Multimodal UNcommonsense: From Odd to Ordinary and Ordinary to Odd

Yejin Son[1*], Saejin Kim[1*], Dongjun Min[1], Youngjae Yu[2]
[1]Yonsei University [2]Seoul National University
Correspondence:
yejinhand@yonsei.ac.kr, jerry0110@yonsei.ac.kr, youngjaeyu@snu.ac.kr

Abstract

Commonsense reasoning in multimodal contexts remains a foundational challenge in artificial intelligence. We introduce **Multimodal UNcommonsense (MUN)**, a benchmark designed to evaluate models' ability to handle scenarios that deviate from typical visual or contextual expectations. MUN pairs visual scenes with surprising or unlikely outcomes described in natural language, prompting models to either rationalize seemingly odd images using everyday logic or uncover unexpected interpretations in ordinary scenes. To support this task, we propose a retrieval-based in-context learning (R-ICL) framework that transfers reasoning capabilities from larger models to smaller ones without additional training. Leveraging a novel **Multimodal Ensemble Retriever (MER)**, our method identifies semantically relevant exemplars even when image and text pairs are deliberately discordant. Experiments show an average improvement of 8.3% over baseline ICL methods, highlighting the effectiveness of R-ICL in low-frequency, atypical settings. MUN opens new directions for evaluating and improving visual-language models' robustness and adaptability in real-world, culturally diverse, and non-prototypical scenarios.

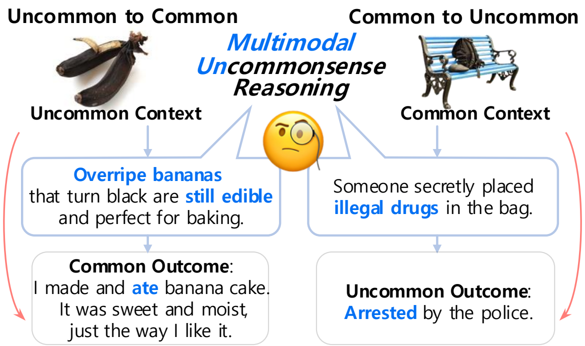


Figure 1: Multimodal UNcommonsense Reasoning aims to produce explanations that make given outcomes appear likely. For example, overripe bananas (an uncommon context) can still be used for baking a sweet, moist banana cake (a common outcome), while a bag on a bench (common context) leads to an arrest (uncommon outcome). This highlights the challenge of bridging visual cues with logical reasoning, as addressed in our Multimodal Uncommonsense (MUN) dataset.

1 Introduction

In everyday life, commonsense functions as an invisible framework, akin to "dark matter" in the universe. Though we cannot directly perceive it, commonsense subtly influences our decisions, such as recognizing social norms or interpreting ambiguous situations (Bosselut et al., 2019; Tao et al., 2024).

While this commonsense often leads to stable reasoning in familiar contexts, it can falter when confronted with visual or textual cues that fall outside typical experience or cultural familiarity (Wang et al., 2023). However, most existing benchmarks for commonsense reasoning evaluate models on frequent or prototypical cases that are well-covered by large-scale English-language corpora and standard vision-language datasets (Brown et al., 2020b; Raffel et al., 2020; Hendricks et al., 2016; Agrawal et al., 2017; Li et al., 2019). As a result, current AI systems exhibit significant brittleness when faced with rare, ambiguous, or culturally-specific phenomena that lie beyond the training distribution.

To address this gap, we introduce the **Multimodal UNcommonsense (MUN)** benchmark: a human-curated dataset specifically designed to evaluate models' ability to reason about **uncommon or counterintuitive outcomes** in visual contexts. Unlike prior datasets that emphasize prototypical commonsense, MUN centers on visual situations that violate typical expectations, such as an overripe banana being preferable for baking, or culturally specific gestures like the Indian head wobble

*Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17586–17609
November 4-9, 2025

Visual Context	Textual Context	Outcome	Explanation
	Black spots in a banana.	Person enjoyed the banana bread without any health concerns.	The overripe bananas were likely used for baking banana bread, which was enjoyed without any health issues.
	Air bubbles in freshly poured concrete.	Person observed the concrete setting well, providing a solid foundation for construction.	The air bubbles in freshly poured concrete, a common occurrence during mixing, did not compromise its integrity, as the concrete set properly to form a solid foundation.
	A blue backpack resting on a wooden bench.	Arrested by the police.	While the owner went to the restroom, someone placed illegal drugs in the bag without their knowledge.
	A laptop and coffee cup on a café table.	Lost money due to fraud.	Someone attempted hacking through an unsecured Wi-Fi connection, stealing personal information and committing fraud.

Figure 2: MUN examples. The first two examples are from MUN-vis and the next two examples come from MUN-lang; explanations are written by human annotators. Note that textual context is only used during dataset generation.

indicating agreement. These cases challenge the model to reconcile visual oddities with logically or culturally grounded explanations.

The necessity of MUN lies in its focus on **low-frequency, multimodal reasoning**, a facet critical for real-world applications where visual inputs and commonsense expectations often diverge. By constructing and evaluating against such examples, MUN serves as a benchmark that complements existing datasets, expanding the scope of commonsense evaluation beyond conventional boundaries.

We collected human-written and LLM-generated explanations for each case, revealing a significant gap in interpretability and diversity. While LLM explanations are often precise, human annotations offer a broader range of perspectives, as noted in prior work (Zhao et al., 2023). We leverage both via augmentation to build a high-quality benchmark that supports rich supervision and evaluation.

To enhance reasoning in visually and contextually atypical scenarios, we adopt a retrieval-based in-context learning (R-ICL) method (Lin et al., 2022, 2023), which improves smaller models by leveraging semantically relevant exemplars generated by larger models. Specifically, to effectively identify these exemplars within scenarios exhibiting visual-textual discordance, we introduce an innovative retrieval framework known as the *Multimodal Ensemble Retriever* (MER). MER independently scores similarity in each modality and fuses them via a tunable weighting mechanism, enabling flexible retrieval in the presence of intentionally discordant image-text pairs in MUN.

Unlike conventional retrievers that assume strong cross-modal alignment, MER accommodates the unaligned nature of our benchmark. To the best of our knowledge, this is the first application of R-ICL in a setting where visual and textual signals are deliberately discordant, enabling abductive reasoning over unaligned multimodal inputs. This approach yields an average 8.3% increase in win rate over a random baseline, demonstrating the effectiveness of R-ICL in boosting nuanced multimodal reasoning.

By connecting intuitive impressions with underlying truths in visually uncommon scenarios, MUN lays the groundwork for building trustworthy AI systems capable of reasoning beyond the obvious across cultures, contexts, and expectations.

2 Related Work

Abductive Reasoning. Abductive reasoning, central to commonsense, involves inferring the most plausible explanations from incomplete observations. While various efforts have explored textual and multimodal abductive reasoning, each existing approach exhibits limitations (Table 1). For example, Abductive-NLI (Bhagavatula et al., 2019) focuses solely on textual input in everyday scenarios without visual grounding. Sherlock (Hessel et al., 2022) integrates real images and text but remains constrained to common situations and unidirectional reasoning. UNcommonsense (Zhao et al., 2023) targets uncommon contexts but lacks visual signals, while NL-Eye (Ventura et al., 2024) employs synthetic images without adequately address-

Dataset	Modality	Real image	Uncommon	Bi-direction Reasoning?
Abductive-NLI	T	✗	✗	✗
Sherlock	I+T	✓	✗	✗
Uncommonsense	T	✗	✓	✗
NL-Eye	I+T	✗	✗	✗
MUN (Ours)	I+T	✓	✓	✓

Table 1: Comparison with Abductive reasoning benchmark. (Bhagavatula et al., 2019; Hessel et al., 2022; Zhao et al., 2023; Ventura et al., 2024) "I" stands for Image and "T" stands for Text. The MUN uniquely supports "Bi-direction UNcommonsense Reasoning," combining unusual contexts, outcomes, and nuanced visual scenarios.

ing non-commonsensical scenarios or bidirectionality. As a result, none of these existing approaches simultaneously incorporate real imagery, handle uncommon contexts, and support bidirectional abductive inference. In contrast, our proposed MUN (Multimodal UNcommonsense) dataset integrates real images and text, actively considers uncommon scenarios, and enables bidirectional reasoning, thereby addressing these gaps and offering a more comprehensive and nuanced abductive reasoning benchmark.

Retrieval-Augmented and In-Context Learning Recent advances in large language models (LLMs)(Brown et al., 2020a; Chowdhery and et al., 2022; OpenAI, 2023) and large vision-language models (VLMs)(Alayrac et al., 2022; Li et al., 2023a) have shown remarkable zero- and few-shot learning capabilities. However, their reasoning often remains tied to patterns entrenched in their training data. Retrieval-augmented paradigms (Liu and et al., 2022b; Thoppilan and et al., 2022) and in-context learning (ICL) techniques (Wei and et al., 2022; Zhou and et al., 2022) represent promising strategies to extend model capabilities beyond memorized knowledge. By dynamically incorporating external documents, exemplars, or contextual cues, models can handle more complex reasoning tasks and adapt to new domains. In visual domains, multi-source retrieval (Zhu and et al., 2020; Liu and et al., 2022a) and retrieval-based image grounding show potential. Our work aligns with this trend by using a retrieval-based ICL approach. We retrieve both textual and visual exemplars from MUN scenarios, guiding model reasoning and distilling complex abductive and cultural logic into accessible formats. This approach assists smaller VLMs in navigating unusual scenarios and producing coherent, contextually rich explanations.

3 Multimodal UNcommonsense (MUN)

To advance research in Visual Uncommonsense Reasoning, we have constructed the benchmark **Multimodal UNcommonsense (MUN)**, created to challenge models with scenarios that diverge from standard visual or contextual expectations. Inspired by prior work (Zhao et al., 2023) on uncommonsense reasoning, specifically abductive reasoning about unusual situations, our dataset adopts a structured *context-results-explanation* paradigm. In this framework, models are required to interpret an image-based context along with a textual scenario (the results) and then generate an explanation that reconciles the two.

3.1 Task Settings

We focus on two complementary task settings that emphasize the delicate interplay between visual cues and textual reasoning.

MUN-vis: Uncommon Image (Context) → Common Results In this task, the model is presented with an image that initially appears visually peculiar or “uncommon,” representing situations that occur with low frequency or probability. Despite this apparent strangeness, the goal is to generate a coherent explanation that normalizes the scenario and demonstrates that it is actually common or perfectly reasonable. For instance, in the first row of Figure 2, a photograph of a blackened banana might initially seem unusual. However, the outcome states, "Person enjoyed the banana bread without any health concerns," indicating that an explanation such as "The bananas were overripe and therefore used for baking banana bread, which was enjoyed without any health issues" is needed to bridge the gap between the context and the outcome. This task involves generating explanations that connect seemingly peculiar visual inputs to familiar and logical everyday contexts.

MUN-lang: Common Image (Context) → Uncommon Results In this scenario, the model is presented with an image that appears completely ordinary but must explain an unusual or "uncommon" textual outcome associated with it. In Figure 2, the context depicted in the third row shows a seemingly typical scene of a blue backpack resting on a wooden bench, while the outcome is "Arrested by the police," which does not naturally align with the given context. The explanation must bridge this gap by uncovering less obvious details, such as "While the owner was in the restroom, someone secretly placed illegal drugs in the bag without their knowledge," providing a surprising yet plausible rationale to make sense of the discordant situation.

3.2 Dataset Creation

We constructed the MUN dataset through a multi-step process, generating diverse “uncommonsense” scenarios that challenge multimodal reasoning models.

Scenario Generation We used GPT-4o to produce a diverse range of textual scenarios(contexts). For MUN-vis, we instructed the model to depict scenes initially appearing visually odd but ultimately normal. For MUN-lang, we asked for ordinary-looking scenes that conceal surprising rationales. Our prompting strategy encouraged the model to analyze hypothetical image-text pairs, classify them as “normal” or “anomalous,” and provide brief explanations. By varying visual and contextual cues and highlighting underlying reasons, we obtained scenarios rich in cultural context, sensory detail, and conceptual twists. This approach guided GPT-4o to produce structured, logically grounded explanations. In MUN-vis entries, seemingly strange images were normalized by uncovering rational backstories. In MUN-lang entries, mundane appearances were reinterpreted through hidden surprises or unconventional practices.

Filtering for Ensuring Diversity To ensure a diverse dataset, we implemented a comprehensive filtering process after generating a large pool of candidate scenarios. Observing numerous similar scenarios, we prioritized removing them to promote diversity and minimize redundancy. Using the Dedupe library,[1] A specialized tool for data deduplication, we effectively eliminated duplicates.

Inspired by diversity filtering (Han et al., 2023), we further enriched the diversity of contexts by identifying a list of specific keywords. Examples were filtered out if the language description of an image contained any of these keywords. To maintain balance, we ensured that the occurrence of these keywords in the contexts remained below 20.

Image Pairing and Selection Process For each textual scenario(context), we first retrieved five candidate images using the Bing Web Search API[2], then manually reviewed them to select the image that best reflected the scenario’s uniqueness or ordinariness. If suitable images were not found through automated searches, we conducted additional manual searches to identify appropriate options.

By incorporating real-world images, the model can achieve more stable and generalizable reasoning capabilities, as demonstrated by research on ALBEF(Li et al., 2021), BLIP(Li et al., 2022), and LLaVA(Liu et al., 2023a), as well as large-scale, diverse image resources like LAION-5B(Schuhmann et al., 2022). Building datasets grounded in authentic visuals enables expansion to cover rare situations and cultural nuances. Through iterative refinement, this approach surpasses existing limitations and supports more nuanced cross-domain reasoning.

Human Explanation Generation We recruited 26 graduate students specializing in computer science and artificial intelligence as annotators to participate in the primary explanation-writing tasks. All participants were proficient in English, and the interface and instructions were provided in English. To ensure a fair and efficient workflow, the tasks were divided into small batches, with the workload evenly distributed among the annotators. This approach prevented any single annotator from being overburdened, thereby maintaining the consistency and quality of the dataset. Additionally, to enhance the contextual reliability of the dataset, annotators were instructed not to write explanations for scenarios they deemed irrelevant or inappropriate. This measure prevented the inclusion of unnecessary or non-essential explanations. Furthermore, annotators were encouraged to logically infer and articulate the reasons behind outcomes that appeared mismatched within the provided visual context.

LLM-Enhanced Human-Written Explanations As shown in subsequent analysis, and consistent with previous studies (Zhao et al., 2023), human-written explanations demonstrate the diversity and broad understanding, while LLM-generated responses tend to be relatively narrow and specific. We aim to combine these complementary strengths to further refine human annotations. Specifically, we use carefully crafted prompts to guide GPT-4o in improving human-written explanations, enabling

[1] `https://github.com/dedupeio/dedupe`.

[2] The Bing Web Search API:`https://www.microsoft.com/en-us/bing/apis/bing-web-search-api`.

LV	Human(%)	LLM(%)	Human+LLM(%)
1	30.5	0.3	1.3
2	40.1	8.9	9.3
3	8.6	35.4	20.5
4	11.9	55.0	62.6
5	8.9	0.3	6.3
Avg.	**2.29**	**3.46**	**3.63**

Table 2: Comparison of the specificity of explanations written by humans (Human), explanations generated by LLMs (LLM), and human-written explanations enhanced by LLMs (LLM+Human). Each value in the table represents the proportion of explanations rated at each specificity level (1 to 5) in percentile.

it to present clearer and more specific logical connections between visual scenarios and the given uncommon outcomes. This process preserves the diversity and nuance of human explanations while leveraging the precision of LLMs, resulting in an improved set of explanations that provide a more informative baseline for comparison.

3.3 Data Analysis

The MUN dataset includes two subtasks: MUN-vis with 515 instances of visually uncommon contexts and common outcomes, and MUN-lang with 500 instances of visually common contexts and uncommon outcomes, totaling 1,015 visual context-outcome pairs. Human explanations were collected for 143 instances from MUN-vis and 156 from MUN-lang, with LLM-generated explanations for all pairs.

Diversity of MUN The MUN dataset spans a broad range of scenarios across various categories, with each example including detailed textual explanations linking visual context to outcomes. While certain categories may be emphasized, individual examples still capture complex, multilayered scenes. For detailed reports on the frequencies of topics and their combinations, see Appendix I.

The t-SNE(Van der Maaten and Hinton, 2008) visualization (Figure 3) reveals that textual contexts cluster into distinct groups, covering a wide array of subjects.

3.4 Comparison Analysis of Explanations

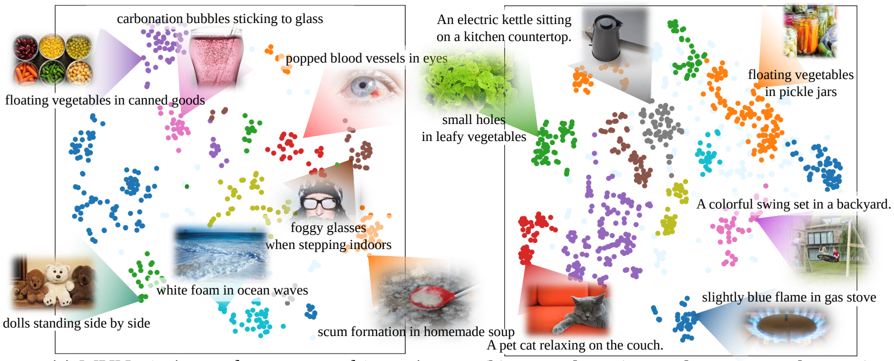
Consistent with Uncommonsense (Zhao et al., 2023), there were noticeable differences in both the length and lexical diversity of explanations generated by LLM, Human, and Human+LLM. Figure 4 illustrates the distribution of explanation lengths. In the **MUN-vis** task, human explanations were relatively long and variable, averaging 32.0 ± 38.5 tokens. LLM explanations, on the other hand, were shorter and more stable at 25.1 ± 4.4 tokens, whereas Human+LLM explanations were longer at 50.7 ± 37.2 tokens, offering more detailed content. In the **MUN-lang** task, humans produced shorter explanations (16.3 ± 7.9 tokens), while LLM outputs were longer and more consistent (44.5 ± 6.4 tokens). This pattern suggests that, in more open-ended tasks like MUN-lang, LLMs produce richer and longer explanations, whereas in more structured tasks like MUN-vis, humans tend to provide longer descriptions. Human+LLM explanations reached 44.6 ± 13.1 tokens, approaching LLM-level length while combining human creativity with LLM stability.

To quantify lexical diversity, we measured n-gram entropy ($n \in \{1, \dots, 5\}$) as shown in Figure 5, conducting 1,000 bootstrap iterations[3]. In **MUN-vis**, human explanations displayed higher n-gram entropy than LLM explanations, and Human+LLM exceeded human entropy, reflecting a synergy where human variability and LLM precision were combined. LLM explanations showed lower entropy, possibly due to the task's structured nature. In **MUN-lang**, LLM entropy was similar to or even higher than Human+LLM's and significantly exceeded that of humans, indicating that LLMs employ more diverse wording in open-ended tasks, whereas human language use is more constrained. Human+LLM still maintained high entropy, effectively blending human creativity and LLM rigor.

Recent work (Stiennon et al., 2020; Liu et al., 2023b) suggests that LLMs can reliably evaluate qualitative aspects of text, such as specificity, given well-structured prompts. Following this approach, we employed GPT-4o to evaluate the specificity (scores 1 to 5).

Table 2 shows that human explanations had a high proportion (70.6%) of low specificity (scores 1 to 2) and a relatively low proportion (20.8%) of high specificity (scores 4 to 5). LLM explanations generally maintained moderate to high specificity (scores 3 to 4), with a large proportion of 4-point ratings (55.0%), but very few achieved the highest specificity (0.3% for score 5). In contrast, Human+LLM had an even higher proportion of 4-point ratings (62.6%) and improved the proportion of 5-point ratings (6.3%), thereby maximizing overall specificity. This demonstrates that LLMs

[3]In each iteration, one explanation was randomly selected per context-outcome pair from each subset.



(a) MUN-vis (normal text, anomal image) (b) MUN-lang (anomal text, normal image)

Figure 3: t-SNE visualization of MUN-vis (a) and MUN-lang (b) based SimCSE (Gao et al., 2021) across categories.

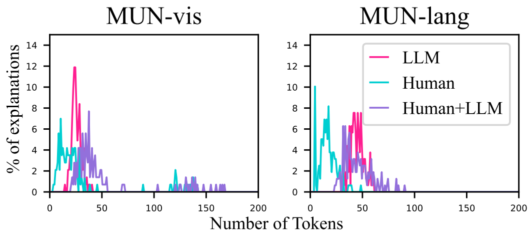


Figure 4: Explanation token length distributions in MUN: The left section represents MUN-vis, while the right section depicts MUN-lang, derived from calculations on the development sets of each data subset.

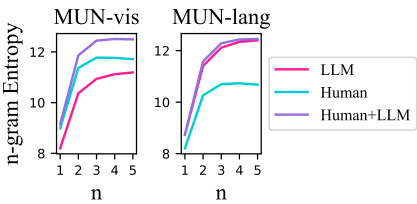


Figure 5: The n-gram distribution entropies for MUN-vis (left) and MUN-lang (right) were calculated based on the development sets for each data subset.

can refine and expand upon human input, achieving a higher level of detail and specificity, and that a Human+LLM approach can combine the strengths of both while compensating for their respective weaknesses. Based on these results, Human+LLM was deemed to be the best and was selected as the baseline for evaluation.[4]

4 Visual Uncommonsense Reasoning with Retrieved In-Context Learning

The core aim of the MUN dataset is to challenge models with atypical, low-frequency visual-text scenarios that resist conventional commonsense interpretations. Unlike standard benchmarks, MUN focuses on “uncommonsense” reasoning, where the model must infer nuanced, often abductive rationales for unusual outcomes. This pushes beyond the straightforward pattern matching that large vision-language models (VLMs) typically excel at due to their extensive pretraining on statistically dominant patterns.

However, when confronted with such atypical scenes, models tend to regress to high-probability patterns learned during pretraining, producing literal or overly generic captions rather than abductive explanations (Hessel et al., 2022). To enable more contextually grounded reasoning in these unfamiliar scenarios, we adopt retrieval-based in-context learning (ICL) to surface relevant yet semantically non-obvious examples.

Conventional retrieval methods, which assume strong alignment between modalities, often struggle with the intentional divergence of image-text pairs in MUN. To address this, we introduce a *Multimodal Ensemble Retriever* (MER) that scores image and text similarities separately and combines them via a tunable fusion mechanism. This approach enables MER to retrieve semantically coherent examples even when the visual and textual cues signal distinct or conflicting commonsense expectations.

Specifically, MER embeds (image, text) pairs using a CLIP-style image encoder and a BERT-based text encoder, and computes cosine similarity between the query and dataset entries for each modality. The two similarity scores are then integrated using a weighting coefficient α that balances the contribution of each modality. This separate-but-aligned retrieval strategy allows MER to flexibly accommodate modality-specific signals, providing a principled mechanism for bridging conceptual

[4]Further analysis of how humans perceive differences between human-written and LLM-generated or LLM-augmented responses is provided in the Appendix D.

Dataset	Model	0-shot	1-shot		3-shot		5-shot	
			Rand.	R-ICL	Rand.	R-ICL	Rand.	R-ICL
MUN vis	*Gemma3*	0.335	0.428	0.457	0.422	**0.491**	0.393	0.382
	InternVL 2.5	0.243	0.121	0.254	0.208	0.312	0.387	**0.434**
	LLaVA OV	0.301	0.439	0.405	0.434	0.376	**0.474**	0.445
	Phi 3.5*v*	0.283	0.324	0.335	0.387	0.387	0.428	**0.445**
	Phi 4*mm*	0.410	0.312	0.272	0.393	0.387	**0.630**	0.618
	Qwen2.5 VL	0.364	0.428	0.387	0.422	0.428	0.439	**0.538**
	Qwen2 VL	0.225	0.399	0.405	0.283	0.376	0.272	**0.486**
MUN lang	*Gemma3*	0.257	0.341	0.418	0.430	0.454	0.498	**0.546**
	InternVL 2.5	0.325	0.273	0.293	0.361	0.357	0.430	**0.470**
	LLaVA OV	0.285	0.301	0.325	0.333	0.369	0.365	**0.369**
	Phi 3.5*v*	0.337	0.353	0.329	0.410	0.390	0.430	**0.442**
	Phi 4*mm*	0.357	0.502	0.582	0.534	0.554	0.651	**0.655**
	Qwen2.5 VL	0.422	0.353	0.329	0.357	0.357	0.410	**0.426**
	Qwen2 VL	0.349	0.349	0.321	0.365	0.341	**0.365**	0.357

Table 3: Comparison of models in different shot settings, measured by winning ratio against human-assisted explanations(higher is better). "Random" indicates randomly chosen examples, and "R-ICL" indicates retrieved examples for in-context learning. Model outputs were compared with Human+LLM explanations, judged using LLM.

Model	LR	LC	LE	CS
Gemma3	3.16 (+0.12)	3.59 (+0.15)	4.12 (+0.14)	3.59 (+0.14)
InternVL 2.5	2.83 (+0.45)	3.40 (+0.52)	3.79 (+0.63)	3.33 (+0.54)
LLaVA-OV	3.21 (+0.07)	3.77 (+0.07)	4.13 (+0.04)	3.84 (+0.11)
Phi 3.5v	3.23 (+0.14)	3.75 (+0.09)	4.16 (+0.11)	3.77 (+0.15)
Phi 4mm	3.31 (+0.21)	3.85 (+0.29)	4.11 (+0.25)	3.82 (+0.28)
Qwen2.5 VL	3.29 (+0.08)	3.87 (+0.13)	4.25 (+0.07)	3.89 (+0.12)
Qwen2 vl	3.28 (+0.11)	3.86 (+0.14)	4.22 (+0.11)	3.94 (+0.15)

Table 4: Effect of retrieval–based in-context selection on flask-based skill metrics(higher is better). LR stands for Logical Robustness, LC for Logical Correctness, LE for Logical Efficiency, and CS stands for Commonsense. Each cell shows the R-ICL score with the gain over the random baseline in parentheses.

gaps in visually grounded abductive reasoning.

To the best of our knowledge, this is the first application of such a dual-scoring retrieval framework in a setting where the visual and textual modalities are intentionally discordant. The full formulation and algorithmic implementation are provided in Appendix B.

5 Experiments

We evaluate the effectiveness of our proposed retrieved in-context learning (ICL) approach for multimodal uncommonsense reasoning using the MUN dataset. Our experimental study is organised to shed light on two core questions:

RQ1. How does the number of in-context examples (shots) affect model performance?

RQ2. What is the impact of retrieval-based examples compared to randomly selected ones?

To establish robust baselines and ensure comprehensive evaluation, we benchmark several state-of-the-art vision-language models (VLMs) and utilize a multimodal ensemble retriever for our retrieval-based ICL approach.

5.1 Models Selection and Retrieval Mechanism

We evaluate seven interleaved VLMs spanning different architecture families and size scales: Qwen2-VL (Wang et al., 2024), Qwen2.5-VL (Bai et al., 2025), Phi-3.5-vision (Abdin et al., 2024), Phi-4-multimodal (Abouelenin et al., 2025), InternVL-2.5 (Chen et al., 2024), Gemma3 (Team et al., 2025), LLaVA-Onevision (Li et al., 2024). To support retrieved ICL, we use a multimodal ensemble retriever combining textual and visual inputs. BERT-based text encoder (Xiao et al., 2023) encodes and retrieves text examples based on query outcomes, while a CLIP-based image encoder (Radford et al., 2021) handles images. The ensemble merges similarity scores from both modalities with hyperparameters α assigned to 0.4. For experiments, we created a database with 372 and 344 image-scenario pairs from MUN-vis and MUN-lang, which lack human label explanations and are not used for testing. For baseline comparisons, we implement standard in-context learning (ICL) prompts where examples are randomly chosen from the MUN dataset, irrespective of their relevance to the query.

5.2 Experimental Setup

Varying the Number of In-Context Examples. To investigate how the number of in-context examples affects model performance, we vary the number of retrieved exemplars (from 1, 3, to 5) provided to the models. This setup allows us to assess the scalability of the ICL approach and determine the optimal number of examples for effective reasoning.

Retrieval-Based vs. Randomly Selected Shots. To evaluate the importance of retrieval quality, we compare our retrieval-based ICL with a baseline ICL approach that uses randomly selected examples from the MUN dataset.

As for the metric, we adopted the Alpaca-Eval framework (Li et al., 2023b) to evaluate the quality of the generated explanations by comparing them against human and LLM-generated explanations. Specifically, we prompt GPT-4o to rank the explanations produced by different models against the Human+LLM explanations. The ranking assesses the coherence, relevance, and abductive reasoning quality of the model-generated explanations.

5.3 Results

Table 3 summarizes the performance of the selected VLMs under various in-context example configurations and retrieval strategies. We report results for two tasks, MUN-vis and MUN-lang, to capture both visual and linguistic reasoning quality.

RQ1. Effect of Shot Scaling. All models generally improve as the number of in-context examples increases from zero to 1, 3, and 5 shots, especially when using R-ICL. Across the seven VLMs the median gain is **+6.1 pp** on *vis* and **+7.2 pp** on *lang*. While MUN-lang also benefits from more examples and R-ICL, the improvement in MUN-vis is generally more pronounced, highlighting that visual reasoning gains more from the effective selection and increased number of in-context examples.

RQ2. Retrieval *vs* Random. Comparing the two columns under each shot size in Table 3 shows that R-ICL beats random selection in **12 of 14** model-dataset combinations. The stronger gains on *vis*(+0.040) over *lang*(+0.023) confirm that supplying semantically aligned image exemplars is particularly helpful for visual reasoning. Notable examples include *InternVL 2.5* (+13 pp at 1-shot) and *Qwen2.5-VL* (+9.9 pp at 5-shot) on *vis*; improvements on *lang* are positive but smaller (e.g., *Phi-4mm* +3.4 pp at 3-shot). Appendix C offers a detailed qualitative analysis of these patterns and examples in the Qualitative Results section.

5.4 Analysis of Model Response

To rigorously quantify the contribution of R-ICL to the logical-reasoning capacities of VLMs, we conducted two complementary analyses: an automated evaluation and a human analysis.

Automatic evaluation In the automated, skill-based evaluation framework proposed by FLASK (Ye et al., 2023), four complementary criteria are considered: Logical Robustness (LR), Logical Correctness (LC), Logical Efficiency (LE), and Commonsense Understanding (CS). Each scored on a 1-to-5 scale by GPT-4o using the rubric in the FLASK frameworks. As summarized in Table 4, retrieval-based in-context selection yields consistent improvements on every metric for every model. The largest absolute gains are observed for *InternVL-2.5* (+0.63 LE; +0.54 CS), while even the smallest gains remain positive across all four skills. These findings indicate that supplying semantically relevant exemplars not only strengthens deductive reasoning but also enhances commonsense inference, further underscoring the role of context quality in in-context learning.

Setting	MUN-vis (%)	MUN-lang (%)
Zero-shot	16	36
Random 5-shot	12	44
Retrieval 5-shot	32	56

Table 5: **Human preference win rates (%).** Percentage of cases (out of 50 samples per modality) where human annotators preferred the model-generated explanation over the LLM+Human.

Human evaluation Table 5 presents the results of a human evaluation comparing the reasoning quality of *phi-4-mm* across different shot settings. For both the vision and language modalities, 50 samples were randomly selected, and human annotators were asked to compare the model's responses, generated under zero-shot, random 5-shot, and retrieved 5-shot settings, against human-assisted explanations. For each sample, the annotators selected the response they judged to be more coherent and convincing. The reported values represent the winning ratio, i.e., the proportion of cases in which the model's output was preferred over the human-assisted explanation. While some variance exists due to limited sample size, the overall trend, particularly in human evaluation, suggests that increasing the number of in-context examples, especially through retrieval, generally leads to improved reasoning performance.

6 Conclusion

We introduce the Multimodal UNcommonsense (MUN) dataset to evaluate how vision-language models handle atypical scenarios that challenge commonsense reasoning. Extensive experiments show that retrieved in-context learning (ICL) examples, rather than randomly chosen ones, enhance model performance. By bridging unexpected visual cues with logical explanations, we successfully guide models to produce more coherent, contextually aligned reasoning. This approach enables more adaptive and reliable multimodal AI systems that are better equipped to understand uncommon events, cultural nuances, and low-frequency phenomena in real-world settings.

7 Acknowledgements

This research was supported by Hyundai Motor Company and Kia Corporation. It was also funded by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program, Yonsei University) and by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (Nos. RS-2024-00354218 and RS-2024-00353125). We also used AI assistants to refine the writing style and for preliminary coding assistance.

8 Limitation

While MUN provides a valuable benchmark for evaluating multimodal uncommonsense reasoning, it is not without shortcomings. First, while the dataset benefits from meticulous human curation that enhances per-sample quality, this comes at the cost of scale, potentially limiting its representation of the broader variability found in real-world scenarios and may not capture the full breadth of cultural, environmental, or domain-specific complexities.

Second, our retrieval-based in-context learning approach, while effective, relies on the quality and diversity of available exemplars; overly domain-specific or homogeneous retrieval sets could limit the generalizability of results.

Additionally, the current approach relies on post hoc evaluations with language models to assess explanation quality, which may introduce biases or yield incomplete metrics for reasoning capabilities. Subsequent efforts might also integrate multi-turn interactive reasoning processes, allowing models to clarify ambiguities before producing their final explanations. Advances in automated evaluation metrics could provide more objective assessments of abductive reasoning quality.

Moreover, combining retrieval-based techniques with model fine-tuning or parameter-efficient adaptation strategies may yield more robust and domain-transferable reasoning systems. Ultimately, pursuing these directions can further strengthen the utility, fairness, and resilience of multimodal AI models in handling complex and atypical scenarios.

9 Ethical Considerations

Dataset Construction. The dataset was constructed using images sourced from the web and carefully filtered to minimize inappropriate, sensitive content. All images were reviewed by annotators following a strict set of guidelines to ensure that the dataset does not propagate bias, stereotypes, or harmful cultural depictions.

Cultural and Contextual Reasoning. The reasoning tasks presented in MUN encourage models to produce abductive explanations grounded in cultural and contextual knowledge. This raises the possibility that models might inadvertently generate content that reflects implicit biases or culturally insensitive narratives. We emphasize the importance of using diverse sets of evaluators and retrieval corpora to mitigate these risks and improve fairness and inclusivity. Researchers, developers, and users are encouraged to apply adversarial testing and ongoing monitoring to identify and address any unintended harm.

Responsible Applications and Safeguards. Lastly, while the improved reasoning capabilities we pursue may have beneficial real-world applications, from more accurate image analysis in healthcare to a better understanding of global cultural phenomena, they also open the door to more sophisticated image and text manipulation. It is crucial that developers implement robust guardrails, transparency measures, and user consent mechanisms to ensure that these advanced reasoning techniques serve the public interest responsibly, respecting privacy, cultural values, and intellectual property rights.

References

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

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel

Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2451–2461.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *Preprint*, arXiv:1906.05317.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and et al. 2020a. Language models are few-shot learners. In *NeurIPS*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Aakanksha Chowdhery and et al. 2022. Palm: Scaling language modeling with pathways. In *arXiv preprint arXiv:2204.02311*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. 2023. Reading books is great, but not if you are driving! visually grounded reasoning about defeasible commonsense norms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 894–914, Singapore. Association for Computational Linguistics.

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision (ECCV)*, pages 3–19.

Jack Hessel, Jingkang Zhao, Ranjay Krishna, Angel Chang, and Yonatan Bisk. 2022. Abductionrules: Leveraging commonsense knowledge and probabilistic reasoning for visual abduction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1451–1463.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

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

Junnan Li, Dongxu Li, Xiaohua Xie, Mark Yatskar, and Steven C.H. Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Junnan Li, Ramakrishna Vedantam Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven C.H. Hoi. 2021. ALBEF: Aligning vision and language with momentum contrast. In *NeurIPS*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang, and Zhenyu Chi. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. 2022. Unsupervised cross-task generalization via retrieval augmentation. *Advances in Neural Information Processing Systems*, 35:22003–22017.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. LLaVA: Large language and vision assistant. *Preprint*, arXiv:2302.06675.

Yanan Liu and et al. 2022a. Cpt: A pre-trained unbalanced transformer for both chinese and english language generation. In *Findings of ACL*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yixin Liu and et al. 2022b. Rer: A retrieval-augmented model for knowledge-intensive nlp tasks. In *ACL*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Niki Matena, Niki Parmar, Yanqi Liu, and Alex Jolicoeur-Martineau. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Christoph Schuhmann, Lyonel Beaumont, Romain Vencu, Cade W. Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Kartikay Katta, Chris Mullis, Roman Kaczmarczyk, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Preprint*, arXiv:2210.08402.

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*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M Ziegler, Ryan Lowe, Casey Voss, Alec Radford, Dario Amodei, and Ilya Sutskever. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

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

Rami Thoppilan and et al. 2022. Lamda: Language models for dialog applications. In *arXiv preprint arXiv:2201.08239*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Mor Ventura, Michael Toker, Nitay Calderon, Zorik Gekhman, Yonatan Bitton, and Roi Reichart. 2024. Nl-eye: Abductive nli for images. *arXiv preprint arXiv:2410.02613*.

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

Xingbo Wang, Renfei Huang, Zhihua Jin, Tianqing Fang, and Huamin Qu. 2023. Commonsensevis: Visualizing and understanding commonsense reasoning capabilities of natural language models. *IEEE Transactions on Visualization and Computer Graphics*.

Jason Wei and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Wenting Zhao, Justin T Chiu, Jena D Hwang, Faeze Brahman, Jack Hessel, Sanjiban Choudhury, Yejin Choi, Xiang Lorraine Li, and Alane Suhr. 2023. Uncommonsense reasoning: Abductive reasoning about uncommon situations. *arXiv preprint arXiv:2311.08469*.

Dian Zhou and et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *NeurIPS*.

Song-Chun Zhu and et al. 2020. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. In *AAAI*.

A Experiment Details and Hyperparameter

Table 6 shows the hyperparameters of the models we used in the experiments and the exact model checkpoints used in the experiments are reported in Table 7. All experiments, except those involving GPT-4o, were conducted using two NVIDIA A6000 GPUs.

Hyperparameter	**Configuration**
Text emb. model	`BAAI/bge-large-en`
Image emb. model	`clip-vit-base-patch16`
Image resolution	512×512
Ensemble ratio α	0.4
Retrieval lib.	`langchain` (`https://python.langchain.com/docs/introduction/`)
Vector DB lib.	FAISS (Douze et al., 2024)
VLLM lib.	VLLM (Kwon et al., 2023)

Table 6: Hyperparameter configurations for the main experiment.

Open-source Models	
Gemma 3	`google/gemma-3-4b-it`
InternVL 2.5	`OpenGVLab/InternVL2_5-8B`
LLaVA-OneVision	`llava-hf/llava-onevision-qwen2-7b-ov-hf`
Phi 3.5-Vision	`microsoft/Phi-3.5-vision-instruct`
Phi 4-Multimodal	`microsoft/Phi-4-multimodal-instruct`
Qwen 2.5-VL	`Qwen/Qwen2.5-VL-7B-Instruct`
Qwen 2-VL	`Qwen/Qwen2-VL-7B-Instruct`
Closed-source Models	
GPT-4o	`gpt-4o-2024-11-20 (via OpenAI API)`

Table 7: Model checkpoints used in our experiments. Open-source models were accessed via Hugging Face, and the closed-source model (GPT-4o) was accessed via the OpenAI API.

B Bi-Encoder Retrieval Mechanism

To retrieve relevant in-context examples for uncommonsense reasoning, we use a bi-encoder retrieval strategy that computes and fuses modality-specific similarity scores. First, we embed the (image, text) pairs stored in the dataset $D_{(i,t)}$ using a CLIP-style image encoder E_I and a BERT-based text encoder E_T, respectively. Given a user query $q = (q_i, q_t)$, we compute cosine similarities between the query vectors (v_{q_i}, v_{q_t}) and stored vectors (v_i, v_t). The final similarity score is obtained by weighting the image and text similarities using a tunable coefficient α that controls the relative contribution of each modality.

Algorithm 1 Ensemble Retrieval Method (Bi-encoder). This computes cosine similarities in visual and textual embedding spaces, fuses them by α, and returns the top k matches.

Input: $q = (q_i, q_t)$; number of retrievals k, weight ratio α
Output: list of top k retrieved (d_i, d_t) pairs
Vector database $D_{(i,t)}$ containing (image, text) pairs
Image encoder E_i and text encoder E_t
Convert query to vectors: $v_q = (v_{q_i}, v_{q_t}) = (E_i(q_i), E_t(q_t))$
Initialize `Results` $\leftarrow$ [], `Indices` $\leftarrow$ []
for each (v_i, v_t) with index j in $D_{(i,t)}$ **do**
    Compute similarity:
    $s = \alpha \cdot \text{cos\_sim}(v_{q_i}, v_i) + (1 - \alpha) \cdot \text{cos\_sim}(v_{q_t}, v_t)$
    Append s to `Results`
    Append j to `Indices`
end for
Combine `Results` and `Indices` into pairs and sort by similarity in descending order
Select top k indices as `TopKIndices`
Initialize `TopK` $\leftarrow$ []
for each j in TopKIndices **do**
    Append (d_{i_j}, d_{t_j}) to `TopK`
end for
return `TopK`

C Qualitative results

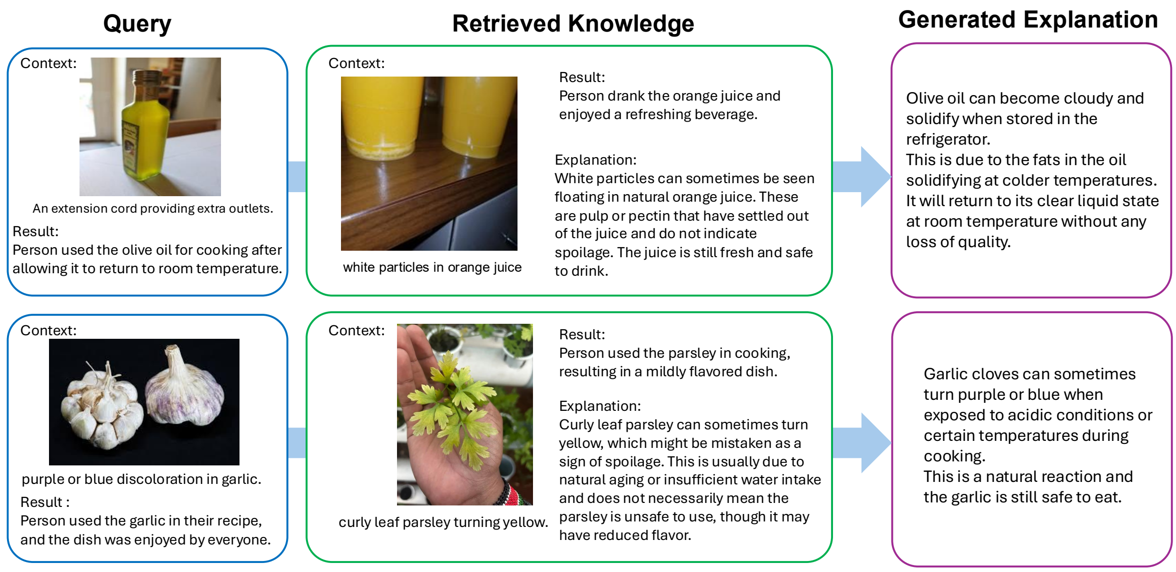


Figure 6: MUN-vis qualitative results.

Figure 6 illustrates the model's capacity to retrieve contextual knowledge and produce precise, explanatory answers in MUN-vis. For the first row, when queried about the haze that develops in refrigerated olive oil, the model draws an analogy to the white flecks that appear in orange juice. In both cases, low temperatures cause constituents to congeal and aggregate: fats solidify in olive oil, while pulp- and pectin-rich particles clump together in orange juice. Once the liquids return to room temperature, they clarify, showing that neither product's quality is compromised. This example demonstrates how the system enhances its explanatory power by juxtaposing uncommon yet analogous phenomena across different contexts.

Figure 7 highlights the model's ability to retrieve contextual knowledge and generate precise, explanatory responses in MUN-lang. For the first row, while the power strip appears normal at first glance, the outcome of 'smoke and emergency evacuation' necessitates the model to abductively infer an 'inherent risk of overheating' within the power strip. To facilitate this inference, the retrieved few-shot examples include scenarios such as 'residents evacuating due to smoke from a cozy fireplace'. Despite depicting

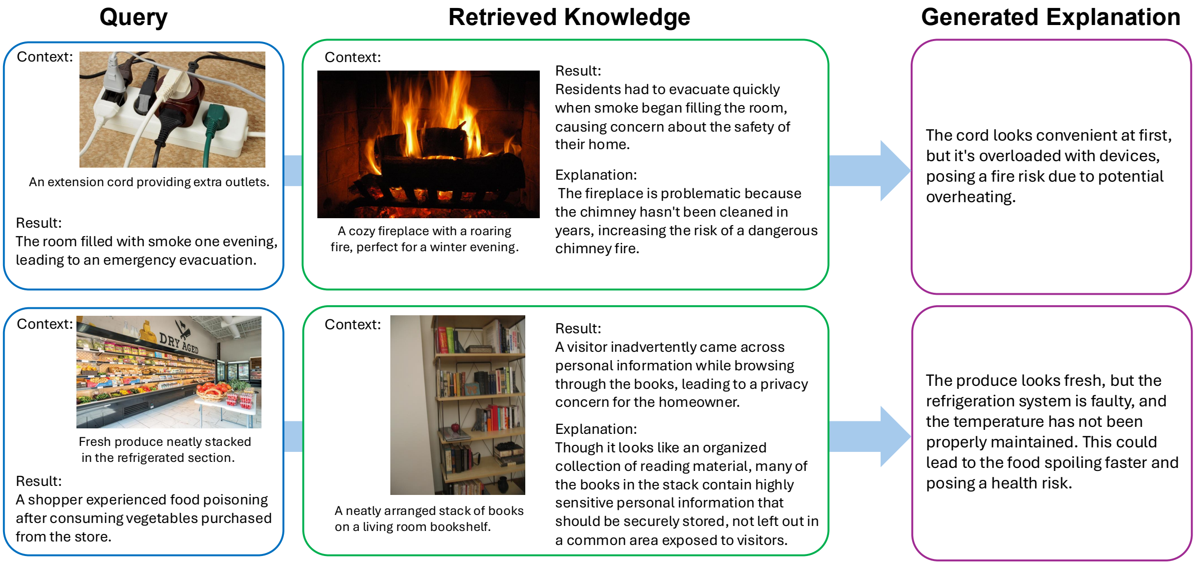


Figure 7: MUN-lang qualitative results.

different visual subjects, it induces a common causal pattern: 'an ostensibly normal object contains a hidden fire hazard'. This connects to everyday experience-based abductive reasoning that 'appliances generating heat pose a fire hazard', helping the model to infer potential dangers beyond what is visually apparent.

As the qualitative examples make clear, MUN-vis and MUN-lang probe two orthogonal yet complementary facets of uncommonsense reasoning. In MUN-vis, an uncommon visual cue must be normalized via specific commonsense knowledge (e.g., "cloudy refrigerated olive oil" $\leftrightarrow$ "white flecks in chilled orange juice"), whereas MUN-lang inverts the challenge: a common visual scene masks an anomalous outcome, demanding abductive reconstruction of hidden risks (e.g., "benign-looking power strip" $\rightarrow$ "concealed fire hazard" $\rightarrow$ "evacuation"). Together, these tracks enforce a balanced assessment of a model's ability to anchor striking images to everyday facts and infer unseen causal mechanisms behind unexpected events. By integrating both dimensions into a single benchmark and leveraging MER's targeted retrieval of concrete analogues or abstract causal templates, MUN provides a comprehensive framework for evaluating models, spanning from concrete commonsense grounding to abstract causal inference.

D Comparison of Human Agreement on Explanations

We conduct a human evaluation comparing human-written explanations against those generated by two types of models: LLM (directly generated by the model) and HLLM (either model-generated or model-augmented based on human-written content). As summarized in Table 12, there is no significant decline in the perceived quality of responses generated by the LLM. Specifically, 70.8% of LLM-generated explanations achieved higher than moderate agreement with human-written explanations, while 71.6% of HLLM explanations reached this level of agreement. These results indicate that model-generated and model-augmented explanations can closely match human-written ones in terms of response quality.

E Evaluation of GPT4o on MUN dataset

We have conducted GPT-4o's performance on our dataset with a similar setup as the sec 5, which shows strong performance across both mun-vis and mun-lang, with generally similar performance improvement trends with open-source models. However, we excluded GPT-4o from our initial experiments due to the well-documented "self-preference bias" where LLMs tend to favor their own generated answers and attach our results in the appendix.

Dataset	1-shot		3-shot		5-shot	
	Rand.	R-ICL	Rand.	R-ICL	Rand.	R-ICL
MUN vis	0.572	0.597	0.604	0.610	0.673	0.704
MUN lang	0.678	0.636	0.671	0.664	0.650	0.692

Table 8: Evaluation of GPT4o on different shot settings, measured by winning ratio against human-assisted explanations(higher is better). "Random" indicates randomly chosen examples, and "R-ICL" indicates retrieved examples for in-context learning. Model outputs were compared with Human+LLM explanations, judged using LLM.

α	0.3	0.4	0.5	0.6	0.7
Winrate	0.572	0.618	0.611	0.611	0.603

Table 9: Ablation study on hyperparameter alpha on MUN vis with Phi4-mm model.

Dataset	Model	0-shot	1-shot		3-shot		5-shot	
			Rand.	R-ICL	Rand.	R-ICL	Rand.	R-ICL
MUN lang	*Gemma3*	0.259	0.364	**0.399**	0.294	0.301	0.259	0.287
	Phi3.5v	0.273	0.329	**0.336**	0.287	0.329	0.315	0.287
	Phi4mm	0.371	**0.497**	0.448	0.455	0.427	0.448	**0.497**
	Qwen2.5VL	0.364	0.322	0.357	0.287	**0.378**	0.273	0.322
MUN vis	*Gemma3*	0.333	0.233	0.289	0.226	**0.365**	0.233	0.314
	Phi3.5v	0.075	0.126	0.176	0.113	**0.201**	0.170	0.157
	Phi4mm	0.195	0.239	0.214	0.132	**0.277**	0.101	0.170
	Qwen2.5VL	0.170	0.164	**0.358**	0.176	0.327	0.208	0.302

Table 10: Comparison of models in different shot settings, measured by winning ratio against human-assisted explanations, judged by opensource LLM. "Random" indicates randomly chosen examples, and "R-ICL" indicates retrieved examples for in-context learning. Model outputs were compared with Human+LLM explanations, judged using opensource LLM(Llama-4-Scout).

F Evaluation of haperparameter alpha

Table 9 shows the effects of different hyperparameter α on performance on the MER on a MUN-vis subset with Phi4-mm model. Based on findings in Table 9, we have used an α value of 0.4 during the main experiments.

G Evaluation with an Open-Source Judge

To verify that the performance benefits of our R-ICL method are robust and not dependent on a single proprietary evaluator, we have evaluated with the state-of-the-art open-source Llama-4-Scout model as a judge model for comparison between model outputs and Human+LLM explanations. As tab 10 confirms that the central trend observed in the main experiments holds. While absolute win rates differ due to the new evaluator's distinct preferences, our Retrieval-Augmented In-Context Learning (R-ICL) consistently outperforms or remains highly competitive with zero-shot and random few-shot baselines across most models and settings (achievements highlighted in bold).

H Evaluation of MER on other opensource benchmarks

To provide empirical evidence for the generalizability of our MER framework, we conducted a preliminary experiment on the A-OKVQA benchmark (Schwenk et al., 2022). We tested the accuracy of Qwen-2.5-VL on 500 randomly selected multiple-choice questions from the validation set with 5000 samples from the training set acting as ICL context. Table 11 demonstrates that R-ICL improves accuracy over both zero-shot and random-shot baselines. The performance gain on A-OKVQA, a task requiring both visual

Sampleing Mode	Accuracy
Zero shot	0.818
Random 1 shot	0.832
R-ICL 1 shot	**0.842**

Table 11: Ablation study on MER method on A-OKVQA datasets with Qwen-2.5-VL.

understanding and external knowledge, strongly suggests that MER's ability to retrieve relevant context is a generalizable principle.

Level	**LLM**		**HLLM**	
	Cnt	**%**	**Cnt**	**%**
1	136	13.6	123	12.3
2	162	16.2	161	16.1
3	194	19.4	204	20.4
4	221	22.1	255	25.5
5	287	28.7	257	25.7
Avg.	**3.36**		**3.36**	

Table 12: Distribution of human agreement levels (out of 1000 samples each) for LLM vs. Human and HLLM vs. Human responses. The average score is computed assuming Level 1 to 5 correspond to scores from 1 to 5.

I Dataset Categories

Categories	**MUN-vis**	**MUN-lang**
Household Items and Furniture	100	300
Beverages	82	22
Fruits and Vegetables	80	8
Tools, Equipment	57	143
Dairy Products and Eggs	54	0
Health and Personal Care	44	15
Canned, Packaged, and Processed Goods	36	5
Meat and Seafood	22	2
Condiments and Sauces	21	0
Grains, Bread, and Baked Goods	19	5
Total	**515**	**500**

Table 13: Comparison of object category counts across textual description of visual context. Total counts for each dataset are provided in the last row.

We selected the top 30 most frequent categories based on the textual context of MUN-vis and MUN-lang. As shown in Table 13, MUN-vis focuses more on food-related elements, while MUN-lang emphasizes household and furniture items. However, the subsets still feature diverse subcategories and context-rich scenes at the example level, as illustrated in Figures 8.

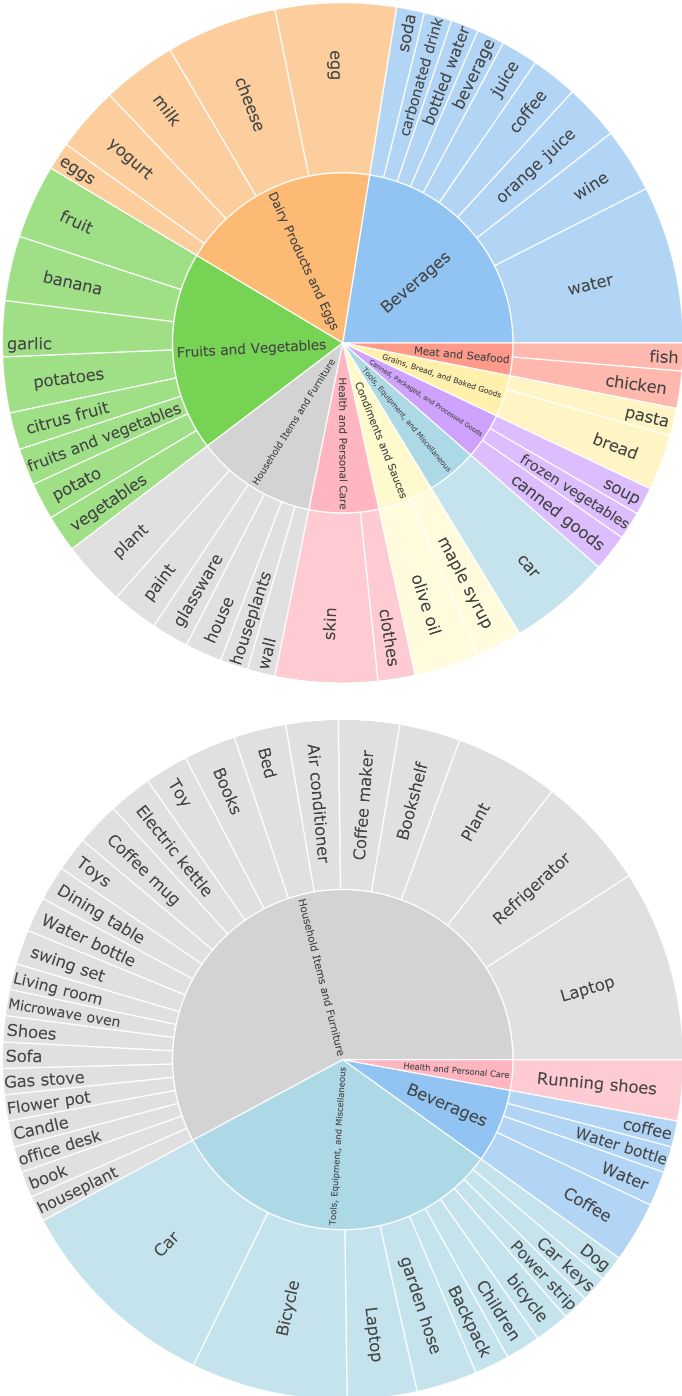


Figure 8: Textual context distribution in (top) MUN-vis and (bottom) MUN-lang.

J Prompt used during Experiments

Figure 12 through 17 illustrate the various prompts used during dataset generation and evaluation.

K Human Annotation Details

K.1 Human Dataset Construction

To construct the human-written dataset, we recruited 26 graduate students to generate contextualized explanations that logically bridge two provided segments. Annotators were instructed to follow a standardized interface that guided the construction of fluent and coherent connecting sentences. Each explanation was written with reference to the surrounding context to ensure narrative consistency. The interface used for collecting human explanations is illustrated in Figure 9.

K.2 Human Evaluation Protocol

We conducted two human evaluation studies via the Prolific platform[5], recruiting participants whose first language is English.

(1) Human Agreement Evaluation. To assess alignment between human and model-generated outputs, we asked annotators to compare two anonymized responses for each of 500 randomly selected samples, across two comparisons: (a) LLM vs. Human and (b) HLLM vs. Human. Each sample was evaluated by two independent annotators, resulting in a total of 2,000 judgments. A total of 141 unique participants were recruited for this task, and workers were compensated at a rate of €7.50 per hour. The interface used for collecting human agreement on explanations is illustrated in Figure 10.

(2) Win Rate Comparison. We further evaluated relative response quality across few-shot prompting variants (zero-shot, random 5-shot, retrieved 5-shot) using a win-rate setup. For each of 50 representative samples, we constructed 3 pairwise comparisons (e.g., retrieved vs. zero-shot), resulting in 150 comparisons per modality. This evaluation was conducted separately for MUN-LANG and MUN-VIS, yielding a total of 300 pairwise comparisons. Each comparison was rated by a single annotator. A total of 20 unique participants were recruited for this task, and they were compensated at a rate of €7.71 per hour. The interface used for collecting win-rate judgments is shown in Figure 11.

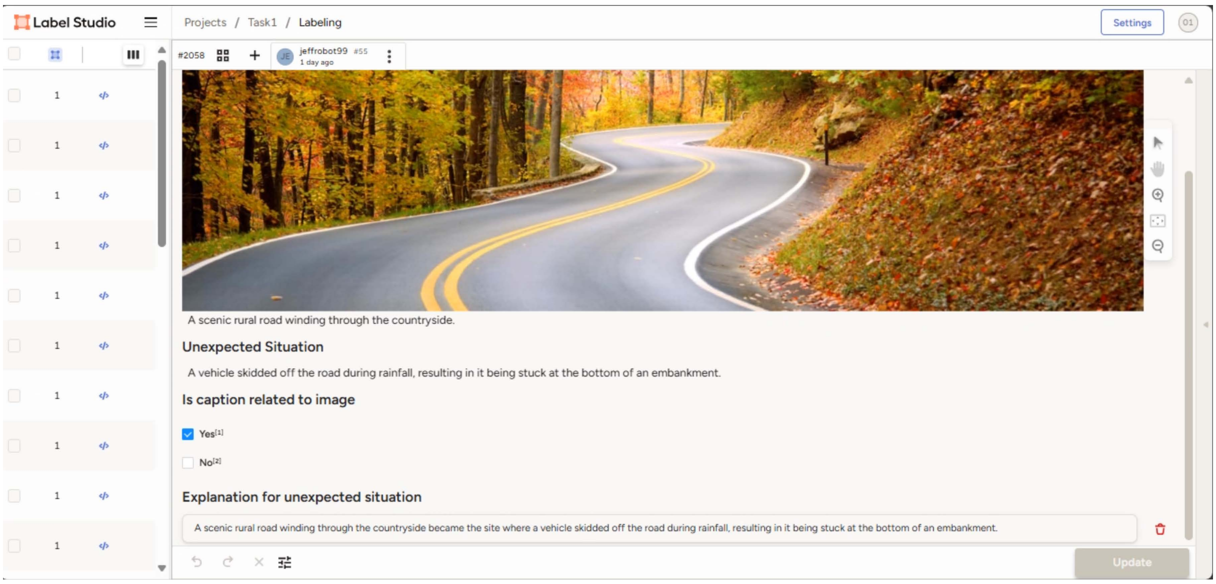


Figure 9: The user interface used for human annotation tasks, designed to facilitate the creation of detailed and contextually relevant explanations in MUN-vis and MUN-lang.

[5] https://www.prolific.com

Common Instruction (applies to all questions): Evaluate how closely the explanations below align with typical human reasoning (logic and quality). Explanation A and Explanation B are randomly assigned between Machine-generated and Human + LLM explanations (human-written, enhanced by AI). Their exact sources are masked. Rate how similar each explanation (A and B) is to human reasoning.

Description of the image: A pantry stocked with various food items and snacks.

Outcome of the image: An individual experienced food poisoning symptoms after consuming items from the pantry.

Human Explanation: The air condition was bad there, too wet and hot inside. The foods inside there got worse only in few days.
A: The pantry's poor air condition, characterized by excessive humidity and heat, caused the food items stored inside to spoil quickly, leading to the individual experiencing food poisoning symptoms after consuming them.
B: The pantry seems well-supplied, but it is problematic because many of the food items are past their expiration dates, posing health risks if consumed.

Below are possible explanations for the image and its outcome. Please rate how much you agree that each explanation is **similar to the human explanation** in terms of logic and quality.

	1 Very Low Agreement	2 Low Agreement	3 Moderate Agreement	4 High Agreement	5 Very High Agreement
A	○	○	○	○	○
B	○	○	○	○	○

Figure 10: Interface used for human agreement evaluation. Annotators were presented with two anonymized responses, one written by a human, and the other either directly generated by an LLM or revised by an LLM based on the human version, and asked to select the more appropriate one.

We are conducting a survey to evaluate the performance of different large-language models. You will be presented with the instructions given to the models and the responses from two different models. Your task is to determine which model's response you believe would be preferred by most people.

Here is the instruction given to the models:

You are tasked with rationalizing an unexpected outcome where each entry consists of the following components:

Image: A single image that contains a scene.

Unlikely Outcome: Unexpected outcome based on the scene of the image.

Rationale: A plausible reasoning explaining why the situation might happen from the image.

Guidelines for Output: The Rationale must provide a plausible reason why the outcome happened from the scene. Use clear and concise language.

Now your task:

Based on the provided Image and Unlikely Outcome, generate the corresponding Rationale following the structure and format above.

Image: slightly wrinkled skin on peppers

Unlikely Outcome: Person used the peppers in a stir-fry and enjoyed a flavorful meal.

Rationale:

Please choose the better explanation between A and B based on the quality of their responses.

- ◯ **Explanation A:** The slightly wrinkled skin on peppers is a natural phenomenon that occurs as moisture evaporates over time. This reduction in freshness makes the peppers less suitable for raw consumption but perfect for stir-frying, which enhances their flavor. As a result, the person used the peppers in a stir-fry, transforming their texture and achieving a delicious and flavorful meal.
- ◯ **Explanation B:** The white spots on the olive oil are likely due to the solidification of fats when stored in cool temperatures, which can be resolved by warming the oil.

Figure 11: Interface presented to human annotators for evaluating pairwise win rates between model responses (e.g., zero-shot vs. retrieved 5-shot). Annotators were shown two anonymized outputs and asked to select the better one based on quality.

```
You are tasked with generating a dataset where each entry consists of the following components:

Caption: A short description of an object or scene in an image.
Rationale: A plausible reasoning explaining why the object or scene might lead to an issue.
Situation: A potential outcome based on the caption and rationale, without explicitly mentioning the cause.

Guidelines for Output:
- The Situation must describe the outcome without directly linking it to the rationale.
- Use clear and concise language.
- Format the output for each entry as follows, enclosed in curly brackets {} to make it easy to parse:

{Caption: "<caption text>"} {Rationale: "<rationale text>"} {Situation: "<situation text>"}

Examples:

Example 1:
{Caption: "red liquid in steak packaging"} {Rationale: "The red liquid found in steak packaging is often mistaken for blood. It is actually a mixture of water and a protein called myoglobin that naturally occurs in muscle tissue. This liquid is perfectly normal and does not indicate that the meat is unsafe."} {Situation: "Person cooked and enjoyed the steak without health issues."}

Example 2:
{Caption: "settling of liquid in yogurt"} {Rationale: "When you open a container of yogurt, you might observe a layer of clear liquid on top, which some may believe signifies spoilage. This liquid is simply whey separating from the yogurt solids, a natural process that doesn't affect the yogurt's quality. Stirring the whey back into the yogurt will restore its creamy texture."} {Situation: "Person enjoyed the yogurt as part of their breakfast."}

Example 3:
{Caption: "green patina on copper cookware"} {Rationale: "Copper cookware may develop a greenish layer called patina. Some people mistake this for harmful corrosion, but patina is natural and can actually protect the copper from further oxidation. The cookware is still usable after proper cleaning."} {Situation: "Person used copper cookware to prepare a delicious meal."}

Example 4:
{Caption: "yellowing leaves on indoor plants"} {Rationale: "Indoor plant leaves may start to turn yellow as a natural part of their growth cycle or due to minor stress factors like overwatering. A few yellow leaves do not necessarily indicate that the plant is dying."} {Situation: "Person continued to care for the plant, and it grew healthy new leaves over time."}

Example 5:
{Caption: "skin peeling after a sunburn"} {Rationale: "After a sunburn, the skin may start to peel. This peeling is part of the natural healing process where the body sheds damaged skin cells. While it might look alarming, it is a normal response to skin damage from ultraviolet light exposure and not a cause for concern."} {Situation: "Person applied moisturizer and supported the skin's healing process comfortably."}

Now your task:
Based on the provided Caption and Rationale, generate the corresponding Situation following the structure and format above.

{Caption: "{INPUT CAPTION HERE}"}{Rationale: "{INPUT RATIONALE HERE}"}
```

Figure 12: Prompt Template for Generating Scenarios for MUN-vis

```
You are tasked with generating a dataset where each entry consists of the
    following components:

Caption: A short description of an object or scene in an image.
Rationale: A plausible reasoning explaining why the object or scene might lead
     to an issue.
Situation: A potential outcome based on the caption and rationale, without
    explicitly mentioning the cause.

Guidelines for Output:
- The Situation must describe the outcome without directly linking it to the
    rationale.
- Use clear and concise language.
- Format the output for each entry as follows, enclosed in curly brackets {}
    to make it easy to parse:

{Caption: "<caption text>"} {Rationale: "<rationale text>"} {Situation: "<
    situation text>"}

Examples:

Example 1:
{Caption: "A coffee maker ready to brew the perfect cup."} {Rationale: "While
    the coffee maker looks functional, its internals are corroded, leading to
    potential contamination of the brewed coffee."} {Situation: "A customer
    experienced stomach discomfort after drinking coffee brewed from the
    machine."}

Example 2:
{Caption: "A sleek sports car parked in the driveway."} {Rationale: "The
    sports car is problematic because it has an undiagnosed mechanical issue,
    making it dangerous to drive."} {Situation: "The driver encountered a
    sudden loss of control while driving, leading to a minor collision."}

Example 3:
{Caption: "A colorful toy ready for playtime."} {Rationale: "This is
    problematic because the toy is a recall item due to safety hazards that
    could pose a choking risk."} {Situation: "A child briefly choked while
    playing with the toy, requiring quick intervention."}

Example 4:
{Caption: "A desktop computer ready for work."} {Rationale: "The computer
    appears functional but is severely infected with malware that could
    compromise sensitive information."} {Situation: "The user faced
    unauthorized access to their private accounts after using the computer for
    online transactions."}

Now your task:
Based on the provided Caption and Rationale, generate the corresponding
    Situation following the structure and format above.

{Caption: "{INPUT CAPTION HERE}"}{Rationale: "{INPUT RATIONALE HERE}"}
```

Figure 13: Prompt Template for Generating Scenarios for MUN-lang

```
Can you improve this explanation so that it becomes more specific to the
    context and makes the outcome more likely to happen?

Context: {INPUT CONTEXT HERE}
Outcome: {INPUT OUTCOME HERE}
Explanation for the outcome: {INPUT EXPLANATION HERE}
```

Figure 14: Prompt Template for improving the human explanation

System Prompt

```
You are a helpful assistant, that ranks models by the quality of their answers
    .

Prompt
I want you to create a leaderboard of different large-language models. To do
    so, I will give you the instructions (prompts) given to the models, and the
     responses of two models. Please rank the models based on which responses
    would be preferred by humans. All inputs and outputs should be Python
    dictionaries.

Here is the prompt:
{
    "instruction": """{instruction}"""
}

Here are the outputs of the models:
[
    {
        "model": "model_1",
        "answer": """{output_1}"""
    },
    {
        "model": "model_2",
        "answer": """{output_2}"""
    }
]

Now please rank the models by the quality of their answers, so that the model
    with rank 1 has the best output. Then return a list of the model names and
    ranks, i.e., produce the following output:
[
    {"model": "model_1", "rank": 1},
    {"model": "model_2", "rank": 2}
]

Your response must be a valid Python dictionary and should contain nothing
    else because we will directly execute it in Python. Please provide the
    ranking that the majority of humans would give.
```

Figure 15: Prompt Template for Assessing Win Rate

```
You are tasked with evaluating the specificity of a given text on a scale of 1
     to 5.
1 (Very Low Specificity): Extremely vague and general.
2 (Low Specificity): Limited details, mostly general.
3 (Moderate Specificity): Includes some details but still general in parts.
4 (High Specificity): Contains clear and detailed information.
5 (Very High Specificity): Extremely detailed and precise, leaving no room for
     ambiguity.

Only output the score as a single number.

Input Text:
[Insert the generated text here]

Output Format:
[Score (1-5)]
```

Figure 16: Prompt Template for Assessing Specificity

```
You are tasked with evaluating the specificity of a given text on a scale of 1
     to 5.
1 (Very Low Specificity): Extremely vague and general.
2 (Low Specificity): Limited details, mostly general.
3 (Moderate Specificity): Includes some details but still general in parts.
4 (High Specificity): Contains clear and detailed information.
5 (Very High Specificity): Extremely detailed and precise, leaving no room for
     ambiguity.

Only output the score as a single number.

Input Text:
[Insert the generated text here]

Output Format:
[Score (1-5)]
```

Figure 17: Prompt Template for Assessing Specificity