

Can Your Agent Read the Source Faithfully? Benchmarking MLLM Agents on Source-Grounded Multimodal Retrieval

Wentao Gao^{1,*‡} Shizhou Huang^{2,*} Junling Zhuang³ Advitya Garg¹
Mohammad Hasibur Rahman¹ Mengliang Zhang¹ Qin Qin^{4, §1} Yu Yu^{5, §2}

¹University of Texas at Arlington ²East China Normal University

³Georgia Institute of Technology ⁴Meta ⁵Hark

wentao.gao@uta.edu 52275901004@stu.ecnu.edu.cn jzhuang48@gatech.edu

{axg0826,mhr9808,mxz3935}@mavs.uta.edu qinqin@meta.com yu.yu@hark.com

^{*}Equal contribution. [‡]Corresponding author.

^{§1}Work done independently of the author’s role at Meta.

^{§2}Work done independently of the author’s role at Hark.

Abstract

Multimodal agents are increasingly evaluated on their ability to search the web, invoke tools, and solve complex tasks autonomously. Yet no existing benchmark systematically evaluates the capability of autonomously searching, parsing, and extracting precise answers from multimodal raw media files as a unified, end-to-end agentic task. We introduce **RDQA (Raw Data Question Answering)**, a benchmark of 1,350 questions across three source modalities (document, video, audio), and 20 real-world task scenarios, where every answer is designed to require source access rather than model knowledge alone. RDQA provides capability-stratified evaluation through 9 capability families organized by modality and cognitive complexity, and introduces the *Parametric Leakage Rate (PLR)*, a process-aware metric that flags correct answers produced without genuine evidence retrieval. We evaluate 12 agent systems spanning 6 backbone models and 2 scaffoldings. Our results show that the the best generic agent reaches only 24.0% adjusted accuracy, while a process-aware baseline (RDQA-Skill) reaches 41.1%. PLR reaches as high as 96.0%, indicating that many ostensibly correct answers are not grounded in retrieved evidence. We release the full dataset, evaluation framework, and a baseline to support future research.¹

1 Introduction

Large language models (LLMs) and LLM-powered agents have made substantial progress in information seeking, web browsing, tool use, and complex task solving. Recent web-agent benchmarks show that agents can navigate websites, click links, scroll pages, fill forms, and perform multi-step interactions in realistic web environments (Zhou et al., 2024; Koh et al., 2024). Meanwhile, multimodal large language models (MLLMs) have also advanced rapidly in document, video, and audio

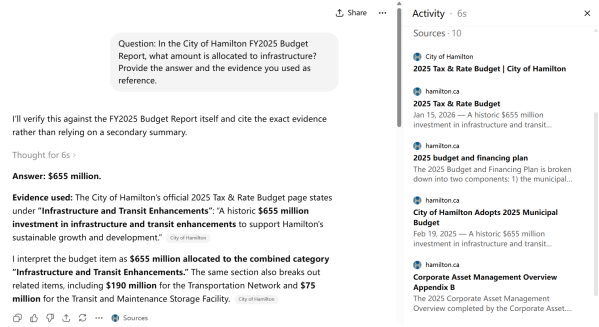


Figure 1: An RDQA failure example.

understanding, providing the perceptual foundation for agents to process heterogeneous information sources (Dong et al., 2025; Sakshi et al., 2025; Wei et al., 2026).

However, current agents still rely on an implicit assumption: the information needed to answer a query is often available from indexed web text, search snippets, or human-curated secondary sources. In many real-world information-seeking scenarios, this assumption does not hold. Critical facts are often embedded inside raw source artifacts, such as government budget PDFs, procurement documents, financial reports, technical manuals, video frames, scrolling captions, and audio recordings. Answering such queries requires more than retrieving a relevant webpage. An agent must identify the correct raw source, invoke appropriate tools to parse it, localize fine-grained evidence within complex structures, and generate an answer strictly grounded in that evidence.

Figure 1 illustrates a typical source-routing failure. For the Hamilton FY2025 budget query, the model claims to verify the official report but instead relies on a high-level Hamilton.ca webpage and returns an aggregate infrastructure-related amount. The correct answer requires locating the City of Hamilton, Ohio FY2025 Budget Report and extracting the standalone “Infrastructure” allocation, 12.5 million. This example shows how an appar-

¹<https://github.com/Dextergao14/RDQA>

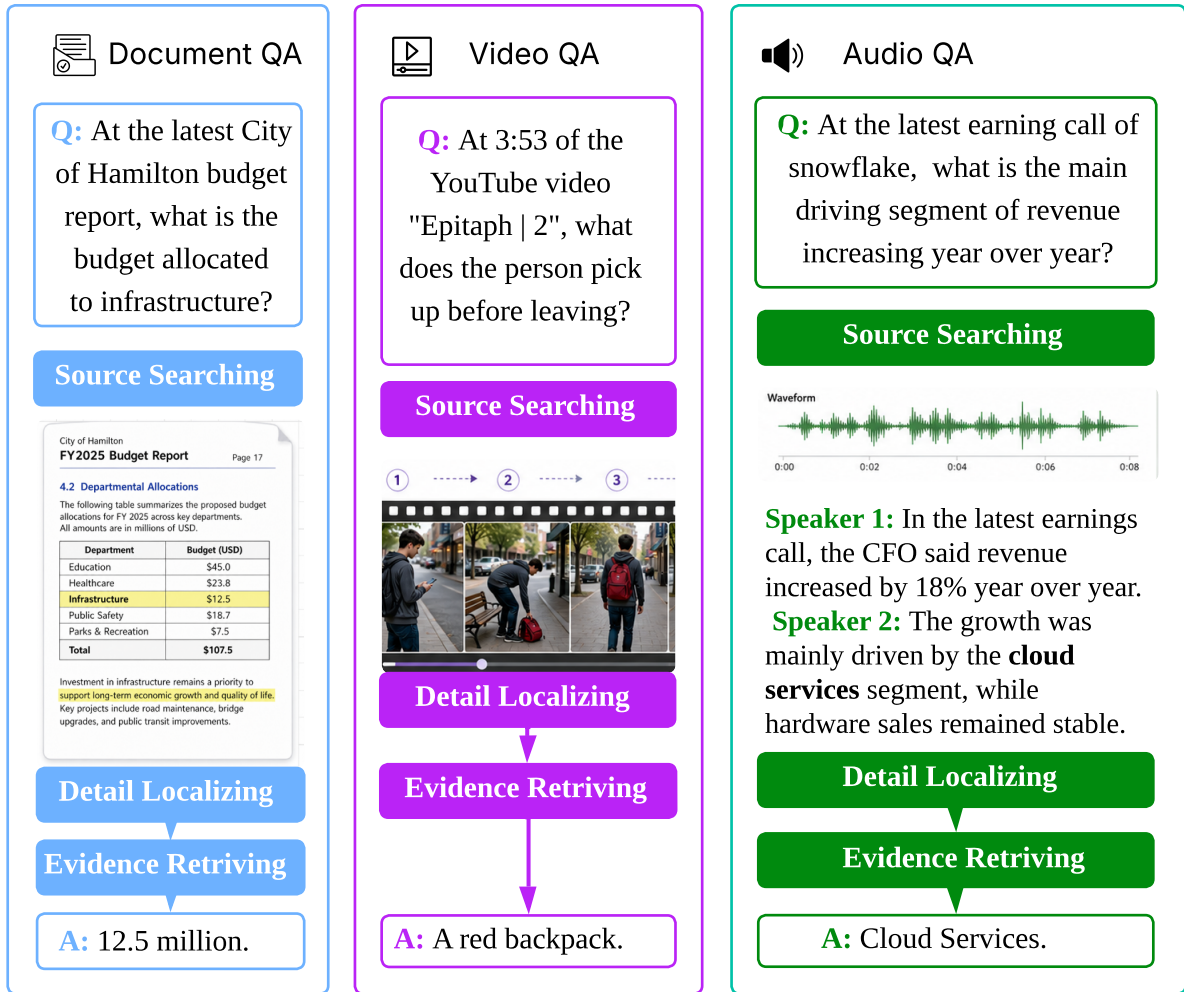


Figure 2: Sample questions of RDQA.

ently relevant webpage can mask failures in entity disambiguation, source routing, evidence localization, and grounding.

To fill this gap, we introduce **RDQA (Raw Data Question Answering)**, a benchmark designed to systematically evaluate agents on this full pipeline. RDQA targets information that is fine-grained, long-tail, and tied to specific factual artifacts for which AI models’ knowledge offers no guarantee, regardless of the modalities and data volume a model was trained on. Reliable answers require live retrieval and grounding in the original source. Therefore every question from RDQA is constructed to be unanswerable from parametric knowledge of a model alone, see § 3.4.

The main contributions of this work are summarized as follows:

1. **RDQA benchmark:** a novel benchmark to systematically evaluate agents on the full raw-data retrieval-and-extraction pipeline, with

1,350 questions across 3 modalities, 20 task scenes, and 9 capability families (§ 3.1).

2. **Empirical findings:** evaluating 12 agent systems spanning 6 backbone models and 2 agentic architectures, we find that current agents remain far from solving reliable source-grounded QA. The best-performing agent reaches only 24.0% PLR-adjusted accuracy, and the Parametric Leakage Rate (PLR) reaches as high as 96.0%. Metric acronyms are defined in § 3.4.
3. **Human and agentic baselines:** we measure average human performance on RDQA, and propose a reference baseline tailored to RDQA-style questions that combines chain-of-thought scope narrowing, targeted tool invocation, and reflective self-verification, establishing a reproducible performance anchor for the community.

2 Related Work

2.1 Question Answering and Multimodal Understanding

A substantial body of work evaluates models on comprehending *given* documents. DocVQA (Mathew et al., 2021) and InfographicsVQA (Mathew et al., 2022) test question answering over single document images. ChartQA (Masry et al., 2022) focuses on chart comprehension. MP-DocVQA (Kang et al., 2024) extends to multi-page settings, and M3DocVQA (Cho et al., 2024) further scales to open-domain retrieval across 3,000+ PDF documents.

On the video side, existing work largely treats video as a modality for semantic understanding rather than as an answer-bearing artifact. MVBench (Li et al., 2024) and VideoMME (Fu et al., 2025) evaluate temporal visual understanding. These benchmarks emphasize high-level semantics, such as recognizing what is happening in a video, rather than fine-grained information extraction, such as reading the amount of a seasoning ingredient shown in a cooking video.

2.2 Retrieval-Augmented Generation Benchmarks

The retrieve-then-generate paradigm is structurally analogous to RDQA’s pipeline: an agent retrieves evidence from external sources before generating a grounded answer. A growing body of RAG benchmarks evaluates this paradigm. CRAG (Yang et al., 2024b) introduces 4,409 factual questions across five domains and varying entity popularity, exposing a substantial gap between LLM accuracy and trustworthy QA even with retrieval. RAG-Bench (Friel et al., 2024) provides explainable evaluation labels across industry corpora such as user manuals.

However, RAG benchmarks share a common assumption that fundamentally separates them from RDQA: the retrieval corpus is *pre-built and pre-indexed*. Documents have already been collected, chunked, and embedded into a vector store, the agent’s task is to *query* this index. RDQA inverts this assumption.

2.3 Autonomous Agent Benchmarks

GAIA (Mialon et al., 2024) proposes 466 real-world questions requiring multi-step reasoning, web browsing, and tool use, and reports a striking performance gap between humans (92%) and

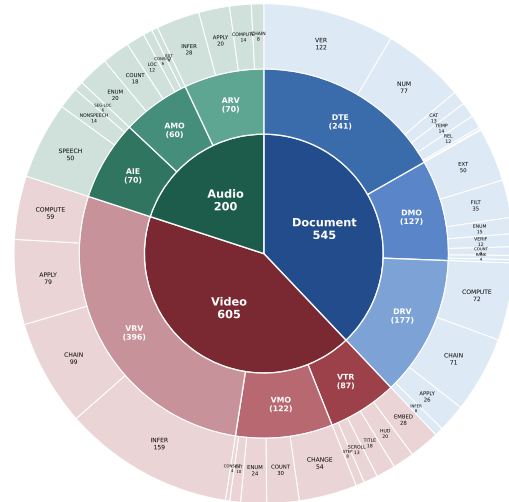


Figure 3: Question distribution across modalities, capability families, and subcategories. Inner ring: modality. Middle ring: capability family. Outer ring: subcategory. See Table 2 for full category and Appendix A.2 for full subcategory definitions.

GPT-4 with plugins (15%). However, GAIA treats raw file processing as one capability among many, and the majority of its tasks can be answered from web text alone. Table 1 compares RDQA with existing benchmarks in further detail.

3 The RDQA Benchmark

3.1 Task Definition

To rigorously assess and foster advancements in MLLM agents against these challenges, RDQA, a comprehensive benchmark is formulated. Figure 2 showcases representative examples, illustrating the specific nature and scope of each task. Formally, each question requires the agent to complete a four-stage pipeline:

- (1) **Source routing:** locate and identify the target raw file on the open web, bypassing secondary summaries, news digests.
- (2) **Tool invocation:** select and execute the appropriate tool(s) to download/buffer and parse the file, e.g. PDF text extractors, video frame samplers, audio transcribers, spreadsheet readers.
- (3) **Detail localization:** perform fine-grained spatial or temporal reasoning to isolate the precise evidence, i.e. the correct page, table row, video timestamp, or audio segment;
- (4) **Parametric knowledge suppression:** ground the final answer strictly in retrieved evidence.

Benchmark	Modality			Source	Tool	Detail	PLR	#Q
	Document	Video	Audio	Routing	Use	Local.	Detect.	
Document understanding benchmarks (source file provided)								
MMLongBench-Doc (Ma et al., 2024)	✓					✓		1K
MMDocIR (Dong et al., 2025)	✓			○		✓		1.7K
Video understanding benchmarks (source file provided)								
Video-MME (Fu et al., 2025)		✓						2.7K
CFVBench (Wei et al., 2026)		✓		○		✓		5K
Audio understanding benchmarks (source file provided)								
AIR-Bench (Yang et al., 2024a)			✓					19K
MMAU (Sakshi et al., 2025)			✓					10K
Agent benchmarks (autonomous web / tool use)								
WebArena (Zhou et al., 2024)				○	✓			812
VisualWebArena (Koh et al., 2024)	○			○	✓			910
AgentBench (Liu et al., 2024)				○	✓			1.4K
OSWorld (Xie et al., 2024)				○	✓			369
AssistantBench (Yoran et al., 2024)	○			✓	✓			214
Dyn-VQA (Li et al., 2025)	○			○	○			1.5K
GAIA (Mialon et al., 2024)	○	○	○	✓	✓	○		466
RDQA (ours)	✓	✓	✓	✓	✓	✓	✓	1.35K

Table 1: Comparison of RDQA with prior benchmarks across the four stages of source-grounded raw-data QA. Source Routing, Tool Use, Detail Localization, and PLR Detection are defined in § 3.1. ✓ = first-class capability tested; ○ = partially or implicitly covered; blank = not addressed.

resisting reliance on parametric memory that may be outdated, incomplete, or hallucinated.

We organize the evaluation along three dimensions: task modality (the source modality of the file: document, video, or audio), task scene (the real-world domain, e.g. government budget, sports broadcast, see full scenes on § A), and capability family that reflects the cognitive complexity required, organized in three levels, see Table 2.

3.2 Benchmark Construction

Source Collection & Filtering. RDQA draws on raw files from three complementary source categories: government and regulatory portals (budget reports, court filings), technical and professional documentation (datasheets, vehicle manuals, syllabi), and publicly available video and audio (YouTube, sports archives, podcasts).

Manual Annotation. To capture real-world distributions of information retrieval from raw media files, we adopt fully manual question authoring, with annotators developing one QA pair per source file. For each item, the annotator records the file’s SHA-256 hash and download URL, the golden evidence with precise localization metadata (page index and snippet for PDFs; start/end timestamps for video and audio), and a few answer variants cov-

ering format differences. The two core authoring steps are:

- (1) identify an answer-worthy target aligned with Table 2 whose evidence is verifiable and unlikely to be memorized, i.e. an L1 single data point, an L2 multi-point operation target, or an L3 multi-step reasoning target.
- (2) craft a natural-language query that embeds the source context such that answering *requires* accessing the raw file.

A second annotator then independently answers each question using only the specified source file, confirming answer correctness and evidence localizability.

Verification. To uphold data quality, a reviewer inspects all crafted items. For each inspected question, the reviewer verifies that: (a) the golden evidence is correctly localized within the source file; and (b) no publicly accessible web page contains the exact answer in plain text, which would undermine the routing challenge. Questions failing any criterion are revised or discarded.

3.3 Evaluation Criteria

Considering the characteristics of answer types across RDQA’s questionset, we adopt a two-stage scoring protocol that balances determinism with

Level	Complexity	Document	Video	Audio
L1	$\Theta(1)$	DTE — Document Text Extraction. <i>E.g.</i> : What is the Infrastructure budget in the Hamilton FY2025 Budget Report?	VTE — Video Text Extraction. <i>E.g.</i> : At 3:53 in the YouTube video “Epitaph 2”, what score is shown on the HUD?	AIE — Audio Information Extraction. <i>E.g.</i> : In Snowflake’s latest earnings call, what YoY revenue growth did the CFO announce in the opening remarks?
<i>Definition: Extract a single data point from the artifact without operating on its value.</i>				
L2	$\Theta(k)$	DMO — Document Multi-Point Operation. <i>E.g.</i> : In the Hamilton FY2025 Budget Report, which department has the highest budget?	VMO — Video Multi-Point Operation. <i>E.g.</i> : In this unboxing video, does the model number announced by the host match the one on the product packaging?	AMO — Audio Multi-Point Operation. <i>E.g.</i> : Is the CEO’s full-year margin guidance in the opening remarks consistent with the figure given during the analyst Q&A?
<i>Definition: Extract multiple data points and/or apply one operation (max, min, count, filter, enumeration, etc.).</i>				
L3	$\Theta(n)+$	DRV — Document Reasoning & Verification. <i>E.g.</i> : Among subsidiaries with carbon intensity below the industry average and ESG rating A+, which has the highest annual revenue?	VRV — Video Reasoning & Verification. <i>E.g.</i> : In this fashion show recording, which designer brand exhibited the largest number of looks featuring the same color palette as the opening look’s dominant color?	ARV — Audio Reasoning & Verification. <i>E.g.</i> : In this press conference recording, how many questions did journalists ask in total?
<i>Definition: Multi-step reasoning over multiple data points through multiple operations, answer must not be verbatim.</i>				

Table 2: The 9 capability families of RDQA arranged as a 3×3 matrix of cognitive complexity (rows) and source modality (columns). Capability acronyms follow the pattern $\{D,V,A\} \times \{TE/IE, MO, RV\}$, where the first letter denotes modality and the suffix denotes operation type. $\Theta(\cdot)$ is used illustratively to describe the cognitive complexity and typical number of evidence-extraction and operation steps required.

robustness to surface-form variations.

Normalized Exact Match. For questions whose ground truth is a well-defined numerical value, date, identifier, or short factual string. We perform normalized exact match against a curated list of answer variants. The variant list is constructed during annotation (Section 3.2) and typically contains several surface forms per question.

LLM-as-a-Judge. Exact match alone is insufficient for answers that vary in phrasing, granularity, or language while remaining semantically equivalent. Answers that fail exact match are forwarded to an LLM judge (Claude Sonnet 4.5 (Anthropic, 2025)) that receives the question, reference answer, full variant list, and the agent’s response, and assigns a binary score of 0 or 1 based on semantic equivalence.

Combined Scoring. The final correctness label c_i for question q_i is

$$c_i = \max \left(\text{EXACTMATCH}(a_i, V_i), \text{LLMJUDGE}(q_i, g_i, a_i) \right), \quad (1)$$

where a_i is the agent’s answer, V_i is the variant set, and g_i is the reference ground truth. An answer is correct if it passes *either* stage.

Judge Reliability. The LLM judge may itself introduce errors, we conduct a human audit on all judge decisions, recording agreement rate, false positive rate, and false negative rate. In cases of disagreement, the human label takes precedence in reported scores.

3.4 Anti-hallucination Design

Priori Prevention. Annotators systematically prefer obscure entities, target first-hand, fine-grained values that demand precise localization, embed multiple similar-looking candidate values within the same source file as distractors, and reference structural coordinates such as specific pages, table cells, or timestamps that cannot be approximated from model training memory.

Detection Metric: Parametric Leakage Rate (PLR). We adopt a *process-aware* detection strategy based on behavioral trace analysis. We require all evaluated agent systems to produce a structured trace log recording every search query, every URL

accessed, every file downloaded, and every tool invocation with its input and output. For each question q_i , let $c_i \in \{0, 1\}$ indicate answer correctness and $r_i \in \{0, 1\}$ indicate whether the trace confirms that the agent downloaded a file matching by SHA-256 OR golden origin URL *and* invoked a parsing tool on it. The Parametric Leakage Rate is defined as:

$$\text{PLR} = \frac{\sum_i c_i \cdot (1 - r_i)}{\sum_i c_i} \quad (2)$$

PLR measures the fraction of correct answers produced *without* genuine evidence retrieval. A well-functioning agent that faithfully follows the retrieve–parse–extract pipeline should exhibit a PLR close to 0%; We additionally report PLR-adjusted accuracy as a complementary metric:

$$\text{Acc}_{\text{adj}} = \text{Acc}_{\text{raw}} \times (1 - \text{PLR}) \quad (3)$$

which credits only those correct answers that are genuinely grounded in retrieved evidence.

Evaluation Philosophy Because RDQA answers are designed to require source access, bare-model accuracy measures leakage rather than capability and is omitted. We instead report agent-level performance structured as a $\{\text{model}\} \times \{\text{scaffolding}\}$ matrix: parametric suppression is solely model-driven, whereas source routing, tool invocation and detail localization exercises both. Since the model makes the decision on calling available static tools including searching, and synthesizing reasoning and tool-using during detail localization.

3.5 Statistics

Through source selection, manual annotation, and verification, we collect a total of 1,350 source files along with the same number of high-quality QA pairs. As illustrated in Figure 3, it encompasses a wide range of real-world information-seeking scenarios.

4 Experiments

We evaluate 12 agent systems (6 backbones $\times$ 2 scaffolds) in a sandboxed environment. As a human reference, four college seniors or graduate students answer all questions under the same open-web setting, with no source files pre-provided. Participants are required to report the final answer, the source URL, and the supporting evidence used to derive the answer. Standard aids such as calculators and notes are allowed. No fixed time limit or session-length constraint is imposed, and participants may

take breaks at will. Additionally, we include an agentic baseline extended from one evaluated architecture and evaluate it under the same setting.

Sandbox architecture. Each agent runs in a clean, isolated Docker container provisioned only with its scaffold’s standard toolset (web search, browsing and download, media parsing, and general file-system utilities). No benchmark assets are present in, or reachable from, the container: golden source files, annotation metadata, answer keys, and download caches reside exclusively on a separate scoring host, which consumes the agents’ structured trace logs *post hoc* to compute correctness and PLR.

4.1 Main Results

Table 3 summarizes agent and human performance on RDQA.

Large human–agent gap. Human participants achieve an overall accuracy of 57.5%, with performance decreasing across complexity levels (76.2% at L1, 63.9% at L2, 42.8% at L3), substantially higher than the best agent result. The strongest agent configuration, Hermes Agent + Gemini 3 Flash, reaches 24.0% PLR-adjusted accuracy. Notably, human accuracy is far from saturated: because participants traverse the same multi-stage pipeline as agents, they remain susceptible to selecting an incorrect source, overlooking fine-grained details, and lapses in attention over long retrieval sequences. This moderate ceiling indicates that RDQA captures intrinsic task difficulty that challenges for human performers.

Best systems are still far from satisfactory. Hermes Agent + Gemini 3 Flash obtains the best adjusted score (24.0%) with a comparatively low PLR of 29.0%. OpenClaw + Kimi K2.5 is the strongest open-source configuration by adjusted accuracy, largely because it has the lowest PLR in the table (20.1%), despite not leading every raw capability column. However, no configuration achieves accuracy sufficient for users to trust an agent with real-world source-grounded retrieval tasks. Audio capabilities remain the weakest across all configurations, with even the best scores staying below 28%.

Scaffolding has mixed effects. The effect of scaffold choice is backbone-dependent. For Qwen3.6 Plus, Hermes improves document reasoning over OpenClaw, but its PLR remains high

Table 3: **Evaluation results on RDQA.** The highest accuracy of each column is shown in **bold**, the second highest is underlined, and the third highest is in red, metric(s) with ↓ is marked the lower the better. "—" indicates non-disclosed information or unsupported modality of model(s).

Agent System	Size	Document			Video			Audio			PLR↓	Acc _{adj}
		DTE	DMO	DRV	VTE	VMO	VRV	AIE	AMO	ARV		
MLLM agents with closed-source backbones												
OpenClaw + Claude Opus 4.6	—	27.2%	34.7%	29.9%	12.1%	9.8%	23.4%	—	—	—	51.0%	12.2%
OpenClaw + GPT-5.2	—	48.8%	41.0%	34.9%	32.1%	29.4%	32.7%	22.6%	27.8%	19.8%	64.0%	12.3%
OpenClaw + Gemini-3 Flash	—	51.4%	26.8%	36.4%	47.4%	38.2%	26.6%	19.8%	21.6%	24.7%	53.4%	14.9%
Hermes Agent + Claude Opus 4.6	—	30.4%	27.6%	31.1%	10.7%	14.5%	27.4%	—	—	—	89.0%	2.8%
Hermes Agent + GPT-5.2	—	42.0%	44.6%	28.2%	35.6%	31.8%	28.0%	18.7%	21.3%	26.0%	57.2%	13.8%
Hermes Agent + Gemini 3 Flash	—	50.0%	22.7%	41.2%	53.3%	40.2%	30.0%	24.8%	18.4%	32.2%	29.0%	24.0%
MLLM agents with open-source backbones												
OpenClaw + Qwen3.6 Plus	27B (Dense)	11.2%	8.6%	13.6%	15.2%	13.7%	16.2%	10.2%	8.8%	12.4%	96.0%	0.5%
OpenClaw + Kimi K2.5	1T (A32B)	37.0%	29.6%	33.0%	29.8%	28.6%	22.1%	17.2%	22.8%	16.4%	20.1%	21.4%
OpenClaw + GLM4.7v	355B (A32B)	22.6%	13.3%	8.6%	14.8%	20.0%	11.5%	4.2%	6.6%	10.4%	62.0%	5.2%
Hermes Agent + Qwen3.6 Plus	27B (Dense)	21.6%	16.5%	30.1%	18.2%	12.0%	22.7%	12.6%	8.0%	5.0%	65.1%	7.4%
Hermes Agent + Kimi K2.5	1T (A32B)	8.8%	11.1%	10.7%	11.9%	18.2%	21.4%	14.8%	12.4%	13.6%	83.3%	2.3%
Hermes Agent + GLM4.7v	355B (A32B)	6.4%	5.8%	10.4%	10.7%	11.8%	14.2%	3.9%	7.2%	11.6%	79.0%	2.2%
Baselines												
Average Human Work	—	74.7%	51.4%	56.2%	88.4%	79.7%	38.4%	66.1%	58.3%	33.7%	0.0%	57.5%
RDQA-Skill + GPT 5.2	—	42.7%	49.9%	41.3%	56.0%	44.6%	34.2%	60.0%	42.5%	32.5%	2.4%	41.1%

(65.1%). For Kimi K2.5 and GLM4.7v, OpenClaw yields substantially higher PLR-adjusted accuracy than Hermes (21.4% vs. 2.3% for Kimi; 5.2% vs. 2.2% for GLM). These results suggest that better scaffolding is not simply more tool access; it must also constrain shortcuts and preserve source-grounded evidence use. The reasoning and decision-making by backbones seem vital in these cases.

PLR is essential for interpreting performance.

Several systems achieve non-trivial raw scores while relying heavily on shortcuts. OpenClaw + Qwen3.6 Plus reaches double-digit accuracy in most document and video categories, but its PLR is 96.0%, reducing adjusted accuracy to 0.5%. Hermes Agent + Claude Opus 4.6 similarly obtains competitive raw document and video scores but drops to 2.8% adjusted accuracy because its PLR reaches 89.0%. Thus, RDQA distinguishes answer correctness from faithful source use: models that appear capable under raw accuracy can be poor at the intended retrieve–parse–localize pipeline.

4.2 How Do Agents Cheat? Decomposing PLR by Leakage Mode

To understand *what* the high PLR values in Table 3 represent, we conduct a per-trajectory inspection on one open-source and the other closed-source backbones: Claude Opus 4.6 and GLM-4.7v. For each correct answer we classify the trajectory into one of three behavioral modes:

Table 4: Decomposition of leakage patterns for the selected backbones. Each cell shows percentage of correct answers in that modality belonging to the discovered pattern.

Model	Mod.	TRUSTED	FILE-LEAK	WEB-LEAK
Opus 4.6	Doc.	5.6%	93.1%	1.2%
	Video	18.2%	8.3%	73.6%
GLM-4.7v	Doc.	5.1%	82.1%	12.8%
	Video	30.6%	0.0%	69.4%
	Audio	15.4%	23.1%	61.5%

- **TRUSTED**: the agent visited the correct source URL *and* did not access local benchmark files;
- **FILE-LEAK**: the trace shows `read_file` or `directory-listing` invocations probing the local file system. Such behaviors is consistent with the agent suspecting it is under evaluation and searching for benchmark files. Because the sandbox contains no benchmark assets (§4), such probing finds nothing. The agent then produces its answer from parametric memory rather than from the retrieved source.
- **WEB-LEAK**: the agent used web/search tools but never fetched the source URL, answers came from second-hand public-indexed information including search snippets, aggregator pages, or YouTube metadata.

Table 4 reveals that the two backbones exhibit *the same two leakage pathways*, differing only in intensity.

Architectural implications. The leakage decomposition implies to two complementary mitigations. File-leak requires stricter agent discipline and/or tool exposure: the scaffold should prevent benchmark-directed goal hacking and block irrelevant file-system access such as `read_file`. Web-leak instead calls for better source-grounded tool development on media files and more disciplined instruction following of backbones. So that agents must inspect frames, timestamps, and audio segments rather than rely on indexed snippets. Although recent MLLMs can reason over media when the source is directly provided (Shi et al., 2025), RDQA exposes the remaining agentic gap: systems are unable to autonomously retrieve, localize, and record relevant evidence. The convergence of the PLR on different backbones leads to similar PLR-adjusted accuracy (2.8% and 2.2%) despite different raw scores shows that much of the observed gain reflects scaffold behavior rather than grounded multimodal competence. It suggests better scaffolding design and development are key to unleash full capability of backbones.

4.3 A Source-Grounded Agentic Baseline

The analysis in § 4.2 indicates that high raw accuracy is frequently obtained through shortcut paths rather than faithful retrieval, and that video and audio are the modalities where agents most often substitute indexed snippets for genuine source inspection. To establish a reproducible anchor that targets these failure modes, we extend Hermes Agent into a source-grounded agentic baseline, denoted **RDQA-Skill**.

Method. RDQA-Skill treats each question as an evidence-localization loop (Figure 4). Given a query, the agent first performs *scope narrowing*: it reasons explicitly about where the answer is most likely to reside in the source, producing a hypothesis about the evidence location. It then issues a *targeted tool invocation* against that location to extract one or more answer candidates grounded in the retrieved span. A *reflection* step then assesses whether the retrieved evidence actually supports a candidate with sufficient confidence on the given query. If confidence is low (e.g. the candidate conflicts with the query constraints), the agent revises its location hypothesis and repeats the loop, examining a different part of the source; if confidence is sufficient, it emits the answer together with the grounding evidence. Crucially, the loop forces the

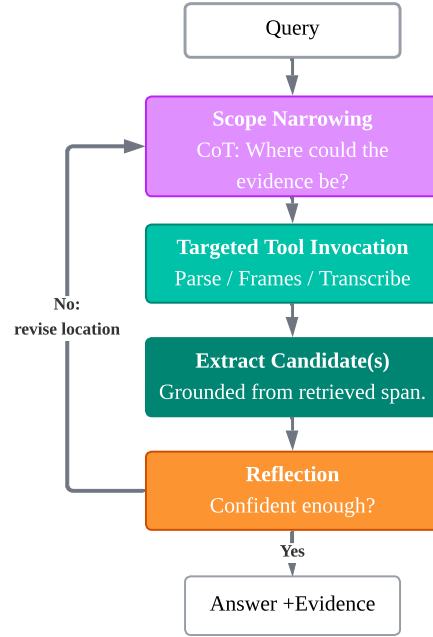


Figure 4: Pipeline of RDQA-Skill.

agent to inspect the raw media before committing, which is the behavior that the PLR metric rewards.

Results. As shown in Table 3, RDQA-Skill+GPT-5.2 reaches 41.1% PLR-adjusted accuracy, while driving PLR down to 2.4%, which the lowest among all agent systems and approaching human baseline. The gains are most pronounced on audio, exactly the modality the design targets: AIE rises to 60.0% and AMO to 42.5%, whereas no other agent configuration exceeds 28% on either column. This shows that the audio gap reported in § 4.1 is not an intrinsic ceiling of the backbone but is largely attributable to scaffolding that fails to enforce source transcription.

RDQA-Skill is a *reference anchor*, not yet a solved solution. Its overall adjusted accuracy still trails the human baseline by more than 16 points, and the hardest reasoning families remain difficult even under disciplined retrieval. The low PLR confirms that these residual errors stem from genuine localization and multi-step reasoning failures rather than from leakage. RDQA-Skill therefore demonstrates that process-aware scaffolding meaningfully narrows the source-grounded gap.

5 Conclusion

In this paper, we introduce RDQA, a benchmark for evaluating whether multimodal agents can faith-

fully answer questions from raw source files on the open web. RDQA targets a practical but underexamined capability: locating the correct raw source, invoking appropriate tools, localizing fine-grained evidence, and grounding answers in the original artifact. Our experiments show that current agents remain far from reliable, with low performance, strong modality variation, and high dependence on agent scaffolding. We further show that raw accuracy can overstate true source-grounded ability, as many correct answers arise from shortcuts such as web summaries, search snippets, or parametric leakage. These findings highlight the need to enforce faithful source grounding rather than relying only on final-answer correctness.

Limitations

First, due to the cost of running full agentic evaluations, each configuration is evaluated only once without variance estimates, so small differences between systems should not be read as significant rankings and “best-performing” refers to this single-run setting. Second, PLR is a conservative estimate: since we credit retrieval only on a SHA-256 or golden-URL match, an agent that faithfully retrieves the same fact from a legitimate alternative source is still counted as leakage; PLR thus over-estimates shortcuts, and PLR-adjusted accuracy should be read as a lower bound on grounded capability. Finally, because RDQA relies on live open-web retrieval, some source URLs may become invalid over time; such inaccessible items are excluded from scoring, and reported scores are computed over the accessible subset.

Ethical Considerations

All file sources included in RDQA are collected from publicly accessible materials, CC-BY licensed resources, or similarly permissive open-use sources, and the benchmark is intended solely for research purposes. To the best of our knowledge, the proposed dataset does not introduce any foreseeable ethical risks.

References

- Anthropic. Claude sonnet 4.5. <https://www.anthropic.com/claude>, 2025. Large language model.
- Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *CoRR*, abs/2411.04952, 2024. doi: 10.48550/ARXIV.2411.04952. URL <https://doi.org/10.48550/arXiv.2411.04952>.

10.48550/ARXIV.2411.04952. URL <https://doi.org/10.48550/arXiv.2411.04952>.

- Kuicai Dong, Yujing Chang, Derrick-Goh-Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. Mmdocir: Benchmarking multimodal retrieval for long documents. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 30971–31005. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1576. URL <https://doi.org/10.18653/v1/2025.emnlp-main.1576>.

- Robert Friel, Masha Belyi, and Atindriyo Sanyal. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *CoRR*, abs/2407.11005, 2024. doi: 10.48550/ARXIV.2407.11005. URL <https://doi.org/10.48550/arXiv.2407.11005>.

- Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL <https://openaccess.thecvf.com/content/CVPR2025/html/Fu-Video-MME-The-First-Ever-Comprehensive-Evaluation-Benchmark-of-Multi-modal-LLMs-in-CVPR-2025-paper.html>.

- Lei Kang, Rubèn Tito, Ernest Valveny, and Dimosthenis Karatzas. Multi-page document visual question answering using self-attention scoring mechanism. In *International Conference on Document Analysis and Recognition*, pages 219–232. Springer, 2024.

- Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multi-modal agents on realistic visual web tasks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 881–905. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.50. URL <https://doi.org/10.18653/v1/2024.acl-long.50>.

- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer*

- Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02095. URL <https://doi.org/10.1109/CVPR52733.2024.02095>.
- Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip S. Yu. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=VvDEuyVXkG>.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=zADUB0aCTQ>.
- Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yugang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ae0e43289bffe40c1fa34633fc608e92-Abstract-Datasets-and-Benchmarks_Track.html.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, volume ACL 2022 of *Findings of ACL*, pages 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.177. URL <https://doi.org/10.18653/v1/2022.findings-acl.177>.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE, 2021. doi: 10.1109/WACV48630.2021.00225. URL <http://doi.org/10.1109/WACV48630.2021.00225>.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE, 2022. doi: 10.1109/WACV51458.2022.00264. URL <https://doi.org/10.1109/WACV51458.2022.00264>.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=fibxvavhs3>.
- S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=TeVAZXr3yv>.
- Yang Shi, Huanqian Wang, Wulin Xie, Huanyao Zhang, Lijie Zhao, Yifan Zhang, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, Zhuoran Zhang, Xinlong Chen, Bohan Zeng, Sihan Yang, Yushuo Guan, Zhang Zhang, Liang Wang, Haoxuan Li, Zhouchen Lin, Yuanxing Zhang, Pengfei Wan, Haotian Wang, and Wenjing Yang. Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/779efd29bc0236b07df1da5a548d0bba-Abstract-Datasets-and-Benchmarks_Track.html.
- Kaiwen Wei, Xiao Liu, Jie Zhang, Zijian Wang, Ruida Liu, Yuming Yang, Xin Xiao, Xiao Sun, Haoyang Zeng, Changzai Pan, Yidan Zhang, Jiang Zhong, Peijin Wang, and Yingchao Feng. Cfvbench: A comprehensive video benchmark for fine-grained multimodal retrieval-augmented generation. In Hakim Hacid, Yoelle Maarek, Francesco Bonchi, Ido Guy, and Emine Yilmaz, editors, *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pages 2501–2512. ACM, 2026. doi: 10.1145/3774904.3792726. URL <https://doi.org/10.1145/3774904.3792726>.

A Appendix

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Airbench: Benchmarking large audio-language models via generative comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 1979–1998. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.ACL-LONG.109. URL <https://doi.org/10.18653/v1/2024.acl-long.109>.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen tau Yih, and Xin Luna Dong. Crag – comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*, 2024b. URL <https://arxiv.org/abs/2406.04744>.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8938–8968. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.505. URL <https://doi.org/10.18653/v1/2024.emnlp-main.505>.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=oKn9c6ytLx>.

A.1 Task Scenes

Table 5 lists the full taxonomy of task scenes in RDQA. Each source file is assigned to exactly one scene during annotation.

Table 5: Full taxonomy of task scenes in RDQA.

Scene	Description
government_budget	Government budget reports and appropriations documents.
tender_procurement	Public procurement tenders, RFPs, and bidding documents.
financial_report	Corporate annual reports, 10-K/10-Q filings, and earnings releases.
esg_report	Environmental, social, and governance disclosure reports.
game_hud.ocr	Video game HUD overlays showing scores, stats, or status indicators.
sports_scoreboard	Sports broadcasts with on-screen scoreboards and statistical overlays.
credits.ocr	End credits and rolling text from films or video productions.
news_ticker	News broadcast tickers and lower-third text overlays.
fashion_product_id	Fashion clips where products are identified from on-screen captions or tags.
unboxing_specs	Product unboxing and review videos with on-screen specs and prices.
architectural_blueprint	Architectural and engineering blueprints with dimensioned plans.
vehicle_manual	Vehicle owner’s and service manuals with technical parameters.
howto_video	Instructional tutorials with step-by-step procedures and annotations.
technical_datasheet	Component datasheets and product spec sheets from manufacturers.
academic_syllabus	University course syllabi with schedules and grading policies.
event_agenda	Conference programs and workshop agendas.
life_fact_check	Everyday fact-checking scenarios from podcasts or recorded statements.
fashion_videos	Fashion show recordings and editorial fashion videos.
video_embedded_docs	Videos in which the answer-bearing artifact is texts filmed on screen.
dynamic_scrolling_text	Continuously scrolling text such as stock tickers or rolling lists.

A.2 Capability Subcategory Taxonomy

Tables 6, 7, and 8 present the full subcategory taxonomy used for each level on each modality. Each of the 9 capability families is decomposed into several subcategories that isolate distinct cognitive or perceptual demands. Subcategories under L1 families (DTE/VTE/AIE) describe *what kind of single value is read out*; subcategories under L2 families (DMO/VMO/AMO) describe *which single aggregation operation is applied to the multiple data points*; subcategories under L3 families (DRV/VRV/ARV) describe *which kind of multi-step reasoning produces the answer*.

Table 6: Document capability subcategories across L1 (DTE), L2 (DMO), and L3 (DRV). DTE is organized along a *Surface* axis that describes the answer’s intrinsic form; L2 isolates which single aggregation operation is applied to multiple data points; L3 isolates which kind of multi-step reasoning derives the answer, with all L3 answers required to be closed-form (a number, named entity, date, identifier, verdict, or selection from an enumerated set).

Family	Sub	Description
<i>DTE — Document Text Extraction (L1)</i>		
	VER	Verbatim Phrase. The answer is a contiguous phrase quoted verbatim, e.g. a warning statement, a defined term, or a disclaimer.
	ID	Identifier. The answer is a formatted identifier such as a case number, contract ID, file reference, or standard designation.
	META	Document Metadata. The answer is a document-level metadata field: author, publication date, version number, or issuing authority.
	NUM	Numeric Attribute. The answer is a single quantitative property of a named entity: capacity, price, percentage, area, weight, or rating.
	CAT	Categorical Attribute. The answer is a single categorical property: model name, product class, certification standard, or grade level.
	TEMP	Temporal Attribute. The answer is a single date or time-bound property: effective date, expiration, filing deadline, or approval timestamp.
	REL	Relational Attribute. The answer identifies a single relational party: responsible person, parent department, approving authority, or contracting party.

Continued on next page

Table 6 (continued)

Family	Sub	Description
<i>DMO — Document Multi-Point Operation (L2)</i>		
	EXT	Extremum Selection. Scan multiple values and return the maximum or minimum (e.g. the highest-budget department).
	COUNT	Counting. Scan multiple candidates and report how many satisfy a stated property (e.g. how many subsidiaries hold rating A).
	FILT	Single-Criterion Filtering. Scan candidates against <i>one</i> stated criterion and return those that satisfy it.
	FILT-MULTI	Compound-Criterion Filtering. Scan candidates in a single filter pass against a conjunctive or disjunctive predicate over <i>two or more</i> stated criteria; the answer is a verbatim entry from the document.
	ENUM	Enumeration. Enumerate all entries of a stated category (e.g. list all clauses in Article III).
	RANK	Top-K / Bottom-K. Sort multiple values by a stated key and return the top or bottom k entries.
	VERIF	Single-Source Verification. Yes/no judgment on whether a proposition is supported by a single passage, table, or stated definition.
<i>DRV — Document Reasoning and Verification (L3)</i>		
	COMPUTE	Derived Value Calculation. Extract ≥ 2 values from non-adjacent locations and apply arithmetic, ratio, weighted aggregation, or temporal delta to derive a value that does <i>not</i> appear verbatim in the document.
	CHAIN	Pipelined Multi-Criterion Selection. Apply ≥ 2 <i>sequential</i> operations over a candidate set (e.g. filter $\rightarrow$ rank, or filter $\rightarrow$ filter $\rightarrow$ extremum) to identify a single entity whose selection requires chained reasoning rather than a single compound predicate.
	APPLY	Cross-Section Rule Application. Extract a rule or specification stated in one section, extract a separate concrete fact stated in another non-adjacent section, and apply the rule to the fact to produce a verdict (e.g. compliant / violates / passes / fails). The rule and the fact must originate from different parts of the document.
	INFER	Closed-Form Inference. Combine ≥ 2 implicit textual cues—relative emphasis, ordering, omissions, framing across non-adjacent sections—to derive a conclusion that is never explicitly stated. The final answer must be closed-form: a named entity, categorical label, or short verdict. Open-ended summaries are out of scope.

Table 7: Video capability subcategories. L1 isolates what single value is read from one frame; L2 isolates which single aggregation or verification operation is applied across multiple frames; L3 isolates which kind of multi-step reasoning derives the answer. All L3 answers are closed-form.

Family	Sub	Description
<i>VTE — Video Text Extraction (L1)</i>		
	HUD	Overlay / HUD Text. Reading text from a semi-transparent heads-up display, score overlay, or on-screen graphic with a complex background, in a single frame.
	EMBED	Embedded Document Text. Reading text on a physical object visible in the frame: a label, panel, paper document, or sign.
	TITLE	Title Card / Caption. Reading post-produced text overlays in a single frame: opening titles, scene captions, lower-thirds, or end-screen text.
	SCROLL	Single-Entry Scrolling Lookup. Reading one specific entry from continuously scrolling content (e.g. a single name in end credits or a single ticker entry).
	STEP	Single-Step Lookup. Reading the content shown at the n -th step of a procedural sequence—one frame, one value.
<i>VMO — Video Multi-Point Operation (L2)</i>		
	ENUM	Cross-Frame Enumeration. Enumerate all on-screen items of a stated category across the video (e.g. all designer brand logos in a fashion show).
	COUNT	Cross-Frame Counting. Count occurrences of a stated event or item across multiple frames (e.g. number of fouls signaled).
	EXT	Cross-Frame Extremum. Select the maximum or minimum among values observed at multiple timestamps (e.g. highest score displayed across the game).
	CHANGE	Change-Point Detection. Identify the timestamp at which a displayed value changes (e.g. when the lead first changes hands). Computing the numeric magnitude of the change belongs to VRV-COMPUTE.

Continued on next page

Table 7 (continued)

Family	Sub	Description
	CONSIST	Cross-Temporal Consistency. Verify in a single comparison whether two on-screen readings observed at different timestamps agree or contradict (e.g. does the half-time score shown at 30:00 match the running tally shown earlier?).
	CROSSCHECK	Audio-Visual Cross-Check. Verify in a single comparison whether spoken narration matches the text, number, or label shown on screen at the same moment (e.g. does the host’s announced model number match the box?).
<i>VRV — Video Reasoning and Verification (L3)</i>		
	COMPUTE	Cross-Frame Derived Value. Extract ≥ 2 on-screen quantitative values at different timestamps and compute a derived number (difference, ratio, trend slope, percentage change) that is not displayed in any single frame (e.g. “by how many points did the lead change between 5:00 and 20:00?”).
	CHAIN	Pipelined Frame-Level Selection. Apply ≥ 2 sequential operations over frame-level evidence (e.g. enumerate items $\rightarrow$ filter by visual property $\rightarrow$ select extremum) to identify a single frame, segment, or entity whose selection requires chained reasoning.
	APPLY	Procedural / Specification Application. Extract a stated procedure or specification from one location in the video (printed instructions, recipe card, manual page, opening title) and a separately demonstrated action elsewhere in the video, then apply the specification to the action to produce a verdict (compliant / matches / violates).
	INFER	Closed-Form Visual Inference. Combine ≥ 2 implicit visual or audio-visual cues—facial reaction, scene ordering, off-screen sound, framing, layout choice—to derive a conclusion that is never explicitly stated. The final answer must be closed-form: a named entity, categorical label from a stated set, or short verdict. Open-ended descriptive answers are out of scope.

Table 8: Audio capability subcategories. L1 isolates what single value is extracted from one utterance; L2 isolates which single aggregation or verification operation is applied across multiple segments; L3 isolates which kind of multi-step reasoning derives the answer. All L3 answers are closed-form.

Family	Sub	Description
<i>AIE — Audio Information Extraction (L1)</i>		
	SPEECH	Spoken Fact Extraction. Extracting a specific factual value—number, name, date, or measurement—from a single utterance or short speech span.
	NONSPEECH	Non-Speech Audio Cue. Identifying an event type or state from a single non-speech audio signal: alarm, tone, environmental sound, or music cue.
	SEG-LOC	Single Boundary Localization. Identifying when a single named topic shift, speaker change, or section transition occurs in the audio timeline.
<i>AMO — Audio Multi-Point Operation (L2)</i>		
	ENUM	Cross-Segment Enumeration. Enumerate all entries of a stated category across audio segments (e.g. all topics raised by panelists).
	COUNT	Cross-Segment Counting. Count occurrences of a stated event across the recording (e.g. number of audience questions).
	EXT	Cross-Speaker Extremum. Select the speaker or segment with the maximum or minimum of a stated property (e.g. longest-speaking participant).
	LOC	Cross-Segment Localization. Locate where a stated entity is first or last mentioned across the recording.
	CONSIST	Cross-Segment Consistency. Verify in a single comparison whether two factual statements made at different points in the recording agree or contradict (e.g. does the headcount figure cited in the prepared remarks match the figure given in Q&A?).
<i>ARV — Audio Reasoning and Verification (L3)</i>		
	COMPUTE	Cross-Segment Derived Value. Extract ≥ 2 numeric or temporal values stated in different audio segments and compute a derived value (delta, ratio, basis-point change, weighted aggregation) that is not announced verbatim in any single segment (e.g. “by how many basis points did the CFO’s full-year guidance change between opening remarks and the analyst Q&A?”).
	CHAIN	Pipelined Segment-Level Selection. Apply ≥ 2 sequential operations over speaker- or segment-level evidence (e.g. enumerate questions $\rightarrow$ filter by topic $\rightarrow$ select longest answer) to identify a single segment, speaker, or entity whose selection requires chained reasoning.

Continued on next page

Tester	DTE	DMO	DRV	VTE	VMO	VRV	AIE	AMO	ARV	Overall
T1	69.3%	57.6%	50.7%	91.2%	74.3%	41.8%	59.4%	64.1%	29.6%	56.8%
T2	82.1%	44.9%	61.5%	84.6%	85.9%	32.7%	72.8%	52.6%	41.3%	58.0%
T3	76.4%	53.8%	58.9%	93.7%	77.6%	45.5%	63.5%	60.8%	31.8%	60.5%
T4	71.0%	49.3%	53.7%	84.1%	81.0%	33.6%	68.7%	55.7%	32.1%	54.7%
Avg.	74.7%	51.4%	56.2%	88.4%	79.7%	38.4%	66.1%	58.3%	33.7%	57.5%

Table 9: Per-tester human accuracy across the nine capability families. All four testers independently answered the same complete question set; the bottom row matches the human baseline reported in Table 3.

Table 8 (continued)

Family	Sub	Description
	APPLY	Cross-Segment Policy / Guidance Application. Extract a stated policy, guidance, or rule from one segment and a separate concrete claim or action stated in another non-adjacent segment, then apply the policy to the claim to produce a verdict (compliant / violates / contradicts / triggers). The policy and the claim must originate from different segments.
	INFER	Closed-Form Acoustic-Dialogic Inference. Combine ≥ 2 implicit acoustic, prosodic, or dialogic cues—tone shifts, hedging, topic avoidance, interruption patterns, audience reaction—to derive a conclusion that is never explicitly stated. The final answer must be closed-form: a named speaker, categorical label from a stated set, or short verdict (e.g. “which analyst’s question received the most evasive response” answered with a single analyst name). Open-ended summaries are out of scope.

A.3 Human Baseline: Protocol and Per-Tester Results

Protocol. Four testers (college seniors or graduate students) each independently answered the complete question set under the same open-web conditions as the evaluated agents: no source files were pre-provided, and for every item the tester was required to report the final answer, the source URL, and the supporting evidence used to derive the answer. Standard aids (calculators, notes) were permitted; no time limit or session-length constraint was imposed, and testers could take breaks at will. Responses were scored under the same two-stage protocol as agent runs (§3.3).

Per-tester results. Table 9 reports each tester’s accuracy on all nine capability families. The four testers show non-uniform strengths (e.g., T2 is strongest on DTE but weakest on DMO), and per-family averages correspond to the “Average Human Work” row of Table 3.

Interpretation. The 57.5% aggregate is intended as a reference point for diligent human performance under agent-identical open-web conditions, not as a saturation ceiling or a precisely reproducible constant: because testers traverse the same multi-stage retrieve–parse–localize pipeline as agents, they remain susceptible to source-selection errors, overlooked fine-grained details, and attention lapses over long retrieval sequences. The paper’s comparative claims—the best generic agent at 24.0% adjusted versus humans in the 57.5%, and L3 reasoning remaining difficult even for humans (42.8%)—are robust to several points of variation in this anchor.

A.4 Closed-Book Control: Testing Parametric Answerability

Section 3.4 claims that RDQA questions are constructed to be unanswerable from parametric knowledge alone. To test this empirically rather than by construction alone, we run a closed-book control: each of the six evaluated backbones answers every question with *all* tools disabled—no web search, no browsing, no file access—under the same answer-extraction and two-stage scoring protocol as the main runs (§3.3).

Table 10 reports the results. Closed-book accuracy is near zero for every backbone (at most 0.2%), and the few correct answers concentrate on L2–L3 items whose answer formats are intuitively guessable (e.g., small counts or binary verdicts) rather than on items whose evidence was memorized. This empirically supports the design claim in §3.4: RDQA items are generally unanswerable from parametric knowledge alone, and correct answers in the main evaluation must therefore derive from retrieval behavior—faithful or otherwise—which is precisely what PLR decomposes.

A.5 Evaluation Prompt

The prompt settings for evaluation, llm-as-a-judge are shown in table 11, 12.

Table 11: Prompt used for evaluating agents on RDQA questions.

Evaluation Prompt
You are an RDQA web-agent benchmark participant.
You are given only a natural-language query and constraints.
You are NOT given the source file, ground-truth answer, golden evidence, page index, timestamp, or benchmark metadata.
Your task:

1. Use web search to locate the relevant public raw source, such as a PDF, webpage, legal opinion, syllabus, manual, report, video page, or audio page.
2. Read or infer from the located source.
3. Return a structured JSON prediction only.

Rules:

- answer_value must be the final answer only, concise and exact.
- For PDF/document tasks:
 - page_index should be the zero-based page index if you can determine it.
 - content_snippet should be a short exact or near-exact supporting snippet from the located source.
 - timestamp_start and timestamp_end should be null.
- For video/audio tasks:
 - timestamp_start and timestamp_end should be evidence times in seconds if determinable.
 - page_index should be null.
 - content_snippet can be a transcript/OCR/audio snippet if available.
- If you cannot determine a field, use null.
- Do not include explanations outside the JSON.
- Do not fabricate evidence. If evidence location is uncertain, use null for the uncertain fields.

Table 12: Prompt used for judging agent answers on RDQA questions.

Judge Prompt

You are an RDQA answer judge.

Compare a model prediction against the ground-truth answer_value.

Judging rules:

- Return true if the prediction and ground-truth answer are semantically consistent.
- Return false if they are inconsistent, contradictory, missing, or materially different.
- Do not require exact string match. Accept equivalent wording, formatting, units, casing, punctuation, and harmless expansions abbreviations.
- For numbers, dates, names, titles, yes/no answers, and NONE/null-style answers, require the same meaning.
- Use the query and constraints only to resolve ambiguity; judge the final answers, not the evidence.
- Return exactly one lowercase token: true or false.

A.6 LLM-Judge Reliability Audit

As described in §3.3, answers that fail normalized exact match are scored by an LLM judge. To quantify judge reliability, we conducted a human audit covering *all* judge decisions across all evaluated agent systems and baselines, recording the agreement rate, false-positive rate, and false-negative rate against human labels, with the human label taking precedence in all reported scores.

The audit yields 100.0% agreement, with a 0% false-positive rate and a 0% false-negative rate; consequently, the human audit changed no reported metric, and judge noise does not propagate into Table 3. We attribute this reliability to two design choices: the judging task is deliberately kept mild—the judge receives the question, the reference answer, the full answer-variant list, and the agent’s response, and makes only a binary semantic-equivalence decision (Table 12); and the judge is a near-frontier model that is not among the evaluated backbones, avoiding self-preference effects. The audit therefore serves as a one-time validation of the scoring pipeline: future users of RDQA can rely on the automated two-stage protocol (§3.3) without repeating the manual audit, and reported scores are reproducible from the released harness alone.

Table 13: One representative example for each (capability family, sub-category) pair in RDQA, selected from the released benchmark.

Family-Sub	ID	Query	Answer
<i>Document modality</i>			
DTE (L1)			
DTE-VER	0082	In this university course catalog PDF, what exact short phrase does the instructor of the autofiction-themed intermediate creative writing workshop use to describe the consequence of using AI?	will count as plagiarism

Continued on next page

Table 13 (continued)

Family-Sub	ID	Query	Answer
DTE-ID	0068	In this technical datasheet PDF for a handheld spectrum analyzer, which military/industry standard is cited as the basis for the instrument's electromagnetic compatibility compliance?	GJB3947A-2009
DTE-META	0891	In the city of Houston FY2025 adopted budget financial letter PDF, what is the exact title of Melissa Dubowski as stated in the letterhead?	Chief Business Officer/Director of Finance
DTE-NUM	1186	In this e-scooter owner's manual PDF, what is the listed Rated Maximum Load in the product specifications?	265 lb
DTE-CAT	0090	In this ocean science meeting agenda PDF, what is the exact name of the restaurant where the SCOR-hosted dinner takes place on the evening of 29 October 2025?	Pescaito Carnes y Mariscos
DTE-TEMP	0207	In this municipal construction solicitation PDF for a Florence, SC project, by what date and time is the deadline for the submission of all inquiries?	March 14,2025 at 5:00 pm
DTE-REL	1151	In this dual-credit syllabus PDF, what is the name and title of the person students are directed to contact if concerns remain after speaking with the instructor?	Leslie Richardson, Mathematics Dept. Chair
DMO (L2)			
DMO-EXT	0070	In this technical datasheet PDF, what is the maximum cable distance specified for the RS232 serial output?	50 ft
DMO-COUNT	0054	In this regulatory floor-plan form PDF, how many numbered instruction items are listed before the first illustrative diagram appears?	8
DMO-FILT	0012	In this annual report PDF, according to the table of contents, what is the title of the section that begins on page 15?	Risk Management of Valartis Group
DMO-FILT-MULTI	0638	In this CSR report PDF, within the Scope 1 & 2 Emissions table, what is the 2024 Absolute Scope 2 figure reported for the Precast Infrastructure and Engineered Systems Business Unit?	1,287
	0745	In this syllabus PDF, what is the fifth item listed under the section enumerating topics to be covered?	Casting and forming (metals, powders, ceramics, plastics)
DMO-RANK	0433	In this syllabus PDF, in the grading-percentages breakdown, which component carries the second-highest weight?	Project
DRV (L3)			
DRV-COMPUTE	0025	In this corporate ESG PDF, comparing the turnover figure and the donations allocated to social and cultural initiatives in the reporting year, what approximate share of turnover did those donations represent?	0.15%
DRV-CHAIN	0014	In this asset management trustee report PDF, how many of the AMC's schemes are reported as having zero unclaimed dividends as of 31 March 2024?	4
DRV-APPLY	0431	In this university course syllabus PDF, if a student earns a final weighted score of 82.5, what letter grade would they receive?	B-
DRV-INFER	0807	In this court ruling PDF, was Justice Barrett's level of agreement with the principal dissent identical to that of Justices Kagan and Jackson?	No
<i>Video modality</i>			
VTE (L1)			
VTE-HUD	0119	In ABYSSUS First Gameplay Demo — New BRUTAL SHOOTER like Doom with Cthulhu Steampunk Monsters, when the player looks toward the glowing green doorway in the cave corridor, what location label appears above the doorway?	Blessing Altar
VTE-EMBED	0821	In this Step Inside the International Space Station (POV Tour) video, what exact caution, instruction, or equipment label is visible on a panel during the station tour?	PORTABLE BREATHING APPARATUS
VTE-TITLE	0121	In INDIKA Gameplay Walkthrough FULL GAME No Commentary + ENDING, during the early title card with the large star-like shape in the center, what word is displayed on screen?	DUTY

Continued on next page

Table 13 (continued)

Family-Sub	ID	Query	Answer
VTE-SCROLL	0035	In How to List Credits on a Short Film : Filmmaking Techniques, what exact person name is readable in the first clear portion of the closing credits?	Zach Poulin
VTE-STEP	0036	In How to add end scrolling credits to your YouTube video or short film video, how do you make the text block scroll in your video project?	Enable the 'Roll' checkbox.
VMO (L2)			
VMO-ENUM	0032	In Transform Your Local Sports Livestreams with this One Device. Yolobox Extreme, what exact score line is shown on the first readable scoreboard graphic?	1-0
VMO-COUNT	0312	In Scoreboard Tutorial - Basketball, When the demonstrator is setting the main clock, how many individual button presses are visible on the numeric keypad to input '10:00'?	5
VMO-EXT	0527	In this How To Create News Tricker/Scrolling Text — After Effects Tutorial — Rolling Text video, what is the final choosen exact size of the 'SHEHAB EDITZ' text in pixels, as shown in the Character panel?	132
VMO-CHANGE	0780	What exact set, period, quarter, or inning indicator is shown when the scoreboard first appears?	2nd
VMO-CONSIST	0497	In this Live Stream Local Cricket Match With CrickPro Overlay video, when the overlay theme is chosen as 'ASHES 2022', what is the Text Font Family?	JetBrains Mono
VMO-CROSSCHECK	0832	According to the audio in the International Space Station (ISS) Tour - NASA Space Science Investigations video, what is the exact name of the first module or section they enter, per the commentator?	Kibo
VRV (L3)			
VRV-COMPUTE	0910	In this Disney BLASTED For Releasing WOKE Clothing Line For Kids! Consumer Confidence PLUMMETS! video, what statistic is highlighted in the embedded tweet about Americans' favorable impression of Disney?	77% to 33%
VRV-APPLY	0540	In this Crispy Bloomin' Potato Recipe video, what material should the bowl be covered with?	Saran
VRV-INFER	0049	In 5 Small YouTube Channels But They Make More Than Big Creators, how much money did the 18-year-old Ivy League foreign exchange student make in a single month after pivoting his niche?	\$80,000
<i>Audio modality</i>			
AIE (L1)			
AIE-SPEECH	1424	According to this audio recording of "KPAY 1290 AM [NewsTalk 1290] : March 01, 2017 09:00PM-10:00PM PST", what was the exact category of the tornado spotted in parts of Illinois?	F3
AIE-NONSPEECH	1465	In this audio "VOA [Voice of America] Africa : March 11, 2019 07:00PM-08:00PM EDT", what is the exact name of the Imagine Dragons song that is played?	Natural
AIE-SEG-LOC	1404	In this audio recording of "TT 184 Q 3 2020 Earnings", what specific news segment is discussed immediately after the UK news?	Worldwide News
AMO (L2)			
AMO-EXT	1508	According to this audio "VOA [Voice of America] Africa : October 17, 2019 02:00AM-03:00AM EDT", what is the exact name of the reporter who tells more about the story of the Venezuelan people walking through fires to...	Arash Arabasadi
AMO-LOC	1479	According to this audio "Air Talk with Harry Gray and Jerry Willis", the speaker states that a few curious characters will show up at what specific location?	Caltech
ARV (L3)			
ARV-COMPUTE	1529	According to this audio "Coca-Cola quarterly earnings- MP3", the cost of goods sold has increased by how much for the year?	9%

Continued on next page

Table 13 (continued)

Family-Sub	ID	Query	Answer
ARV-CHAIN	1371	According to this 3M 2022 Q3 Recording audio, what is the exact job title of Bruce Jermeland?	Senior Vice President of Investor Relations
ARV-APPLY	1375	According to this 3M Q3 2021 Recording audio, under what specific heading on their website is the accompanying presentation posted?	Quarterly Earnings

Backbone	Closed-Book Acc.
Claude Opus 4.6	0.10%
GPT-5.2	0.20%
Gemini 3 Flash	0.07%
Qwen3.6 Plus	0.00%
Kimi K2.5	0.07%
GLM-4.7v	0.00%

Table 10: Closed-book accuracy on RDQA: each backbone answers all 1,350 questions with all tools disabled, scored under the protocol of §3.3. Near-zero accuracy across all backbones confirms that RDQA items are not answerable from parametric knowledge alone.

Table 14: Per-question GPT-5.2 token usage sample and cost of the RDQA-SKILL reference baseline (one sample per modality). *main* columns count the agent’s tool-use loop; *vision* columns count per-frame OCR calls inside *vision_analyze*. Pricing: \$1.75/M input, \$14/M output (OpenRouter, May 2026).

Mod.	Item	main loop		main tokens			vision (per-frame OCR)			Cost (USD)
		calls	ext. tools	in	out	reasoning	calls	in	out	
PDF	RDQA_0012	3	2	7,509	506	129	0	0	0	0.020
Audio	RDQA_1382	3	2	11,083	856	229	0	0	0	0.031
Video	RDQA_0119	8	4	97,375	8,347	1,997	80	39,096	12,234	0.527
Total		14	8	115,967	9,709	2,355	80	39,096	12,234	0.584