KoTextVQA: A Benchmark for Understanding and Reasoning in Korean Text-Rich Visual Question Answering

Anonymous ACL submission

Abstract

In real-world scenarios, text within images plays a crucial role in conveying information across various domains, including documents, everyday environments, and digital interfaces. Understanding text within its visual context remains a fundamental challenge for Vision-Language Models (VLMs), driving the development of text-rich Visual Question Answering (VQA) datasets and evaluation benchmarks. However, low-resource languages remain underexplored, lacking appropriate benchmarks for real-world applications. In the absence of such benchmarks, systematic evaluation becomes challenging, hindering iterative advancements in model performance and the refinement of fine-tuning strategies. To address this, we introduce **KoTextVQA**, a **Ko**rean **Text**-rich **VQA** benchmark for comprehensive VLM evaluation. KoTextVQA enables an in-depth evaluation of visual text understanding (System 1) and reasoning (System 2) capabilities, while also supporting a multifaceted assessment across diverse image types and domains. Additionally, we release an automated VQA generation pipeline that leverages de facto standard models to efficiently construct benchmarks, enabling the scalable creation of high-quality datasets. While our benchmark is designed for Korean, the proposed methodology is highly adaptable and can be extended to other languages, supporting broader multilingual VLM research.

1 Introduction

One of the core challenges in vision-language integration is the effective interpretation of textual content within images. While text often conveys essential information, ranging from structured documents to everyday signage, many existing models struggle to accurately capture and reason about textual elements in realistic settings. Although recent Vision-Language Models (VLMs) (Liu et al., 2023; Bai et al., 2023; Wang et al., 2024) have achieved significant progress in text-rich Visual Question Answering (VQA) (Singh et al., 2019; Mishra et al., 2019), these advances largely concentrate on high-resource languages, benefiting from abundant data and standardized benchmarks that facilitate systematic evaluation. In contrast, low-resource languages lack well-curated datasets, making it difficult for researchers and practitioners to diagnose specific model shortcomings or devise effective training strategies. Without robust benchmarks and well-defined evaluation protocols, refining VLMs for diverse linguistic and cultural contexts continues to be a significant challenge.

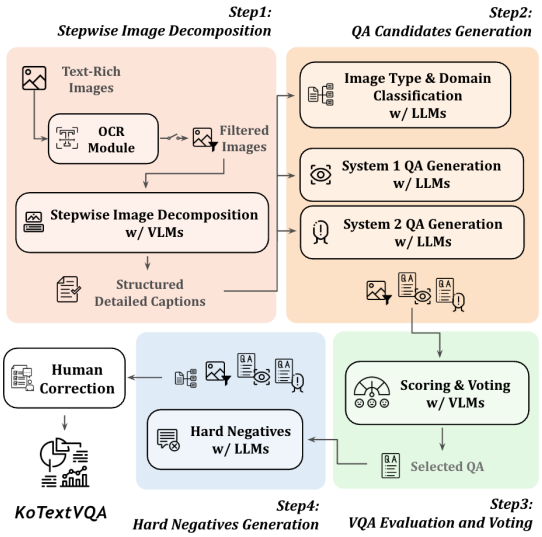

Figure 1: Automated VQA generation pipeline for text-rich images

Recently, several multilingual text-VQA benchmarks (Tang et al., 2024b; Sun et al., 2024) have been proposed; however, covering all languages in depth remains challenging, and Korean is no exception to this limitation. Existing Korean VQA benchmarks (Ju et al., 2024), often focus on document-based tasks or translated English datasets, overlooking the varied real-world scenarios such as

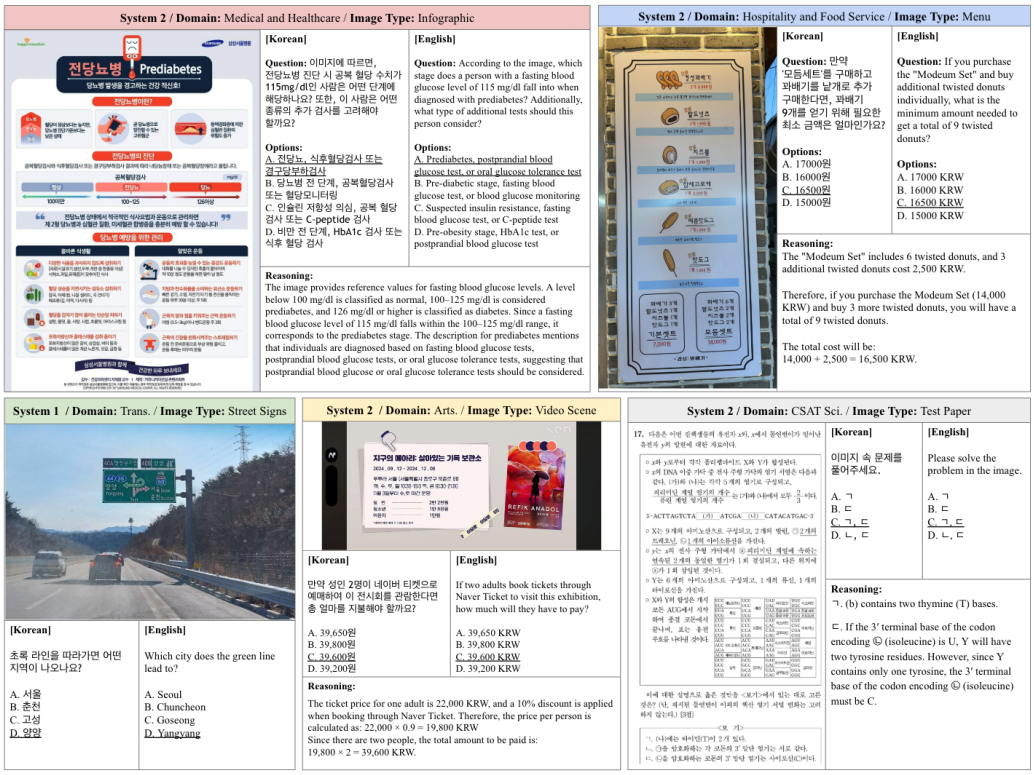

Figure 2: Examples from KoTextVQA, showcasing diverse domains and image types categorized under System 1 and System 2. The model input consists of an image, a Korean question, and multiple-choice options.

infographics, public signage, and digital interfaces, where text frequently serves domain-specific communicative purposes.

For instance, as shown in Figure 2, a prediabetes poster in the Medical & Healthcare domain use an infographic format to convey health guidelines. However, recognizing all textual content does not guarantee correct answers if the model lacks domain knowledge or the ability to interpret image structure. Some tasks further require complex reasoning, adding another layer of difficulty. Despite these challenges, no comprehensive benchmark fully accounts for domain-specific nuances, diverse image types, and varying levels of cognitive demand.

To address these gaps, we introduce KoTextVQA, a benchmark specifically designed to evaluate VLMs on Korean text-rich images. Our contributions are threefold:

1. **A structured and multi-faceted evaluation framework:** We adopt a dual-level reasoning framework (*System 1* for basic understanding and *System 2* for advanced reasoning) to evaluate both visual text recognition and reasoning tasks. Additionally, we classify images by *domain* and *image type* to better reflect real-world contexts where textual content serves diverse functions.

2. **An automated VQA generation pipeline:** We develop a systematic and scalable multi-step pipeline that leverages de facto standard models for dataset construction, incorporating stepwise image decomposition, QA candidate generation, evaluation and voting, and hard negative option generation. This ensures a rigorous benchmark with high data quality and reliability.

3. **A comprehensive benchmark for low-resource language:** By integrating the above approaches, KoTextVQA establishes the text-rich VQA benchmark for Korean. We further release our prompts and code to facilitate adaptation to other low-resource languages, supporting broader multilingual VLM research.

By providing a scalable and culturally adaptive evaluation framework, KoTextVQA offers deeper insights into how VLMs process Korean text-rich images while guiding the development of domain-specific fine-tuning strategies and robust reasoning mechanisms for low-resource languages.

2 Related Work

2.1 Vision-Language Models

Recent advancements in VLMs (Liu et al., 2023; Wang et al., 2024; Dai et al., 2023; Abdin et al., 2024; Wu et al., 2024; Chen et al., 2024; Yao et al., 2024) have broadened their capabilities beyond traditional computer vision tasks, enabling not only contextual interpretation and reasoning across various visual domains but also a deeper integration of language and vision. However, general-purpose VLMs often struggle with text-rich images, as they focus on holistic scene interpretation rather than precise text comprehension. To address this, text-centric VLMs such as LLaVAR (Zhang et al., 2024b), LLaVA-Read (Zhang et al., 2024a), and TextSquare (Tang et al., 2024a) enhance reading abilities by refining text recognition and reasoning. While these models improve performance on text-heavy tasks, they remain largely limited to English, highlighting the need for multilingual VLMs capable of handling diverse linguistic contexts, a critical challenge for Text-Rich VQA benchmarks.

2.2 Text-Rich VQA Benchmarks

Although general VQA benchmarks (Lu et al., 2022; Yue et al., 2024a; Yuan Liu, 2023) assess broad reasoning capabilities, benchmarks focused on text-rich VQA remain limited, especially across diverse languages. TextVQA (Singh et al., 2019) and OCR-VQA (Mishra et al., 2019) primarily address English text, focusing on images with printed content such as billboards and book covers. MTVQA (Tang et al., 2024b) provides multilingual annotations but is constrained in scale, while MUST-VQA (Vivoli et al., 2022) expands existing datasets through automatic translation, which may fail to preserve language-specific nuances. Most text-focused VQA datasets prioritize high-resource languages or rely on translated English benchmarks, with limited support for Korean (Sun et al., 2024; Yue et al., 2024b). This lack of dedicated benchmarks hinders systematic evaluation, fine-tuning, and model improvement for Korean text-rich images, which KoTextVQA aims to address.

3 KoTextVQA Benchmark

As shown in Table 1, the KoTextVQA benchmark is carefully designed to evaluate the ability of VLMs to understand and reason about text appearing in images, spanning a diverse range of real-world contexts. The following subsections detail the dataset statistic and categorization, the data collection process, the automated VQA generation pipeline, and the human annotation refinement process.

Benchmark	Image Type	Forms	Text-Centric Reasoning	Image Source	Samples
KVQA (Kim et al., 2019)	General	Short	-	Ko	100,445
MTVQA-ko (Tang et al., 2024b)	Multi-text	Short	-	Ko	558
K-Viscuit (Baek et al., 2024)	General	MC	-	Ko	657
K-MMB (Ju et al., 2024)	General	MC	-	En	4,329
K-SEED (Ju et al., 2024)	General	MC	-	En	2,971
K-MMSTAR (Ju et al., 2024)	General	MC	✓	En	1,500
K-DTCBench (Ju et al., 2024)	Document	MC	✓	Ko	240
K-LLaVA-W (Ju et al., 2024)	General	Open	-	En	60
KoTextVQA (ours)	Multi-text	MC	✓	Ko	2,577

Table 1: Overview of Korean VQA Benchmarks. The Image Type column distinguishes between *Document* (structured text images) and *Multi-text* (diverse text-rich images). The Forms indicates whether the benchmark uses Open-ended (*Open*), Short answer (*Short*), or Multiple-choice (*MC*) questions. The Image Source column differentiates datasets with images originally in Korean (Ko) from those translated from English (En).

3.1 Data Statistics and Categorization

Our benchmark consists of 2,577 samples, each annotated with corresponding QA pairs. The images cover 26 distinct types across 15 domains. Each image is categorized into one or both reasoning levels: *System 1* (basic recognition and understanding) and *System 2* (advanced reasoning). In total, the dataset includes 1,426 System 1 QA pairs and 1,151 System 2 QA pairs. Beyond the in-depth analysis provided by the reasoning-based categorization, we conduct a multi-faceted analysis of VLM performance by categorizing images along two additional dimensions: *Domain* and *Image Type*.

System 1 vs. System 2 To assess challenges in visual text understanding, we adopt a two-tiered cognitive framework (Kahneman, 2011) distinguishes basic recognition (System 1, fast thinking) from advanced reasoning (System 2, slow thinking). System 1 relies on intuitive and automatic recognition, requiring direct text extraction and straightforward interpretation. In contrast, System 2 demands advanced reasoning, such as contextual understanding, multi-step decisions, numerical reasoning (e.g., mathematical calculations) and integration of external knowledge when necessary. By incorporating both reasoning levels into our benchmark, we provide a comprehensive framework for the in-depth evaluation of VLM capabilities, from fundamental recognition to high-level reasoning.

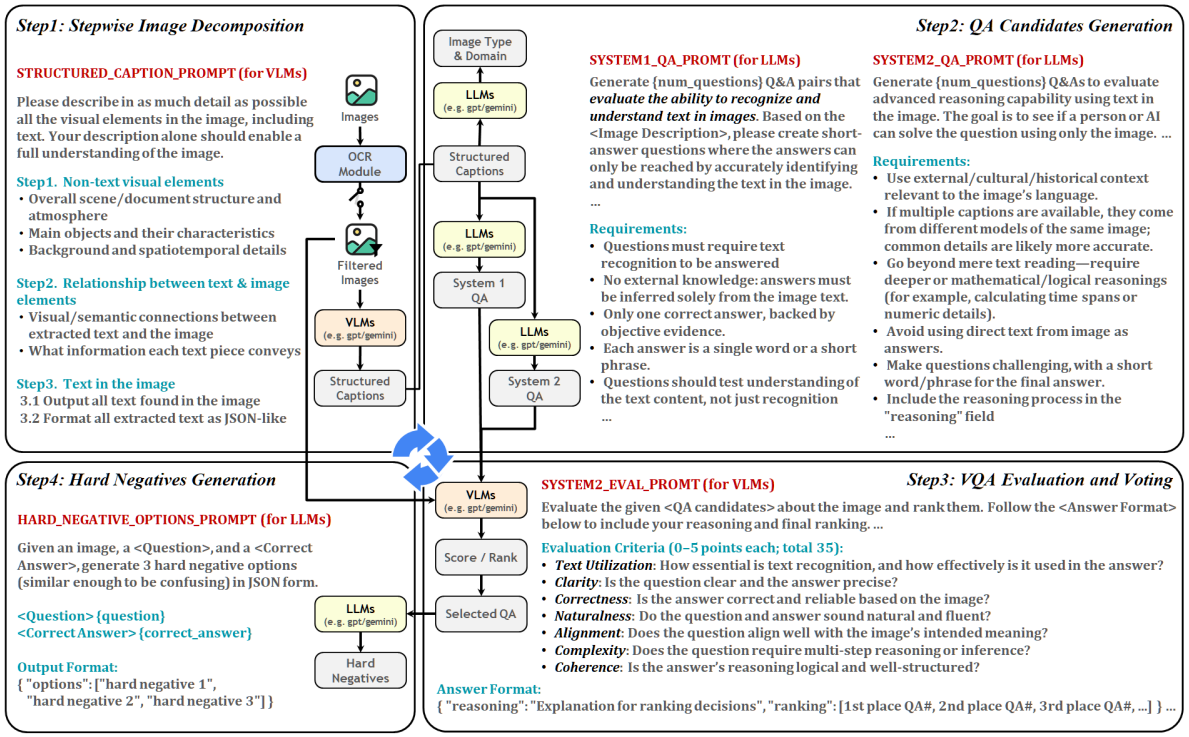

Figure 3: An overview of the automated VQA generation pipeline with prompts. Each step involves data processing using either VLMs or LLMs, with corresponding prompts shown in the figure. The actual data generation process uses Korean prompts. Some prompts, such as examples, are shortened or omitted for readability.

Domain To ensure that our domain classification aligns with real-world industrial applications, we refer to the Korean Standard Industrial Classification (KSIC) (Statistics, 2024) framework. We adapt this framework to suit our image data analysis, following a structured approach similar to the MMMU (Yue et al., 2024a), we define 13 primary domains: Public & Administration, Economics & Finance, Marketing & Advertising, Retail & Commerce, Education & Academia, Medical & Healthcare, Science & Technology, Arts & Humanities, Transportation & Logistics, Travel & Tourism, Hospitality & Food Service, Entertainment & Media, and Personal & Lifestyle.

In addition, we incorporate CSAT (College Scholastic Ability Test) Science and History as separate domains. Unlike other domains generated through our pipeline, CSAT questions are carefully tailored from existing exam materials to ensure authenticity and alignment with real-world assessments. All questions within these domains are categorized as System 2 because they require complex reasoning across diverse disciplines. Collectively, these 15 domains establish a comprehensive framework for evaluating VLMs across diverse contexts and provide insights for fine-tuning or domain-specific training strategies.

Image Type Images are categorized based on their inherent visual structures and the way they convey information. To systematically analyze VLM performance across different visual formats, we classify all images into 26 distinct image types, each representing a specific mode of text presentation. These categories include charts and plots, infographics, posters, mobile/PC screenshots, manuals, receipts, street signs, menus, among others, spanning a spectrum from highly structured formats (e.g., tables, receipts) to more dynamic and visually complex representations (e.g., posters, PC screenshots). By leveraging the image type classification in KoTextVQA, we aim to examine whether VLMs exhibit consistent performance across different text-rich visual formats and to identify weaknesses in processing specific image types.

3.2 Data Collection

For this study, we have compiled a dataset of images by sourcing them from diverse online repositories with no copyright restrictions and by directly capturing original photographs. To ensure comprehensive and balanced coverage of real-world scenarios, we identify domain imbalances and mitigate them by adding more images.

Model	Size	Overall (2,577)	System 1 (1,426)	System 2 (1,151)
closed VLMs				
GPT-4o (OpenAI, 2024)	-	84.6	95.9	**70.5**
GPT-4o-mini (OpenAI, 2024)	-	73.3	88.7	54.1
Gemini-2.0-flash (DeepMind, 2025)	-	**85.4**	**98.0**	69.8
Claude-3.5-Sonnet (Anthropic, 2024)	-	80.5	93.4	64.5
Open-source VLMs				
LLaVA-OneVision (Li et al., 2024)	0.5B	42.3	49.6	33.3
Deepseek-VL2-tiny (Wu et al., 2024)	1B	48.8	60.8	34.0
Deepseek-VL2-small (Wu et al., 2024)	2.8B	53.3	67.3	36.1
Qwen2.5-VL (Wang et al., 2024)	3B	**71.8**	**94.2**	43.9
Ovis1.6-Llama3.2 (Lu et al., 2024)	3B	52.2	62.8	39.1
InternVL2.5 (Chen et al., 2024)	4B	70.7	90.7	**45.9**
Phi-3.5-Vision (Abdin et al., 2024)	4.2B	42.6	52.2	30.8
LLaVA-OneVision (Li et al., 2024)	7B	54.0	65.1	40.1
Qwen2.5-VL (Wang et al., 2024)	7B	68.5	**94.5**	36.1
InternVL2.5 (Chen et al., 2024)	8B	70.8	89.8	47.3
MiniCPM-V-2.6 (Yao et al., 2024)	8B	41.0	50.4	29.4
MiniCPM-o-2.6 (Yao et al., 2024)	8B	64.3	84.1	39.9
Ovis1.6-Gemma2 (Lu et al., 2024)	9B	58.4	68.9	45.4
VARCO-VISION (Ju et al., 2024)	14B	**72.3**	90.9	**49.3**

Table 2: Evaluation results of various VLMs on the KoTextVQA benchmark. This table provides a detailed comparison of closed and open-source models, highlighting their capabilities in tackling Korean text-rich visual question answering tasks based on dual-level reasoning framework: System 1 and System 2.

3.3 Automated VQA Generation Pipeline

To generate comprehensive and high-quality QA pairs for Korean text-rich visual understanding and reasoning, we propose a four-step pipeline, as illustrated in Fig. 3. The pipeline begins with a stepwise image decomposition process, which includes image filtering and textual content extraction to ensure high-quality data input for subsequent steps.

Step 1: Stepwise Image Decomposition In this step, we refine the dataset by filtering out low-quality images. Images with a shortest side of 384 pixels or less are discarded to ensure text readability. To further ensure meaningful textual content, we use PaddleOCR[1] to exclude images with fewer than 10 or more than 1,000 Korean characters.

Following filtering process, multiple VLMs are employed to extract both textual and non-textual elements from each image with minimal hallucination and enhanced thoroughness. The decomposition process first analyzes non-textual visual attributessuch as the overall scene, document layout, key objects, and background detailsto establish contextual understanding. It then examines the structural and semantic relationships between text and visual components before finally extracting and processing all textual content into a structured format. This method preserves contextual relationships between visual and textual elements, leading to higher-quality outputs than direct text extraction.

Step 2: QA Candidates Generation and Classification Using the structured detailed captions from Step 1, this step simultaneously generates question-answer candidates via LLMs and classifies images into their corresponding domain and image type as defined in Section 3.1. The pipeline offers flexible control over the number of QA candidates generated and the number of models used in this process. QA generation follows the System 1 and System 2 framework, with prompts specifically designed to assess different levels of visual text understanding and reasoning. Independently, the classification step assigns each image to its appropriate domain and image type based on the structured captions from Step 1.

Step 3: QA Evaluation and Voting In this step, multiple VLMs evaluate the generated QA candidates to determine the highest-quality question-answer pair for each image. Drawing inspiration from prior LLM evaluation research (Zheng et al., 2023; Fu et al., 2024), the process employs a set of predefined criteria to systematically assess the quality of each candidate.

For System 1 candidates, we evaluate QA pairs using five metrics (Text Utilization, Clarity, Correctness, Naturalness, and Alignment) to ensure they capture textual content accurately and coherently. For System 2 candidates, two additional metrics (Complexity and Coherence) are introduced to account for multi-step reasoning and logical inference, as illustrated in Figure 3. Each VLM assigns a score from 0 to 5 for each metric, and the aggregated scores are used to rank the candidates. A voting mechanism then selects the highest-ranked QA pair, with multiple VLMs helping to minimize individual model biases.

Step 4: Hard Negatives Generation After selecting the final QA pair, an LLM generates three hard negative options that resemble the correct answer while remaining distinct in meaning. These options follow the correct answers structure and context, making the multiple-choice format more challenging. This step enhances the benchmarks ability to assess fine-grained comprehension and prevents models from relying on superficial cues.

[1] https://github.com/PaddlePaddle/PaddleOCR

Model	Size	Overall (2,577)	Gov. (245)	Econ. (104)	Mktg. (145)	Comm. (154)	Edu. (215)	Med. (90)	Tech. (92)	Arts. (83)	Transp. (167)	Tour. (108)	FnB. (264)	Ent. (168)	Life. (204)	*Sci.* (478)	*Hist.* (60)
Closed																	
GPT-4o (OpenAI, 2024)	-	84.6	93.5	92.3	97.2	90.3	**96.7**	91.1	**96.7**	**100.0**	84.4	93.5	**93.6**	**97.0**	95.1	**44.1**	**93.3**
GPT-4o-mini (OpenAI, 2024)	-	73.3	82.4	82.7	85.5	84.4	87.4	83.3	80.4	89.2	80.2	84.3	81.4	86.3	87.3	30.3	45.0
Gemini-2.0-flash (DeepMind, 2025)	-	**85.4**	**95.1**	**95.2**	**99.3**	**96.1**	**96.7**	**92.2**	93.5	98.8	**90.4**	**98.1**	93.2	95.2	**96.6**	**44.1**	78.3
Claude-3.5-Sonnet (Anthropic, 2024)	-	80.5	93.5	91.3	92.4	87.0	93.0	91.1	87.0	91.6	84.4	94.4	89.8	92.3	92.2	37.4	70.0
Open-source																	
LLaVA-OneVision (Li et al., 2024)	0.5B	42.3	51.8	48.1	47.6	44.8	39.5	50.0	44.6	40.9	49.7	51.9	41.7	44.6	46.1	28.0	31.7
Deepseek-VL2-tiny (Wu et al., 2024)	1B	48.8	57.1	55.8	63.4	58.4	51.2	57.8	57.6	45.8	54.5	58.3	43.9	47.0	54.4	30.5	31.7
Deepseek-VL2-small (Wu et al., 2024)	2.8B	53.3	61.6	63.5	66.9	63.0	57.2	64.4	68.5	50.6	59.9	63.0	48.9	56.0	57.4	30.8	36.7
Qwen2.5-VL (Wang et al., 2024)	3B	**71.8**	81.6	**76.9**	85.5	77.9	**87.4**	**80.0**	**79.3**	**85.5**	**75.4**	**84.3**	**76.9**	**87.5**	83.3	**33.9**	36.7
Ovis1.6-Llama3.2 (Lu et al., 2024)	3B	52.2	64.5	69.2	60.7	57.1	55.8	54.4	62.0	51.8	60.5	61.1	56.8	52.4	49.5	30.5	31.7
InternVL2.5 (Chen et al., 2024)	4B	70.7	**82.0**	**76.9**	**87.6**	**83.1**	83.7	78.9	**79.3**	79.5	75.4	77.8	69.3	81.0	**86.3**	**33.9**	**46.7**
Phi-3.5-Vision (Abdin et al., 2024)	4.2B	42.6	53.5	55.8	40.0	49.4	43.3	40.0	53.3	50.6	44.3	46.3	42.8	43.5	44.6	27.6	36.7
LLaVA-OneVision (Li et al., 2024)	7B	54.0	64.1	63.5	63.4	63.6	58.6	55.6	64.1	45.8	68.3	65.7	55.3	55.4	55.9	30.8	33.3
Qwen2.5-VL (Wang et al., 2024)	7B	68.5	80.0	77.9	**85.5**	81.2	**87.4**	76.7	75.0	**89.2**	77.8	82.4	**77.7**	**86.3**	**85.8**	15.1	36.7
InternVL2.5 (Chen et al., 2024)	8B	70.8	**81.6**	76.9	**85.5**	81.8	83.7	81.1	77.2	78.3	76.0	**83.3**	74.2	78.6	**85.8**	**34.1**	**38.3**
MiniCPM-V-2.6 (Yao et al., 2024)	8B	41.0	50.2	54.8	50.3	53.2	44.7	41.1	52.2	33.7	43.7	48.1	43.6	45.8	46.1	18.2	25.0
MiniCPM-o-2.6 (Yao et al., 2024)	8B	64.3	75.9	83.7	79.3	75.9	76.7	65.6	75.0	73.5	69.5	79.6	67.8	77.4	74.0	25.5	25.0
Ovis1.6-Gemma2 (Lu et al., 2024)	9B	58.4	64.1	69.2	71.0	72.7	60.9	71.1	67.4	53.0	68.9	75.9	65.2	58.9	63.2	30.5	28.3
VARCO-VISION (Ju et al., 2024)	14B	**72.3**	**81.6**	**87.5**	83.4	**83.1**	84.2	**86.7**	**84.8**	79.5	**82.6**	**83.3**	76.1	81.5	85.3	33.7	31.7

Table 3: Evaluation results of various VLMs across 15 domains in the KoTextVQA benchmark. This table compares closed and open-source models, highlighting their performance on Korean text-rich visual question answering tasks. Abbreviations: Gov. (Public & Administration), Econ. (Economics & Finance), Mktg. (Marketing & Advertising), Comm. (Retail & Commerce), Edu. (Education & Academia), Med. (Medical & Healthcare), Tech. (Science & Technology), Arts. (Arts & Humanities), Transp. (Transportation & Logistics), Tour. (Travel & Tourism), FnB. (Hospitality & Food Service), Ent. (Entertainment & Media), Life. (Personal & Lifestyle), Sci. (CSAT Science), Hist. (CSAT History).

3.4 Human Annotation Refinement

The final QA pairs undergo a thorough human review process based on the same evaluation criteria used in Step 3. Adjustments are made if a question is answerable solely from textual content without image context (Text Utilization); QA pairs are verified to ensure alignment with the original purpose of the image (Alignment); for System 2, it is confirmed that the question requires at least one inferential step to avoid overly simple responses (Complexity); and the language, grammar, and factual content are reviewed to ensure they are natural, unambiguous, and precise (Naturalness, Correctness, and Clarity). Additionally, to maintain dataset diversity, QA pairs for System 1 and System 2 are selected to cover a range of topics while removing overly repetitive or low-quality candidates.

4 Empirical Analysis

We leverage VLMEvalKit (Duan et al., 2024), an open-source evaluation toolkit designed to facilitate the assessment of VLMs, including both proprietary APIs and open-source models. For fair comparison, we use multiple-choice prompts from MMMU (Yue et al., 2024a), following the format used in the original evaluation of each model.

4.1 Performance across System 1 vs. System 2

Table 2 presents the performance breakdown between System 1 and System 2. Across both open-source and closed models, System 1 accuracy is significantly higher, indicating that most models handle text recognition and simple contextual understanding well. Notably, Gemini-2.0-flash (DeepMind, 2025) achieves 98.0% on System 1, reflecting near-perfect perception.

However, System 2 reveal substantial performance drops, particularly in open-source models. Qwen2.5-VL-7B (Wang et al., 2024) drops sharply from 94.5% in System 1 to 36.1% in System 2, and Deepseek-VL2-small (Wu et al., 2024) falls from 67.3% to 36.1%. In contrast, GPT-4o retain stronger performance, achieving 70.5% in System 2. This suggests that open-source models still struggle with complex reasoning, requiring further enhancements in external knowledge integration and multi-step reasoning capabilities.

4.2 Performance across Domain

Table 3 demonstrates the evaluation results, comparing the performance of closed and open-source models across various domains. Among closed models, Gemini-2.0-flash achieves the highest overall score (85.4%), followed by GPT-4o with 84.6%. Notably, GPT-4o demonstrates superior performance in the CSAT History domain, achieving 93.3%, suggesting a strong capability in leveraging historical and cultural context for reasoning. In contrast, Gemini-2.0-flash exhibits consistently high performance across multiple domains, reflecting robust text recognition and contextual comprehension in a real-world scenario.

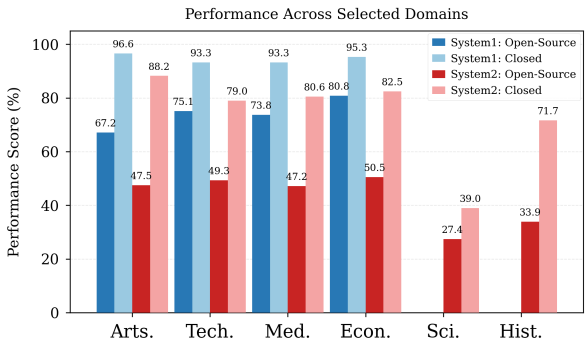

Figure 4: Comparison of open-source and closed models across different domains on KoTextVQA. Bars show the average scores of closed and open-source models separately for System 1 and System 2 in each domain.

Image Type	Closed		Open		Sys1 - Sys2		Closed - Open	
	Sys1	Sys2	Sys1	Sys2	Closed	Open	Sys1	Sys2
Document								
Chart and Plot	94.9	86.7	79.3	48.2	8.2	31.1	15.6	38.5
Table	91.0	75.0	70.9	42.3	**16.0**	28.6	20.1	32.7
Infographic	95.4	81.3	**80.0**	44.1	14.1	**35.9**	15.4	37.2
Slides	96.4	**95.0**	73.0	**61.3**	1.4	11.7	23.4	33.7
Book Cover	95.4	91.0	69.0	52.0	4.4	17.0	**26.4**	**39.0**
Product Detail	94.3	87.5	78.6	51.5	6.8	27.1	15.7	36.0
Poster	94.6	87.3	73.8	54.0	7.3	19.8	20.8	33.3
Mobile Screen	**97.2**	90.7	76.9	54.9	6.5	22.0	20.3	35.8
PC Screen	94.8	83.6	74.8	50.1	11.2	24.7	20.0	33.5
Scene Text								
Street Signs	87.0	**93.1**	75.9	**59.3**	-6.1	16.6	11.1	33.8
Public Signs	88.6	69.4	71.2	42.0	19.2	29.2	17.4	27.4
Store Sign	91.4	85.3	70.6	42.0	6.1	28.6	20.8	43.3
Banner	94.6	91.1	78.2	46.2	3.5	**32.0**	16.4	**44.9**
Signage	**94.7**	85.9	**78.5**	54.3	8.8	24.2	16.2	31.6
Menu	91.9	79.9	69.5	40.3	12.0	29.2	**22.4**	39.6
Manual	91.2	71.1	73.2	42.1	**20.1**	31.1	18.0	29.0

Table 4: Performance comparison across image types for closed and open-source models, showing differences across System 1, System 2, and model categories. Only image types with at least 50 VQA pairs are presented.

Open-source models exhibit a broad range of performance on KoTextVQA, with overall scores varying from 42.3% (LLaVA-OneVision (Li et al., 2024) at 0.5B) to 72.3% (VARCO-VISION (Ju et al., 2024)). Several models, such as Deepseek-VL2-tiny (Wu et al., 2024) (48.8%) and Phi-3.5-Vision (42.6%), fall below the 50% mark, while others like Qwen2.5-VL (3B) and InternVL2.5 (Chen et al., 2024) achieve scores in the low 70s. As such, open-source models exhibit significant variations in both overall and domain-specific performance, making it essential to carefully consider model size, architecture, and domain-specific performance when selecting an appropriate model for a given application.

Examining domain-specific results more closely, Figure 4 illustrates the System 1 and System 2 performance gap between closed and open-source models across different domains on KoTextVQA. The disparity is particularly pronounced in System 2 tasks, where closed models outperform open-source counterparts by up to 40.7 percentage points in *Arts & Humanities*, reflecting stronger cultural understanding. In contrast, the *Science & Technology* domain shows a smaller System 2 gap of 29.7 percentage points, suggesting more consistent handling of technical content. Standardized test settings further highlight this trend, with *CSAT Science* showing an 11.6 percentage point gap, while *CSAT History* reaches a 37.8 percentage point gap, underscoring the importance of background knowledge. These results indicate that when using open-source models for reasoning tasks in domains requiring extensive cultural and historical knowledge, such as *Arts & Humanities* and *History*, additional domain-specific training is necessary to bridge the performance gap with closed models.

A particularly striking case is Qwen2.5-VL (7B), which performs well in practical domains like Marketing (85.5%) yet suffers an extreme drop in CSAT Science (15.1%), even underperforming its 3B counterpart. To better understand this phenomenon, we further analyze the model's reasoning capabilities using Chain-of-Thought prompting in Section 5. More broadly, the overall results presented in Table 3 indicate that domain-specific performance metrics can guide model selection for practical applications and inform the curation of fine-tuning datasets to enhance model adaptability.

4.3 Performance across Image Type

Table 4 presents the performance of closed and open-source models across different image types, highlighting key trends in System 1 and System 2 tasks. Performance varies significantly by image type, reflecting distinct model capabilities. Document-based images, such as Tables and Infographics, achieve the high closed-source model accuracy in System 1 (91.0%, 95.4%) and retain relatively strong performance in System 2 (75.0%, 81.3%). However, open-source models struggle more with these formats, particularly in Tables and Infographics, where System 2 accuracy drops by 28.6.7% and 35.9%, respectively. Notably, Book Covers exhibit the largest performance gap between closed and open models (26.4% in System 1, 39.0% in System 2), likely due to their complex typography and mixed visual elements.

Scene text images present different challenges, with Street Signs showing an unusual trend where System 2 accuracy surpasses System 1 for closed models (93.1% vs. 87.0%). This may be due to motion blur or low resolution from vehicle-captured images, hindering recognition in System 1. In contrast, System 2 may rely on clearer images with better context, reducing the impact of noisy inputs. Open-source models perform particularly poorly in Banners and Store Signs, where the closed-open gap in System 2 reaches 44.9% and 43.3%, respectively, indicating difficulties in handling diverse fonts, occlusions, and unconventional layouts common in real-world signage. These findings highlight the varying complexity of image types, emphasizing the need for targeted improvements in both structured text processing and robust scene text understanding.

5 Discussion

While our benchmark effectively evaluates both basic and advanced reasoning in Korean text-rich VQA, several areas remain for further improvement. First, the current classification of System 2 tasks groups diverse complex features, such as multi-step inference and external knowledge integration, under a single category. A more fine-grained categorization could provide deeper insights into model limitations and inform targeted enhancements. Expanding the scope beyond question answering, for example, by incorporating document summarization or cross-referencing information across sections, would further assess higher-level comprehension and data processing abilities.

Chain-of-Thought (CoT) As mentioned before, we evaluate the impact of Chain of Thought (CoT) prompting on model performance, following the approach demonstrated in MMMU-Pro (Yue et al., 2024c). Figure 5 reveals a pronounced gap between closed and open-source models, not only in terms of baseline performance but also in their ability to follow instructions and maintain structured reasoning. Closed models consistently benefit from CoT strategies, showing notable improvements over default prompting while maintaining stable performance across English and Korean inputs. This suggests a greater capacity for structured reasoning within these models, allowing them to effectively utilize CoT.

In contrast, open-source models often struggle with reasoning-intensive tasks, leading to inconsistent or even degraded performance. Furthermore, their sensitivity to prompt language variations underscores fundamental limitations in structured reasoning and instruction following, often manifesting as boiled response format problems.

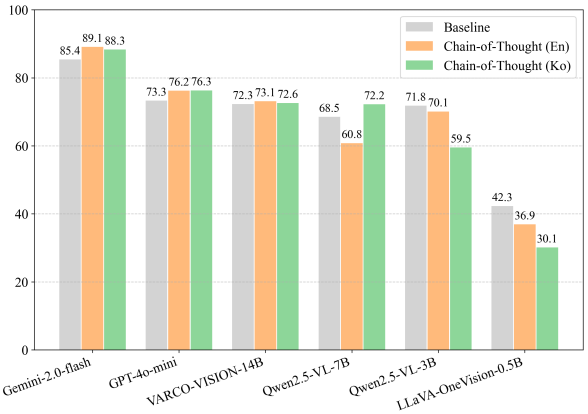

Figure 5: Comparison of two closed and four open-source models of varying sizes on KoTextVQA. The figure shows performance differences across three prompts: Baseline, Chain-of-Thought in English and Korean.

Within our benchmark, System 2 tasks pose complex multi-step reasoning challenges, offering a meaningful testbed for advanced reasoning. While high-reasoning VQA datasets remain scarce, we hope System 2 task contributes to deeper evaluations of testing-time compute in VLM research.

6 Conclusion

In this paper, we present KoTextVQA, a benchmark for evaluating VLMs on Korean text-rich images. By adopting a dual-level reasoning framework, System 1 and System 2, we enable structured analysis of visual text recognition and reasoning. Additionally, we devise a domain and image type classification scheme aligned with real-world contexts, providing meaningful evaluation criteria for practical applications. To support scalable dataset construction, we develop an four-step automated pipeline for generating high-quality VQA datasets. Experimental results show that while models perform well on System 1, System 2 remains challenging, particularly for open-source models lacking domain-specific or cultural knowledge and reasoning ability. These findings underscore the need for domain-specific fine-tuning and advanced reasoning. Beyond Korean, our methodology offers a scalable framework for evaluating VLMs in low-resource languages, fostering more linguistically and culturally adaptive vision-language models.

Limitations

While we provide a comprehensive evaluation of Korean Text-Rich VQA, several limitations highlight areas for future improvements. First, KoTextVQA focuses solely on single-image scenarios, restricting applicability in more complex settings involving multiple images or video-based contexts. Expanding coverage to these scenarios would enhance relevance for real-world applications.

Second, despite covering a diverse range of domains and image types, certain key areas remain underexplored. Mathematical reasoning and handwritten text, which are crucial for robust text understanding, have not been sufficiently addressed. Future iterations could incorporate these aspects to provide a more holistic evaluation.

Lastly, Chain-of-Thought (CoT) reasoning has been shown to improve performance on the System 2 benchmark, but additional strategies for further enhancement have not been explored. Investigating advanced reasoning techniques and optimization methods remains an open challenge for future research. We hope that KoTextVQA serves as a stepping stone for future advancements in this area, driving the development of more effective reasoning strategies and robust vision-language models.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*. Claude-3.5-sonnet-2024-10-22 version [Multimodal Large Language model].

Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration. *arXiv preprint arXiv:2406.16469*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

DeepMind. 2025. Gemini 2.0 flash main page. *https://deepmind.google/technologies/gemini/flash*. Gemini-2.0-flash-exp version [Multimodal Large Language model].

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. Qgeval: A benchmark for question generation evaluation. *arXiv preprint arXiv:2406.05707*.

Jeongho Ju, Daeyoung Kim, SunYoung Park, and Youngjune Kim. 2024. Varco-vision: Expanding frontiers in korean vision-language models. *arXiv preprint arXiv:2411.19103*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY, USA.

Jin-Hwa Kim, Soohyun Lim, Jaesun Park, and Hansu Cho. 2019. Korean localization of visual question answering for blind people. In *SK T-Brain-AI for Social Good Workshop at NeurIPS*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

OpenAI. 2024. Gpt-4o main page. *https://openai.com/index/hello-gpt-4o*. Gpt-4o-2024-11-20 version [Multimodal Large Language model].

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Korea Statistics. 2024. Ksic: The korean standard industrial classification, january, 2024. *URL https://classification.codes/classifications/industry/ksic*.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. 2024. Parrot: Multilingual visual instruction tuning. *arXiv preprint arXiv:2406.02539*.

Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, Wei Shi, Yuliang Liu, Hao Liu, Yuan Xie, Xiang Bai, and Can Huang. 2024a. Textsquare: Scaling up text-centric visual instruction tuning. *arXiv preprint arXiv:2404.12803*.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024b. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Emanuele Vivoli, Ali Furkan Biten, Andres Mafla, Dimosthenis Karatzas, and Lluis Gomez. 2022. Must-vqa: Multilingual scene-text vqa. *https://arxiv.org/abs/2209.06730*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

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*.

Yuanhan Zhang Bo Li Songyang Zhang Wangbo Zhao Yike Yuan Jiaqi Wang Conghui He Ziwei Liu Kai Chen Dahua Lin Yuan Liu, Haodong Duan. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024b. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2024c. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Ruiyi Zhang, Yufan Zhou, Jian Chen, Jiuxiang Gu, Changyou Chen, and Tong Sun. 2024a. Llava-read: Enhancing reading ability of multimodal language models. *arXiv preprint arXiv:2407.19185*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2024b. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.