Benchmarking Visual LLMs Resilience to Unanswerable Questions on Visually Rich Documents

Benchmarking Visual LLMs Resilience to Unanswerable Questions on Visually Rich Documents

Davide Napolitano, Luca Cagliero, Fabrizio Battiloro

Politecnico di Torino, Torino, Italy
name.surname@polito.it

Abstract

The evolution of Visual Large Language Models (VLLMs) has revolutionized the automatic understanding of Visually Rich Documents (VRDs), which contain both textual and visual elements. Although VLLMs excel in Visual Question Answering (VQA) on multi-page VRDs, their ability to detect unanswerable questions is still an open research question. Our research delves into the robustness of the VLLMs to plausible yet unanswerable questions, i.e., questions that appear valid but cannot be answered due to subtle corruptions caused by swaps between related concepts or plausible question formulations. Corruptions are generated by replacing the original natural language entities with other ones of the same type, belonging to different document elements, and in different layout positions or pages of the related document. To this end, we present VRD-UQA (**V**ISUALLY **R**ICH **D**OCUMENT **U**NANSWERABLE **Q**UESTION **A**NSWERING), a benchmark for evaluating VLLMs' resilience to plausible yet unanswerable questions across multiple dimensions. It automatically alters the questions of existing VQA datasets consisting of multi-page VRDs, verifies their unanswerability using a VLLM-as-a-judge approach, and then thoroughly evaluates VLLMs' performance. Experiments, run on 12 models, analyze: (1) The VLLMs' accuracy in detecting unanswerable questions at both page and document levels; (2) The effect of different types of corruption (NLP entity, document element, layout); (3) The effectiveness of different knowledge injection strategies based on in-context learning (OCR, multi-page selection, or the possibility of unanswerability). Our findings reveal VLLMs' limitations and demonstrate that VRD-UQA can serve as an evaluation framework for developing resilient document VQA systems.

Code — https://github.com/DavideNapolitano/VRD-UQA

1 Introduction

Visual Large Language Models (VLLMs) are trained and specialized to produce accurate answers, in textual form, to questions about a mix of visual and textual content (Luo et al. 2024). These models are particularly valuable for analyzing Visually Rich Documents (VRDs) (Wang et al. 2023b), i.e., documents that combine textual content (paragraphs, titles) with structured visual elements (e.g., figures, tables). They encompass various document types, such as PDF files and printed or scanned copies, and cover a variety of domains and sources (e.g., news, financial reports).

Visual Question Answering (VQA) from Visually Rich Documents (VRDs) is particularly challenging because it requires not only an advanced comprehension of the question but also the ability to link the linguistic concepts mentioned in the question to contents, either textual and visual, available in pages with complex layout structures or even in different pages. In this work, we focus on the zero-shot VQA capabilities of VLLMs on multi-page VRDs, which is one the most representative real-world scenarios.

Even though a question on a multi-page VRD may seem plausible and well-formed, its answer can be undetermined. For example, given the question *What is the future projection of sea level in the figure?*, document pages may not contain figures regarding sea level or the information could be embedded in different elements, like a table (Davis 2020).

The ability of VQA models to determine whether a question is answerable or not is at least as important as providing correct and pertinent answers (Guo et al. 2024; Vardi, Nir, and Shamir 2025). In our research, we aim to mimic human questions that are unanswerable due to small errors caused by swaps between related concepts or due to the inherent formulations. They are known to be quite common (Xie et al. 2024) and not as trivial to detect as small typos or meaningless sentences because they pass grammar and semantic checks (Jia and Liang 2017a). Notice that NLP entity swaps may also involve document elements (e.g., *Caption* instead of *Footnote*) or layout information (e.g., *Bottom* instead of *Top*), making their detection even worse.

We verify the VLLMs' robustness in correctly detecting these unanswerable cases both separately for each page and at the document level. To this end, we alter the answerable questions of the multi-page VQA datasets (Tito, Karatzas, and Valveny 2023; Landeghem et al. 2023) with a controlled level of corruption. Specifically, we recognize NLP entities in the original question and replace them with others of the same type, within different multimodal elements, and in different layout positions or pages of the related document.

The purpose of the present work is to address the following Research Questions (RQs):

- RQ1: Are VLLMs capable of accurately detecting question unanswerability due to the entity corruption?

- RQ2: What is the effect of different corruption types on models' performance?
- RQ3: Which in-context learning strategies are able to mitigate the limitations of VLLMs in identifying unanswerable questions?

To address the RQs, we propose VRD-UQA (**V**ISUALLY **R**ICH **D**OCUMENT **U**NANSWERABLE **Q**UESTION **A**NSWERING), a new framework aimed to evaluate VLLMs performance in detecting unanswerable questions. Given a multi-page document VQA dataset and a set of models, VRD-UQA automatically corrupts the questions, verifies their actual unanswerability using a VLLM-as-a-judge approach (Li et al. 2024; Zheng et al. 2023), and then evaluates the document- and page-level accuracies of distinct models by analyzing the separate and combined effects of different corruption types. The experiments carried out on 12 models and 2 datasets showcase:

- The models' performance, underlying the importance of the model pretraining strategy which is paramount even compared to the number of model parameters (RQ1);
- The models' strengths and weaknesses with specific NLP entities, (e.g., fairly robust to perturbations of location entities, weak on document structure-related entities), the variable resilience in handling document elements (e.g., higher resilience with headers and footnotes, lower with tables), and the difficulty to circumvent corruptions in long documents and caused by in-page entities (RQ2);
- The benefits of adopting in-context learning strategies, such as providing the document OCR or stating the possibility of unanswerability, to mitigate the limitations of state-of-the-art models in tackling the unanswerability detection problem (RQ3).

The main paper contributions can be summarized as:

- An open source **new evaluation framework** (VRD-UQA) focused on evaluating VLLMs' robustness to unanswerable questions on multi-page VRDs.
- We present a **pipeline** aimed to alter answerable questions available in VQA benchmark datasets with controlled levels of corruptions regarding **NLP entities**, **document elements**, and **document layout**.
- We **release** the extended and corrupted versions of the established DUDE (Landeghem et al. 2023) and MP-DocVQA (Tito, Karatzas, and Valveny 2023) datasets.
- An extensive empirical **evaluation** carried out on 12 VLLMs, and 2 VQA datasets for multi-page VRDs.

The rest of the paper is organized as follows. Section 2 reports related work. Sections 3 and 4 introduce preliminary notions and corruption strategies. Section 5 describes the VRD-UQA benchmark. Section 6 discusses the results. Section 7 draws conclusions and discusses future works.

2 Related Work

Recently, the research community has released several VQA benchmarks for VRDs (Mathew, Karatzas, and Jawahar 2021; Landeghem et al. 2023; Tito, Karatzas, and Valveny 2023; Mathew et al. 2021; Choi et al. 2018; Deng et al. 2025). A comprehensive taxonomy can be found in (Rogers, Gardner, and Augenstein 2023). Parallel works have focused on assessing the VQA models' capability to detect unanswerable questions using corrupted images and questions (Guo et al. 2024; Whitehead et al. 2022; Zhang, Ho, and Vasconcelos 2023; Akter et al. 2024). Specifically, Reliable-VQA and UNK-VQA (Guo et al. 2024) are designed to handle single images or text without contextual knowledge, whereas our approach (VRD-UQA) is capable of processing documents including multiple images. While VRD-UQA dynamically corrupts the input questions through a mix of NLP and multimodal learning techniques, UNK-VQA applies predefined perturbations, RGQA (Zhang, Ho, and Vasconcelos 2023) applies self-supervised contrastive learning to generate image-question pairs, whereas VisReas (Akter et al. 2024) generates unanswerable queries using Visual Genome data.

Alternative approaches, such as MMLongBench-Doc (Ma et al. 2024), TUBench (He et al. 2024) and LongDocURL (Deng et al. 2025), evaluate VQA models' robustness through natively unanswerable questions. In contrast, our approach generates unanswerable questions by corrupting answerable ones. Furthermore, we explore multiple dimensions (e.g., document elements and layout), either separately or jointly, while preserving question plausibility.

Other studies have examined the frequency and kind of typing errors (Cucerzan and Brill 2004), showing that entity substitution errors occur through mechanisms like autocorrect interference, phonetic similarity, and memory lapses (Shi et al. 2025). Human transcription errors (Hong et al. 2013; Mays and Mathias 2019) are alternative sources of corruption, which potentially preserve coherence and plausibility of the unanswerable question. Previous studies on the robustness of QA models (Belinkov and Bisk 2017; Ribeiro, Singh, and Guestrin 2018) have shown that these models are highly sensitive to corrupted inputs, with minor substitutions causing significant performance degradation in document understanding (Jia and Liang 2017b). This calls for new VQA testing benchmarks aimed to evaluate VLLMs performance with corrupted questions on VRDs.

3 Preliminaries

A VRD D consists of one or multiple pages $p_1, p_2, \ldots, p_{|D|}$. It includes not only textual elements, such as paragraphs and headlines, but also visual elements (e.g., charts and tables). Given a natural language question Q on a D, VQA from VRDs exploits a model to generate the answer A to Q based on the D's content. In this work, we focus on questions Q with no answer, i.e., the *unanswerable questions*. We ask VLLMs to detect these cases and return *No answer* as corresponding response. To evaluate the models' capabilities to accurately identify unanswerable questions, we define the following experimental setting. Firstly, we leverage the VQA model with (1) a specific instruction prompt, where additional information such as OCR and unanswerability information may be included, (2) an unanswerable question and (3) a window sliding over the document pages (the window size w is a user-specified parameter). Then, we verify

the correctness of the provided answer (correct: *No answer*, incorrect: *otherwise*). Next, we repeat the test over different page windows. Finally, we evaluate the model's performance according to the following performance metrics: (1) *Document-Level Accuracy* (Acc_D), i.e., the percentage of unanswerable questions for which *all* the associated document page-level answers are correct. (2) *Page-Level Accuracy* (Acc_P), i.e., the average rate of correct page-level answers for each corrupted question.

4 Question Corruption

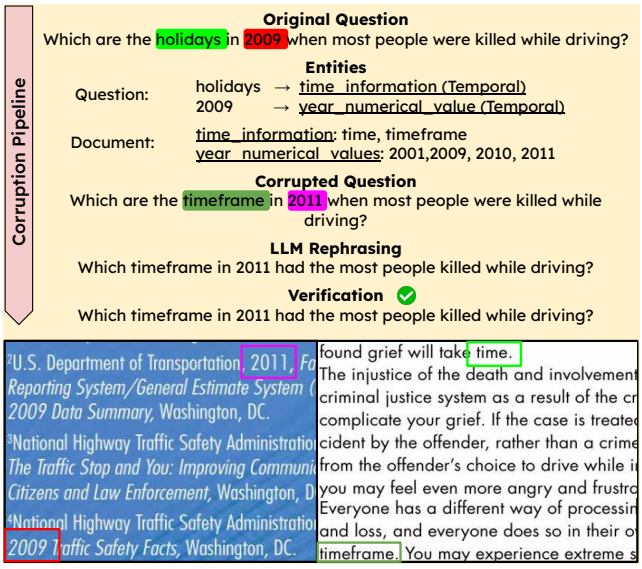


Figure 1: VRD-UQA generates unanswerable questions starting from an answerable question and the reference document. Example from DUDE (Landeghem et al. 2023)).

Questions on VRD may contain errors due to typos, misunderstandings, and memory lapses, or they may be intrinsically unanswerable. In this work, we focus on unanswerable questions built by swapping related concepts, identified by NLP entities. These questions mimic user queries that appear plausible and contextually relevant but cannot be answered based on the VRD content. This allows us not only to evaluate the robustness to the presence of semantically similar but incorrect entity matchings between questions and document contents but also to evaluate the effect of incorrect references to document elements or layout information.

We design a systematic approach to generate plausible yet unanswerable questions. Specifically, we inject a controlled level of corruption into the (answerable) questions of a VQA document dataset. To this end, we consider (separately or mixed) three main corruption types: (1) *NLP entities*, (2) *Document elements*, (3) *Document layout*. Our framework is motivated by the unique challenges of VRDs, which require models to jointly reason over textual semantics, structural composition (i.e., functional elements), and spatial layout to enable effective question answering (Vardi, Nir, and Shamir 2025; Wang et al. 2023a). To evaluate such capabilities, we propose a multi-level corruption framework that systematically analyzes multiple facets of documents and models.

NLP Entities In Natural Language Processing, named entities are well-known concepts that are typically described by one or more words in the document text (Manning, Raghavan, and Schütze 2008). Based on their semantic meaning, entities are usually categorized into predefined types (e.g., numbers, temporal information). For example, in Figure 1, we highlight in red and green the identified temporal entities. A possible typo in writing consists of replacing an entity with another of the same type, such as reporting *2011* instead of *2009* (see Figure 1). Similar human errors in entity value specification are common in document information retrieval (Shi et al. 2025). In most cases, these subtle textual modifications would preserve the plausibility of the question, thus making the detection of unanswerable cases particularly challenging. Hence, we evaluate models' resilience to entity-level corruptions by replacing an entity occurring in the original question with another one of the same type occurring in the document (regardless of its relative position). This approach simulates the most challenging settings for models, as all relevant elements in the question are contained within the document.

Document Elements Elements in VRDs encompass both textual items (e.g., paragraphs, captions) and visual ones (tables, figures). When the entities in question are corrupted, unanswerability becomes particularly difficult to grasp if the substitutes are placed in different document elements. To simulate element-wise corruptions, we replace an NLP entity in the original question with any one appearing in the different elements present in the document. For example, to corrupt the entities in the question, we pick 2 out of 11 temporal entities belonging to *Abandon* elements (i.e., headers, footers, footnotes, and marginal notes) from the infographic appearing in the left-hand side document page in Figure 1.

Document Layouts Question entities can appear in multiple layout positions and pages. While evaluating the unanswerability of a question with a corrupted entity on a given document page, the presence of similar entities within that page, eventually in different positions (i.e., the in-page corruption), can be challenging because the model may struggle to detect the error as layout information becomes diriment. We simulate layout-related errors by replacing an entity in the original question with both in- and out-page entities. For example, in Figure 1, the question is corrupted with the green and red entities belonging to different pages.

5 The Evaluation Benchmark

This section describes the VRD-UQA framework, which generates unanswerable questions to test VQA models. Its architecture is depicted in Figure 2. Given a VQA dataset consisting of VRDs (see Section 3), it performs:

1. *Augumentation*, which extracts auxiliary information from VRDs, like OCR or visual element captions, necessary for the next steps;
2. *Corruption*, which corrupts the questions as described in Section 4, both separately for each entity type and mixed;

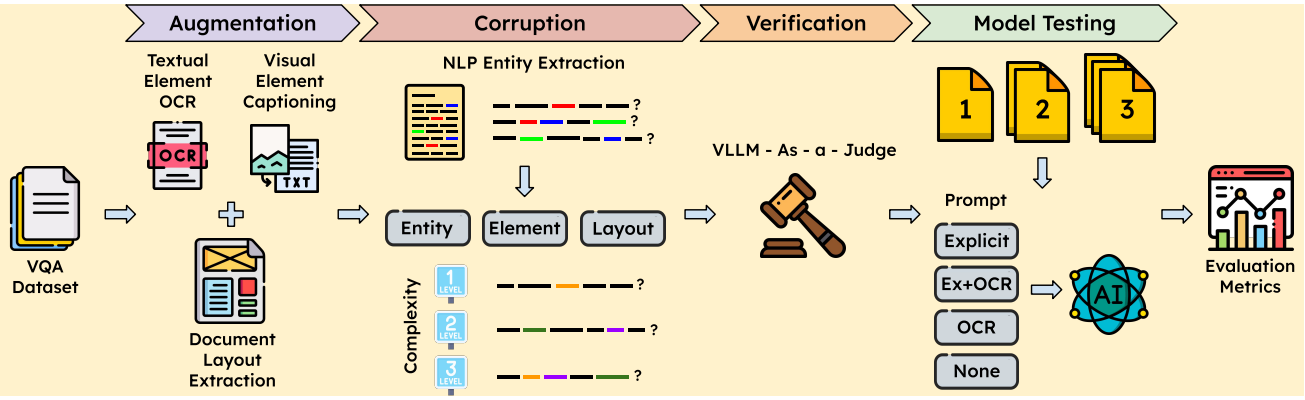


Figure 2: The VISUALLY RICH DOCUMENT UNANSWERABLE QUESTION ANSWERING framework.

3. *Verification*, verifies the actual questions' unanswerability by employing a VLLM-as-a-judge approach;
4. *Evaluation*, which tests the corrupted questions on several VQA models, with and without enriching the prompt with the auxiliary information and collecting the document- and page-level accuracies.

Each pipeline step, thoroughly described in the following sections, is designed to be modular and highly customizable. Additional information, including prompts, on each of the following steps is reported in the Supplementary Material.

Augmentation

This step focuses on enriching the input VQA examples with the following auxiliary information:

1. D*ocument* L*ayout* A*nalysis* (DLA): it extracts document elements (including metadata) within each VRD page;
2. E*lement* C*aptions* (EC): it generates textual captions of visual elements (e.g., plots, diagrams, figures) using the state-of-the-art Qwen 2.5 VL model (Wang et al. 2024). The purpose is to allow the automatic extraction of entities from multimodal document elements;
3. O*ptical* C*haracter* R*ecognition* (OCR): it generates transcriptions of the textual elements present in the document image, such as paragraphs or titles, using the cost-effective GOT-OCR 2 model (Wei et al. 2024).

Our structured pipeline enables a comprehensive textual representation of document pages by systematically processing visual elements, such as tables and figures. This approach enhances document comprehension compared to existing methods, which extract information from visual elements without preserving their structural relationships.

Corruption

This stage transforms each answerable question Q in the original VQA dataset into an unanswerable one $\hat{Q}$ by applying the corruption types described in Section 4.

To detect NLP entities, we leverage the GliNER model (Zaratiana et al. 2024) on a set of user-defined entities. We define the following macro categories: Numerical, Temporal, Miscellaneous, Location, and Structural. To test VLLMs' resilience to varying perturbation levels, we generate corrupted questions where a single corruption is applied (C=1) or two/three corruptions (C=2/3) are present in the generated question. We decided not to exceed C=3 to avoid excessively severe corruptions (uncommon for humans).

To ensure grammatical coherence and semantic plausibility while maintaining specific corrupted elements, we leverage Qwen 2.5 (Yang et al. 2024). Our methodology provides the LLM with a structured framework comprising the original question for contextual reference, the corrupted version requiring refinement, a comprehensive list of corruption items that must be preserved, explicit directives emphasizing readability and linguistic naturalness, and curated examples of adequate and suboptimal refinements.

Verification

To prevent VRD-UQA from collecting results from not truly unanswerable questions, we employ a VLLM-as-a-judge approach (Li et al. 2024; Zheng et al. 2023), i.e., we inquire the established Gemini 2.5 Flash VLLM (Team et al. 2023) to double-check whether each new question is actually unanswerable on each page of the analyzed VRDs and is unlikely to contain hallucinations. To limit evaluation circularity, we deliberately use a model different from those tested during the main experiments. We selected Gemini 2.5 Flash due to its strong trade-off between performance on multimodal document understanding tasks and efficiency. The verification process utilizes a structured prompt that incorporates several critical components to ensure accurate assessment. In detail, the prompt includes a detailed task description, comprehensive OCR output from the document page, and explicit entity mapping that shows the relationship between original and corrupted entities. Questions marked as unanswerable by Gemini ($\sim$30%) are manually reviewed by human experts to evaluate the quality of the model's judgment. On this step, we find an average precision of 96.97%, indicating strong alignment between Gemini's predictions and human assessment. Notably, we observe that the discarded questions are predominantly associated with lower complexity levels, suggesting that simpler corruptions are more prone to accidental answerability.

Evaluation

Given the set of models under evaluation, we prompt each of them with the corrupted and verified questions and collect their outcomes. We test the pre-trained model versions under a zero-shot setting. To enrich the benchmarking phase, beyond the question complexity, we also consider for VLLMs the following parameters: (1) *Page window size* (w), which indicates the number of consecutive pages that are processed (1 to 3), reflecting the multi-page nature of documents; (2) *OCR-inclusion*, which indicates whether the text transcription is included or not (Wei et al. 2024); (3) *Explicit*, which indicates whether the prompt indicates the possibility of unanswerability of the given question or not.

6 Experiments

We run experiments on a machine equipped with NVIDIA A6000 GPUs, 192GB of RAM, and an AMD 7950X CPU. The computational budget was around 90 hours to perform experiment with a single execution. The statistics about the original and sampled datasets, the models settings and additional results are reported in the Supplementary Material[1].

Original Datasets

We analyze two open-source benchmark datasets for multi-page VQA from VRDs: (1) *MPDocVQA* (Tito, Karatzas, and Valveny 2023), which collects 5,131 documents of varying length along with 36,230 question-answer pairs in its train set; (2) *DUDE* (Landeghem et al. 2023), which consists of 5,017 documents and 40,000+ questions. To limit the computational and human efforts, hereafter we will consider a representative sample of 300 questions.

Augmented, Corrupted, and Verified Data

For both datasets' samples we process 424 documents containing 595 questions and generate 2176 potentially unanswerable questions as well as the necessary auxiliary information (more details in the Supplementary Material). Then, we perform verification, identifying 593 genuinely unanswerable questions with variable complexity (318 level-1 questions, 201 level-2 questions, and 74 level-3 questions). In both datasets, we achieve a significant variety in the number of pages per document. The elements of types Abandon and Plain Text are predominant, whereas figures, tables, and titles are relatively rare but with non-negligible peaks. Similar to prior works (Tito, Karatzas, and Valveny 2023; Landeghem et al. 2023), we neglect other elements, such as formulas, as they are statistically irrelevant.

Models Used in the Framework

Due to the nature of the datasets, which primarily consist of scanned documents, including handwritten documents, we consider the DocLayout-YOLO model a reference after an empirical evaluation of over 100 documents for each dataset. We leverage an additional phase to extract textual representations upon identifying document elements. In particular, we employ the lightweight state-of-the-art GOT-OCR 2 (Wei et al. 2024) for OCR on textual elements, while Qwen 2.5 VL 7B (Wang et al. 2024) for visually rich items captioning. These operations are performed at the image patch level to provide reliable results. We employ GliNER Large V2 (Zaratiana et al. 2024) for NER on both VRD elements and questions in order to perform the corruption. Additionally, to post-process corrupted questions, we leverage Qwen 2.5 (Yang et al. 2024). For the verification phase, we employ a state-of-the-art Gemini 2.5 Flash (Team et al. 2023).

Evaluated Models

We test a variety of VLLMs with different sizes, pretraining procedures, and optimize their parameter settings. In detail, we analyze Phi 4 Multimodal (Abdin et al. 2024), Qwen 2.5 VL 7B and 72B (Wang et al. 2024), Molmo 7B (Deitke et al. 2024), InternVL 3 9B and 78B (Zhu et al. 2025), Ovis1.6 9B (Lu et al. 2024), LLama 3.2 11B (Grattafiori et al. 2024), Gemma3 27B (Team et al. 2025), Llava1.6 34B (Liu et al. 2024), GPT-4.1-mini and O3 (Achiam et al. 2023).

Results Discussion

RQ1: Are VLLMs capable of accurately detecting question unanswerability due to the entity corruption? To answer RQ1, we analyze the Acc_D and Acc_P achieved by the tested models (see Table 1). Acc_D performance is consistently lower than Acc_P. This trend is particularly evident in long documents, where the likelihood of misclassifying an unanswerable question is higher (see Figure 3). Due to their highly specialized pretraining, Qwen and Gemma models demonstrate superior performance metrics. Our analysis indicates that model size is not the most discriminant factor influencing the performance, suggesting that architectural features and training strategies are paramount and often yield more substantial gains than the VLLM scale. Our findings show that a comprehensive understanding of the document is crucial to effectively address VQA on VRDs. This is confirmed by the results achieved with complexities 1 and 2, as all models roughly perform similarly. Converseley, at Complexity 3 the overall performance degrades because the corruption severely alters questions' meaning.

RQ2: What is the effect of different corruption types on models' performance? In Figure 3 we analyze the effect of the corrupted entity type (i.e., numeric, temporal, miscellaneous, location, structure), the number/percentage of different visually reach document elements, and the document length on the document-level performance. Furthermore, we also compare the page-level performance across the NLP entity types. For the other two analysis we group the outcomes by the page-level presence of document elements and by the presence of corrupted entities on a page for the layout.

The results show the models' resilience to location and numerical entity corruptions. Oppositely, their performance drop while dealing with structural entity modifications, particularly when structural layout-related information is manipulated (e.g., replacing the word "Figure" with "Table"). The composition of document elements also significantly affects models' performance. As the ratio of visual elements

[1]https://arxiv.org/pdf/2511.11468

		Phi4	Molmo	Ovis	Llama	Llava 34B	Gemma3 27B	Qwen2.5 VL 7B	Qwen2.5 VL 72B	InterVL3 9B	InterVL3 78B	GPT4.1 mini	O3
DUDE													
Acc_D		0.070	0.230	0.241	0.289	0.401	0.503	0.460	**0.599**	0.267	0.326	0.214	0.239
Acc_P		0.248	0.554	0.674	0.680	0.717	0.786	**0.835**	0.754	0.713	0.781	0.638	0.663
Acc_D	C1	0.079	0.254	0.281	0.342	0.377	0.482	0.465	**0.588**	0.281	0.342	0.202	0.227
	C2	0.052	0.190	0.172	0.224	0.483	0.586	0.517	**0.707**	0.259	0.328	0.276	0.301
	C3	0.067	0.200	0.200	0.133	0.267	**0.333**	0.200	0.267	0.200	0.200	0.067	0.092
Acc_P	C1	0.266	0.577	0.723	0.712	0.701	0.810	**0.843**	0.753	0.738	0.805	0.636	0.661
	C2	0.240	0.542	0.615	0.655	0.760	0.764	**0.847**	0.816	0.684	0.760	0.669	0.694
	C3	0.141	0.423	0.513	0.526	0.692	0.679	**0.731**	0.519	0.628	0.667	0.538	0.563
MPDocVQA													
Acc_D		0.037	0.340	0.217	0.325	0.357	0.394	0.490	**0.581**	0.241	0.219	0.264	0.163
Acc_P		0.211	0.780	0.792	0.796	0.708	0.838	**0.881**	0.842	0.782	0.818	0.775	0.738
Acc_D	C1	0.044	0.358	0.221	0.314	0.309	0.402	0.500	**0.613**	0.255	0.275	0.294	0.18
	C2	0.028	0.329	0.189	0.322	0.441	0.420	0.497	**0.538**	0.259	0.175	0.259	0.16
	C3	0.034	0.305	0.271	0.373	0.322	0.305	0.441	**0.576**	0.153	0.136	0.169	0.06
Acc_P	C1	0.225	0.830	0.815	0.845	0.707	0.852	**0.901**	0.855	0.829	0.849	0.827	0.780
	C2	0.188	0.699	0.740	0.724	0.729	0.824	**0.850**	0.791	0.725	0.757	0.691	0.669
	C3	0.205	0.749	0.808	0.749	0.663	0.808	0.865	**0.885**	0.712	0.824	0.741	0.712

Table 1: Document- and Page- Accuracy for each dataset and complexity. Best models are in bold, the second are underlined.

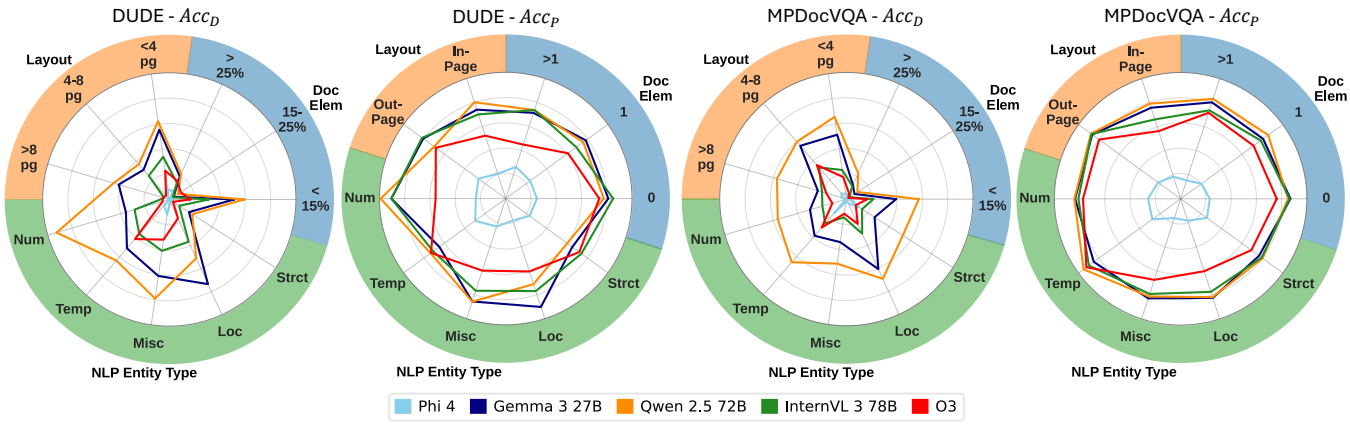


Figure 3: Impact of corruption type on VQA performance across datasets and sizes of the models (representative subset).

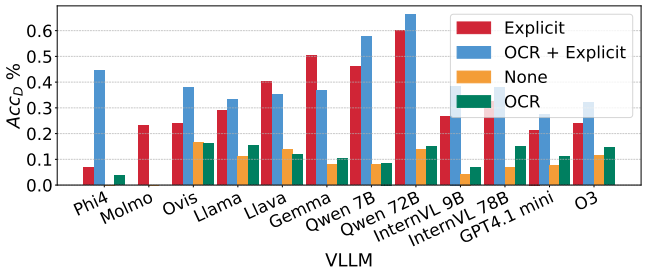
to the total number of elements increases, we observe a consistent decrease in both accuracy metrics. As expected, all the tested models generally perform better on unanswerable questions related to text-only pages. The document length emerges as a critical factor, with longer documents proving to be more challenging to process. Corruptions involving in-page entities turn out to be the most challenging. This difficulty stems from the proximity of misleading information, which creates false contexts that models struggle to discern.

Table 2 deepens the analysis of the Acc_P performance for specific types of element- and layout-wise corruptions. Our findings indicate that Abandon elements have a negligible impact on unanswerability detection capabilities, as models consistently demonstrate higher proficiency in distinguishing this secondary information from core content. Conversely, models exhibit poor performance on title elements on the MPDocVQA dataset, which directly correlates with the diminished accuracy in the top-left quadrant (50.63%, with the remainder distributed between the top-right and bottom-left regions). The performance gap between Abandon and Title elements indicates that the specificity of the document type significantly influences models' behavior. Similarly, we detect a correlation between the presence of tabular elements and the fairly low performance of left-hand side page quarters, motivated by the apperance of almost half of the tables in the top-left quadrant.

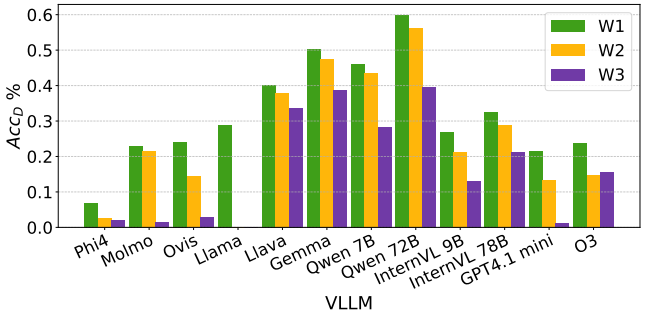
RQ3: Which in-context learning strategies are able to mitigate the limitations of VLLMs in identifying unanswerable questions? We compare different in-context

		Phi4	Molmo	Ovis	Llama	Llava 34B	Gemma3 27B	Qwen2.5 VL 7B	Qwen2.5 VL 72B	InterVL3 9B	InterVL3 78B	GPT4.1 mini	O3
DUDE													
Doc Elem	Title	0.143	0.357	0.643	0.500	0.571	0.714	0.786	**0.846**	0.429	0.786	0.357	0.407
	Text	0.152	0.411	0.601	0.506	0.715	0.766	0.766	**0.835**	0.576	0.728	0.500	0.486
	Figure	0.328	0.406	0.438	0.531	0.625	0.781	0.750	**0.831**	0.656	0.625	0.344	0.409
	Table	0.150	0.483	0.400	0.483	0.567	0.617	0.683	**0.732**	0.417	0.617	0.533	0.686
	Abandon	0.233	0.567	**0.767**	0.667	0.533	0.733	0.733	0.675	0.700	0.733	0.633	0.610
Layout	Top Left	0.072	0.387	0.361	0.428	0.680	0.768	0.655	**0.892**	0.392	0.649	0.670	0.720
	Top Right	0.131	0.377	0.492	0.492	0.705	0.869	0.770	**0.896**	0.525	0.754	0.426	0.412
	Bottom Left	0.131	0.381	0.452	0.548	0.685	**0.750**	0.667	0.742	0.500	0.708	0.435	0.499
	Bottom Right	0.321	0.543	0.514	0.500	0.693	0.664	0.707	**0.752**	0.629	0.643	0.471	0.624
MPDocVQA													
Doc Elem	Title	0.056	0.444	0.389	0.389	0.667	0.583	0.528	**0.682**	0.333	0.361	0.361	0.306
	Text	0.208	0.625	0.709	0.646	0.721	0.800	0.823	**0.828**	0.654	0.723	0.613	0.609
	Figure	0.176	0.576	0.353	0.647	0.694	0.624	**0.800**	0.658	0.600	0.612	0.447	0.388
	Table	0.060	0.641	0.530	0.581	0.607	0.675	**0.735**	0.719	0.350	0.325	0.487	0.462
	Abandon	0.433	0.767	0.700	0.800	0.767	0.833	0.800	0.833	0.667	**0.867**	0.733	0.600
Layout	Top Left	0.105	0.499	0.506	0.491	**0.686**	0.663	0.676	0.670	0.410	0.388	0.469	0.416
	Top Right	0.297	0.669	0.699	0.720	0.665	0.803	**0.845**	0.834	0.628	0.628	0.632	0.573
	Bottom Left	0.246	0.652	0.696	0.672	0.701	0.736	0.812	**0.849**	0.661	0.690	0.559	0.614
	Bottom Right	0.231	0.790	0.706	0.803	0.714	0.824	**0.899**	0.844	0.744	0.832	0.689	0.664

Table 2: Effect of the corruption type on the Page-Level Accuracy. Best results are in bold, the second are underlined.



(a) Effect of augmentation on Acc_D



(b) Effect of window size on Acc_D

Figure 4: Effect of augmented information and window size on Acc_D performance. DUDE dataset.

learning strategies for mitigating VLLM limitations in identifying unanswerable questions (see Figure 4). Regarding the prompt setting, explicitly stating the possibility of unanswerability significantly improves VLLMs' performance. As expected, models demonstrate significantly higher accuracy when explicitly instructed that questions may not be answerable based on the provided context. Additionally, including OCR-extracted text generally improves performance across all experimental settings, suggesting that textual information provides valuable context for determining unanswerability. The combination of two pieces of information yields the best overall performance, indicating a synergistic effect between semantic task understanding and comprehensive information access. We also analyze the effect of varying window sizes, with accuracies consistently decreasing as the window size increases. The tested models struggle to handle larger contexts as they may introduce noise or excessively spread the information. Similar results on MPDocVQA are given in the Supplementary Material.

7 Conclusions and Future Work

The paper presents an evaluation framework for comprehensively analyzing the VLLM's capabilities to detect unanswerable questions. The framework evaluates the robustness of models in realistic scenarios, where questions are built by entity swaps. To fully explore the challenges of the VRDs, we leverage corruptions within different multimodal elements, layout positions, and pages. The results provide insights into models' performance, highlighting gaps between models with different pretraining and size and their resilience to various corruption settings. The main limitations of this work are (1) the zero-shot settings only; (2) the use of general-purpose in-context learning strategies. To address these issues, as future work, we plan to fine-tune VLLMs and to design more advanced mitigation strategies.

Acknowledgments

This study was carried out within the FAIR (Future Artificial Intelligence Research) and received funding from NextGenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1555.11-10-2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akter, S. N.; Lee, S.; Chang, Y.; Bisk, Y.; and Nyberg, E. 2024. VISREAS: Complex Visual Reasoning with Unanswerable Questions. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 6735–6752. Bangkok, Thailand: Association for Computational Linguistics.

Belinkov, Y.; and Bisk, Y. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; tau Yih, W.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. QuAC : Question Answering in Context. arXiv:1808.07036.

Cucerzan, S.; and Brill, E. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 293–300.

Davis, E. 2020. Unanswerable questions about images and texts. *Frontiers in Artificial Intelligence*, 3: 51.

Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Deng, C.; Yuan, J.; Bu, P.; Wang, P.; Li, Z.-Z.; Xu, J.; Li, X.-H.; Gao, Y.; Song, J.; Zheng, B.; and Liu, C.-L. 2025. LongDocURL: a Comprehensive Multimodal Long Document Benchmark Integrating Understanding, Reasoning, and Locating. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1135–1159. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, Y.; Jiao, F.; Shen, Z.; Nie, L.; and Kankanhalli, M. S. 2024. UNK-VQA: A Dataset and a Probe Into the Abstention Ability of Multi-Modal Large Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12): 10284–10296.

He, X.; Zhang, Q.; Jin, A.; Yuan, Y.; Yiu, S.-M.; et al. 2024. TUBench: Benchmarking large vision-language models on trustworthiness with unanswerable questions. *arXiv preprint arXiv:2410.04107*.

Hong, M. K.; Yao, H. H.; Pedersen, J. S.; Peters, J. S.; Costello, A. J.; Murphy, D. G.; Hovens, C. M.; and Corcoran, N. M. 2013. Error rates in a clinical data repository: lessons from the transition to electronic data transfer—a descriptive study. *BMJ open*, 3(5): e002406.

Jia, R.; and Liang, P. 2017a. Adversarial Examples for Evaluating Reading Comprehension Systems. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031. Copenhagen, Denmark: Association for Computational Linguistics.

Jia, R.; and Liang, P. 2017b. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Landeghem, J. V.; Powalski, R.; Tito, R.; Jurkiewicz, D.; Blaschko, M. B.; Borchmann, L.; Coustaty, M.; Moens, S.; Pietruszka, M.; Anckaert, B.; Stanislawek, T.; Józiak, P.; and Valveny, E. 2023. Document Understanding Dataset and Evaluation (DUDE). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 19471–19483. IEEE.

Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; and Liu, Y. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural Embedding Alignment for Multimodal Large Language Model. *arXiv preprint arXiv:2405.20797*.

Luo, C.; Shen, Y.; Zhu, Z.; Zheng, Q.; Yu, Z.; and Yao, C. 2024. LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15630–15640.

Ma, Y.; Zang, Y.; Chen, L.; Chen, M.; Jiao, Y.; Li, X.; Lu, X.; Liu, Z.; Ma, Y.; Dong, X.; et al. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37: 95963–96010.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to information retrieval*. Cambridge University Press. ISBN 978-0-521-86571-5.

Mathew, M.; Bagal, V.; Tito, R. P.; Karatzas, D.; Valveny, E.; and Jawahar, C. V. 2021. InfographicVQA. arXiv:2104.12756.

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

Mays, J. A.; and Mathias, P. C. 2019. Measuring the rate of manual transcription error in outpatient point-of-care testing. *Journal of the American Medical Informatics Association*, 26(3): 269–272.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, 856–865.

Rogers, A.; Gardner, M.; and Augenstein, I. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10): 1–45.

Shi, D.; Zhu, Y.; Fernandes Junior, F. E.; Zhai, S.; and Oulasvirta, A. 2025. Simulating Errors in Touchscreen Typing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–13.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Tito, R.; Karatzas, D.; and Valveny, E. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144: 109834.

Vardi, B.; Nir, O.; and Shamir, A. 2025. CLIP-UP: CLIP-Based Unanswerable Problem Detection for Visual Question Answering. *arXiv preprint arXiv:2501.01371*.

Wang, D.; Raman, N.; Sibue, M.; Ma, Z.; Babkin, P.; Kaur, S.; Pei, Y.; Nourbakhsh, A.; and Liu, X. 2023a. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z.; Zhou, Y.; Wei, W.; Lee, C.; and Tata, S. 2023b. VRDU: A Benchmark for Visually-rich Document Understanding. In Singh, A. K.; Sun, Y.; Akoglu, L.; Gunopulos, D.; Yan, X.; Kumar, R.; Ozcan, F.; and Ye, J., eds., *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, 5184–5193. ACM.

Wei, H.; Liu, C.; Chen, J.; Wang, J.; Kong, L.; Xu, Y.; Ge, Z.; Zhao, L.; Sun, J.; Peng, Y.; et al. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model.

Whitehead, S.; Petryk, S.; Shakib, V.; Gonzalez, J.; Darrell, T.; Rohrbach, A.; and Rohrbach, M. 2022. Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, 148–166. Springer.

Xie, G.; Zhang, K.; Duan, L.; Zhang, W.; and Huang, Z. 2024. Typos Correction Training against Misspellings from Text-to-Text Transformers. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16907–16918. Torino, Italia: ELRA and ICCL.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Zaratiana, U.; Tomeh, N.; Holat, P.; and Charnois, T. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5364–5376. Mexico City, Mexico: Association for Computational Linguistics.

Zhang, Y.; Ho, C.; and Vasconcelos, N. 2023. Toward Unsupervised Realistic Visual Question Answering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 15567–15578. IEEE.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.