2B (50%). This case demonstrates that chart-only information extraction remains exceptionally challenging even for advanced MLLMs, with most models either refusing to answer or extracting incorrect values. Unlike text-only scenarios where parameter size strongly predicted performance, chart interpretation capabilities show inconsistent correlation with model scale, suggesting fundamental limitations in current visual processing architectures rather than mere parameter efficiency issues.

.3 Case Study: Intra-Document Text-only Question

Multi-hop reasoning from text demonstrates high performance ceiling across model scales, though smaller models still exhibit information integration failures. This Intra-Document Text-only example (Figure 19) required models to integrate information from two separate paragraphs about federal judge appointments. Most models performed impressively well on this task. Claude-4.5-Sonnet, Gemini-2.5-Pro, GPT-5, GPT-4o, Llama-3.2-90B-Vision, and MiniCPM-V-2.6 all correctly identified that 47% of appellate judges were appointed by Democratic presidents and 31% by Obama specifically. This high success rate suggests that text-only multi-hop reasoning has reached a significant level of maturity across various model scales. However, Qwen2-VL-7B-instruct demonstrated a partial failure, correctly identifying the 47% Democratic appointment figure but claiming the Obama-specific percentage was not provided, despite it being clearly stated in the second retrieved paragraph. SAIL-VL-2B provided the correct figures but padded its response with unnecessary framing statements. This case reveals that while text-only reasoning capabilities are generally strong across models, information integration across multiple paragraphs remains challenging for smaller parameter models, which may struggle to maintain coherent attention across longer contexts even when the relevant information is explicitly provided.

.4 Case Study: Intra-Document Chart-only Question

Chart-only multi-hop reasoning reveals severe inconsistency across models, with most models misrepresenting at least one key statistic despite access to ground truth charts. This example (Figure 20) required integrating information from two separate charts about Indonesian religious attitudes. Claude-4.5-Sonnet performed exemplarily, accurately reporting all relevant statistics: 41% seeing no impact from diversity, 54% viewing Muslim extremists as threats, and 35% concerned about growing Christian numbers. GPT-5 and GPT-4o also provided complete and accurate responses. However, other models showed significant failures: Gemini-2.5-Pro omitted statistics about Christian numbers, Llama-3.2-90B-Vision incorrectly claimed "nearly all adults" see diversity positively rather than 41% seeing no impact, MiniCPM-V-2.6 reported entirely incorrect figures (61% instead of 41%, and 41% instead of 35%), Qwen2-VL-7B-instruct claimed it couldn't find information about the "no impact" percentage despite it being clearly present, and SAIL-VL-2B completely refused to answer. This case highlights that chart-only multi-hop reasoning remains exception-

Source-Constrained and Modality-Constrained Question-Answer Categories	
Single-Point Text-Only	Questions that require reasoning about an individual textual keypoint ($k_i^t \in K^T$), focusing on discrete factual validation within a single text segment.
Single-Point Chart-Only	Questions centered on an individual chart-only keypoint ($k_i^c \in K^C$), examining specific data points or visual elements within a single chart.
Intra-Document Text-Only	Questions that necessitate integrative reasoning across multiple textual keypoints ($k_i^t, k_j^t \in K^T$) within the same document ($d_i \in D$).
Intra-Document Chart-Only	Questions requiring comparative analysis of multiple chart-only keypoints ($k_i^c, k_j^c \in K^C$) from a single document ($d_i \in D$).
Intra-Document Text-Chart	Questions involving cross-modal reasoning between textual and chart-only keypoints ($k_i^t \in K^T, k_j^c \in K^C$) within the same document ($d_i \in D$).
Inter-Document Text-Only	Questions demanding associative reasoning between textual keypoints ($k_i^t, k_j^t \in K^T$) from distinct documents ($d_i, d_j \in D, i \neq j$).
Inter-Document Chart-Only	Questions requiring comparative analysis of chart-only keypoints ($k_i^c, k_j^c \in K^C$) across different documents ($d_i, d_j \in D, i \neq j$).
Inter-Document Text-Chart	Questions involving cross-modal and cross-document reasoning, integrating textual and chart keypoints ($k_i^t \in K^T, k_j^c \in K^C$) from different documents ($d_i, d_j \in D, i \neq j$).

Table 8: Taxonomy of question-answer pairs in Chart-MRAG, categorized by source constraints (Single-Point/Intra-Document/Inter-Document) and modality constraints (Text-only/Chart-only/Text-Chart).

ally challenging even with ground truth available. The inconsistent performance across models suggests fundamental limitations in current visual processing architectures rather than simple parameter scaling issues, with even relatively large models struggling to extract and integrate multiple statistics accurately from chart-only content.

.5 Case Study: Intra-Document Text-Chart Question

Cross-modal integration reveals critical failures across most models, with even large proprietary models struggling to correctly synthesize information from text and charts. This Intra-Document Text-Chart example (Figure 21) required models to integrate percentage information from both a chart about tipping habits and text about vacation time usage. Only Claude-4.5-Sonnet correctly reported that 2% of adults tip more than 20%, while 46% don't use all their allowed time off. Gemini-2.5-Pro also provided correct figures and made an appropriate comparison. However, all other models showed significant failures: GPT-5 and Qwen2-VL-7B-instruct claimed inability to answer despite having the necessary information, GPT-4o incorrectly reported 22% for tipping (which actually represents those who tip exactly 20%, not more), Llama-3.2-90B-Vision made the same error and claimed the vacation information wasn't provided, MiniCPM-V-2.6 also reported the incorrect 22% figure while correctly identifying the 46% vacation statistic, and SAIL-VL-2B simply produced an incorrect "5%" answer with no context. This case reveals that cross-modal integration between text and charts represents an exceptional challenge for current MLLMs. The failure of even powerful models like GPT-4o on this task suggests that cross-modal reasoning remains a frontier challenge, with current architectures struggling to correctly interpret relationships between numeric information across different modalities, even when the required information is explicitly provided in the ground truth context.

.6 Case Study: Inter-Document Text-only Question

Inter-document text reasoning reveals considerable inconsistency across proprietary models, with Claude-4.5-Sonnet unexpectedly producing a complete information mismatch. This example (Figure 22) required synthesizing information from two separate documents about Americans' views on political rhetoric. Surprisingly, Claude-4.5-Sonnet completely failed this task, providing information about climate scientists' influence rather than political rhetoric—a complete topic mismatch that suggests serious retrieval or context integration failures. In contrast, Gemini-2.5-Pro, GPT-5, GPT-4o, Llama-3.2-90B-Vision, MiniCPM-V-2.6, Qwen2-VL-7B-instruct, and SAIL-VL-2B all successfully extracted and integrated the key information that Americans hold similar views about heated political language and prioritize shared political views when evaluating candidates. This unusual pattern—where a typically high-performing model fails completely while smaller models succeed—suggests that inter-document reasoning introduces unique challenges that don't simply correlate with parameter count. The failure may indicate that Claude's processing of multiple documents creates vulnerabilities to context confusion that smaller, more specialized models avoid through simpler document handling. This case demonstrates that even state-of-the-art proprietary models can experience catastrophic failures in multi-document contexts, highlighting the need for robust evaluation across diverse cross-document reasoning scenarios.

.7 Case Study: Inter-Document Chart-only Question

Chart reasoning across documents reveals crucial challenges even for high-performing models, with data interpretation capabilities varying widely and unpredictably. This example (Figure 23) required extracting and relating percentage information from two different charts about climate change opinions. Several models performed well: Claude-4.5-Sonnet, GPT-5, GPT-4o, and Llama-3.2-90B-Vision all correctly reported that 31% of U.S. adults believe climate scientists have poor understanding, with 23% saying the government is doing an adequate job. However, Gemini-2.5-Pro correctly identified the 31% figure but mistakenly suggested the second percentage referred to all adults rather than specifically those who doubted scientists. MiniCPM-V-2.6 reported entirely incorrect figures (37% and 56%), Qwen2-VL-7B-instruct claimed inability to answer despite having the necessary information, and SAIL-VL-2B produced completely fabricated statistics (49% and 11%). This pattern reveals that interpreting multiple charts across different documents represents a significant challenge for current MLLMs. Unlike text-only reasoning, chart interpretation abilities don't correlate straightforwardly with model size, as even some larger models struggle while others succeed. The unpredictable performance suggests that current visual processing architectures lack robust mechanisms for reliable chart data extraction and cross-reference, highlighting an important frontier for improvement in multimodal reasoning systems.

.8 Case Study: Inter-Document Text-Chart Question

Cross-modal reasoning across documents reveals significant variability in performance, with integration abilities not strictly correlating with model size. This example (Figure 24) required extracting information from both a chart about Taoist connections in Taiwan and text about emotional connections to China. Several models performed well: Claude-4.5-Sonnet, GPT-5, GPT-4o, and SAIL-VL-2B all correctly identified that 48% of Taiwanese adults felt connected to Taoism, while 46% of those aged 35 and over felt emotional connections to China. However, other models demonstrated various failures: Gemini-2.5-Pro reported an incorrect "22%" for Taoist connection while correctly identifying the 46% China connection, Llama-3.2-90B-Vision incorrectly claimed the image didn't contain Taoist connection information, MiniCPM-V-2.6 reported an incorrect figure (35% instead of 48%), and Qwen2-VL-7B-instruct claimed inability to answer the first part while misinterpreting the second part (attributing the 48% figure to Chinese connections rather than Taoist ones). This case demonstrates that cross-modal reasoning across different documents represents a particular challenge that doesn't strictly correlate with model size or general capability. The inconsistent performance suggests that current multimodal architectures lack robust mechanisms for reliably integrating information across both modality boundaries (text vs. chart) and document boundaries. Notably, even smaller models like SAIL-VL-2B occasionally succeed at these complex tasks, indicating that architectural design choices may be as important as parameter count for these challenging cross-modal, cross-document reasoning scenarios.

Setup and Metrics

.9 Retrieval Setup

For retrieval system, we designed three distinct configurations to evaluate different approaches to multimodal information retrieval:

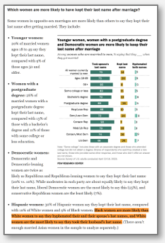
Unified Multimodal Embedding and Single Vector Store. We directly embedded charts and text into a unified embedding space using vision-language models CLIP, JINA-CLIP, and SigLIP. This approach maps all content to same-dimensional vectors in a single vector store, en-

Single-Point_Text-only_00607

Query: Based on data from mid-April 2023, how do racial differences affect married women's surname preferences?

GT_Answer: Black women are more likely to prefer hyphenated last names, while white women are more likely to adopt their husband's last name.

GT Retrieval content



Paragraph_00031

"Black women are more likely than White women to say they hyphenated their and their spouse's last names, and White women are the most likely to say they took their husband's last name."

Paragraph_00031_02_07

Answering

Claude-4.5-Sonnet

Black women are more likely than White women to say they hyphenated their and their spouse's last names, and White women are the most likely to say they took their husband's last name.

Correctness: 1.0 *Coverage: 1.0*

Llama-3.2-90B-Vision

Based on the information provided, racial differences appear to influence married women's surname preferences. Black women are more likely to choose to hyphenate their and their spouse's last names. In contrast, White women are the most likely to adopt their husband's last name.

Correctness: 1.0 *Coverage: 1.0*

Gemini-2.5-Pro

Black women are more likely than White women to say they hyphenated their and their spouse's last names, and White women are the most likely to say they took their husband's last name.

Correctness: 1.0 *Coverage: 1.0*

MiniCPM-V-2.6

Black women are more likely than White women to hyphenate their and their spouse's last names, while White women are more likely to take their husband's last name.

Correctness: 1.0 *Coverage: 1.0*

GPT-5

Black women are more likely than White women to hyphenate their and their spouse's last names, while White women are more likely to take their husband's last name.

Correctness: 1.0 *Coverage: 1.0*

Qwen2-VL-7B

Black Americans have a higher preference for getting news on TV compared to people of other racial or ethnic backgrounds. 38% of Black Americans prefer TV as their primary news source. 41% of Black adults report consuming news and information from local news outlets either extremely or fairly often.

Correctness: 0.0 *Coverage: 0.0*

GPT-4o

Black women are more likely than White women to say they hyphenated their and their spouse's last names, and White women are the most likely to say they took their husband's last name.

Correctness: 1.0 *Coverage: 1.0*

SAIL-VL-2B

Yes, based on data from mid-April 2023, racial differences do affect married women's surname preferences. Married Black women are more likely than White women to say they hyphenated their and their spouse's last names, as compared to White women who are the most likely to say they took their husband's last name.

Correctness: 1.0 *Coverage: 1.0*

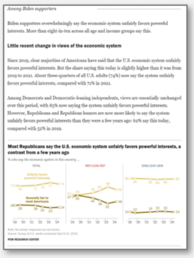
Figure 17: A sample case of Single-Point Text-only question answering.

Single-Point_Chart-only_00472

Query: In 2023, what percentage of Republicans and Republican leaning voters believe that the U.S. economic system is generally fair for most Americans?

GT_Answer: In 2023, 32% of Republicans and Republican leaners believe that the U.S. economic system is generally fair for most Americans.

GT Retrieval content



Paragraph_00123

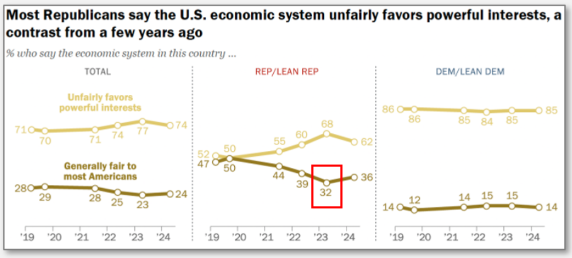


chart_00123_04

Answering

Claude-4.5-Sonnet

Based on the graph, in 2023, 32% of Republicans/Republican-leaning voters believed the U.S. economic system was generally fair to most Americans.

Correctness: 1.0 *Coverage: 1.0*

Llama-3.2-90B-Vision

36%

Correctness: 0.0 *Coverage: 0.0*

Gemini-2.5-Pro

32%

Correctness: 1.0 *Coverage: 1.0*

MiniCPM-V-2.6

In 2023, 47% of Republicans and Republican-leaning voters believe that the U.S. economic system is generally fair for most Americans.

Correctness: 0.0 *Coverage: 0.0*

GPT-5

Sorry, based on the given retrieved content, I cannot answer this question.

Correctness: 0.0 *Coverage: 0.0*

Qwen2-VL-7B

Based on the given retrieved content, I cannot answer this question.

Correctness: 0.0 *Coverage: 0.0*

GPT-4o

In 2023, 39% of Republicans and Republican leaning voters believe that the U.S. economic system is generally fair for most Americans.

Correctness: 0.0 *Coverage: 0.0*

SAIL-VL-2B

50

Correctness: 0.0 *Coverage: 0.0*

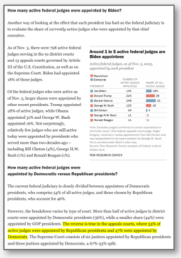
Figure 18: A sample case of Single-Point Chart-only question answering.

Intra-Document_Text-only_00407

Query: What percentage of current appellate court judges were appointed by Democratic presidents? And what percentage were appointed by President Obama?

GT_Answer: Forty-seven percent of current appellate court judges were appointed by Democratic presidents, with Obama appointing 31 percent of current federal judges.

GT Retrieval content



Paragraph_00098

Paragraph_00098_04_04:

The reverse is true in the appeals courts, where 53% of active judges were appointed by Republican presidents and 47% were appointed by Democrats.

Paragraph_00098_03_04:

Trump appointed 28% of active judges, while Obama appointed 31% and George W. Bush appointed 16%.

Answering

Claude-4.5-Sonnet

47% of active appeals court judges were appointed by Democratic presidents.31% of active judges were appointed by President Obama.

Correctness: 1.0 *Coverage: 1.0*

Llama-3.2-90B-Vision

47% of the current appellate court judges were appointed by Democratic presidents, and 31% were appointed by President Obama.

Correctness: 1.0 *Coverage: 1.0*

Gemini-2.5-Pro

47% of current appellate court judges were appointed by Democratic presidents. 31% of active judges were appointed by President Obama.

Correctness: 1.0 *Coverage: 1.0*

MiniCPM-V-2.6

47% of current appellate court judges were appointed by Democratic presidents, and 31% were appointed by President Obama

Correctness: 1.0 *Coverage: 1.0*

GPT-5

47% of current appellate court judges were appointed by Democratic presidents, and 31% were appointed by President Obama.

Correctness: 1.0 *Coverage: 1.0*

Qwen2-VL-7B

47% of current appellate court judges were appointed by Democratic presidents. The percentage of judges appointed by President Obama was not specified in the given content.

Correctness: 0.0 *Coverage: 0.5*

GPT-4o

47% of current appellate court judges were appointed by Democratic presidents, and 31% were appointed by President Obama.

Correctness: 1.0 *Coverage: 1.0*

SAIL-VL-2B

percentage of current appellate court judges appointed by Democratic presidents was 47%. According to the information given, Barack Obama appointed 31% of active judges.

Correctness: 1.0 *Coverage: 1.0*

Figure 19: A sample case of Intra-Document Text-only question answering.

Intra-Document_Chart-only_00407

Query: In Indonesia, how do people perceive the impact of having diverse religions, ethnicities, and cultures on their lives? Among those living in such a diverse environment, what are seen as the main threats to Islam by the country's Muslims?

GT_Answer: 41% of Indonesians believe that having diverse religions, ethnicities and cultures will not have any impact on their lives. At the same time, 54% of Indonesian Muslims believe that Muslim extremists are a threat to Islam, ,33% see the growing number of Buddhists as a threat, and 54% see Muslim extremists as a threat.

GT Retrieval content

Paragraph_00091

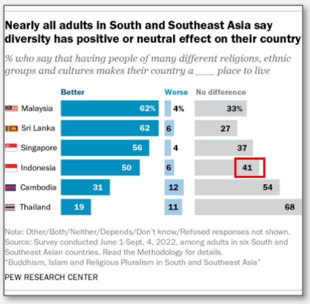


chart_00091_02

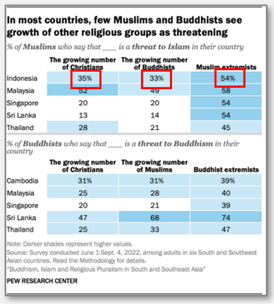


chart_00091_05

Answering

Claude-4.5-Sonnet

In Indonesia, 41% of …, 35% see the growing number of Christians as a threat\n- 33% see the growing number of Buddhists as a threat\n- 54% ...

Correctness: 1.0 ***Coverage: 1.0***

Llama-3.2-90B-Vision

Nearly all adults in Indonesia believe that …, and cultures makes their country a better place to live. …35% of Muslims in …, 33% believe …, and 54% consider Muslim extremists to be a threat."

Correctness: 0.0 ***Coverage: 0.75***

Gemini-2.5-Pro

41% of people in Indonesia believe that diversity has no impact on their lives; the factors considered the biggest threat to Islam in Indonesia are Muslim extremists (54%).

Correctness: 0.0 ***Coverage: 0.5***

MiniCPM-V-2.6

Based on the first image, in Indonesia, 61% of people believe that ... As for the second image, it shows that in Indonesia, 41% of Muslims see the growing number of Christians as a threat to Islam in their country.

Correctness: 0.0 ***Coverage: 0.0***

GPT-5

In Indonesia, 41% of people believe that … the growing number of Christians (35%), the growing number of Buddhists (33%), and Muslim extremists (54%).

Correctness: 1.0 ***Coverage: 1.0***

Qwen2-VL-7B

… 41% of ... However, the retrieved content does not provide information about the percentage of people who believe that diverse religions, ethnicities, and cultures have no impact on their lives in Indonesia. …, 54% …, 74% of ….

Correctness: 0.0 ***Coverage: 0.25***

GPT-4o

41% of people …, the growing number of Christians (35%), the growing number of Buddhists (33%), and Muslim extremists (54%).

Correctness: 1.0 ***Coverage: 1.0***

SAIL-VL-2B

Sorry, based on the given retrieved content, I cannot answer this question.

Correctness: 0.0 ***Coverage: 0.0***

Figure 20: A sample case of Intra-Document Chart-only question answering.

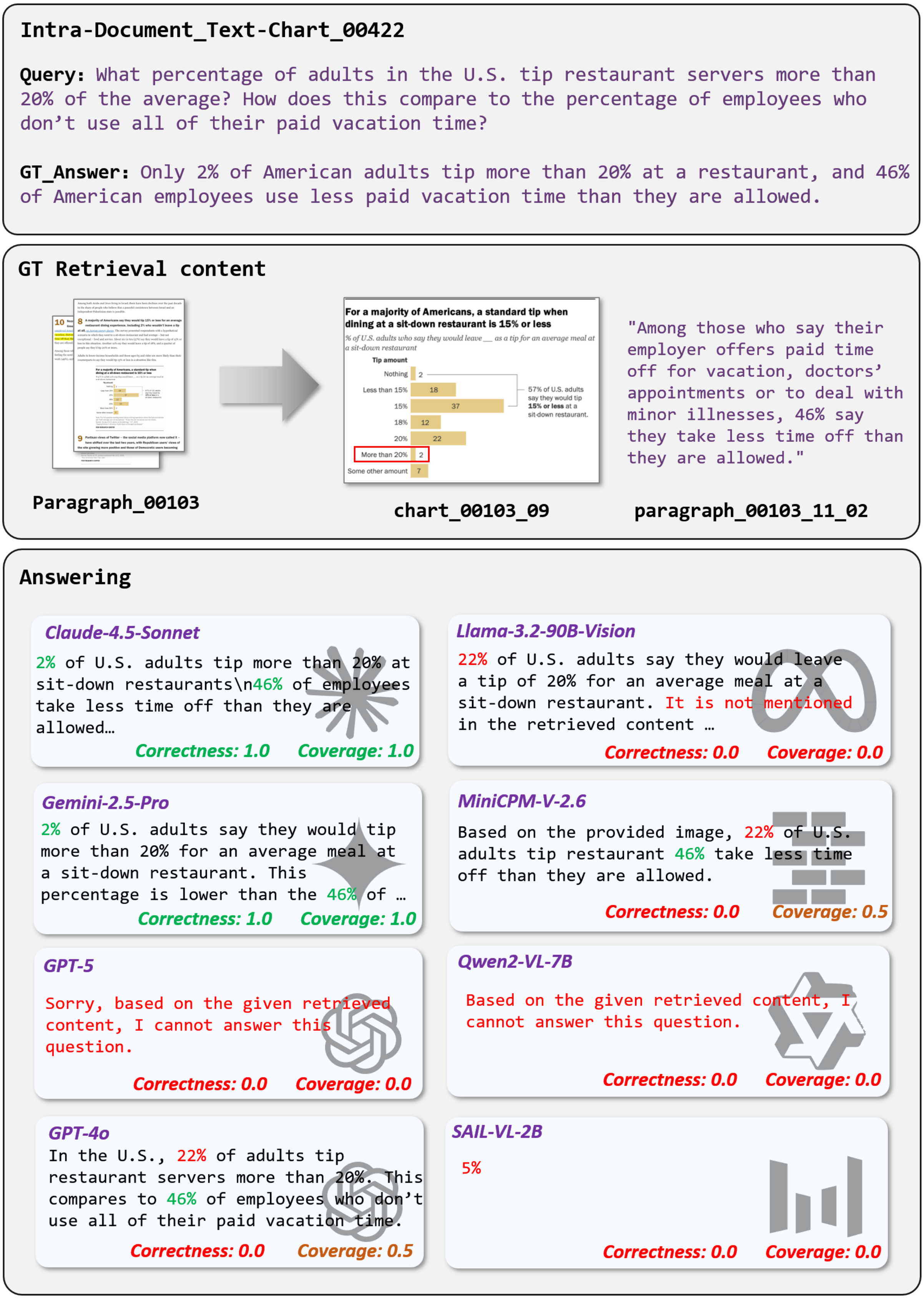


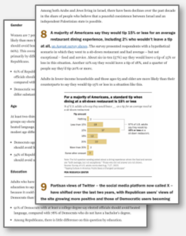
Figure 21: A sample case of Intra-Document Text-Chart question answering.

Inter-Document_Text-only_00281

Query: To what extent do Americans consider it acceptable for political candidates and elected officials to use heated or aggressive language? Given that most Americans find such heated rhetoric acceptable, what primary factor ultimately shapes their opinions of those political figures?

GT_Answer: Americans are almost uniformly open to the use of heated or aggressive language by political candidates and elected officials. Their opinions of political candidates tend to focus more on whether the candidate shares the same political views as them.

GT Retrieval content



Paragraph_00061
Paragraph_00225

paragraph_00061_05_01:
Americans have nearly identical views about the acceptability of political candidates and elected officials using heated or aggressive language.

paragraph_00225_03_08:
The most important factor for Americans is that candidates share their political views.

Answering

Claude-4.5-Sonnet
17% of Conservative Republicans … n\n2. … states that \"Republicans' rating of scientists' understanding of ...
Correctness: 0.0 *Coverage: 0.0*

Llama-3.2-90B-Vision
Americans view the acceptability of ... This perception is closely related to the candidates sharing their political views, which is...
Correctness: 1.0 *Coverage: 1.0*

Gemini-2.5-Pro
Americans hold nearly identical views on the acceptability of political candidates ... Sharing their political views is ...
Correctness: 1.0 *Coverage: 1.0*

MiniCPM-V-2.6
Americans view the acceptability of ... The most important factor for them is that candidates share their political views.
Correctness: 1.0 *Coverage: 1.0*

GPT-5
Americans hold nearly identical views about the acceptability of …. What's most important to them is that candidates share their political views.
Correctness: 1.0 *Coverage: 1.0*

Qwen2-VL-7B
… It states that Americans hold nearly identical views ... is that candidates share their political views.
Correctness: 1.0 *Coverage: 1.0*

GPT-4o
Americans hold nearly identical views about the acceptability of …. What's most important to them is that candidates share their political views.
Correctness: 1.0 *Coverage: 1.0*

SAIL-VL-2B
Americans generally believe that can and should use "heated or aggressive language". they should "share their political views"
Correctness: 0.0 *Coverage: 0.5*

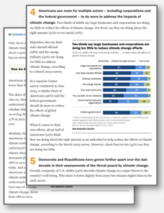
Figure 22: A sample case of Inter-Document Text-only question answering.

Inter-Document_Chart-only_00301

Query: What percentage of U.S. adults in 2023 say that climate scientists have a poor or no understanding of whether climate change is occurring? And what percentage of these adults say that the federal government is doing an adequate or just enough job to help reduce the effects of global climate change?

GT_Answer: Thirty-one percent of U.S. adults say climate scientists have a poor or poor understanding of whether climate change is occurring. Of these people, about 23% say the federal government is doing an adequate or good job of helping reduce the impacts of global climate change.

GT Retrieval content



Paragraph_00077

Paragraph_00004

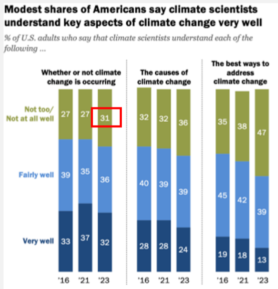


chart_00077_01

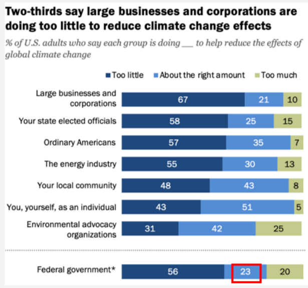


chart_00004_05

Answering

Claude-4.5-Sonnet

Based on … 31% of U.S. adults say climate scientists understand \"Not at all/Not too well\" … , 23% of U.S. adults say …

Correctness: 1.0 ***Coverage: 1.0***

Llama-3.2-90B-Vision

Based on the provided images, 31% of U.S. adults in 2023 say that … . 23% of these adults say that ….

Correctness: 1.0 ***Coverage: 1.0***

Gemini-2.5-Pro

In 2023, 31% of U.S. adults say that …, while 23% say they are doing an adequate or just enough job.\n

Correctness: 0.0 ***Coverage: 0.5***

MiniCPM-V-2.6

Based on the first image, in 2023, 37% of U.S. adults say that … 56% of U.S. adults say that ... .

Correctness: 0.0 ***Coverage: 0.0***

GPT-5

In 2023, 31% of U.S. adults say that … , 23% say that the federal government is doing an adequate …."

Correctness: 1.0 ***Coverage: 1.0***

Qwen2-VL-7B

Based on the given retrieved content, I cannot answer this question.

Correctness: 0.0 ***Coverage: 0.0***

GPT-4o

31% of U.S. adults in 2023 say that … . 23% of these adults say that …."

Correctness: 1.0 ***Coverage: 1.0***

SAIL-VL-2B

The relevant information for answering the question is as follows:\n\n In 2023, 49% of U.S. adults believed that .... 11% of adults said it had the federal ...

Correctness: 0.0 ***Coverage: 0.0***

Figure 23: A sample case of Inter-Document Chart-only question answering.

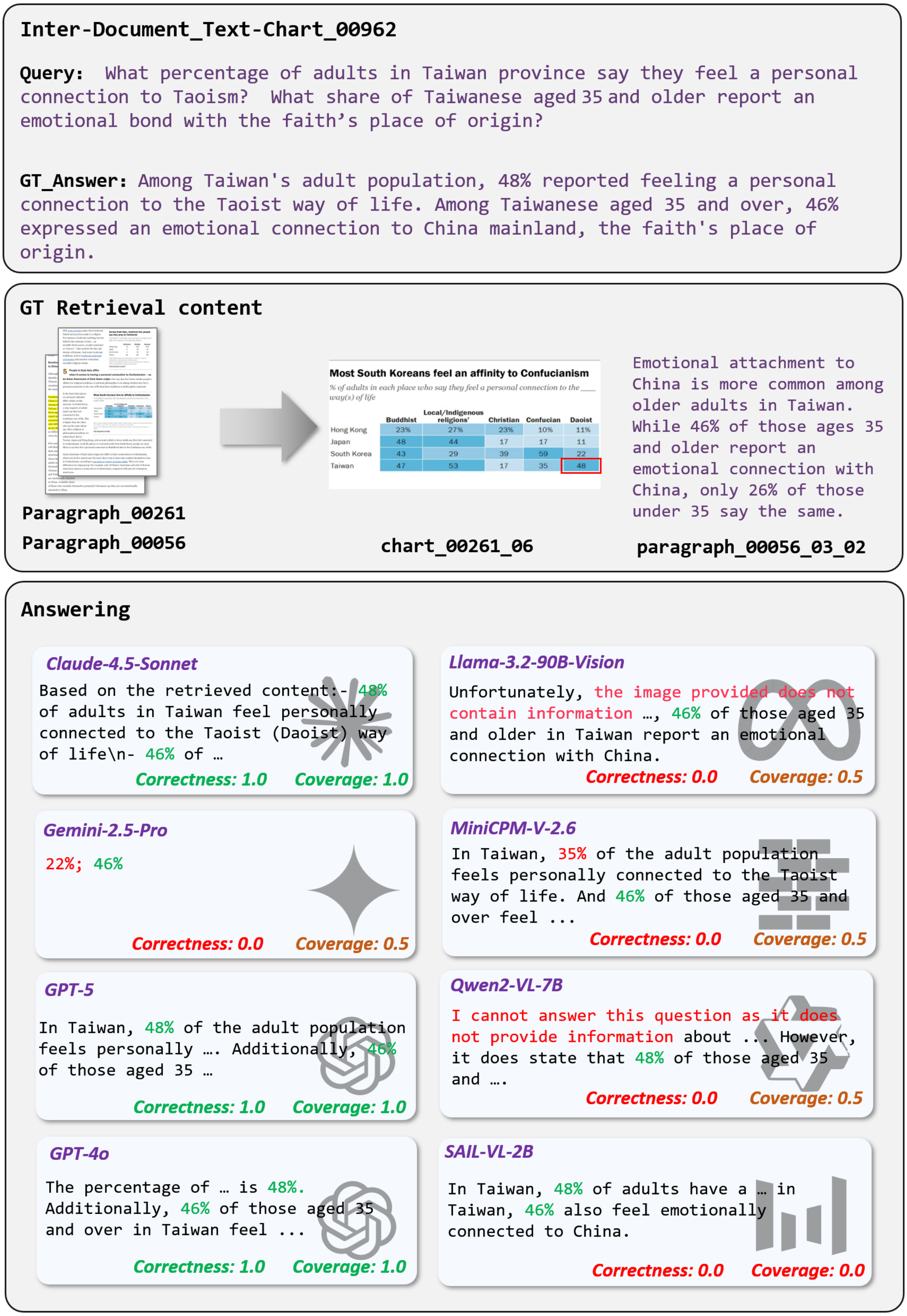


Figure 24: A sample case of Inter-Document Text-Chart question answering.

abling cross-modal matching between queries and documents regardless of their original modality.

Multimodal Embeddings and Combined Vector Stores. In this approach, charts are first converted to text summaries using *GPT-4o*. Both these summaries and PDF-extracted text are then embedded using sparse BM25 and dense embedding models BGE-M3-base/large, E5-base/large into their respective vector stores. Similarity search in these embedding spaces retrieves relevant documents across both modalities.

Multimodal Embeddings and Separate Vector Stores. This approach maintains distinct embedding spaces for different modalities, leveraging specialized models for optimal representation. Charts are encoded using vision-language models (CLIP, JINA-CLIP, SigLIP), while textual content is processed through both sparse retrieval (BM25) and dense embedding models (BGE-M3-base/large, E5-base/large). The retrieval process operates in parallel across separate vector stores, with the final results aggregated using a weighted combination scheme.

.10 Retrieval Metrics

We segment text into semantic chunks with an average length of 24.97 words, while treating each chart as an individual retrieval unit. We employ Recall@5 and Recall@10 as primary retrieval metrics. To ensure balanced representation, we implement a text-to-chart ratio of 3:2 in the final retrieval results.

Given that Chart-MRAG bench primarily consists of multi-hop questions requiring both textual and visual information, the comprehensive retrieval of all relevant sources is crucial for accurate answers. We employ Recall@5 and Recall@10 to evaluate the effectiveness and efficiency of the retrieval stage.

Multimodal Recall. We introduce a Multimodal RAG Retrieval Recall metric to evaluate the effectiveness of cross-modal retrieval process. For textual content, we perform sentence-level retrieval, while for charts, we treat each visualization as an individual reference unit. The Recall is formally defined as

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(M(G_i, \mathcal{R})), \tag{3}$$

where n is the total number of ground truth references (including both text chunks and charts), G_i denotes the i-th ground truth reference, $\mathcal{R} = \{R_1, R_2, \ldots, R_k\}$ represents the set of retrieved references, $M(G_i, \mathcal{R})$ is a boolean function that returns true if (1) for textual references, all constituent sentences in G_i are found in at least one reference in $\mathcal{R}$, or (2) for chart references, the exact chart is present in $\mathcal{R}$, and $\mathbb{1}(\cdot)$ is the indicator function.

This metric assesses the cross-modal alignment between retrieved and ground truth references, where successful retrieval is determined by modality-specific criteria: sentence-level matching for text and exact matching for charts.

.11 Generative Setup

Backbone MLLMs: The backbone of our generative setup comprises several advanced MLLMs, including GPT-4o (version 2024-11-20) (Radford et al., 2021), GPT-5 (Radford et al., 2021), Gemini-2.5-Pro (Team et al., 2024), Claude-4.5-Sonnet (version 2025-09-29) (Awadalla et al., 2023), SAIL-VL-2B (Team, 2024), Qwen2-VL-7B-instruct (Wang et al., 2024), MiniCPM-V-2.6 (8B) (Yao et al., 2024), and Llama-3.2-90B-Vision (Dubey et al., 2024).

To ensure the integrity of our experiments, all closed-source models are accessed via their respective official APIs, utilizing default parameters consistent with the chat mode settings. In contrast, all open-source models are deployed on an 8*A100 GPUs configuration, with precision, temperature, and other parameters strictly adhering to the specifications outlined in the official documentation. This methodological framework guarantees that our experimental conditions are equitable, thereby facilitating a valid comparison across different model architectures.

.12 Generative Metrics

Motivation for New Evaluation Metrics. Traditional question-answering (QA) tasks commonly employ generation evaluation metrics such as ROUGE-L and BLEU-4, which have proven effective in assessing response quality through sentence similarity. However, these metrics exhibit significant limitations when applied to the Chart-MRAG benchmark, where responses primarily focus on numerical accuracy and definitive conclusions rather than linguistic similarity. To illustrate this limitation, we present a representative example from Chart-MRAG benchmark 27.

As shown in Table 9, while both Claude-4.5-Sonnet and GPT-5 correctly identified the 19%

Model	Response	ROUGE-L	BLEU-4
Ground Truth	19% of Democrats or those who lean Democratic believe climate scientists have the appropriate influence.	–	–
Claude-4.5-Sonnet	19% believe scientists have "about the right amount of influence" in debates.	0.54	0.21
GPT-5	19% think scientists have the appropriate amount of influence in policy debates.	0.74	0.42
MiniCPM-V-2.6-8B	75% think scientists have the appropriate amount of influence.	0.74	0.42
SAIL-VL-2B	Cannot answer based on the given content.	0.06	0.00

Table 9: Comparison of model responses on a representative Chart-MRAG example.

figure, MiniCPM-V-2.6-8B provided an incorrect value of 75%. Surprisingly, despite this critical numerical error, MiniCPM-V-2.6-8B achieved the same BLEU-4 score (0.42) as the correct response from GPT-5. This discrepancy clearly demonstrates that traditional metrics fail to capture the essential aspects of chart-only question answering:

- **Numerical Precision**: Traditional metrics may assign high scores to responses with incorrect numerical values if the surrounding text is similar.
- **Factual Accuracy**: Sentence similarity metrics cannot effectively distinguish between correct and incorrect factual conclusions drawn from charts.
- **Response Completeness**: Simple rejection responses (as shown by SAIL-VL-2B) receive low scores under traditional metrics, but fail to reflect whether such responses are appropriate given the chart context.

These observations motivate our introduction of new evaluation metrics specifically designed for chart-only question answering, focusing on response Correctness and Coverage. Our proposed metrics directly address these limitations by emphasizing numerical accuracy and completeness of information extraction from charts.

Generative Metrics Setup. Having highlighted the shortcomings of traditional similarity-based scores for chart-centric QA, we now formalize an evaluation framework that directly targets the two key properties we care about: numerical accuracy and answer completeness. To this end, we introduce two complementary metrics—Correctness and Coverage—that together provide both a strict assessment of exact keypoint matching and a graded measure of how much of the ground-truth information is recovered. In the subsections below, we first define our Correctness metric and then present the continuous Coverage metric, illustrating how each contributes to a more faithful evaluation on the Chart-MRAG benchmark.

Correctness. It measures the exact match between response and ground truth keypoints. Given a question-answer pair $\{Q, A, K^{gt}\}$ with ground truth keypoints $K^{gt} = \{k_1^{gt}, ..., k_n^{gt}\}$, we extract keypoints $K^r = \{k_1^r, ..., k_m^r\}$ from the model's response using an LLM. The score is defined as:

$$\text{Correctness}(K^r, K^{gt}) = \mathbb{1}[K^r \equiv K^{gt}], \quad (4)$$

where $K^r \equiv K^{gt}$ implies complete keypoint matching and equal cardinality. This binary metric requires perfect accuracy, with zero tolerance for missing information or errors.

Coverage. It quantifies the proportion of correctly captured ground truth keypoints:

$$\text{Coverage}(K^r, K^{gt}) = \frac{|K^m|}{|K^{gt}|}, \quad (5)$$

where K^m represents matched ground truth keypoints. This continuous metric in [0,1] enables granular evaluation.

Text-Over-Visual Modality Bias Case

Fig 25 presents a comprehensive analysis of Text-Over-Visual Modality Bias, revealing a systematic preference for text-only processing across different model scales. Our experiments, rigorously verified by human experts, demonstrate that multimodal language models consistently favor text-only responses, even in scenarios where visual elements (particularly charts) contain more precise and relevant information. This bias raises important questions about the effective integration of multiple modalities in current AI systems.

Notably, our investigation reveals a clear correlation between model scale and the ability to handle multimodal information effectively. Larger MLLMs, particularly *GPT-4o*, demonstrate sophisticated capabilities in detecting and managing information redundancy across modalities, proactively acknowledging such overlaps in 23% of their responses. This behavior suggests a more nuanced understanding of the complementary nature of different information sources.

In contrast, smaller models exhibit significant limitations in processing multimodal inputs. For instance, SAIL-VL-2B (2B parameters) shows a stark inability to integrate information across modalities, highlighting the critical role of model scale in achieving effective multimodal reasoning.

Model Prompts

CHARGE framework encompasses multiple stages, each guided by specific prompts designed to facilitate different aspects of the process. We detail these prompts according to their respective stages:

In the Extract Keypoints stage, we employ two specialized prompts: one for document keypoint extraction (Fig. 34) and another for chart keypoint extraction (Fig. 35). These prompts are designed to identify and extract crucial information points from both textual and visual components.

The Cross-modal Verification stage utilizes two key prompts: a keypoint classification prompt (Fig. 36) and a cross-modal information verification protocol (Fig. 37). These prompts work in tandem to ensure the consistency and accuracy of information across different modalities.

For Question-Answer Pair Generation, we implement two distinct protocols: a single-point generation protocol (Fig. 38) for straightforward questions, and a multi-hop generation protocol (Fig. 39) for complex questions requiring multiple reasoning steps.

The Response stage features two prompts: one designed for generating responses without retrieved information (Fig. 40), and another for responses incorporating retrieved information (Fig. 41). This dual approach enables flexible response generation based on available context.

Finally, the Evaluation stage employs two metric calculation prompts: one for assessing correctness (Fig. 42) and another for measuring coverage (Fig. 43). These prompts ensure comprehensive evaluation of the generated responses.

Text-Over-Visual Modality Bias_00011

Query
What percentage of the world's population lives in UTC+8 time zones?

GT Answer - Text
About one-fifth of the world's population lives in UTC+8.

Source – Text
About one-fifth of the world's population lives in this time zone, including all of China. UTC+8 also includes parts of other populous countries or regions, such as the Philippines, Indonesia, and Australia. Another suitable option for holding a meeting is UTC+5.5, which includes India—the most populous country in the world—and Sri Lanka. This time zone accounts for approximately 18% of the global population.

GT Answer - Chart
21% of the world's population lives in UTC+8.

Source – Chart

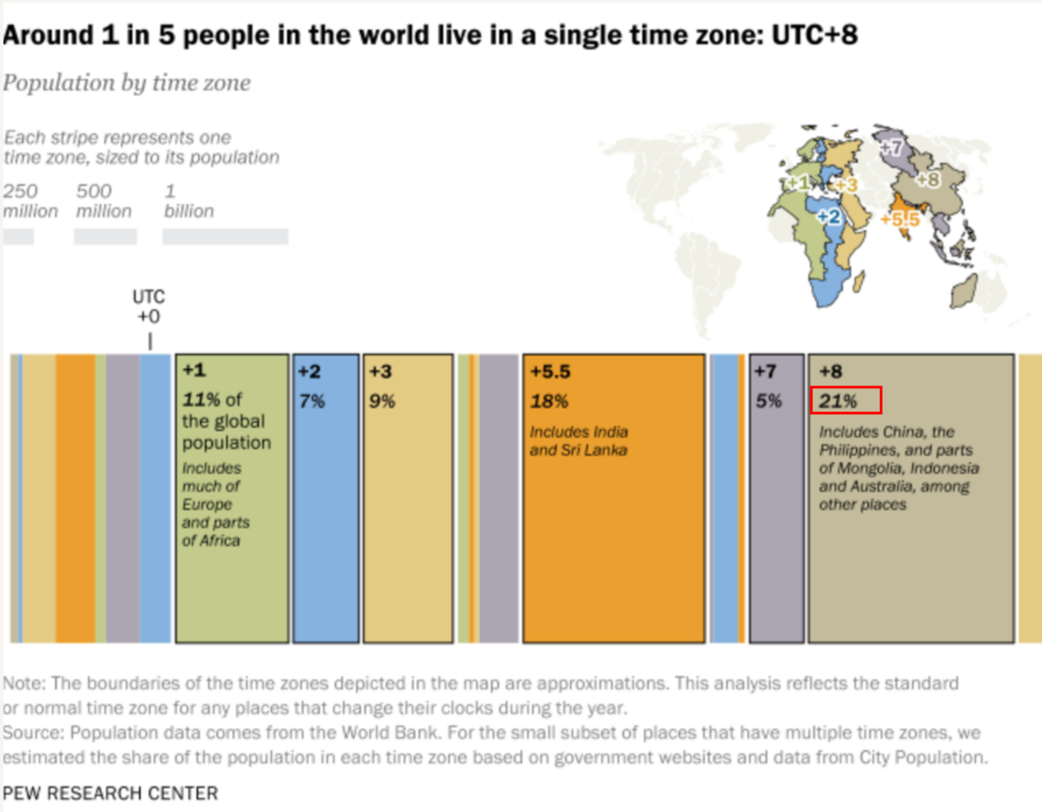


Figure 25: A Sample Case of Text-Over-Visual Modality Bias

Single-Point_Text-only_00316

Query
In 2024, what percentage of U.S. adults will support prioritizing renewable energy over fossil fuel expansion?

Answer
67% of American adults support prioritizing renewable energy over expanding oil, coal and natural gas production.

Source

Source 1
"id": "paragraph_00189_07",
"text": "Democrats are much more likely than Republicans to say protecting the environment (63% vs. 23%) and dealing with climate change (59% vs. 12%) should be top policy priorities for 2024. In fact, addressing climate change ranks last on Republicans' list of priorities this year. Views of the Biden administration's current climate policies also differ sharply by party. Eight-in-ten Democrats say the federal government is doing too little to reduce the effects of climate change, compared with 29% of Republicans, according to a Center survey from spring 2023. Overall, a majority of U.S. adults (67%) support prioritizing the development of renewable energy, such as wind and solar, over expanding the production of oil, coal and natural gas. But Democrats are far more likely than Republicans to prefer this (90% vs. 42%). Still, the public overall is hesitant about a full energy transition: Just 31% say the U.S. should phase out fossil fuels completely."

keypoints
"paragraph_00189_07_04": "Overall, a majority of U.S. adults (67%) support prioritizing renewable energy sources, such as wind and solar, over expanding oil, coal and natural gas production."

Figure 26: A sample case of single-point text-only question answering.

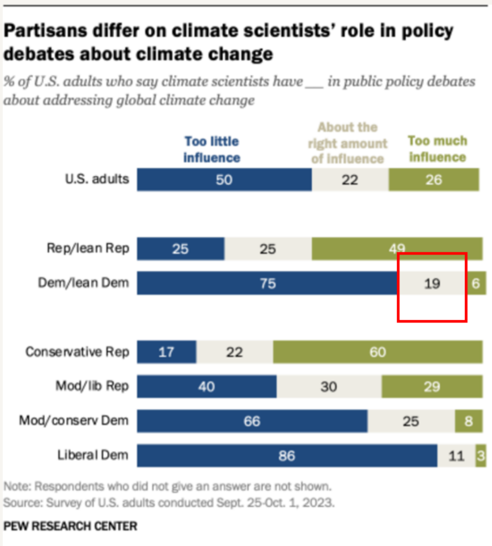
Single-Point_Chart-only_ 00317

Query
In the public policy debate about climate change, what percentage of Democrats or those who lean Democratic think climate scientists have the appropriate amount of influence?

Answer
Nineteen percent of Democrats or those who lean Democratic believe climate scientists have the appropriate influence in the public policy debate over addressing global climate change.

Source

Source 1
"id": "chart_00077_04"



keypoints
"chart_00077_04": "19% of Dem/lean Dem who say climate scientists have About the right amount of influence in public policy debates about addressing global climate change"

Figure 27: A sample case of single-point chart-only question answering.

```
# Intra-Document_Text-only_00571

## Query
In Japan, what percentage of adults say their family owns a cemetery? Among
these adults, how many of those with no religious affiliation have made
offerings to their ancestors in the past 12 months?

## Answer
85% of Japanese adults claim that their family owns a cemetery. Among these
adults, 59% of those with no religious affiliation have prepared food, water, or
beverages for their ancestors in the past 12 months.

## Source

### Source 1
"id": "paragraph_00128_01"
"text": "People in Japan are preparing to celebrate Obon – a festival devoted to
celebrating ancestors that features lighting lanterns and maintaining family
gravesites. In Japan, 85% of adults say their family has such a gravesite, and
79% say they have looked after this gravesite by sweeping or cleaning it in the
past year, according to a recent Pew Research Center survey. Obon …"

### Source 2
"id": "paragraph_00128_02"
"text":" We also asked survey respondents … For instance, 59% of Japan's
religiously unaffiliated adults say they have offered food, water or drinks in
the past 12 months to care for their ancestors. Christians generally are less
likely to engage in these sorts of activities. However, many Vietnamese
Christians have burned incense, offered flowers or lit candles to care for
ancestors in the last year. "

## keypoints
"paragraph_00128_01_02": "85% of adults in Japan say their family has a
gravesite",

"paragraph_00128_02_07": "59% of Japan's religiously unaffiliated adults have
offered food, water, or drinks in the past 12 months to care for their
ancestors"
```

Figure 28: A sample case of intra-document text-only question answering.

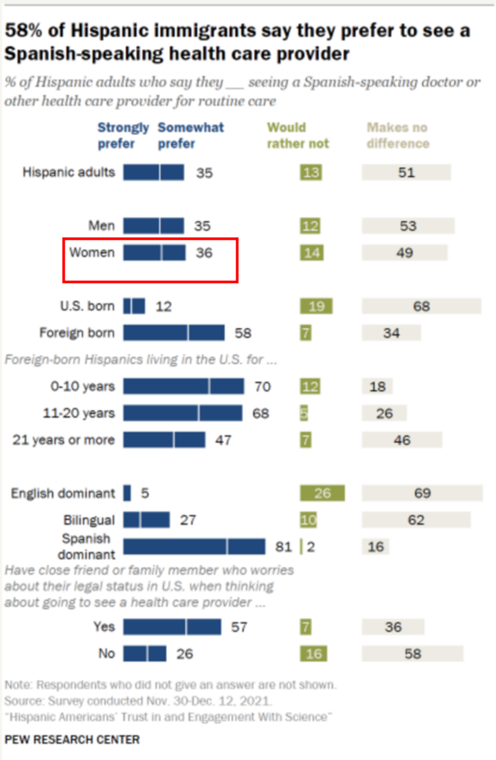
Intra-Document_Chart-only_00257

Query
Among Hispanic women, what percentage strongly prefers a Spanish-speaking physician for routine care? Among these women, what percentage of those with a bachelor's degree or higher have a primary care provider?

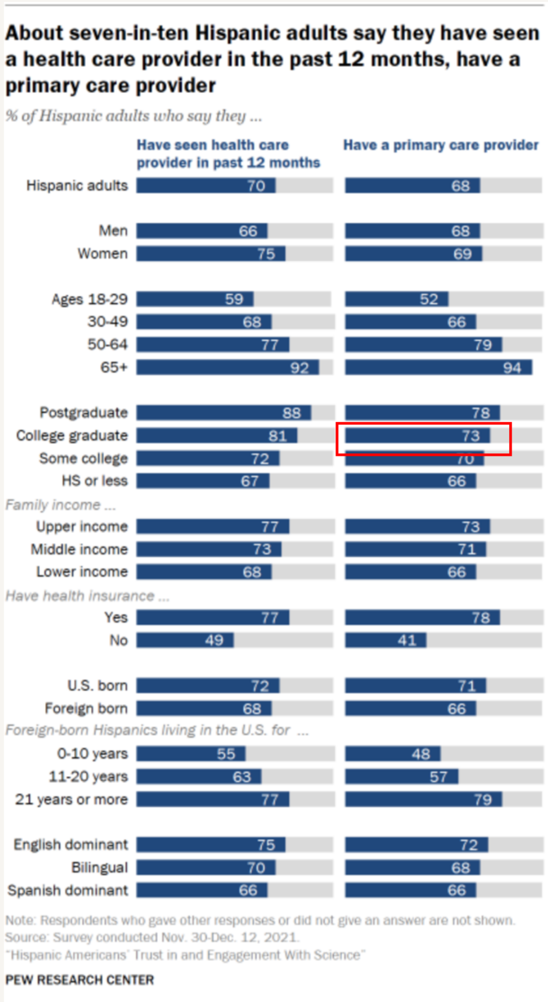
Answer
The data showed that 36% of Hispanic women strongly preferred to have a Spanish-speaking physician for their routine care. Of these women, 73% of college graduates had a primary care provider.

Source

Source 1
chart_00080_05



Source 2
chart_00080_02



keypoints
"chart_00080_05": "36% of Hispanic adults who say they Strongly prefer seeing a Spanish-speaking doctor or other health care provider for routine care among Women",
"chart_00080_02": "73% of Hispanic adults who are College graduate have a primary care provider"

Figure 29: A sample case of intra-document chart-only question answering.

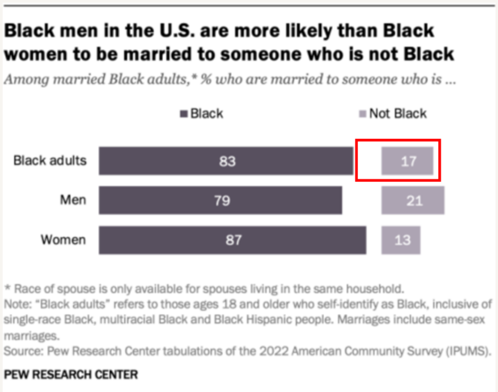
Intra-Document_Text-Chart_00225

Query
Among married black adults, what percentage of them have a non-black spouse? How does this compare to the percentage of black women who are divorced?

Answer
Among married black adults, 17% have a non-black spouse. In comparison, 14% of black women are divorced. This indicates that the rate of interracial marriage among black adults is slightly higher than the divorce rate among black women.

Source

Source 1
"id": "chart_00058_09"



Source 2
"id": "paragraph_00058_08"
"text": "About a third of Black adults (32%) are currently married. That compares with 53% of adults who are not Black. Among Black adults, 36% of men are married, compared with 29% of women. Black women, in turn, are slightly more likely than Black men to be divorced (14% vs. 10%) or widowed (8% vs. 2%). "

keypoints
"chart_00058_09": "17% of married Black adults who are Black adults married to Not Black",

"paragraph_00058_08_02": "14% of Black women are divorced"

Figure 30: A sample case of intra-document text-chart question answering.

Inter_Document_Text-only_00406

Query
How important is spending time with family in the lives of American adults? In comparison, how many adults say having children or being married is equally important to life satisfaction?

Answer
Nine-in-ten Americans say spending time with family is very or one of the most important things to them personally, compared to just 26% who say having children and 23% who say being married is extremely or very important to life satisfaction.

Source

Source 1
"id": "paragraph_00090_02_03"
"text": "According to a Center survey from this spring, a large majority of U.S. adults (73%) say family time is one of the most important aspects of their life – and none of the other priorities we asked about comes close. Overall, nine-in-ten Americans say spending time with family is either very important or one of the most important things to them personally, regardless of how much time they actually devote to it. What's more, spending time with family is the top priority for Americans regardless of political affiliation. About nine-in-ten Republicans and Democrats (including those who lean to each party) say family time is either very important or one of the most important aspects of their life. "

Source 2
"id": "paragraph_00103_02_04"
"text": "According to … A Center survey conducted in April found that relatively few Americans see marriage as essential for people to live a fulfilling life compared with factors like job satisfaction and friendship. While majorities say that having a job or career they enjoy (71%) and having close friends (61%) are extremely or very important for living a fulfilling life, far fewer say this about having children (26%) or being married (23%). "

keypoints
"paragraph_00090_02_03": "Nine-in-ten Americans say spending time with family is either very important or one of the most important things to them personally",

"paragraph_00103_02_04": "Far fewer Americans say having children (26%) or being married (23%) are extremely or very important for living a fulfilling life"

Figure 31: A sample case of inter-document text-only question answering.

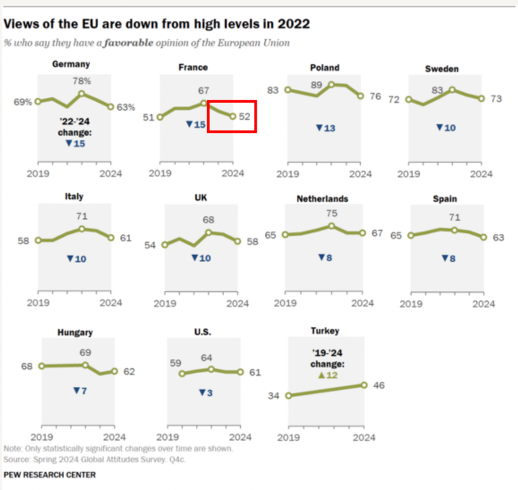
Inter_Document_Chart-only_00513

Query
What percentage of people in France will have a positive view of the EU in 2024? In comparison, what percentage did this hold in Germany in 2019?

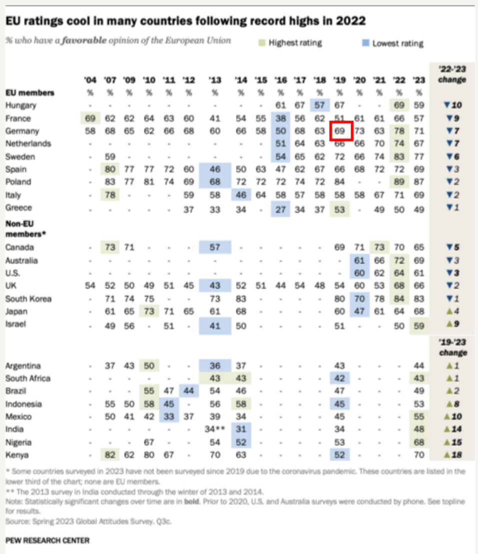
Answer
In 2024, 52% of people in France will have a positive view of the EU. In 2019, this figure was 69% in Germany.

Source

Source 1
"id": "chart_00112_02"



Source 2
"id": "chart_00076_02",



keypoints
"chart_00112_02": "52% of people in France who say they have a favorable opinion of the European Union in 2024",

"chart_00076_02": "69% who have a favorable opinion of the European Union among Germany in 2019"

Figure 32: A sample case of inter-document chart-only question answering.

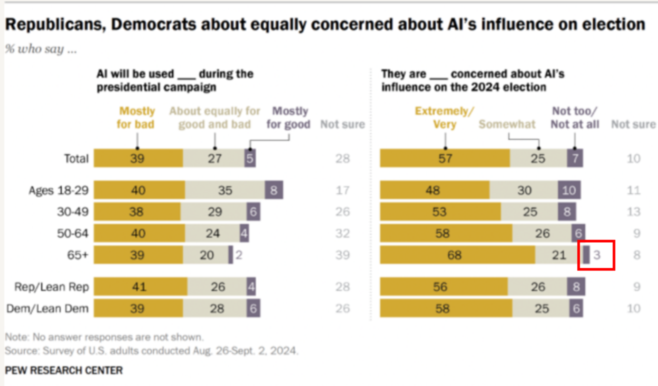
Inter_Document_Text-Chart_01016

Query
Among people aged 65 and older, what percentage are not too concerned or not at all concerned about the impact of AI in the 2024 election? And among these same people, what percentage are mostly concerned about the growing use of AI in daily life?

Answer
Among those 65 and older, 3% are not too concerned or not at all concerned about the impact of AI in the 2024 election. Meanwhile, about 61% of this age group are mostly concerned about the growing use of AI in daily life.

Source

Source 1
"id": "chart_00257_04"



Source 2
"id": "paragraph_00025_01_06"
"text": " A growing share of Americans express concern about the role artificial intelligence (AI) is playing in daily life, according to a new Pew Research Center survey. Overall, 52% of Americans say they feel more concerned than excited about the increased use of artificial intelligence. Just 10% say they are more excited than concerned, while 36% say they feel an equal mix of these emotions. The share of Americans who are mostly concerned about AI in daily life is up 14 percentage points since December 2022, when 38% expressed this view. Concern about AI outweighs excitement across all major demographic groups. Still, there are some notable differences, particularly by age. About six-in-ten adults ages 65 and older (61%) are mostly concerned about the growing use of AI in daily life, while 4% are mostly excited. "

keypoints
"chart_00257_04": "3% of 65+ who are Not too/Not at all concerned about AI's influence on the 2024 election",

"paragraph_00025_01_06": "About 61% of adults ages 65 and older are mostly concerned about the growing use of AI in daily life"

Figure 33: A sample case of inter-document text-chart question answering.

```
# Document Keypoints Extraction

## Task Description
- In this task, you will receive a paragraph and need to summarize definitive key
  points based on the paragraph
- List multiple points as needed, with each key point existing independently,
  supplementing pronouns and subjects.

## Example
Paragraph: Among Americans under 30, 60% have positive views of Palestinians, while
46% have positive views of Israelis. In contrast, older Americans tend to have more
favorable views of Israelis. In recent years, young Americans' favorability toward
Israelis has declined. Since 2019, the percentage of adults under 30 who have
positive views of Israelis has dropped by 17 percentage points, while views of
Palestinians have remained unchanged during this period. Older Americans' views of
both Israelis and Palestinians have remained largely stable.

## Output Example
"60% of American adults under 30 have positive views of Palestinians",
"46% of American adults under 30 have positive views of Israelis",
"Older Americans tend to have more favorable views of Israelis",
"In recent years, young Americans' favorability toward Israelis has declined",
"Since 2019, the percentage of adults under 30 who have positive views of Israelis
has dropped by 17 percentage points",
"Since 2019, views of Palestinians among adults under 30 have remained unchanged",
"Older Americans' views of both Israelis and Palestinians have remained largely
stable “

**Note:**
Output keypoints directly following the example format, without any explanation or
code markup.

## Test
Paragraph:
{paragraph}
```

Figure 34: Document keypoints extraction prompt details.

```
# Chart Keypoints Extraction

## Task Description
- In this task, you will receive nested key-value pairs obtained from OCR and the
  original chart. There are multiple sets of key-value pairs separated by "---"
- Following the specific steps, combine the chart information and key-value
  information to extract the chart's keypoints
- Output the final keypoints directly

## Specific Steps
1. Identify the data_meaning in the chart (gray subtitle, often starting with "%
of" )
2. Step by step, insert each key-value pair into the data_meaning to form
complete keypoints. Each keypoint should contain only one core information, i.e.,
one numerical interpretation
3. Note that keys may have root_key and nested_key structures, which need to be
flexibly inserted into the data_meaning
4. Remove duplicate keypoints, check that all keypoints are complete without
omissions or duplications
5. Output the keypoints

## Output Example
"75% of Israelis who have a Very unfavorable opinion of Prime Benjamin Netanyahu
among Arabs",
"13% of Israelis who have a Somewhat unfavorable opinion of Prime Benjamin
Netanyahu among Arabs",
"7% of Israelis who have a Somewhat favorable opinion of Prime Benjamin Netanyahu
among Arabs",
"2% of Israelis who have a Very favorable opinion of Prime Benjamin Netanyahu
among Arabs",
"24% of Israelis who have a Very unfavorable opinion of Prime Benjamin Netanyahu
among Jews“

**Note:**
1. Some charts may not have nested structures, percentage structures, or
data_meaning. Please observe the chart carefully and construct keypoints flexibly
while ensuring clear reference and complete structure
2. The completeness and clarity of keypoint semantic expression take precedence
over generation strategy requirements.

## Test
key_value:
{key_value}
```

Figure 35: Chart keypoints extraction prompt details.

```
# Keypoint Classification Task

## Task Description
- In this task, you will receive two sets of keypoints: paragraph keypoints and chart keypoints
- You need to carefully analyze and classify all keypoints into two categories: text_only and chart_only
- text_only refers to keypoints that are only expressed in paragraphs and not shown in charts
- chart_only refers to keypoints that only appear in charts and are not mentioned in paragraphs

## Important Notes
- Mixed category keypoints appear in pairs, with one from the paragraph and one from the chart expressing the same information
- For numerical keypoints, if they describe the same object with only differences in numerical precision, they should be classified
as mixed
- If a chart keypoint can lead to a conclusion stated in a paragraph keypoint, it should be classified as mixed
- For keypoints that are truly difficult to classify, maintain their original category
- All output must be in English, and double quotes in keypoints should be converted to single quotes'

## Example
### Input Example
Paragraph keypoints:
'74% of Americans support U.S. participation in international efforts to reduce the effects of climate change',
'36% of Americans think the U.S. is doing more than other large economies to reduce the effects of global climate change',
'Almost one-third of Americans say the U.S. is doing less than other large economies to reduce climate change effects',
'32% of Americans think the U.S. is doing about as much as other large economies in reducing climate change effects',
'The U.S. is the second-largest carbon dioxide emitter, contributing about 13.5% of the global total',
'Half of Americans say the U.S. should do about as much as other large economies to address climate change',
'27% of Americans think the U.S. should do more than others to address climate change',
'In January 2022, 59% of Americans said the U.S. does not have a responsibility to provide financial assistance to developing
countries for renewable energy',
'The UN conference on climate change has addressed how wealthier nations should assist developing countries with climate change',
'COP27 established a "loss and damage" fund for vulnerable countries impacted by climate change'

Chart keypoints:
'30% of U.S. adults who say U.S. is doing Less than others',
'32% of U.S. adults who say U.S. is doing About as much',
'36% of U.S. adults who say U.S. is doing More than others',
'15% of U.S. adults who say U.S. should be doing Less than others',
'56% of U.S. adults who say U.S. should be doing About as much',
'27% of U.S. adults who say U.S. should be doing More than others'

### Output Example

"text_only": [
"74% of Americans support U.S. participation in international efforts to reduce the effects of climate change",
"The U.S. is the second-largest carbon dioxide emitter, contributing about 13.5% of the global total",
"In January 2022, 59% of Americans said the U.S. does not have a responsibility to provide financial assistance to developing
countries for renewable energy",
"The UN conference on climate change has addressed how wealthier nations should assist developing countries with climate change",
"COP27 established a 'loss and damage' fund for vulnerable countries impacted by climate change"
],
"chart_only": [
"15% of U.S. adults who say U.S. should be doing Less than others"
]


## Test
Please output the classification results in JSON format directly, without any additional content like '''json''' or explanations!

### Input
Paragraph keypoints:
{paragraph_keypoints}

Chart keypoints:
{chart_keypoints}

### Output:
```

Figure 36: Keypoint classification task prompt details.

Cross-Modal Information Verification Protocol

Task Overview
This protocol establishes a systematic approach for verifying whether given question-answer pairs can be validated using provided background information (text paragraph or visual chart).

Methodology

Primary Verification Process
1. Given: Background information (textual content or data visualization) and a question-answer pair (Q, A_1)
2. Generate: Independent answer (A_2) using only the background information
3. Evaluate: Compare A_1 and A_2 to determine answer derivability from background information

Verification Criteria
1. **Positive Verification**:
   - Question is verifiable if the background information supports the same conclusion
   - Minor variations in reasoning or expression are acceptable if conclusions align

2. **Negative Verification**:
   - Question is unverifiable if:
     a. Background information is insufficient for answer derivation
     b. Numerical values cannot be computed from provided data
     c. Verification certainty cannot be established

Response Protocol
- Verifiable: Output "The background data can answer this question."
- Unverifiable: Output "The background data cannot answer this question."

Test

Input
- Background Information: {background_info}
- Question-Answer Pair: {qas}

Output

Figure 37: Cross-modal information verification protocol prompt details.

```
# Single-Point Question-Answer Generation Protocol

## Objective
Generate deterministic question-answer pairs from background information and key points with precise
correspondence to source material.

## Methodology

### 1. Context Analysis

Analyze background information for:
- Subject matter and core assertions
- Temporal, spatial, and entity parameters

### 2. Construction Process

- Generate specific, closed-ended questions directly linked to key points
- Prioritize quantitative data where available
- Maintain numerical precision (2 decimals for calculations, whole numbers for percentages)
- Ensure natural language flow and clear reference resolution

## Examples

Background: "This article examines U.S. public attitudes toward digital privacy and data usage,
focusing on trust and concerns about corporate AI use."

Keypoint: "paragraph_xxxxx_xx": "62% of those who have heard of AI say companies using it to analyze
personal details could make life easier"

## Output Example

{
    "question": "What percentage of people who have heard of AI believe corporate AI analysis of
personal data could improve convenience?",
    "answer": "62% of people who have heard of AI believe that companies using AI to analyze personal
details could make life more convenient.",
    "keypoints": {
        "paragraph_xxxxx_xx": "62% of those who have heard of AI say companies using it to analyze
personal details could make life easier"
    }
}

## Test
Input:
- Background: "{background}"
- Key Point: "{keypoint_1}"

## Output
Direct JSON without code markers
```

Figure 38: Single-point question-answer generation protocol prompt details.

```
# Multi-Hop Question-Answer Generation Protocol

## Objective
Generate high-quality question-answer pairs by combining two related key points to:
- Build knowledge base question-answer data
- Train models to understand inter-point relationships
- Generate natural conversational content

## Methodology

### 1. Key Point Selection
Select optimal key point combinations based on relationship types:
- Entity coherence (shared subjects across points)
- Causal relationships
- Complementary data (numbers and percentages)
- Comparative relationships (temporal/spatial)

### 2. Construction Parameters
- Questions must require both key points for complete answers
- Natural progression in questioning sequence
- Closed-ended questions with clear point traceability

## Example

Background: Study of Venezuelan immigrant demographics in the United States

Key Point 1: "paragraph_xxxxx_xx": "The number of Venezuelan immigrants who have lived in the United
States for five years or less increased from 40,000 in 2010 to 215,000 in 2021"

Key Point 2: "chart_xxxxx_xx": "62% of Venezuelan immigrants have lived in the United States for less
than 10 years"

## Output Example
{
   "question": "Among Venezuelan immigrants in the US, what percentage lived there less than 10 years,
and how many specifically had resided there for 5 years or less by 2021?",
   "answer": "62% of Venezuelan immigrants had lived in the United States for less than 10 years.
Within this group of recent immigrants, 215,000 had resided in the country for 5 years or less as of
2021.",
   "keypoints": {
       "paragraph_xxxxx_xx": "The number of Venezuelan immigrants who have lived in the United States
for five years or less increased from 40,000 in 2010 to 215,000 in 2021",
       "chart_xxxxx_xx": "62% of Venezuelan immigrants have lived in the United States for less than
10 years"
   }
}

## Test
Input:
- Key Point 1: "{keypoint_1}"
- Key Point 2: "{keypoint_2}"

## Output
Direct JSON without code markers
```

Figure 39: Multi-hop question-answer generation protocol prompt details.

```
# Response without Retrieved Information
You are a rigorous AI assistant. Please answer questions directly and concisely.

## Requirements
- I will provide you with a question. Please answer directly without additional
explanation
- Each question contains 1-2 key points that need to be addressed
- Each key point should be answered with a single, concise sentence

## Example
- Query: How many Americans felt anxious about the COVID-19 pandemic in 2020? And how many
expressed confidence in the government's response?

- Answer: 67% of Americans reported feeling anxious about the pandemic. Only 23% of
Americans expressed confidence in the government's handling of COVID-19.

## Test
- Based on the following question, provide your direct answer

### Input
- Question: {query}

### Output
```

Figure 40: Response without retrieved information prompt details.

```
# Response with Retrieved Information
You are a rigorous AI assistant. Please answer questions based on the following retrieved
images and text content.

## Requirements
- Provide answers based on the retrieved content to the best of your ability.
- If the retrieved content is insufficient to answer the question, please respond with
"Sorry, based on the given retrieved content, I cannot answer this question".
- Your answers should be accurate, concise, and directly address the key points without
additional explanations or comments.

## Test

### Image Content
Please refer to the image I provided

### Text Content
{text_content}

### Input
- Question: {query}

### Output
```

Figure 41: Response with retrieved information prompt details.

```
# Correctness Metric Calculation

## Metric Definition
Assesses answer accuracy by comparing numerical values against key point references. Requires
complete precision in all values with no partial credit allowance. This metric evaluates answer
exactness.

## Calculation Method
- Correctness can only be 0.0 or 1.0
- All numerical values match key points: Correctness = 1.0
- Any incorrect or missing value: Correctness = 0.0

## Calculation Rules
1. Numeric Precision
  - Values must exactly match key points
  - No rounding or approximation allowed
  - Units must be consistent
2. Completeness
  - All requested values must be present
  - Missing any value results in 0.0 score
3. Value Validation
  - Compare each number against key points
  - Check for exact matches only
  - No tolerance for variation

## Examples
### Basic Example
Question: What is A? What is B?
Key Points: A=8, B=22
Answer 1: A=8, B=22  -> Correctness = 1.0 (all correct)
Answer 2: A=8, B=25  -> Correctness = 0.0 (contains error)
Answer 3: A=8        -> Correctness = 0.0 (incomplete)

### Complex Example
{
   "keypoints": [
       "Arab-Jewish conflict rated 'very strong' by 46%",
       "Left-right political divide rated 'very strong' by 32%",
       "Inter-party supporter conflict rated 'very strong' by 19%",
       "Arab-Jewish tension highest at 46%"
   ],
   "model_answer": "Proportion rating conflict as 'very strong': Arab-Jewish: 46%; Left-right
political: 32%; Inter-party supporters: 19%. Highest is Arab-Jewish at 46%."
}

Output:
{
   "correctness": 1.0
}

## Test

### Input
{test_input}

## Output
Direct JSON format response without explanation or code tags
```

Figure 42: Correctness metric calculation prompt details.

```
# Coverage Metric Calculation

## Metric Definition
Measures the ratio of correctly answered key points to total required key points, allowing partial
credit. Note: Values must be exactly correct to count towards Coverage.

## Calculation Method
- Coverage = Number of correctly matched key points / Total number of required key points
- All key points present in answer: Coverage = 1.0
- M out of N key points present (M<=N): Coverage = M/N
- No key points present: Coverage = 0.0

## Calculation Rules
1. Value Matching
  - Each key point must be exactly matched
  - No partial credit for approximate values
  - Units must be consistent
2. Score Calculation
  - Count correctly matched points
  - Divide by total required points
  - Result is always between 0.0 and 1.0
3. Key Point Validation
  - Compare each value independently
  - Only exact matches count
  - Missing or incorrect values excluded

## Examples

### Basic Example
Question: What is A? What is B?
Key Points: A=8, B=22
Answer 1: A=8, B=22  -> Coverage = 2/2 = 1.0 (full coverage)
Answer 2: A=8, B=25  -> Coverage = 1/2 = 0.5 (one correct)
Answer 3: A=8        -> Coverage = 1/2 = 0.5 (only one correct)
Answer 4: A=5, C=22  -> Coverage = 0/2 = 0.0 (none correct)

### Complex Example
{
   "keypoints": [
       "Arab-Jewish conflict rated 'very strong' by 46%",
       "Left-right political divide rated 'very strong' by 32%",
       "Inter-party supporter conflict rated 'very strong' by 19%",
       "Arab-Jewish tension highest at 46%"
   ],
   "model_answer": "Proportion rating conflict as 'very strong': Arab-Jewish: 46%; Left-right
political: 23%; Inter-party supporters: 19%. Highest is Arab-Jewish at 46%."
}

Output:
{
   "coverage": 0.75
}

## Test

### Input
{test_input}

## Output
Direct JSON format response without explanation or code tags
```

Figure 43: Coverage metric calculation prompt details.