

A EXAMPLE OF TRAINING PAIRS FOR RETRIEVAL MODULE

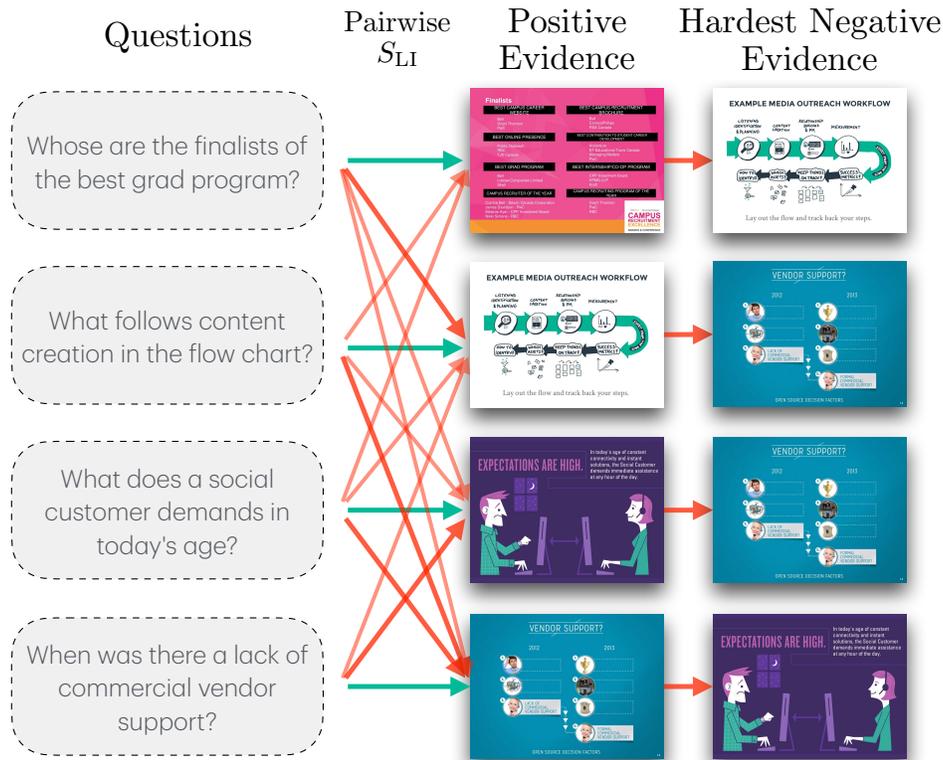


Figure A.1: Example of training pairs within a batch (batch size: 4) for contrastive training, using samples from the SlideVQA dataset.

B EVALUATION METRICS

We evaluate the model’s performance on evidence retrieval and question-answering using five metrics explained as follows:

Top-k Accuracy In our experiment, we focus on questions that have evidence from a single page. We use top-k accuracy to evaluate retrieval methods, which measures the percentage of times the evidence image appears within the top k most similar images.

Exact Match Following (Tanaka et al., 2023), we report exact match (EM) frequency between generated answers and the ground truth, allowing for case insensitivity and extra spaces. While effective for fine-tuned models, this metric is less suited for LLM responses, which often include full sentences. Correct answers with extra context may thus be unfairly penalized.

Generalized Accuracy We report generalized accuracy (G-Acc) from MMLongBench-Doc (Ma et al., 2024d), a GPT-dependent, rule-based evaluation protocol. Model responses are simplified using GPT-4o and scored based on answer-type-specific rules. However, G-Acc has two limitations: it introduces randomness from GPT’s stochastic outputs and relies on answer-type annotations, limiting its applicability across datasets.

ANLS Average Normalized Levenshtein Similarity (ANLS) (Biten et al., 2019) measures the similarity between predicted and ground truth text using the Levenshtein distance, normalized by the longer string’s length. It outputs a similarity score between 0 and 1. ANLS allows mismatches, insertions, and deletions making it useful for OCR and document understanding tasks when exact matches are not required.

PNLS The *partial normalized Levenshtein similarity* (PNLS) (Chen et al., 2024b) generalizes ANLS by not penalizing extra prefixes or suffixes while allowing mismatches, insertions, and deletions within the matched region. This makes it more suitable for evaluating LLM responses, which are often verbose to improve user experience.

The PNLs metric is formally defined as follows: String $\mathcal{T}_{1,m} = t_1 \dots t_m$ represents the true answer and $\mathcal{S}_{1,n} = s_1 \dots s_n$ is a model generated string. We first use using the approximate string matching algorithm (Sellers, 1980) to identify the sub-string of \mathcal{S} that has the minimum edit distance to \mathcal{T} . Specifically, we first construct a scoring matrix \mathbf{F} of size $(m + 1) \times (n + 1)$, where $F_{i,j}$ stores the smallest edit distance between the i -prefix $\mathcal{T}_{1,i}$ and any sub-string $\mathcal{S}_{x,j}, \forall x \in \{1, \dots, j - 1\}$ that ends at position j . The scoring matrix can be computed recursively

$$F_{i,j} = \begin{cases} 0 & \text{if } i = 0 \\ m & \text{if } j = 0 \\ \min \begin{pmatrix} F_{i-1,j-1} + c(t_i, s_j) \\ F_{i-1,j} + 1 \\ F_{i,j-1} + 1 \end{pmatrix} & \text{otherwise,} \end{cases}$$

where c is the substitution cost that takes a value of 0 if $t_i = s_j$ and 1 otherwise. Once \mathbf{F} is computed, the minimum value in the last row is the optimal edit distance and the end index of the matched sub-string $j' = \arg \min_j (F_{m+1,j})$. The start index i' can be found by tracing back the the computation of Eq.(B) using $\arg \min$ operation. Finally, the PNLs is computed as: $m / (m + j' - i' + 1)$. In our experiments we use binary cost function: $c(t_i, s_j) = 0$ if $t_i = s_j$ else $c(t_i, s_j) = 1$

C EXAMPLE OF INFERENCE FAILURE SCENARIO

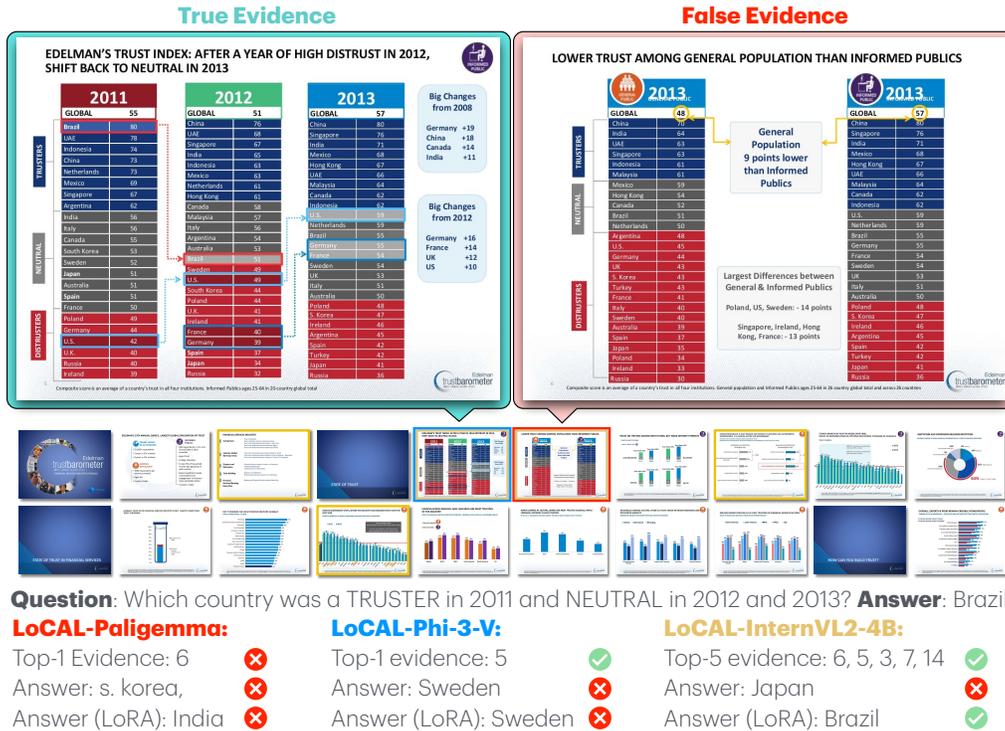


Figure C.1: Inference example of a challenging case in the SlideVQA dataset. SV-RAG-Paligemmma retrieved the wrong evidence page due to limitations in its retrieval module, leading to an incorrect answer. SV-RAG-Phi-3-V retrieved the correct page but provided a wrong answer due to limitations in its QA module. Meanwhile, SV-RAG-InternVL2-4B also assigned the highest relevance score to an incorrect page. However, since it processes multiple pages (top 5), the correct evidence page was included in the input, allowing its fine-tuned QA module to deliver the correct answer.

D QAULITATIVE RESULTS IN QUESTION-ANSWERING

Question: What is Indonesia's GDP in billions of dollars? **Answer:** 868.3

GPT-4o: \$868.3 billion

Phi-3-V: Indonesia's GDP is \$868.3 billion.

InternVL2-4B: Indonesia's GDP is \$868.3 billion.

Paligemma: \n868.3

Phi-3-V (LoRA): 868.3 \n

InternVL2-4B (LoRA): \$868.3 billion



True Evidence Page

Question: What are five audience targeting methods?

Answer: Contextual, Behavioral, Demographic, Geographic, Retargeted

GPT-4o: 1. Contextual\n2. Behavioral\n3. Demographic\n4. Geographic\n5. Retargeted

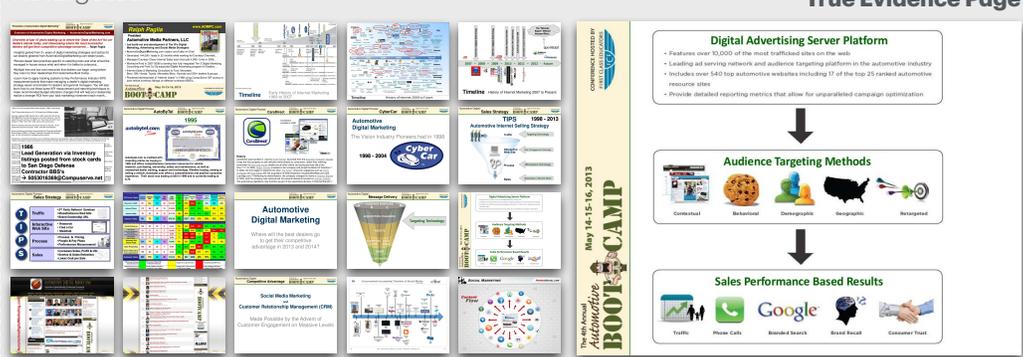
Paligemma: \ncontextual, behavioral, demographic, geographic, retargeting

Phi-3-V: \n1. Traffic\n2. Phone Calls\n3. Branded Search\n4. Brand Recall\n5. Consumer Trust

Phi-3-V (LoRA): Contextual, Behavioral, Demographic, Geographic, Retargeted \n

InternVL2-4B: The five audience targeting methods are:\n1. Demographic\n2. Geographic\n3. Behavioral\n4. Psychographic\n5. Geographic

InternVL2-4B (LoRA): Contenxtual, Behavioral, Demographic, Geographic, Retargeted



True Evidence Page

Figure D.1: Question answering examples on the SlideVQA dataset using different QA modules. Models without fine-tuning, such as Phi-3-V and InternVL2-4B, tend to produce verbose and error-prone responses. However, in the second example, fine-tuning with the LoRA adapter significantly improves the accuracy of Phi-3-V and InternVL2-4B.

10.3. COMMUNICATIONS MODEL 95

services such as name services, task and sessions management, and distributed data services for inventory management. The main feature of the generic approach is the common distributed applicability of framework functions as well as framework executables from the standard user interface.

The provided interfaces for runtime-integration supports in-process-interfaces for plugins and inter-process interfaces for framework tools and wrapper calls.

Figure 10.6: ctyc distributed components

Thus an integrated custom application is capable for distributed operations by default, where a basic version management for distributed compatibility issues is integrated.

10.3 Communications Model

The communications between entities within the UnifiedSessionsManager is modeled around the basic idea of transparent access to user desktops for local and remote login with extensions for pure command line sessions. This results in **IO-stream based communications** which is limited but quite simple to implement.

The consequence of this approach is a resulting 2-type category of communications. The first is the forwarding of the display to the user for remote actions, either complete remote desktops or simple remote shells as the managing entity for user sessions. The second is the local execution of the user interface for the management of current session, where the local sessions client opens communications to remote services.

The resulting architecture is a layered architecture modeled as an independent communications service layer and a sessions management layer which is basically peer-to-peer oriented only. Thus the communications is handled on high level abstraction with an addressing abstraction called `<machine-address>`. The remaining knowledge of communications type for the sessions layer are the two types as basic pattern

DISPLAYFORWARDING

Question: In the provided communications model, what enables the communication between the workstation and the server or remote workstation?

True Answer: The communication between the workstation and the server or remote workstation is enabled by **SSH**. ✓

Text only answer: In the provided communications model, communication between the workstation and the server or remote workstation is enabled by **IO-stream based communications**. ✗

Figure E.2: Example question-and-answer pair from the VisR-Bench dataset, highlighting the reliance on both image and surrounding text for accurate responses.

F COMPARISON OF RETRIEVAL METHOD EFFICIENCY

Text Extraction		Text Encoding		Multimodal Encoding	
PaddleOCR	0.275	BM25	0.0001	CLIP	0.022
		BGE-m3	0.131	SigLip	0.109
PDF Parser	0.762	BGE-large	0.137	Col-Paligemma	0.140
		NV-embed-v2	0.117	Col-Phi-3-V	0.230
				Col-InternVL2	0.581

Table F.1: Per-page time cost of retrieval methods: The left table presents time cost (seconds) of text-based methods that rely on text extraction techniques, such as OCR models, followed by text encoders to compute page embeddings. The right table presents time cost (seconds) of multi-modal methods that encode the entire page as an image.

G ADDITIONAL EXPERIMENT RESULTS

We compare our method with text-only baselines using a document parser³ to highlight the advantages of MLLMs in multi-modal understanding. QA results are reported for the VisR-Bench and MMLongBench-Doc datasets, where PDF files are available.

Table G.1 presents QA results on VisR-Bench and MMLongBench-Doc datasets. To evaluate answer quality for VisR-Bench, where true answers are long and detailed, we introduce the Mean GPT Score (MGS), as string-matching methods often penalize variations in wording for lengthy answers. Instead, we prompt GPT-4o to compare a model’s answer with the ground truth and assign a binary score based on detail alignment.

QA Module	Retrieval Module	Evidence	VisR-B MGS	MMLong G-Acc
<i>Text only QA methods</i>				
Phi-3 + parser	Col-Phi-3-V	R5	14.1	29.2
GPT-4o + parser	Col-Phi-3-V	R5	24.9	43.2
GPT-4o + parser	-	A	27.6	42.4
<i>MLLM QA models</i>				
PaliGemma	Col-PaliGemma	R1	12.2	23.9
Phi-3-V	Col-Phi-3-V	R1	24.2	30.7
SV-RAG-InternVL2	Col-InternVL2	R5	25.2	33.2
GPT-4o	Col-Phi-3-V	R5	47.2	55.1
GPT-4o	-	A	43.2	54.5

Table G.1: parser results

Our results indicate that using image evidence consistently outperforms text-only evidence. On VisR-Bench, text-only baselines showed a significant performance drop, emphasizing the dependency of questions on both image and text. However, the MGS of text-only baselines is not zero, likely because the model leverages text from a broader context rather than relying solely on the surrounding text, enabling it to extract relevant information even in the absence of visual input.

Additionally, reducing input pages with the retrieval module improved GPT-4o’s performance with image evidence, aligning with the findings in Table 2. In contrast, retrieval did not enhance GPT-4o’s performance on VisR-Bench in the text-only setting, likely because the evidence pages lacked sufficient information to fully address the questions. Including additional context in such cases might yield better results.

³Adobe Extract API: <https://developer.adobe.com/document-services/apis/pdf-extract/>