Worldwide LiveVQA: Real-Time Visual Knowledge Seeking and Updating Across Languages

Xuanao Huang[1*], Xingjia Liu[1*], Yuyang Peng[1], Zetong Zhou[1], Yao Wan[1‡], Dongping Chen[2‡]

[1]Huazhong University of Science and Technology, [2]University of Maryland,

Correspondence: wanyao@hust.edu.cn, dongping@umd.edu

Figure 1: LiveVQA-W: a new dataset for worldwide visual knowledge seeking and updating, comprising 10 different languages, 171,105 visual entities, 234,286 images, and 872,896 questions. Each instance contains a representative image, mostly serving as visual knowledge from Aug. 2025 to Dec. 2025.

Abstract

Knowledge about the visual world is not only constantly evolving but also inherently happening all over the world: breaking news in Tokyo, political events in São Paulo, and cultural phenomena in Cairo are first reported in Japanese, Portuguese, and Arabic, carrying regional context that English-centric resources cannot fully capture. Yet existing resources for visual knowledge remain confined

Findings of the Association for Computational Linguistics: ACL 2026, pages 39819–39894
July 2-7, 2026

to English, creating a *"Worldwide Knowledge Gap"* that hinders developing truly global assistants. To quantify this gap, we introduce LIVEVQA-WORLDWIDE (LIVEVQA-W), the first dynamic-updating dataset for real-time, multilingual visual knowledge seeking and updating across ten major languages. Drawing from worldwide news outlets, YouTube videos, and academic platforms during August–December 2025, LIVEVQA-W comprises 234K images, 873K questions, and 171K visual entities with hierarchical evaluation: Level 1 for visual entity recognition and Level 2 for multi-hop cross-lingual reasoning. Our comprehensive benchmarking of 15 state-of-the-art MLLMs reveals that models without search achieve near-random performance, while search-augmented models exhibit severe linguistic bias, with English accuracy nearly double that of other languages. Furthermore, we explore visual knowledge updating through large-scale training, finding that injected knowledge improves recall but remains fragile under prompt rephrasing and image perturbations such as rotation and flipping. We release the fully replicable data collection pipeline and raw dataset to support continuous community-driven expansion. The benchmark, code, and related resources are available at: worldwide-livevqa.github.io.

1 Introduction

In the era of information globalization, the aspiration of *"knowing the world without leaving home"* is rapidly materializing. Multimodal Large Language Models (MLLMs) are emerging as universal portals that aggregate and deliver knowledge from every corner of the globe, promising to serve as omniscient assistants capable of understanding and reasoning about visual information across cultures, languages, and temporal contexts. However, this vision remains fundamentally constrained by a critical asymmetry: while real-world knowledge is produced and disseminated across diverse linguistic communities, the up-to-date resources that shape these models remain English-centric.

The production and dissemination of knowledge exhibit a distinct trend toward decentralization and linguistic fragmentation. Breaking news, social events, cultural phenomena, and technological breakthroughs are often first reported by local media in their native languages, carrying deep regional context that cannot be fully captured through translation alone. Prior research has demonstrated that culture-specific knowledge is best learned and evaluated in its corresponding language (Manvi et al., 2024; Huang et al., 2024c; AlKhamissi et al., 2024). However, current research for visual factuality seeking and updating remains fixed and confined to English contexts with often unverified answers, making it challenging to measure global knowledge acquisition (Fu et al., 2025; Narayan et al., 2025). This creates a severe *"Worldwide Knowledge Gap"*: real-time developments in non-English regions are systematically excluded from mainstream knowledge bases and evaluation frameworks, while MLLMs trained primarily on English data struggle to objectively perceive and understand factual evolutions that fall outside their English-centric pre-training corpora.

To bridge these gaps, we propose LIVEVQA-WORLDWIDE (LIVEVQA-W), a dynamic-updating dataset comprising 234,286 images, 872,896 questions, and 171,105 unique visual entities with full multilingual context for global, real-time visual knowledge seeking and updating. Our automated framework features an AI Agent-driven automated pipeline that continuously crawls multimodal content from worldwide news, videos, and academic platforms across ten major languages. This pipeline is capable of autonomously identifying and integrating new information sources, enabling dynamic expansion of multilingual channels without human intervention.

Based on LIVEVQA-W, we conduct large-scale evaluations of 15 state-of-the-art MLLMs on their basic visual factuality seeking capability, revealing a systemic language bias: even the most capable search-augmented models perform substantially better in English than in other languages, highlighting their suffering from the Internet bias and overreliance on high-resource linguistic data, a key limitation that our benchmark aims to address.

To further demonstrate the utility of LIVEVQA-W beyond evaluation, we also investigate its effectiveness as a training resource for large-scale visual knowledge updating. Specifically, we validate that the SFT stage is more suitable for visual entity knowledge injection than the pretraining phase. We further conduct experiments on several representative MLLMs and analyze how exposure to globally diverse, fact-grounded vision-language data influences models' general capabilities. Moreover, our experiments also demonstrate that post-trained visual knowledge is fragile, which can be disturbed through rephrasing and image disturbance like ro-

Name	Images	Questions	Entities	Language	Dynamic?	Replicable?[‡]	Web search?	Multi-hop?	For update?
Benchmarks									
SimpleVQA (Cheng et al., 2025)	2,025	2,025	–		✘	✘	✘	✘	✘
LiveVQA-Benchmark (Fu et al., 2025)	1,500	3,000	–		✔	✘	✔	✔	✘
Dyn-VQA (Li et al., 2024)	1,452	1,452	–		✔	✘	✔	✔	✔
BrowseComp-VL (Geng et al., 2025)	399	399	–		✘[†]	✘	✔	✔	✘
MM-BrowseComp (Li et al., 2025)	224	224	–		✘[†]	✘	✔	✔	✘
MMSearch (Jiang et al., 2024)	171	300	–		✘	✘	✔	✔	✘
MMSearch-Plus (Tao et al., 2025)	441	279	–		✘	✘	✔	✔	✘
Datasets									
KVQA (Shah et al., 2019)	24k	183k	18.8k		✘	✘	✘	✔	✘ (outdated)
A-OKVQA (Schwenk et al., 2022)	23,692	24,903	–		✘	✘	✘	✘	✘ (outdated)
InfoSeek (Chen et al., 2023)	1.35M	1.35M	11,481		✘	✘	✘	✘	✘ (outdated)
E-VQA (Yang et al., 2023)	2,690	9,088	182		✔	✘	✘	✔	✘ (outdated)
Encyclopedic VQA (Mensink et al., 2023)	1M	221k	–		✘	✘	✘	✘	✘ (outdated)
EDIS (Liu et al., 2023)	1.04M	32,493	–		✘	✘	✘	✘	✘
FVQA (Wu et al., 2025)	6,800	6,800	–		✘	✘	✔	✔	✘
DeepMMSearchVQA (Narayan et al., 2025)	10,000	10,000	–		✘	✘	✔	✔	✘
LiveVQA-Dataset (Fu et al., 2025)	26,988	104,143	–		✔	✘	✘	✔	✔
LiveVQA-W(orldwide)[*]	234k	873k	171k		✔	✔	✔	✔	✔

[†] BrowseComp-VL/MM-BrowseComp require web browsing but are not inherently time-sensitive.
[‡] Fully open-sourced and replicable data collection, preprocessing and curation pipeline.
[*] Our dataset features an automated scalable pipeline that can easily incorporate new languages and data sources.

Table 1: Comparison of visual info-seeking benchmarks and visual knowledge update datasets. Columns indicate whether tasks are time-sensitive (*Dynamic?*), have a fully replicable data pipeline (*Replicable?*), require online search to answer (*Web search?*), involve multi-hop reasoning, and whether they can be used to study *multimodal knowledge update*.

tation and noising.

We will release the fully replicable data collection pipeline and whole dataset to support continuous community-driven expansion to worldwide visual knowledge seeking and updating.

2 Related Works

Visual knowledge. Visual (world) knowledge links images to facts, concepts, and relations about the real world, combining commonsense scene understanding with externally verifiable facts (Marino et al., 2019; Schwenk et al., 2022). In this work, we focus on ***visual factual knowledge***—entity-centric and fact-seeking queries—where models must identify entities and retrieve or verify facts beyond the pixels by grounding to KBs and the web (Chen et al., 2022; Zhao et al., 2023; Caffagni et al., 2024; Caron et al., 2024; Jiang et al., 2024). Early efforts curated static resources (celebrities, landmarks, bio/encyclopedic attributes) to tie visual entities to factual signals (tonyassi, 2024; Weyand et al., 2020; Yang et al., 2024; Mensink et al., 2023). More recently, work shifts from static to synchronizing multimodal assistant with *up-to-date* visual knowledge: answering entity/fact questions against evolving news, events, and cultural phenomena via retrieval-augmented or browsing-based pipelines (Jiang et al., 2024; Fu et al., 2025; Nayak et al., 2024; Huang et al., 2024a). This presents the nature that as we automate aspects of our tasks, we would want our assistants to remain as up-to-date as we are. (Geng et al., 2025; Li et al., 2025).

Worldwide Visual Knowledge Seeking and Updating. Motivated by building native multilingual assistants, prior work argues that culture-specific knowledge should be learned in the corresponding language (Manvi et al., 2024; Huang et al., 2024c; AlKhamissi et al., 2024). A broad ecosystem of multilingual vision-language resources now spans captioning, VQA, and instruction tuning from web-scale image-text corpora to multilingual VQA benchmarks and instruction suites (Liu et al., 2021; Romero et al., 2024; Li et al., 2023; Sun et al., 2024). **For InfoSeek**, *i.e.*, retrieving and updating factual visual knowledge via browsing/grounding, recent evaluations in Chinese and Hungarian reveal steep cross-lingual performance drops, highlighting the fragility of multilingual info-seeking and web-grounded training setups (He et al., 2024; Tan et al., 2024; Gu et al., 2025; Yang et al., 2025b). *In this work*, we take the first step to *visual knowledge seeking and updating* at *worldwide* scale to test whether similar multilingual bias persists, and we analyze how a domestic-lingual (in-language) setting and large-scale knowledge updates affect visual entity/fact seeking.

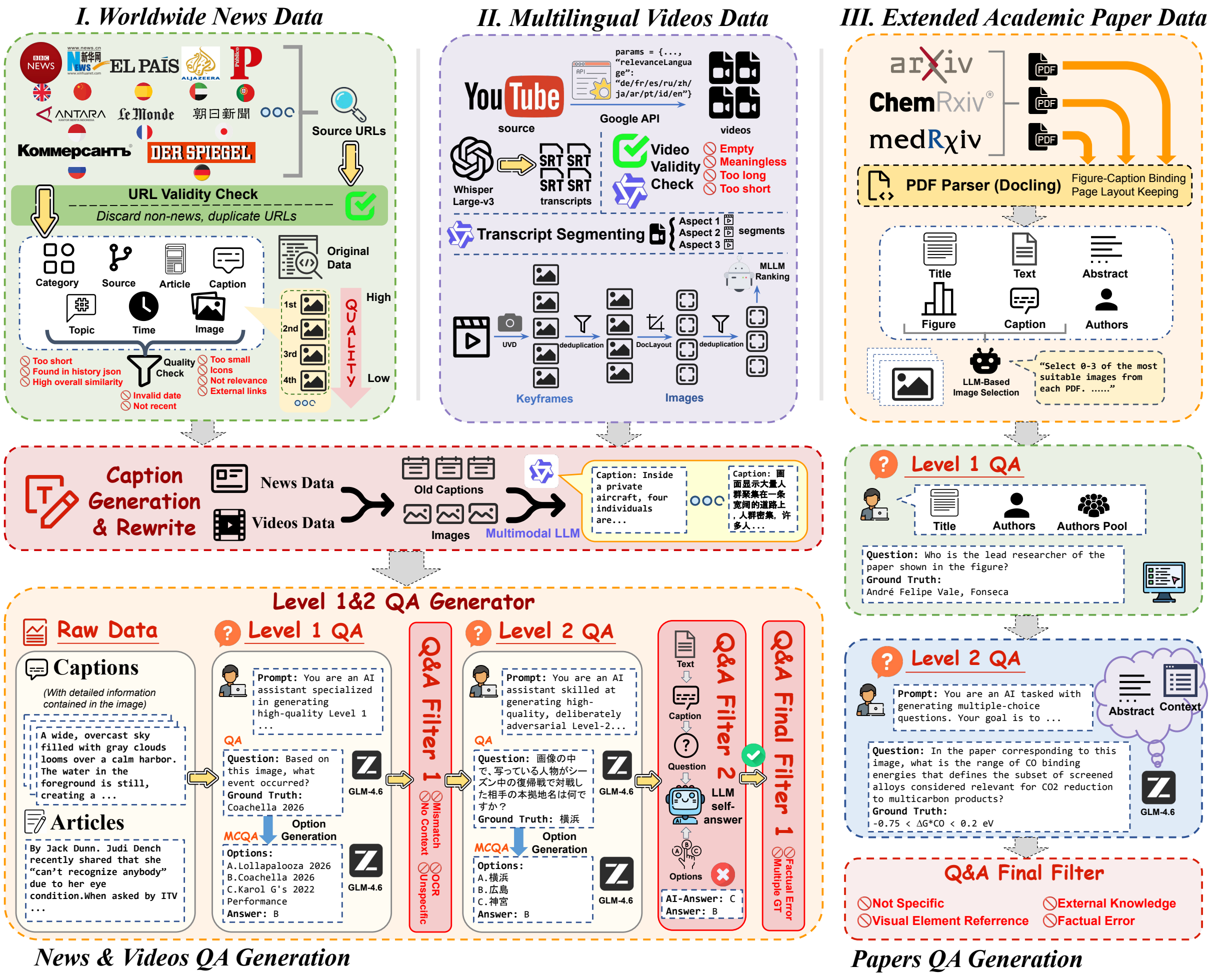


Figure 2: Overview of LIVEVQA-TOOLBOX, serving as an automated pipeline to collect fresh and real-time visual entity knowledge from the Internet. It collects worldwide visual data from multiple domains (*i.e.*, News articles, YouTube videos, and Papers), conducts multi-level data filtering, and generates info-seeking (Level 1) and multi-hop (Level 2) Q&A pairs.

3 LIVEVQA-TOOLBOX: Facilitating Visual Knowledge Collection at Scale

We develop LiveVQA-Toolbox, a comprehensive toolkit that continuously harvests fresh visual content from diverse online sources to facilitate automated visual knowledge collection. As shown in Figure 2, the toolbox automates the entire data acquisition workflow: from crawling prominent news outlets across different countries, to processing complex image-text interleaved documents, to extracting the most representative visual content from news articles, video platforms, and academic.

A key innovation is our agent-driven crawler generation system, which significantly reduces manual effort when integrating new data sources. As illustrated in Figure 3, when adding a new source, the system first analyzes the page structure and models content elements (title, date, text, images, captions). A coding agent then generates modular extraction scripts with appropriate configurations, which are deployed into a sandboxed environment for trial-run validation. The system automatically verifies successful access, article scraping, timestamp retrieval, and image downloading. If errors occur, the agent iteratively debugs based on error messages until all checks pass, after which the crawler is seamlessly integrated with regular monitoring.

For news articles, multi-level filtering ensures quality through URL screening, image selection by relevance and size, and LLM-based semantic filtering. For video content, the pipeline segments videos based on subtitles, extracts keyframes using perceptual hashing and layout detection to remove UI elements, then employs an MLLM to select top-K frames by topical relevance. For academic papers, the system crawls arXiv across domains, prioritizing distinctive diagrams over common vi-

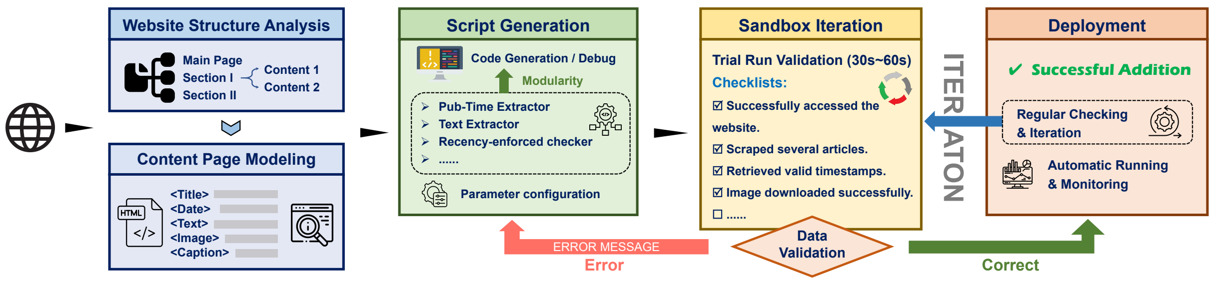


Figure 3: Automated adaptation for integrating new source via coding agent workflow, featuring modular script generation, sandbox testing with feedback loops, and live deployment.

sualizations. After collecting raw data from all sources, the pipeline unifies the processing to generate two-level QA pairs: Level-1 questions focus on visual entity recognition (*e.g.*, people, locations, events), filtered to retain only substantive answers; Level-2 questions require multi-hop cross-modal reasoning, validated by checking whether an LLM can correctly answer given the full context. Both levels include chain-of-thought reasoning trajectories and are available in open-ended and multiple-choice formats. See Appendix B for more details. Our collection process implements a multi-stage processing via a fully *open-sourced* pipeline and undergoes strict human validation with a greater than 90% pass rate, as detailed in Appendix B.3.

Positioning vs. LiveVQA. We extend LiveVQA automated pipeline with *(i)* multilingual adaptations for worldwide sources collection and preprocessing, *(ii)* fully open-sourced components (Whisper-large-v3 for ASR; Qwen3-VL for caption; GLM-4.6 for question curation) to reduce API cost while preserving replicability, *(iii)* an agentic onboarding workflow that easily scales news and videos scraping to new sources across languages, and *(iv)* using MLLMs for captioning and more powerful LLMs for pairing figures with captions, filtering, and question generation to enable high-quality visual entity knowledge collection.

4 LiveVQA-W(orldwide)

Leveraging LiveVQA-Toolbox, we present LiveVQA-W, a first-of-its-kind multilingual visual entity knowledge dataset containing 171,105 unique visual entities, 234,286 images, and 872,896 corresponding questions with full multilingual context. LiveVQA-W expands LiveVQA with 10 frequently used languages, designed for bridging frontier multimodal assistants from English-centric visual knowledge to worldwide visual knowledge seeking and updating. LiveVQA-W is dynamically collected from worldwide News, Video, and Academia platforms, featuring recent worldwide visual content spanning August 2025 to early December 2025 and continuously collecting to be released annually.

Category	Images	#Question	Level 1	Level 2
		Dataset		
News	172,981	657,936	162,720	495,216
Videos	45,612	167,881	41,149	126,732
Papers	15,693	47,079	15,693	31,386
Avg. per Picture	1	3.73	0.94	2.78
Overall	234,286	872,896	219,562	653,334
Test Split	500	500	250	250
Training Split	9,605	35,266	8,949	26,317
		Benchmark		
News	196	200	100	100
Videos	195	200	100	100
Papers	95	100	50	50

Table 2: Overall statistics of LiveVQA-W.

LiveVQA-W is structured as: *(1)* A visually distinctive image depicting a specific visual entity with democratic linguistic content. *(2)* Basic visual factuality seeking question focusing on a specific visual entity (*e.g.* celebrity, landmark, event, time). *(3)* Harder questions requiring web browsing and multi-hop cross-modality reasoning capability curated from the corresponding context. *(4)* Both types of questions are available in multiple-choice questions for pretrained model assessments and verified reward training (Chen et al., 2025; Wu et al., 2025; Ren et al., 2025), as well as open-ended format for info-seeking capability assessments (Wei et al., 2024, 2025; Haas et al., 2025).

For evaluation, we construct a benchmark of 500 instances sampled from LiveVQA-W, including 200 multilingual News instances, 200 multilingual Videos instances, and 100 English-only Papers instances. For the multilingual News and Videos portions, we ensure balanced coverage across lan-

Language	Benchmark		Dataset Shard	
	Flu.	Rel. (%)	Flu.	Rel. (%)
Arabic	4.55	100.0	4.39	100.0
Chinese	4.31	97.5	4.34	92.5
English	4.55	100.0	4.48	97.5
French	4.30	90.0	4.41	92.5
German	4.00	100.0	4.05	90.0
Indonesian	4.28	100.0	4.39	97.5
Japanese	4.25	97.5	4.38	100.0
Portuguese	4.41	90.0	4.60	90.0
Russian	3.96	92.5	4.02	95.0
Spanish	4.41	100.0	4.38	95.0

Table 3: Statistics of fluency (1-5) and culture relevance in benchmark and 500 sampled dataset shard.

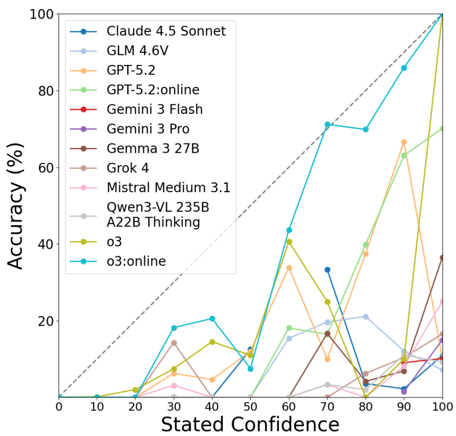


Figure 4: Distribution of model accuracy across different confidence intervals.

guages and difficulty levels, such that each language and each level contains the same number of samples. We use this benchmark to evaluate different models on worldwide visual knowledge seeking, and report the corresponding results in the following sections.

To further assess the quality of the generated VQA pairs, we evaluate all 400 instances in the benchmark (News and Videos) and additionally sample 400 instances from the dataset, using GPT-5.2 as an automatic evaluator. Specifically, we measure *(1)* fluency, rated on a 1–5 scale, and *(2)* whether each question is grounded in locally relevant language or cultural context. As shown in Table 3, the generated VQA pairs achieve consistently high fluency scores and strong cultural relevance across languages, further validating the overall quality of our multilingual data pipeline.

5 Can MLLMs Seek for Worldwide Visual Knowledge?

Keeping pace with the latest visual knowledge across cultures and languages is crucial for developing global assistants that are deeply integrated into human life and capable of solving users' problems across the world. Previous research has demonstrated that visual assistants show promise in handling English content. Here, we investigate *how well current MLLMs acquire visual knowledge across diverse linguistic cultures* and evaluate their effectiveness in leveraging external tools to access this knowledge. We also measure calibration (Wei et al., 2024), *i.e.*, whether models *"know what they know"* about visual content.

5.1 Experiment Setups

Models. Our evaluation encompasses 15 state-of-the-art MLLMs, including *Offline Models* GPT-5.2 (OpenAI, 2025a), GPT-o3 (OpenAI, 2025b), Qwen3-VL (Bai et al., 2025), Grok 4 (xAI, 2025), GLM-4.6V (Team, 2025), Mistral Medium 3.1 (Mistral AI, 2025), Gemma3-27B (Gemma Team, Google DeepMind, 2025), Gemini-3-Pro/Flash (Google DeepMind, 2025), Claude-Sonnet-4.5 (Anthropic, 2025), and *Search-Augmented Models* MMSearch-R1 (Wu et al., 2025), WebWatcher (Geng et al., 2025). The latter refers to systems equipped with text and image search capabilities.

Evaluation. In this section, we mainly evaluate the model's visual factuality recall and info-seeking capability. Therefore, we use free-form QA style and use prompts following LiveVQA (Fu et al., 2025) to collect model responses and confidence scores to evaluate their performance and calibration. We also adapt grading metrics by GPT-5-mini, following SimpleQA (Haas et al., 2025) to measure *Correct*, *Not Attempted*, and *Incorrect Responses*, along with the final F-score (detailed in Appendix C). All results are reported as averages over three independent evaluations.

5.2 Results and Analysis

As shown in Table 11, among models without web access, **Gemini-3** family achieves the highest accuracy on both Level 1 and Level 2 tasks, significantly outperforming other non-retrieval baselines. However, even the best offline model remains far behind search-augmented systems, highlighting the limitations of purely parametric knowledge in com-

Model	Correct	Not attempted	Incorrect	Correct & given attempted	F-score
		***w.o.* Search**			
Claude Sonnet 4.5	3.4	**82.2**	14.4	19.1	6.0
Gemini 3 Flash Preview	16.8	3.2	**80.0**	17.4	17.4
Gemini 3 Pro Preview	19.8	2.8	<u>77.4</u>	20.4	20.6
Gemma 3 27B	5.4	32.8	61.8	8.0	6.3
Mistral Medium 3.1	6.8	57.0	36.2	15.8	9.3
GPT-5.2	5.2	73.2	21.6	19.4	8.6
GPT-o3	9.6	50.6	39.8	19.4	13.0
Qwen3 VL 235B A22B Thinking	6.0	51.0	43.0	12.2	8.0
Grok 4	8.8	40.4	50.8	14.8	10.9
GLM 4.6V	4.2	<u>82.0</u>	13.8	23.3	6.7
		***w.* Text & Image Search**			
GPT-5.2:online	**35.0**	34.6	30.4	**53.5**	**42.1**
GPT-o3:online	<u>30.8</u>	34.4	34.8	<u>47.0</u>	<u>37.4</u>
MMSearch-R1	13.0	20.6	66.4	16.4	14.8
WebWatcher-7B	25.8	18.0	56.2	31.5	28.4
WebWatcher-32B	26.6	12.2	61.2	30.3	28.1

Table 4: Detailed breakdown of non-search and search models' performance.

Model	Avg.	Arabic	Chinese	English	French	German	Indonesian	Japanese	Portuguese	Russian	Spanish
					***w.o.* Search**						
Claude Sonnet 4.5	3.75	2.50	2.50	5.00	5.00	2.50	2.50	2.50	5.00	7.50	2.50
Gemini 3 Flash Preview	17.50	**15.00**	17.50	15.00	15.00	20.00	**15.00**	**20.00**	**22.50**	22.50	12.50
Gemini 3 Pro Preview	**19.50**	7.50	**25.00**	**17.50**	**20.00**	**25.00**	**15.00**	**20.00**	**22.50**	**25.00**	**17.50**
Gemma 3 27B	4.00	5.00	0.00	2.50	2.50	5.00	0.00	2.50	7.50	7.50	7.50
Mistral Medium 3.1	6.25	10.00	10.00	10.00	5.00	2.50	5.00	0.00	7.50	5.00	7.50
GPT-5.2	4.75	5.00	10.00	0.00	5.00	5.00	2.50	5.00	2.50	5.00	7.50
GPT-o3	9.00	10.00	7.50	7.50	17.50	5.00	7.50	7.50	7.50	7.50	12.50
Qwen3 VL 235B A22B Thinking	4.75	5.00	7.50	2.50	7.50	0.00	0.00	0.00	5.00	12.50	7.50
Grok 4	8.75	5.00	10.00	10.00	7.50	15.00	5.00	10.00	10.00	10.00	5.00
GLM 4.6V	3.75	2.50	2.50	5.00	5.00	2.50	0.00	5.00	5.00	7.50	2.50
					***w.* Text & Image Search**						
GPT-5.2:online	**26.25**	25.00	**32.50**	40.00	22.50	20.00	22.50	15.00	22.50	27.50	**35.00**
GPT-o3:online	25.75	**30.00**	25.00	35.00	20.00	22.50	17.50	**22.50**	**25.00**	**30.00**	30.00
MMSearch-R1	10.50	15.00	7.50	22.50	0.00	17.50	2.50	2.50	20.00	7.50	10.00
WebWatcher-7B	23.00	20.00	22.50	47.50	**27.50**	20.00	22.50	12.50	20.00	12.50	25.00
WebWatcher-32B	24.75	17.50	15.00	**55.00**	25.00	**30.00**	**25.00**	17.50	**25.00**	10.00	27.50

Table 5: Accuracy (%) of different models on the News and Videos parts of the benchmark, evaluated on open-ended questions in 10 languages. See Appendix D for more results in different axes.

plex visual factuality tasks. Models equipped with search capabilities demonstrate a dramatic performance gain. For instance, **GPT-o3:online** attains an accuracy of **39.2%** on Level 2 tasks, nearly four times higher than the best offline model. This underscores the critical role of external retrieval in handling real-world, time-sensitive queries.

> ***Finding 1:*** Search-augmented models outperform offline models across languages.

Cross-lingual evaluation reveals a stark disparity: offline models maintain relatively balanced performance, whereas search-enabled models perform significantly better in **English** and **Spanish** than in other languages. For example, GPT-5.2:online scores 40.0% in English but falls below 25% in low-resource languages. This indicates that current multimodal retrieval pipelines are heavily optimized for English-centric web content.

To pinpoint the exact bottleneck, we conducted an error attribution analysis on 301 failed queries from WebWatcher-32B across the News and Videos domains. We categorized errors into Retrieval Failures (search tools missing ground-truth sources) and Answer Failures (MLLMs failing to reason over correctly retrieved context). For English queries, errors were split between retrieval (77.8%) and answering (22.2%). However, non-English errors were overwhelmingly dominated by retrieval failures (94.0%). This demonstrates that the cross-lingual performance drop stems primarily from the search engines' inability to surface local-language sources, rather than the MLLMs' intrinsic reasoning deficits.

> ***Finding 2:*** Multilingual curse remains in visual factuality seeking, where search-augmented models exhibit strong language bias toward English primarily due to fundamental retrieval bottlenecks.

Model	Level	Arabic	Chinese	English	French	German	Indonesian	Japanese	Portuguese	Russian	Spanish	Average
Gemini 3 Pro Preview	1	0.1667	0.3591	0.0876	0.1500	0.2000	0.2258	0.2000	0.3500	0.3530	0.2333	0.2038
Gemini 3 Pro Preview	2	0.0500	0.2000	0.2174	0.3275	0.2833	0.1000	0.2184	0.1333	0.2933	0.2500	0.2087
GPT-5.2:online	1	0.2527	0.4571	0.5393	0.2677	0.2308	0.1667	0.2963	0.3333	0.3333	0.4529	0.3819
GPT-5.2:online	2	0.4000	0.3397	0.7244	0.3429	0.3529	0.4381	0.2498	0.2377	0.3750	0.3636	0.4581

Table 6: F-scores of Gemini 3 Pro Preview and GPT-5.2:online across different levels and languages.

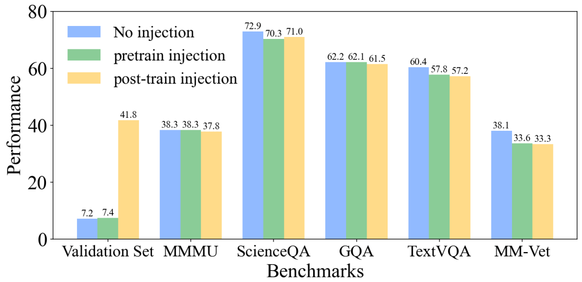


Figure 5: Performance (%) for three Tinyllava variants: the original TinyLLaVA-Phi-2-SigLIP-3.1B, and two derivatives obtained by injecting the dataset during pre-training or during post-training, respectively.

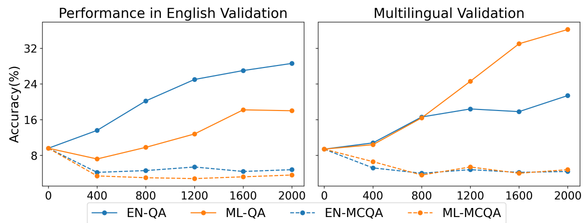


Figure 6: Accuracy over training steps on English and multilingual open-ended QA validation sets for models trained on different data formats (EN-QA, ML-QA, EN-MCQA, and ML-MCQA).

Figure 4 demonstrates a positive correlation between stated confidence and accuracy across models, though with significant calibration issues. As shown in Table 4, search-augmented models show better calibration compared to their counterparts, especially GPT-o3:online, reaching 37.4 F-score and only slightly below the ideal $y = x$ line, indicating that using the search tool releases overconfidence in visual factuality seeking and underscoring substantial opportunities for improving MLLM calibration. To provide a finer-grained analysis of multilingual calibration, we further report the F-scores of Gemini 3 Pro Preview and GPT-5.2:online across all 10 languages in Table 6.

> ***Finding 3:*** Search-augmented models exhibit strong calibration compared to offline models.

6 Updating MLLMs with Live Worldwide Visual Knowledge

While search-augmented models can effectively incorporate external visual factuality knowledge, they introduce latency and often fail to resolve semantically similar visual inputs. Building on prior work (Ravaut et al., 2024; Zeng et al., 2024; Chen et al., 2024; Fu et al., 2025), we dive deeper into ***how*** and ***when*** of *updating MLLMs with worldwide visual knowledge.* We also investigate ***how much*** and ***how stable*** updated knowledge is under cross-lingual and cross-modality disturbance settings.

6.1 Experiment Setups

We leverage the TinyLLaVA (Zhou et al., 2024) codebase to investigate the impact of knowledge injection. All experiments are conducted on 8×H100 servers. See Appendix C for detailed setups.

- For ***when***, we primarily investigate knowledge injection at different stages, *i.e.*, pre-training alignment injection and SFT injection. Phi-2-SigLIP-3.1B is selected as the base model.
- For ***how***, we analyze which data format is best suited for visual knowledge injection. We post-train both Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct for 2000 steps with knowledge to be injected with LoRA (Hu et al., 2022) across English and Multilingual content, as well as free-form and multiple-choice QA format. After training, we evaluate the model on languages different from those used during training to assess whether the injected visual knowledge can transfer across languages. To further demonstrate this correlation, we added a comparison showing the performance on questions in different languages across various training steps, using datasets of different languages and formats.
- To assess ***how much*** visual knowledge is injected, we construct a 500-sample validation set sampled from the training data in a *"[User]: Question; [Assistant]: Ground Truth"* format, with each question rephrased by Qwen3-30B-A3B to test knowledge retention. We also include general benchmarks such as MMMU (Yue et al., 2024), ScienceQA (Lu et al., 2022), GQA (Hud-

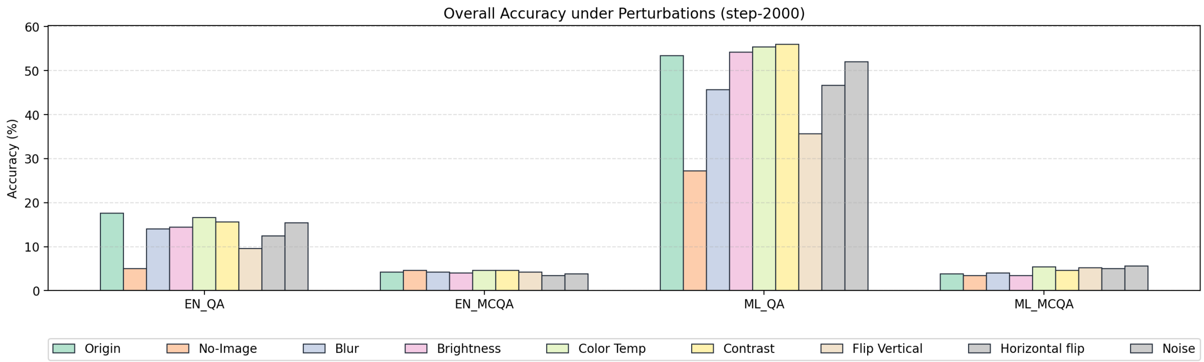


Figure 7: Performance of Qwen3-VL-4B-Instruct (trained on different data formats for 2000 steps) under various image perturbations.

son and Manning, 2019), TextVQA (Singh et al., 2019), and MM-Vet (Yu et al., 2024) to determine whether knowledge-intensive training harms general performance.

- To assess ***how stable*** visual knowledge is updated, we tested the model finetuned on four dataset formats on Multilingual QA with different question styles (EN/ML) and multimodal disturbance, including no image information, image blurring, brightness/color/contrast changing, noise, and vertical/horizontal flip.

6.2 Results and analysis

As shown in Figure 5 and Table 14, we observe that injecting visual entity knowledge during the pretraining phase yields minimal learning gains and may even slightly degrade the model's general performance. In contrast, during post-training (i.e., the supervised fine-tuning stage), the model demonstrates substantial memorization of visual factual information even after a single pass through the data (+480%). However, incorporating such data during post-training also introduces a notable degradation in general performance (-5.3% on TextVQA and -12.6% on MM-Vet). We hypothesize that this may be attributed to a domain gap between short, factuality-seeking QA pairs and other training data.

As shown in Figure 6, we find that although models exhibit improved memorization of visual factual knowledge, their ability to establish cross-lingual associations remains notably weak, where the corresponding training language significantly outperforms that in other languages. Multilingual visual factuality injection does not generalize well across languages, and substantial gaps between languages persist (>8%). This remains an open challenge that warrants further investigation. Furthermore, we observe that training with the MCQA format contributes minimally to factuality memorization; the model fails to internalize these knowledge entities or establish meaningful associations.

Finding 4: Multilingual visual factuality injection suffers from poor cross-lingual generalization, and MCQA training fails to facilitate effective knowledge memorization.

As shown in Figure 7, perturbation experiments reveal that vertical flipping causes the most severe drop in accuracy compared to other spatial transformations. Since the dataset focuses on visual entities without physical dynamics, this sensitivity suggests the model does not learn robust, rotation-invariant features. Instead, its recognition capability is anchored to the standard upright views and absolute spatial configurations present in the training data, leading to failure when the input deviates from this fixed layout.

Finding 5: The model degrades a lot under linguistic and query style changing, and removing images and flipping also harm visual entity recognition.

7 Conclusion

We present LiveVQA-W, the first large-scale dataset for real-time, multilingual visual knowledge seeking and updating. Our evaluation reveals that current MLLMs suffer from severe *"world-wide knowledge gap"* in visual knowledge seeking, and that injected visual knowledge remains fragile under perturbations, highlighting critical challenges for building truly global assistants. We release the fully replicable pipeline to support community-driven expansion to new languages and knowledge sources.

Acknowledgment

We thank Prof. Tianyi Zhou for his invaluable support. Dongping Chen is supported by Modal Academic Funding in this project.

Limitations

Due to copyright restrictions on certain websites and news sources, we are unable to release the VQA dataset processed from their news articles. Therefore, we will release the benchmark, toolkit, visual entity training/validation set and all raw data under the `CC-BY-NC-4.0` license in the camera-ready phase. While our work provides the community with a pipeline for large-scale visual factuality knowledge collection and processing, several limitations remain. Due to copyright restrictions on the crawled content, our dataset is released for non-commercial use only. Furthermore, constrained by computational resources, our experiments primarily focused on relatively small models with the conventional encoder-projector-LLM architecture. In future work, we plan to investigate whether models of varying scales and alternative architectures exhibit different capabilities in visual factuality memorization, to make further contributions to the research community.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Anthropic. 2025. Claude Sonnet 4.5. `https://www.anthropic.com/news/claude-sonnet-4-5`.

Shuai Bai and 1 others. 2025. Qwen3-VL technical report. `https://arxiv.org/abs/2511.21631`. *Preprint*, arXiv:2511.21631.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.

Mathilde Caron, Alireza Fathi, Cordelia Schmid, and Ahmet Iscen. 2024. Web-scale visual entity recognition: An llm-driven data approach. *Advances in Neural Information Processing Systems*, 37:34533–34560.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Qizhou Chen, Chengyu Wang, Dakan Wang, Taolin Zhang, Wangyue Li, and Xiaofeng He. 2024. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. *arXiv preprint arXiv:2411.15432*.

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*.

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen-tau Yih. 2025. Learning to reason for factuality. *arXiv preprint arXiv:2508.05618*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, and 1 others. 2025. Simplevqa: Multimodal factuality evaluation for multimodal large language models. *arXiv preprint arXiv:2502.13059*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*.

Yuntao Du., Kailin Jiang, Zhi Gao, Chenrui Shi, Zilong Zheng, Siyuan Qi, and Qing Li. 2025. MMKE-bench: A multimodal editing benchmark for diverse visual knowledge. In *The Thirteenth International Conference on Learning Representations*.

Yujie Feng, Xu Chu, Yongxin Xu, Zexin Lu, Bo Liu, Philip S Yu, and Xiao-Ming Wu. 2024. Kif: Knowledge identification and fusion for language model continual learning. *arXiv preprint arXiv:2408.05200*.

Mingyang Fu, Yuyang Peng, Benlin Liu, Yao Wan, and Dongping Chen. 2025. Livevqa: Live visual knowledge seeking. *arXiv preprint arXiv:2504.05288*.

Xingyu Fu, Ben Zhou, Ishaan Preetam Chandratreya, Carl Vondrick, and Dan Roth. 2022. There is a time and place for reasoning beyond the image. *arXiv preprint arXiv:2203.00758*.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, and 1 others. 2025. Webwatcher: Breaking new frontiers of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*.

Google DeepMind. 2025. Gemini 3 Flash. https://deepmind.google/models/gemini/flash/.

Jihao Gu, Yingyao Wang, Pi Bu, Chen Wang, Ziming Wang, Tengtao Song, Donglai Wei, Jiale Yuan, Yingxiu Zhao, Yancheng He, and 1 others. 2025. " see the world, discover knowledge": A chinese factuality evaluation for large vision language models. *arXiv preprint arXiv:2502.11718*.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. 2025. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*.

Guoxiu He, Xin Song, and Aixin Sun. 2025. Knowledge updating? no more model editing! just selective contextual reasoning. *arXiv preprint arXiv:2503.05212*.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, and 1 others. 2024. Chinese simpleqa: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2024a. Vlkeb: A large vision-language model knowledge editing benchmark. *arXiv preprint arXiv:2403.07350*.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024b. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 2024c. 1+ 1> 2: Can large language models serve as cross-lingual knowledge aggregators? *arXiv preprint arXiv:2406.14721*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2491–2498.

Pengyue Jia, Yingyi Zhang, Xiangyu Zhao, and Yixuan Li. 2025. Geoarena: An open platform for benchmarking large vision-language models on worldwide image geolocalization. *arXiv preprint arXiv:2509.04334*.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, and 1 others. 2024. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*.

Mladjan Jovanovic and Peter Voss. 2024. Towards incremental learning in large language models: A critical review. *arXiv preprint arXiv:2404.18311*.

Aditya Kumar and Manjinder Kaur. 2024. Live code sync. In *Advances in AI for Biomedical Instrumentation, Electronics and Computing*, pages 554–559. CRC Press.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, and 1 others. 2023. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, and 1 others. 2025. Mm-browsecomp: A comprehensive benchmark for multimodal browsing agents. *arXiv preprint arXiv:2508.13186*.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S Yu, Fei Huang, and 1 others. 2024. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. 2023. Edis: Entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4877–4894.

Zeyu Leo Liu, Shrey Pandit, Xi Ye, Eunsol Choi, and Greg Durrett. 2024. Codeupdatearena: Benchmarking knowledge editing on api updates. *arXiv preprint arXiv:2407.06249*.

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. *Preprint*, arXiv:2209.09513.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

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.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124.

Mistral AI. 2025. Mistral Medium 3.1. `https://mistral.ai/news/mistral-medium-3`.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. 2025. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, and 1 others. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

OpenAI. 2025a. GPT-5.2 system card. `https://openai.com/index/gpt-5-system-card-update-gpt-5-2/`.

OpenAI. 2025b. OpenAI o3 and o4-mini system card. `https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf`.

Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R Fung, and 1 others. 2025. Scaling laws of synthetic data for language models. *arXiv preprint arXiv:2503.19551*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. How much are large language models contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*.

Baochang Ren, Shuofei Qiao, Wenhao Yu, Huajun Chen, and Ningyu Zhang. 2025. Knowrl: Exploring knowledgeable reinforcement learning for factuality. *arXiv preprint arXiv:2506.19807*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Amit Rozner, Barak Battash, Lior Wolf, and Ofir Lindenbaum. 2024. Knowledge editing in language models via adapted direct preference optimization. *arXiv preprint arXiv:2406.09920*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

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.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and 1 others. 2024. Parrot:

Multilingual visual instruction tuning. *arXiv preprint arXiv:2406.02539*.

Yingshui Tan, Boren Zheng, Baihui Zheng, Kerui Cao, Huiyun Jing, Jincheng Wei, Jiaheng Liu, Yancheng He, Wenbo Su, Xiangyong Zhu, and 1 others. 2024. Chinese safetyqa: A safety short-form factuality benchmark for large language models. *arXiv preprint arXiv:2412.15265*.

Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. 2025. Mmsearch-plus: Benchmarking provenance-aware search for multimodal browsing agents. *arXiv preprint arXiv:2508.21475*.

Deep Search Team. 2024. Docling technical report. Technical report.

V Team. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Lukas Thede, Karsten Roth, Matthias Bethge, Zeynep Akata, and Tom Hartvigsen. 2025. Understanding the limits of lifelong knowledge editing in llms. *arXiv preprint arXiv:2503.05683*.

tonyassi. 2024. Celebrity-1000.

Chenlong Wang, Zhaoyang Chu, Zhengxiang Cheng, Xuyi Yang, Kaiyue Qiu, Yao Wan, Zhou Zhao, Xuanhua Shi, and Dongping Chen. 2025a. Codesync: Synchronizing large language models with dynamic code evolution at scale. *arXiv preprint arXiv:2502.16645*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. 2025b. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. 2025. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*.

xAI. 2025. Grok 4. `https://x.ai/`. Released July 9, 2025. xAI does not publish formal technical documentation. Accessed: 2026-01-06.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yibin Yan and Weidi Xie. 2024. Echosight: Advancing visual-language models with wiki knowledge. *arXiv preprint arXiv:2407.12735*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, and 1 others. 2024. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37:102101–102120.

Haote Yang, Xingjian Wei, Jiang Wu, Noémi Ligeti-Nagy, Jiaxing Sun, Yinfan Wang, Zijian Győző Yang, Junyuan Gao, Jingchao Wang, Bowen Jiang, and 1 others. 2025b. Openhueval: Evaluating large language model on hungarian specifics. *arXiv preprint arXiv:2503.21500*.

Zhenguo Yang, Jiale Xiang, Jiuxiang You, Qing Li, and Wenyin Liu. 2023. Event-oriented visual question answering: The e-vqa dataset and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10210–10223.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

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, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Zhen Zeng, Leijiang Gu, Xun Yang, Zhangling Duan, Zenglin Shi, and Meng Wang. 2024. Visual-oriented fine-grained knowledge editing for multimodal large language models. *arXiv preprint arXiv:2411.12790*.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and 1 others. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *Preprint*, arXiv:2402.14289.

A Full Related Works

Visual knowledge. Visual knowledge, also known as world knowledge, refers to the ability to connect visual information with broader facts, concepts, and relationships about the real world (Marino et al., 2019; Schwenk et al., 2022). This knowledge encompasses both commonsense understanding about objects, their interactions, and contextual relationships in visual scenes (Xie et al., 2019; Wang et al., 2015, 2017; Jain et al., 2021) and factual information from external sources (Caron et al., 2024; Yan and Xie, 2024; Jiang et al., 2024; Cheng et al., 2025). **This paper mainly focuses on visual factual knowledge, also known as visual entity and fact seeking.** The acquisition of visual knowledge involves leveraging external knowledge bases and structured repositories that ground visual elements in their broader conceptual context (Chen et al., 2022; Zhao et al., 2023; Caffagni et al., 2024; Yan and Xie, 2024; Abootorabi et al., 2025). Previous research curated static visual knowledge dataset with celebrity (tonyassi, 2024; Shah et al., 2019), landmarks (Weyand et al., 2020; Jia et al., 2025), bioinfo knowledge (Yang et al., 2024) and detailed properties (Mensink et al., 2023). Recent developments are expanding visual knowledge from static data to *"live"* visual content, addressing challenges in being more helpful as real-time and real-life multimodal assistants (Jiang et al., 2024; Fu et al., 2025; Geng et al., 2025; Li et al., 2025). This live visual knowledge specifically pertains to understanding and reasoning about current news (Fu et al., 2022), emerging events (Yang et al., 2023), cultural phenomena (Nayak et al., 2024; Romero et al., 2024), and temporally relevant information that constantly evolves (Du. et al., 2025; Huang et al., 2024a). Such live knowledge allows multimodal assistants to provide timely, relevant, and contextually appropriate responses to visual queries about ongoing situations and events.

Worldwide Multilingual Multimodal Dataset. Originating from the development of native multilingual assistants, researchers have argued that culture-specific knowledge should be trained with corresponding linguistic content (Manvi et al., 2024; Huang et al., 2024c; AlKhamissi et al., 2024). The same intuition also applies to multimodal models. Recent years have seen the release of numerous multilingual vision-and-language datasets spanning image captioning, VQA, and instruction tuning. Early benchmarks like Multi30K provided bilingual image captions (English–German), while the WIT dataset introduced over 11 million image–text examples across more than 100 languages (Srinivasan et al., 2021). Large-scale collections such as LAION-5B (5.85 billion image–text pairs with only 40% English) (Schuhmann et al., 2022), the video-and-language dataset VATEX (Wang et al., 2019), and Google's WebLI corpus (1 billion alt-text image-text pairs in 100 languages) (Wang et al., 2025b) have further expanded multilingual coverage. In VQA, MaRVL (Liu et al., 2021) and MAXM (Changpinyo et al., 2022) provide comprehensive benchmarks and datasets for visual reasoning across diverse languages. For vision–language instruction tuning, recent resources like M3IT (Li et al., 2023) and Parrot (Sun et al., 2024) compile dozens of tasks (*e.g.*, captioning and VQA) with data translated into multiple languages, enabling broad multilingual multimodal training. For visual factuality seeking and updating, benchmarks in Chinese (He et al., 2024; Tan et al., 2024; Gu et al., 2025) and Hungarian (Yang et al., 2025b) highlight the multilingual curse in factuality seeking, where model performance drops drastically across languages, uncovering the fragility of information-seeking and browsing-based environments and training.

Synthetic data for knowledge update. Knowledge updating, also known as continual learning, focuses on rapidly injecting the latest knowledge into pretrained models so they immediately recognize emerging concepts while retaining prior competencies (De Cao et al., 2021; Zhang et al., 2023; Huang et al., 2024a; Chen et al., 2024; Jovanovic and Voss, 2024; He et al., 2025). A major challenge in this process is catastrophic forgetting, where models lose previously acquired knowledge when learning new information, necessitating carefully-constructed high-quality data and specialized updating techniques (Luo et al., 2023; Huang et al., 2024b; Feng et al., 2024). To address these challenges, synthetic datasets have emerged as a critical solution for continuous knowledge infusion without extensive retraining (Thede et al., 2025; Abdin et al., 2024). For textual knowledge, frameworks like SynthLLM generate diverse, high-quality synthetic datasets by transforming existing corpora (Qin et al., 2025), while techniques such as Knowledge Direct Preference Optimization (KDPO) leverage synthetic examples for targeted factual updates (Rozner et al., 2024). Previous research in the language and code domain has successfully build up an automatic synthetic framework for code api knowledge synchronizing (Liu et al., 2024; Wang et al., 2025a; Kumar and Kaur, 2024). Our work extends LiveVQA's synthetic engine to a fully replicable pipeline with open-source models and incorporates worldwide visual entity knowledge, which automatically collects the latest visual knowledge in their domestic language from online sources and synthesize high-quality datasets for visual knowledge updating.

B Dataset Construction Details

B.1 Raw Data Collection & Preprocessing

Worldwide News.

We build a multilingual news corpus spanning **10 languages**, curated via an AI Agent-driven pipeline that automates source onboarding and content extraction, as illustrated in Figure 3. The full metadata collection and pipeline is detailed as follows:

- **AI-Agent-Driven Multilingual Source Onboarding.** We track ten languages: English, Chinese, Russian, Arabic, Japanese, Spanish, German, Indonesian, French, and Portuguese. For each language, we curate at least five news websites as information sources based on traffic volume, site stability, and HTML structural regularity. Building on this, we develop an AI Agent–driven automation pipeline: the agent analyzes the target website's page structure and automatically generates a site-specific configuration along with a modular crawler script comprising dedicated components for URL discovery, content extraction, and robust date parsing. After self-validation and iterative refinement, the agent deploys the solution, enabling fully automated onboarding of new sites.
- **Image extraction, deduplication and filtering.** We extract images from articles using CSS selectors that cover common content management system (CMS) templates, and discard those with excessively small dimensions (e.g., website icons). Subsequently, we perform visual deduplication based on histogram correlation: for any pair of images whose similarity score exceeds **0.85**, only the one with the larger pixel area is retained. To further ensure relevance to the article content, we filter out potentially advertisement-like images by determining whether they are embedded within hyperlinks that point to external domains. Finally, at most **4** high-quality images are preserved per article.
- **Initial caption collection.** We prioritize article-provided figure captions as image titles; if unavailable, we fall back to the `img@alt` text, and use it as the initial caption field.

Multilingual Videos. We crawl YouTube across ten languages using a hierarchical language-identification (LID) chain, transcribe audio with Whisper-large-v3 (Radford et al., 2022), and apply an early validity filter with Qwen3-A3B-30B (Yang et al., 2025a) to eliminate noise-only or low-information content; language-aware prompt templates then generate QA pairs that are verified for extractability and safety.

- **Data Acquisition.** We extend video retrieval into a configurable crawling paradigm that supports multiple languages and cross-region coverage. For each video, we perform language identification in a prioritized manner: we first rely on metadata signals; if unavailable, we query the subtitle track language, preferring human-curated captions over ASR-generated ones; as a fallback, we clean the title and description text and infer the language from the processed content, which mitigates language drift and improves the stability of cross-lingual dataset construction.

- **ASR without auto-caption dependency.** Whisper-large-v3 transcribes audio into time-aligned segments even when auto-captions are disabled; we chunk audio into fixed windows with overlap to preserve context, perform VAD to skip silence, and normalize casing, punctuation, numerals, and non-speech events while retaining per-segment timestamps for later alignment with QA spans.
- **Keyframe Mining and Quality-Controlled Frame Preparation.** To process video content at scale, we first run UVD to propose representative keyframe candidates, thereby reducing temporal redundancy. We then apply a representative-frame selection strategy to suppress large volumes of near-duplicate frames, retaining only the most informative instance within each visually similar cluster. Next, we use DocLayout-YOLO to crop each image and remove overlaid text and other irrelevant regions. Finally, to prevent residual near-identical content from being over-retained, we perform an additional deduplication step following a similar criterion to the previous filtering stage, further distilling the selected images into a more compact and representative set.
- **Cost-Efficient verification and selection.** We use Qwen3-30B-A3B for transcript-validity checking and video-topic cleaning, and Qwen3-VL-30B-A3B for final keyframe selection. This design reduces reliance on paid APIs while maintaining strong reasoning performance across the filtering pipeline. We further batch ASR and verification to improve GPU utilization and amortize per-sample context construction overhead.

Extended Academia Platform We ingest preprints from arXiv, ChemRxiv, and medRxiv by downloading PDFs and parsing them with `Docling` (Team, 2024) to recover figures and surrounding text; captions are paired to figures via proximity heuristics, and a lightweight scorer then selects up to three representative images per paper. This PDF-first design avoids the brittleness of HTML scraping and expands coverage beyond arXiv.

- **PDF-based structural parsing and figure–caption alignment.** Each PDF is parsed using `Docling` to recover page layout, figure bounding boxes, and nearby text; captions are bound to figures using keyword cues (*Figure*, *Fig.*, *Image*, etc.) and spatial proximity, and the resulting pairs are stored with page indices and character offsets to preserve traceability back to the original document.
- **Repository expansion and failure handling.** In addition to arXiv, we ingest ChemRxiv and medRxiv; failed downloads trigger scheduled retries with mirrored endpoints and error logging that captures HTTP codes and redirection chains so that flaky sources do not block the pipeline.
- **Representative-image selection.** Images below a minimum area threshold are discarded outright. For the remaining images, a classifier that incorporates caption semantics assigns scores on a 0–10 scale, giving higher weight to conceptual or architectural diagrams and penalizing standard experimental plots. The top three images with scores ≥ 8 are retained per paper as representatives.

B.2 Multilingual Visual Questions and Answers Generation

We generate QA pairs from the worldwide News and Videos data through a three-stage pipeline. First, we use a multimodal large language model to produce image captions, thereby replacing raw images with textual descriptions that preserve the salient visual content. Second, we construct Level-1 questions that directly query worldwide visual entities and apply strict filtering to remove ambiguous, low-quality, or ill-formed items. Third, building on the Level-1 questions, we generate Level-2 questions that require multi-hop reasoning, where the model must effectively integrate visual information from the caption with external web knowledge to perform deeper inference. Finally, we conduct careful quality control for both levels, including model self-answering and rule-based validity checks, to ensure that the resulting questions are of high quality.

Caption Generation. We use Qwen3-VL-30B to produce fine-grained image descriptions, aiming to capture as many visual details as possible in text. For the News subset, the original images are accompanied by human-written captions; we provide these captions as additional input to the model and regenerate a consolidated caption accordingly. Since caption generation primarily requires accurate visual grounding rather than long-context understanding or rule-based reasoning, we find that a large vision-language model is not strictly necessary to obtain high-quality captions.

QA Generation. Leveraging the generated image captions, we can rely on a text-only large language model to focus on question generation. High-quality question writing is governed by numerous constraints and criteria, which demand strong reasoning ability; under comparable compute budgets, text-only models typically deliver better generation quality than vision-language models for this stage. We categorize Level-1 questions into six types: person, object, organization, location, event, and time, which usually query visual entities explicitly mentioned or implied by the image. In contrast, Level-2 questions are defined on top of Level-1 but exclude the object category and add a non-OCR count category, enabling reasoning from visual entities to more abstract concepts and deriving counts through multi-hop inference. For the Papers subset, Level-1 questions are further divided into two categories, authors and title, while Level-2 questions are divided into data and conclusion. All QA-generation prompts are written in the same language as the target QA, to ensure linguistic consistency between instructions and outputs, minimize cross-lingual variability, and avoid confounding effects from translation or language mismatch during generation and evaluation.

QA Filter. During QA pair generation, we apply a three-step filtering process to meet our dataset's quality requirements. QA Filter 1 targets Level-1 questions, checking visual grounding, non-OCR reliance, answer uniqueness, and the validity of the question stem, with particular emphasis on language consistency in the multilingual setting. QA Filter 2 targets Level-2 questions and uses LLM self-answering as a verification signal: given the article text and other context, the model must be able to answer the corresponding Level-2 question correctly. Since QA Filter 2 is relatively coarse-grained, we finally apply a Final Filter to both Level-1 and Level-2 questions, performing rule-based checks such as screening for factual errors and enforcing well-formed, standardized question phrasing.

B.3 Human Annotation Details

Every experiment using LLM/MLLM is validated with human-annotated ground truth and agreement. We provide detailed instructions and annotation environments. The annotation is conducted by 5 authors of this paper independently. All the annotations are conducted under Streamlit[1]. As acknowledged, the diversity of annotators plays a crucial role in reducing bias and enhancing the reliability of the benchmark. These annotators have rich knowledge in this domain, with different genders, ages, and educational backgrounds. To ensure the annotators can proficiently mark the data, we provide them with detailed tutorials, teaching them how to evaluate model responses more objectively. Specifically, they are required to give judgments without bias, like answer lengths, and certain names of the response. All processes using LLM/MLLM are listed as follows:

1. **Video Data - subtitle parsing and event segmentation with Qwen3-30B-A3B.** We parse subtitle files into timestamped word-level tokens and then use Qwen3-30B-A3B to group them into coherent event segments. The segmentation leverages textual cues, accordingly, we introduce segment boundaries only when the model identifies a clear topic shift. To validate the structural quality of the resulting segments, we build a lightweight annotation interface (Figure 8); manual verification shows that 93% of the inspected segments are judged to be appropriately split.
2. **Video Data - images selecting with Qwen3-VL-30B-A3B.** We exploit Qwen3-VL-30B-A3B as an automated visual curator to filter large pools of video frames into a compact, QA-ready subset. The pipeline is explicitly two-stage: (i) batch scoring, where Qwen3-VL-30B-A3B rates each image (1–10) conditioned on the topic and content description, prioritizing relevance, clarity, information richness, and minimal textual interference that could leak answers; and (ii) set-level selection, where the model chooses up to five images from the top candidates while enforcing diversity by rejecting near-duplicates and redundant viewpoints. To assess the filtering quality of Qwen3-VL-30B-A3B, we developed a lightweight annotation interface (Figure 9) with a passing rate 94%.
3. **Academic paper - key image selection with GLM-4.6.** We use GLM-4.6 to select representative images from each article by providing the model with image captions as a proxy for

[1] https://streamlit.io/

Language	News Sources and URLs
English	**BBC** (https://www.bbc.com/), **CNN** (https://edition.cnn.com/), **Variety** (https://variety.com/), **Forbes** (https://www.forbes.com/), **AP News** (https://apnews.com/)
Chinese	**Xinhua** (https://www.news.cn/), **People's Daily** (https://www.people.com.cn/), **The Paper** (https://www.thepaper.cn/), **Southern Weekly** (http://www.infzm.com/), **Phoenix** (https://www.ifeng.com/)
Spanish	**El País** (https://elpais.com/), **El Mundo** (https://www.elmundo.es/), **La Nación** (https://www.lanacion.com.ar/), **El Universal** (https://www.eluniversal.com.mx/), **ABC** (https://www.abc.es/)
Arabic	**Al Jazeera** (https://www.aljazeera.net/), **Al-Watan** (https://www.elwatannews.com/), **Al Mayadeen** (https://www.almayadeen.net/), **Al-Riyadh** (https://www.alriyadh.com/), **Sabq** (https://sabq.org/)
Portuguese	**Público** (https://www.publico.pt/), **Diário de Notícias** (https://www.dn.pt/), **O Globo** (https://oglobo.globo.com/), **Folha de S.Paulo** (https://www.folha.uol.com.br/), **Correio da Manhã** (https://www.cmjornal.pt/)
Indonesian/Malay	**ANTARA** (https://www.antaranews.com/), **Kompas** (https://www.kompas.com/), **Detik** (https://www.detik.com/), **Utusan Malaysia** (https://www.utusan.com.my/), **The Star** (https://www.thestar.com.my/)
French	**Le Monde** (https://www.lemonde.fr/), **L'Express** (https://www.lexpress.fr/), **France-info** (https://www.franceinfo.fr/), **Libération** (https://www.liberation.fr/), **RT France** (https://francais.rt.com/), **L'Obs** (https://www.nouvelobs.com/), **20 Minutes** (https://www.20minutes.fr/)
Japanese	**Asahi** (https://www.asahi.com/), **Sankei** (https://www.sankei.com/), **Mainichi** (https://mainichi.jp/), **NHK** (https://www3.nhk.or.jp/news/), **Nikkei** (https://www.nikkei.com/)
Russian	**RIA Novosti** (https://ria.ru/), **Kommersant** (https://www.kommersant.ru/), **Gazeta.ru** (https://www.gazeta.ru/), **Rossiyskaya Gazeta** (https://rg.ru/), **TASS** (https://tass.ru/)
German	**Der Spiegel** (https://www.spiegel.de/), **FAZ** (https://www.faz.net/), **Die Zeit** (https://www.zeit.de/), **SZ** (https://www.sueddeutsche.de/), **Deutsche Welle** (https://www.dw.com/de/), **BILD** (https://www.bild.de/), **Tagesschau** (https://www.tagesschau.de/)

Table 7: News websites serving as information sources in ten languages.

the images, i.e., We define key images as those whose captions most clearly and uniquely identify the referenced research paper, distinguishing it from other publications. We further assess the plausibility of the GLM-4.6's choices with a Streamlit-based labeling interface, achieving a 95% accuracy rate.

4. **News article - image filter with GLM-4.6.** We filter article images using caption-only inputs—the model receives the generated captions as a textual proxy, rather than the raw images themselves. We keep only visuals that meaningfully capture the article's ongoing social topic, and remove static objects, logos, near-duplicates, generic/decorative illustrations, or any image that does not provide distinctive evidence about the core story. An image is retained only if it contributes information that cannot be recovered from the text alone and is necessary for understanding the report. We then generate factual captions restricted to explicitly supported details. Manual assessments demonstrate that GLM-4.6 achieved a 96% accuracy rate.

5. **Image caption generation with Qwen3-VL-30B-A3B.** We use the Qwen-VL-30B-A3B model to generate a caption for each image in the Videos and News subsets, where the caption provides a detailed description of the image and serves as a surrogate for the original image in subsequent stages. We employ Qwen3-VL-30B-A3B to finish this job and design the annotation tool in Figure10. Manual assessments show Qwen3-VL-30B-A3B achieved a 99% accuracy rate.

6. **Level-1 question generation with GLM-4.6.** We construct basic QA pairs by prompting GLM-4.6 with an image caption (rather than the image itself) as a textual proxy for visual content, and ask it to generate a question Based on the caption. The answer must be explicitly supported by the accompanying article. We exclude undesirable cases such as celebrity identification, counting, and shortcuts based on OCR cues (e.g., visible text or logos). We use GLM-4.6 for this generation step and provide an annotation interface for inspection and quality control (Figure 10). Manual assessments show GLM-4.6 achieved a 93% accuracy rate.

7. **QA Filter 1 with GLM-4.6.** We discard articles lacking meaningful context (e.g., only noting a broadcast question). We always keep questions whose answers are non–major-figure

names, specific named locations, named events, specific products, or organizations. We employ GLM-4.6 to remove any unqualified QA pairs and use an annotation tool to evaluate. The annotation tool is shown in Figure 11. The accuracy of GLM-4.6 is 93%.

8. **Level-2 question generation with GLM-4.6.** We construct Level-2 multi-hop QA pairs, restricted to six categories (person, organization, location, event, time, and non-OCR count; excluding object). Prompts match the target language to avoid cross-lingual confounds. We use GLM-4.6 for generation; self-answering reaches 95% accuracy.
9. **QA Filter 2 with GLM-4.6.** We verify Level-2 questions via GLM-4.6 self-answering and retain only those for which GLM-4.6, given the article text and the image caption as context, produces an answer that matches the original ground truth. The accuracy of GLM-4.6 is 98%.
10. **Final filter with GLM-4.6.** We apply a unified, rule-based quality check to all final Level-1 and Level-2 questions using GLM-4.6, following a comprehensive set of criteria to ensure quality, including screening for factual errors and validating the well-formedness of question stems. The accuracy of GLM-4.6 is 99%.

To ensure the timeliness and relevance of our video data, we perform manual verification with controlled labeling environment. The authors independently review a random sample of 500 entries, checking that the selected news and videos depict current events rather than historical footage. In this audit, all inspected examples are judged to be recent, indicating that outdated content is effectively mitigated by our curation process.

C Detailed Experiment Setups

Models. The detailed specifications of all models evaluated in the experiments of Section 5 are provided in Table 10.

Metric: Guessing strategy and F-score. Originate from Wei et al. (2024), while F-score is a good metric in some ways, the issue with it is that it incentivizes the model to always guess when it is at least 50% sure that it can get the correct answer. To understand why this is the case, consider the following expression for the F-score:

$$F\text{-}score = \frac{2}{\frac{c+i}{c} + \frac{c+i+n}{c}} = \frac{2c}{2c+2i+n},$$

where:

- c is the number of correct answers,
- i is the number of incorrect answers, and
- n is the number of non-answered questions.

If you have a greater than $\frac{1}{2}$ chance of being correct, your expected score from guessing is better than the score from not guessing, regardless of the specific values for c, i, and n. This is because the following inequality always holds:

$$\frac{2c+1}{2c+2i+n+2} > \frac{2c}{2c+2i+n+1}.$$

The left-hand side represents the expected F-score from guessing, assuming a 50/50 chance of correctness, while the right-hand side is the score from not answering the additional question. Since the denominators are adjusted similarly whether the guess is correct or incorrect, guessing with a probability $> \frac{1}{2}$ yields a better score.

Model pretraining details. We conduct pretraining using the TinyLLaVA training pipeline with DeepSpeed across 8*H100 GPUs. Training is performed in FP16 with FlashAttention-2. The hyperparameters we used during pretraining are listed in Table 8. We augment the original LLaVA-1.5-558K pretraining dataset by injecting 5,000 image–caption pairs as visual knowledge from LiveVQA-W, with 2,500 samples drawn from the News subset and 2,500 from the Videos subset. We further adopt a data replay strategy, repeating these 5,000 injected pairs 10 times within the mixed training data to increase their exposure during pretraining.

Model fine-tuning details. For TinyLLaVA, we conduct fine-tuning using the TinyLLaVA training pipeline with DeepSpeed across 8*H100 GPUs. Training is performed in FP16 with FlashAttention-2. The hyperparameters we used during fine-tuning are listed in Table 9. We augment the original LLaVA-1.5-mix-665k fine-tuning dataset by injecting 5,000 QA pairs instead of image-caption pairs as visual knowledge from LiveVQA-W, with the same composition and data replay strategy as above.

D Additional Experiment Results

See Table 11 for the overall performance for Level 1 and 2 questions. See Table 12 for the performance of Qwen3-VL-8B trained on different data

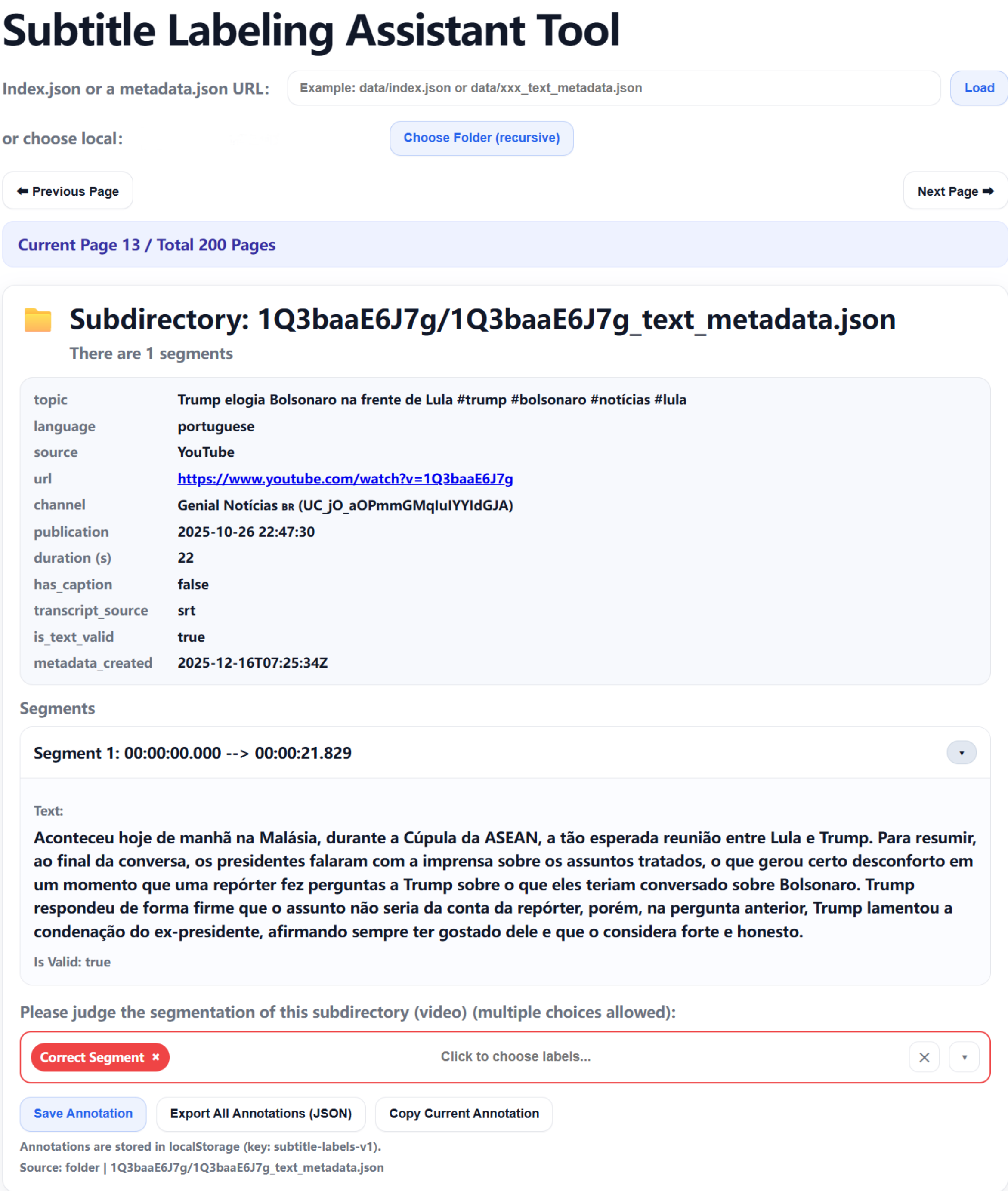


Figure 8: Human Annotation - Video Data - Parsing Subtitle by Qwen3-VL-30B-A3B.

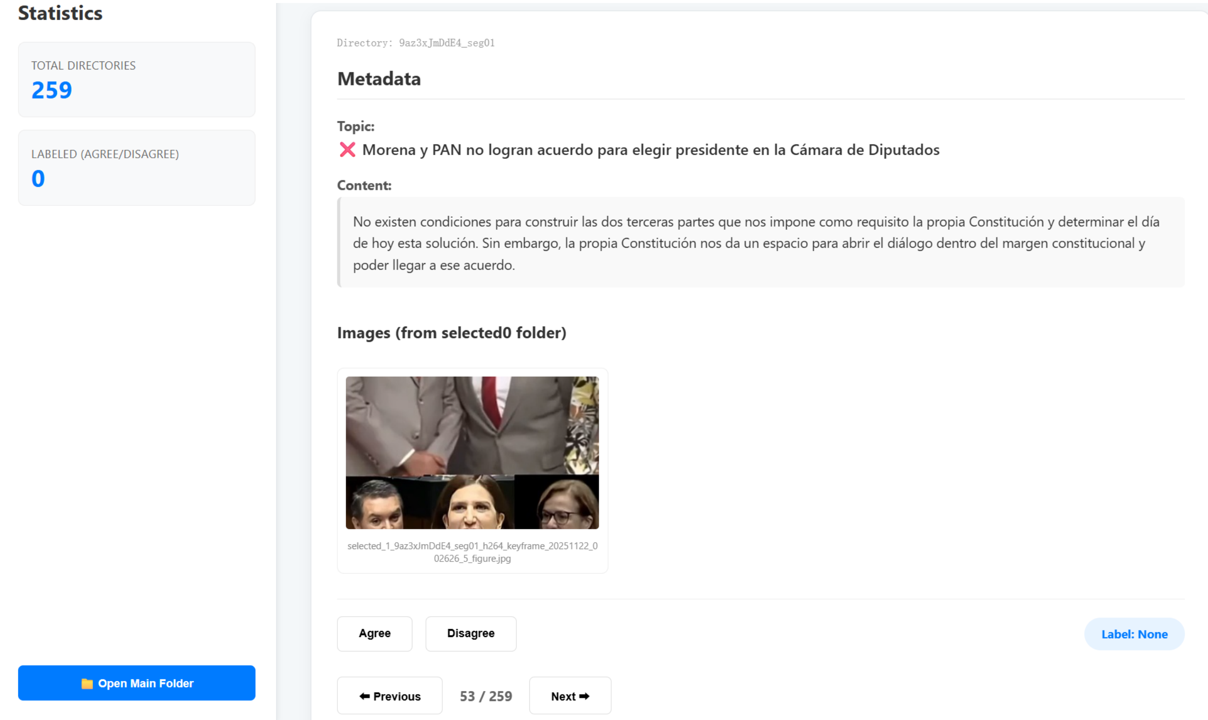


Figure 9: Human Annotation - Video Data - Selecting Images for Video Data.

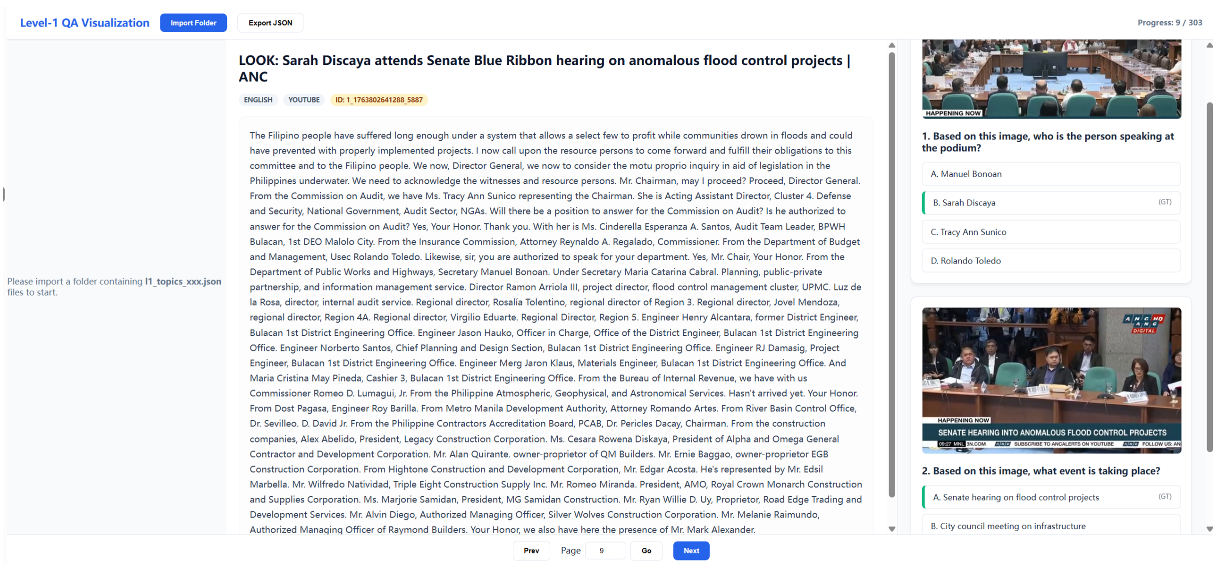


Figure 10: Human Annotation - Level-1 QA generation.

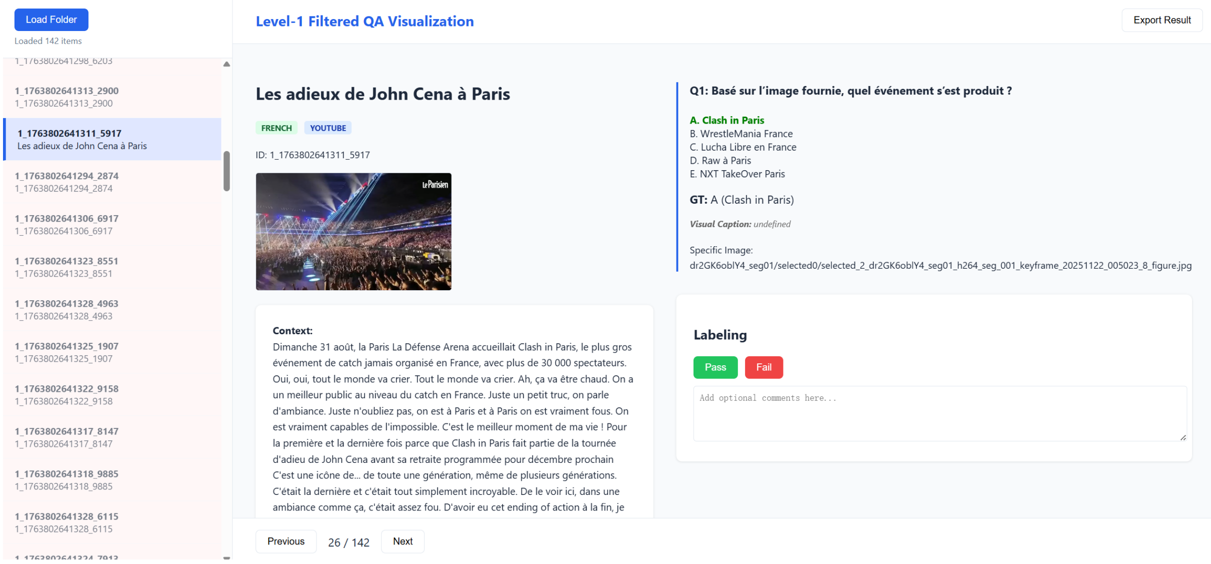


Figure 11: Human Annotation - QA Filter 1.

Table 8: Overview of the pretraining (stage1) hyperparameters.

Hyperparameter	Value
lr	1e-3
lr_scheduler	cosine
warmup ratio	0.03
global batch size	256
epoch	1
optimizer	AdamW

Table 9: Overview of the finetune (stage2) hyperparameters.

Hyperparameter	Value
lr	2e-5
lr_scheduler	cosine
warmup ratio	0.03
global batch size	128
epoch	1
optimizer	AdamW

formats under various perturbations. See Table 6 for F-scores across different levels and languages. See Table 13 for performance under category-wise breakdown. See Table 14 for the performance of visual knowledge update under different training strategies.

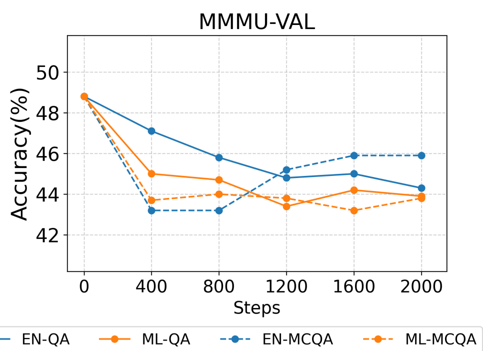


Figure 12: Overall accuracy over training steps on MMMU-Validation sets for models trained on different data formats.

E LLM Utilization

The use of large language models (LLMs) in this work is strictly limited to auxiliary text editing, such as correcting spelling and improving grammar, and dataset generation. During dataset synthesis, LLMs are extensively used for video and transcript cleaning and preprocessing, as well as for generating and filtering two levels of question-answer pairs and their multiple-choice options. All conceptual and technical contributions are the original work of the authors. We are transparent about this limited usage.

F Case Study and Prompts

See Figure 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36 for case study.

Model	**Creator**	**Version**	**Knowledge Cutoff**	**License**
Claude Sonnet 4.5	Anthropic	claude-sonnet-4-5-20250929	2025.7	Proprietary
Gemini 3 Flash Preview	Google	gemini-3-flash-preview	2025.1	Proprietary
Gemini 3 Pro Preview	Google	gemini-3-pro-preview	2025.1	Proprietary
Gemma 3 27B	Google	gemma-3-27b-it	2024.8	Open-source
Mistral Medium 3.1	Mistral AI	mistral-medium-2508	2025.6	Proprietary
GPT-5.2	OpenAI	gpt-5.2	2025.8	Proprietary
o3	OpenAI	o3-2025-04-16	2024.6	Proprietary
Qwen3-VL-235B-A22B-Thinking	Alibaba	Qwen3-VL-235B-A22B-Thinking	Unknown	Open-source
Grok 4	xAI	grok-4	2024.11	Proprietary
GLM-4.6V	Zhipu AI	GLM-4.6V	Unknown	Open-source
		w. Text & Image Search		
GPT-5.2:online	OpenAI	gpt-5.2	2025.8	Proprietary
o3:online	OpenAI	o3-2025-04-16	2024.6	Proprietary
MMSearch-R1	ByteDance	MMSearch-R1	-	Open-source
WebWatcher-7B	Alibaba-NLP	WebWatcher-7B	-	Open-source
WebWatcher-32B	Alibaba-NLP	WebWatcher-32B	-	Open-source

Table 10: Model names, Creators, Version, Knowledge Cutoff and License.

Model	**Cutoff**	**Level 1**				**Level 2**			
		News	Videos	Papers	Avg.	News	Videos	Papers	Avg.
				***w.o.* Search**					
Claude Sonnet 4.5	Jul. 2025	5.0	5.0	0.0	4.0	1.0	4.0	2.0	2.4
Gemini 3 Flash Preview	Jan. 2025	**28.0**	**19.0**	0.0	**18.8**	9.0	14.0	16.0	12.4
Gemini 3 Pro Preview	Jan. 2025	27.0	17.0	**2.0**	18.0	**15.0**	**19.0**	**18.0**	**17.2**
Gemma 3 27B	Aug. 2024	4.0	4.0	**2.0**	3.6	3.0	5.0	6.0	4.4
Mistral Medium 3.1	Jun. 2025	11.0	5.0	0.0	6.4	2.0	7.0	4.0	4.4
GPT-5.2	Aug. 2025	3.0	4.0	0.0	2.8	2.0	10.0	14.0	7.6
GPT-o3	Jun. 2024	6.0	8.0	0.0	5.6	6.0	16.0	**18.0**	12.4
Qwen3 VL 235B A22B Thinking	Unknown	5.0	6.0	0.0	4.4	3.0	5.0	8.0	4.8
Grok 4	Nov. 2024	11.0	12.0	0.0	9.2	5.0	7.0	10.0	6.8
GLM 4.6V	Unknown	8.0	3.0	0.0	4.4	1.0	3.0	4.0	2.4
				***w.* Text & Image Search**					
GPT-5.2:online	Aug. 2025	23.0	**23.0**	**46.0**	**27.6**	**31.0**	28.0	**76.0**	38.8
GPT-o3:online	Jun. 2024	20.0	17.0	38.0	22.4	29.0	**37.0**	64.0	**39.2**
MMSearch-R1	-	16.0	9.0	6.0	11.2	8.0	9.0	36.0	14.0
WebWatcher-7B	-	28.0	21.0	36.0	26.8	17.0	26.0	28.0	22.8
WebWatcher-32B	-	**29.0**	22.0	32.0	26.8	22.0	26.0	26.0	24.4

Table 11: Accuracy (%) of visual factuality seeking benchmark in open-ended format across different models across difficulty levels and data sources.

Dataset	**Origin**	**Image Perturbations**							
		No-Image	Blur	Brightness	Color Temp	Contrast	Flip Vertical	Horizontal Flip	Noise
EN_QA	0.238	0.114	0.190	0.236	0.246	0.224	0.146	0.176	0.240
EN_MCQA	0.042	0.048	0.038	0.038	0.054	0.044	0.048	0.050	0.042
ML_QA	0.488	0.252	0.378	0.414	0.432	0.470	0.326	0.416	0.466
ML_MCQA	0.050	0.046	0.038	0.036	0.032	0.036	0.040	0.038	0.040

Table 12: Performance of Qwen3-VL-8B-Instruct trained on different data formats for 2000 steps under various image perturbations.

Model	Avg.	Authors	Conclusion	Count	Data	Event	Location	Object	Organization	Person	Time	Title
						***w.o.* Search**						
Claude Sonnet 4.5	3.20	0.00	0.00	1.72	0.00	1.85	5.41	0.00	4.84	5.08	2.22	0.00
Gemini 3 Flash Preview	15.60	0.00	37.50	5.17	14.29	7.41	13.51	30.00	16.13	**33.05**	6.67	0.00
Gemini 3 Pro Preview	17.60	2.63	50.00	13.79	19.05	7.41	17.57	30.00	12.90	32.20	11.11	0.00
Gemma 3 27B	4.00	2.63	12.50	1.72	4.76	5.56	2.70	0.00	6.45	4.24	4.44	0.00
Mistral Medium 3.1	5.40	0.00	12.50	0.00	0.00	1.85	5.41	20.00	4.84	12.71	2.22	0.00
GPT-5.2	5.20	0.00	62.50	5.17	4.76	5.56	9.46	0.00	4.84	2.54	2.22	0.00
GPT-o3	9.00	0.00	37.50	6.90	14.29	9.26	17.57	10.00	4.84	7.63	8.89	0.00
Qwen3 VL 235B A22B Thinking	4.60	0.00	25.00	5.17	4.76	0.00	6.76	0.00	1.61	8.47	2.22	0.00
Grok 4	8.00	0.00	37.50	1.72	9.52	7.41	5.41	10.00	9.68	16.10	0.00	0.00
GLM 4.6V	3.40	0.00	25.00	0.00	0.00	1.85	2.70	0.00	4.84	7.63	0.00	0.00
						***w.* Text & Image Search**						
GPT-5.2:online	**33.20**	**50.00**	**100.00**	32.76	**85.71**	25.93	25.68	40.00	**40.32**	16.95	**35.56**	33.33
GPT-o3:online	30.80	44.74	87.50	**37.93**	71.43	22.22	**35.14**	30.00	29.03	13.56	**35.56**	16.67
MMSearch-R1	12.60	5.26	37.50	10.34	33.33	9.26	12.16	0.00	14.52	11.02	17.78	8.33
WebWatcher-7B	24.80	28.95	37.50	13.79	23.81	25.93	22.97	**50.00**	22.58	24.58	24.44	**58.33**
WebWatcher-32B	25.60	26.32	25.00	20.69	23.81	**33.33**	24.32	20.00	22.58	23.73	28.89	50.00

Table 13: Accuracy (%) of different models breakdown by question type.

Benchmark	Subject/Type/Strategy	Accuracy (%)		
		M0	M1	M2
Validation Set	Event	2.2	1.1	15.1
	Location	10.0	14.2	55.8
	Object	18.6	11.6	44.2
	Organization	9.3	10.5	59.3
	Person	3.3	2.5	40.2
	Time	5.6	5.6	25.0
	Avg	**7.2**	**7.4**	**41.8**
MMMU	Art & Design	55.8	55.8	53.3
	Business	34.0	32.7	34.7
	Science	28.0	26.7	28.7
	Health & Medicine	46.0	41.3	42.7
	Humanities & Social Science	47.5	50.8	48.3
	Tech & Engineering	28.1	31.4	28.1
	Avg	**38.3**	**38.3**	**37.8**
ScienceQA(multimodal)	Natural	72.4	70.0	70.5
	Language	79.6	77.2	79.6
	Social	73.4	70.4	71.3
	Avg	**72.9**	**70.3**	**71.0**
GQA	Attribute	68.4	68.6	67.5
	Category	54.2	54.8	54.0
	Global	59.2	59.9	60.5
	Object	86.1	86.3	85.7
	Relation	54.4	53.7	53.8
	Avg	**62.2**	**62.1**	**61.5**
TextVQA	**Avg**	**60.4**	**57.8**	**57.2**
MM-Vet	Recognition	41.1	37.5	38.3
	OCR	33.3	25.9	24.1
	Knowledge	28.6	25.2	26.4
	Language Generation	28.4	24.5	26.2
	Spatial Awareness	36.1	26.7	25.6
	Math	11.9	6.9	13.8
	Avg	**38.1**	**33.6**	**33.3**

Table 14: Detailed benchmark accuracies (%) for three TinyLLaVA variants: the original TinyLLaVA-Phi-2-SigLIP-3.1B (M0), and two derivatives obtained by injecting the dataset during pretraining (M1) or during post-training (M2), respectively.

LLM Backbone	Vision Tower	Validation Set			MMMU			ScienceQA			TextVQA			MM-Vet		
		None	Pre	Post	None	Pre	Post	None	Pre	Post	None	Pre	Post	None	Pre	Post
Qwen3-0.6B	siglip-so400m-patch14-384	3.0	4.0	15.0	34.1	33.1	33.1	60.8	60.8	59.6	52.1	51.6	51.4	25.2	28.2	24.9
Qwen3-1.7B	siglip-so400m-patch14-384	5.4	4.8	15.8	39.8	39.8	37.7	69.1	70.8	68.5	55.5	56.6	55.7	31.1	31.9	30.6
Phi-2	siglip2-so400m-patch14-384	6.6	6.8	42.8	37.3	36.0	38.1	70.8	71.6	70.9	58.7	61.6	60.4	35.8	34.0	34.8
Phi-2	metaclip-2-worldwide-b16-384	5.0	5.8	30.8	35.9	34.6	36.1	69.1	69.7	68.0	48.4	49.4	48.5	26.9	25.1	29.4

Table 15: Performance (%) for various MLLM architectures based on the TinyLLaVA framework under no injection, pre-training injection and post-training injection settings across visual knowledge validation set and general benchmarks.

Prompt: Transcript Validity Check

Analyze whether the following transcribed text is valid and meaningful. Check for:

1. Empty content, or only background noise/music descriptions.
2. Only filler words or meaningless utterances.
3. Completely incoherent or nonsensical content.
4. Text length < 100, and the information lacks timely content, is outdated, or contains too little valid information.

Text sample:

{sample_text}

Respond in JSON with the following structure:

```
{
  "is_valid": true/false,
  "reason": "Brief explanation if invalid"
}
```

Prompt: Timestamped Transcipt Segmentation

You are an expert at segmenting timestamped transcripts into coherent paragraphs based on **TOPIC CONTENT**.
Segmentation Principles:

1. **MINIMIZE the number of segments — this is CRITICAL!** Aim for as few segments as possible.
2. **TOPIC CONTINUITY is the PRIMARY criterion** — keep all content about the same event/topic together.
3. If multiple sentences contain **SHARED KEYWORDS** or related concepts, they **MUST** be grouped together.
4. If adjacent or nearby sentences mention the same entities (people, places, events), they **MUST** be merged.
5. Changes in speaker or dialogue format should **NOT** create new segments if the topic remains related.
6. **Only** create a new segment for a **COMPLETE TOPIC CHANGE** to an unrelated subject.

Special Notes:

- If content discusses different aspects of the same general topic (e.g., different angles or consequences of one event), keep it all in **ONE** segment.
- Look for **semantic relationships** between sentences, not just superficial connections.
- The goal is to create **COMPREHENSIVE segments** that cover complete topics, not short fragments.

- **(New) Avoid creating excessively short segments:** If a segment's duration (end_time − start_time) is too short (e.g., less than 0.5 seconds), carefully verify if the segmentation is correct. Unless it is a very brief, distinct utterance representing a complete topic, try to merge it with adjacent, topically related segments.
- If the content is a TV program outro, credits, music, or thank-you message that is distinct and lengthy, always segment it separately; these segments can be flagged as non-content.

Additional Guidance:

- You MAY use large time gaps (e.g., > 3 seconds) between subtitles as a **secondary** clue for segmentation, but **ALWAYS prioritize topic continuity over timing**.
- **PRIORITIZE content similarity over timestamp gaps** — related content should stay together even with pauses.
- **(New) Tendency to merge on short intervals:** As a **secondary signal**, if the time gap between the end of one line and the start of the next is **very short** (e.g., less than 1 second), and their topics are **related or continuous**, then they are **more likely** to belong to the same segment. This supports merging when topic continuity is present, but should not override a clear topic change.

Example — Should be ONE segment (same policy topic):

[00:00:01.000 –> 00:00:10.000] Content about Taiwan policy by different speakers or at different times.

Example — All this should be ONE segment:

[00:00:01.000 –> 00:00:05.000] Prime Minister says Malaysia will adopt a whole of nation approach to address the tariffs.
[00:00:05.000 –> 00:00:10.000] Criminal elements and negligence are factors in the probe into the gas pipeline explosion.
[00:00:10.000 –> 00:00:15.000] Gas supply disruptions are expected to last until April 20th.

Timestamp Rules (Very Important):

- **(Refined)** For each segment in the final output:
  - The `start_time` **MUST** be the **earliest `start_time`** among all original lines included in that segment.
  - The `end_time` **MUST** be the **latest `end_time`** among all original lines included in that segment.
- All timestamps must come directly from the original input lines.
- **Time Validity Check:** Ensure that for every segment, the `start_time` is strictly earlier than its `end_time`.
- **Segment Ordering and Gaps:** Ensure segments in the final JSON are ordered chronologically by `start_time`. The next segment's `start_time` should be $\geq$ the preceding segment's `end_time`, possibly with a gap.

Validity Check (per segment): For every segment, analyze if the text is valid and meaningful. Check for:

1. Empty content or only background noises/music descriptions.
2. Only filler words or meaningless utterances.
3. Completely incoherent or nonsensical content.
4. Text length < 100, and the information lacks timely content, is outdated, or contains too little valid information.
5. A TV program outro, credits, music, or thank-you message that is distinct and lengthy.

If any of the above conditions are met, mark `"is_valid"` as `false`.
Format your response as JSON:

```
{
  "segments": [
    {
      "start_time": "Earliest start time from included lines",
      "end_time": "Latest end time from included lines",
      "content": "Full text content of segment",
      "is_valid": true/false
    }
    // ... more segments
  ]
}
```

The following is a timestamped transcript from a video. Each line follows this format:

[START_TIME –> END_TIME] content

Your task:

1. Segment this transcript into as **FEW** coherent segments as possible based on **topic content**.
2. Keep all content discussing the same topic/event together in **ONE** segment.
3. Look for shared keywords and semantic relationships between sentences to determine which should be merged.
4. For segments that appear to be program endings, thank-you messages, or credits (e.g., "Thanks for watching", "See you tomorrow", "This has been News at 9"), mark them with `"is_valid": false`.
5. For each segment, **strictly adhere** to all rules regarding timestamp generation, validity, and segment spacing outlined above.

Here's the transcript:

{formatted_text}

Respond strictly in the JSON format described above.

Prompt: Image Selection for Videos

Prompt Objective: You are an expert image analyst tasked with selecting images for a Question-Answering (QA) generation system. Your selections will be used to test a Large Language Model's

(LLM) visual understanding, so images with minimal textual clues are paramount.
Core Task: Evaluate **EACH** image provided in the current batch based on the Topic and Content Description below. Assign a score from 1 to 10 (10 is best) and provide a concise justification, focusing on its suitability for QA generation and the level of textual interference.
IMPORTANT SCORING GUIDANCE:

- Assign 8–10 to images that perform strongly on most criteria and do not have major flaws. Minor imperfections (e.g., small background text, mild quality issues, or faint watermarks/media logos) can still receive scores in the 7–9 range if overall relevance and informativeness are high.
- Images with some visual or contextual issues may still score 6–7 if they are otherwise useful for question generation.
- Only assign very low scores (1–3) to images that are blurry, of extremely poor quality, or have large overlaid text that clearly reveals answers or dominates the content.
- **News-style captions, watermarks, or channel graphics** are acceptable as long as they do not contain direct answers or overwhelm the main visual content.

General Advice: When in doubt, favor moderate to high scores for images that are clearly useful for QA purposes. Extreme scores (1 or 10) should be reserved for clearly unusable or exceptional cases.
Topic:
`"{topic}"`
Content Description:
`"{content}"`
Evaluation Criteria (Score each image from 1–10):

1. **High Content Relevance (Weight: High):**
   - MUST be strongly related to the Topic and Content Description.
   - Focus: Does the image offer rich visual context for generating insightful questions about the topic?
2. **Visual Clarity & Quality (Weight: High):**
   - MUST be clear, well-focused, and well-composed. Reject blurry or very low-quality images (assign score 1–2).
   - Focus: Are visual details easily discernible for LLM interpretation?
3. **Information Richness & Element Diversity (Weight: Medium-High):**
   - Prioritize images showing varied scenes, multiple relevant objects, interactions, or activities. Avoid overly simplistic or empty images.
   - Focus: Does the image provide multiple distinct visual elements or sub-topics for questioning?
4. **Minimal Textual Interference (Weight: CRITICAL — Low score for significant text):**
   - **CRITICAL:** Images with significant overlay text (captions, large logos, direct answers) that could "give away" information to the LLM should be scored very low (e.g., 1–3). The goal is to test visual understanding, not text reading.
   - Acceptable: Incidental background text (e.g., a distant street sign) is usually fine if not prominent or central to understanding the core content.

- Focus: Does the image primarily convey information visually, or does text play a major role that would simplify QA for an LLM? Less text is better.

5. **No Personal/Sensitive Identifiers (Weight: High — Reject if present):**
   - MUST NOT contain visible PII (names, faces of non-public figures unless anonymized/-consented), or private organizational details. Score 1 if present.
   - Focus: Is the image safe and appropriate for general use?

6. **Context over Sole Presenter (Weight: Medium):**
   - Avoid images SOLELY of a speaker/presenter unless their specific action/expression is key and described in the content. Prefer images with more contextual elements.
   - Focus: Does the image offer more than just a portrait?

Output Format (STRICTLY FOLLOW — Your entire response MUST be a single, valid JSON object as described below):
Your response must be a single JSON object. This object must contain one top-level key: `"image_evaluations"`. The value of `"image_evaluations"` must be a JSON array. Each element in this array must be a JSON object representing one image, with the following fields:

- `"image_number"`: (Integer) The 1-based index of the image as it was presented in the batch.
- `"score"`: (Float or Integer) The score assigned, from 1 to 10.
- `"justification"`: (String) A concise justification for the score, specifically mentioning relevance, visual quality, and especially the level/impact of any text.
- `"contains_problematic_text"`: (Boolean) true if the image contains significant overlay text, captions, or labels that could directly provide answers or make QA too easy; false otherwise.

Example of the EXACT JSON output format (for a batch of 1 image):

```
{"image_evaluations": [{
    "image_number": 1,
    "score": 8.5,
    "justification": "High relevance, excellent clarity. Minimal
    ↪  non-distracting background text.",
    "contains_problematic_text": false
}]}
```

Prompt: Image Caption Generation

You are an expert skilled at converting visual information into detailed linguistic descriptions. Based on the image below, generate a passage that thoroughly describes the content of the picture.
Original image caption (for reference): `{caption}`
Requirements:

1. The description must be detailed and specific, covering the main elements and scenes within the image.
2. The description must be written in **{language}**.
3. The description must strictly reflect the visual content actually shown in the image; hallucinated or fabricated details are strictly prohibited.

4. Do not use phrases such as "The image shows…". Provide the description directly.

You must answer in **{language}**.

Prompt: Level-1 QA Generation

(This prompt has multilingual versions; only the English version is shown here.)
You are an AI assistant skilled at generating high-quality "Level 1" multi-hop questions. Your task is to create questions based on "image and text." These questions should primarily test objective factual knowledge, rather than generalized reasoning or answers that come purely from describing the image. Based on the news article and the image below, generate one "Level 1" multi-hop question. This Q&A should test "social/factual knowledge," not something answerable only from the image description.
The question must involve specific social knowledge, such as a clearly identified person's role/identity by name, or the exact time and place of an event, rather than coarse information like simply recognizing objects or crowds.
The question must have a unique answer (the answer must be specific, unambiguous, and consistent with the article).
The image is provided only in the form of a textual image description. You should act as if you were given an actual image based on that description; any mention of "the image" or "the picture" below refers to a picture that shows what is depicted in the image description.
Article Title: `{title}`
Article Body: `{text}`
Image Caption: `{caption}` `{used_types_info}{used_questions_info}`
Required Requirements:

1. The question must begin with `"Based on this image"`.
2. The answer must be explicitly found in the article body.
3. The answer must be a short phrase or a few words (not a complete sentence).
4. The question type must be one and only one of the following categories: `location`, `person`, `organization`, `time`, `object`, `event`.
5. If asking about an event (`event`), the question stem may only ask: `"What event occurred?"`
6. For time and location questions, the stem **must** specify the required precision level, e.g., `"When did xxxx happen (precise to the day)?"` / `"Where did xxxx happen (precise to the city)?"`
7. For time and location questions, the ground-truth answer must be uniquely determined. For example, the answer cannot be `"Friday"` or `"September 28"` (because multiple years could match), but should be uniquely specific like `"September 28, 2025."`
8. **Factual Consistency:** The question stem and `Ground_Truth_List` must not contain factual or geographic common-sense errors (e.g., `"Luoyang, Hunan"`); if it conflicts with the article, the question is invalid.
9. During evaluation, the person answering will only see "image + question stem" and cannot access any article, body text, or image-caption text. Therefore, the stem must not contain words like `"article"`, `"text"`, or `"description"`, and must not reference the image description in any way.

10. The generated question must not be answerable using only the image description (i.e., not answerable from visual information alone). Deriving the answer must require social/factual knowledge.
11. Keep the stem concise and avoid leaking too many article details. The answer **must never** be obtainable from the stem alone. For example, `"Who is the Fed Chair speaking in the image?"` is unacceptable because `"Fed Chair"` gives away the answer; it should be changed to `"Who is the person speaking in the image?"` Therefore, entities referenced in the stem should be described as "the one in the image" rather than with detailed factual qualifiers.
12. Each question may contain only one interrogative sentence.
13. The ground-truth answer must be a definite fact from the article body, not an uncertain outcome (e.g., speculation, prediction).

Key Quality Constraints:

1. Do not ask "Who is he/she?" questions about extremely famous public figures (e.g., Donald Trump).
2. The answer must be specific and uniquely identifiable (e.g., `"Nike factory in Vietnam"`, not the generic `"factory"`).
3. Do not generate questions for images that lack temporal context (e.g., food close-ups, generic product photos).
4. Counting questions are strictly prohibited (e.g., `"How many people/objects are in the image. . . "`).
5. Avoid book-cover-type questions and overly generic answers such as `"dust jacket"`.
6. Do not ask questions whose answers can be directly read from visible text or logos in the image. (You may infer whether this applies from the image description.)
7. Location answers must be a specific place, not generic types like `"shopping mall/clothing store"`.
8. Event answers must be a specifically named event, not a generic event type like `"protest/fashion show"`.
9. Do not ask questions about visible chart/graph data in the image.
10. Person answers must be a specific full name, not an occupational label like `"police officer/doctor"`.
11. Time answers must be a specific absolute point in time, not a vague range like `"morning / 2020 / yesterday"`.

Additional Key Constraints:

12. The stem should be concise and should not reveal too much information from the article.
13. The stem should not include specific names, dates, or unique details from the article.
14. The stem must stand on its own when only the image is shown (the benchmark displays only the image and the stem).
15. Focus on the image caption, while ensuring the answer exists in the article body.

16. Avoid leaking the answer in the stem or providing too much context.
17. Very important: your question must be “significantly different” from other questions already generated for other images under the same topic.
18. Do not repeat questions about people, objects, or places that have already been asked.
19. You may not generate questions that can be answered solely from the image description without requiring social/factual knowledge, and you may not ask only about visible attributes (clothing color, number of people, etc.).
20. Avoid ambiguous or subjective questions.

Error Pattern Examples (avoid):

- `"Based on the provided image, who is the person speaking at the podium?"` → `"President Donald Trump"` (too obvious)
- `"Based on the provided image, what type of footwear is shown?"` → `"designer sneakers"` (too broad)
- `"Based on the provided image, what dish is being prepared?"` → `"pizza"` (food close-up, lacks context)
- `"Based on the provided image, how many protesters are visible?"` → `"24"` (counting question)
- `"Based on the provided image, what is shown on the book's cover?"` → `"dust jacket"` (generic answer)
- `"Based on the provided image, which company's logo is shown?"` → `"Google"` (visible text/logo)
- `"Based on the provided image, what type of factory is depicted?"` → `"garment factory"` (generic location)
- `"Based on the provided image, what event is taking place?"` → `"a protest"` (generic event)
- `"Based on the provided image, what does the chart show?"` → `"stock price increase"` (chart data)
- `"Based on the provided image, who is the person in uniform?"` → `"police officer"` (generic label)

Return strictly in the following JSON format:

```
{
  "question": "Based on this image, [provide your concise and clear
  ↪  question]?",
  "question_type": "[Category: location/person/organization/time/object/event
  ↪  - must be exactly one of these]",
  "Ground_Truth_List": ["[Ground-truth answer]", "[Equivalent phrasing 1]",
  ↪  "[Equivalent phrasing 2]"]
}
```

Formatting Notes:

1. `Ground_Truth_List` must include multiple acceptable answer phrasings (up to 10). The answer must be highly specific. For example, for a “Who is this person?” question, the answer must be the person’s name, not their title (e.g., `"Russia's foreign minister"`) or a description like `"the person wearing red."` For a “What event occurred?” question, the answer should include as much information as possible such as subject, time, and place (inferred from the news content and the image), rather than generic descriptions like `"a protest"` or `"postponing an election."`
2. If you cannot generate an appropriate question, return: `{"error": "Unable to generate an appropriate question"}`.
3. Ensure that all content is in English except the `question_type` field.

Respond strictly in the JSON format described above.

Prompt: QA Filter 1 (Language Check)

You are a language consistency checker for multilingual QAs. All text in these fields **SHOULD BE MAINLY** in `{language}`, with at least ∼80% of the words in `{language}`.

- Minor inclusions such as digits, punctuation, symbols, and a few foreign terms or abbreviations (e.g., AI, GDP) are tolerated.
- Option prefixes `"A."`, `"B."`, `"C."`, `"D."`, `"E."` should be ignored.
- If more than ∼20% of the text is clearly in another language → **DISCARD**.

Fields to check:

- **Question:** `{questions_text}`
- **Options:** `{options_text}`
- **Ground_Truth_List:** `{answers_str}`

RULES (RELAXED CONSTRAINTS):

1. Text must be mostly `{language}`, but small amounts of foreign words are allowed.
2. Ignore common proper nouns and domain-specific abbreviations.
3. If a field is empty or missing → **DISCARD**.
4. If highly uncertain → still accept (bias toward **YES**).

OUTPUT: Only output **YES** or **NO** to indicate whether the text is mainly in `{language}`. **DO NOT** explain anything.

Prompt: QA Filter 1 (Complete Check)

You are a specialized AI assistant responsible for evaluating and filtering news-related visual questions. Your expertise lies in identifying high-quality questions that require temporal and socio-cultural background knowledge. Carefully review each news article and its associated questions to determine whether they meet our strict quality standards.
YOUR TASK: For each question related to a news topic, evaluate it based on the strict quality standards below and identify which questions should be discarded. Note: Each question’s corre-

sponding image is provided in the form of a **CAPTION**. You must pretend you have the image itself, and all visual information you can get from the image is fully contained in the CAPTION's description.

EVALUATION CRITERIA — If a question meets any of the following conditions, it MUST BE DISCARDED:

1. TEMPORAL CONTEXT: Prefer questions that reflect a modern or time-sensitive context.

- For example, in food-related news, if the image is just a close-up of an object with no temporal context, such questions should be discarded.

2. VISUAL CLUE (non-OCR only): The answer needs visual information in the image (i.e., information from the CAPTION).

- If the question can be answered without looking at the CAPTION, discard it.
- If the question can be answered solely by reading visible text/signs described in the image (e.g., the CAPTION says there is a sign or subtitle showing some text in the image, and the answer can be obtained directly from the text), discard it.
- However:
  - Do **not** discard a question just because the CAPTION itself doesn't contain much information.
  - If the question requires combining the CAPTION with external news/context, this still counts as using visual understanding of the image described in the CAPTION, so the question is valid and should be **kept**.
- Bad Question Examples:
  - A stadium name is visible on a sign; the question asks for the name of the stadium.
  - A logo on a product shows the brand/company name; the question asks for the brand.
  - A name tag / caption shows a person's name; the question asks for that person's name.
- Good Question Example:
  - The CAPTION shows two politicians shaking hands, and the question asks which event this represents. (Even if the answer requires temporal/news background, understanding the scene visually is still necessary.)

3. AMBIGUOUS ANSWERS:

- If the Ground_Truth answer is too vague or ambiguous to be reliably determined, discard.
- Discard examples: `"Designer sneakers"`, `"high-end sneakers"` (vague categories).
- Good examples: `"Nike Air Force 1"` or `"Louis Vuitton Trainers"` (specific and identifiable).

4. SIMPLE COUNTING:

- Discard questions that only require counting visible objects (e.g., `"How many X are in the picture?"` expecting just a number).

5. BOOK COVERS:

- If the CAPTION describes the image as a book cover and the answer is just `"book cover"`, `"memoir cover"`, `"dust jacket"`, etc., discard the question.

---

6. VISIBLE TEXT / LOGO ANSWERS (HARD REJECTION — APPLY FIRST):

- If the correct answer string (or a clear abbreviation/variant) appears as visible text or a readable logo in the image described by the CAPTION, discard.
- This includes stadium/venue names on signs, city/country names on banners, organization names on podiums/backdrops, product/brand logos, lower-third name captions, etc.
- If the answer can be obtained simply by reading visible text/numbers described in the CAPTION (e.g., a year, label, subtitle, score, identifier), discard.
- Note: This does **not** mean that every question should be discarded whenever there is any text/logo in the CAPTION. If the text/logo does **not** reveal the answer or is just part of the context, that is **not** a reason to discard.
- Keep example:
  - `"Which person is standing under the banner that says 2020?"` → Keep.
- Discard examples:
  - The sign reads `"AMERICAN FAMILY FIELD"`, and the question asks `"What is the name of the stadium?"` → Discard.
  - The jersey clearly says `"LAKERS"`, and the question asks `"Which team is this?"` → Discard.
  - The image shows the number `"2020"`, and the question asks `"Which year is shown?"` → Discard.

---

7. GENERIC LOCATION / ESTABLISHMENT TYPES:

- Discard if the answer is just a generic type of place (e.g., `"textile factory"`, `"clothing factory"`, `"shopping mall"`, `"clothing store"`) without identifying a specific location.
- Good examples: `"Nike factory in Vietnam"` or `"Galeries Lafayette department store"` (specific identifiable locations).

---

8. GENERIC EVENT DESCRIPTIONS:

- Discard if the answer is just a generic event type (e.g., `"stunt show"`, `"protest"`, `"fashion show"`) without identifying a specific event.
- Good examples: `"Paris Fashion Week 2023"` or `"Black Lives Matter protest in Portland"` (specific identifiable events).

---

9. CHART DATA:

- Discard if the CAPTION describes the image as a scientific chart/graph and the question merely asks about data that is clearly displayed in the chart.

---

10. INCOMPLETE CONTENT:

- Discard if the topic has no question or no CAPTION.

11. GENERIC PERSON DESCRIPTIONS:

- Discard if the answer about a person is overly generic (e.g., `"police officer"`, `"protester"`, `"doctor"`) and does not identify a specific individual.
- Good examples: `"Emmanuel Macron"` or `"Taylor Swift"` (specific identifiable people).

12–17. Additional strict checks:

12. If the question contains multiple sub-questions, discard it.
13. If the content of the correct option is different from every item in the `ground_truth_list`, discard it.
14. Discard it if the answer is not uniquely determined. For example, answers like `"Friday"` or `"September 28"` should be discarded (because multiple dates across different years could match).
15. Discard it if the question stem includes `"in the text/description"` or if it mentions an `"image description"`.
16. Discard it if the answer can be obtained solely from the question stem (e.g., `"Who is the Fed Chair speaking in the image?"`).
17. Discard it if the answer does not address the question. For example, if the question asks `"What event happened?"` but the answer is not an event, discard it.

INSTRUCTIONS:

1. Analyze the given news article, its associated image (via CAPTION), and each related question.
2. For each question, decide whether it violates any of the criteria above.
3. Return a JSON-formatted response specifying:
   - Which questions should be discarded and why;
   - Which questions can be kept.
4. Apply the criteria very strictly — when in doubt, choose to discard.

Please evaluate this news topic and its associated questions:
TOPIC ID: `{topic_item.get('id', 'Unknown')}`
IMAGE CAPTIONS ({len(img_paths)} total):
IMAGE {i}: `{img_path}` **CAPTION:** `{caption}`

Prompt: Level-2 QA Generation

(This prompt has multilingual versions; only the English version is shown here.)
You are an AI assistant skilled at generating high-quality, deliberately adversarial Level-2 multi-hop question answering items. Your task: based on a given "news article + image description + Level-1 QA," generate 3–5 Level-2 questions to stress-test a multimodal large language model's (MLLM's) visual social knowledge and reasoning.

Image and input setup (must follow)

- The image is provided only as **image-caption text**. You must assume a real image exists and that its content matches the caption exactly; any mention of "the image/photo/in the picture" in the question refers to this real image.

Test setup (must follow)

1. In the final test, the model can only see: **the image + the question stem**.
2. The model cannot see: the article title/body, the image-caption text, the Level-1 QA, or the answer.
3. Therefore, the question stem must **not** contain or imply words like `"in the text/body/article/description/ below/above"`, and must not mention `"image caption"`.
4. **Do not** rely on reading **any text/slogans/logos/on-screen text** in the image to answer (even if the image might contain text, it cannot be used as a key clue).
5. The question stem and `Ground_Truth_List` **must not contain factual or geographic common-sense errors**; if they conflict with the article body, the question is invalid.

How to reference the image (required; must be neutral)

- The question stem must include the phrase **"In the image"**.
- The **"In the image, . . . "** clause must be neutral and contain **no proper nouns** (no person/-place/organization/event names, no specific dates, etc.). Only abstract references are allowed, e.g.:
  "`In the image, the main subject / another person / the scene / the background building / the location / the ongoing occasion . . .` "
- Proper nouns may appear outside the "In the image, . . . " clause, but you must still satisfy "no leakage" (see below) and must not let the answer be deducible from the stem alone.
- **Do not describe specific visual details** (e.g., tents, national flags, colors, clothing details, architectural style, number of people, prominent text, etc.).
- No subjective/speculative wording is allowed, such as "`seems/might/probably/obviously/ suggests/implies/solemn/crisis`", etc.

Inputs

- Article URL: `{url}`
- Article title: `{title}`
- Article body: `{text[:5000]}`
- Image caption: `{caption}`
- Level-1 question: `{level1_question}`
- Level-1 question type: `{level1_type}`
- Level-1 answer: `{level1_answer_text}` `{excluded_types_text}`
- Output language: `{language}`

Task: Generate 3–5 Level-2 questions (deliberately adversarial)

Each question must satisfy all of the following:

1. **Reliance on the Level-1 answer:** Solving must require the Level-1 answer, but the question stem must not mention or imply the Level-1 question, the Level-1 answer, or any equivalent expression, and must not contain any wording related to `"Level-1"`.
2. **Multi-hop reasoning:** At least 2–3 steps of reasoning are required (e.g., use the image to pin down the object/context $\rightarrow$ combine with the Level-1 answer $\rightarrow$ locate the unique fact in the article body).
3. **The image is a necessary input:** External common knowledge may be used, but the question must not be uniquely answerable from the stem text alone; the image must play a key role in disambiguating/locking onto the target entity.
4. **Closed-loop and definite answer:** The gold answer must be a definite, objective fact stated in the article body (not speculation/prediction/opinion).
5. **The answer must be short:** The answer should be a phrase or a few words (not a complete sentence).
6. **No leakage in the stem:** Do not include unique details from the article body in the stem, such as specific names, specific dates, or exact percentages/vote counts; avoid making the answer inferable from the stem alone.
7. **Concise stem:** Aim for ≤ 35 Chinese characters / ≤ 25 English words; do not pad background with `"given/based on/in the context of/taking into account. . . "`; lists of same-type entities must not exceed 2.
8. **Single question:** Only one question sentence is allowed; it must be an interrogative (must include a question mark or a clear interrogative structure) and must not be an instruction (e.g., `"Answer. . . "`).
9. **Time expressions:** Do not use relative time (e.g., `"past 24 hours/recently/yesterday/this week"`); if it is a time-type question, the stem must state the required precision (e.g., `"to the year/month/day/minute"`).
10. **Not answerable from image description alone:** The question must not degrade into pure visual recognition/pure image description; do not ask about text/logos that can be directly read from the image; do not ask about data visibly shown in charts/graphs.
11. **Do not mention "reason":** The stem must not contain `"reason"` or related phrasing, because questions about reasons carry subjective elements.

Difficulty-raising strategies (deliberately increase difficulty)

For each question, you may use the following strategies to raise difficulty, without using tongue-twisters or flashy phrasing:

1. **Fine-grained exclusion / disambiguation:** among similar candidates, use the image anchor plus article facts to eliminate wrong options and lock a unique answer.
2. **Time / causality / ordering:** build multi-hop reasoning using temporal order, causality, or process relations (while keeping the stem concise).
3. **Implicit clue linkage:** provide only a neutral anchor in the stem, forcing the solver to link "the object/occasion in the image" with the Level-1 answer to a specific fact in the article.

4. **Precise distinction among near-synonyms:** distinguish very similar organization/event/place concepts precisely; the answer must be unique and verifiable.

Language requirement: the stem must be natural and direct; difficulty should come from the reasoning structure and disambiguation, not from complex syntax, stacked context, or language tricks.

Key quality constraints

- Do not ask "Who is he/she?" about extremely famous public figures (e.g., globally well-known political leaders); avoid overly obvious identity-recognition items.
- Avoid questions tied to images lacking time/event context (e.g., food close-ups, generic product photos).
- Location/event/person answers must be specific and uniquely identifiable:
  - Location cannot be a generic type like `"a shopping mall/clothing store"`; it must be a specific place-name level.
  - Event must be a specifically named event, not a generic description like `"a protest/fashion show."`
  - Person must be a specific full name, not an occupation label like `"a police officer/doctor."`
- Avoid vague or subjective questions; avoid hypothetical premises like `"If . . . then . . . "`.
- Generated questions should differ substantially from other questions under the same theme; do not repeatedly ask about entities already asked (person/place/organization/object).

question_type constraints (strict)

- Each question must provide `question_type`, and it can only be one of:
  `["location","person", "organization","time","event", "count"]`
- Across the 3–5 questions, all `question_type` values must be **pairwise distinct** (no repeats).

Answer format requirements by type (strict)

- **location:** a uniquely identifiable **absolute place name** (no `"nearby/northern/within"`); the stem must declare the precision level (e.g., `"to the exact city"`).
- **person: full name (given name + surname)**, uniquely mapped; do not replace the name with a title/occupation.
- **organization:** official full name or commonly used abbreviation; if it appears for the first time, provide the full name (optionally with abbreviation).
- **time:** must be an **absolute, uniquely determined time expression**, consistent with the precision declared in the stem; no relative time.
  - If neither the article body nor the image caption provides any "uniquely determined" absolute time information, do not generate time questions (sports may be an exception allowing "the Xth minute").
  - For time-type questions that ask about a specific point in time, the standard answer must include a coarse-grained time component to ensure the answer is uniquely determined (e.g., `"YYYY year M month D day . . . "`). If the coarse-grained time information cannot be obtained from the article body, then this time question should not be created.

- **event:** must be a **unique, specifically named event** (write the full event name).
- **count:** an exact Arabic numeral (no ranges/words). Count must not be direct counting of salient objects/people in the image.

Ground truth rules

1. First determine the single correct answer (a definite fact from the article body), then provide `Ground_Truth_List` ($\leq$ 10 items).
2. Every item in `Ground_Truth_List` must match the answer type (all persons are names, all locations are same-level place names, times share the same precision, etc.).
3. The answer must be specific (avoid generic references like `"CEO/official residence/building/researcher /microphone"`).

Self-check (for verification only; must not appear in the final JSON)

After generating each question, write a step-by-step reasoning chain to self-check:

- If the image is hidden and only the stem is given: can it still be uniquely answered? If yes → invalid, must rewrite.
- Does the "In the image, . . . " clause contain any proper noun? If yes → invalid.
- Does it include any banned words (in the text/body/description/implies/seems/might/past 24 hours, etc.)? If yes → invalid.
- Does it rely on image text/logo/chart-visible data? If yes → invalid.
- Does it include hypothetical premises (`"if/suppose/if . . . then . . . "`) or subjective judgment? If yes → invalid.
- Does `question_type` repeat another question's type? If yes → invalid.
- If it is a time question: do the article body / image caption provide "uniquely determined" absolute time info, and does the answer granularity match the stem? If no → invalid.

Final output (the only allowed format; do not add any other explanation)

```
{
  "level2_qas": [
    {
      "question": "...?",
      "question_type": "location|person|organization|time|event|count",
      "Ground_Truth_List": ["Correct answer", "Equivalent 1", "Equivalent 2"],
      "reasoning": "[Detailed reasoning process: Start with “The correct
       ↪  answer is [correct answer text]. The source is [article title]
       ↪  [article URL]”. You need to explain that the answer must first be
       ↪  found by locating this source through the image, then obtaining the
       ↪  news/video transcript from that source, and finally explaining step
       ↪  by step how the transcript and the image lead to the correct
       ↪  answer.]"
    }
  ]
}
```

Generate 3–5 questions; if you cannot generate any qualified questions, output:

```
{ "level2_qas": [] }
```

Language requirement: `question` and `Ground_Truth_List` must be written in **{language}**.

Prompt: QA Filter 2 (Self-answering)

You are a **STRICT** language consistency checker for multilingual QAs. Your **ONLY TASK** is to verify that the following fields are written in pure `{language}`:

- `"Question"`: `{question}`
- `"Options"`: `{options_content}`
- `"Ground_Truth_List"`: `{answers_str}`

RULES (HARD CONSTRAINTS):

1. All text in these fields **MUST** be in pure `{language}`.
   - Ignore the option prefixes `"A."`, `"B."`, `"C."`, `"D."`, `"E."` (they are neutral labels).
   - If even **ONE** word is in another language → **DISCARD**.
2. Partial compliance, code-switching, transliteration, or mixing with English or other languages = **STRICT VIOLATION**.
3. Globally recognized proper nouns (e.g., `"UN"`, `"NASA"`, `"COVID-19"`, `"Paris"`, `"Macron"`) **ARE ALLOWED** and do not count as violations.
4. If any field is empty or missing → **DISCARD**.
5. When uncertain → **DISCARD** (default to stricter filtering).

OUTPUT: You only need to output **YES** or **NO** to indicate whether the text in the fields is written in pure `{language}`. **DO NOT** explain anything.

Prompt: QA Filter 2 (Self-Answering)

Please answer the following multiple-choice question based on the provided image and text context.
Text: `{text}`
Question: `{question}`
Image Description: `{caption}`
Options: `{options_text}`
Please provide only the letter of your answer (`A`, `B`, `C`, `D`, or `E`). Do not provide any explanation.

Prompt: Final Filter

You are given a single-answer multiple-choice QA:

```
"question": "{question}",
"Ground_Truth_List": {gt_list_json},
"Options": {options_text}
```

Your task:

1. Check if the **question text** or the **Ground_Truth** itself contains factual errors (e.g., "Hunan Luoyang" is invalid because Luoyang is not in Hunan Province).
   - Ignore factual correctness of **non-correct options**: they can be wrong, misleading, or even fictional.
2. Check if the Ground_Truth is **specific enough** to uniquely identify the correct option.
3. Ensure the Ground_Truth exactly matches **one and only one option** semantically.

Additional strict checks:

4. **QA-centered uniqueness:** The QA must have a **single, uniquely correct** answer.
    - The Ground_Truth_List may contain multiple **variants** of the same core answer (up to 10).
    - Variants are acceptable if they all **effectively answer the question** and **unambiguously converge on the same core entity/event/time**.
    - Variants may include synonyms, abbreviations, expanded or elaborated forms, sub-units of the same organization, or commonly used aliases (e.g., "RSF" and "Hemedti's forces").
    - If the Ground_Truth_List contains **entries pointing to genuinely different answers** (different entities, different reasons, or different times), it is invalid.
    - For **person-type questions**, all entries must be the person's proper name. Titles, roles, or vague references (e.g., "the singer," "the person in red," "Foreign Minister of Russia") are not acceptable.
    - **Special case for dates:** If the calendar system is not explicitly specified, treat different formats (e.g., "August 15") as equivalent. Only when the calendar system is explicitly stated (Gregorian, Lunar, Islamic, etc.) should they be treated as different.

5. **Person-name rule:** For person-type questions, the Ground_Truth_List must include the person's proper name.
    - Acceptable: Variants that contain the proper name plus additional titles/roles (e.g., "President Emmanuel Macron", "Ministro do STF Alexandre de Moraes").
    - Invalid: Entries that only include titles/roles without the name (e.g., "Foreign Minister", "a police officer", "the coach").
    - Invalid: Vague references (e.g., "the singer", "the person in red").

6. **Event specificity:** For event questions, reject **vague/generic** Ground_Truth (e.g., "a protest", "a ceremony") if it cannot uniquely match one option. Require named/specific events.

7. **Time absoluteness:** For time questions, the Ground_Truth must be an **absolute time expression**, with the minimum granularity determined by the question wording:
    - If the question explicitly asks for "which year" → year alone is acceptable (e.g., "2024").
    - If the question does not limit to "year" only → the Ground_Truth must include at least **year+month** (e.g., "2024-08") **or month+day** (e.g., "August 15") if the year is clearly implied.
    - If the question explicitly asks for "which day/date" → Ground_Truth must include a full date, preferably **year+month+day** (e.g., "2024-08-15"); but **month+day** alone is acceptable if the year is obvious from context.
    - **Exception:** in sports/match contexts, relative time markers (e.g., "minute 31", "stoppage time first half") are valid.
    - **7.a Strict ban on relative phases (non-sports):** Reject Ground_Truth (and matching option) if it uses only relative/phase expressions without an absolute date or date+time, e.g., "evening", "night", "later that night", "this morning/afternoon", "before/after the storm", "at the peak intensity", "as the waters receded", "when a second round was forecasted". These are **invalid** unless accompanied by an absolute timestamp (e.g., "2024-06 evening" is still invalid; use "2024-06-12 20:00").
    - **7.b Valid examples:** Year only: "2024"; Year+Month: "2024-06"; Month+Day (year implied): "August 15"; Year+Month+Day: "2024-06-12".

- **Invalid examples (non-sports):** “evening hours”, “later that night”, “during the peak”, “after receding”, “before landfall”.

8. **Option set sanity:** There must be exactly one option semantically matching the Ground_Truth.
   - Non-correct options do **not** need to be factually plausible; they may be wrong or invented.
   - Validity only depends on whether Ground_Truth uniquely matches one option.

9. **Question style rule:**
   - The question must be phrased as a **standalone natural QA**, not as a meta-question about options.
   - Strictly forbid meta-phrases such as:
     - English: “which of the following”, “choose from the options below”, “select all that apply”
   - The question must stand alone naturally without referencing options.
   - You are creating a QA-style question; the multiple-choice options are added **afterward** only as distractors and the correct answer.

10. **Social-knowledge specificity filter:** The question must involve concrete social knowledge rather than simple visual recognition.
    - **Person-type:** Must identify a specific individual by proper name. Variants may include name + title (e.g., “President Emmanuel Macron”) but **do not** accept entries that are only a title/role or vague descriptions (e.g., “a police officer”, “the singer”, “the person in red”).
    - **Event-type:** Must be a **named, specific** event (e.g., “2025 Bilibili Top 100 UP Awards”), not a generic type (“a ceremony”, “a protest”, “a press conference”).
    - **Location-type:** Must be a **specific, identifiable place** (venue/city/institution with a proper name), not generic types (“a mall”, “a park”, “a factory”).
    - **Organization-type:** Must be a **named organization/unit** (e.g., “Kowloon City Police District”), not generic groups (“the police”, “the committee”).
    - **Object-type:** Prefer **specific models/brands** when the benchmark context requires it (e.g., “Neumann U87 studio microphone”), not broad categories (“a microphone”), unless the article/image only supports that level.
    - Any QA whose Ground_Truth refers **only** to generic groups/crowds/roles, vague activities, or broad categories **must be discarded** as lacking social knowledge specificity.

Decision procedure:

A. Normalize Ground_Truth entries (handle synonyms, surface variations).

B. Check if they converge to one unique core answer.

C. Verify exactly one option matches that answer.

D. Enforce social-knowledge specificity (Rule 10): reject QAs whose Ground_Truth is generic (group/role/type) and does not pinpoint a concrete person/event/location/organization/object as required by the question type.

Output format (JSON only):

```
{
  "valid": true/false,
  "reason": "Explain clearly why this QA pair is valid or invalid. If invalid,
  specify which rule(s) failed, e.g., factual error, non-unique answer, vague
  event, role-not-name, non-absolute time.",
  "question": "...",
  "Ground_Truth_List": "...",
  "Options": "..."
}
```

Prompt: Automate Crawler Agent

Task Objective

You need to write a crawler for a specified website to scrape news content and related images. The specific URL will be provided later. You need to debug and verify the results after the crawling is completed.
Remember: Add only one website's crawler at a time, and stop working after completion.

Development Process

1. **Get the homepage source code of the target website:**
   - You can obtain the homepage source code of the target website through online search or using `tools/fetch_html.py`. If access is not possible, please output relevant error information and stop all subsequent programs.
2. **Analyze homepage structure and generate configuration:**
   - Refer to the existing format in `config.py`.
   - Extract the website's section information from HTML as the `SECTION` field.
3. **Update `config.py` and `config_sections.py`:**
   - First, thoroughly read `config.py` and `config_sections.py` to understand where configurations are stored.
   - Imitate other website configurations, add a new `SECTION` configuration for this website in the `SECTION` configuration part of `config_sections.py`.
   - At the bottom of `config.py`, set `ENABLE_XXX = False` (don't comment this out) and `ENABLE_XXX = True` (comment this out by default) respectively (replace `XXX` with the website's abbreviation).
   - Comment out other websites' `ENABLE_xxx = True` at the bottom of `config.py`, and uncomment the `ENABLE_XXX = True` configuration for the currently configured website.
4. **Analyze news link structure:**
   - Read the HTML source code of any news article through online search or `tools/fetch_html.py`.
   - Analyze the image storage format and time information storage format.
   - Read `utils_xxx` to determine whether existing tools can correctly extract the website's images and publication time information.
   - If not extractable correctly, make corresponding changes to `utils_xxx`.
5. **Write crawler script:**

- Create a new file in the `collectors/xxx` directory, where `xxx` is the language of the website.
- Implement `scrape_xxx_news` and other functions, ensuring consistency with existing crawler styles.
- Follow the function implementation of `folha_collector.py`, simplifying code by calling integrated functions in `utils_generic_scraper.py`.
- Do not refer to other crawlers in the same language.
- **Requirements:**
  - Debug output and comments **must only use English**.
  - Output format should be completely consistent with existing crawlers.
  - Add necessary comments, ensuring consistency with current style.
  - Avoid introducing any new utility functions.
  - If there is no definitive information indicating pagination functionality exists, do not add it.

6. **Modify `logging_manager.py`, `task_manager.py`, `file_manager`:**
   - Search for "BBC" and completely imitate its related configuration for the new website configuration.
   - Adapt according to the new website to ensure logging and scheduling functions work properly.

7. **Image processing:**
   - If image scraping fails, handle the issue in the `utils_xx` module first.
   - Avoid adding special image processing logic in the crawler script.

8. **Check the above steps.** Ensure the following files have been modified:
   - `config.py`
   - `config_sections.py`
   - `task_manager.py`
   - `logging_manager.py`
   - `file_manager.py`
   - `xxx_collector.py`

Testing Process

1. In `config.py`, make sure that at the bottom of the file, only the newly added website's `ENABLE_XXX = True` is active, and comment out other `ENABLE = True` configurations.

2. **You must always use the following command to run the program:** `python run.py`

3. **Check terminal output** to confirm whether titles, links, timestamps, and images were successfully scraped.
   - Pay attention to whether timestamps are `null`, links are invalid, or images are missing.

4. If problems occur:
   - **Add debug output**, re-run and analyze the reasons for failure.
   - You can choose a random news page and check why problems occur.
   - After successful debugging, comment out or delete all unnecessary debug output added.

5. If encountering 403 problems:
   - First try adding request headers to avoid this issue.
   - If it fails, revert all changes and report that the website has unresolvable anti-crawling measures.
5. **Adjust `utils_xx`** (if needed), then re-run.
6. Verify `hot_topics` file:
   - Check whether timestamps and other data in the file are correct.
   - Randomly access several URLs to check if timestamps are extracted accurately.
7. After successful testing:
   - Add a `done` mark for the website in `websites.md`.
8. Task completion indicator:
   - Ensure news articles, images, and publication dates are all successfully scraped.
   - Delete all debug output to ensure code is clean and tidy.

Note: When testing, add a runtime limit of **30–60 seconds**, then terminate it and check terminal output to avoid excessively long execution times.

Important Notes (Strict)

- Please be sure to mark as "done" after confirming successful crawler implementation.
- Must run all process instructions step by step.
- Do not try to create terminal tasks/processes; run code directly in the terminal.
- Limit maximum runtime (generally 30–60 seconds) for any run.
- **No new files allowed** — only existing files can be modified.
- **No test scripts allowed** — only use `test1.py` and `test_utils_images.py` when errors occur; otherwise use `run.py` (with time limit).
- All new code must include comments and follow existing comment styles.
- Any code containing "BBC" must be copied and adapted for new sites.
- Debug output and logs must use English.
- If needed, you can search online or use `tools/fetch_html.py` to understand structure or anti-crawling measures.
- Maintain consistency and maintainability; avoid temporary/workaround solutions.
- Ensure commands can run automatically without manual intervention.
- Always use `python run.py` when testing.
- All websites are in `websites.md`; skip those marked "done". Implement one website at a time; add "done" afterward.
- When naming images, use the website's full name as prefix to avoid abbreviation conflicts.
- If unsolvable, revert all changes and report the problems encountered.
- After completing all modifications, revert ENABLE settings for other websites in `config.py` to original versions.

Prompt: Title Distractor Generation

You are an AI assistant specialized in generating academic paper title distractors.
Given the real title of a research paper, create four alternative titles that sound plausible but are clearly different from the original.
These should be believable as academic paper titles in a similar field, but not actual existing papers.

Original Title: {title}

Your response should be in JSON format:

```
{
  "distractors": [
    "First title",
    "Second title",
    "Third title",
    "Fourth title"
  ]
}
```

Prompt: Paper Level-1 QA Reasoning Generation

You are an AI assistant specialized in academic papers.
Given the information about a research paper, create a detailed explanation for a question-answer pair.

Paper Title: {title}
Paper Abstract: {abstract}
Question: {question}
Correct Answer: {answer}

Your response should:

1. First clearly state the correct answer.
2. Then provide a brief summary of the paper based on its abstract.
3. Maximum length: 100 words.
4. Format: "`The correct answer is [correct answer]. [Paper summary]`"

Your response:

Prompt: Paper Summary Generation

You are an AI assistant specialized in academic papers.
Given the information about a research paper, create a concise summary.

Paper Title: {title}
Paper Abstract: {abstract}

Your response should:

1. Provide a brief summary of the paper based on its abstract.
2. Maximum length: 80 words.
3. Format: clear, concise summary in 1–2 sentences.

Your response:

Prompt: Paper Level-2 QA Generation

You are an AI tasked with generating multiple-choice questions. Your goal is to create questions that appear to be based solely on an image from a scientific paper.

I will provide you with the full textual content related to this image, including the paper's title, abstract, and the caption of the image:

{content_for_qa}

You will use this information to craft your questions and answers. However, your generated questions and explanations must be framed as if the end-user was only initially provided with the image itself and no other information.

Please generate 1–2 multiple-choice questions. For each question, adhere to these specific instructions:

1. **Challenge:**
   - Craft a question that pushes the limits of search and reasoning for both humans and AI.
   - It should not be answerable by a simple keyword lookup; it must require careful reading and inference.
2. **Focus on Text Details Only:**
   - Target one specific, simple detail from the provided text.
   - Do NOT mention or describe any visual content from the image.
   - Refer to the paper's methods abstractly (e.g., "the method described in the paper").
   - Each question begins with the following sentence: "`In the paper corresponding to this image,`"
3. **Questions Format:**
   - You must ensure that the question can be both a multiple-choice question and an open-ended question, and that the answer is appropriate for an open-ended question.
   - For example, you cannot ask which option is correct or which of the following meets the requirements, because this violates the requirements of an open-ended question.
4. **Answer Types (choose exactly one):**
   a. A specific data value (e.g., a number or percentage) from the paper. `[question_type: data]`
   b. A precise time mentioned in the paper. `[question_type: time]`
      - When asking these types of questions, make sure you provide the time precision that matches the answer in the question (e.g., which year/which month/. . .).
   c. An objective research result statement that appears verbatim in the text. `[question_type: conclusion]`
5. **Uniqueness & Definiteness:**
   - The answer must be unambiguous and unique in the text.
6. **Non-Visual:**

- The correct answer cannot be derived from any visual element; it must depend on the text.

7. **Self-Contained Answerability:**
   - The question must be answerable using only the given abstract or context, without external knowledge.

8. **Note:**
   - When asking a question with the `question_type` set to `time`, make sure you provide the time precision that matches the answer in the question (e.g., which year/which month/...).

For each question, provide the following:

- A clear, concise question text.
- Five options (labeled A through E).
- The correct answer's letter (this letter should be randomly chosen from A–E for each question).
- A list containing the correct answer phrased in one or more ways (e.g., `["The primary finding was X.", "X was identified as the main result."]`).
- Detailed reasoning process to get the correct answer. MUST NOT mention about other options, they are not needed.

Format your entire response as a single JSON object. Do not include any markdown formatting or any text outside of this JSON object.

```
{
  "level2_qas": [
    {
      "question": "[Your question text here]",
      "question_type": "data/time/conclusion",
      "options": [
        "A. [Option A text]",
        "B. [Option B text]",
        "C. [Option C text]",
        "D. [Option D text]",
        "E. [Option E text]"
      ],
      "Ground_Truth": "[Correct letter]",
      "Ground_Truth_List": ["[The correct answer phrased as in the text]",
      ↪  "[An alternative phrasing of the correct answer]"],
      "reasoning": "[Detailed reasoning process: Start with 'The correct
      ↪  answer is [correct answer string]. The source paper is [the paper]'.
      ↪  Explain step-by-step how the correct answer is derived from the
      ↪  specific details within the provided abstract or contextual
      ↪  information of that identified paper. This reasoning should not
      ↪  suggest the answer comes directly from the abstract or context you
      ↪  were given but rather from the text *of the paper found via the
      ↪  image*]"
```

```
    },
    { ... more questions in the same format ... }
  ]
}
```

Prompt: News Caption Formatting

You are a text formatting assistant. Your task is to clean and format image captions by:

1. Removing special characters, symbols, and formatting artifacts
2. Fixing broken words and spacing issues
3. Ensuring proper capitalization and punctuation
4. Maintaining the original meaning and technical terms
5. Keeping the text concise and readable
6. Removing line break symbol, formatting symbol, such as `\n`, `\r`
7. If there exist a formula, please convert it into plain text form by inferring the original format

Return only one line of the processed caption text in plain text format, without any additional commentary or instructions.

Prompt: Paper Image Selection

Objective:
You are tasked with analyzing the provided **paper abstract** and **image captions** to **identify, score, and rank** the figures that best serve as memorable **"paper identifiers"**. A **"paper identifier" figure** is one that **represents the unique contributions** and **core concepts** of this paper, making it visually distinct and memorable based solely on its caption and content.

Inputs You Will Receive:

1. **Paper Abstract:** A concise summary of the paper's **research**, **methodology**, and **findings**.
   - Only used for inferring the theme of the article, **DO NOT** use it for inferring the content of image!!!
2. **Image Captions:** A list of **captions** for each figure in the paper.
   - If an image does not have a caption, DO NOT attempt to complete it through speculation.

Your Task:

1. **Understand the Core Contributions:**
   - **Read the Abstract carefully** to identify:
     - The **primary contributions** of the paper.
     - The **methodologies** used.
     - Specific **datasets**, **key theoretical concepts**, and **distinct results**.
   - The goal is to determine **what makes this paper stand out** among others.
2. **Evaluate Each Image Caption for Memorability and Uniqueness:**

- **Analyze each image caption** for:
  - **Distinctiveness** and **memorable qualities**.
  - Alignment with the **core contributions** of the paper and its **unique concepts**.
- **Consider the overall contribution of the figure** to the paper's identity.

3. **Score and Rank All Figures:**

   - **For each image:**
     (a) Infer the image type based ONLY on its caption, without abstract.
       - **If an image has an empty caption, NEVER infer its type and leave the caption blank!**

     (b) **Assign a score from 0 to 10** based on the following criteria.
       a) These are some things you **MUST** ensure:
         - **If an image has no caption**, you **MUST** assign it 0 (Zero) score.
         - **If the caption contains parameters, numbers, formulas, or equations**, you **MUST** assign it a **low score** (this applies to data analysis images, such as graphs or statistical visualizations).
         - **If you are unsure** about the image type, you **MUST** give it a **low score**.
         - **If a picture contains multiple sub-pictures**, select the **one that best meets the criteria**.

       b) **Low Scores:** Assign low scores to images of the following types (or those likely to be such):
         - Snapshots.
         - Data visualizations (e.g., bar charts, line plots, scatter plots, heatmaps).
         - Statistical charts, maps, coordinate images, or distribution maps.
         - Generic images of **people, animals, or objects** that are not unique to the paper.
         - Images with heavy **text content** (e.g., summaries, conclusions, limitations).
         - **Tables**, **equations**, or **algorithm boxes/pseudocode** presented as images.

       c) **High Scores:** Assign higher scores to images that meet the following descriptions **but do not meet the low-score criteria**:
         - **Flowcharts** → **strongly recommended**.
         - **Block diagrams** → **strongly recommended**.
         - **Principle images** → **strongly recommended**.
         - **Framework/architectural diagrams**.
         - **Highly distinctive scientific illustrations**.

       d) **For each figure**, provide a **brief reason** for its score, addressing:
         - **Memorability**.
         - **Uniqueness**.
         - **Connection to the core novelties** of the paper.
         - NOTE: **If an image has no caption, NEVER infer its type and leave the caption blank!**

4. **Ranking Figures:**

   - **Rank all figures** from highest to lowest score.
   - **Ensure that all images are sorted in descending order according to their scores**.
   - **Tie-breakers:** If multiple figures share the same score, apply the following tie-breakers:
     (a) **Closer alignment** to the paper's **core novelty**.
     (b) **Higher conceptual uniqueness** (e.g., preferred framework/concept diagrams).
     (c) **Broader memorability**, independent of the paper's text.

5. **Recommend the Selection Count (0–3):**
   - Recommend **0 to 3 figures** to select based on the following:
     - Only recommend figures with a **score** $>$ **8** (i.e., 9–10).
     - **Never** recommend any figure with a score $\leq$ **8**.
     - **If more than 3 figures** score $>$ 8, select the **top 3 by score**, applying tie-breakers if necessary.
     - **If 0–2 figures** score $>$ 8, select exactly that many.
     - If **no figure** scores $>$ 8 (e.g., all figures are weak or statistical visualizations), recommend **0** figures.
6. **Explain the Recommendation for Count:**
   - Provide a **concise explanation** of why you recommend selecting the specific number of figures, based on:
     - The **score threshold** ($>$ 8).
     - The **overall quality distribution** of figures.
     - How the top candidates meet the **identifier criteria** (distinctiveness, core contributions, etc.).

Output Format (must match exactly):
Return a **JSON object** containing the following:

```
{
  "ranking": [
    {
      "index": figure_index_start_from_1,
      "reason": "Score: (x/10), Type: (xxx). (Brief explanation of why this
      ↪  caption content ranks here.)"
    }
  ],
  "recommended_count": number_of_figures_to_select,
  "selection_reason": "(Brief explanation of why you recommend selecting this
  ↪  many figures based on caption analysis and the >8 score threshold.)"
}
```

Important Notes (must be strictly followed):

- The `"ranking"` array must contain **ALL figures** from the input JSON file, sorted by their scores (highest first).
- The number of figures in the `"ranking"` array must match exactly the number of figures in the input JSON file.
- Be brave to give low scores to figures that do not qualify as memorable identifiers, even if 0 images are selected.

Case Study

Question: استنادًا إلى هذه الصورة، ما هي الجمعية التي يقدم البنك الدعم لها في هذا الإطار؟
Based on this image, which association is the bank supporting in this context?

Ground Truth: جمعية الأشخاص ذوي الإعاقة بالأحساء
Al-Ahsa Association for Persons with Disabilities

Level: 1
Question Type: Organization

Model Answer

Gemini3 Pro: الجمعية التي يقدم البنك الدعم لها في هذه الصورة هي جمعية الأطفال ذوي الإعاقة.
The association the bank is supporting in this image is the Disabled Children's Association.

GPT-5.2:online: الجمعية التي يقدّم لها البنك الدعم في هذا الإطار هي جمعية «كفو للتوظيف والتأهيل».
The association the bank is supporting in this context is the "Kafoo for Employment and Rehabilitation" Association.

Claude Sonnet 4.5: لا يمكنني تحديد اسم الجمعية المحددة التي يقدم لها البنك الدعم بناءً على الصورة المعروضة.
I cannot determine the specific name of the association the bank is supporting based on the displayed image.

Figure 13: Case study for Arabic & Level 1 Question.

Case Study

Question: في الصورة، نقاش حول الخطة الاستثمارية. ما هو العام المالي المحدد لذلك؟
In the image, there is a discussion about the investment plan. What is the specified fiscal year for it?

Ground Truth: 2025-2026

Level: 2
Question Type: Time

Model Answer

Gemini3 Pro: العام المالي المرتبط بأحدث وأبرز هذه النقاشات الرسمية هو العام المالي 2023 / 2024.
The fiscal year associated with the latest and most prominent of these official discussions is 2023/2024.

GPT-5.2: العام المالي الخاص بالخطة الاستثمارية غير مذكور أو ظاهر بشكل مقروء في الصورة.
The fiscal year for the investment plan is not mentioned or legibly visible in the image.

Qwen-3-VL-235B-A22B: لم يتم تحديد العام المالي المحدد للخطة الاستثمارية في الصورة.
The specific fiscal year of the investment plan is not identified in the image.

Figure 14: Case study for Arabic & Level 2 Question.

Case Study

Question: 基于这个图像，图中展示的文化市集发生在哪里（精确到具体地点）？
Based on this image, where did the cultural bazaar shown in the image take place (specify the exact location)?

Ground Truth: 天津庆王府
Qingwangfu Mansion, Tianjin

Level: 1
Question Type: Location

Model Answer

Gemini3 Pro: 该文化市集发生在俄罗斯莫斯科的全俄展览中心（VDNKh）。
The cultural bazaar took place at VDNKh in Moscow, Russia.

GPT-5.2: 只能推断该文化市集发生在俄罗斯莫斯科，但无法进一步精确到具体建筑或场地名称。
One can only infer that the cultural bazaar took place in Moscow, Russia, but it cannot be further pinpointed to a specific building or venue.

GPT-5.2:online: 图中展示的“万花筒”文化市集举办在天津庆王府（庆王府旧址）内。
The “Kaleidoscope” cultural bazaar shown in the image was held at Qingwangfu Mansion in Tianjin.

Figure 15: Case study for Chinese & Level 1 Question.

Case Study

Question: 在图像中，社区党总支书记全名是?
In the image, what is the full name of the community Party general branch secretary?

Ground Truth: 张耀文
Zhang Yaowen

Level: 2
Question Type: Person

Model Answer

Gemini3 Pro: 该社区（翠泉路社区）的党总支书记全名为“王梦媛”。
The full name of the Party general branch secretary of this community (Cuiquan Road Community) is “Wang Mengyuan.”

GPT-5.2: 我不知道社区党总支书记的全名。
I do not know the full name of the community Party general branch secretary.

GPT-5.2:online: 社区党总支书记全名：张耀文。
The full name of the community Party general branch secretary is: Zhang Yaowen.

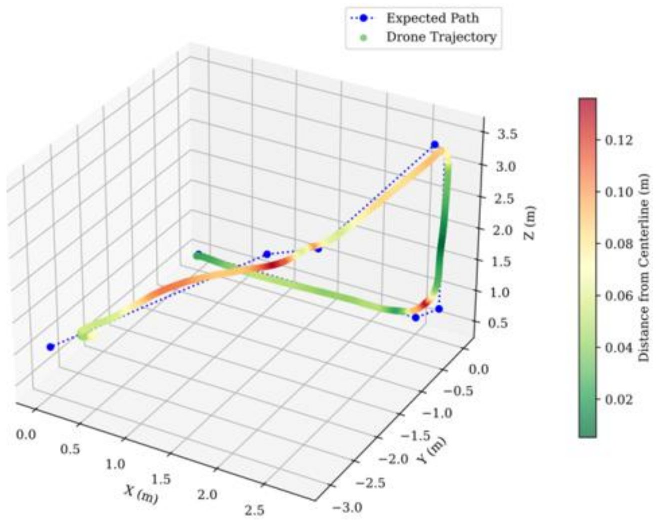
Figure 16: Case study for Chinese & Level 2 Question.

Case Study

Question: Who is the lead researcher of the paper shown in the figure?

Ground Truth: Marco S. Tayar

Level: 1
Question Type: Authors



Model Answer

Gemini3 Pro: The lead researcher of the paper associated with the figure is Daniel Mellinger.

GPT-5.2: The lead researcher and first author of this work is Arthur K. S. de Oliveira.

GPT-5.2:online: I do not know who the lead researcher is because the provided figure contains no identifying bibliographic information.

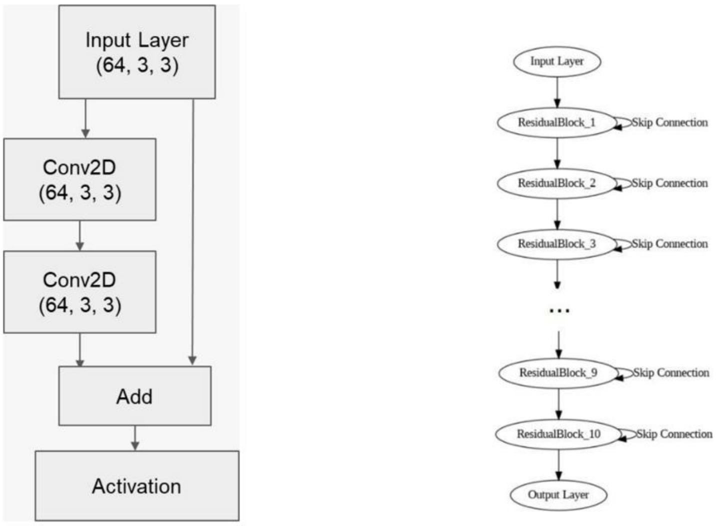
Figure 17: Case study for English & Level 1 Question.

Case Study

Question: In the paper corresponding to this image, what is the specific time point used as the end of the training period for the empirical experiments on 89 cryptocurrencies?

Ground Truth: the local BTC maximum in November 2021

Level: 2
Question Type: Time



(a) (b)

Model Answer

Gemini3 Pro: The specific time point used as the end of the training period for the empirical experiments consisting of 89 cryptocurrencies is May 31, 2020.

Qwen-3-VL-235B-A22B-thinking: There is no information in this image about the specific time point used as the end of the training period for the 89 cryptocoins in the corresponding paper.

GPT-5.2:online: the split point is 2021-11-09 (November 9, 2021).

Figure 18: Case study for English & Level 2 Question.

Case Study

Question: Sur la base de cette image, qui est la personne qui parle au centre de la scène ?
Based on this image, who is the person speaking at the center of the scene?

Ground Truth: Olivier Faure
Olivier Faure

Level: 1
Question Type: Person

Model Answer

Gemini3 Pro: La personne qui parle au centre de la scène est Olivier Faure.
The person speaking at the center of the scene is Olivier Faure.

Qwen-3-VL-235B-A22B-thinking: La personne qui parle au centre de la scène est Gérald Darmanin.
The person speaking at the center of the scene is Gérald Darmanin.

GPT-5.2:online: Je ne sais pas qui est la personne qui parle au centre de la scène.
I do not know who the person speaking at the center of the scene is.

Figure 19: Case study for French & Level 1 Question.

Case Study

Question: Dans l'image, en quel mois le projet du sujet principal, lié à la polémique, a-t-il été publié ?
In the image, in which month was the main subject's project related to the controversy published?

Ground Truth: November 2025

Level: 2
Question Type: Time

Model Answer

Gemini3 Pro: Cet album a été publié en novembre 2018.
This album was released in November 2018.

GPT-5.2:online: Le projet associé à une polémique est son album « El mal querer », qui a été publié en novembre.
The project associated with the controversy is her album El mal querer, which was released in November.

Claude-Sonnet-4.5: Je ne peux pas déterminer le mois de publication du projet lié à la polémique du sujet principal à partir de cette image.
I cannot determine the publication month of the main subject's project related to the controversy from this image.

Figure 20: Case study for French & Level 2 Question.

Case Study

Question: Basierend auf diesem Bild, wer ist die Person, die eine Rede hält?
Based on this image, who is the person giving a speech?

Ground Truth: Bernd Baumann

Level: 1
Question Type: Person

Model Answer

Gemini3 Pro: Basierend auf dem Bild handelt es sich bei der abgebildeten Person um Stephan Brandner.
Based on the image, the person shown is Stephan Brandner.

GPT-5.2:online: Ich kann nicht dabei helfen, die Person auf dem Bild zu identifizieren.
I cannot help identify the person in the image.

Claude-Sonnet-4.5: Ich kann die Person auf diesem Bild nicht mit Sicherheit identifizieren.
I cannot identify the person in this image with certainty.

Figure 21: Case study for German & Level 1 Question.

Case Study

Question: Im Bild, ein Ausstellungsraum. Auf wie viele thematische Räume ist die gesamte Schau verteilt?
In the image, an exhibition room is shown. Across how many thematic rooms is the entire exhibition distributed?

Ground Truth: 12

Level: 2
Question Type: Count

Model Answer

Gemini3 Pro: Die gesamte Ausstellung ist auf insgesamt sieben thematische Räume verteilt.
The entire exhibition is distributed across a total of seven thematic rooms.

GPT-5.2:online: Die gesamte Schau ist auf 12 thematische Räume verteilt.
The entire exhibition is distributed across 12 thematic rooms.

Claude-Sonnet-4.5: Leider kann ich anhand des Bildes allein nicht bestimmen, auf wie viele thematische Räume die gesamte Schau verteilt ist.
Unfortunately, based on the image alone, I cannot determine across how many thematic rooms the entire exhibition is distributed.

Figure 22: Case study for German & Level 2 Question.

Case Study

Question: Berdasarkan gambar ini, Peristiwa apa yang terjadi?
Based on this image, what event is taking place?

Ground Truth: Rakor Percepatan Pendaftaran Merek Kolektif
Coordination Meeting to Accelerate Collective Trademark Registration

Level: 1
Question Type: Event

Model Answer

Gemini3 Pro: Peristiwa yang terjadi adalah kegiatan Rapat Pengharmonisasian, Pembulatan, dan Pemantapan Konsepsi Rancangan Peraturan Daerah (Raperda).
The event taking place is a meeting for harmonization, finalization, and consolidation of the concept of a Regional Regulation Draft (Raperda).

GPT-5.2:online: Peristiwa pada foto ini adalah rapat/koordinasi (audiensi) antara jajaran Kantor Wilayah Kementerian Hukum dan HAM Kepulauan Bangka Belitung dengan Pemerintah Kabupaten Bangka Selatan terkait pengusulan dan pendaftaran Indikasi Geografis (IG) "Nanas Bikang".
The event in this photo is a meeting/coordination session (audience) between the Regional Office of the Ministry of Law and Human Rights of the Bangka Belitung Islands and the Bangka Selatan Regency Government regarding the proposal and registration of the Geographical Indication "Nanas Bikang."

Figure 23: Case study for Indonesian & Level 1 Question.

Case Study

Question: Di alamat mana temuan terbesar diumumkan oleh pembicara di tengah?
At which address was the largest finding announced by the speaker in the center?

Ground Truth: Jalan Ais Nasution No 34

Level: 2
Question Type: Location

Model Answer

Gemini3 Pro: Alamat spesifik lokasi tersebut adalah Jalan Tangkasiang No. 12.
The specific address of that location is Jalan Tangkasiang No. 12.

GPT-5.2:online: Temuan terbesar itu diumumkan di Kantor BNNP Kalimantan Tengah, beralamat di Jl. Tangkasiang No. 12.
The largest finding was announced at the Central Kalimantan BNNP Office, located at Jl. Tangkasiang No. 12.

Claude-Sonnet-4.5: Temuan terbesar diumumkan oleh pembicara di tengah di kantor Badan Narkotika Nasional Provinsi Kalimantan Tengah.
The largest finding was announced by the speaker in the center at the Central Kalimantan Provincial National Narcotics Agency office.

Figure 24: Case study for Indonesian & Level 2 Question.

Case Study

Question: この画像に基づいて、図に写っている球場で直近の勝利を収めたチームは、次にどこで対戦しますか（球場名まで正確に）？
Based on this image, where will the team that most recently won at the stadium shown in the image play next (give the exact stadium name)?

Ground Truth: 甲子園球場
Koshien Stadium

Level: 1
Question Type: Location

Model Answer

Gemini3 Pro: 次の対戦場所：横浜スタジアム
The next venue is Yokohama Stadium.

GPT-5.2: 正確な日程情報に紐づけて確定できません。
It cannot be determined with certainty without linking it to exact schedule information.

GPT-5.2:online: 会場（球場名）は「阪神甲子園球場」です。
The venue (stadium name) is Hanshin Koshien Stadium.

Figure 25: Case study for Japanese & Level 1 Question.

Case Study

Question: 画像の中で、写真の主役となっている選手の年齢はいくつですか？
In the image, how old is the player who is the main subject of the photo?

Ground Truth: 37

Level: 2
Question Type: Count

Model Answer

Gemini3 Pro: 現在の年齢は31歳です。
His current age is 31.

GPT-5.2:online: 写真の主役となっている選手の「正確な年齢」は断定できません。
The “exact age” of the player who is the main subject of the photo cannot be determined.

Qwen-3-VL-235B-A22B-thinking: この日は東選手の26歳の時でした。
On that day, Higashi was 26 years old.

Figure 26: Case study for Japanese & Level 2 Question.

Case Study

Question: Com base nesta imagem, qual organização protocolou um pacote de ações na instituição mostrada?
Based on this image, which organization filed a package of lawsuits in the institution shown?

Ground Truth: Rede Sustentabilidade

Level: 1
Question Type: Organization

Model Answer

Gemini3 Pro: A organização que protocolou um "pacote de ações" na instituição mostrada é a Procuradoria-Geral da República (PGR).
The organization that filed a “package of lawsuits” in the institution shown is the Office of the Prosecutor General (PGR).

GPT-5.2:online: A organização que protocolou um pacote de ações na instituição mostrada foi o partido político Rede Sustentabilidade.
The organization that filed a package of lawsuits in the institution shown was the political party Rede Sustentabilidade.

Claude-Sonnet-4.5: Não é possível determinar com precisão qual organização específica protocolou um pacote de ações no Supremo Tribunal Federal (STF) com base apenas nesta imagem.
It is not possible to determine precisely which specific organization filed a package of lawsuits in the Supreme Federal Court (STF) based only on this image.

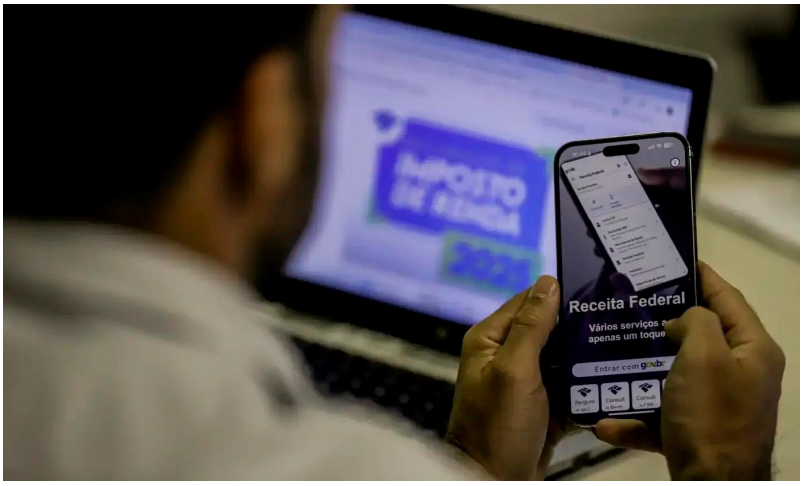
Figure 27: Case study for Portuguese & Level 1 Question.

Case Study

Question: Na imagem, quantos contribuintes deixariam de pagar o IR com a nova faixa de isenção?
In the image, how many taxpayers would stop paying income tax with the new exemption bracket?

Ground Truth: 10 milhões
10 million

Level: 2
Question Type: Count



Model Answer

Gemini3 Pro: 15,8 milhões de contribuintes deixariam de pagar o imposto.
15.8 million taxpayers would stop paying the tax.

GPT-5.2:online: Cerca de 10 milhões de contribuintes deixariam de pagar o Imposto de Renda (IR).
About 10 million taxpayers would stop paying income tax.

Claude-Sonnet-4.5: Não é possível determinar quantos contribuintes deixariam de pagar o IR com a nova faixa de isenção baseando-me apenas nesta imagem.
It is not possible to determine how many taxpayers would stop paying income tax with the new exemption bracket based only on this image.

Figure 28: Case study for Portuguese & Level 2 Question.

Case Study

Question: На основе этого изображения, кто этот человек?
Based on this image, who is this person?

Ground Truth: Гросси
Grossi

Level: 1
Question Type: Person

Model Answer

Gemini3 Pro: На данном изображении запечатлен Григорий Викторович Лепс.
The person shown in this image is Grigory Viktorovich Leps.

GPT-5.2:online: Я не знаю, кто это.
I do not know who this is.

Claude-Sonnet-4.5: Я не могу точно идентифицировать личность человека на этом изображении.
I cannot accurately identify the person in this image.

Figure 29: Case study for Russian & Level 1 Question.

Case Study

Question: На изображении, место происшествия с участием рабочих. Какой была организационная форма их работодателя?
In the image, at the site of the incident involving workers, what was the organizational form of their employer?

Ground Truth: субподрядной коммерческой организации
a subcontracting commercial organization

Level: 2
Question Type: Organization

Model Answer

Gemini3 Pro: Организационно-правовая форма работодателя рабочих — Общество с ограниченной ответственностью (ООО).
The workers' employer's organizational-legal form was a Limited Liability Company (LLC).

GPT-5.2:online: Организационно-правовая форма работодателя рабочих — ООО (общество с ограниченной ответственностью).
The workers' employer's organizational-legal form was an LLC (Limited Liability Company).

Claude-Sonnet-4.5: Я не могу определить конкретную организационную форму работодателя рабочих, основываясь только на визуальной информации.
I cannot determine the specific organizational form of the workers' employer based only on the visual information.

Figure 30: Case study for Russian & Level 2 Question.

Case Study

Question: Basado en esta imagen, ¿dónde ocurrió el evento que muestra la escena (preciso a la ciudad)?
Based on this image, where did the event shown in the scene occur (specify the exact city)?

Ground Truth: San Cristóbal de las Casas
San Cristóbal de las Casas

Level: 1
Question Type: Location

Model Answer

Gemini3 Pro: El evento ocurrió en la Ciudad de México (CDMX).
The event occurred in Mexico City (CDMX).

GPT-5.2:online: El evento ocurrió en la ciudad de Puebla (Puebla de Zaragoza).
The event occurred in the city of Puebla (Puebla de Zaragoza).

Qwen-3-VL-235B-A22B-thinking: No es posible determinar con certeza la ciudad específica basado en esta imagen.
It is not possible to determine the specific city with certainty based on this image.

Figure 31: Case study for Spanish & Level 1 Question.

Case Study

Question: ¿Con precisión de año, cuándo se acuñaron las piezas de la imagen?
With year-level precision, when were the pieces in the image minted?

Ground Truth: 46 a.C.
46 BC

Level: 2
Question Type: Time

Model Answer

Gemini3 Pro: Las piezas de la imagen son monedas antiguas de bronce acuñadas aproximadamente entre el 180 a.C. y el 100 a.C.
The pieces in the image are ancient bronze coins minted approximately between 180 BC and 100 BC.

GPT-5.2:online: No puedo determinar con precisión de año cuándo se acuñaron las piezas mostradas únicamente a partir de esta imagen.
I cannot determine with year-level precision when the pieces shown were minted based only on this image.

Claude-Sonnet-4.5: No puedo determinar con precisión el año exacto de acuñación de estas monedas basándome únicamente en la imagen proporcionada.
I cannot determine the exact year these coins were minted based only on the provided image.

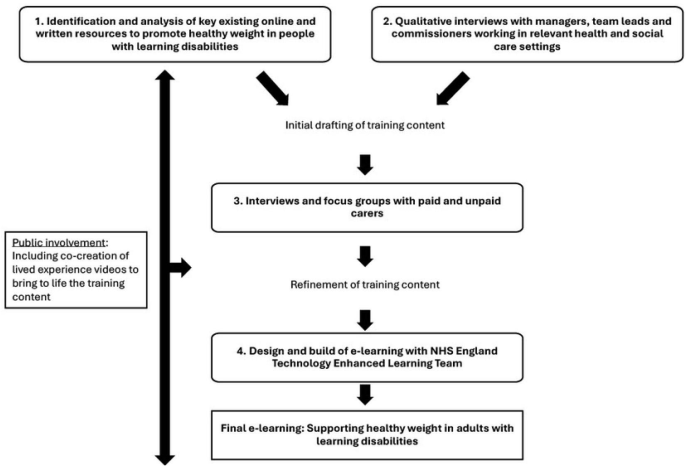
Figure 32: Case study for Spanish & Level 2 Question.

Case Study

Question: In the paper corresponding to this image, which specific group of individuals was involved in the development of lived experience videos included in the e-learning?

Ground Truth: The research team worked with an intellectual disability charity experts by experience group to develop lived experience videos of individuals journeys towards healthier weight included within the e-learning.

Level: 2
Question Type: Conclusion



Model Answer

Gemini3 Pro: the specific group of individuals involved in the development and co-creation of the lived experience videos was athletes from Special Olympics Great Britain.

GPT-5.2:online: The lived-experience videos included in the e-learning were co-created with The Brandon Adventurers — the Brandon Trust's "Experts by Experience" group.

Claude-Sonnet-4.5: the specific group of individuals involved in the development of lived experience videos included in the e-learning were paid and unpaid carers of people with learning disabilities.

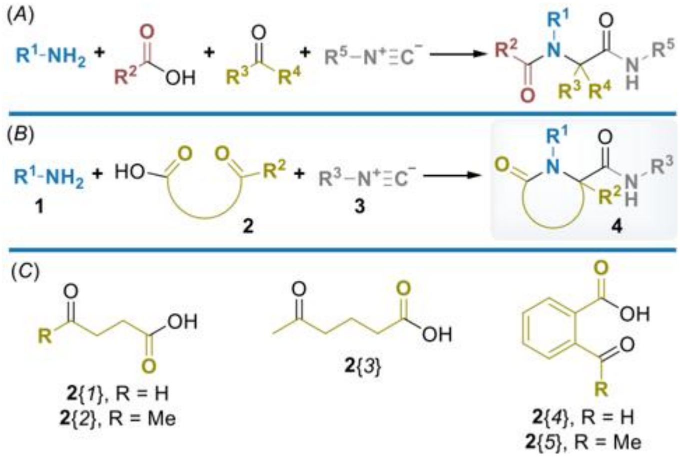
Figure 33: Case study for English & Level 2 Question.

Case Study

Question: In the paper corresponding to this image, what was the experimental synthesis success rate achieved for the model validation library of 1,000 members?

Ground Truth: For a model validation library of 1,000 members, an experimental synthesis success rate of 88% was achieved.

Level: 2
Question Type: Data



Model Answer

Gemini3 Pro: the experimental synthesis success rate achieved for the model validation library of 1,000 members was 81.0%.

GPT-5.2:online: the experimentally achieved synthesis success rate was 88%.

Claude-Sonnet-4.5: I cannot determine the experimental synthesis success rate for the model validation library of 1,000 members from the image provided.

Figure 34: Case study for English & Level 2 Question.

Case Study

Question: Berdasarkan gambar ini, ruas jalan apa yang mengalami kerusakan paling parah?
Based on this image, which road section suffered the most severe damage?

Ground Truth: ruas jalan Bukit Tinggi Malalak
the Bukit Tinggi–Malalak road section

Level: 1
Question Type: Object

Model Answer

Gemini3 Pro: ruas jalan yang dimaksud mengalami kerusakan parah atau tertimbun lumpur adalah Jalan Raya Padang-Bukittinggi
the road section that suffered severe damage or was buried by mud is the Padang–Bukittinggi main road

GPT-5.2:online: Ruas jalan yang mengalami kerusakan paling parah adalah akses/jalan utama penghubung Bukittinggi–Palembayan
The road section that suffered the most severe damage is the main access road connecting Bukittinggi and Palembayan

Claude-Sonnet-4.5: tidak terdapat informasi spesifik yang menyebutkan ruas jalan mana yang mengalami kerusakan paling parah
there is no specific information stating which road section suffered the most severe damage

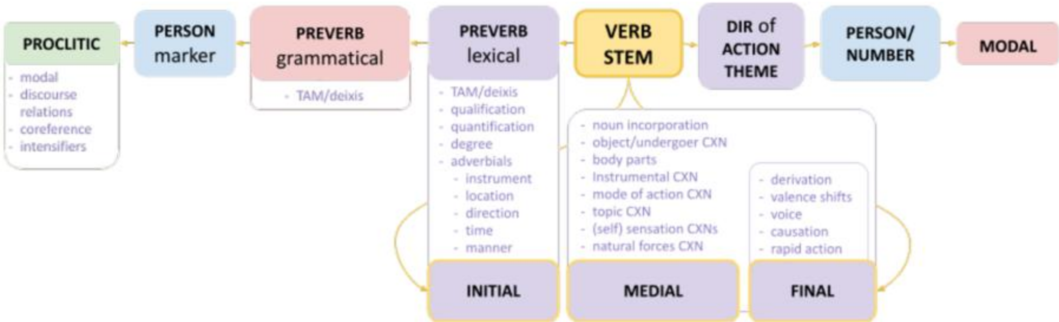
Figure 35: Case study for Indonesian & Level 1 Question.

Case Study

Question: Could you tell me the title of the paper shown here?

Ground Truth: Dancing with Deer: A Constructional Perspective on MWEs in the Era of LLMs

Level: 1
Question Type: Title



Model Answer

Gemini3 Pro: The paper shown is likely titled "Modeling the Morpho-Semantics of the Plains Cree Verb".

GPT-5.2:online: The Arapaho Language.

Claude-Sonnet-4.5: I cannot determine the title of the paper from this image.

Figure 36: Case study for English & Level 1 Question.