Benchmarking Deflection and Hallucination in Large Vision-Language Models

Nicholas Moratelli[†*], **Christopher Davis**[‡], **Leonardo F. R. Ribeiro**[‡], **Bill Byrne**[‡,§], **Gonzalo Iglesias**[‡],
[†]University of Modena and Reggio Emilia , [‡]Amazon AGI, [§]University of Cambridge
Correspondence: davisjnh@amazon.com

Abstract

Large Vision–Language Models (LVLMs) increasingly rely on retrieval to answer knowledge-intensive *multimodal* questions. Existing benchmarks overlook conflicts between visual and textual evidence and the importance of generating *deflections* (*e.g.*, "Sorry, I cannot answer...") when retrieved knowledge is incomplete. These benchmarks also suffer from rapid obsolescence, as growing LVLM training sets allow models to answer many questions without retrieval. We address these gaps with three contributions. First, we propose a dynamic data curation pipeline that preserves benchmark difficulty over time by filtering for genuinely retrieval-dependent samples. Second, we introduce **VLM-DeflectionBench**, a benchmark of 2,775 samples spanning diverse multimodal retrieval settings, designed to probe model behaviour under conflicting or insufficient evidence. Third, we define a fine-grained evaluation protocol with four scenarios that disentangle parametric memorization from retrieval robustness. Experiments across 20 state-of-the-art LVLMs indicate that models usually fail to deflect in the presence of noisy or misleading evidence. Our results highlight the need to evaluate not only what models know, but how they behave when they do not, and serve as a reusable and extensible benchmark for reliable KB-VQA evaluation. All resources will be publicly available upon publication.

1 Introduction

Large Vision–Language Models (LVLMs) are rapidly moving into real-world applications where reliability is critical (Caffagni et al., 2024a; Liu et al., 2024). Users expect not only accurate answers to complex multimodal queries, but also trustful behavior when knowledge is missing or contradictory. Indeed, LVLMs prompted to strictly ground their answers in retrieved evidence should state that the user request cannot be fulfilled when said knowledge does not support a reliable answer. Instead, models might generate an answer that cannot be traced back to correctly retrieved evidence. In this paper, we will refer to these two types of answer as **deflection** and **hallucination**, respectively. We note that both types of answers are strictly speaking incorrect in that the user request is not fulfilled. However, when faced with insufficient evidence, deflection is the preferable failure mode.

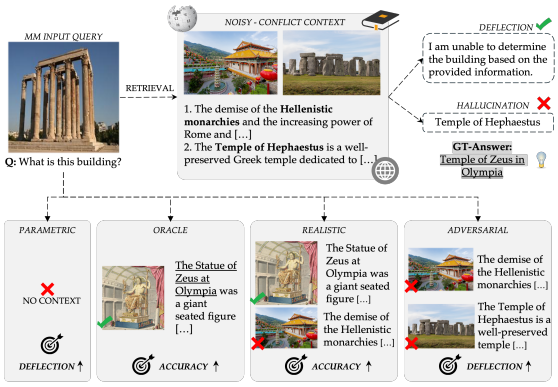


Figure 1: Overview of **VLM-DeflectionBench**. Top: LVLMs often hallucinate instead of abstaining when context is misleading. Bottom: **VLM-DeflectionBench** evaluates calibration across four scenarios-**Parametric**, **Oracle**, **Realistic**, and **Adversarial** to test whether models align their behavior with available knowledge.

Knowledge-based Visual Question Answering (KB-VQA) provides a natural testbed to study these behaviors. In KB-VQA, models must integrate both visual inputs and retrieved textual or visual passages to answer open-ended questions. Several benchmarks (Mensink et al., 2023; Chen et al., 2023; Chang et al., 2022; Lerner et al., 2022; Hu et al., 2025) have addressed these problems, but they face two limitations. First, they suffer from *rapid obsolescence*: many questions that once re-

[*]Work done during an internship at Amazon.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28348–28370

quired retrieval can now be answered directly from parametric knowledge. Second, they focus on *accuracy*, but do not distinguish failure modes.

To close this gap we present **VLM-DeflectionBench**, a new benchmark with a strict retrieval-augmented generation evaluation. As illustrated in Figure 1, models are judged not only on what they know, but also on how they behave when knowledge is missing or misleading. An ideal system should abstain when reliable evidence is absent, rather than hallucinate unsupported answers.

With this benchmark we can address a central question: *can LVLMs deflect when they do not have enough information?* Our results show that the answer is still no. Experiments with 20 state-of-the-art LVLMs indicate that even the strongest proprietary systems fail to deflect reliably, often hallucinating when provided incomplete or distracting knowledge.

We note that **VLM-DeflectionBench** is designed as a reusable and extensible framework rather than a one-off dataset. Future versions can rerun the curation pipeline to incorporate additional KB-VQA sources and include stronger models to filter parametrically-answerable questions. This ensures **VLM-DeflectionBench** will continue to provide controlled conditions for measuring hallucination and deflection behavior under realistic retrieval noise, while preserving comparability through fixed, versioned releases. To our knowledge, it is the first KB-VQA benchmark to combine multimodal retrieval with explicit evaluation of hallucination and deflection, while also providing a dynamic foundation that can evolve alongside LVLMs. In summary, our contributions are threefold:

- We propose a dynamic curation pipeline that filters out parametrically solvable samples, ensuring that retained questions remain genuinely retrieval-dependent as models improve.
- We use this pipeline to build **VLM-DeflectionBench**, a collection of 2,775 samples covering diverse multimodal retrieval scenarios, each paired with both gold and distractor contexts.
- We define a fine-grained evaluation protocol with four complementary scenarios: parametric, oracle, realistic, and adversarial that disentangle memorization from retrieval robustness and explicitly measure hallucination versus deflection.

2 Related Work

Knowledge-Based VQA Benchmarks. Research on Knowledge-based Visual Question Answering (KB-VQA) (Qiu et al., 2024; Wu et al., 2025; Shah et al., 2019; Su et al., 2025) has progressed from early entity-centric datasets such as OK-VQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), and ViQuAE (Lerner et al., 2022), which assumed every question was answerable and discouraged abstention. InfoSeek (Chen et al., 2023) and Encyclopedic-VQA (Mensink et al., 2023) introduced retrieval pipelines, while WebQA (Chang et al., 2022) emphasized multi-hop reasoning. However, these benchmarks remained primarily text-centric, where external knowledge was supplied only as passages, without visual retrieval or multimodal distractors. Evaluation thus reduced to text-grounding accuracy, treating abstention as incorrect. More recent multimodal retrieval benchmarks (Wu et al., 2025; Dong et al., 2025) such as MRAG-Bench (Hu et al., 2025) incorporate visual contexts, moving toward realistic multimodal information needs. However, they still lack explicit mechanisms for unanswerable queries or for distinguishing hallucination from deflection.

Hallucination and Deflection Benchmarks. Parallel efforts have begun probing reliability beyond accuracy. Vision-only datasets such as HaloQuest (Wang et al., 2024) and AMBER (Wang et al., 2023) diagnose hallucinations from mis-grounded visual evidence, while text-only sources (Yang et al., 2024; Wei et al., 2024) such as MultiHop-RAG (Tang and Yang, 2024) and GaRaGe (Sorodoc et al., 2025) include unanswerable queries and explicitly reward abstention. However, they remain unimodal and cannot capture conflicts between retrieved text and images.

Retrieval-Augmented Multimodal Models. Recent advances in KB-VQA have focused on developing specialized architectures (Deng et al., 2025; Yuan et al., 2025; Xuan et al., 2024; Chen et al., 2022) – including hierarchical retrieval (Caffagni et al., 2024b), re-ranking (Yan and Xie, 2024), and adaptive control mechanisms (Cocchi et al., 2025; Asai et al., 2024). These works highlight the centrality of retrieval for LVLMs but generally assume

Benchmark	Evaluation Scope				Evidence Configuration			
	Hallucination	Deflection	Scenarios	Dynamic	MM-Query	MM-Context	Pre-retrieved	Negatives
OK-VQA (Marino et al., 2019)	×	×	1	×	✓	×	×	×
A-OKVQA (Schwenk et al., 2022)	×	×	1	×	✓	×	×	×
E-VQA (Mensink et al., 2023)	×	×	1	×	✓	×	×	×
InfoSeek (Chen et al., 2023)	×	×	1	×	✓	×	×	×
WebQA (Chang et al., 2022)	×	×	1	×	×	✓	✓	✓
MRAG-Bench (Hu et al., 2025)	×	×	1	×	✓	×	✓	×
ViQuAE (Lerner et al., 2022)	×	×	1	×	✓	×	×	×
SK-VQA (Su et al., 2025)	×	×	1	×	✓	×	✓	×
MMDocRAG (Dong et al., 2025)	×	×	1	×	×	✓	✓	✓
VLM-DeflectionBench	✓	✓	**4**	✓	✓	✓	✓	✓

Table 1: Comparison of **VLM-DeflectionBench** with current KB-VQA benchmarks. Our benchmark is the only one that explicitly evaluates hallucination and deflection across four retrieval scenarios: parametric (no additional context), oracle (gold evidence only), realistic (mixed positive and negative evidence), and adversarial (only negative contexts).

high retrieval quality and measure only accuracy, leaving reliability under noisy or adversarial contexts underexplored.

Table 1 compares **VLM-DeflectionBench** with other representative KB-VQA benchmarks. Existing resources focus narrowly on accuracy, omit distractors, or rely on uncontrolled online retrieval, which, while realistic, makes evaluation difficult to control and reproduce over time. In contrast, **VLM-DeflectionBench** explicitly evaluates hallucination and deflection across four complementary scenarios, with pre-retrieved gold and distractor contexts that preserve benchmark stability as LVLMs evolve. To our knowledge, it is the first KB-VQA framework to address hallucination and deflection in multimodal retrieval.

3 Multi-Modal Deflection Benchmark

Unlike static datasets, **VLM-DeflectionBench** is built through a dynamic pipeline parameterized by strong open-weight models and an external evaluator, ensuring adaptability as model capabilities evolve. As shown in Figure 2, the pipeline consists of three stages: (i) filtering parametrically-answerable samples, (ii) augmenting retained samples with gold and distractor knowledge contexts, and (iii) quality control to guarantee both solvability and difficulty.

3.1 Stage I: Filtering samples that can be answered parametrically

Let $\mathcal{S}_0 = \{x\}$ denote the pool of raw samples collected from heterogeneous KB-VQA sources. A KB-VQA instance is $x = (q, v, \mathcal{K}, a)$, where q is a question, v is an optional image, $\mathcal{K}$ is candidate knowledge, and a is the gold answer. We denote by $\mathcal{G} = \{G_1, \ldots, G_n\}$ a set of *gating models*[1]. Not all sources provide visual inputs, so v may be empty. This reflects the diversity of existing KB-VQA datasets, which range from purely text-based queries to fully multimodal questions requiring both text and images.

Source	Samples	Filtered (%)	Query	Context
InfoSeek	1646	5.6%	Multi	Textual
WebQA	826	16.6%	Textual	Multi
E-VQA	185	4.9%	Multi	Textual
MMDocRAG	58	1.5%	Textual	Multi
MRAG-Bench	45	3.3%	Multi	Visual
ViQuAE	15	1.2%	Multi	Textual
Total	2,775	5.1%	Multi	Multi

Table 2: Per-source breakdown of our dataset. We report the number of retained samples, filtering rate, and dominant modality for each source. The "Filtered" column indicates the percentage of samples retained from the original source after applying our filtering criteria.

Statistic	Value
Total samples	2,775
Unique questions	1,246
Unique query images	2,717
Avg. question length (words)	12.5
Avg. gold contexts per sample	1.9
Avg. negative contexts per sample	11.8

Table 3: Global statistics of **VLM-DeflectionBench**. The dataset combines gold and negative contexts across text and vision, ensuring multimodal coverage and non-triviality.

To eliminate parametrically solvable instances, each G_j is queried in the **parametric setting with visual query** (hereinafter, the parametric setting), where the model receives the question q together with the associated image v when available, but no external knowledge: $\hat{a}_j =$

[1]We set $\mathcal{G} = \{$GEMMA3-27B, QWEN-2.5-VL-32B, INTERNVL3-38B, VL-RETHINKER-72B$\}$

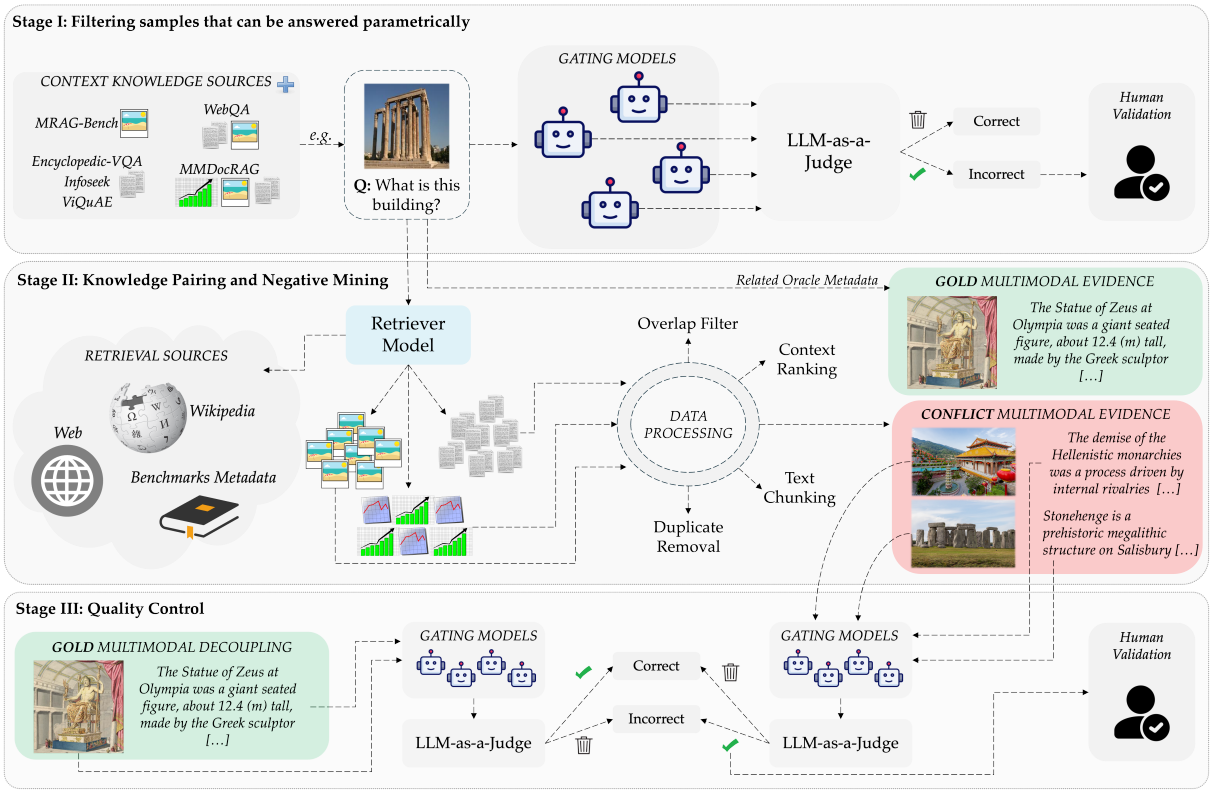


Figure 2: Pipeline of **VLM-DeflectionBench**. Starting from 6 benchmarks, we apply parametric filtering to remove query-solvable samples (STAGE I), retrieve negative (multimodal) contexts via different indices (STAGE II), then perform oracle filtering to eliminate false positive contexts (where gating models fail) and unreliable negative contexts (where gating models succeed) (STAGE III).

$G_j(q, v)$. Responses are judged by an external evaluator E: we use GPT-4o as a judge with the SIMPLEQA prompt (Wei et al., 2024), which assigns one of three labels: $E(q, \hat{a}_j, a) \in \{\texttt{CORRECT}, \texttt{INCORRECT}, \texttt{NOT ATTEMPTED}\}$[2].

We retain only samples unsolvable by all gating models:

$$\mathcal{R}(x) = 1 \iff \forall G_j \in \mathcal{G},\ E(q, G_j(q, v), a) = \texttt{INCORRECT}. \quad (1)$$

This produces $\mathcal{S}_1 \subseteq \mathcal{S}_0$, a set requiring external knowledge to solve.

3.2 Stage II: Knowledge Pairing and Negative Mining

Each retained sample $x \in \mathcal{S}_1$ is paired with gold knowledge $K(x)^+$ and distractors $K(x)^-$. If negatives exist in the source dataset, we adopt them; otherwise, we mine additional distractors via retrieval. For textual negatives, we construct a Wikipedia index $\mathcal{W}_\mathcal{T}$ and retrieve candidate pages using EVA-CLIP (denoted as R^ϕ, where ϕ represents the similarity function). These pages are then segmented into passages and re-ranked using Contriever (Izacard et al., 2021); for $x = (q, v, \mathcal{K}, a)$:

$$\begin{aligned} C(x) &= \text{Chunk}(\text{top-10}(R^\phi(v; \mathcal{W}_\mathcal{T}))) \\ K(x)^{-\text{txt}} &= \text{Rerank}(C(x)) \setminus K(x)^+ \end{aligned} \quad (2)$$

For visual negatives, we build an index $\mathcal{W}_\mathcal{I}$ of query and metadata images, retrieving the top-10 most similar non-gold images:

$$K(x)^{-\text{img}} = \text{top-10}(R^\phi(v; \mathcal{W}_\mathcal{I})) \setminus K(x)^+. \quad (3)$$

3.3 Stage III: Quality Control

We apply two checks. First, solvability: if all gating models fail even with gold evidence, the sample is discarded:

$$\forall G_j \in \mathcal{G}, \quad E(q, G_j(q, v, K^+), a) = \texttt{INC}. \quad (4)$$

Second, distractor validity: if any gating model exploits a distractor k^- to answer correctly, that distractor is removed:

$$\exists G_j \ \text{ s.t. } \ E(q, G_j(q, v, k^-), a) = \texttt{CORRECT}. \quad (5)$$

[2]We map these labels to our evaluation metrics as follows: CORRECT → Accuracy, INCORRECT → Hallucination, NOT ATTEMPTED → Deflection.

Finally, we enforce a minimum of $K_{\min} = 5$ negatives per sample, yielding the final benchmark:

$$\mathcal{S}^* = \{s, (K(x)^+, K(x)^-)\}, \quad (6)$$

and we randomly shuffle the order of positive and negative contexts before presenting them to models during evaluation, to simulate realistic retrieval scenarios.

3.4 Benchmark Definition

The final benchmark comprises 2,775 curated samples, each paired with gold and distractor passages. All contexts are pre-retrieved, ensuring reproducibility without external databases while reflecting real-world retrieval noise.

Table 2 summarizes **VLM-DeflectionBench**, reporting per-source retention, filtering ratio, dominant modality, and average of multimodal contexts. Complementary statistics are shown in Table 3, including unique questions and images, average question length, and prevalence of multimodal distractors. Full distributions are provided in Appendix C.

3.5 Evaluation Scenarios

We define four scenarios by controlling the context $\mathcal{Z}(x)$: **Parametric** ($\mathcal{Z}_P(x) = \emptyset$) forces reliance on parametric knowledge only, while still grounding on the input image v when available. **Oracle** ($\mathcal{Z}_O(x) = K^+$) supplies only gold knowledge. **Realistic** ($\mathcal{Z}_R(x) = K^+ \cup K^-$) mixes gold and distractors. **Adversarial** ($\mathcal{Z}_A(x) = K^-$) includes only distractors. Ideal behavior varies by scenario: **Parametric** should yield minimal accuracy, confirming our controlled strict-RAG setup; **Oracle** should achieve high accuracy given perfect evidence; **Realistic** should maintain high accuracy despite distractors; **Adversarial** should maximize deflection. Across all conditions, robust models should exhibit low hallucination rates. We acknowledge that some parametric accuracy is expected due to models' inherent knowledge.

Together, these scenarios span the full spectrum from knowledge-free to perfectly grounded to noisy and misleading contexts. Crucially, our framework implements a rigorous strict retrieval-augmented generation evaluation: we assess systems based on their factual accuracy and their response when confronted with inadequate or contradictory retrieved context. A trustworthy RAG system should recognize when to withhold answers due to poor evidence quality, rather than generate confident but unsubstantiated claims.

4 Benchmark Evaluation

We evaluate across all scenarios using GPT-4o as a judge with the SIMPLEQA protocol, following prior works (Wei et al., 2024; Haas et al., 2025). To verify robustness, we also conducted a human validation study, which showed $> 92\%$ agreement and Cohen's κ=0.91 across models of different scales (Appendix E).

Since models are explicitly instructed to rely solely on provided evidence (*i.e.*, a strict RAG setting), we assume all incorrect responses as hallucinations. We acknowledge that in the Realistic scenarios some INCORRECT predictions may correspond to misgrounding (*e.g.*, selecting a distractor) rather than unsupported fabrication. Nevertheless, we follow prior practice in grouping them under hallucination for consistency and comparability across scenarios. This unified formulation provides a reproducible framework for knowledge-based VQA.

4.1 Experimental Setup

We design our experiments to address four research questions. First, we ask whether current LVLMs can deflect reliably under different knowledge conditions (Sec. 4.2). Second, we analyze how modality interacts with reliability (Sec. 4.3).

Third, we study how prompting strictness influences the deflection–hallucination trade-off (Sec. 4.4). Finally, we evaluate robustness to retrieval noise (Sec. 4.5).

We evaluate 20 representative LVLMs, spanning both open-weight and proprietary systems. On the open-weight side, we include multiple variants from major families: LLAVA-ONEVISION (Li et al., 2024a), MINICPM-4.5-8B (Yao et al., 2024), KEYE-1.5 (Yang et al., 2025), INTERNVL3 (9B, 38B) (Zhu et al., 2025), GLM-4.1V-9B (Hong et al., 2025), PIXTRAL (Agrawal et al., 2024), OVIS2 (16B, 34B) (Lu et al., 2024), MISTRAL-SMALL-3.1-24B-INSTRUCT, ARIA (Li et al., 2024b), GEMMA3-27B (Team et al., 2025), QWEN-2.5-VL-32B (Bai et al., 2025) and VL-RETHINKER (Wang et al., 2025), ranging from 7B to 72B parameters.

For proprietary systems, we test CLAUDE-4 (SONNET, OPUS), GPT-5 and GEMINI-2.5 (FLASH, PRO) (Comanici et al., 2025), representing the current frontier of commercial LVLMs. All open-weight models are run with the vLLM

Model	Parametric			Oracle			Realistic			Adversarial		
	Acc ↑	Defl ↑	Hall ↓	Acc ↑	Defl ↓	Hall ↓	Acc	Defl ↓	Hall ↓	Acc ↑	Defl ↑	Hall ↓
Open Source Models												
LLaVA-OneVision	4.7	4.0	91.3	54.9	**3.5**	41.6	34.0	10.8	55.2	**4.5**	14.5	81.0
MiniCPM-V-4.5-8B	4.6	3.2	92.2	65.3	8.9	25.8	44.3	16.3	39.4	1.9	46.2	51.9
Keye-1.5-8B	3.0	23.9	73.1	63.4	9.3	27.3	47.6	11.9	40.5	2.8	43.4	53.8
InternVL3-9B	3.1	9.6	87.3	61.2	7.7	31.1	45.1	10.3	44.6	2.3	35.9	61.8
GLM-4.1V-9B-Thinking	4.9	7.7	87.4	55.2	31.5	13.3	37.4	40.2	22.4	1.2	76.3	22.5
Pixtral-12B	4.6	19.4	76.0	62.7	6.8	30.5	42.6	12.6	44.8	2.3	40.9	56.8
Ovis2-16B	4.0	25.2	70.8	64.4	4.2	31.4	**61.1**	**3.7**	35.2	3.9	24.2	71.9
Mistral-Small-3.1-24B-Instruct	2.5	**70.9**	**26.6**	42.6	47.1	**10.3**	23.5	61.6	**14.9**	0.6	**83.8**	**15.6**
Aria-25B	4.4	31.0	64.6	62.3	10.2	27.5	39.5	21.4	39.1	2.9	40.7	56.4
Gemma3-27B⋆	4.6	7.1	88.3	59.5	16.1	24.4	42.9	22.6	34.5	1.5	59.1	39.4
Qwen2.5-VL-32B-Instruct⋆	3.2	4.5	92.3	61.0	5.1	33.9	45.2	5.3	49.5	2.4	13.7	83.9
Ovis2-34B	4.5	21.7	73.8	**66.5**	6.7	27.8	49.1	7.6	43.3	3.2	38.7	58.1
InternVL3-38B⋆	4.8	1.4	93.8	64.2	13.2	22.6	50.3	15.8	33.9	2.5	55.2	42.3
VL-Rethinker-72B⋆	**5.0**	1.9	93.1	62.6	12.9	24.5	47.2	11.9	40.9	2.3	42.3	55.4
Closed Source Models												
Claude-Sonnet-4	9.7	31.8	58.5	47.5	42.6	9.9	33.3	56.2	10.5	0.9	87.0	12.1
Claude-Opus-4	6.7	**65.0**	**28.3**	49.1	41.7	**9.2**	32.1	59.4	**8.5**	0.7	**88.3**	**11.1**
Gemini-2.5-Flash	17.5	11.7	70.8	58.8	27.9	13.3	47.0	32.3	20.7	1.4	74.5	24.1
Gemini-2.5-Pro	**27.0**	7.5	65.5	59.8	26.3	13.9	51.0	28.5	20.5	1.7	76.1	22.2
GPT-5	23.7	1.5	74.8	**73.1**	**14.3**	12.6	**59.5**	**15.0**	25.5	**4.1**	61.2	34.7

Table 4: Performance of LVLMs across evaluation scenarios. We report Accuracy (Acc), Deflection Rate (Defl), and Hallucination Rate (Hall). The number of negatives is fixed at 2 to match the average number of gold contexts per sample, ensuring a balanced ratio of positive and negative evidence. ⋆ marks selected gating models.

framework[3], using top-$p = 1.0$, repetition penalty 1.05, and temperature 0.2. Closed-source APIs are queried with their default decoding settings. Unless otherwise specified, all experiments use the moderate prompting strictness (see Appendix I for full prompt details).

4.2 Deflection under Knowledge Conditions

Table 4 compares open- and closed-weight models across the four scenarios. In the **parametric** setting, where only the question and image are available, accuracy is near-zero for open-weight systems, confirming that **VLM-DeflectionBench** successfully filters out parametrically answerable questions. The majority of models attempt to answer most questions rather than abstain, with deflection rates below 35%. The only exception is MISTRAL-SMALL-3.1, which deflects on 70.9% of queries. Proprietary systems behave differently: GEMINI-2.5-PRO and GPT-5 obtain accuracies of 27.0% 23.7%, respectively, likely reflecting training-set contamination.

Providing **oracle** knowledge boosts accuracy substantially, with several open-weight models surpassing 60% (OVIS2-34B 66.5%), and even outperforming CLAUDE-OPUS-4 (49.1%). Hallucination, however, remains strikingly high: LLAVA-ONEVISION hallucinates on 41.6% of queries despite being given the correct evidence, showing that grounding, not retrieval, is the primary bottleneck.

In the **realistic** setting, where gold evidence is mixed with distractors, accuracy drops by 10–20 points and hallucination often exceeds 40%. For instance, PIXTRAL-12B falls from 62.7% to 42.6% accuracy, while even GPT-5 hallucinates on 25.5% of queries.

Finally, in the **adversarial** setting with only distractors where the desired behavior is near-total abstention, CLAUDE-OPUS-4 achieves 88.3% deflection and MISTRAL-SMALL-3.1 reaches 83.8%. Some accuracy remains possible as models may still possess parametric knowledge to answer correctly despite misleading context. Notably, consistent with the parametric setting, most open-weight systems attempt to answer questions rather than deflect, with QWEN-2.5-VL providing incorrect responses on 83.9% of queries.

Overall, no model achieves balanced reliability: MISTRAL over-deflects, OVIS2 excels only with clean evidence, and CLAUDE sacrifices oracle accuracy for adversarial robustness. These contrasts illustrate the unique value of **VLM-DeflectionBench** in exposing hallucination–deflection trade-offs that remain invisible

[3] https://docs.vllm.ai

when measuring accuracy alone.

Model	Context		Acc ↑	Defl ↓	Hall ↓
	Positive	Negative			
Keye-1.5-8B	Textual	MM	80.1	3.3	16.6
		Visual	82.1	2.7	15.2
		Textual	82.5	3.3	14.2
	Visual	MM	1.0	82.5	16.5
		Visual	9.4	19.2	71.4
		Textual	0.9	82.8	16.3
Mistral-Small-3.1	Textual	MM	60.2	31.5	8.3
		Visual	69.4	23.6	7.0
		Textual	53.8	39.4	6.8
	Visual	MM	0.8	98.1	1.1
		Visual	3.8	93.2	3.0
		Textual	1.1	97.4	1.5
Claude-Opus-4	Textual	MM	82.8	9.8	7.4
		Visual	84.3	9.6	6.0
		Textual	80.8	13.0	6.2
	Visual	MM	18.8	72.9	8.3
		Visual	35.3	50.8	13.9
		Textual	28.9	63.2	7.9

Table 5: Effect of negative-context modality in the **Realistic** scenario. Results reveal a language-over-vision bias: models handle visual distractors well when positives are textual, but collapse when textual distractors interfere with visual positives, demonstrating that misleading text systematically overrides correct visual evidence.

4.3 Impact of Negative Context

In Table 5 we analyze model behavior in the **Realistic** scenario when the modality of negative contexts varies across textual, visual, or both. We evaluate 530 examples with textual positives and 266 examples with visual positives[4]. Notably, we observe a consistent drop in overall performance when moving from textual to visual positive contexts across all models, indicating that LVLMs struggle more with visual knowledge grounding than textual knowledge. When positives are textual, models generally handle visual distractors well: CLAUDE-OPUS-4 maintains 84.3% accuracy with visual noise, while KEYE-1.5-8B even improves slightly (82.1% vs. 80.1% with mixed noise). In contrast, MISTRAL-SMALL-3.1 becomes more conservative, dropping from 69.4% accuracy with visual distractors to 60.2% with mixed, while deflection rises from 23.6% to 31.5%. The picture reverses when positives are visual: textual distractors dominate. Both KEYE-1.5-8B and MISTRAL-SMALL-3.1 collapse to near-zero accuracy ($\leq 1\%$) with deflection exceeding 80%, effectively refusing to answer once misleading text is present. Even CLAUDE-OPUS-4 falls from 35.3% accuracy with visual distractors to 28.9% with textual ones. Multimodal distractors produce intermediate results but still degrade reliability.

Crucially, this asymmetry reverses the expected grounding hierarchy: text dominates vision even when vision carries the gold evidence. The presence of misleading text systematically overrides correct visual signals, reducing accuracy across all architectures. Multimodal distractors yield intermediate degradation, reflecting partial—but still language-weighted—fusion. These results expose a persistent **language-over-vision bias** in current LVLMs: they rely on textual priors even under oracle visual grounding. This failure mode highlights the limits of current alignment training and underscores the need for *retrieval-aware grounding* strategies that weight evidence by modality reliability rather than surface plausibility.

4.4 Effect of Prompt Strictness and Trade-offs

We test whether increasing the strictness of refusal instructions affects the balance between deflection and hallucination. Table 6 reports results for soft, moderate, and severe prompts (see Appendix I for full prompt text) in the realistic and adversarial scenarios. These prompts differ only in how forcefully they direct the model to abstain when uncertain: the soft version politely suggests caution; the moderate version explicitly recommends abstention if unsure; and the strict version mandates refusal unless the answer is fully supported. Results show a clear but uneven effect. In the adversarial setting, where the target behavior is near-total abstention, stronger instructions substantially increase deflection. For instance, MISTRAL-SMALL-3.1 jumps from 57.6% to 98.2% deflection with severe prompting, nearly eliminating hallucinations. Proprietary systems such as CLAUDE-OPUS-4 also benefit, though less dramatically ($87.0 \rightarrow 92.2\%$). By contrast, in the realistic setting where gold evidence is mixed with distractors, severity introduces a trade-off: hallucination decreases, but accuracy collapses. For example, KEYE-1.5-8B drops from 58.7% to 36.2% accuracy, while MISTRAL falls from 57.7% to 9.2% accuracy under severe prompting, reflecting excessive caution. Overall, prompt engineering is a powerful but blunt tool for the prompts we tested. Strong refusal instructions enforce abstention in adversarial cases but cause over-

[4]Sample sizes differ because not all examples support both multimodal positive and negative contexts due to source limitations and question formulation constraints.

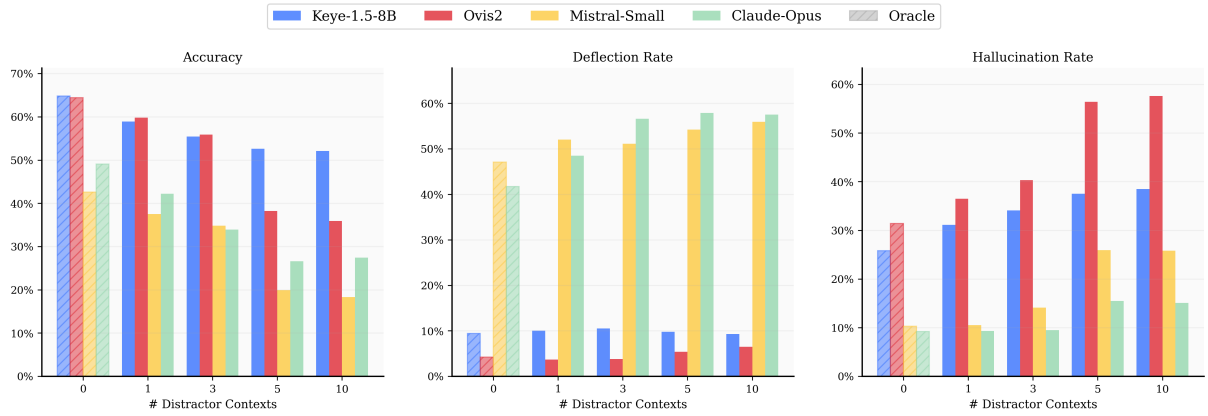


Figure 3: Effect of distractor quantity in the **Realistic** scenario, including **Oracle** results (0 distractors) for reference. Accuracy declines and hallucination rises as negatives increase, while deflection shows limited improvement.

deflection in realistic ones, undermining accuracy even when valid evidence is present. While model-specific prompt optimization might achieve better accuracy-deflection trade-offs, our results suggest that distinguishing genuine uncertainty from noisy retrieval remains difficult and cannot be easily solved by prompting alone.

4.5 Robustness to Retrieval Noise

To test how LVLMs cope with noise, we vary the number of negatives from 1 to 10 in the **realistic** scenario (Figure 3). Results reveal a sharp vulnerability to distractor density. Accuracy steadily declines while hallucination rises, even for strong systems. For instance, OVIS2-16B drops from 59.8% accuracy with one distractor to 35.9% with ten, with hallucination nearly doubling (36.5% → 57.6%). Proprietary models are more resilient but not immune: the accuracy of CLAUDE-OPUS-4 still falls from 42.2% to 27.4%, while its hallucination remains between 9–15%. Crucially, deflection does not increase enough to compensate. MISTRAL-SMALL-3.1-24B, for example, raises its deflection rate only marginally (52.0 → 55.9%) while accuracy collapses from 37.5 to 18.3%). This pattern shows that models prefer hallucination over abstention when retrieval is cluttered.

Overall, this experiment demonstrates that retrieval noise, not just retrieval quality, is a fundamental bottleneck. As soon as distractor density increases, both open-weight and proprietary models struggle to separate signal from noise, defaulting to hallucinations instead of deflection, even when explicitly prompted to deflect properly.

Model	Severity	Realistic			Adversarial		
		Acc ↑	Defl ↓	Hall ↓	Acc ↑	Defl ↑	Hall ↓
Keye-1.5-8B	None	58.7	4.7	36.6	4.1	21.0	74.9
	Soft	56.9	8.8	34.3	3.2	37.2	59.7
	Moderate	56.0	9.8	34.2	3.1	43.8	53.1
	Severe	36.2	42.9	20.9	1.5	77.2	21.3
Ovis2-16B	None	57.9	1.0	41.1	4.6	5.7	89.7
	Soft	59.4	2.4	38.2	3.8	13.7	82.5
	Moderate	61.1	3.7	35.2	3.9	24.2	71.9
	Severe	50.9	21.4	27.7	1.7	62.3	36.0
Mistral-Small-3.1	None	57.7	16.9	25.4	2.2	57.6	40.2
	Soft	51.0	27.3	21.7	1.7	69.3	29.0
	Moderate	23.5	61.6	14.9	0.6	83.8	15.6
	Severe	9.2	89.9	0.9	0.2	98.2	1.6
Claude-Opus-4	None	37.7	52.0	10.3	0.6	87.5	11.9
	Soft	38.1	52.0	9.9	0.8	87.0	12.2
	Moderate	32.1	59.4	8.5	0.7	88.3	11.0
	Severe	32.0	62.3	5.7	0.7	92.2	7.1

Table 6: Effect of prompt strictness across **Realistic** and **Adversarial** scenarios for four representative LVLMs. Higher strictness increases deflection rates and reduces hallucination while decreasing accuracy in realistic scenarios.

5 Conclusion

We introduced **VLM-DeflectionBench**, a benchmark for measuring how LVLMs balance accuracy, deflection, and hallucination under varying knowledge conditions. Results show that even top-performing systems hallucinate when evidence is incomplete and over-deflect when refusal cues are too strong, revealing a lack of calibrated confidence. Beyond static datasets, **VLM-DeflectionBench** provides a flexible and model-agnostic framework that can evolve with stronger judging and gating models. By unifying KB-VQA and retrieval-augmented evaluation under a common protocol, it offers a step toward trustworthy multimodal reasoning—where models learn not only to answer correctly but also to know when not to answer.

6 Limitations

We acknowledge several limitations of our work. First, our evaluation relies on GPT-4o with the validated SIMPLEQA protocol as an automatic judge. This enables scalability and consistency, but it cannot fully replace human annotations, especially for nuanced multimodal reasoning. Second, the benchmark inherits biases from its source datasets, which are predominantly text-centric and may underrepresent knowledge that is primarily visual or domain-specific.

Third, our experiments are limited to license-friendly or API-accessible models, which excludes certain families such as some LLaMA-based derivatives. Fourth, LVLMs may show stochastic variability across runs; due to inference costs we report results on a single pass per query, while acknowledging that repeated sampling could yield small fluctuations. Fifth, **VLM-DeflectionBench** focuses on short-form question answering; extending it to long-context reasoning or interactive dialogue remains open for future work. These limitations highlight opportunities for further research without diminishing our central finding that hallucination and deflection remain persistent challenges, especially for Retrieval Augmented Generation models.

Finally, due to the fact that the benchmark is constructed from pre-existing publicly available datasets, we cannot guarantee that offensive and harmful language and images are not included from the source datasets.

Acknowledgments

Bill Byrne holds concurrent appointments as an Amazon Scholar and as Professor of Information Engineering at the University of Cambridge. This paper describes work performed at Amazon.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024a. The revolution of multimodal large language models: A survey. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

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? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Lianghao Deng, Yuchong Sun, Shizhe Chen, Ning Yang, Yunfeng Wang, and Ruihua Song. 2025. Muka: Multimodal knowledge augmented visual information-seeking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. Benchmarking retrieval-augmented multimomal generation for document question answering. *arXiv preprint arXiv:2505.16470*.

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*.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2025. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *Proceedings of the International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, and 1 others. 2024b. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

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*.

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*.

Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A Crook, Yifan Xu, Babak Damavandi, Xin Luna Dong, Christos Faloutsos, Lei Li, and Seungwhan Moon. 2024. Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision*.

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*.

Ionut-Teodor Sorodoc, Leonardo FR Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. 2025. Garage: A benchmark with grounding annotations for rag evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Xin Su, Man Luo, Kris W Pan, Tien Pei Chou, Vasudev Lal, and Phillip Howard. 2025. Sk-vqa: Synthetic knowledge generation at scale for training context-augmented multimodal llms. In *Proceedings of the International Conference on Machine Learning*.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*.

Kelly Tang, Wei-Lin Chiang, and Anastasios N. Angelopoulos. 2025. Arena explorer: A topic modeling pipeline for llm evals & analytics.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. 2024. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *Proceedings of the European Conference on Computer Vision*.

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*.

Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. 2025. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *arXiv preprint arXiv:2502.16636*.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. Lemma: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*.

Yibin Yan and Weidi Xie. 2024. Echosight: Advancing visual-language models with wiki knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, and 1 others. 2025. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, and 1 others. 2024. Crag-comprehensive rag benchmark. In *Advances in Neural Information Processing Systems*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Xu Yuan, Liangbo Ning, Wenqi Fan, and Qing Li. 2025. mkg-rag: Multimodal knowledge graph-enhanced rag for visual question answering. *arXiv preprint arXiv:2508.05318*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix

This appendix provides supplementary materials and extended analyses to complement the main paper. We first present a detailed comparison between our benchmark and prior multimodal question-answering datasets. We then discuss guidelines for refreshing the benchmark, analyze data composition and pipeline convergence, and compare GPT-4o–based automatic judgments with human annotations. Further sections include model-wise performance breakdowns, sensitivity to prompt strictness, implementation details, full prompt templates, and qualitative examples illustrating common failure modes. These materials are intended to ensure transparency, reproducibility, and a deeper understanding of our proposed benchmark.

A Extended Comparison with existing benchmarks

Our benchmark draws on six publicly available datasets, each targeting complementary aspects of knowledge-intensive multimodal reasoning.

InfoSeek (Chen et al., 2023) is a large-scale knowledge-seeking VQA dataset built from Wikipedia. It combines a high-quality human-written set with over 1.3M automatically generated image–question–answer pairs, designed to test fine-grained factual queries about images that require external knowledge rather than surface-level recognition.

Encyclopedic-VQA (Mensink et al., 2023) emphasizes fine-grained, encyclopedic knowledge. It contains over 220k unique question–answer pairs linked to 16.7k categories, each paired with multiple images. Questions are grounded in ∼2M Wikipedia articles, many requiring multi-hop reasoning and providing evidence attribution to specific passages.

WebQA (Chang et al., 2022) simulates real-world web search QA. It consists of ∼46k user-like queries paired with noisy pools of text snippets and images retrieved from the web, including hard distractors. Queries often require multi-hop reasoning across modalities, making it a challenging open-domain multimodal QA benchmark.

MMDocRAG (Dong et al., 2025) targets retrieval-augmented generation over long multimodal documents. It comprises ∼4k annotated QA pairs grounded in 222 documents across diverse domains, interleaving text with tables, charts, and figures. The benchmark evaluates whether models can retrieve fine-grained evidence and generate answers that may integrate both text and visual content.

MRAG-Bench (Hu et al., 2025) is a diagnostic benchmark for vision-centric retrieval-augmented QA. It includes 1.3k multiple-choice questions and a curated corpus of ∼9.6k images. Scenarios are designed around perspective changes and transformations, requiring retrieval of supporting images to answer questions that textual knowledge alone cannot resolve.

ViQuAE (Lerner et al., 2022) is a smaller-scale benchmark (∼3.7k examples) focusing on entity-centric knowledge-based VQA. Questions originate from TriviaQA, with the entity mention replaced by an image. Each example links to the relevant Wikipedia page and section, requiring models to identify the entity visually and retrieve the corresponding fact.

These datasets span encyclopedic, web-based, document-level, vision-centric, and entity-centric settings, providing a diverse testbed for evaluating retrieval-augmented multimodal language models.

B Benchmark Evolution and Maintenance

Although this release of **VLM-DeflectionBench** is fixed and versioned for comparability, the underlying pipeline is designed for **controlled evolution**. Its modular structure supports the inclusion of new models, data sources, and evaluation criteria as the field advances. When stronger models reach high parametric accuracy, they can be incorporated into the gating pool to reapply filtering and remove samples that no longer require retrieval. This procedure keeps **VLM-DeflectionBench** retrieval-dependent and behaviorally informative, while explicit versioning guarantees reproducibility across updates. Implementation details are provided in Section H.

Beyond stability, a central strength of our framework lies in its **flexibility**. The pipeline is not tied to a fixed ensemble or annotation protocol but can be adapted to any retrieval-augmented benchmark. In practice, researchers can substitute the gating models with the strongest available candidates and use the most reliable automatic judge for evaluation. As new LVLMs and RAG systems emerge, this design remains future-proof, since it makes no assumptions about specific architectures or knowledge sources.

This flexibility also enables integration across di-

verse knowledge-grounded tasks. Any KB-VQA or RAG benchmark can be incorporated by re-running the same pipeline with updated models, ensuring consistent retrieval dependence and evaluation semantics. Moreover, because the framework explicitly enforces retrieval grounding, it harmonizes datasets that have historically differed in annotation style, distractor construction, or evaluation criteria.

We view **VLM-DeflectionBench** as a **foundation for unified evaluation** in the KB-VQA community: a shared judge, a principled deflection-aware metric, and an ensemble of retrieval-dependent samples that evolve transparently with model progress. By decoupling benchmark content from model implementation, **VLM-DeflectionBench** remains both reproducible and extensible, able to grow with the field while preserving comparability across generations of LVLMs.

C Data Distribution Analysis

To assess whether our filtering pipeline alters the topical composition of the data, Figure 4 compares sample distributions before and after curation, following Tang et al. (2025). We use BERTOPIC to cluster questions based on embeddings from OPENAI TEXT-EMBEDDING-3-LARGE, reduced via UMAP and grouped with HDBSCAN using cluster probability scores. For interpretability, we select ten representative questions per cluster and use GPT-4o to assign category names and brief descriptions (each defined by the query image and question). The left panel aggregates all six source benchmarks (INFOSEEK, ENCYCLOPEDIC-VQA, WEBQA, MMDOCRAG, VIQUAE, and MRAG-BENCH); the middle panel shows the resulting distribution in **VLM-DeflectionBench**; and the right panel overlays the two. While relative weights shift slightly (*e.g.*, *Identifying Features* and *Food and Drinks* become marginally more frequent), the overall long-tailed shape is preserved. In particular, core categories such as *Object Identification* and *Conservation Status* remain within one to two percentage points of their original prevalence. This analysis confirms that our filtering pipeline does not bias **VLM-DeflectionBench** toward a few sources or topics but maintains a broad and diverse question distribution while enforcing stricter retrieval dependence and grounding fidelity.

D Pipeline Effectiveness and Convergence

Figure 5 illustrates how dataset composition evolves as successive filters are applied in our pipeline. To assess convergence, we monitor three representative LVLMs: KEYE-1.5-8B, OVIS2-16B, and MISTRAL-SMALL-3.1-24B-INSTRUCT under the parametric setting. As expected, accuracy decreases monotonically with each stage, indicating that questions solvable through parametric recall are progressively eliminated. At the same time, deflection increases, showing that the remaining samples require grounding in external evidence rather than model memory. Hallucination rates also rise, reflecting the removal of parametric instances and the retention of more challenging, knowledge-intensive ones. The rightmost panel shows that over 90% of the initial ∼45.9k pooled samples are filtered out, converging to approximately 2.7k. This reduction stems not only from eliminating parametrically answerable items but also from pruning residual noise—ambiguous queries, unanswerable cases even under gold evidence, and false negatives where distractors were mislabeled as positives. Convergence stabilizes after four models, as the inclusion of GPT-4O produces negligible additional filtering. We therefore define this point as the convergence threshold, balancing dataset compactness and source coverage. Despite this aggressive pruning, the final dataset preserves the relative distribution of sources from the original pool (see Figure 4), ensuring representativeness. Overall, these trends confirm that the pipeline effectively suppresses parametric leakage, removes residual noise, and converges to a compact yet diverse dataset where success requires genuine retrieval-based reasoning rather than memorization. The complete workflow is formalized in Algorithm 1.

E Human Validation vs. GPT-4o Judgments

To evaluate the reliability of GPT-4o as an automatic judge, we conducted a human validation study. We sampled 100 responses per model, balanced across the four evaluation scenarios, from four representative LVLMs spanning different scales: KEYE-1.5-8B (small, open-weight), OVIS2-16B (mid-size, open-weight), MISTRAL-3.1-24B-INSTRUCT (large, open-weight with refusal-oriented prompting), and CLAUDE-OPUS-4 (frontier-level proprietary). In total, 400 responses were annotated by a human rater using the

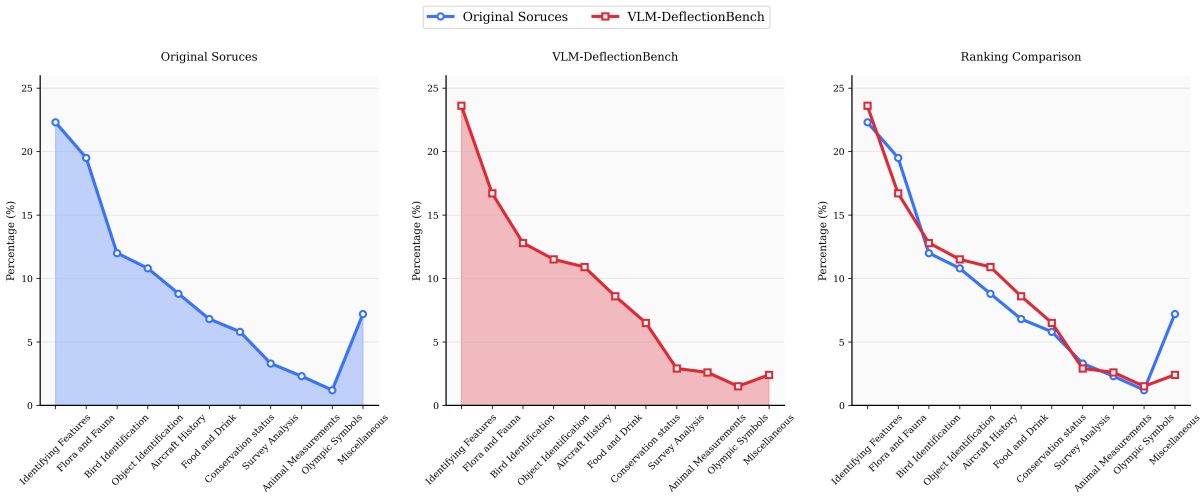


Figure 4: Comparison of category distributions before and after curation. (Left) Aggregate distribution of the six original source benchmarks. (Middle) Final distribution in **VLM-DeflectionBench**. (Right) Overlay showing relative shifts across categories.

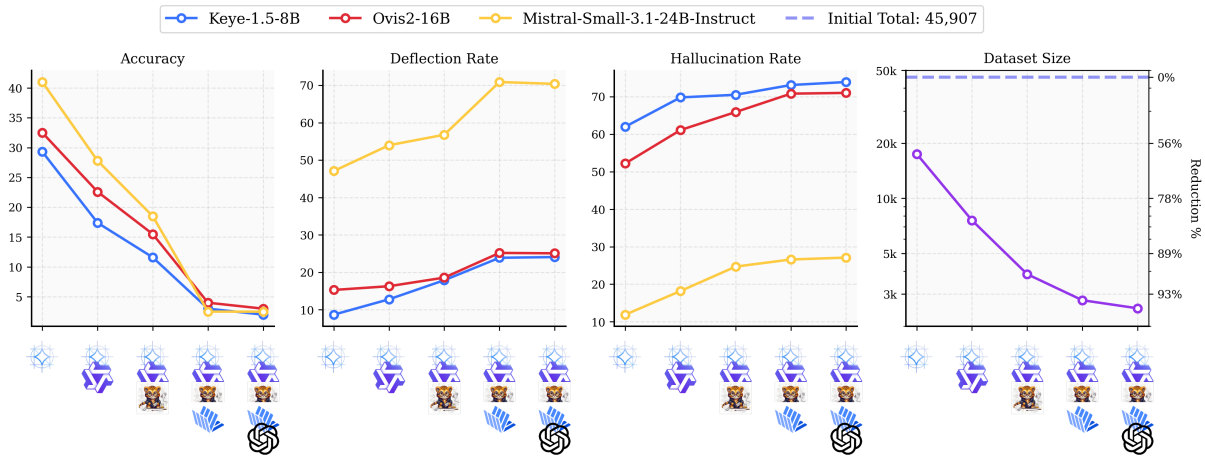


Figure 5: Performance on **Parametric Scenario** and dataset size as models are incrementally added to the filtering pipeline. Convergence is reached after four models, with over 90% of samples removed.

same SIMPLEQA rubric with three labels: CORRECT, INCORRECT, and NOT_ATTEMPTED.

We compared human and GPT-4o labels using two complementary metrics: (i) raw agreement (%), the proportion of identical labels, and (ii) Cohen's κ, which accounts for chance agreement given the label distribution (κ=1 indicates perfect, κ=0 chance-level agreement). Results in Table 7 show consistently high alignment across models, with an average agreement of 92.6% and κ=0.91, corresponding to *almost perfect* agreement. Discrepancies were rare and largely involved borderline cases such as partially correct answers or cautious refusals. These findings confirm that GPT-4o provides a reliable and reproducible basis for large-scale automatic evaluation in **VLM-DeflectionBench**.

F Performance Breakdown by sources

While the main paper reports results averaged across sources, Table 8 provides a more detailed analysis, showing per-dataset performance for four representative LVLMs: KEYE-1.5-8B, OVIS2-16B, MISTRAL-SMALL-3.1-24B-INSTRUCT, and CLAUDE-OPUS-4. Three consistent patterns emerge.

First, performance varies substantially across datasets, revealing heterogeneous difficulty. INFOSEEK is particularly challenging in the parametric setting, where accuracy remains near zero across all models, but improves markedly once gold evidence is available in the oracle scenario. In con-

Algorithm 1: Dynamic Curation Pipeline

Input: Raw samples S_0, Gating models G, Textual KB K_{txt}, Visual KB K_{img}, Threshold $K_{\min}$

Output: Curated benchmark S^*

$S_1 \leftarrow \emptyset$;
for $s \in S_0$ **do**
  $decisions \leftarrow \{J(g(I_s, q_s)) : g \in G\}$;
  if $all(decisions = \texttt{INCORRECT})$ **then**
    $S_1 \leftarrow S_1 \cup \{s\}$;

for $s \in S_1$ **do**
  $K_s^+ \leftarrow$ **gold contexts** ;
  $K_s^- \leftarrow$ **retrieve negatives from** $K_{\text{txt}}, K_{\text{img}}$;
  if $\nexists g \in G : J(g(I_s, q_s, K_s^+)) =$ CORRECT **then**
    discard s; **continue**;
  for $n \in K_s^-$ **do**
    if $\exists g \in G : J(g(I_s, q_s, n)) =$ CORRECT **then**
      remove n ;
  if $|K_s^-| < K_{\min}$ **then**
    discard s ;
  else
    retain s with (K_s^+, K_s^-) ;

return S^* ;

Model	Agreement (%)	Cohen's κ
KEYE-1.5-8B	91.1	0.90
OVIS2-16B	92.4	0.91
MISTRAL-3.1-24B	94.0	0.92
CLAUDE-OPUS-4	93.0	0.92
Average	**92.6**	**0.91**

Table 7: Human validation of GPT-4o judgments (100 samples per model, balanced across scenarios). We report raw agreement (%) and Cohen's κ. Averages are computed across the four models.

trast, WEBQA and ENCYCLOPEDIC-VQA yield higher accuracy under both oracle and realistic conditions, while smaller datasets such as VIQUAE exhibit instability due to limited sample size, producing large metric fluctuations across models and scenarios.

Second, hallucination patterns closely mirror the availability of evidence. In the parametric scenario, hallucination is consistently high as models rely on internal memory. Once oracle passages are supplied, hallucination decreases across all sources, confirming that relevant grounding effectively constrains model generations. However, in adversarial settings, hallucination resurges sharply, demonstrating that carefully designed distractors remain highly effective at misleading LVLMs.

Finally, model ranking remains largely stable across sources. CLAUDE-OPUS-4 and OVIS2-16B maintain higher accuracy and lower hallucination than KEYE-1.5-8B and MISTRAL-SMALL-3.1-24B, though all systems exhibit vulnerabilities in adversarial contexts. Even for strong models, gains on INFOSEEK under oracle grounding are modest, underscoring the intrinsic difficulty of encyclopedic reasoning.

Overall, this breakdown confirms that benchmark difficulty is not uniform: datasets differ both in absolute difficulty and in how models trade off accuracy, deflection, and hallucination. These variations highlight the importance of combining multiple KB-VQA sources to achieve a comprehensive and robust evaluation of retrieval-grounded reasoning.

G Further Effect of Prompt Strictness

Table 9 extends the main analysis of prompt severity by reporting results in the **parametric** and **oracle** scenarios. While the main paper highlighted clear trade-offs in the **realistic** and **adversarial** settings, these additional results reveal how refusal instructions behave in knowledge-free and perfectly grounded contexts. In the **parametric** case, where the target behavior is near-complete deflection, stronger prompts do encourage abstention. For example, CLAUDE-OPUS-4 rises from 32.7% to 97.8% deflection when moving from no prompt to severe prompting, almost eliminating hallucination (54.6% $\rightarrow$ 1.6%). A similar trend holds for MISTRAL-SMALL-3.1, which achieves 99.5% deflection under severe prompting. These results confirm that explicit refusal instructions are highly effective when models truly lack knowledge. In the **oracle** setting, however, severe prompting leads to systematic over-deflection, even when gold knowledge is provided. For instance, OVIS2-16B accuracy falls from 65.0% (soft) to 52.3% (severe), while deflection rises sharply from 2.3% to 25.1%. Similarly, CLAUDE-OPUS-4 loses over 10 points of accuracy under severe prompting, despite having access to perfect evidence. This pattern highlights

Model	Source	# Samples	Parametric			Oracle			Realistic			Adversarial		
			Acc ↑	Defl ↑	Hall ↓	Acc ↑	Defl ↓	Hall ↓	Acc ↑	Defl ↓	Hall ↓	Acc ↑	Defl ↑	Hall ↓
Keye-1.5-8B	Infoseek	1,646	1.6	17.1	81.3	56.7	11.7	31.6	47.1	10.8	42.1	0.9	38.4	60.7
	WebQA	826	5.6	28.6	65.9	79.5	4.4	16.1	70.7	8.6	20.7	6.5	52.7	40.8
	Encyclopedic-VQA	185	3.8	27.0	69.2	77.3	11.9	10.8	73.5	9.7	16.8	2.2	66.5	31.4
	MMDocRAG	58	0	31.0	69.0	69.0	0	31.0	60.3	0	39.7	15.5	12.1	72.4
	MRAG-Bench	45	11.1	4.4	84.4	37.8	8.9	53.3	35.6	6.7	57.8	8.9	26.7	64.4
	ViQuAE	15	6.7	33.3	60.0	46.7	40.0	13.3	53.3	20.0	26.7	6.6	46.7	46.7
	Total	**2,775**	**3.1**	**21.4**	**75.5**	**64.8**	**9.4**	**25.8**	**56.0**	**9.8**	**34.2**	**3.1**	**43.8**	**53.1**
Ovis2-16B	Infoseek	1646	2.2	27.0	70.8	53.4	5.1	41.5	49.3	5.0	45.7	0.5	20.9	78.6
	WebQA	826	7.6	19.1	73.3	86.1	1.0	12.9	82.6	0.2	17.2	8.7	24.9	66.4
	Encyclopedic-VQA	185	3.8	35.7	60.5	79.5	4.9	15.6	80.0	4.9	15.1	5.4	40.5	54.1
	MMDocRAG	58	0	36.2	63.8	62.1	6.9	31.0	65.5	8.6	25.9	5.2	60.3	34.5
	MRAG-Bench	45	11.1	20.0	68.9	11.1	24.4	64.5	13.3	8.9	77.8	11.1	11.3	75.6
	ViQuAE	15	0	13.3	86.7	60.0	0	40	66.7	0	33.3	6.7	6.6	86.7
	Total	**2,775**	**4.0**	**25.2**	**70.8**	**64.4**	**4.2**	**31.4**	**61.1**	**3.7**	**35.2**	**3.9**	**24.2**	**71.9**
Mistral-Small-3.1	Infoseek	1646	1.2	68.3	30.5	38.0	49.7	12.3	26.4	58.9	14.7	0.1	76.4	23.5
	WebQA	826	5.1	73.1	21.8	48.1	45.6	6.3	42.6	52.3	5.1	0.8	96.6	2.6
	Encyclopedic-VQA	185	2.2	81.6	16.2	64.9	28.1	7.0	62.2	30.3	7.5	1.1	88.1	10.8
	MMDocRAG	58	0	79.3	20.7	50.0	37.9	12.1	39.7	44.8	15.5	1.7	91.4	6.9
	MRAG-Bench	45	8.9	66.7	24.4	6.7	66.7	26.6	2.2	91.1	6.7	4.4	91.1	4.5
	ViQuAE	15	0	80.0	20.0	53.3	46.7	0	53.3	46.7	0	13.3	80.0	6.7
	Total	**2,775**	**2.5**	**70.9**	**26.6**	**42.6**	**47.1**	**10.3**	**23.5**	**61.6**	**14.9**	**0.6**	**83.8**	**15.6**
Claude-Opus-4	Infoseek	1646	4.4	59.1	36.5	36.5	53.0	10.5	27.3	64.1	8.6	0.2	84.9	14.9
	WebQA	826	11.3	71.9	16.8	71.2	21.1	7.7	63.8	28.5	7.7	0.8	94.9	4.3
	Encyclopedic-VQA	185	2.7	85.9	11.4	69.2	26.5	4.3	68.2	27.5	4.3	1.1	91.3	7.6
	MMDocRAG	58	3.4	79.3	17.3	53.4	31.0	15.6	51.7	27.6	20.7	3.4	82.8	13.8
	MRAG-Bench	45	20.0	48.9	31.1	11.1	82.2	6.7	11.1	71.1	17.8	6.7	91.1	2.2
	ViQuAE	15	13.3	80.0	6.7	53.3	46.7	0	53.3	33.3	13.4	6.6	66.7	26.7
	Total	**2,775**	**6.7**	**65.0**	**28.3**	**49.1**	**41.7**	**9.2**	**32.1**	**59.4**	**8.5**	**0.7**	**88.3**	**11.1**

Table 8: Breakdown of performance across individual sources for four representative LVLMs. Results are reported in terms of accuracy, deflection, and hallucination under parametric, oracle, realistic, and adversarial settings. The table highlights heterogeneous dataset difficulty, systematic effects of external grounding on hallucination, and consistent relative model ranking across sources.

a key limitation: while severity enforces abstention in knowledge-sparse settings, it also undermines performance when knowledge is available, effectively biasing models toward excessive caution. Overall, these results strengthen our main conclusion that prompt severity is a blunt tool: it can enforce deflection in parametric and adversarial contexts, but at the cost of suppressing accuracy in oracle and realistic ones. Reliable calibration therefore requires methods beyond simple refusal instructions, since current models cannot selectively adapt their abstention threshold to the evidence provided.

Model	Severity	Parametric			Oracle		
		Acc ↑	Defl ↓	Hall ↓	Acc ↑	Defl ↑	Hall ↓
Keye-1.5-8B	None	4.6	4.5	90.9	68.0	2.9	29.1
	Soft	4.1	11.4	84.5	65.8	6.9	27.3
	Moderate	3.1	21.4	75.5	64.8	9.4	25.8
	Severe	0.6	82.3	17.1	37.5	47.0	15.5
Ovis2-16B	None	5.3	9.6	85.1	63.0	1.2	35.8
	Soft	4.0	24.8	71.2	65.0	2.3	32.7
	Moderate	4.0	25.2	70.8	64.4	4.2	31.4
	Severe	0.7	89.5	9.8	52.3	25.1	22.6
Mistral-Small-3.1	None	6.4	33.3	60.3	66.2	14.1	19.7
	Soft	4.5	48.6	46.9	61.0	22.5	16.5
	Moderate	2.5	70.9	26.6	42.6	47.1	10.3
	Severe	0.1	99.5	0.4	13.3	85.8	0.9
Claude-Opus-4	None	12.7	32.7	54.6	52.4	37.0	10.6
	Soft	9.1	49.0	41.9	51.6	37.0	11.4
	Moderate	6.7	65.0	28.3	49.1	41.7	9.2
	Severe	0.6	97.8	1.6	40.0	53.1	6.9

Table 9: Deflection prompt strictness effects in the **Parametric** and **Oracle** scenarios across four LVLMs. Increasing strictness successfully raises deflection rates and reduces hallucination in strong models, but comes at the cost of significant accuracy degradation, revealing fundamental trade-offs in reliability tuning.

H Implementation Details

We employ a CLIP-based encoder (*i.e.*, EVA-CLIP-18B (Sun et al., 2024)) to encode either textual metadata or visual content into dense vectors $w_i \in \mathbb{R}^d$, which are stored in a similarity-search index $W = \{w_i\}_i$. Retrieval is carried out via cosine similarity between query embeddings and indexed features. For each benchmark, we systematically explore both retrieval modalities: (i) **Textual**, where an input image is mapped to Wikipedia text (either titles only (T) or titles plus summaries (T+S)); and (ii) **Visual**, where similarity is computed directly between images (I2I). The retrieval modality with highest performance on each dataset is selected. For **InfoSeek**, we adopt image-to-text retrieval (I2T). The knowledge base is built from

the provided English Wikipedia dump, containing roughly 6M entities. After retrieving the top-10 candidate pages, we further segment each page into 250-token passages with 32-token overlap using CONTRIEVER (Izacard et al., 2021), ensuring well-formed sentence boundaries and more targeted retrieval. For **Encyclopedic-VQA**, we instead rely on image-to-image retrieval (I2I). The knowledge base here is the dataset-provided collection of Wikipedia pages (about 2M entries), which differs from the full 6M-page dump used for InfoSeek. Query images are compared against the first image of each page, and the top-5 retrieved pages are subsequently segmented into 250-token passages with 32-token overlap using the same CONTRIEVER-based procedure. For **MRAG-Bench**, which supplies its own knowledge base, we construct an index from the provided images and perform I2I retrieval of top-20 examples, excluding those overlapping with gold evidence. **ViQuAE** follows the same text-based strategy as InfoSeek (I2T with page segmentation). In contrast, **WebQA** and **MMDocRAG** require no external retrieval, as positive and negative candidates are already provided. Notably, in principle the retrieval process could be reduced to a simple overlap filtering step with gold evidence. However, we design a more complex pipeline to prioritize the generation of challenging and plausible negative contexts, rather than maximizing retrieval precision. This choice ensures that the benchmarks serve their primary purpose: evaluating model robustness under difficult distractor conditions.

I Evaluation Prompts

Figure 6 illustrates the **system prompt** used to constrain models to rely only on the provided context. All evaluations were conducted with the **moderate prompt** (Figure 8), which instructs models to answer only when sufficiently confident. For the ablation experiments on prompt strictness (Tables 6 and 9), we additionally considered the **soft** and **severe** variants (Figures 7 and 9), corresponding to looser and stricter abstention requirements, respectively. These variants allow us to analyze how LVLMs trade off coverage and reliability under different answering conditions.

J Qualitative Analysis and Failure Modes

Figures 10–16 show representative **Oracle** cases (gold evidence provided). The examples surface three recurring behaviors under perfect evidence: (i) *correct grounding* (*e.g.*, OVIS2-16B, CLAUDE-OPUS-4 in Figures 10, 14); (ii) *over-deflection* despite sufficient evidence (*e.g.*, MISTRAL-3.1-24B in Figures 10, 11, and 15); and (iii) *hallucination with gold context* (*e.g.*, OVIS2-16B in Figures 12, 13, and CLAUDE-OPUS-4 in Figure 14, 15).

Textual gold (Figures 10, 11, 14) often enables precise answers, yet some models still refuse (over-deflect) or produce unsupported numbers. **Visual gold** (Figures 12, 13, 15, 16) exposes calibration gaps: models either misidentify fine-grained categories or abstain even when the image suffices. These cases confirm that remaining errors are not due to retrieval quality but to *evidence utilization*: models must better distinguish when gold evidence is adequate to answer versus when abstention is warranted.

System Prompt

```
You are a helpful assistant in a Retrieval-Augmented Generation (RAG) scenario.
You may receive:
- Textual information from retrieved documents
- Visual information from images
- Multimodal content combining both text and images
```

Figure 6: System prompt used in our RAG experiments showing the initial instructions given to the model.

SOFT Uncertainty Level Prompt

```
Answer the question based on the available information. You may abstain if
necessary. Answer:
```

Figure 7: Soft prompt variant encouraging models to answer based on available information with optional deflection when necessary.

MODERATE Uncertainty Level Prompt

```
Please provide an answer only if you feel confident in its accuracy. If you are
uncertain or the available information is insufficient to provide a reliable
answer, please abstain rather than guessing. Answer:
```

Figure 8: Moderate prompt variant instructing models to answer only when sufficiently confident in accuracy, with deflection preferred over uncertain responses.

SEVERE Uncertainty Level Prompt

```
You must provide an answer ONLY if you are completely certain of its accuracy.
If there is ANY doubt or uncertainty, you MUST refuse to answer and state "I
cannot answer this question." Do not attempt partial or uncertain responses.
Answer:
```

Figure 9: Severe prompt variant requiring models to answer only when completely certain, mandating deflection for any uncertainty or doubt.

Source: ENCYCLOPEDIC-VQA

Context: The primary diet of *M. tener* consists of other snakes, primarily earth snakes, and other small fossorial species. It is cannibalistic. It also occasionally eats small lizards, but the consumption of rodents by coral snakes is rare.

Question: What does this reptile eat?
GT Answer: Earth snakes, fossorial species.

OVIS2-16B: This reptile, which appears to be a coral snake, primarily eats other snakes, particularly earth snakes and other small fossorial species. It is also known to be cannibalistic and occasionally consumes small lizards. The consumption of rodents by coral snakes is rare. JUDGE: CORRECT

MISTRAL-3.1-24B: I cannot provide an answer based on the given information because the context provided refers to a different species, *M. tener*, and does not specify the species of the snake in the image. JUDGE: NOT_ATTEMPTED

CLAUDE-OPUS-4: Based on the retrieved information provided, this reptile (which appears to be a coral snake based on its distinctive banding pattern) primarily eats other snakes, especially earth snakes and other small fossorial species. It is also cannibalistic, meaning it will eat other coral snakes. Additionally, it occasionally consumes small lizards, though rodent consumption is rare. JUDGE: CORRECT

Figure 10: Example from ENCYCLOPEDIC-VQA (**Oracle**). All models are given gold textual evidence about the coral snake's diet. OVIS2-16B and CLAUDE-OPUS-4 ground their answers correctly, while MISTRAL-3.1-24B deflects despite sufficient information, illustrating conservative behavior under clear evidence.

Source: ENCYCLOPEDIC-VQA

Context: The agile frog prefers light deciduous mixed forests with plentiful water. The open land around a forest is often also populated, as long as it is connected to the forest by shrubs. In dry, warm forests, this species often also lives far away from the water. Of the three Middle European *Rana* species, this frog likes warmth and dryness the most.

Question: What is the habitat of this animal?
GT Answer: Mixed forests with plentiful water.

OVIS2-16B: The habitat of the agile frog includes light deciduous mixed forests with plentiful water, open land connected to the forest by shrubs, and dry, warm forests. JUDGE: CORRECT

MISTRAL-3.1-24B: I cannot confidently determine the habitat of the organism in the image based on the provided information. The retrieved context discusses the habitat preferences of the agile frog, but there is no indication that the organism in the image is an agile frog. Therefore, I cannot provide an accurate answer. JUDGE: NOT_ATTEMPTED

CLAUDE-OPUS-4: I cannot provide information about the habitat of the animal in the image based on the retrieved context provided. The context discusses the habitat preferences of the agile frog, but I cannot confirm whether the animal in the image is an agile frog or another species. Without being able to definitively identify the animal in the image as an agile frog, I cannot reliably apply the habitat information from the context. JUDGE: NOT_ATTEMPTED

Figure 11: Example from ENCYCLOPEDIC-VQA (**Oracle**). Gold text specifies the habitat; OVIS2-16B answers correctly while MISTRAL-3.1-24B and CLAUDE-OPUS-4 over-deflect under gold evidence.

Source: MRAG-BENCH

Context:

Question: Can you identify the particular model of this car?
GT Answer: Aston Martin Virage Coupe 2012.

OVIS2-16B: Based on the visual characteristics such as the sleek design, the shape of the taillights, and the overall silhouette, the car appears to be an Aston Martin DB9. This identification is made by comparing the visible features with known models from Aston Martin's lineup. However, without a clear view of specific badges or additional identifying marks, this identification cannot be guaranteed with absolute certainty. JUDGE: INCORRECT

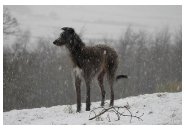
MISTRAL-3.1-24B: I cannot definitively identify the exact model of the car from the image alone. However, based on the design and features visible, it appears to be a high-performance sports car, possibly from a luxury brand. For precise identification, additional details or a clearer view of specific badges or emblems would be necessary. JUDGE: NOT_ATTEMPTED

CLAUDE-OPUS-4: Based on the retrieved images provided, I cannot definitively identify the particular model of this car. While the retrieved images show similar orange/copper-colored luxury sports cars with comparable styling, no text or captions are provided with these images that explicitly identify the specific model. Without additional contextual information such as model names, badges, or descriptive text in the retrieved content, I cannot provide a reliable answer about the particular model of this vehicle. JUDGE: NOT_ATTEMPTED

Figure 12: Example from MRAG-BENCH (**Oracle**, visual gold). OVIS2-16B hallucinates a wrong car model; MISTRAL-3.1-24B and CLAUDE-OPUS-4 abstain. A grounding failure with gold images.

Source: MRAG-BENCH

Context:

Question: Can you identify what creature this is?
GT Answer: Scottish Deerhound.

OVIS2-16B: The creature in the image is a dog, specifically a breed that resembles an Afghan Hound, known for its long, silky coat and distinctive appearance. This identification is based on the visible physical characteristics of the dog, such as its long fur, slender build, and the shape of its ears and muzzle. The setting suggests that the dog may be participating in a dog show, which is a common context for displaying purebred dogs. JUDGE: INCORRECT

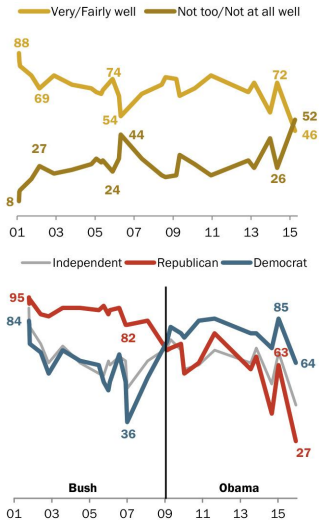
MISTRAL-3.1-24B: The creature in the image is a dog. Specifically, it appears to be a Deerhound, which is a breed known for its shaggy coat and tall, slender build. JUDGE: CORRECT

CLAUDE-OPUS-4: Based on the retrieved images provided, I can see several dogs that appear to be similar to the creature in the input image. The retrieved images show various dogs with long, shaggy, grey coats that match the characteristics of the animal in the question image. However, without any text or specific breed information provided in the retrieved context, I cannot identify the specific breed of dog this is. I can only confirm based on the visual similarity to the retrieved images that this appears to be a dog with a distinctive long, grey, shaggy coat. JUDGE: NOT_ATTEMPTED

Figure 13: Example from MRAG-BENCH (**Oracle**, visual gold). MISTRAL-3.1-24B identifies the breed correctly; OVIS2-16B hallucinates; CLAUDE-OPUS-4 over-deflects. Calibration varies despite gold images.

Source: MMDocRAG

Context (Images):



Context (Text):

1. The latest national survey by Pew Research Center, conducted Dec. 8–13 among 1,500 adults, finds that since the start of this year, the share of Americans who say the government is doing well in reducing the threat of terrorism has fallen by 26 percentage points – from 72% to 46%.

2. Compared to early 2015, assessments of government efforts to combat terrorism are more negative across the political spectrum. Democrats are now the only partisan group in which a majority (64%) say the government is doing at least fairly well (down from 85% in January). Independents' positive ratings have dropped 25 points, from 69% to 44%. And just 27% of Republicans now say the government is doing very or fairly well reducing the terrorist threat, down from 63% at the beginning of the year.

Question: In 2015, how many percentage of surveyed adults, Republicans, and Democrats believed that the government was doing very/fairly well in reducing the threat of terrorism? Please write the answer in list format, *e.g.*, [3, 2]

GT Answer: ["46", "27", "64"].

OVIS2-16B: ["46%", "27%", "64%"] JUDGE: CORRECT

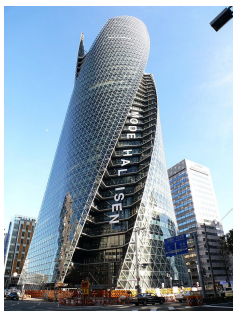
MISTRAL-3.1-24B: I cannot provide an answer based on the given information because the data provided is from December 2015, not from earlier in the year. JUDGE: NOT_ATTEMPTED

CLAUDE-OPUS-4: Based on the retrieved information, I can provide the percentages for early 2015 (January): 72%, 63%, and 85%. Therefore, the answer in list format is: ["72", "63", "85"]. JUDGE: INCORRECT

Figure 14: Example from MMDocRAG (**Oracle**). Gold text supports exact numeric extraction: OVIS2-16B is correct; MISTRAL-3.1-24B over-deflects; CLAUDE-OPUS-4 hallucinates outdated figures.

Source: WEBQA

Context (Images):



Context (Text):
NO CONTEXT

Question: How many letters are found on the outside of the Nagoya Mode Academy Spiral Towers?
GT Answer: 11 letters are found on the outside of the Nagoya Mode Academy Spiral Towers.

OVIS2-16B: There are 12 letters found on the outside of the Nagoya Mode Academy Spiral Towers.
JUDGE: INCORRECT

MISTRAL-3.1-24B: I am unable to determine the number of letters on the outside of the Nagoya Mode Academy Spiral Towers based on the provided information.
JUDGE: NOT_ATTEMPTED

CLAUDE-OPUS-4: Based on the retrieved image, I can see letters displayed vertically on the outside of the spiral tower building. The visible letters spell out "MODE GAKUEN." Counting these letters, there are 10.
JUDGE: INCORRECT

Figure 15: Example from WEBQA (**Oracle**, visual gold). Counting letters from the gold image: OVIS2-16B and CLAUDE-OPUS-4 hallucinate, MISTRAL-3.1-24B over-deflects.

Source: WEBQA

Context (Images):

Context (Text):
NO CONTEXT

Question: Which is taller: the '50 Ford F-Series or Lotus Elise?
GT Answer: The '50 Ford F-Series is taller.

OVIS2-16B: Based on the images provided, the '50 Ford F-Series appears to be taller than the Lotus Elise. The Ford F-Series has a higher roofline and overall height compared to the Lotus Elise, which has a lower profile typical of sports cars.
JUDGE: CORRECT

MISTRAL-3.1-24B: I cannot determine which vehicle is taller based solely on the images provided.
JUDGE: NOT_ATTEMPTED

CLAUDE-OPUS-4: Based on the provided images, I cannot determine which vehicle is taller. While I can see that the Ford F-Series is a truck with a higher roofline and the Lotus Elise is a low-slung sports car, the lack of scale or shared reference prevents reliable comparison.
JUDGE: NOT_ATTEMPTED

Figure 16: Example from WEBQA (**Oracle**, visual gold). With two gold images, OVIS2-16B correctly infers the taller object; MISTRAL-3.1-24B and CLAUDE-OPUS-4 over-deflect.