# MRMR: A REALISTIC AND EXPERT-LEVEL MULTIDISCIPLINARY BENCHMARK FOR REASONING-INTENSIVE MULTIMODAL RETRIEVAL

**Siyue Zhang**[*NS] **Yuan Gao**[*J] **Xiao Zhou**[*J] **Yilun Zhao**[Y] **Tingyu Song**[A]
**Arman Cohan**[Y] **Anh Tuan Luu**[NV] **Chen Zhao**[SC]

[N]Nanyang Technological University [Y]Yale University [S]NYU Shanghai
[J]Shanghai Jiao Tong University [A]University of the Chinese Academy of Sciences
[C]Center for Data Science, New York University [V]VinUniversity

## ABSTRACT

We introduce MRMR, the first expert-level multidisciplinary multimodal retrieval benchmark requiring intensive reasoning. MRMR contains 1,435 queries spanning 23 domains, with positive documents carefully verified by human experts. Compared to prior benchmarks, MRMR introduces three key advancements. First, it challenges retrieval systems across diverse areas of expertise, enabling fine-grained model comparison across domains. Second, queries are reasoning-intensive, with images requiring deeper interpretation such as diagnosing microscopic slides. We further introduce Contradiction Retrieval, a novel task requiring models to identify conflicting concepts. Finally, queries and documents are constructed as image–text interleaved sequences. Unlike earlier benchmarks restricted to single images or unimodal documents, MRMR offers a realistic setting with multi-image queries and mixed-modality corpus documents. We conduct an extensive evaluation of 4 categories of multimodal retrieval systems and 14 frontier models on MRMR. The text embedding model Qwen3-Embedding with LLM-generated image captions achieves the highest performance, highlighting substantial room for improving multimodal retrieval models. Although latest multimodal models such as Ops-MM-Embedding perform competitively on expert-domain queries, they fall short on reasoning-intensive tasks. We believe that MRMR paves the way for advancing multimodal retrieval in more realistic and challenging scenarios. [1]

## 1 INTRODUCTION

LLM-based agents, such as DeepResearch (OpenAI, 2024; Qiao et al., 2025), have been widely applied in domains including science, engineering, medicine, and finance (Zhao et al., 2025; Tang et al., 2024; Barry et al., 2025; Phan et al., 2025). These systems move beyond the intrinsic knowledge of LLMs by actively retrieving and integrating external information, making a strong and robust retrieval component essential (Chen et al., 2025). In practice, many expert-domain applications rely on multimodal information, underscoring the need for retrieval methods that can handle queries and documents spanning both visual and textual modalities, or even interleaved image–text content (Zhang et al., 2021; Liu et al., 2021; 2023). For instance, given a medical image, the agent system should retrieve similar cases or guidelines to support clinical decisions.

While existing multimodal retrieval benchmarks have made progress, they are insufficient to capture the complexity of agentic scenarios. We identify three key limitations: (1) **Multidisciplinary expert domains**: most multimodal benchmarks are built on Wikipedia text and images, focusing on general-domain knowledge (Hu et al., 2023; Chen et al., 2023; Zhang et al., 2025b). However, state-of-the-art LLMs already demonstrate strong capabilities in handling such knowledge (Team et al., 2025), making it essential to develop benchmarks for high-stakes expert domains such as medicine, science,

[1]Our data is available at https://huggingface.co/datasets/MRMRbenchmark.

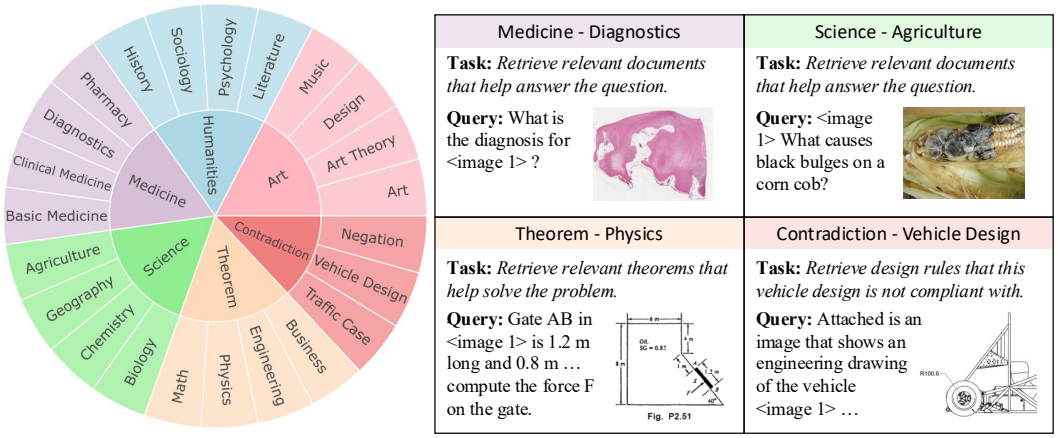



Figure 1: Overview of the MRMR benchmark. MRMR includes 1,435 expert-annotated examples, covering 23 domains across 6 disciplines. It is specifically designed to assess multimodal retrieval models in expert-level, reasoning-intensive tasks. Notably, we originally introduce the *Contradiction Retrieval* task in the multimodal setting, which requires retrieving documents that conflict with the user query and features deeper logical reasoning.

and engineering. (2) **Reasoning intensity**: existing benchmarks primarily target semantic matching and information-seeking tasks, whereas real-world queries often involve expert-domain images and require deeper understanding and logical reasoning over them. (3) **Image-text interleaving**: prior benchmarks mostly support single-image queries with supplementary text, yet real-world queries and documents typically consist of interleaved text and multiple images (Zhang et al., 2025b).

To address these gaps, we introduce **MRMR**, a comprehensive benchmark measuring retrieval models in expert-level **M**ultidisciplinary and **R**easoning-intensive **M**ultimodal **R**etrieval. Figure 1 presents an overview of our benchmark. MRMR consists of 1,435 expert-annotated examples, categorized into three types of retrieval tasks: (1) *Knowledge* for retrieving web pages related to queries involving multiple expert-domain images; (2) *Theorem* for retrieving theorems involved in solving multimodal math problems; and (3) *Contradiction* for retrieving contradictory statements or rules given a case description. Specifically, we derive complex multidisciplinary queries from established Visual Question Answering (VQA) benchmarks (Yue et al., 2024; 2025) and assign expert annotators to collect positive documents from Internet. To build a sizable corpus, we additionally include negative documents from knowledge-intensive collections (Wang et al., 2024a; Su et al., 2025). To further elevate the reasoning challenge, we originally introduce *Contradiction Retrieval*, which requires models not only to detect semantic relevance but also to perform logical reasoning to identify conflicting concepts. To foster a deeper integration of visual and textual content, we represent both queries and documents in an interleaved multimodal format.

We conduct an extensive evaluation on MRMR across four main categories of multimodal retrieval approaches and 14 representative models. The results reveal that current multimodal retrieval systems consistently underperform text-only retrievers with image captioning on knowledge- and reasoning-intensive multimodal queries. The highest score of 54.1 is achieved by the text embedding model Qwen3-Embedding (Zhang et al., 2025d) combined with LLM-based image captioning. The best-performing multimodal model, Ops-MM-Embedding (OpenSearch-AI, 2025), trails by 6.0 points, mainly due to its limited reasoning capabilities rather than domain expertise. Its performance drops from 67.4 on *Knowledge* tasks to 37.4 and 36.6 on *Theorem* and *Contradiction* tasks, even though the corpora for these two tasks are much smaller than that of *Knowledge*. More importantly, the multidisciplinary setup in MRMR reveals substantial performance differences across models and domains. For instance, Ops-MM-Embedding surpasses the second-best model, MM-Embed (Lin et al., 2025), in the Art discipline, whereas their performances are comparable in the Medicine discipline. We hope our benchmark and findings will help progress in multimodal retrieval.

Table 1: Comparison of multimodal retrieval benchmarks and MRMR. In the "Modality" column, "T → I" indicates retrieving image documents using a text query. The "#Domain" column reports the number of domains; "Open" denotes datasets built from Wikidata with general domains. The "Expert", "Reason", and "Interleaved" columns indicate whether expert knowledge is required, whether intensive reasoning is involved, and whether data are in the interleaved image-text format.

| Benchmarks | Modality | Retrieval Type | #Domain | #Query | Expert? | Reason? | Interleaved? |
|---|---|---|---|---|---|---|---|
| NIGHTS | I → I | Visual Similarity | Open | 20K | ✗ | ✗ | ✗ |
| SciMMIR | T ↔ I | Image Caption | 11 | 530K | ✓ | ✗ | ✗ |
| EDIS | T → IT | Image Caption | Open | 3,241 | ✗ | ✗ | ✗ |
| Wiki-SS | T → I | Document QA | Open | 3,610 | ✗ | ✗ | ✗ |
| WebQA | T → IT | Document QA | Open | 2,511 | ✗ | ✗ | ✗ |
| ViDoRe | T → IT | Document QA | 10 | 3,810 | ✓ | ✗ | ✗ |
| MMDocIR | T → IT | Document QA | 10 | 1,658 | ✓ | ✗ | ✗ |
| FashionIQ | IT → I | Composed Image | 1 | 12,238 | ✗ | ✗ | ✗ |
| CIRR | IT → I | Composed Image | Open | 4,148 | ✗ | ✗ | ✗ |
| CIRCO | IT → I | Composed Image | Open | 1,020 | ✗ | ✗ | ✗ |
| InfoSeek | IT → IT | VQA | Open | 1.35M | ✗ | ✗ | ✗ |
| OVEN | IT → IT | VQA | Open | 18,341 | ✗ | ✗ | ✗ |
| wikiHow-TIIR | IT → IT | VQA | Open | 7,654 | ✗ | ✗ | ✓ |
| MRMR | IT → IT | VQA | 23 | 1,435 | ✓ | ✓ | ✓ |

## 2 Related Work

**Benchmarking multimodal retrieval.** As illustrated in Table 1, existing multimodal retrieval datasets mainly focus on semantic matching or information-seeking tasks. Early semantic matching benchmarks are built from paired image–text data, where the text is semantically aligned with the image (Liu et al., 2023; Wu et al., 2024; Xiao et al., 2025; Jiang et al., 2025c), and the task is to retrieve the corresponding modality. Composed Image Retrieval (CIR) emerges as a challenging task that allows users to search for target images using a multimodal query, comprising a reference image and a modification text specifying the user's desired changes to the reference image (Zhang et al., 2021; Liu et al., 2021; Baldrati et al., 2023; Zhang et al., 2024). Information-seeking benchmarks either retrieve supporting evidence for visual questions (Hu et al., 2023; Chen et al., 2023) or retrieve multimodal documents for textual queries (Ma et al., 2024; Macé et al., 2025; Dong et al., 2025). As all prior studies focus on single-image inputs, TIIR (Zhang et al., 2025b) proposes a more realistic setup in which the query and document consist of interleaved text–image sequences supporting multiple images. However, it is limited to searching general-domain wikiHow tutorials. To further advance multimodal retrieval, we construct MRMR, the first benchmark comprising complex multidisciplinary queries that require in-depth reasoning in the interleaved text–image format.

**Multimodal retrieval models and multimodal retrieval augmented generation.** State-of-the-art multimodal retrieval models commonly rely on large pre-trained encoders such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2023), which map images and texts into a shared embedding space. Their outputs are often combined using fusion strategies (*e.g.,* score fusion) to integrate information across modalities (Wei et al., 2024). More recent works adapt multimodal large language models (MLLMs) for universal multimodal embeddings by fine-tuning them on diverse retrieval tasks (Jiang et al., 2025b; Zhang et al., 2025c; Jiang et al., 2025c; Lin et al., 2025). In these approaches, multimodal queries are processed through the MLLM, and the hidden states from the final transformer layer, typically the last token representation, are used as the dense embedding for retrieval. In this work, we benchmark a diverse set of multimodal retrieval approaches, including text retrievers with image captioning, text and image two-stream models with vector fusion, and multimodal retrievers. Additionally, thanks to advances in both retriever and generative models, multimodal retrieval-augmented generation (MM-RAG) has emerged as a key application (Hu et al., 2025; Jiang et al., 2025a; Wu et al., 2025b; Zhan et al., 2025; Wasserman et al., 2025). While various MM-RAG benchmarks have been introduced, most focus on evaluating response generation and lack evidence-level relevance annotations, making it impractical to assess retrieval performance and its contribution within MM-RAG (Chen et al., 2025).

Table 2: Data statistics of MRMR. For each dataset, we show the number of queries ($Q$) and documents ($D$), the average number of positive documents ($D_+$) per example, the average number of text tokens of queries and documents (measured by the GPT-2 tokenizer (Radford et al., 2019), not including task instruction text), the average number of images in queries and documents, and sources of queries and documents. *Knowledge* datasets share a common retrieval corpus, while *Theorem* datasets share another. Examples for each dataset can be found in Appendix G.

| | Total Number | | | Avg. #Text | | Avg. #Images | | Source | | Ex. |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | $Q$ | $D$ | $D_+$ | $Q$ | $D$ | $Q$ | $D$ | $Q$ | $D$ | |
| *Knowledge* | | | | | | | | | | |
| Art | 157 | 26,223 | 1.8 | 15.4 | 421.6 | 1.1 | 0.72 | MMMU-Pro knowledge question | PIN-14M, Web pages | Fig. 13 |
| Medicine | 167 | 26,223 | 2.2 | 32.0 | 421.6 | 1.1 | 0.72 | | | Fig. 14 |
| Science | 137 | 26,223 | 1.8 | 32.1 | 421.6 | 1.2 | 0.72 | | | Fig. 15 |
| Humanities | 94 | 26,223 | 1.9 | 54.5 | 421.6 | 1.2 | 0.72 | | | Fig. 16 |
| *Theorem* | | | | | | | | | | |
| Math | 60 | 14,257 | 2.1 | 64.6 | 364.3 | 1.1 | 0.001 | MMMU-Pro calculation question | BRIGHT, Web pages | Fig.17 |
| Physics | 104 | 14,257 | 2.1 | 56.2 | 364.3 | 1.0 | 0.001 | | | Fig.18 |
| Engineering | 190 | 14,257 | 2.0 | 53.5 | 364.3 | 1.0 | 0.001 | | | Fig.19 |
| Business | 158 | 14,257 | 3.2 | 64.2 | 364.3 | 1.0 | 0.001 | | | Fig.20 |
| *Contradiction* | | | | | | | | | | |
| Negation | 200 | 4 | 1.0 | 0.0 | 12.8 | 1.0 | 0.00 | COCO | Synthetic | Fig.21 |
| Vehicle Design | 88 | 700 | 1.0 | 152.5 | 107.5 | 1.0 | 0.04 | DesignQA | Design Rules | Fig.22 |
| Traffic Case | 80 | 796 | 1.8 | 19.5 | 123.3 | 1.0 | 0.58 | Synthetic | Driving Handbook | Fig.23 |

**Reasoning-intensive retrieval.** Beyond keyword- and semantic-based information retrieval, BRIGHT (Su et al., 2025) has introduced the first benchmark in the text domain that requires intensive reasoning to identify relevant documents. For example, given a new math or physics problem, the retrieval system is expected to provide previously solved problems using the same theorems or relevant theorem statements. To tackle this challenge, recent methods train the text retrievers using synthetic datasets containing complex queries and hard negatives (Weller et al., 2025; Das et al., 2025; Zhang et al., 2025a; Long et al., 2025; Shao et al., 2025; FlagEmbedding, 2025). Our work extends reasoning-intensive retrieval into the multimodal domain. MRMR is constructed by sourcing expert-level queries from the multimodal understanding and reasoning benchmark MMMU-Pro (Yue et al., 2025), collecting image-text interleaved documents from web pages, and obtaining relevance annotations from human experts.

## 3 MRMR BENCHMARK

### 3.1 TASK FORMULATION

We define the task of multimodal retrieval as follows. Let $Q = \{q_1, \ldots, q_n\}$ be the set of queries and $D = \{d_1, \ldots, d_m\}$ the document corpus. Each query $q$ and document $d$ is represented as a sequence of segments $(x_1, \ldots, x_k)$, where each segment $x$ can be either text or an image. For a query $q$, a document can be either a positive document $d_+$ (relevant) or a negative document $d_-$ (non-relevant). In reasoning-intensive retrieval, a document $d$ is considered relevant if it provides principles or theorems that support the reasoning chain required to answer query $q$ (Su et al., 2025). Unlike prior studies (Xiao et al., 2025; Dong et al., 2025), we do not constrain the corpus to uniform data types, reflecting more realistic retrieval scenarios. To evaluate diverse reasoning capabilities, we design three types of retrieval tasks in MRMR:

- ***Knowledge*.** It emphasizes reasoning over broad expert domain knowledge. For a multimodal query, a document is relevant if expert annotators confirm that it contributes to reasoning about the query by providing critical concepts or theoretical foundations.
- ***Theorem*.** It targets the theorem-based reasoning over calculation problems. For a multimodal calculation query, a document is relevant if it conveys the same underlying theorem or formula needed to solve the problem.

- ***Contradiction*.** It requires logical reasoning to detect conflicting or inconsistent concepts. For a multimodal case description query, a document is relevant if it provides the rule or requirement that the query violates.

### 3.2 KNOWLEDGE: RETRIEVING WEB PAGES THAT HELP ANSWER QUESTIONS

MMMU (Yue et al., 2024) is one of the most widely used benchmark for evaluating multi-discipline multimodal understanding in MLLMs. Its robust version, MMMU-Pro (Yue et al., 2025), excludes questions solvable by text-only models, expands the candidate options, and provides verified correct answers. We repurpose the knowledge- and reasoning-intensive questions in MMMU-Pro as queries $Q$ and construct a corpus $D$ of image–text interleaved documents. The positive documents $D_+$ are scraped from relevant websites referenced by the GPT-Search[2] model (OpenAI, 2024) and verified by human experts; while negative documents $D_-$ are augmented by sampling from the multimodal collection PIN-14M (Wang et al., 2024a) (see Figure 2).

**Selecting questions.** We prompt GPT-5[3] to categorize MMMU-Pro questions into two groups, *i.e.,* knowledge-based and calculation questions. We adopt calculation questions for the *Theorem* subset in Section 3.3. For knowledge questions, we then instruct GPT-5 to filter out questions that require only superficial reasoning over text and images, without reliance on external domain expertise. For the remaining questions, we generate detailed descriptions for each associated image using GPT-5, which we include as part of the input context for subsequent steps.

**Constructing positive and hard negative documents.** Unlike keyword- or semantic-based multimodal retrieval benchmarks, collecting positive documents for our queries is more time-consuming because it requires identifying and validating multimodal sources that support the query's answer. To address this, we design a semi-automated pipeline with human expert annotators. Specifically, for each query, given the GPT-5-generated image descriptions and ground-truth answer, we prompt GPT-Search to reason over the question and generate an explanation for the correct answer with reference web links pointing to diverse materials such as Wikipedia, books, academic papers, and blogs. To preserve the completeness of multimodal content, we capture these webpages as PDFs, apply MonkeyOCR (Li et al., 2025) to extract interleaved text and images, and split the content into chunks while preserving image references. Resulting documents are then screened by GPT-5 and validated by human experts about whether they support the correct answer. Documents with GPT-human agreement on relevance are retained as positives, those agreed irrelevant as hard negatives, while ambiguous cases (30–60% across domains) are discarded. In cases where GPT-Search fails to retrieve relevant documents (38.2% of questions), expert annotators are instructed to search the web and create one supporting document, optionally including image links within the text. Due to the complexity of the questions, the number of positive documents per query is typically fewer than four. We annotate data anonymously through the Turkle platform (HLT-COE@JHU, 2025), with detailed guidelines provided in Appendix B.

**Constructing additional negative documents.** After the previous step, we obtain 993 cleaned and annotated documents for 555 queries. To construct a sizable retrieval corpus comparable to (Xiao et al., 2025; Su et al., 2025), we supplement these with negative documents sampled from the large-scale multimodal collection PIN-14M (Wang et al., 2024a), which contains knowledge-intensive resources such as medical articles from PubMed Central (PMC)[4] and web content from OBELICS (Laurençon et al., 2023). Given the wide topic coverage and large number of documents in PIN-14M, we assume a low probability of false negatives for our sampled documents. We validate this assumption through manual error analysis in Section 5.1. In total, we curate a corpus of 26,223 documents, including text only, image only, and text-image interleaved.[5]

[2]GPT-Search refers to the version `gpt-4o-search-preview-2025-03-11` throughout this work.

[3]GPT-5 refers to the version `gpt-5-2025-08-07` throughout this work.

[4]`https://www.ncbi.nlm.nih.gov/pmc/`

[5]The corpus could be further expanded by sampling additional expert-domain documents, which naturally increases retrieval difficulty and the probability of false negatives. We leave it as future work.

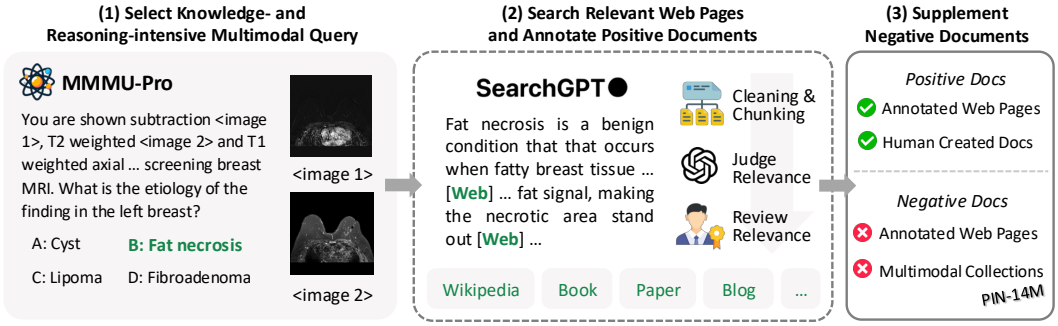



Figure 2: An overview of the data construction workflow for MRMR (*Knowledge*). We select and convert knowledge- and reasoning-intensive questions from MMMU-Pro (Yue et al., 2025) into retrieval queries. Web pages such as Wikipedia, blogs, and papers referenced by the GPT-Search model during reasoning are processed into documents through screen capturing, OCR (Li et al., 2025), and chunking. The relevance of resulting documents is first evaluated by GPT and then verified by expert annotators. Lastly, we source negative documents from the knowledge-intensive multimodal collection PIN-14M (Wang et al., 2024a) to construct a sizable corpus.

### 3.3 THEOREM: RETRIEVING RELEVANT THEOREMS THAT SOLVE PROBLEMS

As introduced by Su et al. (2025), retrieving relevant theorem statements can assist users in solving new math or physics problems. We extend this formulation to the multimodal domain by leveraging challenging calculation problems from MMMU-Pro. In this setting, the query $q$ is a image-centric calculation problem, and the corpus $D$ consists of theorem descriptions across domains such as mathematics, physics, engineering, and business. A document $d$ is regarded as positive if it describes a theorem applicable to solving the query problem.

**Selecting questions.** From the calculation questions in MMMU-Pro, we first use GPT-5 to exclude questions that explicitly state the required theorem in the text. The remaining questions are then organized into four major domains: Math, Physics, Engineering, and Business. The Engineering domain further includes areas such as Mechanical Engineering, Computer Science, and Electronics, while the Business domain covers Finance, Economics, Marketing, and related fields. Then, we prompt GPT-5 to reason through each multimodal question, produce a final answer, and summarize the key theorems used in the solution. We exclude questions for which GPT-5 produces incorrect answers, with a final set of 512 questions.

**Constructing positive and negative documents.** We adopt the theorem statements from BRIGHT (Su et al., 2025) as the primary retrieval corpus (∼13.8k documents), reflecting the realistic setting where most theorems are expressed in text. For each question, the summarized key theorems are used as queries to retrieve the top-10 candidate statements from the corpus with the Qwen3-Embedding model (Zhang et al., 2025d). Among these candidates, GPT-5 identifies the most relevant theorem statements, which are retained as positive documents, while the rest serve as negatives.

**Constructing additional positive documents.** Not all theorems have relevant counterparts in BRIGHT. To address this, we scrape additional theorem statements, optionally accompanied by illustrative images, from webpages such as Wikipedia, following the OCR pipeline described in Section 3.2. GPT-5 then rewrites these theorems to match the format of BRIGHT statements. Finally, we deduplicate the scraped documents to ensure a consistent and complete retrieval corpus. Consequently, 63.6% of the positive documents are sourced from webpages, with the remainder drawn from the BRIGHT corpus. More details are presented in Appendix C.

### 3.4 CONTRADICTION: RETRIEVING CONTRADICTORY RULES AND REQUIREMENTS

Most existing datasets emphasize retrieving positively supporting evidence for a query (Xiao et al., 2025; Chen et al., 2023; Dong et al., 2025). However, retrieving contradictory information could be of great importance especially in expert domains. For example, a user may provide a case description

and seek evidence of violation of laws, policies, or guidelines, as shown in Figure 23. In this setting, the query $q$ is a case description (*e.g.,* traffic or vehicle design cases), while the corpus $D$ comprises mandated rules (*e.g.,* driving theory handbooks or design requirements). A document $d$ is considered positive if it contains the statement or rule contradicting the query case. Unlike traditional retrieval tasks, this new formulation requires not only semantic matching between query and document but also deep logical reasoning to identify conflicting concepts.

**Negation.** To study contradiction retrieval, we first design a synthetic task inspired by the negation benchmark NegBench (Alhamoud et al., 2025). Given an image from COCO (Lin et al., 2014) with ground truth object annotations, we synthesize four candidate text descriptions: three accurately reflecting the objects and one containing a contradiction, either by asserting the existence of a non-existent object or the absence of an existent one. The models are required to pinpoint the text description conflicting with the given image in a multi-choice setup. For example, in Figure 21, the query image shows a keyboard on the table, while the positive document explicitly states that none is present, revealing a contradiction. More details are provided in Appendix D.1.

**Vehicle Design.** To evaluate contradiction retrieval in engineering documents, we construct a vehicle design task by leveraging the Formula SAE Rulebook and design cases from the DesignQA dataset (Doris et al., 2025). In industrial product design, designers must review hundreds of pages of requirement documents to ensure their designs comply with specifications. To assist designers, retrieval systems are expected to identify the specific sections that a design case fails to satisfy. For example, in Figure 22, the vehicle's wheelbase in the design is shorter than the required minimum, indicating a contradiction. During data preparation, we introduce variations to the design cases and chunk the lengthy design document, as detailed in Appendix D.2.

**Traffic Case.** Retrieval systems have been applied to legal documents to assist legal professionals in preparing arguments and citations (Feng et al., 2024). To evaluate this capability in multimodality, we construct a traffic case task to assess whether retrievers can identify which driving rules are violated in traffic cases. We build the corpus by chunking official driving handbooks (Singapore Police Force, 2017) into sections. Meanwhile, we build the query set by selecting dozens of driving rules, each linked to several annotated violation cases. We augment these violation cases by replacing key textual elements with AI-generated images using Qwen-Image (Wu et al., 2025a). For example, as shown in Figure 23, a car is driving only 3 meters behind the vehicle ahead — significantly less than the required safe distance. Further details are provided in Appendix D.3.

# 4 Experiments

## 4.1 Experimental setup

We evaluate 4 types of multimodal retrieval setups with 14 frontier models, as follows: (1) **Text models with image caption (T2T)**: We assess text retrievers, namely BGE-M3 (Chen et al., 2024), NV-Embed-V2 (Lee et al., 2025), and Qwen3-Embedding-8B (Zhang et al., 2025d), by pairing with MLLM-generated image captions (see Appendix E.1 for details). (2) **Text and image two-stream models with vector fusion (IT2IT)**: We evaluate CLIP-style two-stream models, including EVA-CLIP (Sun et al., 2023), SigLIP (Zhai et al., 2023), OpenCLIP (Cherti et al., 2023), and JinaCLIP (Koukounas et al., 2024), by a simple vector-fusion strategy. Given an input sequence, we concatenate all text chunks for one text embedding $t$, while all images are concatenated vertically for another image embedding $i$. Following MTEB (Xiao et al., 2025), the final score is computed using the fused embedding $e = t + i$. (3) **Multimodal models with merged image (IT2IT)**: We evaluate multimodal retrievers including VISTA (Zhou et al., 2024), E5-V (Jiang et al., 2025b), MM-Embed (Lin et al., 2025), VLM2Vec (Jiang et al., 2025c), Ops-MM-Embedding (OpenSearch-AI, 2025) and GME-Qwen2-VL (Zhang et al., 2025c). Since these models support only single-image input, multiple images are concatenated in the same way as for two-stream models. (4) **Multimodal models with document as image (T2I)**: We also include the document retrieval paradigm that receives text-only query and encode entire multimodal documents as screenshot images, such as ColPali (Faysse et al., 2025). Because these models are trained for text queries, query images are replaced with LLM-generated captions, similar to the text retriever setup. Besides, we note that a native image–text interleaved model, TIIR (Zhang et al., 2025b), has been introduced and is expected to

Table 3: The performance of retrieval models on MRMR. We report nDCG@10 for all subtasks except Negation, for which we use Hit@1: Art, Medicine (Med.), Science (Sci.), Humanities (Hum.), Math, Physics (Phy.), Engineering (Eng.), Business (Bus.), Negation (Neg.), Design, and Traffic. Avg. denotes the average score across 11 subtasks. The best score on each subtask is highlighted in **bold**, and the second best is underlined.

| | Knowledge | | | | Theorem | | | | Contradiction | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Art | Med. | Sci. | Hum. | Math | Phy. | Eng. | Bus. | Neg. | Design | Traffic | |
| *Text Models with Image Caption* (T2T) | | | | | | | | | | | | |
| BGE-M3 | 48.6 | 30.0 | 42.4 | 45.6 | 16.5 | 19.5 | 21.6 | 39.3 | 16.0 | 25.9 | 17.4 | 29.3 |
| NV-Embed-v2 | 70.7 | 46.8 | 65.7 | 66.6 | 26.4 | 35.2 | 32.9 | 52.2 | 12.5 | 42.1 | 42.2 | 44.8 |
| Qwen3-Embedding | 71.9 | **53.2** | **72.5** | **74.4** | **37.7** | **50.2** | **42.9** | **58.3** | 12.0 | **67.8** | **54.2** | **54.1** |
| *Text and Image Two-Stream Models with Vector Fusion* (IT2IT) | | | | | | | | | | | | |
| EVA-CLIP | 10.2 | 13.5 | 26.1 | 12.9 | 6.2 | 12.2 | 10.7 | 17.4 | 8.5 | 4.4 | 5.4 | 11.6 |
| SigLIP | 26.7 | 14.7 | 26.7 | 12.3 | 7.4 | 6.5 | 5.9 | 12.5 | 13.5 | 4.9 | 9.6 | 12.8 |
| OpenCLIP | 56.0 | 17.9 | 33.2 | 22.0 | 7.5 | 6.6 | 7.3 | 14.0 | 13.0 | 8.1 | 12.4 | 18.0 |
| JinaCLIP | 21.4 | 16.8 | 27.1 | 10.7 | 10.9 | 7.5 | 9.1 | 13.7 | 10.5 | 16.5 | 9.7 | 14.0 |
| *Multimodal Models with Merged Image* (IT2IT) | | | | | | | | | | | | |
| VISTA | 21.3 | 27.8 | 32.6 | 17.0 | 18.8 | 17.1 | 17.3 | 28.6 | 20.0 | 20.2 | 9.4 | 20.9 |
| E5-V | 25.1 | 11.7 | 16.6 | 10.8 | 2.1 | 3.4 | 2.5 | 5.2 | 11.5 | 3.7 | 2.1 | 8.6 |
| MM-Embed | 65.6 | 53.0 | 63.5 | 62.8 | 23.6 | 30.8 | 27.4 | 44.9 | 7.0 | 23.8 | 34.9 | 39.8 |
| VLM2Vec | 53.5 | 22.4 | 36.7 | 24.0 | 2.1 | 2.8 | 2.8 | 2.9 | 11.5 | 5.6 | 18.3 | 18.1 |
| GME-Qwen2-VL | 54.3 | 40.1 | 46.8 | 45.6 | 28.8 | 36.0 | 30.2 | 45.1 | 15.0 | 26.3 | 29.6 | 36.2 |
| Ops-MM-Embedding | **79.3** | 52.5 | 70.0 | 67.8 | 27.7 | 39.5 | 30.1 | 52.3 | 8.0 | 55.9 | 45.8 | 48.1 |
| *Multimodal Models with Document as Image* (T2I) | | | | | | | | | | | | |
| GME-Qwen2-VL | 54.0 | 40.7 | 59.0 | 50.1 | 21.2 | 22.1 | 27.0 | 45.3 | 14.5 | 56.1 | 40.1 | 39.1 |
| Ops-MM-Embedding | 67.7 | 48.8 | 67.7 | 63.9 | 25.0 | 34.0 | 29.2 | 49.0 | 10.5 | 59.8 | 46.3 | 45.6 |
| ColPali | 36.1 | 29.9 | 42.7 | 29.2 | 7.3 | 17.5 | 13.5 | 34.6 | **28.5** | 19.4 | 18.2 | 25.2 |

best fit the interleaved format of MRMR; however, it is not publicly available. We provide details of each model in Appendix E.1. Following prior work (Xiao et al., 2025; Su et al., 2025), we use nDCG@10 as the main evaluation metric except Negation. Since each query in Negation has exactly one gold document among four candidates, we adopt Hit@1 as the main metric for this task.

### 4.2 Main Results

**Multimodal retrieval systems lag behind text retrieval-based approaches on knowledge- and reasoning-intensive images.** As shown in Table 3, the text retriever Qwen3-Embedding combined with LLM-based image captioning achieves the highest performance (54.1 nDCG@10). Although captions may omit certain visual details, they provide rich contextual descriptions and additional knowledge that substantially benefit retrieval. In contrast, multimodal systems struggle with the expert-level query images in MRMR, which often require deep reasoning, such as diagnosing microscopic tissue sections (Figure 1). CLIP-style two-stream models are particularly limited, as their training emphasizes alignment of superficial text–image semantics and model sizes are relatively small. The most recent MLLM-based embedding models, such as Ops-MM-Embedding, show promising results under both interleaved text–image and document-as-image paradigms, indicating the effectiveness of unified training on diverse retrieval tasks.

**Multimodal retrieval systems perform particularly poorly on reasoning-intensive tasks.** While Ops-MM-Embedding achieves a solid 67.4 nDCG@10 on *Knowledge* subtasks, its performance drops sharply to 37.4 and 36.6 on *Theorem* and *Contradiction*, respectively. Models such as E5-V and VLM2Vec perform even worse, essentially failing on these tasks. This gap highlights the difficulty of extracting abstract concepts from practical problems, for example linking an image-based physics question to the relevant theorem in Figure 1. Notably, Hit@1 scores for most models on the synthetic *Contradiction* task Negation remain below 25%—equivalent to random guessing given four candidates per query. As illustrated in the Negation example Figure 21, humans can readily detect conflicting concepts embedded within supporting evidence, yet retrieval models struggle even for strong text embedding models. Although the candidate corpora for the Design and Traffic

subtasks are much smaller than those of standard knowledge bases (Su et al., 2025; Dong et al., 2025), models still struggle to identify the underlying contradictions. Nevertheless, surface-level semantic matching remains useful in these settings, as it allows models to locate relevant documents without fully resolving the conflicting concepts (*e.g.,* a query concerning driving speed matched with a document specifying the speed limit). These findings suggest that current retrieval models possess strong capabilities in semantic matching and information seeking, but remain fundamentally limited in their reasoning ability.

**Substantial differences in performance are evident across models and domains.** Across all four multimodal retrieval settings, we observe a wide performance difference between models. For instance, among multimodal models with merged image, the weakest model, E5-V, achieves only 8.6 nDCG@10, whereas Ops-MM-Embedding reaches 48.1 nDCG@10, revealing substantial methodological differences. As MRMR is the first multidisciplinary multimodal retrieval benchmark, it enables fine-grained domain-level evaluation. For example, as shown in the breakdown performance Table 7, MM-Embed performs competitively with Ops-MM-Embedding in medical domains such as Clinical Medicine and Diagnostics, yet lags behind in art-related tasks. We also observe pronounced variation in retrieval difficulty across domains. In the Art subtasks, systems can often succeed by matching query images to visually identical or similar artworks, which narrows the search space. However, in medical imaging, such overlap is rare, and models are required to identify underlying pathological and radiological features rather than relying on superficial visual similarity.

# 5 Analysis

## 5.1 Qualitative analysis

To understand model limitations, we conduct 30 case studies by manually reviewing their top-10 documents retrieved by Ops-MM-Embedding. There are two major failure patterns that we have observed. (1) **Visual bias over contextual relevance**: in the Agriculture case as shown in Figure 11, the model ranks a negative document higher because it contains a nematode SEM image resembling the earthworm image in the query, even though the positive document provides a detailed discussion of the key topic Fauna. Similar errors occur in Medicine, where visually similar eye images from different diseases mislead the model. (2) **Failure of higher-level deduction**: in the Traffic case as shown in Figure 12, the model assigns a higher score to a negative document than to a positive one because both depict cars, tunnels, and lane markings. However, it fails to infer that the car is crossing the line, which contradicts the positive document's instruction to "Stay in lane". Although multimodal retrievers exhibit these shortcomings and lag behind text-only retrievers with image captions, we believe they remain essential because many real-world queries inherently span across modalities. Fundamentally, textual descriptions alone cannot fully capture the nuanced information in images, especially when MLLMs lack the required visual knowledge.

## 5.2 Test-time scaling in retrieval

Query expansion is a widely used technique, recently framed as test-time scaling in retrieval (Shao et al., 2025). Prior work (Su et al., 2025) demonstrates that incorporating explicit reasoning substantially improves performance on reasoning-intensive text retrieval tasks. Motivated by this, we have conducted comparative experiments to evaluate the effectiveness for multimodal retrieval. Specifically, we prompt MLLMs, including Qwen2-VL-2B-Instruct (Wang et al., 2024b) and Qwen2.5-VL-72B-Instruct (Bai et al., 2025), to generate reasoning traces, including question summarization and chain-of-thought reasoning, following (Su et al., 2025). As shown in Table 9, replacing the original queries with MLLM-generated reasoning traces leads to substantial performance improvements: $+5.1$ for Qwen2-VL-2B and $+14.8$ for Qwen2.5-VL-72B. The improvements are particularly pronounced on *Knowledge* tasks, whereas *Theorem* tasks benefit to a lesser extent. Meanwhile, we observe that, without constraining output length, the larger model Qwen2.5-VL-72B produces on average 20% and 66% more tokens than Qwen2-VL-2B in *Knowledge* and *Theorem* respectively, trading higher inference cost for larger performance gains (see more details in Appendix F.3).

## 6 CONCLUSION

We introduce MRMR, a realistic, multidisciplinary, reasoning-intensive multimodal retrieval benchmark. We leverage knowledge- and reasoning-intensive questions from MMMU-Pro and build a sizable multimodal corpus with positive documents verified by human experts. In addition, we introduce Contradiction Retrieval for evaluating models' logical reasoning capabilities to identify conflicts. Comprehensive evaluation shows that multimodal retrieval systems lag behind their text-retrieval counterparts, indicating substantial room for improvement. Although state-of-the-art multimodal models excel in *Knowledge* domains, they drop nearly 30 points on reasoning-intensive tasks. We hope MRMR facilitates identifying model limitations and advancing multimodal retrieval.

## CODE OF ETHICS AND ETHICS STATEMENT

All data used in constructing MRMR are sourced from publicly available materials and are employed solely for academic research, not commercial use. We have carefully ensured that the dataset contains no private information or harmful content, such as discriminatory, violent, or unethical material. Our goal is to support socially beneficial research. Following the practice of MMMU (Yue et al., 2024), our annotators and validators are instructed to avoid using materials from websites that prohibit copying or redistribution when reviewing MRMR documents. Consequently, most documents are derived from sources that are free of copyright restrictions, such as Wikipedia pages, government reports (e.g., National Institutes of Health and Singapore Police Force), and PubMed Central (PMC). The datasets we build upon also carry permissive public licenses, including MMMU (Apache-2.0), PIN-14M (CC-BY-4.0), COCO (CC-BY-4.0), and BRIGHT (CC-BY-4.0). For test-time scaling, we primarily focus on text expansion rather than image resizing and process as the text expansion has shown more significant impacts.

## REPRODUCIBILITY

Our datasets and annotation process are introduced in Section 3, and the experimental settings are described in Section 4. Specific implementation details are provided in Appendix E.1. To facilitate the reproduction of our experiments, the data are available at https://huggingface.co/datasets/MRMRbenchmark , and the evaluation code is provided at https://github.com/rebeccaz4/MRMR .

## ACKNOWLEDGMENTS

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). Siyue Zhang and Chen Zhao were supported by NYU Shanghai Center for Data Science. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

## REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Alhamoud_Vision-Language_Models_Do_Not_Understand_Negation_CVPR_2025_paper.pdf.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Baldrati_Zero-Shot_Composed_Image_Retrieval_with_Textual_Inversion_ICCV_2023_paper.pdf.

Mariam Barry, Gaetan Caillaut, Pierre Halftermeyer, Raheel Qader, Mehdi Mouayad, Fabrice Le Deit, Dimitri Cariolaro, and Joseph Gesnouin. GraphRAG: Leveraging graph-based efficiency to minimize hallucinations in LLM-driven RAG for finance data. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, 2025. URL https://aclanthology.org/2025.genaik-1.6/.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL https://huggingface.co/BAAI/bge-m3.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.925/.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent, 2025. URL https://arxiv.org/abs/2508.06600.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://github.com/mlfoundations/open_clip.

Chroma. Chromadb: An open-source vector embedding database, 2025. URL https://github.com/chroma-core/chroma. Apache 2.0 license.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Debrup Das, Sam O' Nuallain, and Razieh Rahimi. Rader: Reasoning-aware dense retrieval models, 2025. URL https://arxiv.org/abs/2505.18405.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. Mmdocir: Benchmarking multi-modal retrieval for long documents, 2025. URL https://arxiv.org/abs/2501.08828.

Anna C Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation. *Journal of Computing and Information Science in Engineering*, 2025.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

Yi Feng, Chuanyi Li, and Vincent Ng. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL https://aclanthology.org/2024.acl-long.350/.

FlagEmbedding. Bge-reasoner: Towards end-to-end reasoning-intensive information retrieval. https://github.com/FlagOpen/FlagEmbedding/tree/master/research/BGE_Reasoner, 2025. Accessed: 2025-09-12.

Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Xin Xia, Xuefeng Xiao, Linjie Yang, Zhonghua Zhai, Xinyu Zhang, Qi Zhang, Yuwei Zhang, Shijia Zhao, Jianchao Yang, and Weilin Huang. Seedream 2.0: A native chinese-english bilingual image generation foundation model, 2025. URL https://arxiv.org/abs/2503.07703.

HLT-COE@JHU. Turkle: An open-source clone of amazon mechanical turk. https://github.com/hltcoe/turkle, 2025.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2302.11154.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *Proceedings of The International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Usklli4gMc.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://arxiv.org/abs/2409.12959.

Ting Jiang, Shaohan Huang, Minghui Song, Zihan Zhang, Haizhen Huang, Liang Wang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, deqing wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models, 2025b. URL https://openreview.net/forum?id=rD6LQagatR.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *Proceedings of The International Conference on Learning Representations (ICLR)*, 2025c. URL https://openreview.net/forum?id=TE0KOzWYAF.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images, 2024. URL https://arxiv.org/abs/2412.08802.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=SKN2hflBIZ.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2025. URL https://arxiv.org/abs/2405.17428.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm, 2025. URL https://arxiv.org/abs/2506.05218.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=i45NQb2iKO.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014. URL https://cocodataset.org/images/coco-paper.png.

Siqi Liu, Weixi Feng, Tsu jui Fu, Wenhu Chen, and William Yang Wang. EDIS: Entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2305.13631.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Liu_Image_Retrieval_on_Real-Life_Images_With_Pre-Trained_Vision-and-Language_Models_ICCV_2021_paper.pdf.

Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yue Shen, Jian Wang, Peng Wei, Jinjie Gu, and Jiahai Wang. Diver: A multi-stage approach for reasoning-intensive information retrieval, 2025. URL https://arxiv.org/abs/2508.07995.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://aclanthology.org/2024.emnlp-main.373/.

Quentin Macé, António Loison, and Manuel Faysse. Vidore benchmark v2: Raising the bar for visual retrieval, 2025. URL https://arxiv.org/abs/2505.17166.

MediaWiki. Api:search — mediawiki,, 2024. URL https://www.mediawiki.org/w/index.php?title=API:Search&oldid=6905053. [Online; accessed 25-September-2025].

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.eacl-main.148/.

OpenAI. Introducing chatgpt search. https://openai.com/index/introducing-chatgpt-search/, 2024. Accessed: 2025-09-17.

OpenSearch-AI. Opensearch-ai/ops-mm-embedding-v1-7b, 2025. URL https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-7B.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, et al. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, Rui Min, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents, 2025. URL https://arxiv.org/abs/2509.13309.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamila Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 1988. URL https://www.sciencedirect.com/science/article/pii/0306457388900210.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen tau Yih, Pang Wei Koh, and Luke Zettlemoyer. Reasonir: Training retrievers for reasoning tasks. *Proceedings of Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.20595.

Singapore Police Force. Basic theory of driving, 2017. URL https://www.police.gov.sg/~/media/spf/files/tp/online%20learning%20portal/bt%20eng%209th%20edition%20130717.pdf. Accessed: 2025-09-21.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=ykuc5q381b.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. URL https://arxiv.org/abs/2303.15389.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://aclanthology.org/2024.findings-acl.33/.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Junjie Wang, Yuxiang Zhang, Minghao Liu, Yin Zhang, Yatai Ji, Weihao Xuan, Nie Lin, Kang Zhu, Zhiqiang Lin, Yiming Ren, Chunyang Jiang, Yiyao Yu, Zekun Wang, Tiezhen Wang, Wenhao Huang, Jie Fu, Qunshu Lin, Yujiu Yang, Ge Zhang, Ruibin Yuan, Bei Chen, and Wenhu Chen. PIN: A knowledge-intensive dataset for paired and interleaved multimodal documents. 2024a. URL https://huggingface.co/datasets/m-a-p/PIN-14M.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL https://arxiv.org/abs/2409.12191.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. Real-mm-rag: A real-world multi-modal retrieval benchmark, 2025. URL https://arxiv.org/abs/2502.12342.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *The European Conference on Computer Vision (ECCV)*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/11927.pdf.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. Rank1: Test-time compute for reranking in information retrieval, 2025. URL https://arxiv.org/abs/2502.18418.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search, 2025b. URL https://arxiv.org/abs/2506.20670.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://aclanthology.org/2024.findings-acl.746/.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark, 2025. URL https://arxiv.org/abs/2504.10471.

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

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL https://aclanthology.org/2025.acl-long.736/.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. URL https://arxiv.org/pdf/2303.15343.

Zaifu Zhan, Jun Wang, Shuang Zhou, Jiawen Deng, and Rui Zhang. Mmrag: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning, 2025. URL https://arxiv.org/abs/2502.15954.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *The Forty-first International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2403.19651.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. Diffusion vs. autoregressive language models: A text embedding perspective, 2025a. URL https://arxiv.org/abs/2505.15045.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/papers/Wu_Fashion_IQ_A_New_Dataset_Towards_Retrieving_Images_by_Natural_CVPR_2021_paper.pdf.

Xin Zhang, Ziqi Dai, Yongqi Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, Jun Yu, Wenjie Li, and Min Zhang. Towards text-image interleaved retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025b. URL https://aclanthology.org/2025.acl-long.214/.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2025c. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Zhang_Bridging_Modalities_Improving_Universal_Multimodal_Retrieval_by_Multimodal_Large_Language_CVPR_2025_paper.pdf.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025d. URL https://arxiv.org/abs/2506.05176.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot, 2025. URL https://arxiv.org/abs/2502.04413.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL https://aclanthology.org/2024.acl-long.175/.

## Appendix Contents


**A The Use of Large Language Models** **18**

**B Dataset Construction: Knowledge** **18**
- B.1 Annotator Biography . . . 18
- B.2 Annotation Guideline and Interface . . . 18
- B.3 Data Annotation Payment . . . 18
- B.4 Dataset Construction Prompts . . . 19

**C Dataset Construction: Theorem** **19**
- C.1 Theorem Database Construction . . . 19
- C.2 Wikipedia Content Processing Pipeline . . . 19
- C.3 Document Deduplication . . . 19

**D Dataset Construction: Contradiction** **20**
- D.1 Negation . . . 20
- D.2 Vehicle Design . . . 20
- D.3 Traffic Case . . . 20

**E Experiment Details** **23**
- E.1 Models and Instructions . . . 23
- E.2 Implementations and Machines . . . 23
- E.3 Detailed Results . . . 24

**F Analysis Details** **25**
- F.1 False Positive and False Negative Analysis . . . 25
- F.2 Error Case Studies . . . 25
- F.3 Test-Time Scaling in Retrieval . . . 29

**G Data Examples** **30**

# A The Use of Large Language Models

In this work, large language models (LLMs) are employed solely as tools for data generation, as described in the main paper. Importantly, no parts of the manuscript are generated by LLMs. Hence, there are no concerns of plagiarism or scientific misconduct related to text generation.

# B Dataset Construction: Knowledge

## B.1 Annotator Biography

The detailed biographies of the annotators involved in MRMR construction are presented in **Table 4**. All annotators are from universities ranked in the Top 500 of the 2025 QS Global Rankings[3] and are fluent in English. Annotators assess document–query relevance by judging whether a document facilitates answering the query. To ensure quality, independent validators conduct an additional round of verification.

Table 4: Biographies of 24 annotators involved in MRMR construction (Author biographies are hidden to protect identity confidentiality).

| ID | Year | Major | Assigned Subject(s) | Author? | Validator? |
|---|---|---|---|---|---|
| 1 | 3rd year Undergraduate | Biological Engineering | Biology | ✗ | ✗ |
| 2 | 1st year Master | Biological Engineering | Biology | ✗ | ✓ |
| 3 | 1st year Master | Biomedical Engineering | Biology, Pharmacy | ✗ | ✗ |
| 4 | 2nd year Master | Biomedical Engineering | Biology, Pharmacy | ✗ | ✗ |
| 5 | 1st year Master | Biomedical Engineering | Biology, Pharmacy | ✗ | ✗ |
| 6 | 1st year PhD | Chemistry | Chemistry | ✗ | ✗ |
| 7 | 2nd year Master | Chemistry | Chemistry | ✗ | ✓ |
| 8 | 3rd year PhD | Medicine | Basic Medicine | ✗ | ✗ |
| 9 | 3rd year Undergraduate | Clinical Medicine | Clinical Medicine, Diagnostics | ✗ | ✗ |
| 10 | 3rd year Undergraduate | Medicine | Basic Medicine | ✗ | ✓ |
| 11 | 2nd year Master | Clinical Medicine | Clinical Medicine, Diagnostics | ✗ | ✗ |
| 12 | 2nd year Master | Clinical Medicine | Clinical Medicine, Diagnostics | ✗ | ✓ |
| 13 | 3rd year Undergraduate | Pharmacology | Pharmacy | ✗ | ✓ |
| 14 | 4th year Undergraduate | Pharmacology | Pharmacy | ✗ | ✗ |
| 15 | 1st year Master | Music | Music | ✗ | ✗ |
| 16 | 1st year Master | Clinical Medicine | Clinical Medicine | ✗ | ✗ |
| 17 | 1st year PhD | Sociology | Sociology, Psychology | ✗ | ✗ |
| 18 | 1st year Master | Bioinformatics | Biology | ✗ | ✗ |
| 19 | 2nd year PhD | Agricultural and Biosystems Engineering | Agriculture | ✗ | ✗ |
| 20 | 4th year Undergraduate | Literature | History, Literature | ✗ | ✗ |
| 21 | 3rd year Undergraduate | Geography and Environmental Studies | Geography | ✗ | ✗ |
| 22 | 4th year PhD | Computer Science | - | ✓ | ✓ |
| 23 | 4th year Undergraduate | Computer Science | - | ✓ | ✓ |
| 24 | 3rd year Undergraduate | Electronic Engineering | - | ✓ | ✓ |

## B.2 Annotation Guideline and Interface

To facilitate data annotation, we develop the following interface based on Turkle (HLT-COE@JHU, 2025), an open-source clone of Amazon's Mechanical Turk. The annotation guideline and interface is detailed in Figure 3, Figure 4, and Figure 5.

## B.3 Data Annotation Payment

The annotation and validation process for MRMR spanned three months. Each annotator was assigned approximately **50 questions** aligned with their academic major. After annotation, validators independently assessed the quality of the labels. We provided a *base rate* of **7 USD per hour**, with a quality adjustment of about 10%. On average, annotating a single question required **10 minutes**, while validation took **4 minutes**. This compensation scheme ensured that annotators received wages competitive with the average teaching assistant salary at their universities. To maintain a manageable workload and reduce pressure, we recommended a maximum of **10 questions per day**.

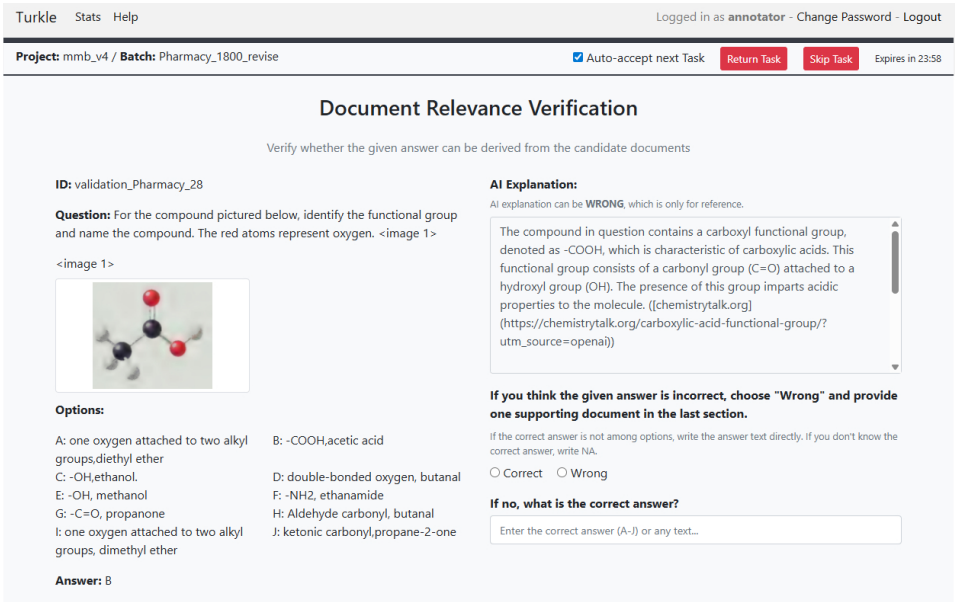



Figure 3: **Annotation Interface - Step 1: Question Understanding.** Annotators are first shown the question, associated images, candidate options, the correct answer, and an AI-generated explanation. The explanation is provided to aid understanding, though annotators are informed it may be incorrect. In this step, they judge whether the given answer is correct based on their own knowledge.

### B.4 Dataset Construction Prompts

The dataset construction prompts are presented in Figure 6, Figure 7, Figure 8, and Figure 9.

## C Dataset Construction: Theorem

### C.1 Theorem Database Construction

The BRIGHT theorem corpus was embedded using Qwen3-Embedding (Zhang et al., 2025d) and indexed in ChromaDB, which supports efficient semantic search via HNSW (Chroma, 2025). Each entry retains a unique `theorem_id` and the original `text`, enabling fast, semantics-aware retrieval with full traceability to the source.

### C.2 Wikipedia Content Processing Pipeline

We retrieved Wikipedia content by querying the MediaWiki Search API (MediaWiki, 2024) using theorem names as search keys. For supplementary sources in PDF format, we employed MonkeyOCR (Li et al., 2025) to convert scanned documents into Markdown. The resulting text was then processed through a structured extraction prompt (Figure 10) using GPT-5 to perform final cleaning, normalization, and precise theorem statement extraction.

### C.3 Document Deduplication

All theorems extracted from Wikipedia were deduplicated prior to inclusion in the corpus. Deduplication was performed in two stages: first by theorem name, and then by semantic content using TF-IDF–based cosine similarity (Salton & Buckley, 1988). Specifically, we employed `TfidfVectorizer` to compute `TF-IDF` vectors for all theorem statements (Pedregosa et al., 2011), followed by pairwise cosine similarity. Entries with near-identical content (cosine similarity $\geq 0.85$) were collapsed into a single representative instance.

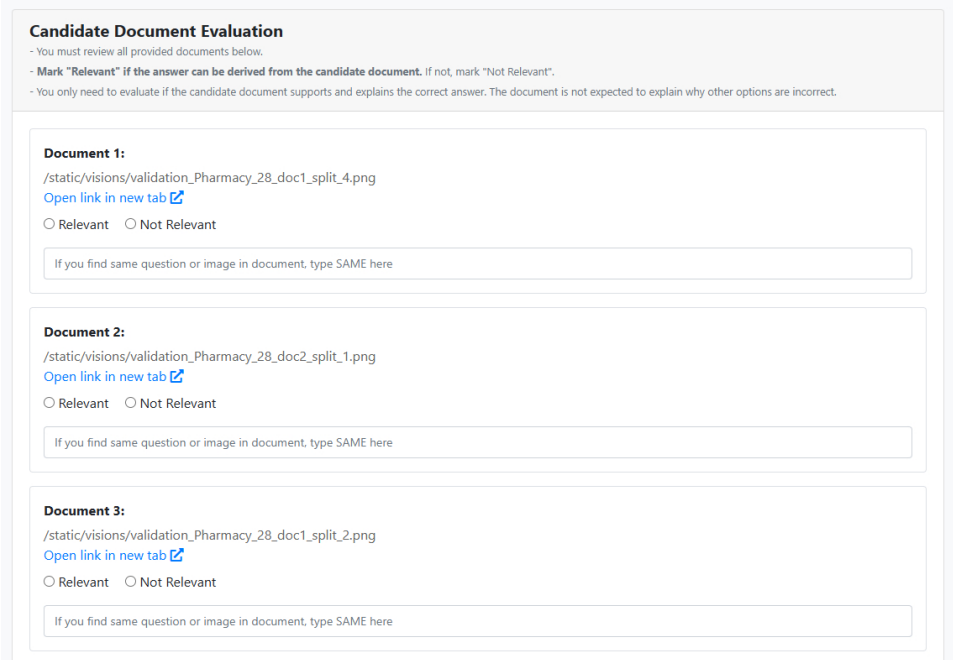



Figure 4: **Annotation Interface — Step 2: Candidate Document Evaluation.** After understanding the question, annotators are instructed to review candidate documents individually and judge whether each can facilitate correctly answering the question. Documents are shown in image format, with up to eight candidates presented. Document relevance definition has been explained to annotators before the annotation process.

# D Dataset Construction: Contradiction

## D.1 Negation

First, we randomly select 200 samples from the COCO (Lin et al., 2014) dataset, each containing at least three positive objectives. For each entry, we construct a description using the template, "The image includes $a$, $b$, $c$, but no $d$." In the positive description, we randomly select three positive objectives to replace $a$, $b$, and $c$, and select one negative objective to replace d. For the negative description, we generate two variations: one where all four objectives ($a$, $b$, $c$, $d$) are selected from the positive objectives, and another where one of $a$, $b$, or $c$ is replaced by $a$ randomly selected negative objective. The image from each sample is used as the query, and the three positive descriptions and one negative description are used as the corpus. Finally, we manually review the 200 queries and corresponding gold documents to ensure that the contradictory descriptions are identifiable by humans, and revise any ambiguous queries for clarity. No LLM prompting is involved in constructing the Negation task.

## D.2 Vehicle Design

On one hand, to construct the queries, we use design cases from the DesignQA dataset (Doris et al., 2025) and augment them through appropriate modifications, such as altering numerical values and introducing variations in image elements. On the other hand, to construct the corpus, we apply MonkeyOCR (Li et al., 2025) to extract and segment the Formula SAE Rulebook into 700 files, organized by rule ID. Finally, we review all the queries to ensure they represent incorrect designs.

## D.3 Traffic Case

First, we select a set of traffic rules and, based on these rules, create traffic violation cases by crafting relevant stories. These stories are then used as prompts to generate 12 images for each story using

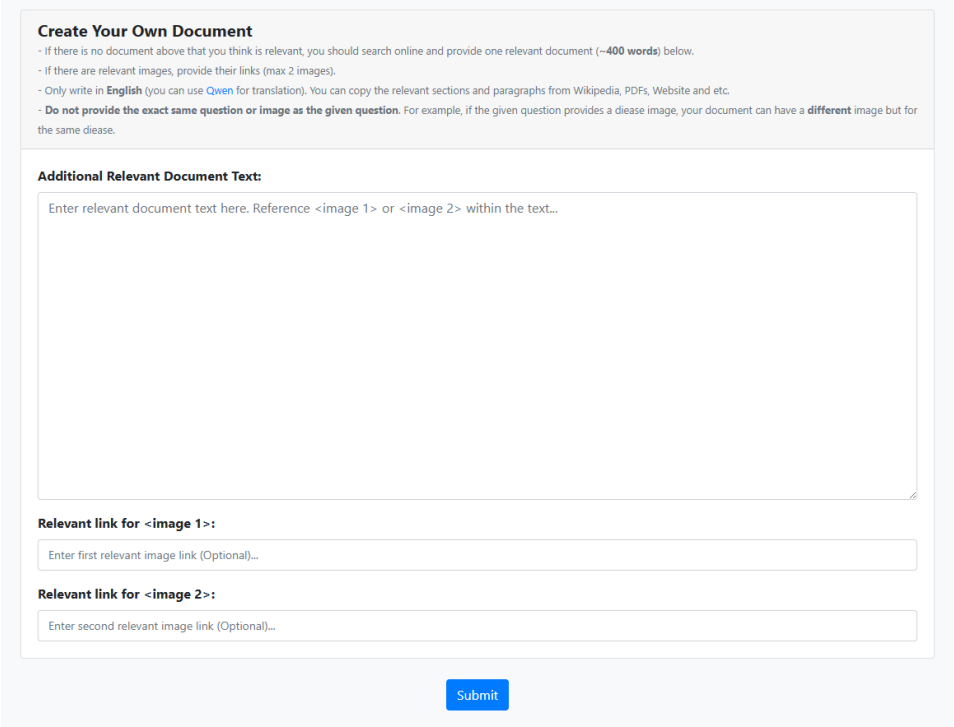



Figure 5: **Annotation Interface — Step 3: Create Relevant Document.** If none of the candidate documents are deemed relevant, annotators are required to search for a suitable web page and provide the gold evidence content. They are encouraged to include images from the source, and the final document is written in an interleaved image–text format.

Your task is to determine whether a question with images requires expert knowledge, such as about a historical event, scientific concept, economic theory, or medical disease. The last line of your response should be of the following format: "Result: YES_OR_NO" (without quotes). If the answer can be obtained easily by reading the question text and image content alone without the need of expert knowledge, say NO. Think step by step before answering. Here are some examples:

`{example_1}`

`{example_2}`

Now please determine whether this new question requires expert knowledge:

`{question_and_answer}`

Figure 6: The prompt for determining whether the question is knowledge-based.

GPT-5. Afterwards, we manually review all the generated images and use Doubao (Gong et al., 2025) to refine and enhance them for better clarity and relevance. Additionally, we leverage Doubao to generate specific objectives from the queries in order to construct image–text interleaved queries. For the corpus, we use MonkeyOCR to split Basic Theory of Driving and Final Theory of Driving (Singapore Police Force, 2017), two official driving handbooks in Singapore, into separate files, which are then organized and used as the corpus. Finally, we conduct a manual review of all the queries, ensuring that any additional corpus IDs caused by excessive image details are properly incorporated into the queries.

Is the multimodal question testing a theorem, formula, equation, or algorithm in domains such as physics, economics, finance, computer science, and math? Answer YES or NO directly.

`{question_and_answer}`

Figure 7: The prompt for determining whether the question is theorem-based.

Explain the answer to the question in a clear and detailed manner. Include citation web links to support the explanation — use relevant Wikipedia pages whenever possible. If a Wikipedia page is not available, use other reliable sources.

`{question_and_answer}`

Figure 8: The prompt for searching relevant web pages using GPT-Search.

Question:

`{question_and_answer}`

Document:

`{document}`

You are a document analysis assistant. Your task is to determine whether the given document above answers the given question and supports the given answer.
**Instructions:**

1. If the answer can be derived or inferred from the text and images in the document, respond YES; otherwise, respond NO.
2. If the document discusses related topics but does not directly answer the given question, respond NO.
3. If the document only provides reference paper titles without substantive content that supports the answer, respond NO.

First, think step by step and explain your reasoning. In the last separate line, directly respond YES or NO without quotes.

Figure 9: The prompt for judging whether the document is relevant to the question.

You are given a markdown document. Your task is to extract the specific theorem, formula, equation, algorithm, or concept named "`{theorem_name}`" from this document.
**Instructions:**

1. Carefully locate the section that describes the theorem "`{theorem_name}`".
2. Extract the complete definition, explanation, and any associated formulas or equations.
3. Remove all reference citations.
4. If there are referenced images in the content, preserve the image references exactly as they appear.
5. Your response MUST follow the following LaTeX-style format:

```
\begin{definition}[{theorem_name}]
Complete definition and explanation,
preserving mathematical notation.
Include examples if present.
\end{definition}
```

Here is the document content:
`{markdown_content}`

Figure 10: The prompt for cleaning the theorem content.

# E Experiment Details

## E.1 Models and Instructions

Table 5: Details of the multimodal retriever models evaluated in MRMR.

| Model | Size | Version |
|---|---|---|
| BGE-M3 (Chen et al., 2024) | 600M | BAAI/bge-m3 |
| NE-Embed-V2 (Lee et al., 2025) | 8B | nvidia/NV-Embed-v2 |
| Qwen3-Embedding (Zhang et al., 2025d) | 8B | Qwen/Qwen3-Embedding-8B |
| EVA-CLIP (Sun et al., 2023) | 400M | QuanSun/EVA02-CLIP-L-14 |
| SigLIP (Zhai et al., 2023) | 650M | google/siglip-large-patch16-256 |
| JinaCLIP (Koukounas et al., 2024) | 860M | jinaai/jina-clip-v2 |
| OpenCLIP (Cherti et al., 2023) | 1.4B | laion/CLIP-ViT-g-14-laion2B-s34B-b88K |
| VISTA (Zhou et al., 2024) | 200M | BAAI/bge-visualized-m3 |
| VLM2Vec (Jiang et al., 2025c) | 4B | TIGER-Lab/VLM2Vec-Full |
| GME-Qwen2-VL (Zhang et al., 2025c) | 7B | Alibaba-NLP/gme-Qwen2-VL-7B-Instruct |
| Ops-MM-Embedding (OpenSearch-AI, 2025) | 7B | OpenSearch-AI/Ops-MM-embedding-v1-7B |
| E5-V (Jiang et al., 2025b) | 8B | royokong/e5-v |
| MM-Embed (Lin et al., 2025) | 8B | nvidia/MM-Embed |
| ColPali (Faysse et al., 2025) | 3B | vidore/colpali-v1.3 |

Following TIIR, we evaluate text retrievers on multimodal retrieval tasks by replacing images with captions generated by an LLM. To simulate real-time inference, we apply the standardized prompt "Describe the image" and use Qwen2-VL-2B-Instruct to produce the captions.

## E.2 Implementations and Machines

The MRMR dataset is constructed following the conventions of MTEB (Muennighoff et al., 2023), including data format and evaluation pipeline, with modifications to support mixed-modality inputs during evaluation. All experiments are conducted on NVIDIA A100, A6000, or H100 GPUs. The runtime of a full evaluation depends on the model, but with the limited corpus size for efficiency, one complete run can be completed within 4 hours on a single A100 GPU for open-source dense models. To further accelerate dense model evaluation, we employ FlashAttention (Dao et al., 2022).

Table 6: Instruction prompts used during model evaluation in MRMR.

| **Task** | **Modality** | **Prompt** |
|---|---|---|
| Knowledge | Multimodal<br>Text | Retrieve relevant documents that help answer the question. |
| Theorem | Multimodal<br>Text | Retrieve relevant theorems that are involved in solving the problem. |
| Negation | Multimodal<br>Text | Given an image, retrieve descriptions that have contradictory information with the image.<br>Given an image caption, retrieve descriptions that have contradictory information with the image caption. |
| Vehicle Design | Multimodal<br>Text | Given a vehicle design, retrieve the design requirements that it violates.<br>Given a vehicle design description, retrieve the design requirements that it violates. |
| Traffic Case | Multimodal<br>Text | Given a traffic case, retrieve the driving rule documents that it violates.<br>Given a traffic case description, retrieve the driving rule documents that it violates. |

### E.3 Detailed Results

Table 7: Detailed performance of retrieval models on MRMR (*Knowledge*).

| | Knowledge | | | | | | | | | | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Music | Design | Theo. | Art | Hist. | Soci. | Psy. | Lit. | Pharm. | Diag. | Clinic. | Basic. | Agri. | Geo. | Chem. | Bio. | |
| *Text Models with Image Caption* | | | | | | | | | | | | | | | | | |
| BGE-M3 | 43.4 | 44.0 | 49.4 | 57.2 | 47.7 | 39.5 | 52.2 | 15.8 | 58.5 | 11.2 | 28.2 | 36.2 | 38.7 | 48.6 | 37.6 | 48.3 | 41.0 |
| NV-Embed-v2 | 63.8 | 61.8 | 70.1 | 86.8 | 70.6 | 64.3 | 59.7 | 95.8 | 78.0 | 19.8 | 46.0 | 59.0 | 65.3 | 63.3 | 70.0 | 63.6 | 64.9 |
| Qwen3-Embedding | 62.8 | 62.1 | 74.8 | 87.3 | 76.1 | 74.0 | 69.3 | 97.8 | 83.1 | 34.8 | 47.0 | 64.0 | 69.5 | 76.5 | 74.0 | 72.6 | 70.4 |
| *Text and Image Two-Stream Models with Vector Fusion* | | | | | | | | | | | | | | | | | |
| EVA-CLIP | 30.5 | 1.5 | 3.5 | 7.5 | 16.7 | 5.5 | 16.3 | 0.0 | 22.7 | 10.3 | 10.0 | 16.4 | 41.6 | 15.4 | 20.4 | 18.5 | 14.8 |
| SigLIP | 25.0 | 25.6 | 26.2 | 30.0 | 16.7 | 1.4 | 14.7 | 22.7 | 13.8 | 9.7 | 15.6 | 19.6 | 30.2 | 18.3 | 26.7 | 27.3 | 20.2 |
| OpenCLIP | 20.9 | 50.7 | 62.9 | 86.4 | 35.8 | 10.2 | 15.1 | 22.7 | 11.1 | 10.6 | 20.8 | 25.8 | 34.1 | 45.8 | 23.9 | 34.3 | 31.9 |
| JinaCLIP | 18.5 | 11.0 | 23.0 | 33.1 | 14.2 | 0.0 | 17.1 | 0.0 | 17.8 | 6.1 | 21.7 | 21.1 | 35.1 | 24.7 | 30.4 | 15.4 | 18.1 |
| *Multimodal Models with Merged Image* | | | | | | | | | | | | | | | | | |
| VISTA | 39.3 | 3.5 | 17.2 | 27.5 | 12.3 | 13.9 | 28.0 | 0.0 | 48.9 | 18.2 | 23.9 | 31.2 | 33.6 | 22.0 | 36.9 | 33.1 | 24.3 |
| E5-V | 13.0 | 23.4 | 17.6 | 46.1 | 15.6 | 4.3 | 10.8 | 7.7 | 12.5 | 7.1 | 13.5 | 13.7 | 18.3 | 13.1 | 23.3 | 10.0 | 15.6 |
| MM-Embed | 51.6 | 60.8 | 68.3 | 80.5 | 57.5 | 69.4 | 59.5 | 94.1 | 63.8 | 35.1 | 50.9 | 68.9 | 60.9 | 76.0 | 62.1 | 60.7 | 63.8 |
| VLM2Vec | 34.4 | 44.0 | 49.6 | 84.8 | 36.4 | 12.3 | 19.3 | 19.2 | 17.4 | 13.6 | 23.7 | 33.1 | 39.0 | 40.7 | 37.5 | 30.8 | 33.5 |
| GME-Qwen2-VL | 55.1 | 40.4 | 57.1 | 64.8 | 39.2 | 50.6 | 51.1 | 32.9 | 57.2 | 20.6 | 32.1 | 62.2 | 38.9 | 48.4 | 63.6 | 39.6 | 47.1 |
| Ops-MM-Embedding | 58.5 | 75.6 | 84.2 | 96.8 | 71.4 | 71.1 | 59.7 | 73.7 | 76.1 | 30.9 | 50.7 | 64.5 | 58.7 | 78.5 | 80.4 | 69.0 | 68.7 |
| *Multimodal Models with Document as Image* | | | | | | | | | | | | | | | | | |
| GME-Qwen2-VL | 58.2 | 46.5 | 53.6 | 58.4 | 52.5 | 48.5 | 48.2 | 52.1 | 72.9 | 16.8 | 31.7 | 40.2 | 49.7 | 69.0 | 53.8 | 45.4 | 49.8 |
| Ops-MM-Embedding | 60.6 | 59.0 | 68.4 | 82.4 | 68.3 | 63.0 | 58.6 | 68.3 | 74.3 | 31.2 | 39.3 | 65.9 | 57.2 | 69.3 | 76.1 | 71.9 | 63.4 |
| ColPali | 25.1 | 27.7 | 46.4 | 43.7 | 31.7 | 19.4 | 38.5 | 0.0 | 64.1 | 10.6 | 23.0 | 60.1 | 36.7 | 32.6 | 67.6 | 56.3 | 36.5 |

# F Analysis Details

## F.1 False Positive and False Negative Analysis

Although queries and their relevant documents were carefully validated by human annotators, false positives and false negatives may still arise when aggregating documents across queries or sampling from external corpora. We explicitly instructed annotators and validators to identify similar or related queries and to cross-annotate documents accordingly. As a result, some queries share the same relevant documents.

To quantitatively assess the prevalence of such labeling errors, we conducted a human audit of the top-retrieved documents retrieved by the best-performing model Ops-MM-Embedding. As shown in Table 8, the audit revealed zero false positives and a false negative rate of only 2.5% for *Knowledge* tasks for sampled 120 documents. Similarly, for *Theorem* tasks, the combined error rate was minimal at approximately 5.8%, comprising 3.3% false negatives and 2.5% false positives. These results suggest that label noise is insignificant, thereby supporting the reliability of the benchmark.

For *Contradiction* tasks, the dataset is relatively small and predominantly constructed manually by annotators and validators. Given the quality control in the construction process, no additional human evaluation was deemed necessary.

Table 8: Human audit of document relevance annotations for the top-retrieved documents produced by the best-performing multimodal model, Ops-MM-Embedding. A *false negative* is a relevant document incorrectly labeled as irrelevant by our method, and a *false positive* is an irrelevant document incorrectly labeled as relevant.

| | | False Negatives | | False Positives | |
|---|---|---|---|---|---|
| Dataset | Documents Checked | Count | Ratio | Count | Ratio |
| *Theorem* | 120 | 4 | 3.3% | 3 | 2.5% |
| *Knowledge* | 120 | 3 | 2.5% | 0 | 0.0% |

## F.2 Error Case Studies

In this section, we present case studies for the Ops-MM-Embedding model in different domains such as Biology (Figure 11) and Traffic Case (Figure 12). The error case analysis for *Theorem* tasks are exemplified as follows.

A recurring issue in Engineering and Geometry tasks is the model's tendency to perform coarse-grained matching based on shape and keywords, while ignoring the specific geometric conditions or physical constraints defined in the query.

**Case Study: Engineering (Geometry)**
*Query ID: validation_Architecture_and_Engineering_5*

- **Query Content:** An image showing a pentagon with internal angles and a specific coordinate bearing angle ($\alpha_{12} = 30°$), asking to calculate other bearing angles.
- **Retrieved Negative (Top-1):** "Inscribing Circle in Regular Pentagon".
- **Analysis:** This represents a **keyword and shape hallucination**. The model correctly identifies the visual object (a pentagon) and the domain (geometry/angles). However, it retrieves a document about inscribing circles—likely because the dense geometric keywords and the visual of a polygon strongly correlate in the embedding space. The model fails to attend to the specific logical task (calculating bearing angles) and instead prioritizes the dominant visual features (the pentagon shape).

**Case Study: Engineering (Statics)**
*Query ID: test_Architecture_and_Engineering_214*

- **Query Content:** A floor plan asking to compute the "tributary areas" for a specific floor beam B1.
- **Retrieved Negative (Top-1):** "Static equilibrium".
- **Analysis:** The retrieved document discusses the static equilibrium of beams. While semantically related to the domain (structural engineering and beams), it is a conceptual mismatch. The model retrieves a theoretical concept (statical indeterminacy) rather than the procedural knowledge required for area calculation. This suggests that the retriever struggles to distinguish between *theoretical concepts* and *practical calculation tasks* when visual cues (schematic diagrams of beams) are similar.

In scientific domains, the model often exhibits "partial understanding", where it correctly identifies the strict sub-domain or topic but fails to retrieve the document addressing the specific variable or relationship queried.

**Case Study: Physics (Thermodynamics)**
*Query ID: test_Physics_74*

- **Query Content:** A $P - V$ (Pressure-Volume) graph showing a cyclic process, asking to identify the point of highest temperature.
- **Retrieved Negative (Top-1):** "Isothermal process".
- **Analysis:** The retrieved document explains isothermal processes (constant temperature), which frequently utilize $P - V$ diagrams similar to the query image. The retrieval is plausible but incorrect; the model latched onto the visual graph type ($P - V$ curve) but failed to deduce the specific relationship ($PV = nRT$) required to find the maximum temperature. This confirms the limitation regarding **higher-level deduction**: the model recognizes the graphical language but not the specific physical implication.

**Case Study: Math (Data Interpretation)**
*Query ID: test_Math_469*

- **Query Content:** A histogram/bar chart showing student distances from school, asking for the percentage of students whose distance falls within the 5 km to 10 km range.
- **Retrieved Negative (Top-1):** "Generic statistical methods".
- **Analysis:** The retrieved document discusses general statistical techniques and data representation, which often involve histograms similar to the query image. The retrieval appears contextually related due to the presence of a histogram but is ultimately incorrect; the model associated the visual format with a broad statistical category but failed to extract or interpret the specific numerical data (bin counts and ranges) needed to compute the required percentage. This highlights a deficiency in **visual-numerical alignment**: the model recognizes the chart type but does not connect it to the mathmatical reasoning.

Science – Biology

**ID:** validation_Agriculture_6

**Task:** *Retrieve relevant documents that help answer the question.*

**Query:** The picture below shows a common soil organism. How should this organism be classified in terms of Flora vs. Fauna and by its size category? <image 1>

**Answer:** macro-fauna.



Negative Documents

**Soybean cyst nematode**

The soybean cyst nematode (SCN), Heterodera glycines, is the most devastating pest to soybean crop yields in the U.S. targeting the roots of soybean and other legume plants. When infection is severe SCNs cause stunting, yellowing, impaired canopy development, and yield loss. The symptoms caused by SCNs can go easily unrecognized by farmers—in some cases there are no warning symptoms before a loss of 40% of the yield. Due to the slight stunting and yellowing, many farmers may mistake these symptoms as environmental problems when in fact they are SCNs. Another symptom of SCNs that may affect farmers' yields is stunted roots with fewer nitrogen-fixing nodules. Due to the fact that soybean cyst nematodes can only move a few centimeters in the soil by themselves, they mostly are spread via tillage or plant transplants. This area of infection will look patchy and nonuniform making diagnosis more difficult for farmers. They can be seen in the roots of summer soybean plants if the roots are taken out very carefully and gently washed with water. The egg masses should be seen as bright white or yellow "pearls" on the roots. The later the roots are pulled the harder it will be to diagnose due to the SCNs female dying and turning a much darker color, forming a "cyst". The best way to know if a field is infected by soybean cyst nematodes is to take a soil sample to a nematologist.



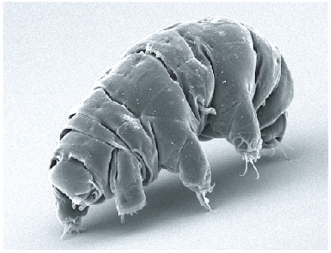
Soybean cyst nematode
Soybean cyst nematode and egg

| | |
|---|---|
| Kingdom: | Animalia |
| Phylum: | Nematoda |
| Class: | Secernentea |
| Order: | Tylenchida |
| Family: | Heteroderidae |
| Genus: | Heterodera |
| Species: | *H. glycines* |

Positive Documents

Soil harbours a huge number of animal species (30% of arthropods live in soil), whether over their entire life or at least during larval stages. Soil offers protection against environmental hazards, such as excess temperature and moisture fluctuations, in particular in arid and cold environments, as well as against predation. Soil provisions food over the year, especially since omnivory seems the rule rather than the execption, and allows reproduction and egg deposition in a safe environment, even for those animals not currently living belowground. Many soil invertebrates, and also some soil vertebrates, are tightly adapted to a subterranean concealed environment, being smaller, blind, depigmented, legfree or with reduced legs, and reproducing asexually, with negative consequences on their colonization rate when the environment is changing at landscape scale. It has been argued that soil could have been a crucible for the evolution of invertebrate terrestrial faunas, as an intermediary step in the transition from aquatic to aerial life.

Soil fauna have been classified, according to increasing body size, in soil microfauna (20 µm to 200 µm), mesofauna (200 µm to 2 mm), macrofauna (2 mm to 2 cm) and megafauna (more than 2 cm). The size of soil animals determines their place along soil trophic networks (soil foodwebs), bigger species eating smaller species (predator-prey interactions) or modifying their environment (nested ecological niches). Among bigger species, soil engineers (e.g. earthworms, ants, termites, moles, gophers) play a prominent role in soil formation and vegetation development, giving them the rank of ecosystem engineers.



SEM image of Milnesium tardigradum in active state

Figure 11: Error case example in Agriculture where the multimodal embedding model Ops-MM-Embedding prioritizes the negative document in the left over the positive document in the right.

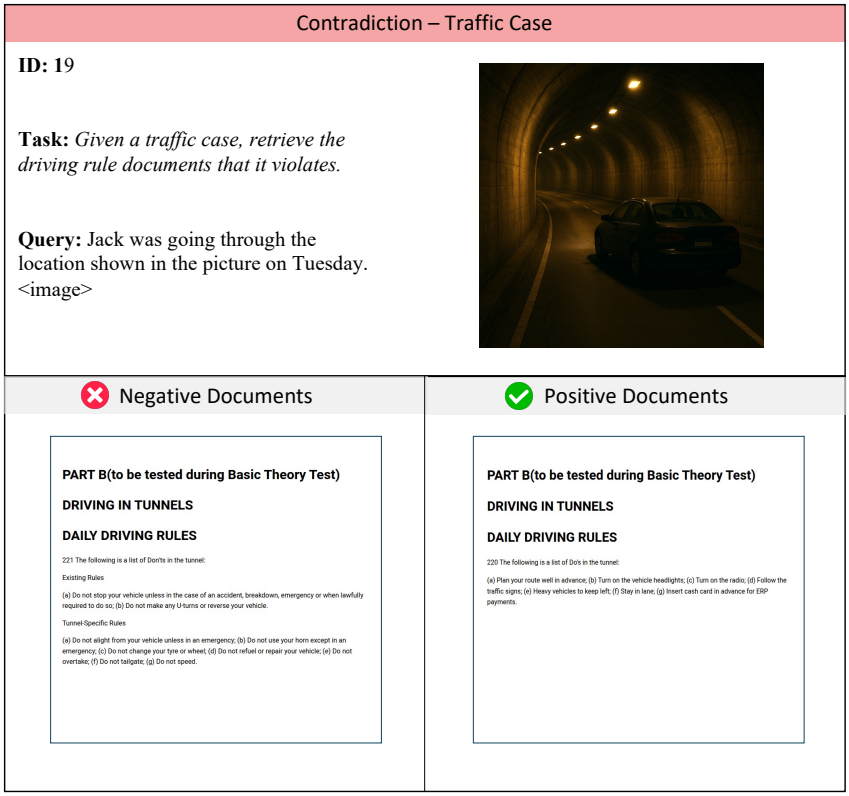



Figure 12: Error case example in Traffic where the multimodal embedding model Ops-MM-Embedding prioritizes the negative document in the left over the positive document in the right.

### F.3 TEST-TIME SCALING IN RETRIEVAL

We conducted query expansion experiments using both weak and strong vision-language models (VLMs)—namely, Qwen2-VL-2B and Qwen2.5-VL-72B—for weak and strong multimodal retrievers (i.e., GME-Qwen2-VL and Ops-MM-Embedding). As shown in Tables 9 and 10, query expansion is generally effective for weak retriever models. However, for stronger retrievers, the quality of the query expansion becomes critical: expansions generated by the weaker VLM actually degrade the performance of the stronger retriever.

With query expansion by a strong LLM (Qwen2.5-VL-72B), the expansion technique is effective for improving both strong and weak retriever models. However, they are still far from perfect on this benchmark. For example, the best retriever with strong query expansion only achieves 55.9 for medical queries and 31.8 for math queries.

Table 9: nDCG@10 scores of the multimodal retriever GME-Qwen2-VL on MRMR *Knowledge* and *Theorem* tasks, comparing the original queries with query expansions generated by Qwen2-VL-2B-Instruct and Qwen2.5-VL-72B-Instruct. The average query length ($Q$ #Text) before and after expansion is reported as the number of tokens measured by the GPT-2 tokenizer.

| | Knowledge | | | | | Theorem | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | $Q$ #Text | Art | Med. | Sci. | Hum. | $Q$ #Text | Math | Phy. | Eng. | Bus. | |
| Original | 31.4 | 54.3 | 40.1 | 46.8 | 45.6 | 58.6 | 28.8 | 36.0 | 30.2 | 45.1 | 40.9 |
| Qwen2-VL-2B | 699.6 | 64.9 | 49.6 | 64.6 | 48.9 | 735.9 | 23.5 | 36.5 | 31.5 | 48.3 | 46.0 |
| Qwen2.5-VL-72B | 843.8 | 76.9 | 61.8 | 77.0 | 72.2 | 1218.4 | 33.3 | 36.9 | 32.7 | 55.0 | 55.7 |

Table 10: nDCG@10 scores of the multimodal retriever Ops-MM-Embedding on MRMR *Knowledge* and *Theorem* tasks, comparing the original queries with query expansions generated by Qwen2-VL-2B-Instruct and Qwen2.5-VL-72B-Instruct. The average query length ($Q$ #Text) before and after expansion is reported as the number of tokens measured by the GPT-2 tokenizer.

| | Knowledge | | | | | Theorem | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | $Q$ #Text | Art | Med. | Sci. | Hum. | $Q$ #Text | Math | Phy. | Eng. | Bus. | |
| Original | 31.4 | 79.3 | 52.5 | 70.0 | 67.8 | 58.6 | 27.7 | 39.5 | 30.1 | 52.3 | 52.4 |
| Qwen2-VL-2B | 699.6 | 77.2 | 45.7 | 67.1 | 58.0 | 735.9 | 24.7 | 35.6 | 29.5 | 50.2 | 48.5 |
| Qwen2.5-VL-72B | 843.8 | 80.5 | 55.9 | 73.1 | 64.6 | 1218.4 | 31.8 | 39.4 | 35.3 | 53.9 | 54.3 |
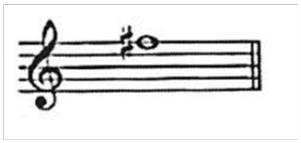
## G DATA EXAMPLES

Art – Music

**ID:** test_Music_327

**Task:** *Retrieve relevant documents that help answer the question.*

**Query:** Determine True or False: This is the dominant in B minor. <image 1>
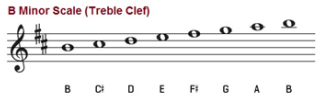


**Answer:** True

Positive Document

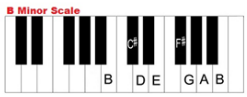
B Minor Scale

This lesson is all about the B minor scale. We will take a look at the three types of minor scale, the natural minor, melodic minor and harmonic minor scales.

Here's a diagram of the B minor scale (Bm scale) on the treble clef.




Here's the B minor scale on piano.




Scale Degrees:

Tonic: B Supertonic: C# Mediant: D Subdominant: E Dominant: F# Submediant: G Subtonic: A Octave: B

The relative major of B minor is D major. Minor keys and their relative major make use of the same notes. The notes of the B minor scale as we've seen are B, C♯, D, E, F♯, G, and A. For the D major scale, it's D, E, F♯, G, A, B and C♯. The difference is the root note of the two scales. The sixth note of a major scale becomes the root note of its relative minor.

You can memorize this formula to form any natural minor scale: whole step – half step – whole step – whole step – half step – whole step – whole step or w – h – w – w – h – w – w. (A whole step skips a key while a half step moves to the next key.) Let's try this with the B minor scale. Let's start on B and move a whole step to C#. From C# move a half step to D. Next, we move a whole step from D to E. From E, let's move a whole step to F#. Next, we go up a half step from F# to G. From G, we move up one whole step to A. Finally, we move a whole step from A to B.

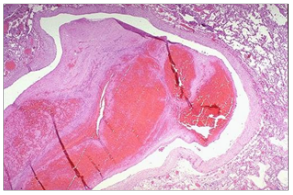
Figure 13: Music example.

Medicine – Clinical Medicine

**ID:** test_Clinical_Medicine_283

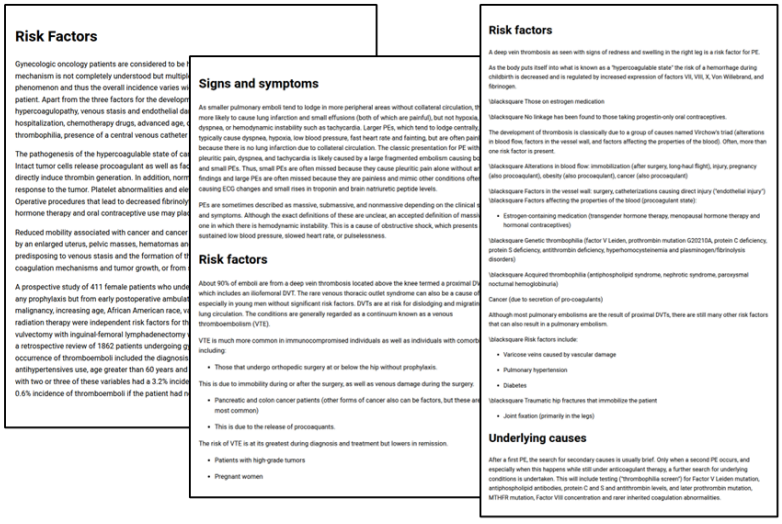
**Task:** *Retrieve relevant documents that help answer the question.*

**Query:** A 61-year-old woman is in the hospital for 2 weeks with bronchopneumonia following surgery for endometrial adenocarcinoma. She then becomes suddenly short of breath. This microscopic appearance from her lung is most typical for which of the following pathologic abnormalities? <image 1>



**Answer:** Thromboembolism

Positive Document



**Risk Factors**

Gynecologic oncology patients are considered to be h
mechanism is not completely understood but multipl
phenomenon and thus the overall incidence varies wi
patient. Apart from the three factors for the developm
hypercoagulopathy, venous stasis and endothelial da
hospitalization, chemotherapy drugs, advanced age, o
thrombophilia, presence of a central venous catheter

The pathogenesis of the hypercoagulable state of ca
Intact tumor cells release procoagulant as well as fac
directly induce thrombin generation. In addition, norm
response to the tumor. Platelet abnormalities and ele
Operative procedures that lead to decreased fibrinoly
hormone therapy and oral contraceptive use may plac

Reduced mobility associated with cancer and cancer
by an enlarged uterus, pelvic masses, hematomas an
predisposing to venous stasis and the formation of th
coagulation mechanisms and tumor growth, or from

A prospective study of 411 female patients who unde
any prophylaxis but from early postoperative ambulat
malignancy, increasing age, African American race, v
radiation therapy were independent risk factors for th
vulvectomy with inguinal-femoral lymphadenectomy
a retrospective review of 1862 patients undergoing gy
occurrence of thromboemboli included the diagnosis
antihypertensives use, age greater than 60 years and
with two or three of these variables had a 3.2% incide
0.6% incidence of thromboemboli if the patient had n

**Signs and symptoms**

As smaller pulmonary emboli tend to lodge in more peripheral areas without collateral circulation, th
more likely to cause lung infarction and small effusions (both of which are painful), but not hypoxia,
dyspnea, or hemodynamic instability such as tachycardia. Larger PEs, which tend to lodge centrally,
typically cause dyspnea, hypoxia, low blood pressure, fast heart rate and fainting, but are often pain
because there is no lung infarction due to collateral circulation. The classic presentation for PE with
pleuritic pain, dyspnea, and tachycardia is likely caused by a large fragmented embolism causing bo
and small PEs. Thus, small PEs are often missed because they cause pleuritic pain alone without ar
findings and large PEs are often missed because they are painless and mimic other conditions ofte
causing ECG changes and small rises in troponin and brain natriuretic peptide levels.

PEs are sometimes described as massive, submassive, and nonmassive depending on the clinical s
and symptoms. Although the exact definitions of these are unclear, an accepted definition of massiv
one in which there is hemodynamic instability. This is a cause of obstructive shock, which presents
sustained low blood pressure, slowed heart rate, or pulselessness.

**Risk factors**

About 90% of emboli are from a deep vein thrombosis located above the knee termed a proximal DV
which includes an iliofemoral DVT. The rare venous thoracic outlet syndrome can also be a cause of
especially in young men without significant risk factors. DVTs are at risk for dislodging and migratin
lung circulation. The conditions are generally regarded as a continuum known as a venous
thromboembolism (VTE).

VTE is much more common in immunocompromised individuals as well as individuals with comorbi
including:

- Those that undergo orthopedic surgery at or below the hip without prophylaxis.

This is due to immobility during or after the surgery, as well as venous damage during the surgery.

- Pancreatic and colon cancer patients (other forms of cancer also can be factors, but these are
most common)
- This is due to the release of procoagunts.

The risk of VTE is at its greatest during diagnosis and treatment but lowers in remission.

- Patients with high-grade tumors
- Pregnant women

**Risk factors**

A deep vein thrombosis as seen with signs of redness and swelling in the right leg is a risk factor for PE.

As the body puts itself into what is known as a "hypercoagulable state" the risk of a hemorrhage during childbirth is decreased and is regulated by increased expression of factors VII, VIII, X, Von Willebrand, and fibrinogen.

\blacksquare Those on estrogen medication

\blacksquare No linkage has been found to those taking progestin-only oral contraceptives.

The development of thrombosis is classically due to a group of causes named Virchow's triad (alterations in blood flow, factors in the vessel wall, and factors affecting the properties of the blood). Often, more than one risk factor is present.

\blacksquare Alterations in blood flow: immobilization (after surgery, long-haul flight), injury, pregnancy (also procoagulant), obesity (also procoagulant), cancer (also procoagulant)

\blacksquare Factors in the vessel wall: surgery, catheterizations causing direct injury ("endothelial injury")
\blacksquare Factors affecting the properties of the blood (procoagulant state)

- Estrogen-containing medication (transgender hormone therapy, menopausal hormone therapy and hormonal contraceptives)

\blacksquare Genetic thrombophilia (factor V Leiden, prothrombin mutation G20210A, protein C deficiency, protein S deficiency, antithrombin deficiency, hyperhomocysteinemia and plasminogen/fibrinolysis disorders)

\blacksquare Acquired thrombophilia (antiphospholipid syndrome, nephrotic syndrome, paroxysmal nocturnal hemoglobinuria)

Cancer (due to secretion of pro-coagulants)

Although most pulmonary embolisms are the result of proximal DVTs, there are still many other risk factors that can also result in a pulmonary embolism.

\blacksquare Risk factors include:

- Varicose veins caused by vascular damage
- Pulmonary hypertension
- Diabetes

\blacksquare Traumatic hip fractures that immobilize the patient

- Joint fixation (primarily in the legs)

**Underlying causes**

After a first PE, the search for secondary causes is usually brief. Only when a second PE occurs, and especially when this happens while still under anticoagulant therapy, a further search for underlying conditions is undertaken. This will include testing ("thrombophilia screen") for Factor V Leiden mutation, antiphospholipid antibodies, protein C and S and antithrombin levels, and later prothrombin mutation, MTHFR mutation, Factor VIII concentration and rarer inherited coagulation abnormalities.

Figure 14: Clinic Medicine example.

Science – Biology

**ID:** test_Agriculture_207

**Task:** *Retrieve relevant documents that help answer the question.*

**Query:** <image 1> With which group are leaf galls like the ones on grapevines, caused by the aphid-like insect called Grape phylloxera - which has root and leaf feeding stages in its lifecycle and is considered an extremely atypical symptom for this insect group - more commonly associated?



**Answer:** Mites

Positive Document



**Dipteran flies**

Some dipteran flies such as the cecidomyiid gall midges Dasineura investita and Neolasioptera boehmeriae, and some Agromyzidae leaf-miner flies cause galls.

Gall of *Japanagromyza inferna* (Agromyzidae) in *Centrosema virginianum*

Nettle gall caused by *Dasineura investita* (Cecidomyiidae)

False nettle stem gall caused by gall midge *Neolasioptera boehmeriae* (Cecidomyiidae)

*Schizomyia impatientis* (Cecidomyiidae) jewelweed flower gall

*Phytoliriomyza pittosporophylli* (Agromyzidae) galls in leaf of *Pittosporum undulatum*

**Mites**

Mites in the family Eriophyidae often cause galls to form on their hosts. The family contains more than 3,000 described species which attack a wide variety of plants.

Lime nail galls caused by the mite Eriophyes tiliae

Galls on purple coneflower caused by an undescribed mite species

Poison ivy leaf galls caused by Aculops rhois

**Nematodes**

Nematodes are microscopic worms that live in soil. Some nematodes (Meloidogyne species or root-knot nematodes) cause galls on the roots of susceptible plants. The galls are often small.



Figure 15: Biology example.

Humanities – Psychology

**ID:** validation_Sociology_1

**Task:** *Retrieve relevant documents that help answer the question.*

**Query:** In 1946, the person in <image 1> was arrested for refusing to sit in the blacks-only section of the cinema in Nova Scotia. This is an example of_____.

**Answer:** A conflict crime



Positive Document

**Conflict criminology**

Largely based on the writings of Karl Marx, conflict criminology holds that crime in capitalist societies cannot be adequately understood without a recognition that such societies are dominated by a wealthy elite whose continuing dominance requires the economic exploitation of others, and that the ideas, institutions and practices of such societies are designed and managed in order to ensure that such groups remain marginalised, oppressed and vulnerable. Members of marginalised and oppressed groups may sometimes turn to crime in order to gain the material wealth that apparently brings equality in capitalist societies, or simply in order to survive. Conflict criminology derives its name from the fact that theorists within the area believe that there is no consensual social contract between state and citizen.

**Discussion**

Conflict theory assumes that every society is subjected to a process of continuous change and that this process creates social conflicts. Hence, social change and social conflict are ubiquitous. Individuals and social classes, each with distinctive interests, represent the constituent elements of a society. As such, they are individually and collectively participants in this process but there is no guarantee that the interests of each class will coincide. Indeed, the lack of common ground is likely to bring them into conflict with each other. From time to time, each element's contribution may be positive or negative, constructive or destructive. To that extent, therefore, the progress made by each society as a whole is limited by the acts and omissions of some of its members by others. This limitation may promote a struggle for greater progress but, if the less progressive group has access to the coercive power of law, it may entrench inequality and oppress those deemed less equal. In turn, this inequality will become a significant source of conflict. The theory identifies the state and the law as instruments of oppression used by the ruling class for their own benefit.

There are various strands of conflict theory, with many heavily critiquing the others. Structural Marxist criminology, which is essentially the most 'pure' version of the above, has been frequently accused of idealism, and many critics point to the fact that the Soviet Union and such states had as high crime rates as the capitalist West. Furthermore, some highly capitalist states such as Switzerland have very low crime rates, thus making structural theory seem improbable.

Figure 16: Psychology example.

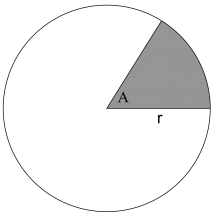
Theorem – Math

**ID:** test_Math_16

**Task:** *Retrieve relevant theorems that are involved in solving the problem.*

**Query:** <image 1> The radius of the circle above is 4 and ∠A=45°. What is the area of the shaded section of the circle?

**Answer:** $2\pi$




Positive Document

## Area of Sector

**Tags**: Circles, Geometry, Area of Sector, Area Formulas

**Theorem**
Let $\mathcal{C} = ABC$ be a circle whose center is $A$ and with radii $AB$ and $AC$.
Let $BAC$ be the sector of $\mathcal{C}$ whose angle between $AB$ and $AC$ is $\theta$.

Then the area $\mathcal{A}$ of sector $BAC$ is given by:

$$\mathcal{A} = \frac{r^2\theta}{2} \tag{1}$$

where:

- $r = AB$ is the length of the radius of the circle,
- $\theta$ is measured in radians.

**Proof**

> ***Note**: To be replaced with something rigorous, based on calculus.*

From *Area of Circle*, the area of $\mathcal{C}$ is $\pi r^2$.
From *Full Angle measures* $2\pi$ *Radians*, the angle within $\mathcal{C}$ is $2\pi$.
The fraction of the area of $\mathcal{C}$ within the sector $BAC$ is therefore:

$$\pi r^2 \times \frac{\theta}{2\pi} \tag{2}$$

Hence the result.
■

## Area of Circle

**Tags**: Circles, Area of Circle, Euclidean Geometry, Area Formulas

**Theorem**
The area $A$ of a circle is given by:

$$A = \pi r^2 \tag{1}$$

where $r$ is the radius of the circle.

**Proof**

We start with the equation of a circle:

$$x^2 + y^2 = r^2 \tag{2}$$

Thus $y = \pm\sqrt{r^2 - x^2}$. From the geometric interpretation of the definite integral, the area $A$ is:

$$\begin{aligned} A &= \int_{-r}^{r} \left[\sqrt{r^2 - x^2} - \left(-\sqrt{r^2 - x^2}\right)\right] dx \\ &= \int_{-r}^{r} 2\sqrt{r^2 - x^2}\, dx \\ &= \int_{-r}^{r} 2r\sqrt{1 - \frac{x^2}{r^2}}\, dx \end{aligned} \tag{3}$$

Let $x = r\sin\theta$ (note that this substitution is valid because $-r \le x \le r$).
Then $\theta = \arcsin\left(\frac{x}{r}\right)$ and $dx = r\cos\theta\, d\theta$.

Applying integration by substitution:

$$\begin{aligned} A &= \int_{\arcsin(-1)}^{\arcsin(1)} 2r^2\sqrt{1 - \frac{(r\sin\theta)^2}{r^2}}\cos\theta\, d\theta \\ &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2r^2\sqrt{1 - \sin^2\theta}\cos\theta\, d\theta \\ &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2r^2\sqrt{\cos^2\theta}\cos\theta\, d\theta \\ &= r^2\left[\theta + \frac{1}{2}\sin(2\theta)\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \\ &= \pi r^2 \end{aligned} \tag{4}$$

■

Figure 17: Math example.

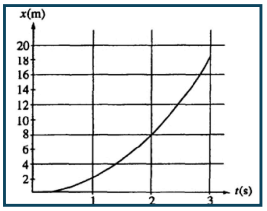
Theorem – Physics

**ID:** test_Physics_265

**Task:** *Retrieve relevant theorems that are involved in solving the problem.*

**Query:** <image 1>The graph above represents position $x$ versus time $t$ for an object being acted on by a constant force. The average speed during the interval between 1s and 2s is most nearly

**Answer:** 6 m/s




Positive Document

## Length of a Real Interval

**Definition**

Let any of the following denote a real interval:

- Closed interval: $[a, b]$
- Half-open interval (right): $[a, b)$
- Half
- Oper

The **length** of

This definitio

## Displacement

**Definition**

The **(physical) displacement** of a body is a measure of its position relative to a given point of reference within a specific frame

Displacement i
(the orientatio

In **Cartesian c**
example, in th

represents a po
origin—thus fu

## Speed

**Definition**

The **speed** of a body is a measure of the magnitude of its velocity, independent of direction.

Because it disregards direction, speed is a **scalar quantity**.

Mathematically, if $\mathbf{v}$ is the velocity vector of a body, then its speed $s$ is given by:

$$s = \|\mathbf{v}\|$$

where $\|\cdot\|$ denotes the magnitude (or norm) of the vector.

Figure 18: Physics example.

Theorem – Engineering

**ID:** validation_Electronics_20

**Task:** *Retrieve relevant theorems that are involved in solving the problem.*

**Query:** In <image 1>. $v_c = \sin(2\pi T)$ Find an expression for $i$ and calculate $i$ at the instants t = 0.

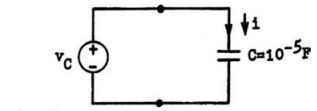
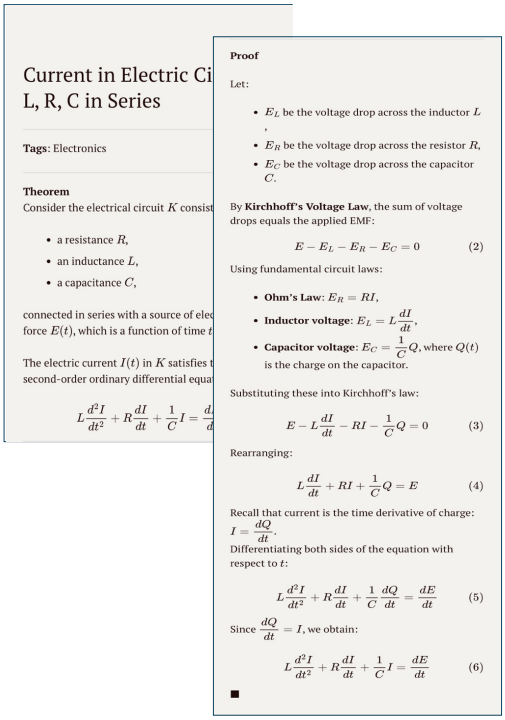



Fig. 1

**Answer:** $2\pi\times 10^{-5}$ A

Positive Document



Current in Electric Ci
L, R, C in Series

**Tags**: Electronics

**Theorem**
Consider the electrical circuit $K$ consist

- a resistance $R$,
- an inductance $L$,
- a capacitance $C$,

connected in series with a source of elec
force $E(t)$, which is a function of time $t$

The electric current $I(t)$ in $K$ satisfies t
second-order ordinary differential equa

$$L\frac{d^2I}{dt^2} + R\frac{dI}{dt} + \frac{1}{C}I = \frac{d}{d}$$

**Proof**

Let:

- $E_L$ be the voltage drop across the inductor $L$,
- $E_R$ be the voltage drop across the resistor $R$,
- $E_C$ be the voltage drop across the capacitor $C$.

By **Kirchhoff's Voltage Law**, the sum of voltage drops equals the applied EMF:

$$E - E_L - E_R - E_C = 0 \tag{2}$$

Using fundamental circuit laws:

- **Ohm's Law**: $E_R = RI$,
- **Inductor voltage**: $E_L = L\frac{dI}{dt}$,
- **Capacitor voltage**: $E_C = \frac{1}{C}Q$, where $Q(t)$ is the charge on the capacitor.

Substituting these into Kirchhoff's law:

$$E - L\frac{dI}{dt} - RI - \frac{1}{C}Q = 0 \tag{3}$$

Rearranging:

$$L\frac{dI}{dt} + RI + \frac{1}{C}Q = E \tag{4}$$

Recall that current is the time derivative of charge: $I = \frac{dQ}{dt}$.
Differentiating both sides of the equation with respect to $t$:

$$L\frac{d^2I}{dt^2} + R\frac{dI}{dt} + \frac{1}{C}\frac{dQ}{dt} = \frac{dE}{dt} \tag{5}$$

Since $\frac{dQ}{dt} = I$, we obtain:

$$L\frac{d^2I}{dt^2} + R\frac{dI}{dt} + \frac{1}{C}I = \frac{dE}{dt} \tag{6}$$

■

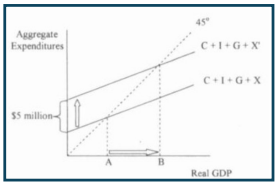
Figure 19: Engineering example.

Theorem – Business

**ID:** test_Economics_157

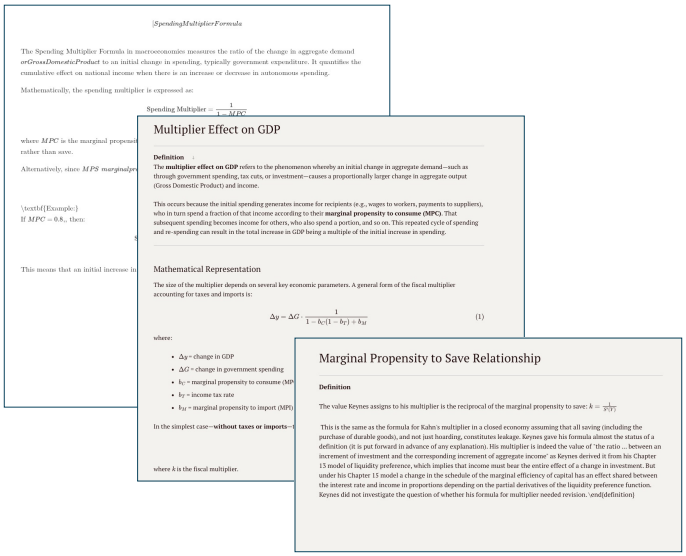
**Task:** *Retrieve relevant theorems that are involved in solving the problem.*

**Query:** Presuming that the Marginal Propensity to Consume (MPC) is 0.8, could you calculate the monetary value signified by the interval A-B in the aforementioned diagram <image 1>?




**Answer:** $25 million.

Positive Document



\SpendingMultiplierFormula

The Spending Multiplier Formula in macroeconomics measures the ratio of the change in aggregate demand or Gross Domestic Product to an initial change in spending, typically government expenditure. It quantifies the cumulative effect on national income when there is an increase or decrease in autonomous spending.

Mathematically, the spending multiplier is expressed as:

$$\text{Spending Multiplier} = \frac{1}{1 - MPC}$$

where $MPC$ is the marginal propensit
rather than save.

Alternatively, since $MPS$ $marginalpr$

\textbf{Example:}
If $MPC = 0.8$,, then:

This means that an initial increase in

Multiplier Effect on GDP

**Definition**

The **multiplier effect on GDP** refers to the phenomenon whereby an initial change in aggregate demand—such as through government spending, tax cuts, or investment—causes a proportionally larger change in aggregate output (Gross Domestic Product) and income.

This occurs because the initial spending generates income for recipients (e.g., wages to workers, payments to suppliers), who in turn spend a fraction of that income according to their **marginal propensity to consume (MPC)**. That subsequent spending becomes income for others, who also spend a portion, and so on. This repeated cycle of spending and re-spending can result in the total increase in GDP being a multiple of the initial increase in spending.

Mathematical Representation

The size of the multiplier depends on several key economic parameters. A general form of the fiscal multiplier accounting for taxes and imports is:

$$\Delta y = \Delta G \cdot \frac{1}{1 - b_C(1 - b_T) + \lambda_M} \quad (1)$$

where:

- $\Delta y$ = change in GDP
- $\Delta G$ = change in government spending
- $b_C$ = marginal propensity to consume (MP
- $b_T$ = income tax rate
- $b_M$ = marginal propensity to import (MPI

In the simplest case—**without taxes or imports**—t

where $k$ is the fiscal multiplier.

Marginal Propensity to Save Relationship

**Definition**

The value Keynes assigns to his multiplier is the reciprocal of the marginal propensity to save: $k = \frac{1}{s \cdot (r)}$

This is the same as the formula for Kahn's multiplier in a closed economy assuming that all saving (including the purchase of durable goods), and not just hoarding, constitutes leakage. Keynes gave his formula almost the status of a definition (it is put forward in advance of any explanation). His multiplier is indeed the value of 'the ratio ... between an increment of investment and the corresponding increment of aggregate income' as Keynes derived it from his Chapter 13 model of liquidity preference, which implies that income must bear the entire effect of a change in investment. But under his Chapter 15 model a change in the schedule of the marginal efficiency of capital has an effect shared between the interest rate and income in proportions depending on the partial derivatives of the liquidity preference function. Keynes did not investigate the question of whether his formula for multiplier needed revision. \end{definition}

Figure 20: Business example.

Contradiction – Negation

**ID:** 340894

**Task:** *Retrieve the text that has contradictory information to the image.*

**Query:** <image 1>



Positive Document

This image includes mouse, dining table, book but no keyboard.

Negative Documents

This image includes mouse, tv, laptop but no bottle.

This image includes mouse, cell phone, laptop but no refrigerator.

This image includes cell phone, person, chair but no bottle.

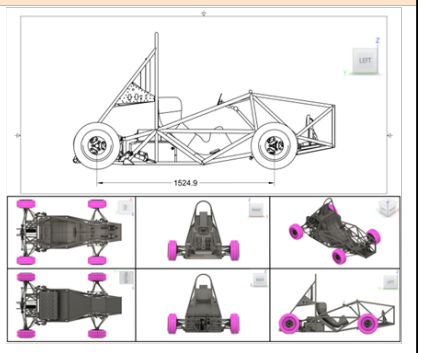
Figure 21: Negation example.

Contradiction – Vehicle Design

**ID:** 0

**Task:** *Given a vehicle design as the query, retrieve the design requirements that it violates,.*

**Query:** Attached is an image that shows an engineering drawing of the vehicle … All units displayed in the engineering drawing have units of mm. <image>




Positive Document

**V - VEHICLE REQUIREMENTS**

**V.1 CONFIGURATION**

The vehicle must be open wheeled and open cockpit (a formula style body) with four wheels that are not in a straight line.

**V.1.2 Wheelbase**

The vehicle must have a minimum wheelbase of 1525 mm

Figure 22: Vehicle Design example.

Contradiction – Traffic Case

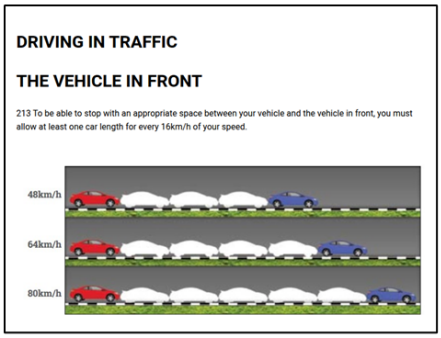
**ID:** 27

**Task:** *Given a traffic case as the query, retrieve the driving rule document that it violates.*

**Query:** In Singapore, Ginny was driving with the speed of 64 km/h, keeping a 3-meter gap behind the silver car, as shown in the picture. <image>




Positive Document



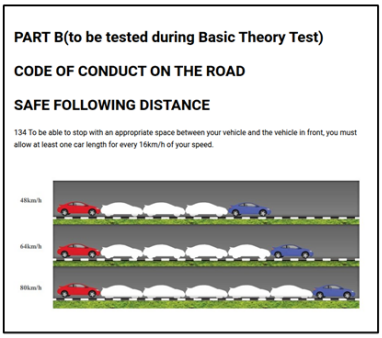




Figure 23: Traffic Case example.