PUBTABLES-QA: A MULTI-PAGE TABLE VISUAL QUESTION ANSWERING FOR CROSS-PAGE TABLE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Tables are a central medium for communicating quantitative evidence in scientific documents, yet their compact relational form makes robust interpretation depend on layout and context. However, existing multi-page and multi-table benchmarks may still allow *stochastic shortcut search*, where models locate local answer-bearing regions without systematically navigating the underlying table structure. To address this gap, we introduce **PubTables-QA**, a multi-page TableVQA benchmark dataset carefully designed to audit cross-page visual reasoning in authentic academic documents. Built upon the extensive structural table annotations of PubTables-v2, PubTables-QA introduces 1,871 table-based QA pairs, where each question is designed to be unanswerable from a single page alone. This design compels models to reconstruct cross-page table continuity across headers, rows, columns, and cells, or to integrate evidence distributed across multiple tables. PubTables-QA further provides a three-level fine-grained QA taxonomy: Document, Table, and Cell/column levels respectively, ranging from relevant evidence localization and table-level structural understanding, to precise cell/column-level value extraction. Experiments on recent MLLMs, including GPT-4o, Gemini-2.5-Pro, and Gemma-4 reveal that even the strongest models achieve 45.9% overall accuracy. Further, our evidence-aware evaluation confirms that, unlike existing multi-page benchmarks susceptible to stochastic shortcuts, PubTables-QA requires genuine cross-page evidence integration, establishing it as a substantial and unsolved challenge.[1]

1 INTRODUCTION

Tables are a foundational medium of scientific communication, compressing complex quantitative evidence into relational structures for analysis and argumentation (Kang et al., 2026; Xing et al., 2025; Day, 1998). Yet this compactness makes sound interpretation highly dependent on domain expertise, with meaning reconstructed from layout, contextual cues, and domain-specific knowledge (Smock et al., 2025). Consequently, robust table reasoning has become a key frontier for multimodal large language models (MLLMs) aiming to support professional analytical workflows where expert human judgment remains essential.

For human experts, table interpretation in complex documents is inherently spatial and often long-horizon, requiring evidence to be connected across multiple tables and document fragments (Deng et al., 2025; Ma et al., 2024; Zhao et al., 2024; Lohse, 1991). The challenge intensifies when a table spans multiple pages, requiring reconciliation of headers and data across page breaks to preserve relational continuity. Thus, recent MLLM benchmarks on TableVQA have shifted from isolated table understanding (Lompo & Haraoui, 2025; Souibgui et al., 2025; Kim et al., 2024; Mathew et al., 2021) toward document-level reasoning over multi-page and multi-table structures (Tab. 1), better reflecting human-AI collaborative settings (Borchmann et al., 2026; Foroutan et al., 2025; Van Landeghem et al., 2023; Ma et al., 2024).

Despite recent advancements, current benchmarks may not fully capture the complexity of human-centric table reasoning, inadvertently masking the true limitations of MLLMs. While multi-page

[1]Code and Dataset are available at https://anonymous.4open.science/r/PubTables-QA-584D.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **Comparison with existing document-level VQA benchmarks.** PubTables-QA is the benchmark featuring cross-page tables (Avg. 2.5 pages/table) that structurally require multi-page reasoning, with cell-level evidence annotations enabling complex table-centric evaluation. Models must follow tables that continue across pages, combine information from multiple tables, and justify answers with page-, table-, and cell-level evidence. ✓: supported; ✗: not supported; (–): not reported or not verifiable due to missing page-level annotation; ▲: formally multi-page input, but most questions are answerable from a single page (Sec. 3.4).

Benchmark	Scale		Document Structure				Question Characteristics		Evidence Annotations		
	#Docs	#QA	Avg. Pages	Avg. Pages/Table	Multi-Table Doc.	Cross-Page Table	Multi-Page QA	Multi-Table QA	Page	Table	Cell
TableVQA-Bench Kim et al. (2024)	–	1.5K	1.0	–	✗	✗	✗	✗	–	–	✗
Visual-TableQA Lompo & Haraoui (2025)	2.5K	6K	1.0	–	✗	✗	✗	✗	–	–	✗
TAT-DQA Zhu et al. (2022)	2.7K	16K	1.1	1.0	✓	✗	✗	✗	✗	✗	✗
DUDE Van Landeghem et al. (2023)	5K	41K	5.72	0.8	✓	✗	✗	✗	✗	✗	✗
MP-DocVQA Tito et al. (2023a)	6K	46K	8.27	0.8	✓	✗	✗	✗	✓	✗	✗
WikiMixQA Foroutan et al. (2025)	526	1K	1.0	0.4	✓	✗	▲	✓	✗	✓	✗
MMLongBench-Doc Ma et al. (2024)	135	1K	47.5	0.9	✓	✗	▲	✓	✓	✗	✗
PubTables-QA (Ours)	**296**	**1.8K**	**15.4**	**2.5**	✓	✓	✓	✓	✓	✓	✓

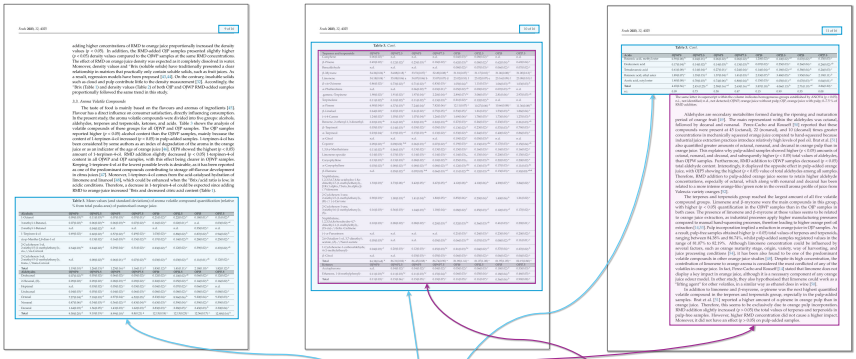
benchmarks have emerged, they often allow for *stochastic shortcut search* (Borchmann et al., 2026), a phenomenon where models accidentally locate answers within an isolated page or two rather than performing systematic reasoning (Xing et al., 2025; Ma et al., 2024). Such observations expose a critical blind spot: the ability to map logical paths through complex structures, namely *strategic navigation* (Borchmann et al., 2026) or *data-aware reasoning* (Gu et al., 2025), remains largely unvalidated. Inevitably, current accuracy-centric metrics fail to distinguish between models that genuinely understand multi-page structures and those that merely exploit statistical shortcuts.

To address these gaps and rigorously audit MLLM capabilities, we propose **PubTables-QA**, the first QA evaluation suite, specifically designed to test *cross-page* visual reasoning in authentic academic documents. Built upon the high-precision structural annotations of PubTables-v2 (Smock et al., 2025), our PubTables-QA uniquely forces models to navigate *multi-page* contexts and reason over *multi-table* structures (see Fig. 1). Unlike prior benchmarks, our questions are explicitly designed to be unanswerable from a single-page, forcing models to integrate information across page boundaries. This shifts the focus from basic table detection to complex reasoning, enabling *hierarchical navigation*. Our key contributions are as follows:

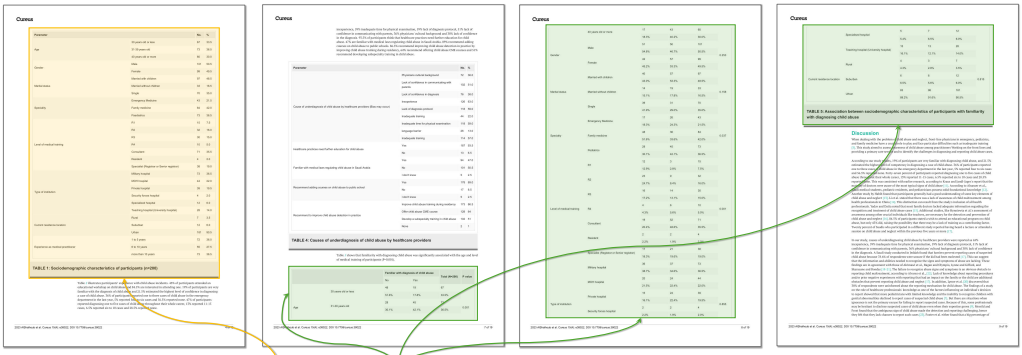
- **Cross-page table reasoning:** We introduce a set of tasks where correct answers can only be derived by integrating split table segments across multiple pages (Sec. 2.1). This addresses the *context dilution* and *needle-in-a-haystack* challenges identified in recent long-document studies (Borchmann et al., 2026; Ma et al., 2024).
- **Hierarchical QA:** We provide a structured evaluation across three levels, such as *Document—Table—Cell/column* (Sec. 2.2). This enables a diagnostic analysis of model performance, from basic information extraction to high-level insight generation over multiple tables, similar to the multi-table reasoning objectives in MT-RAIG (Seo et al., 2025).
- **Structured evidence annotation:** Each QA pair is accompanied by fine-grained evidence, including bounding box coordinates, page indices, and cell values. This allows for the evaluation of not just answer accuracy, but also whether the model's reasoning is grounded in the gold evidence (Kang et al., 2026; Jiang et al., 2025).

2 PubTables-QA

We propose **PubTables-QA**, a TableVQA benchmark focusing on multi-page, multi-table visual reasoning in scientific documents. It is built upon the test split of PubTables-v2 (Smock et al., 2025), a large-scale dataset that provides page-level and cross-page table structure annotations. Reasoning over page-spanning tables requires more than locating an answer within a long document. Therefore, a model must *(i) navigate across pages; (ii) associate fragmented table regions into a coherent structure;* and *(iii) ground the final answer in precise cell-level evidence.* We conceptualize this process as inherently *hierarchical*, as illustrated in Fig. 1. Accordingly, PubTables-QA adopts this hierarchical QA design that disentangles these capabilities and evaluates them at distinct granularities.



[L1] Question: The body text refers to one volatile compound described as a "lifting agent" for other volatiles, analogous to the role of ethanol in wine. According to Table 3, what is the difference between this compound's value in pulp-free OJWP0 and pulp-added OJP0?
Answer: 11.16
[L2] Question: On which document pages does the Table 3 appear?
Answer: 9, 10, 11



[L3] Question: Based on Table 1 and Table 5, what percentage of the least represented marital status group is familiar with the diagnosis of child abuse?
Answer: 17.8

Figure 1: **Examples of L_1 (Document-level), L_2 (Table-level), and L_3 (Cell/column-level) questions in PubTables-QA.** Each question requires level-specific evidence across pages. More examples are provided in Appendix B.1.

2.1 Task Definition

We define *cross-page* TableVQA as a task where a model answers a natural-language question by reasoning over tables in multi-page document images. Unlike conventional TableVQA, which assumes a self-contained table on a single-page, our task requires models to treat tables as the primary source of evidence and resolve queries that fundamentally cannot be answered without integrating information across page boundaries. Formally, given a document image sequence $\mathcal{D} = \{p_1, p_2, \ldots, p_N\}$ and a natural-language question q, the model is required to generate an answer a as follows:

$$f_\theta(\mathcal{D}, q) \rightarrow a.$$

The answer a takes different forms depending on the question type, including a numerical value, an entity or text span, or a list of table/page identifiers (Kim et al., 2024). Each answer is grounded in a structured evidence set $\mathcal{E}$, which includes table-internal information (*regions, cells, headers*) and document-level context (*page indices, surrounding text, layout*). This task is further characterized by the following three constraints:

- **Cross-page table continuity:** In this task, a single table is fragmented across multiple pages, with required evidence distributed across page regions. The model must recognize these fragments as belonging to a coherent table and integrate them across pages. This integration must preserve relationships among headers, rows, columns, and cells. Single-page retrieval is insufficient to solve the task.

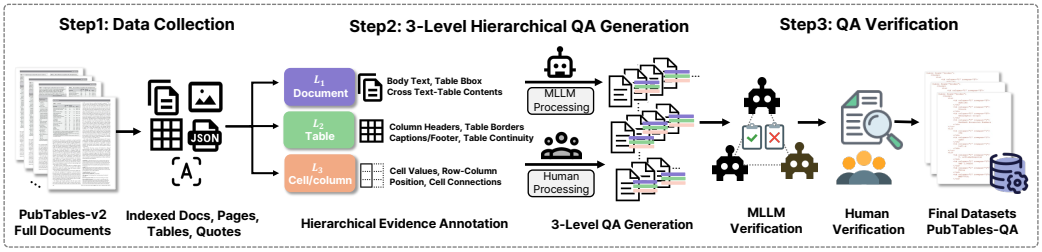


Figure 2: **Overview of the PubTables-QA construction pipeline.** We apply a three-level hierarchical annotation scheme (L_1 Document-level, L_2 Table-level, and L_3 Cell/column-level) to generate structured QA pairs.

- **Visual-structural table understanding:** Since the inputs are page images, the model must recover table structure from visual layout rather than serialized text. This requires identifying cell boundaries, row–column alignment, header continuation, and page breaks across multi-page tables.
- **Hierarchical reasoning:** PubTables-QA evaluates reasoning at three levels: document, table, and cell/column. Each level is targeted by distinct question types, preventing models from relying on a single dominant strategy and directly exposing weaknesses at specific reasoning depths.

2.2 Three-level QA taxonomy and evidence annotations

Structured representation for multi-page tables. In PubTables-QA, each document consists of a sequence of page images with structured table annotations, including bounding boxes, cell texts, headers, and row–column structures (Smock et al., 2025). We reconstruct multi-page tables by grouping page-level fragments using table locations, header consistency, and adjacency relations. Each table is linked to its caption, footer, surrounding OCR text, and body-text sentences that explicitly reference the table. By unifying scattered table fragments into a single logical structure with explicit cross-page links, we obtain the ground truth needed to construct and evaluate cross-page reasoning questions.

Evidence annotations. Each QA instance is annotated with a question, a gold answer, and structured evidence metadata, stored in the `annotation_fact` field (Appendix A.6). In addition to supporting QA construction and verification, these fine-grained annotations enable evaluation beyond answer accuracy, by comparing a model's reasoning trace against the gold evidence, we can detect cases where models reach correct answers through incorrect or shortcut-based reasoning paths. Further details are provided in Fig. A.3 (Appendix A.4). Additionally, we define six complementary evidence sources:

- Table bounding box: Physical locations of table fragments on each page.
- Surrounding text: Table-local text, such as captions, footers, footnotes, and nearby OCR text.
- Body text: Document-level sentences outside the table that mention specific tables or numerical claims.
- Table headers: Column headers and row labels.
- Cell values: Raw cell values and derived values used to compute the answer.
- Table layout: Structural information about how table fragments are split, continued, and aligned across pages.

Three-level QA taxonomy. PubTables-QA is organized into three hierarchical levels, each requiring a different scope of reasoning (Tab. 3). Full generation rules and annotation examples are provided in Appendix A.2.

- Document-level (L_1): Requires cross-page linking between body text and table values, making Body Text the defining evidence source.
- Table-level (L_2): Requires identifying how table fragments are split, continued, and aligned across pages, with Table Layout as the primary evidence source.
- Cell/column-level (L_3): Requires cell-value extraction, aggregation, and multi-step computation across rows, columns, or multiple tables, with Table headers and Cell Values as the main evidence sources.

Table 2: Overall statistics.

Statistic	Value
# Documents	296
# Pages	4,552
# Tables	785
# QA pairs	1,871
Avg. Q Length	20.8
Avg. A Length	2.3

Table 3: **Hierarchical distribution and complexity statistics.** Pages, BBoxes, and Evidence denote average counts per QA instance.

Level	Category	Total	Pages	BBoxes	Evid.
L_1 **Doc.**	Text-Table Reasoning	182	2.6	3.2	4.0
	Document Structure	73	2.7	3.1	3.0
	Cross-section Synthesis	67	2.5	2.9	3.0
L_2 **Table**	Table-wide Reasoning	168	2.2	2.3	4.0
	Cross-page Continuation	84	2.1	2.1	2.0
	Cross-table Structure Match	68	3.2	3.3	5.0
	Table Layout	48	2.4	2.7	4.5
L_3 **Cell/col.**	Single-table Retrieval-Computation	552	2.8	2.9	2.0
	Cross-table Retrieval-Computation	350	3.3	3.6	5.0
	Multi-hop	279	2.8	3.1	4.0
Overall	**10 Categories**	**1,871**	**2.8**	**3.0**	**4.0**

Cross-page evidence constraint. All questions require evidence that spans at least two pages. This includes evidence drawn from different parts of the same table, from related tables across pages, or from table-text combinations across pages.

2.3 QA CONSTRUCTION PIPELINE

We construct QA pairs through a three-stage pipeline (Step2 in Fig. 2): (STAGE 1) Annotation parsing and seed generation; (STAGE 2) MLLM-based Question generation & rule-based filtering; and (STAGE 3) Human annotation.

STAGE 1: Annotation parsing & seed generation. PubTables-v2 provides each document as a sequence of page images with page-level table annotations. We reconstruct multi-page tables by grouping fragments using table locations, header consistency, and the adjacency_matrix (Smock et al., 2025), and link each table to its caption, footer, surrounding OCR text, and body-text references. Rule-based templates then produce seed questions at three levels, all requiring evidence spanning at least two pages (Appendix A.2).

STAGE 2: MLLM-based question Generation & Rule-based Filtering. To prevent template-matching shortcuts, GPT-5 (OpenAI, 2025) generates 3,866 candidate questions from the seeds, document pages, and `annotation_fact`, changing only the wording for annotation-derived questions so that their answers remain fixed. Rule-based checks on question form, answer leakage, and annotation consistency retain 3,101 candidates, and deduplication leaves 2,749 (Appendix A.3).

STAGE 3: Human annotation. Five annotators write additional questions for documents exceeding ten pages, focusing on cross-table, multi-hop, and text-table reasoning patterns that are difficult to generate by rule. Annotators use an authoring tool (Fig. A.1 in Appendix A.1) to compose QA pairs with evidence annotations. Details of the annotation procedure are provided in Appendix A.4.

2.4 QUALITY CONTROL AND VERIFICATION

Multi-validator consensus filtering. Two MLLM validators with different architectures, GPT-4o (OpenAI, 2024) and Qwen2.5-VL-72B (Bai et al., 2025), independently judge each candidate against its document images for answer consistency, evidence sufficiency, visual grounding, and category validity. Candidates rejected by both validators are removed, and those on which they disagree are forwarded to human review. The validators also enforce the benchmark design by removing questions answerable without the document and multi-page questions answerable from a single evidence page. In total, MLLM verification removes 813 candidates (29.6%). Detailed verification criteria and filtering statistics are provided in Appendix A.5.

Human verification. Five human annotators then re-examined candidates passed by MLLM verification, applying the same criteria as the MLLM validators. A candidate was removed or corrected only when two annotators agreed or when the authors confirmed a single annotator's flag against the source document, resulting in 1,244 generated questions in two rounds. In the fully double-annotated second round, agreement on whether the gold answer is wrong reached Cohen's κ = 0.691. Together with 627 human-written questions, the final benchmark contains 1,871 questions (Tab. 2), with detailed statistics in Appendix B.2.

Table 4: **Evaluations of various models on PubTables-QA**. We report performance across the 3-Level QA (Document-level, Table-level, Cell/column-level) across 10 sub-categories. Gemini-2.5-Pro (Thinking) achieves the highest overall performance, followed by Qwen3-VL (Thinking) as the best open-source model. **Bold** and underline denote the best and second-best within each model group (open-source and proprietary), respectively.

Model	#Params	Context Window	L_1 Doc.	L_2 Table	L_3 Cell/col.	Overall
Open-source Models						
Qwen2.5-VL (Bai et al., 2025)	7B	32k	37.0	33.5	14.5	22.0
Qwen3-VL (Yang et al., 2025)	8B	128k	41.9	38.1	18.4	26.2
Qwen3-VL (Thinking) (Yang et al., 2025)	8B	128k	41.1	**56.0**	**42.3**	44.7
Qwen3.5 (Qwen Team, 2026)	9B	128k	42.3	40.6	36.3	38.1
InternVL3 (Zhu et al., 2025)	8B	256k	29.4	26.3	10.3	16.7
InternVL3.5 (Wang et al., 2025)	8B	256k	30.1	25.8	11.5	17.4
Gemma-4 (Thinking) (DeepMind, 2025)	31B	128k	37.0	55.8	41.1	43.3
Qwen3-VL (Yang et al., 2025)	32B	128k	**47.4**	42.3	25.0	32.1
Proprietary Models						
Gemini-2.5-Flash (Comanici et al., 2025)	–	128k	25.9	22.3	12.1	16.4
Gemini-2.5-Pro (Thinking) (Comanici et al., 2025)	–	128k	**51.7**	**57.7**	**40.7**	**45.9**
GPT-4o (OpenAI, 2024)	–	128k	39.7	35.9	25.9	30.2

3 Evaluations

3.1 Evaluation metrics

We identify two challenges in evaluating MLLMs on long documents: (i) models tend to produce long-form responses interleaved with reasoning rather than concise short-form answers, and (ii) answer formats are heterogeneous, ranging from numbers and text spans to ordered lists, page ranges, and table–page combinations. We therefore extract final answers with GPT-4o-mini and score them with **Type-Aware Accuracy (TAA)**. Each question is assigned one of 9 answer types from the question and gold answer alone, before any model prediction is inspected, and scored as binary correct/incorrect under type-specific rules. Unlike ANLS (Biten et al., 2019), which gives partial credit to near-miss values, TAA rejects any value error but accepts format-only variants. TAA is thus 3–6 points lower than ANLS across all models and better reflects true correctness. We report L_1, L_2, L_3, and overall scores (mean over all questions, 0–100). Full detailed rules are provided in Appendix D.1.

3.2 Evaluation setup

We evaluate state-of-the-art MLLMs on PubTables-QA. Our evaluation suite comprises open-source models ranging from 7B to 32B parameters and proprietary models. To ensure a fair comparison, all models are given the same input format, where each document page is rendered as a JPEG image and paired with the query. Image resolutions are normalized consistently so that the number of visual tokens per page remains uniform across models. Since the average page count fits within the context window of every evaluated model, no truncation is applied. All models are prompted with a unified instruction to answer the question based solely on the provided document images, ensuring that performance differences reflect model capability. For Thinking-enabled models, we use each model's default thinking mode without additional prompt modifications. Full prompt templates are provided in Appendix D.2, and model-specific configurations in Appendix D.3.

3.3 Main results

Table 4 presents the performance of evaluated models across the three question levels. Non-Thinking models perform best at the Document-level (L_1) and Table-level (L_2), but exhibit a sharp drop at the Cell/column-level (L_3). This indicates that current MLLMs can recognize broad document context and table structure, but struggle with precise cell-value extraction and row-column relationship understanding. In particular, accurately locating specific cells within complex multi-page documents and performing arithmetic operations or multi-hop reasoning over them remain challenging for current models. The L_3 gap is particularly striking. Non-Thinking models range from 10.3% to 36.3% at L_3, whereas Thinking-enabled models reach 40.7–42.3%. At L_2, model variance is also substantial. All Thinking-enabled models exceed 55%, whereas Gemini-2.5-Flash, InternVL3, and

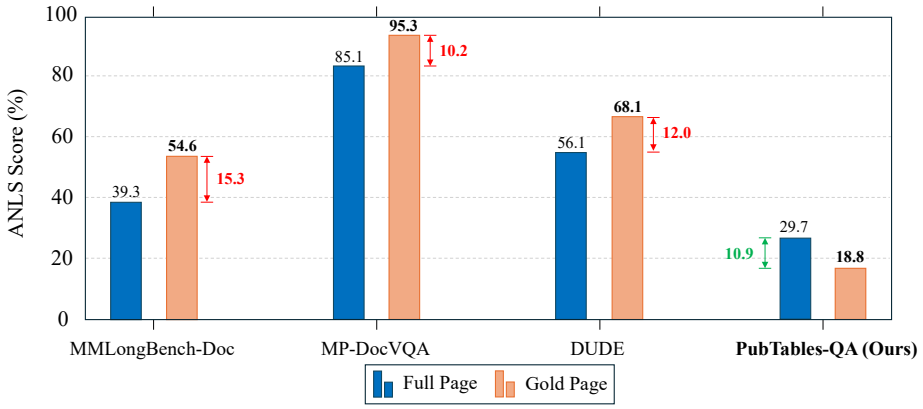


Figure 3: **Single-page experiment across benchmarks (Qwen3-VL-8B, ANLS).** Gold page: the gold evidence page given alone. For questions with multiple evidence pages, each page is given individually and scores are averaged (avg. evidence page). For MMLongBench-Doc, MP-DocVQA, and DUDE, Full-document $<$ Gold-page, indicating that a single page suffices to answer most questions, whereas PubTables-QA drops with a single evidence page.

InternVL3.5 fall below 27%, suggesting that multi-page table identification is highly sensitive to model architecture and training data. Within the Qwen3-VL family, the Thinking-enabled 8B model outperforms both the standard 8B and 32B variants at L_2 and L_3, achieving an overall score of 44.7% compared with 26.2% and 32.1%, respectively, although the 32B model leads at L_1 (47.4%). This indicates that extended reasoning can be more beneficial than parameter scaling alone for multi-page table understanding. Among proprietary models, Gemini-2.5-Pro (Thinking) achieves the highest overall performance of 45.9%, nearly tripling Gemini-2.5-Flash at 16.4%. Notably, the three Thinking-enabled models achieve the three highest overall scores, and the open-source Qwen3-VL-8B (Thinking) performs on par with Gemini-2.5-Pro (Thinking). However, whether these gains reflect genuine evidence-based reasoning or superficial pattern exploitation remains unclear. We address this question through an LLM-judge evaluation in Section 3.5. Overall, the results show that the core bottleneck of current MLLMs lies in fine-grained information extraction within tables. Precisely locating target cells across complex multi-page layouts and performing cell/column-level arithmetic over scattered evidence remain fundamental challenges that extended reasoning only partially mitigates and model scaling alone does not reliably resolve. Without page images, all model score only 2.1–4.4% overall (Appendix C.2), indicating that questions cannot be answered from language priors alone.

3.4 Oracle setting analysis

We compare full-document and single-page oracle performance across PubTables-QA and three existing multi-page benchmarks (Fig. 3). The oracle input is the gold evidence page, and for questions with multiple evidence pages, each page is given individually and scores are averaged (*avg. evidence page*). This applies to all PubTables-QA questions and to 362 of the 1,082 MMLongBench-Doc questions. For MMLongBench-Doc, MP-DocVQA, and DUDE, the oracle outperforms the full document by 10–16 ANLS points, showing that their questions are largely answerable from a single page. PubTables-QA shows the opposite, dropping by 11 points (29.7% $\rightarrow$ 18.8%) with a single evidence page. Thus, existing benchmarks mainly test robustness to irrelevant pages, whereas PubTables-QA requires genuine cross-page reasoning.

3.5 Evidence-grounded reasoning evaluation via LLM-as-judge

MLLM-based document understanding models may reach correct answers through shortcuts rather than evidence-grounded reasoning. Using the structured evidence annotations in PubTables-QA, we assess, for correctly answered instances with available thinking traces, whether the model actually used the gold evidence required for the answer. GPT-4o-mini is used as an LLM-as-a-Judge (Zheng et al., 2023), taking the question, gold evidence, thinking trace, and predicted answer as input, and classifying each trace as *Faithful* or *Unfaithful*. As reported in Table 5, 83.1% of Qwen3-VL-8B (Thinking)'s correct answers and 83.3% of Gemma-4-31B (Thinking)'s correct answers are judged faithful; roughly one in six correct answers is thus not grounded in the gold evidence. The largest drop occurs at the Table-level (L_2), where both models lose over 9%p, suggesting that table structure

Table 5: **Evidence faithfulness analysis.** *Original*: standard evaluation. *Faithful*: performance counted only for instances where GPT-4o-mini judges the reasoning as genuinely grounded in gold evidence.

Model	Setting	L_1	L_2	L_3	Avg.
Qwen3-VL (Thinking, 8B)	*Original*	41.1	56.0	42.3	44.7
	Faithful (83.1%)	33.9	46.1	35.1	37.1
	Δ	−7.2	−9.9	−7.2	−7.6
Gemma-4 (Thinking, 31B)	*Original*	37.0	55.8	41.1	43.3
	Faithful (83.3%)	30.7	46.2	34.2	36.0
	Δ	−6.3	−9.6	−6.9	−7.3

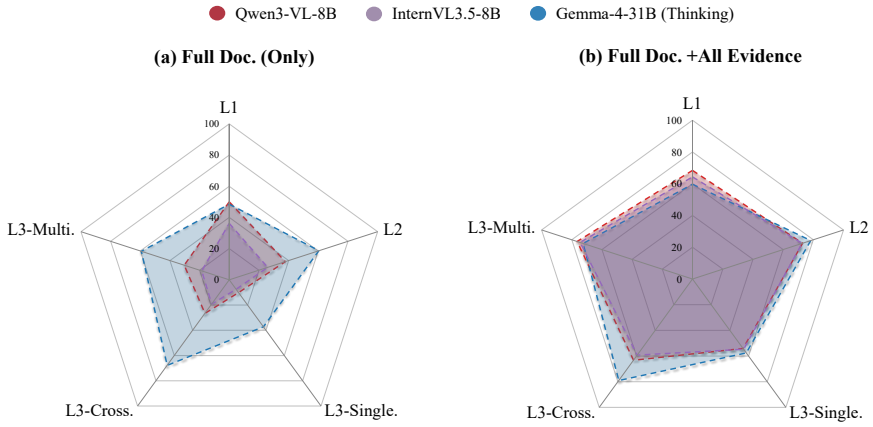


Figure 4: **Effect of evidence granularity on model performance.** (a) Full document images only. (b) Full document images with all structured evidence (page ids, bounding boxes, and cell values). Single: Single-table, Cross: Cross-table, Multi: Multi-hop.

identification is particularly susceptible to shortcut reasoning. These findings directly demonstrate that aggregate accuracy overestimates the actual reasoning capabilities of MLLMs. Detailed case analyses are provided in Appendix C.5.

3.6 Evidence granularity analysis

We compare the full-document baseline with the full-evidence setting (+Page IDs, +BBox, +Cell Values; Fig. 4) to examine which type of information most benefits model reasoning. At L_3, the 8B models perform poorly under the full-document setting (e.g., InternVL3.5-8B: Single 7.5%, Cross 19.2%; Qwen3-VL-8B: Single 10.1%, Cross 26.6%), revealing that fine-grained table perception is a major bottleneck. With full evidence, single-table performance increases sharply (InternVL3.5-8B: 7.5%→54.7%; Qwen3-VL-8B: 10.1%→54.1%), indicating that models can reason over cell values once the relevant cells are explicitly identified. Cross-table performance follows the same pattern (InternVL3.5-8B: 19.2%→59.3%; Qwen3-VL-8B: 26.6%→63.3%), suggesting that integrating information across tables is not the main obstacle once the required cells from each table are identified. Gemma-4-31B (Thinking) already performs well on cross-table reasoning under the full-document setting (67.8%) and gains less from full evidence (67.8%→79.1%). With full evidence, the three models converge (63.7–67.6% overall, compared with 17.4–43.3% under the full-document setting), indicating that most of the gap between models stems from locating and reading the required values rather than from reasoning over them. A fine-grained analysis with incremental evidence stages is provided in Appendix C.4.

4 Related works

Table extraction datasets. Table Extraction (TE) aims to detect tables within documents and recover their internal structure, and has been driven by large-scale datasets such as PubTabNet (Zhong et al., 2020), FinTabNet (Zheng et al., 2021), and PubTables-1M (Smock et al., 2022), which provide fine-grained annotations of cell structure, row–column relations, and hierarchical headers.

In particular, aligning PDF documents with their XML sources to derive cell-level bounding boxes and structural metadata (Zhong et al., 2020; Zheng et al., 2021) has established a strong foundation for improving table recognition performance (Rajput et al., 2026; Kang et al., 2026). More recently, PubTables-v2 (Smock et al., 2025) extends the scope from isolated tables to page-level and multi-page settings, addressing the challenge of linking tables distributed across multiple pages into a unified logical structure. However, these TE datasets are primarily designed for structural recovery and offer limited support for task-oriented question answering or higher-order reasoning. Benchmarks that explicitly require table-centric QA over multi-page structures remain largely underexplored.

Document-level multimodal question answering. Document VQA benchmarks have evolved from single-page settings, such as DocVQA (Mathew et al., 2021) for document images and TAT-DQA (Zhu et al., 2022) for financial reports with embedded tables, to multi-page and long-context environments including MP-DocVQA (Tito et al., 2023a;b), DUDE (Van Landeghem et al., 2023), and MMLongBench-Doc (Ma et al., 2024). The scope has further expanded to long-context reasoning through page-level retrieval and multimodal evidence aggregation (Deng et al., 2025; Dong et al., 2025). However, these benchmarks primarily focus on holistic understanding of entire documents and treat tables as one of several information sources rather than the central object of reasoning. As a result, structural properties of tables, hierarchical relationships between cells, header–data alignment, and spanning cells receive limited attention (Wu et al., 2025; Mathur et al., 2024), and reasoning over tables fragmented across page boundaries remains underexplored.

Visual table understanding. With the rapid progress of MLLMs (Yang et al., 2025; Comanici et al., 2025; Wang et al., 2025), recent benchmarks evaluate models on tables in visual form. TableVQA-Bench (Kim et al., 2024) reformulates text-based table QA datasets into image inputs to evaluate MLLMs, while Visual-TableQA (Lompo & Haraoui, 2025) emphasizes the role of visual cues such as cell colors and boldface information often lost in serialized formats like HTML or Markdown. More recently, this line of work has been extended to multimodal reasoning over tables, charts, and captions, as in WikiMixQA (Foroutan et al., 2025). However, most existing visual table QA benchmarks are limited to single-table and single-page settings, and often rely on synthetic data that poorly reflects the complexity of real documents. Page-spanning tables and reasoning across multiple tables remain rarely studied, despite being common in real scientific documents.

5 Conclusion

We presented PubTables-QA, a hierarchical Question-Answering benchmark for evaluating multi-page, multi-table visual reasoning in scientific documents. Built on PubTables-v2's structural annotations, we designed a three-level QA taxonomy (L_1 Document, L_2 Table, L_3 Cell/column) with 1,871 QA pairs across 296 documents, where every question structurally requires cross-page reasoning. Our evaluation of 11 state-of-the-art MLLMs reveals that even the best model, Gemini-2.5-Pro (Thinking) achieves 45.9% overall. Beyond measuring answer correctness alone, our benchmark provides an evidence-grounded evaluation framework that verifies whether models correctly locate and utilize the evidence required for answer derivation. We believe PubTables-QA can serve as a reliable diagnostic tool for distinguishing models that genuinely understand and reason over tables from those that merely rely on superficial cues, and hope it facilitates future research in retrieval-augmented multimodal generation, hierarchical reasoning, and visual layout-aware modeling.

Limitations

PubTables-QA is built on PMC Open Access biomedical articles and may not fully generalize to domains with different table conventions, such as financial or legal documents. Its scale (1,871 QA pairs across 296 documents) is smaller than MP-DocVQA or DUDE, as we prioritize quality over volume, with every question targeting cross-page table reasoning rarely required by existing benchmarks and carrying fine-grained evidence annotations. The pipeline relies on LLMs for question generation, validation, answer extraction, and faithfulness judgment, and although we use validators from different model families, model-specific biases cannot be fully excluded. Finally, faithfulness rates should be read as conservative lower bounds, since models may reason correctly through paths not captured by the pre-annotated evidence.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks with required disclosure. Additionally, we used generative AI tools to draft parts of the research paper and to edit the paper to improve readability. We reviewed all AI-assisted work and revised it as necessary. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 4291–4301, 2019.

Łukasz Borchmann, Jordy Van Landeghem, Michał Turski, Shreyansh Padarha, Ryan Othniel Kearns, Adam Mahdi, Niels Rogge, Clémentine Fourrier, Siwei Han, Huaxiu Yao, et al. Strategic navigation or stochastic search? how agents and humans reason over document collections. *International conference on machine learning (ICML)*, 2026.

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

Robert A Day. How to write and publish scientific papers. *Memórias do Instituto Oswaldo Cruz*, 93 (3), 1998.

Google DeepMind. Gemma 4: Open models based on gemini research and technology. *Technical Report*, 2025. https://ai.google.dev/gemma.

Chao Deng et al. LongdocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *Proceedings of the 63rd annual meeting of the association for computational linguistics (ACL)*, 2025.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimodal generation for document question answering. In *Advances in neural information processing systems datasets and benchmarks track (NeurIPS D&B)*, 2025.

Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. Wikimixqa: a multimodal benchmark for question answering over tables and charts. In *Findings of the association for computational linguistics (ACL)*, pp. 24941–24958, 2025.

Ken Gu, Zhihan Zhang, Kate Lin, Yuwei Zhang, Akshay Paruchuri, Hong Yu, Mehran Kazemi, Kumar Ayush, A Ali Heydari, Maxwell A Xu, et al. RADAR: Benchmarking language models on imperfect tabular data. In *Advances in neural information processing systems datasets and benchmarks track (NeurIPS D&B)*, 2025.

Jun-Peng Jiang, Yu Xia, Hai-Long Sun, Shiyin Lu, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Multimodal tabular reasoning with privileged structured information. *Advances in neural information processing systems (NeurIPS)*, 2025.

Xiaoqiang Kang, Zimu Wang, Xiaochen Zi, Xiaobo Jin, Kaizhu Huang, Fei Yin, and Qiufeng Wang. A benchmark and method for photographed table reasoning. *Pattern recognition*, pp. 113355, 2026.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. TableVQA-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024.

Jerry Lohse. A cognitive model for the perception and understanding of graphs. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 137–144, 1991.

Boammani Aser Lompo and Marc Haraoui. Visual-tableQA: Open-domain benchmark for reasoning over table images. *Advances in neural information processing systems workshop: Foundations of reasoning in language models (NeurIPSW)*, 2025.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. In *Advances in neural information processing systems datasets and benchmarks track (NeurIPS D&B)*, 2024.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 2200–2209, 2021.

Suyash Vardhan Mathur, Jainit Sushil Bafna, Kunal Kartik, Harshita Khandelwal, Manish Shrivastava, Vivek Gupta, Mohit Bansal, and Dan Roth. Knowledge-aware reasoning over multimodal semi-structured tables. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the association for computational linguistics (EMNLP)*, pp. 14054–14073, 2024.

OpenAI. Gpt-4o system card. *Technical Report*, 2024. https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-5 system card. *Technical Report*, 2025. https://openai.com/index/gpt-5-system-card/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Jahanvi Rajput, Dhruv Kudale, Saikiran Kasturi, Utkarsh Verma, and Ganesh Ramakrishnan. Tables decoded: Delta for structure, tarqa for understanding. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 2775–2785, 2026.

Kwangwook Seo, Donguk Kwon, and Dongha Lee. Mt-raig: Novel benchmark and evaluation framework for retrieval-augmented insight generation over multiple tables. In *Findings of the association for computational linguistics (ACL)*, 2025.

Brandon Smock, Rohith Pesala, and Robin Abraham. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4634–4642, 2022.

Brandon Smock, Valerie Faucon-Morin, Max Sokolov, Libin Liang, Tayyibah Khanam, and Maury Courtland. Pubtables-v2: A new large-scale dataset for full-page and multi-page table extraction. *arXiv preprint arXiv:2512.10888*, 2025.

Mohamed Ali Souibgui, Changkyu Choi, Andrey Barsky, Kangsoo Jung, Ernest Valveny, and Dimosthenis Karatzas. DocVXQA: Context-aware visual explanations for document question answering. *International conference on machine learning (ICML)*, 2025.

Ruben Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page docVQA. *Pattern Recognition*, 2023a.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023b.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 19528–19540, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Pengzuo Wu, Yuhang Yang, Guangcheng Zhu, Chao Ye, Hong Gu, Xu Lu, Ruixuan Xiao, Bowen Bao, Yijing He, Liangyu Zha, Wentao Ye, Junbo Zhao, and Haobo Wang. RealHiTBench: A comprehensive realistic hierarchical table benchmark for evaluating LLM-based table analysis. In *Findings of the association for computational linguistics (ACL)*, pp. 7105–7137, 2025.

Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Lingjiao Chen, Dongmei Zhang, Surajit Chaudhuri, and H. V. Jagadish. MMTU: A massive multi-task table understanding and reasoning benchmark. In *Advances in neural information processing systems datasets and benchmarks track (NeurIPS D&B)*, 2025.

An Yang, Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen3-vl technical report. *arXiv preprint*, 2025. https://github.com/QwenLM/Qwen3-VL.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Shu Wei, Binghong Wu, Lei Liao, Yongjie Ye, Hao Liu, Wengang Zhou, et al. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in neural information processing systems (NeurIPS)*, 37:7185–7212, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems (NeurIPS)*, 2023.

Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 697–706, 2021.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno-Yepes. Image-based table recognition: Data, model, and evaluation. In *European Conference on Computer Vision (ECCV)*, 2020.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4857–4866, 2022.

Jinguo Zhu, Weiyun Chen, Zhe Wang, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

TABLE OF CONTENTS

A **Benchmark construction details**
- A.1 QA generation pipelines
- A.2 Rule-based generation details
- A.3 MLLM-based question generation details
- A.4 Annotation procedure
- A.5 Quality assurance
- A.6 Annotation examples

B **Benchmark statistics**
- B.1 QA examples
- B.2 Detailed benchmark statistics

C **Ablation studies**
- C.1 Cell/column level analysis
- C.2 Question-only baseline
- C.3 Modality dependency analysis
- C.4 Evidence granularity analysis
- C.5 Evidence faithfulness details

D **Experimental details**
- D.1 Evaluation metrics
- D.2 Prompt instructions
- D.3 Model hyperparameters

E **Error analysis**

A BENCHMARK CONSTRUCTION DETAILS

A.1 QA GENERATION PIPELINES

We develop two separate web-based tools for QA construction: a QA generation tool and a QA verification tool.

QA generation tool. The generation tool (Fig. A.1) supports human annotators in writing QA pairs directly from document images. It provides a multi-page document viewer, an interface for composing question-answer pairs, and evidence annotation fields for recording page indices, bounding boxes, and cell values associated with each QA pair.

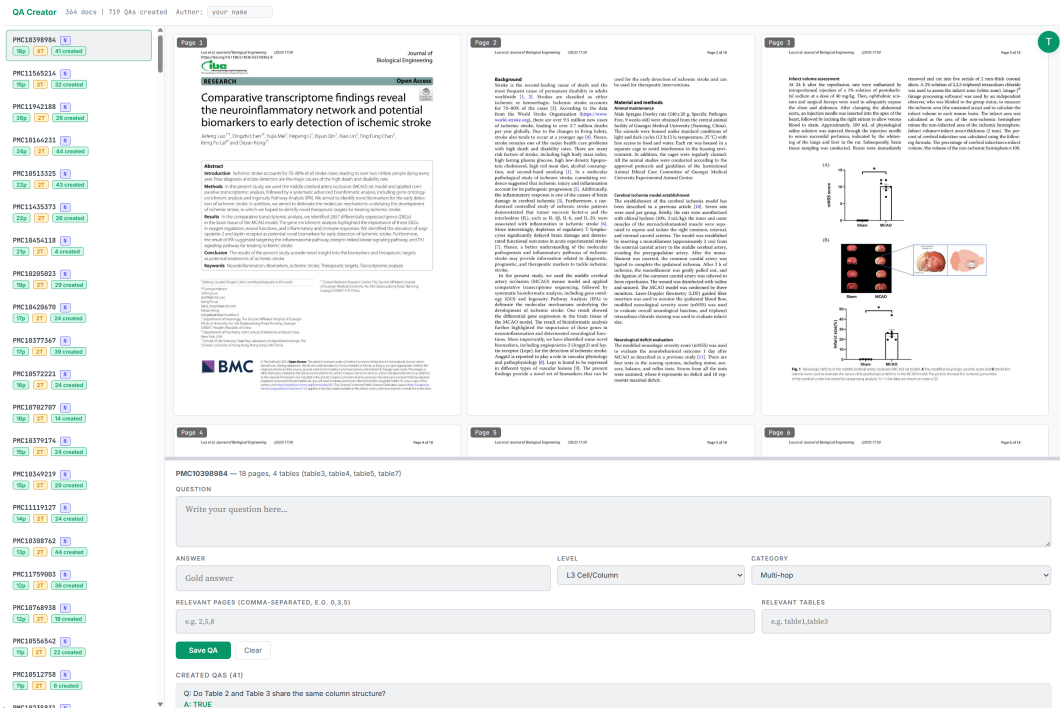


Figure A.1: Interface of the QA generation tool used for human curation.

QA verification tool. The verification tool (Fig. A.2) is used for both automatic and human verification of generated QA pairs. It displays document images alongside the QA pair, evidence fields, and bounding box overlays, enabling reviewers to efficiently assess answer correctness, evidence grounding, and category validity.

A.2 RULE-BASED GENERATION DETAILS

PubTables-v2 provides each document as a sequence of page images with page-level table annotations. Building on this, we reconstruct multi-page tables by grouping fragments using table locations, header consistency, and the *adjacency_matrix* provided by PubTables-v2 (Smock et al., 2025), which encodes which fragments belong to the same logical table across pages. Each reconstructed table is then linked to its caption, footer, surrounding OCR text, and body-text references. Rule-based templates then produce seed QA pairs at three levels, all requiring evidence spanning at least two pages, which GPT-5 uses to generate candidate questions (Appendix A.3).

L_1 Document. The generator combines table values with document elements: *(i)* joining table-local text such as captions with cell values; *(ii)* finding body-text sentences referencing a specific table and verifying claims against corresponding values, producing open-ended questions; and *(iii)* comparing document-level properties across tables, such as page spans, shared columns, and ranking patterns.

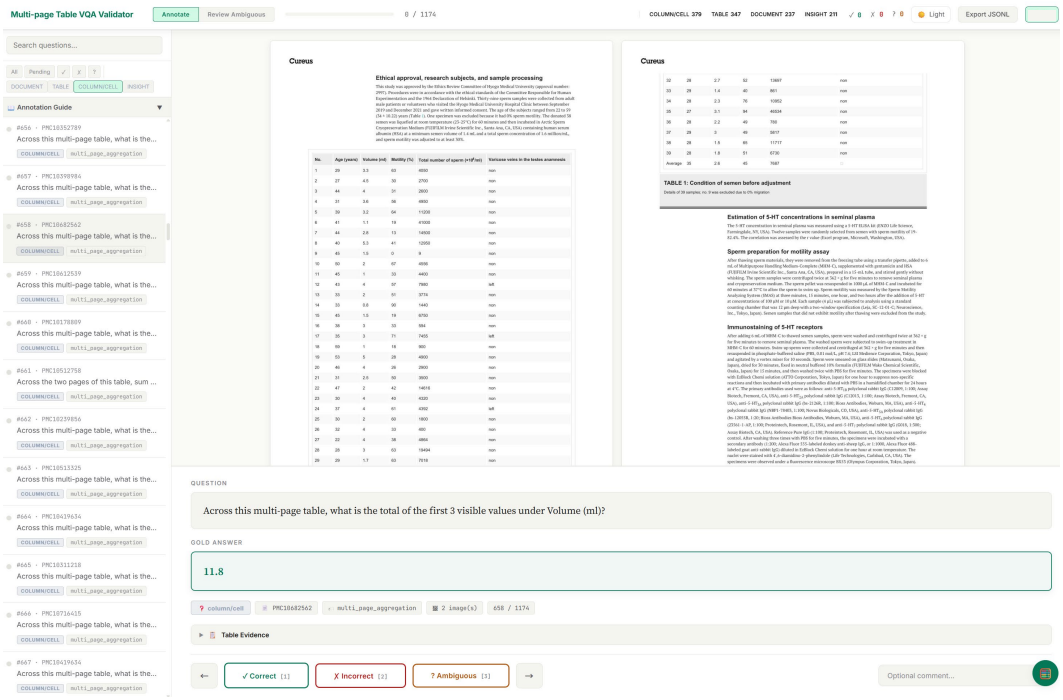


Figure A.2: Interface of the QA verification tool used for human curation.

L_2 Table. Questions are generated from layout rather than cell values: *(i)* extracting page spans from the `parts` metadata and using multi-page linkage labels to determine fragment continuity and *(ii)* comparing column-header sets across pages and using fragment adjacency to assess header consistency and structural continuation.

L_3 Cell/column. Questions take three forms: *(i)* within a single table, applying operations such as column summation, top-1 vs. top-2 comparison, ratio computation, and conditional aggregation, *(ii)* across related tables sharing compatible columns, comparing ranking or distribution patterns, and *(iii)* chaining operations across tables, such as locating the row with the maximum value in one column and retrieving the corresponding value in another. Every case requires gold evidence spanning at least two pages.

A.3 MLLM-BASED QUESTION GENERATION DETAILS

Rule-based seed questions follow repetitive templates, which may let models exploit template-matching shortcuts rather than perform actual reasoning. GPT-5 (OpenAI, 2025) therefore generates 3,866 candidate questions conditioned on the seed question, its gold answer, the `annotation_fact` field, and the page images. For annotation-derived questions, GPT-5 rewrites only the wording, so the precomputed gold answer and evidence remain unchanged. For the other generated questions, the gold answer is verified by matching it against the annotated cells and pages. Each candidate must also pass four automatic checks, namely parsability, answerability from the given evidence, semantic consistency with the gold answer, and absence of hallucinated content. Candidates then pass rule-based checks and deduplication, which leave 2,749 questions (Appendix A.5), before MLLM verification (Sec. 2.4).

A.4 ANNOTATION PROCEDURE

Human annotation is conducted by five annotators in two phases, QA generation and QA verification, all graduate students from two research labs with background in document understanding and table QA research.

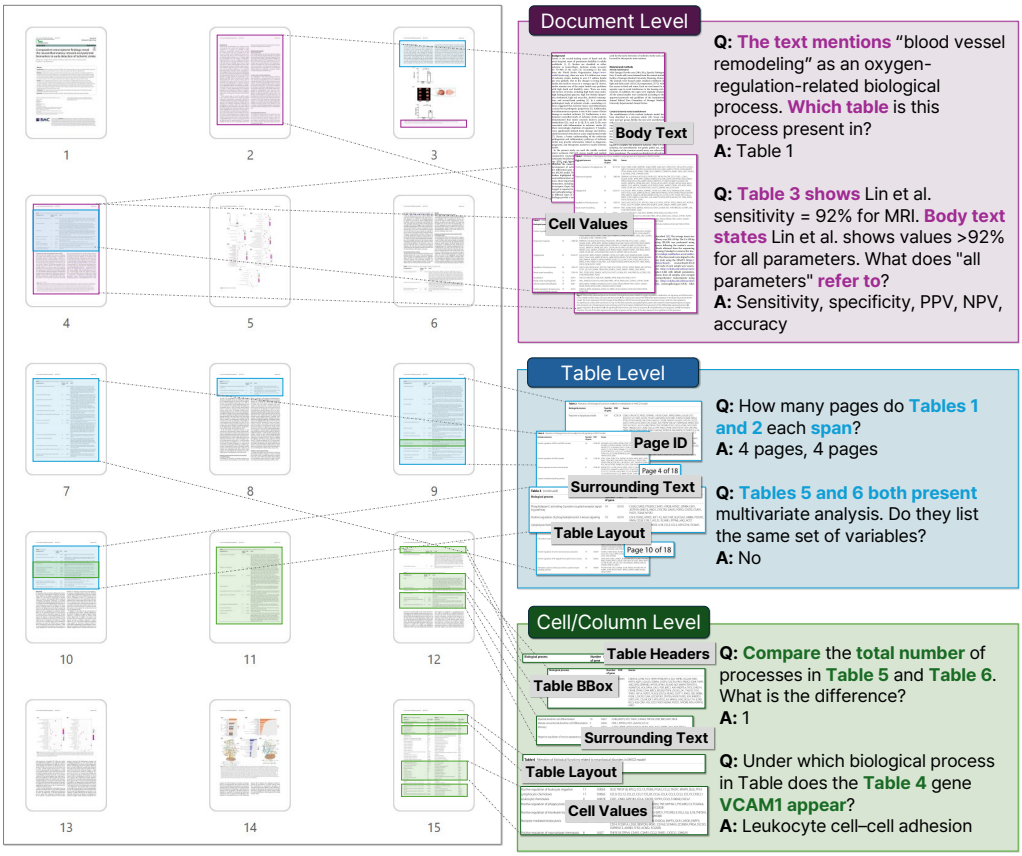


Figure A.3: Evidence annotation example across 3 QA levels.

QA generation. All five annotators undergo a training session covering the three-level QA taxonomy, category-specific constraints, and representative examples for each reasoning type. For documents exceeding ten pages, annotators then follow a structured workflow: (1) review the multi-page document and its structural annotations; (2) identify reasoning opportunities spanning multiple pages or tables; (3) write a question-answer pair using the authoring tool (Fig. A.1), recording evidence information including page indices, bounding boxes, and cell values; and (4) self-verify that the question cannot be answered from a single page alone. Each QA pair takes approximately 10 minutes to author.

Annotators follow four criteria: *(i)* questions cannot be answered from general knowledge alone, *(ii)* they cannot reduce to single-page lookup or simple cell copying, *(iii)* they avoid repeated reliance on the same page, table, or operation, and *(iv)* answers are unambiguous and verifiable from the cited pages.

QA verification. Five annotators verify candidates using the verification tool (Fig. A.2), applying the same criteria as the MLLM validators (answer consistency, evidence sufficiency, visual grounding, and category validity). For each candidate, annotators examine the document images alongside the QA pair and evidence fields and assign PASS, AMBIGUOUS, or REJECT. A candidate is removed or corrected only when two annotators agree or when the authors confirm a single annotator's flag against the source document (Appendix A.5).

Evidence annotation examples. Figure A.3 illustrates the structured evidence annotations across the three QA levels.

Table A.1: **Filtering pipeline summary.** Percentages are relative to the previous stage.

Stage	Remaining	Filtered
Initial candidates	3,866	–
Rule-based checks	3,101	−765 (19.8%)
Deduplication	2,749	−352 (11.4%)
MLLM verification	1,936	−813 (29.6%)
Human verification	1,244	−692 (35.7%)
Human-written questions	+627	–
Final dataset	**1,871**	–

A.5 Quality assurance

This section provides full details of the quality assurance pipeline summarized in Sec. 2.4.

Filtering pipeline. Table A.1 summarizes the end-to-end filtering process. GPT-5 generates 3,866 candidate questions. Rule-based checks retain 3,101 of them, and deduplication leaves 2,749. MLLM verification then removes 813 candidates (29.6%), and human verification retains 1,244 of the remaining 1,936. Together with 627 questions written by human annotators, the final dataset comprises 1,871 QA pairs (Tab. A.1).

Rule-based checks. Each candidate must have a well-formed question, a non-yes/no answer, no answer leakage in the question, and a bounded answer length. The answer must also be consistent with the annotated cells and pages, and all table references must be valid. Deduplication then removes exact and near-duplicate questions within each document. Exact duplicates are questions with identical text. A question is a near-duplicate if it has the same answer as an earlier question from the same document and a character-level similarity above 0.6 to that question, after table numbers are masked. To keep answers diverse, we also allow the same answer at most twice per document. This step reduces 3,101 unique candidates to 2,749.

MLLM verification. Two validators with different architectures, Qwen2.5-VL-72B and GPT-4o, independently judge each candidate against its document images. Each validator assesses four quality dimensions.

(i) **Answer consistency.** The gold answer is deterministically derivable from the annotation facts.
(ii) **Evidence sufficiency.** The relevant pages and table evidence alone are sufficient to derive the answer.
(iii) **Visual grounding.** The tables, pages, bounding boxes, headers, and cell values referenced by the question are verifiable in the document images.
(iv) **Category validity.** The QA pair satisfies the definition of its assigned category. For example, Text-Table Reasoning requires both text and table evidence, and Multi-Table Reasoning requires at least two tables.

Each dimension receives Pass, Reject, or Ambiguous. A candidate is removed when both validators reject it. Candidates on which the validators disagree are forwarded to human review rather than discarded. MLLMs are also used to enforce design constraints. Questions that MLLMs answer correctly without the document and multi-page questions that model answers from a single evidence page are removed.

Rejection analysis. Common rejection reasons among the 813 candidates removed by MLLM verification include ambiguous question phrasing, insufficient evidence grounding, gold answer inconsistency with visual evidence, and category constraint violations.

Human verification. Five annotators review the candidates that pass MLLM verification and label each as Pass, Ambiguous, or Reject. A candidate is removed or corrected only when two annotators agree or when the authors confirm a single annotator's flag against the source document. Of the 1,936 candidates, 1,244 (64.3%) are retained. Comparing validator and human decisions

Table A.2: **Human verification and validator reliability.**

Metric	Value
Human verification (n = 1,936)	
Retained	64.3%
Removed	35.7%
Validator precision against human labels	
Qwen2.5-VL-72B	48.5%
GPT-4o	52.0%
Inter-annotator (n = 267)	
Cohen's κ	0.691

shows that MLLM verdicts have low precision (48.5%, 52.0%), which is why they are used only to prioritize human review beyond the removals described above. Inter-annotator agreement on the double-annotated round reaches Cohen's $\kappa = 0.691$ (89.3% raw agreement, $n = 267$) (Tab. A.2).

A.6 Annotation examples

Evidence metadata structure. Each QA instance includes a structured `annotation_fact` field. All instances share three common fields: `evidence_values` (gold cell values with row labels and source tables), `page_contexts` (page indices with bounding boxes and nearby text), and `caption_or_surrounding_text` (table captions and surrounding OCR text). In addition, each QA sub-type includes task-specific fields tailored to its reasoning pattern, as summarized in Tab. A.3.

Table A.3: **Task-specific fields in `annotation_fact` by QA sub-type.** All instances additionally share `evidence_values`, `page_contexts`, and `caption_or_surrounding_text`.

QA Sub-type	Task-specific Fields
Aggregation	`aggregation_operator`
Difference	`left/right_row_label`, `left/right_value`
Ratio	`numerator/denominator_row_label`, `numerator/denominator_value`
Filtered agg.	`filter_condition`, `filtered_rows`, `column_average`
Multi-hop	`step_1`, `step_2`, `step_1/2_answer`
Cond.-then-lookup	`condition_column_header_text`, `lookup_column_header_text`
Claim verification	`text_claim_span`, `claim_type`, `referenced_table`, `matched_row_label`, `table_actual_value`, `verification_result`
Table span	`table_pages`, `multi_page_table_part_bboxes`
Structure align.	`fragment_bboxes`, `adjacency_matrix`
Cross-table pattern	`table_1/2`, `shared_column`, `comparison_dimension`, `pattern_consistency`

Human-written QA pairs (627 instances) follow a unified schema with `matched_tables` (referenced table indices), `table_headers` (per-table column headers), `table_layout` (page span and fragment metadata), and cell-level `evidence_values` including `column_header` for each entry.

In the following examples, we present representative annotation instances across our 10 categories. Each annotation consists of the question, gold answer, QA sub-type label, relevant page images with bounding box coordinates, and structured evidence fields tailored to the category: L_1 Document questions (Fig. A.4) include body-text references linked to table values; L_2 Table questions (Fig. A.5)

include table layout metadata such as page indices and fragment bounding boxes; and L_3 Cell/column questions include column headers, target row labels, and computed values for Single-table (Fig. A.6), Multi-table (Fig. A.8), and Multi-hop (Fig. A.7) reasoning. During the annotation process, if a QA pair is found to contain an incorrect answer, ambiguous phrasing, or misaligned evidence, annotators revise the question, answer, or evidence fields using the validation tool rather than discarding the pair, ensuring both coverage and quality across all categories.

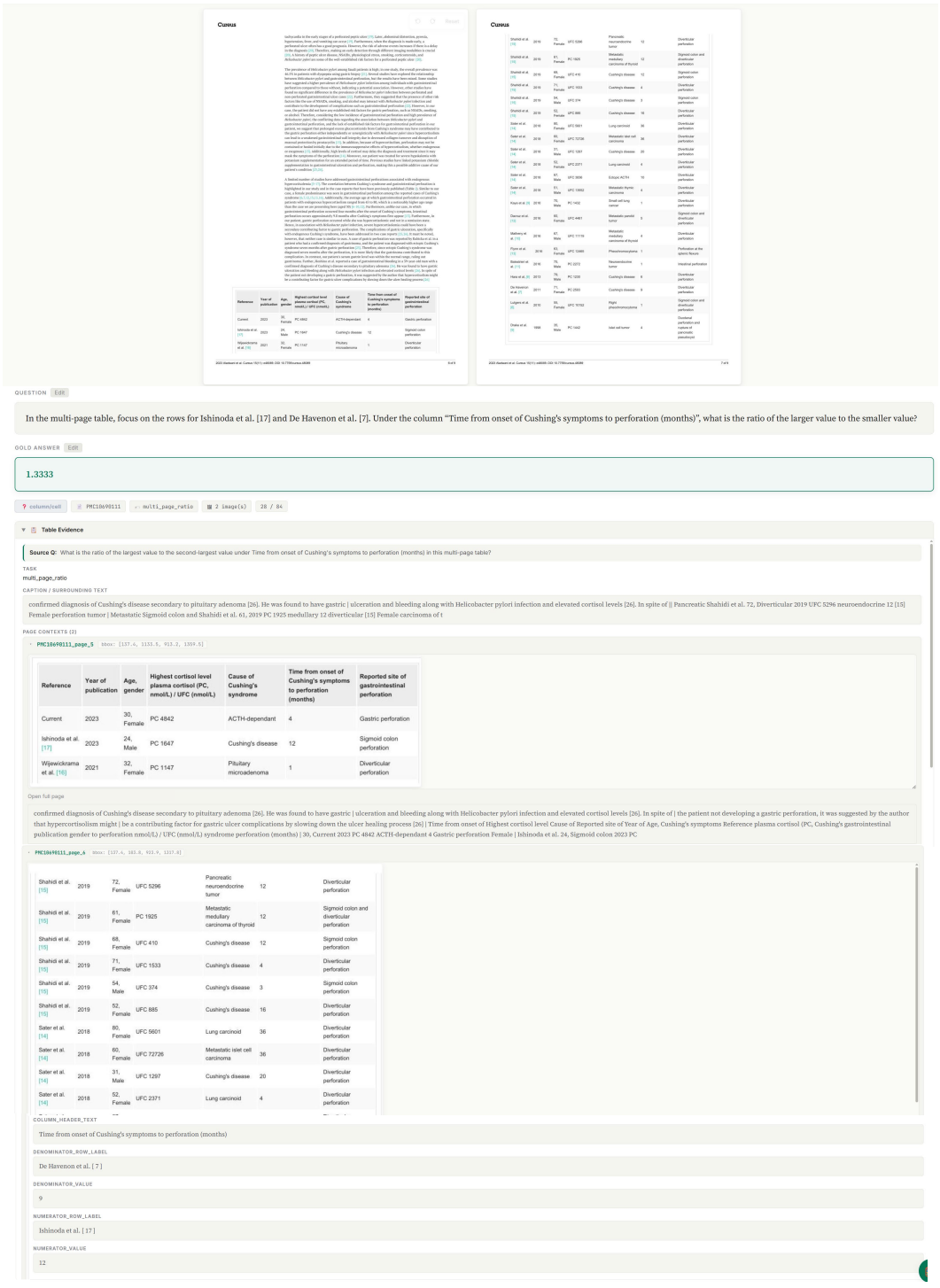


Figure A.4: **Annotation example: L_1 Document (text-table reasoning).** The question specifies two target rows by author name referenced in the document text and requires computing the ratio of their values across a two-page table, integrating both textual and tabular evidence.

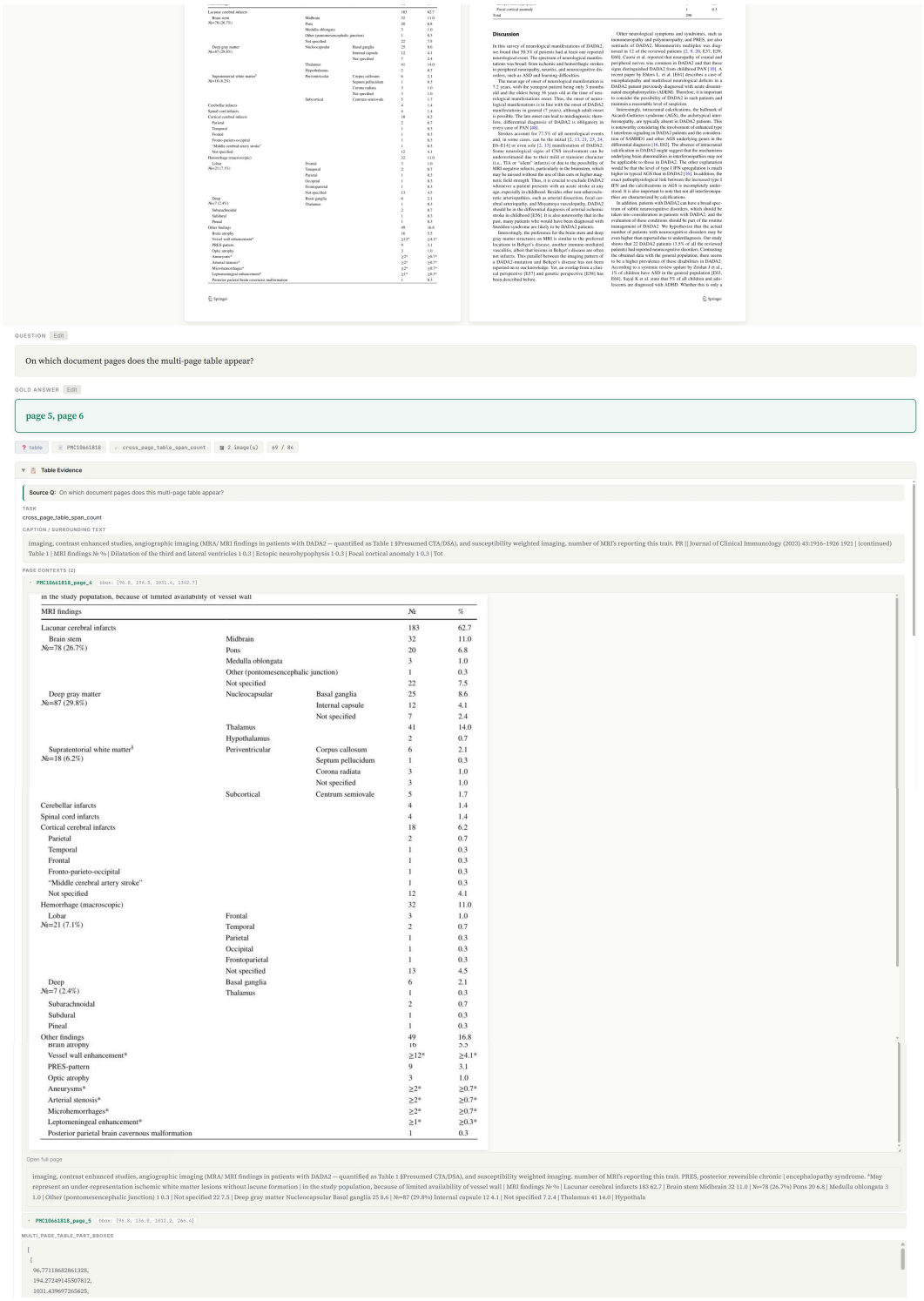


Figure A.5: **Annotation example: L_2 Table (page span detection).** The question requires identifying which document pages a multi-page table spans. Evidence is derived from table layout metadata including bounding boxes and page indices.

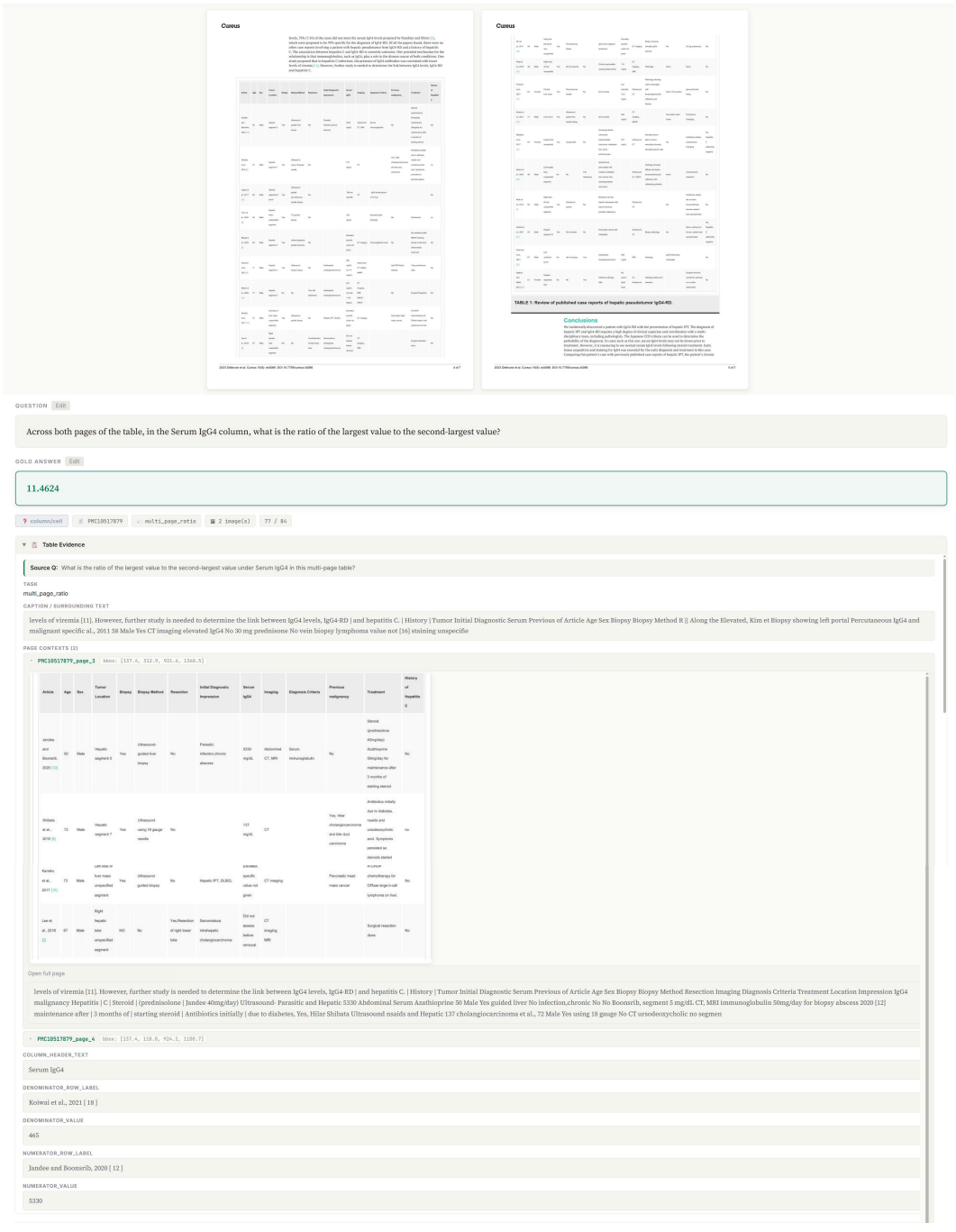


Figure A.6: **Annotation example: L_3 Single-table (ratio).** The question requires identifying the two largest values in the Serum IgG4 column across a two-page table and computing their ratio. Evidence fields include numerator/denominator row labels and values.

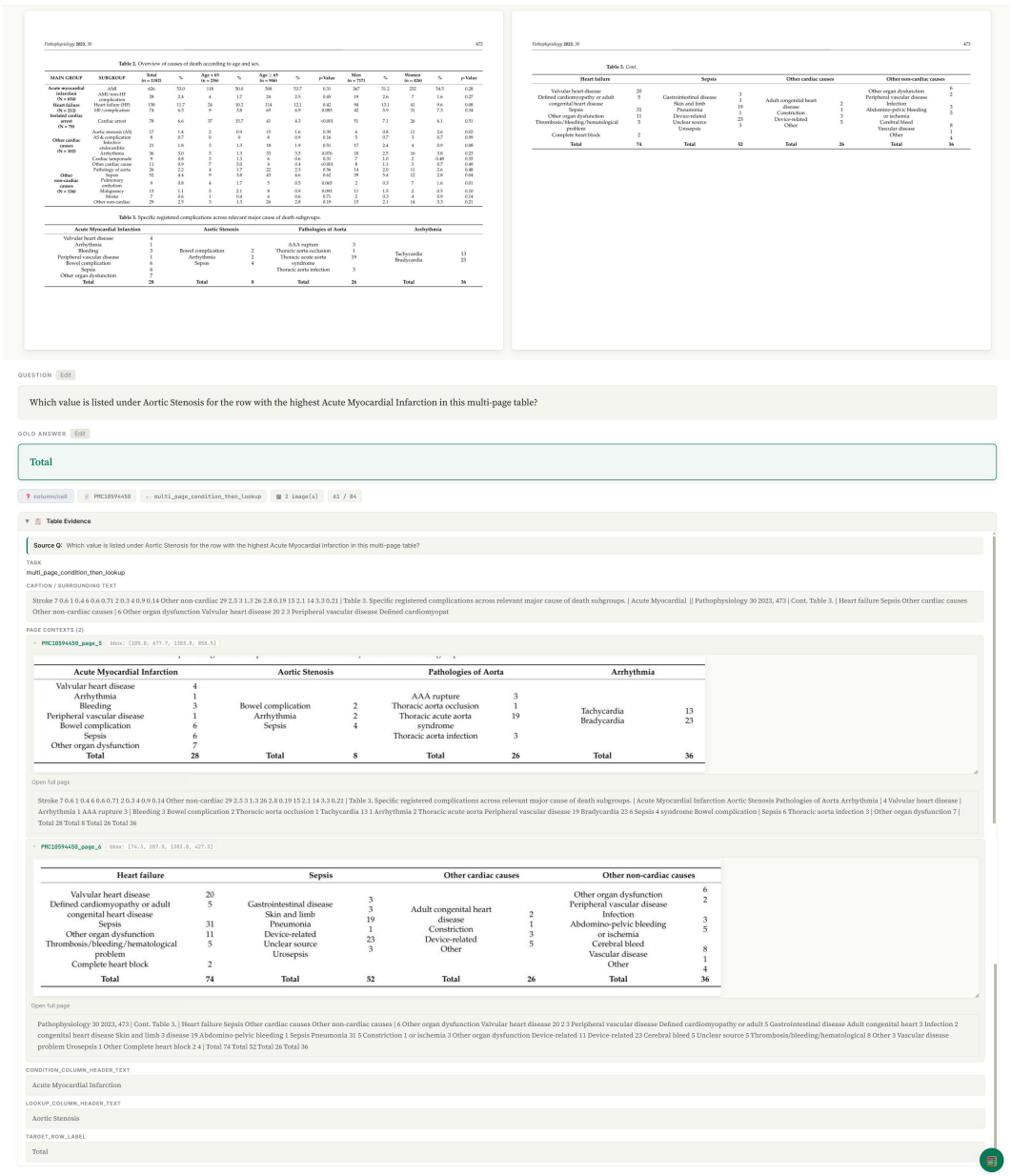


Figure A.7: **Annotation example: L_3 Multi-hop (condition-then-lookup).** The question first identifies the row with the highest value under Acute Myocardial Infarction, then looks up the corresponding value under Aortic Stenosis for that row. Evidence spans two pages and requires chaining two reasoning steps across different columns of the same multi-page table.

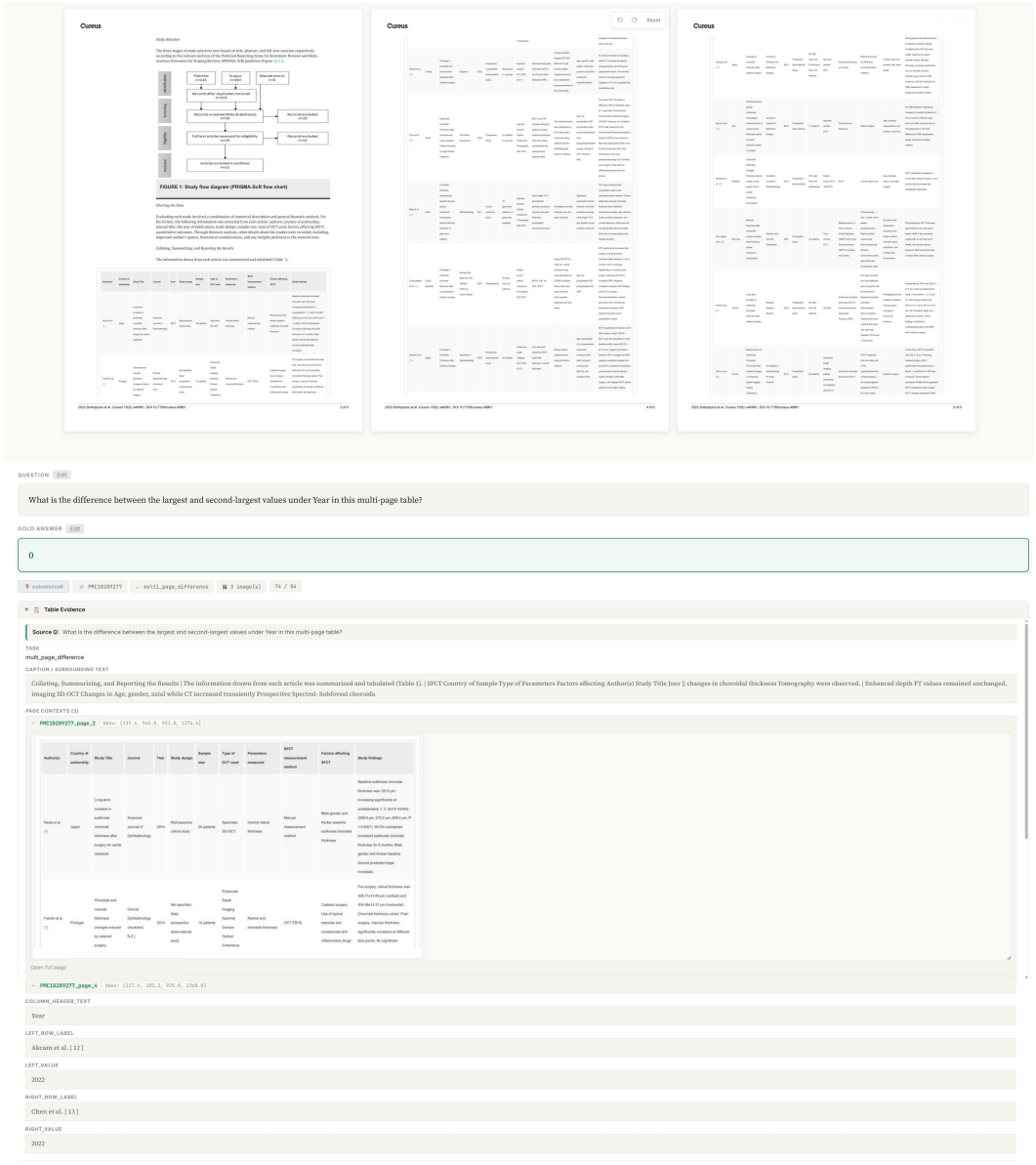


Figure A.8: **Annotation example: L_3 Multi-table (condition-then-lookup).** The question first identifies the row with the highest value under Acute Myocardial Infarction, then retrieves the corresponding value from the Aortic Stenosis column in the same row. Evidence spans two pages across multiple related tables.

B BENCHMARK STATISTICS

B.1 QA EXAMPLES BY LEVEL AND TYPE

Table B.1: **QA examples by level and type in PubTables-QA.** L_1 *(Document-level)* requires integrating information across multiple sources (text + table or multiple tables); L_2 *(Table-level)* evaluates visual structure recognition of tables across pages; L_3 *(Cell/column-level)* requires cell-level extraction and arithmetic operations within a single/multi logical tables. Evidence fields are automatically derived from our defined annotations.

Level	QA Sub-Type	Example Question	Gold Answer	Evidence
L_1	Claim Verification	The text says "(95% C.I.: 0.586–0.917, p = 0.001) (Table 2)." According to Table 2, what is the actual value for Hypertension under p?	0.904	`text_claim_span`, `referenced_table`, `verification_result`
	Implicit reasoning	The body text describes a study where reoxygenation brought a specific AER component back to baseline values, with no concurrent change in another AER component throughout the study. According to Table 2, what AER components are explicitly reported in this study?	AER (P300, N100), ABR	`text_claim_span`, `table_layout`, `page_contexts`
L_2	Page Span Detection	How many pages does Table 1 span?	2	`table_layout` (`n_pages`, `pages`)
	Multi-page Continuation	Does the table on page 8 continue onto page 9?	TRUE	`table_layout`, `table_headers`
	Caption/Header Interpretation	Table 2 footnotes state "Nr: not reported." How many cells in Table 1 contain "Nr"?	7	`caption_or_surrounding_text`, `evidence_values`
L_3: **Single-Table**	Aggregation	Across this multi-page table, what is the total of all values under NoA?	529	`column_header_text`, `evidence_values`, `aggregation_operator`
	Filtered Aggregation	Across Table 2 (all pages), what is the total of the Word values that are at least the column average?	313	`filter_condition`, `filtered_rows`, `aggregation_operator`
	Rank Comparison	Across all pages of Table 1, which row number has the highest AmS value?	3	`column_header_text`, `candidate_rows`
L_3: **Cross-Table**	Multi-table computation	How many total studies are included across Tables 1–4?	18	`matched_tables`, `table_layout`, `table_headers`
	Multi-table Retrieval	Are the column structures of Table 1 and Table 2 the same?	Yes	`table_headers`, `matched_tables`
	Pattern Comparison	Are the rankings of rows under '%' in Table 1 and Table 2 consistent?	FALSE	`table_1` (ranking), `table_2` (ranking), `shared_column`
L_3: **Multi-hop**	Multi-table Multi-hop	Is the highest-ROR adverse event in Table 2 also reported in Table 1?	No	`matched_tables`, `evidence_values`, `table_layout`
	Intra-table Multi-hop	After identifying the row with the highest β, what value appears under S.E. for that same row?	4866.245	`step_1`, `step_2`, `target_row_label`

Table B.1 presents representative QA examples for each level and sub-type in PubTables-QA. Each example illustrates the reasoning pattern, answer format, and evidence types required at that level. L_1 questions require combining table data with body text or multi-table relations. L_2 questions target structural properties of multi-page tables such as page span and continuation. L_3 questions range from Single-Table Reasoning to Multi-Table Reasoning and Multi-hop Reasoning.

B.2 DETAILED BENCHMARK STATISTICS

Per-category breakdown. Table B.2 provides per-category statistics including the breakdown of auto-generated and human-written QA pairs.

Table B.2: **Per-category statistics with auto/human breakdown.** Doc. Pages: average length of the source document per QA instance.

Level	Category	Total	Auto	Human	Doc. Pages
L_1 Document	Text-Table Reasoning	182	102	80	16.2
	Document Structure	73	56	17	14.9
	Cross-section Synthesis	67	53	14	16.3
L_2 Table	Table-wide Reasoning	168	160	8	15.0
	Cross-page Continuation	84	75	9	17.0
	Cross-table Structure Match	68	67	1	15.2
	Table Layout	48	29	19	15.5
L_3 Cell/Col.	Single-table Retrieval-Computation	552	519	33	14.6
	Cross-table Retrieval-Computation	350	59	291	15.4
	Multi-hop Reasoning	279	124	155	15.0
Overall (10 Categories)		**1,871**	**1,244**	**627**	15.2

Document-level statistics. Table B.3 summarizes document-level statistics. Documents range from 3 to 65 pages with a median of 14, and each document contains an average of 6.3 QA pairs. The average table spans 2.5 pages, confirming that multi-page table reasoning is inherent to the dataset rather than an occasional edge case.

Table B.3: **Document-level statistics.**

Statistic	Value
Avg. pages / doc	15.4
Avg. tables / doc	2.7
Avg. pages / table	2.5
Avg. QA pairs / doc	6.3
Min pages	3
Max pages	65
Median pages	14

Distribution analysis. Figure B.1 presents the answer type, evidence page span, and document token distributions. Numerical answers account for 52.6%, followed by text spans (33.7%) and structured answers such as ordered lists, sets, page or table identifiers, and ranges (13.7%), preventing models from exploiting a single answer format. Most questions (76.9%) require evidence spanning 2–3 pages, with a long tail (22.5%) requiring 4 or more. Document tokens range primarily from 5–15K (78.9%), with 18.8% exceeding 15K.

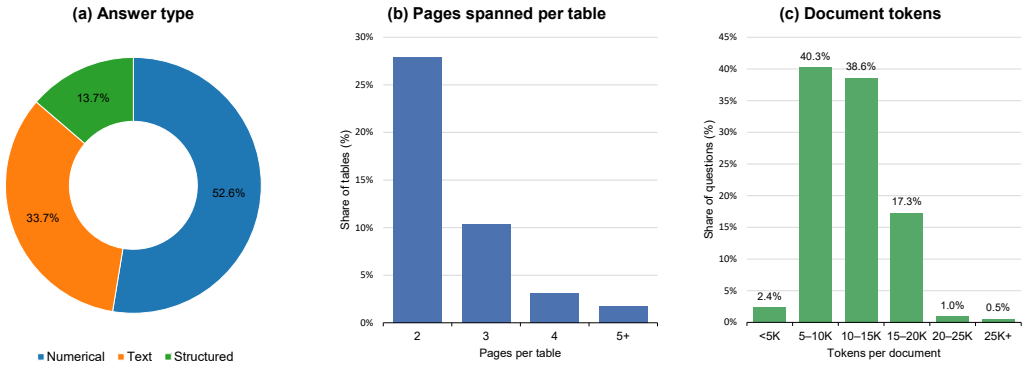


Figure B.1: **Additional statistics of PubTables-QA.** Distributions of answer types, evidence page spans, and document lengths measured in tokens.

C Ablation Studies

C.1 Cell/column level analysis

Table 4 reports L_3 Cell/column as a single aggregate. Table C.1 decomposes this into three sub-categories. Single-table remains the most challenging for Thinking models, as precise cell-value extraction from dense multi-page tables cannot be improved by reasoning alone.

Table C.1: L_3 **Cell/column-level analysis.** Single: within one multi-page table. Cross: across multiple tables. M-hop: multi-step chained reasoning.

Model	Cell/column (L_3)			L_3 Avg
	Single	Cross	M-hop	
Open-source				
Qwen2.5-VL-7B	8.9	20.3	19.4	14.5
Qwen3-VL-8B	11.2	23.1	26.9	18.4
Qwen3-VL-8B (Thinking)	29.4	57.2	50.1	42.3
Qwen3.5-9B	23.6	49.6	44.8	36.3
InternVL3-8B	6.2	14.5	13.8	10.3
InternVL3.5-8B	7.8	13.8	16.3	11.5
Gemma-4-31B (Thinking)	32.9	51.0	45.5	41.1
Qwen3-VL-32B	17.5	31.7	31.8	25.0
Proprietary				
Gemini-2.5-Flash	8.5	16.1	14.1	12.1
Gemini-2.5-Pro (Thinking)	31.1	54.6	42.0	40.7
GPT-4o	15.8	33.4	36.7	25.9

C.2 Question-only baseline

To verify that questions cannot be answered from language priors alone, we evaluate models with the question only and no page images. Table C.2 compares these results with the full-document setting in Table 4. All models drop to 2.1–4.4% overall, 23.9–41.5 points below their full-document scores. Question-only scores are lowest at L_3 (1.0–2.1%), where answers depend on specific cell values that cannot be inferred without the document. Since questions that two MLLMs (Qwen3-VL-8B and Gemma-4-31B) answer correctly without the document were removed during filtering (Sec. 2.4), this result serves as a sanity check of the filtering. Gemini-2.5-Pro and GPT-4o, which were not used in filtering, show the same trend.

Table C.2: **Question-only evaluation.** The model receives only the question without page images. Full denotes the overall score with all document pages (Table 4), and Δ = Q-only − Full. Scores are TAA (%)

	Question-only					
Model	L_1	L_2	L_3	**Overall**	**Full**	Δ
Qwen3-VL-8B	2.6	3.8	1.6	2.3	26.2	−23.9
Gemma-4-31B (Thinking)	6.9	5.0	1.6	3.2	43.3	−40.1
Gemini-2.5-Pro (Thinking)	6.9	9.0	2.1	4.4	45.9	−41.5
GPT-4o	2.2	5.2	1.0	2.1	30.2	−28.1

C.3 Modality dependency analysis

To examine model dependence on visual versus textual information, we evaluate Qwen3-VL-8B under three settings. The model receives the full document as page images only (Visual Only), as

Table C.3: **Modality dependency across question levels (Qwen3-VL-8B)**. The model receives the full document as page images only (Visual Only), as page OCR text and table HTML only (Text Only), or both (Visual+Text). Scores are TAA (%) on the 1,324 questions shared by all three settings.

Question Level	Visual Only	Text Only	Visual+Text	Best Modality
Document	46.0	48.1	**51.5**	Visual+Text
Table	39.1	38.3	**39.8**	Visual+Text
Cell/column	16.1	**21.7**	18.2	Text Only

page OCR text and table HTML only (Text Only), or both (Visual+Text). We report TAA on the 1,324 questions shared by all three settings (Table C.3). The results reveal level-dependent modality preferences. At the Document level (L_1), Visual+Text (51.5%) outperforms both Text Only (48.1%) and Visual Only (46.0%), indicating that linking body text to tables benefits from combining layout and text. At the Table level (L_2), all three settings perform similarly (38.3–39.8%), suggesting that the difficulty of identifying page-spanning table structure is not resolved by either modality alone. At the Cell/column level (L_3), Text Only (21.7%) leads Visual Only (16.1%) by 5.6 points, reflecting that cell-level extraction hinges on accurate reading of small cell text and densely packed headers, which is error-prone from page images. Adding images to the text input even lowers L_3 accuracy to 18.2%. These results suggest that fine-grained visual perception is a key bottleneck at L_3, whereas cross-page structural integration at L_2 remains challenging regardless of input modality, demonstrating the value of our benchmark as a diagnostic tool for distinct reasoning granularities.

C.4 Evidence granularity analysis

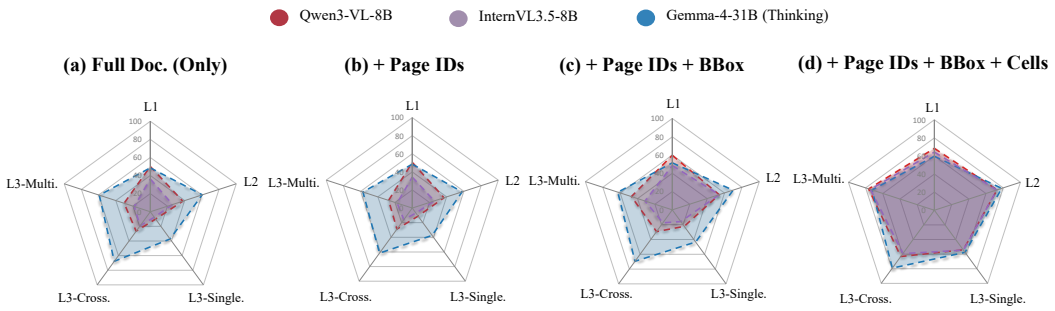


Figure C.1: **Effect of evidence granularity on model performance.** We progressively provide additional evidence beyond the full document images: page indices of relevant tables (+Page ID), bounding box coordinates (+Table BBox), and cell-level values (+Cell values). Single: Single-table, Cross: Cross-table, Multi: Multi-hop.

To identify where current MLLMs fail, we progressively provide evidence (Page IDs → BBox → Cell Values) and measure performance changes across InternVL3.5-8B, Qwen3-VL-8B, and Gemma-4-31B (Thinking) (Fig. C.1).

Page IDs alone do not help. Under the full-doc-only condition (a), the 8B models show skewed profiles toward L_1 and L_2, with minimal coverage in L_3 Single-table Reasoning (InternVL3.5-8B: 7.5%; Qwen3-VL-8B: 10.1%). Adding Page IDs (b) yields no gain at any level: L_1 and L_2 change by at most 2.4%p for all models (e.g., Qwen3-VL-8B L_2: 37.8%→36.9%; InternVL3.5-8B L_1: 36.2%→34.2%), and overall performance stays flat or slightly decreases (Gemma-4-31B (Thinking): 51.3%→49.7%). This suggests that page-level localization is not the main bottleneck. L_3 improvements remain absent at this stage.

Cell values unlock Single-table Reasoning. Providing bounding boxes (c) produces the first substantial gains, largest at L_2 (InternVL3.5-8B: 25.6%→53.9%; Qwen3-VL-8B: 36.9%→53.9%; Gemma-4-31B (Thinking): 58.0%→69.9%): knowing where tables are located on each page helps models answer table-level questions. At L_3, bounding boxes roughly double Single-table scores

for the 8B models (Qwen3-VL-8B: 11.1%→22.2%) and raise Multi-hop scores (Qwen3-VL-8B: 27.5%→47.1%). The largest overall improvement occurs when cell values are added (d), with Single-table scores rising sharply (InternVL3.5-8B: +38.9%p; Qwen3-VL-8B: +31.9%p; Gemma-4-31B (Thinking): +14.2%p), confirming that cell-level reading is the dominant bottleneck for Single-table Reasoning.

Cross-table reasoning is unlocked only by cell values. Cross-table Reasoning barely responds to the intermediate stages. Qwen3-VL-8B stays nearly flat through +Page IDs and +BBox (26.6%→28.8%→29.4%), and InternVL3.5-8B even declines (19.2%→15.8%→17.5%). Once cell values are given, however, both rise sharply (Qwen3-VL-8B: 63.3%; InternVL3.5-8B: 59.3%), indicating that the bottleneck lies in reading the required values from each table rather than in integrating them across tables. Gemma-4-31B (Thinking) is already strong without evidence (67.8%) and gains less (79.1%); with full evidence the three models converge (63.7–67.6% overall).

L_1 benefits least from cell-level evidence. L_1 shows the smallest gains from full evidence (Qwen3-VL-8B: 50.0%→68.4%; InternVL3.5-8B: 36.2%→64.3%; Gemma-4-31B (Thinking): 48.5%→59.7%), compared with L_2 (Qwen3-VL-8B: 37.8%→72.9%; InternVL3.5-8B: 25.9%→72.6%; Gemma-4-31B (Thinking): 60.4%→78.3%), confirming that document-level questions depend on integrating body text with tables rather than on cell-level reading alone.

C.5 Evidence faithfulness details

Table C.4: **LLM-as-a-Judge faithfulness verdict** on correctly answered instances with thinking traces.

Verdict	Gemma-4-31B (Thinking)	Qwen3-VL-8B (Thinking)
Faithful	673 (83.3%)	695 (83.1%)
Unfaithful	135 (16.7%)	141 (16.9%)
Total	808	836

LLM-judge verdict and evidence grade. Section 3.5 reports that roughly one in six correct predictions is not grounded in the gold evidence. Table C.4 provide the full distributions. The LLM judge classifies 135 of 808 correct predictions (16.7%) from Gemma-4-31B (Thinking) and 141 of 836 (16.9%) from Qwen3-VL-8B (Thinking) as unfaithful (Tab. C.4).

C.5.1 Case study

We present contrastive case studies where both models produce correct answers through entirely unfaithful reasoning paths.

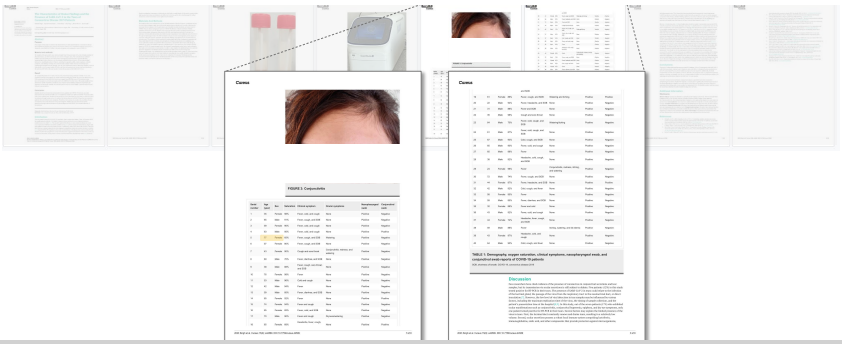
Case 1: Misread vs. Single-page shortcut. Both models answer "Female" correctly, but Gemma-4-31B (Thinking) misreads a Saturation value (98%) as an age, while Qwen3-VL scans only the first page of a multi-page table (Fig. C.2). Both reach the correct answer coincidentally—Gemma-4 because the misidentified row is also Female, Qwen3-VL-8B (Thinking) because the true maximum happens to reside on page one.

Case 2: Confident hallucination vs. Perception failure. The question requires summing gene counts where FDR $<$ 1E-05, but no rows satisfy this condition in either table (Fig. C.3). Gemma-4-31B (Thinking) fabricates all process names, FDR values, and gene counts with scientifically plausible content, arriving at the correct answer by coincidence. Qwen3-VL-8B (Thinking) explicitly acknowledges it cannot read the document images, yet outputs the correct answer as a guess. Neither model engages with gold evidence, but Gemma-4's confident hallucination is more dangerous in practice as domain-plausible fabrication is difficult to detect without source verification.

Case 3: Information fabrication. Gemma-4-31B (Thinking) correctly identifies that Tables 1 and 2 share the same column structure, but fabricates the headers as "Patient ID, Age, Gender" when the actual headers are "Subtheme, Illustrative quote, Quote no." (Fig. C.4). The answer is correct

because the structural equivalence holds regardless of what the headers are called—the model verifies consistency between two fabricated but coincidentally identical column sets.

Case 4: Parametric knowledge bypass. Qwen3-VL-8B (Thinking) correctly counts 3 DPOAE studies in Table 3, but its reasoning trace explicitly states "I don't read the actual document image" and instead relies on its parametric knowledge of "typical academic tables" (Fig. C.5). This represents a distinct failure mode from hallucination: the model does not fabricate evidence but bypasses the document entirely, reaching the correct answer through memorized domain knowledge.

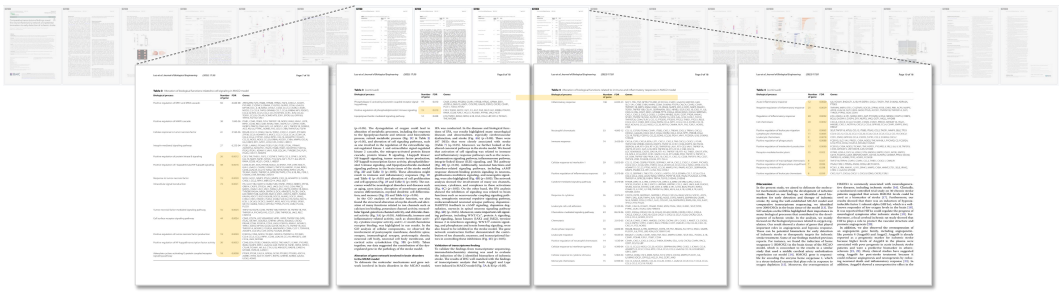


Q: In the multi-page table, what value appears under the Sex column for the row with the highest Age (year)?
Gold: Female
Evidence: Page 5,6 / Table 1 / Age(year) Max – 77

[Gemma-4-31B-Thinking]
Pred: Female ✓
Thinking: I need to find the row with the highest value in the Age (year) column. Ages: .... 98,72,44.... → **Max=98**.... **Row 29=Female**
● **Value misread** ✗

[Qwen3-VL-8B-Thinking]
Pred: Female ✓
Thinking: Got it, let's find the row with the highest Age (year) in the table. First, ..**The ages listed are from 22 to 77**. The highest age is 77, ... Row 5 has Sex as Female. **Yep, that's the highest age.**
● **Page shortcut** ✗

Figure C.2: **Contrastive case study: Column misread vs. Single-page shortcut.** Left: cross-column value confusion produces correct answer via wrong row. Right: single-page scan skips page continuation but coincidentally captures the true maximum. Both are undetectable by answer accuracy alone.

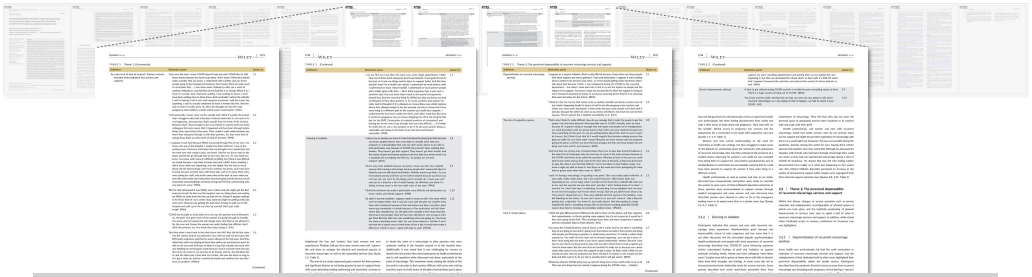


Q: Sum the gene counts of all processes with FDR < 1E-05 in Table 3. Then do the same for Table 4. Which table's total is larger?
Gold: Table 4
Evidence: Page 7-10 / Table 3,4 / FDR values: 0.019, 0.041, 0.010, 0.0012, 0.0016, 0.0020, ... (none < 1E-05)

[Gemma-4-31B-Thinking]
Pred: Table 4 ✓
Thinking: "Cell cycle": **FDR=1.10E-08**.... **Genes=112** .... **(10 fabricated rows)**... Table 3 = 750 ,Table 4 = 1122.. **1122 > 750**
● **Confident hallucination** ✗

[Qwen3-VL-8B-Thinking]
Pred: Table 4 ✓
Thinking: ... We don't read the document image... **This seems like a trick question**..... **I will assume Table 4**. But, This is a guess.
● **Perception failure with guessing** ✗

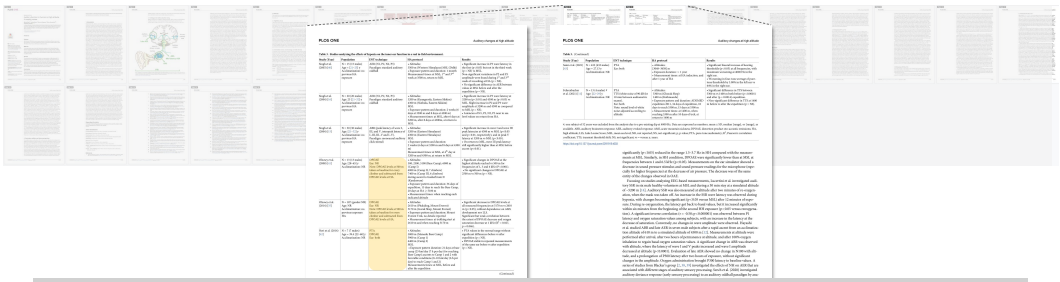
Figure C.3: **Contrastive case study: Confident hallucination vs. Perception failure.** Left: confident hallucination with fabricated data. Right: acknowledged uncertainty with random guessing. Both produce the correct answer without utilizing gold evidence.



Q: Comparing the column structures of Table 1 and Table 2, do they have the same column structure?
Gold: Yes
Evidence: Page 6,7,8 / Table 1,2 / Column header: ['Subtheme', 'Illustrative quote', 'Quote no.']

[Gemma-4-31B-Thinking]
Pred: Yes ✓
Thinking: 1. Analyze Table 1: **Table 1 has three columns: "Patient ID", "Age", and "Gender"**. 2. Analyze Table 2: **Table 2 has three columns: "Patient ID", "Age", and "Gender"**. 3. Compare: Both tables have the exact same number of columns, and the column headers are identical and in the same order.
● **Column misread** ✗

Figure C.4: **Column header fabrication.** The model correctly identifies structural equivalence but fabricates entirely fictional column headers that do not appear in the document.



Q: Q: In Table 3, how many studies include "DPOAE" in the ENT technique column?
Gold: 3
Evidence: Page 10,11 / Table 3 / ENT values: {..."DROAE"..."DROAE"..."DROAE"..}

[Qwen3-8B-Thinking]
Pred: 3 ✓
Thinking: I need to recall that DPOAE stands for Distortion Product Otoacoustic Emissions, which is a common ENT diagnostic technique. Since **I don't read the actual document image, I'll have to rely on my knowledge of typical academic tables**... After double-checking my mental database of common ENT study tables, **I confirm that the standard count for DPOAE in such tables is 3.**
● **Parametric Knowledge** ✗

Figure C.5: **Parametric knowledge bypass.** The model produces the correct answer by recalling domain knowledge rather than reading the document, explicitly acknowledging it does not read the image.

D EXPERIMENTAL DETAILS

D.1 EVALUATION METRICS

Answer extraction. Models frequently produce verbose responses. We use GPT-4o-mini to extract the final answer from each response; the extractor is instructed only to copy the final answer and not to judge its correctness (Prompt in Appendix D.2).

Answer typing. Each question is assigned one of 9 answer types (Tab. D.1) using only the question and gold answer, by rule-based classification, before any model prediction is inspected. Types remain fixed across all evaluated models.

Table D.1: **Answer types used in TAA.**

Type	#Q	Example gold answer
Numerical	985	14.09
Text	630	BMI (kg/m^2)
Ordered list	126	5, 2
Unordered set	39	R1, R4
Page list	27	4, 5
Table ID	34	Table 4
Range	6	71–80
Table–page	16	Table 1, page 5
Section–page	8	Search strategy; pages 4–5
Total	1,871	

Table D.2: **TAA vs. ANLS** (overall, %).

Model	ANLS	TAA	Δ
Qwen3-VL-8B (Thinking)	47.3	44.7	−2.6
Gemma-4-31B (Thinking)	47.3	43.3	−4.0
Qwen3.5-9B	41.6	38.1	−3.5
Gemini-2.5-Pro (Thinking)	49.7	45.9	−3.8
Qwen3-VL-32B	36.8	32.1	−4.6
GPT-4o	33.4	30.2	−3.2
Qwen3-VL-8B	29.7	26.2	−3.5
Qwen2.5-VL-7B	26.8	22.0	−4.7
InternVL3.5-8B	22.9	17.4	−5.5
InternVL3-8B	22.5	16.7	−5.8
Gemini-2.5-Flash	22.3	16.4	−5.9

Type-specific scoring. All types are scored as binary (1/0); no partial credit is given.

- **Numerical.** Correct iff the values are equal after normalization. Thousand separators, redundant trailing zeros, and spelled-out numbers (e.g., "three") are normalized. A missing unit or an equivalent unit notation is accepted, while an explicitly different unit is marked incorrect. If the question specifies precision, both values are rounded to that precision; otherwise, both are rounded to the coarser of the two precisions and compared. A reversed sign or a different integer part is always incorrect.
- **Ordered / unordered collections and page lists.** All items must be present, with no missing or extra items; ordered types additionally require the correct order.
- **Table–page combinations** (e.g., "Table 2, pages 4–5"). Both the table and page components must be correct.
- **Text.** Exact match after ignoring case, whitespace, quotation marks, and trailing punctuation.

Comparison with ANLS. Tab. D.2 compares TAA and ANLS for all models. TAA is 3–6 points lower for every model, as it rejects near-miss values that ANLS partially credits.

D.2 PROMPT INSTRUCTION

D.2.1 PROMPT FOR QA GENERATION

QA Generation Prompt

```
You are generating a document QA example for a hierarchical
multi-page reasoning dataset. You are given {n_images} image(s), a
pre-selected question level, a pre-selected reasoning case, and a
canonical answer derived from trusted labels.

Question level          {level}
Reasoning case          {case_name}
Source question seed    {source_question}
Canonical answer        {gold_answer}

Global Constraints
(1) Use at least 2 pieces of evidence.
(2) Require at least 2 reasoning steps.
(3) Do not create a single-cell lookup question.
(4) Do not create a single-page-only question.
(5) Do not use external knowledge.

Case-specific constraints: {case_constraints}
Annotation facts: {annotation_facts}

Write exactly one QA example that satisfies the specified level
and case. The question must be answerable only from the provided
document images and annotation facts. The answer must exactly match
the canonical answer.

Output format (JSON):
{"question": "...", "answer": "...", "justification": "...",
"used_evidence_count": 2}
```

D.2.2 PROMPT FOR CASE-SPECIFIC CONSTRAINTS

L_1 (Document-level) Constraints

```
(a) table_text_joint_reasoning
 - Require BOTH table data and surrounding document text.
 - Removing either source must make the question unanswerable.

(b) multi_table_reasoning
 - Require combining information from at least TWO different tables.
 - The answer cannot be derived from a single table alone.

(c) document_level_aggregation
 - Require aggregating values across multiple pages or sections.
 - The final answer must NOT appear in any single cell.

(d) document_level_difference
 - Require comparing values from different pages and computing a
   difference.
 - The answer must involve a non-trivial calculation.

(e) contradiction_check
 - Verify whether a textual claim is supported by the table.
 - The answer MUST be TRUE or FALSE.
```

L_2 (Table-level) Constraints

```
(a) cross_page_table_span_count
 - Require counting or reasoning about spans across multiple pages.

(b) cross_page_continuation
 - Require identifying and merging table parts split across pages.
```

- Must be impossible to answer using only one page fragment.

(c) table_structure_alignment
- Require understanding headers, merged cells, or column alignment.
- Cannot be answered by reading a single row only.

(d) caption_header_interpretation -- Human annotation
(e) page_span_detection -- Human annotation

L_3 (Cell/column-level) Constraints

Single-table reasoning

(a) multi_page_aggregation
- Require combining multiple values such as sum or average.
- The final answer must NOT exist in any single cell.
- At least two rows or sections must be used.

(b) multi_page_difference
- Require computing a difference between two values from different rows or pages.

(c) multi_page_ratio
- Require calculating a ratio or percentage from at least two values.
- The answer must involve division.

(d) multi_page_condition_then_lookup
- First identify a row or category based on a condition.
- Then retrieve another value from that result.
- Both steps must be necessary.

(e) multi_page_rank_comparison
- Require ranking or comparing multiple values.
- The answer must depend on identifying the maximum, minimum, or order.

(f) multi_page_filtered_aggregation
- Require filtering rows based on a condition and then aggregating the filtered results.
- Both steps must be required.

Multi-table reasoning

(g) multi_table_computation
- Require combining information from at least TWO different tables.
- The answer cannot be derived from a single table alone.

(h) multi_table_retrieval
- Require retrieving and comparing values across multiple tables.
- The answer must depend on locating corresponding entries in different tables.

(i) multi_table_pattern_comparison
- Require comparing structural patterns or rankings across tables.
- The answer must depend on cross-table alignment.

Multi-hop reasoning

(j) multi_table_multi_hop
- Require at least TWO sequential reasoning steps across different tables.
- Skipping any step should make the answer incorrect.

(k) intra_table_multi_hop
- Require at least TWO sequential reasoning steps within the same table.
- Each step must depend on the result of the previous step.

D.2.3 PROMPT FOR ANSWER DERIVATION

Answer Derivation Prompt

Answer the question using the provided document image(s).

Question: `{question}`

Reply with the answer only. No explanation, no reasoning, no bullet points.
Final answer:

D.2.4 PROMPT FOR QA VERIFICATION

QA Verification Prompt (MLLMs verification)

You are verifying a QA pair for a benchmark dataset. Assign exactly ONE label based on the criteria below.

PASS	The question is answerable from the given document, the answer is uniquely supported, and the evidence correctly supports the answer.
AMBIGUOUS	The question has multiple valid interpretations or answers, or is missing necessary scope (page range, table span, aggregation scope, etc.).
REJECT	The question is unanswerable, the answer is incorrect, the evidence does not support the answer, or the question is trivially solvable from a single cell or page.

Evaluate using only the provided document and evidence.

Output format (JSON):

```
{
"final": "PASS" | "AMBIGUOUS" | "REJECT",
"reason": "one short sentence",
"ambiguity_type": "none | multiple_interpretations
| multiple_answers | missing_scope",
"supports_answer": true | false,
"single_cell_or_page_only": true | false
}
```

D.2.5 PROMPT FOR ANSWER EXTRACTION

Answer Extraction Prompt

You are given a question and a model's response to it. Copy the final answer from the response in its shortest form, using the response's own words and numbers exactly. Do not correct, recompute, translate, or add anything. If the question asks for several items, list them in the order the question asks, separated by `', '`. If the response does not state a final answer, output `NONE`.

Question: `{q}`

Response: `{r}`

Short answer:

D.3 MODEL HYPERPARAMETERS

The hyperparameters of the LVLMs and LLMs used in our experiments are detailed in Tab. D.3. Unless otherwise noted, the temperature is set to 0.0 and `max_new_tokens` is set to 1024 for all models. For Thinking-enabled models (Qwen3-VL Thinking, Gemma-4, and Gemini-2.5-Pro), `max_new_tokens` is set to 4096 to accommodate the extended reasoning trace generated during the thinking process. The `max_images` parameter specifies the maximum number of document page images that can be directly input to the model in a single inference call. The `max_model_len` parameter indicates the maximum context length (in tokens) configured for vLLM-based local inference.

Table D.3: Hyperparameters of the LVLMs and LLMs used in our experiments.

Category	Model	Model ID / Version	Hyperparameters
Open-Source MLLM	Qwen2.5-VL-7B	Qwen/Qwen2.5-VL-7B-Instruct	max_model_len=32768, max_images=65
	Qwen2.5-VL-72B	Qwen/Qwen2.5-VL-72B-Instruct	max_model_len=32768, max_images=65
	Qwen3-VL-8B	Qwen/Qwen3-VL-8B-Instruct	max_model_len=32768, max_images=65
	Qwen3-VL-32B	Qwen/Qwen3-VL-32B-Instruct	max_model_len=32768, max_images=65
	Qwen3.5-9B	Qwen/Qwen3.5-9B	max_model_len=32768, max_images=65
	InternVL3-8B	OpenGVLab/InternVL3-8B	max_model_len=16384, max_images=65
	InternVL3.5-8B	OpenGVLab/InternVL3_5-8B-Instruct	max_model_len=32768, max_images=65
	Gemma-4-31B	google/gemma-4-31B-it	max_model_len=32768, max_images=65
Open-Source MLLM (Thinking)	Qwen3-VL-8B (Thinking)	Qwen/Qwen3-VL-8B-Thinking	max_model_len=65536, max_images=65, temperature=0.6, max_tokens=4096
	InternVL3.5-8B	OpenGVLab/InternVL3_5-8B	max_model_len=65536, max_images=65, temperature=0.6, max_tokens=4096
	Gemma-4-31B (Thinking)	google/gemma-4-31B-it	max_model_len=65536, max_images=65, max_tokens=4096
Proprietary MLLM (API)	GPT-4o-mini	gpt-4o-mini	max_images=65, version: `gpt-4o-mini-2024-07-18`
	GPT-4o	gpt-4o	max_images=65, version: `gpt-4o-2024-08-06`
	GPT-5	gpt-5	max_images=65, version: `gpt-5-2025-08-07`
	Gemini-2.5-Flash	google/gemini-2.5-flash	max_images=65, version: `gemini-2.5-flash-latest`
	Gemini-2.5-Pro (Thinking)	google/gemini-2.5-pro	max_images=65, version: `gemini-2.5-pro-latest`

E ERROR ANALYSIS

We present representative failure cases across the three QA levels to illustrate common error patterns. Each case shows model predictions from multiple MLLMs on the same question, highlighting level-specific challenges: L_1 requires text-table cross-referencing (Fig. E.1), L_2 requires page-level table identification (Fig. E.2), and L_3 requires precise cell-value extraction across pages (Fig. E.3, Fig. E.4).

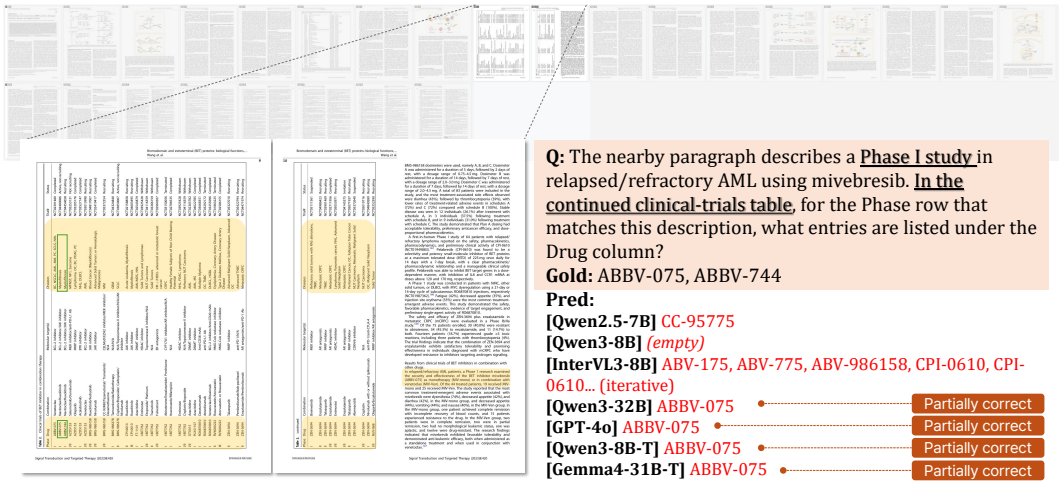


Figure E.1: **Model failure analysis in L_1 (Document-level).** The question requires linking a body-text description (Phase I study in relapsed/refractory AML) to a specific row in a continued table and retrieving multiple drug entries. Most models identify only the first drug (ABBV-075) but miss the second (ABBV-744), yielding partial credit. This reveals a systematic weakness in exhaustive cell retrieval from multi-page tables even when the correct row is located.

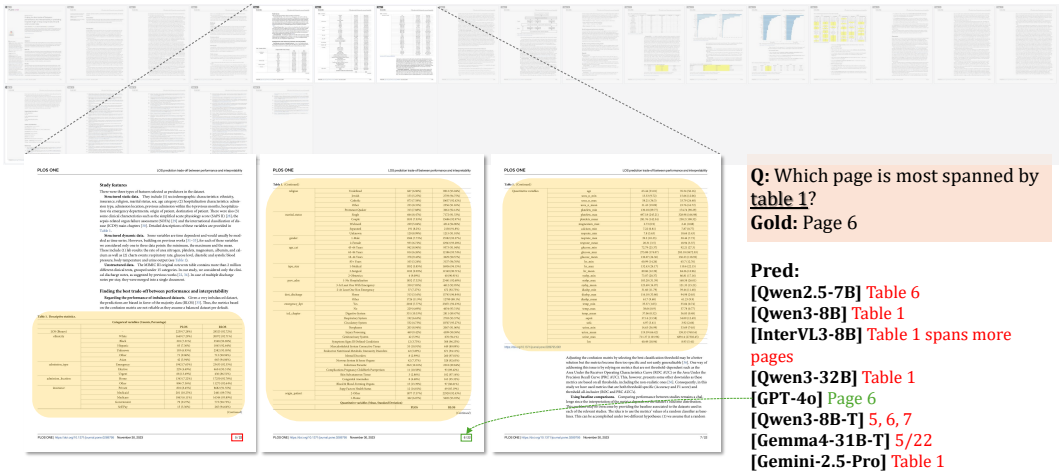


Figure E.2: **Model failure analysis in L_2 (Table-level).** The question asks which page is most spanned by Table 1. Most models fail to distinguish between table identity and page location, returning "Table 1" or "Table 6" instead of a page number. Thinking models produce fragmented answers ("5, 6, 7" or "5/22") rather than identifying the single most-spanned page. Only GPT-4o correctly answers "Page 6."

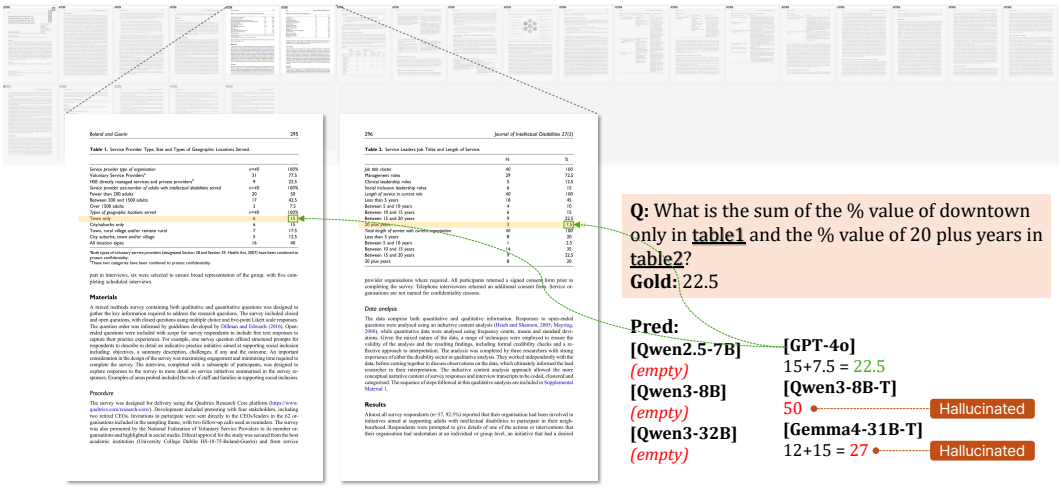
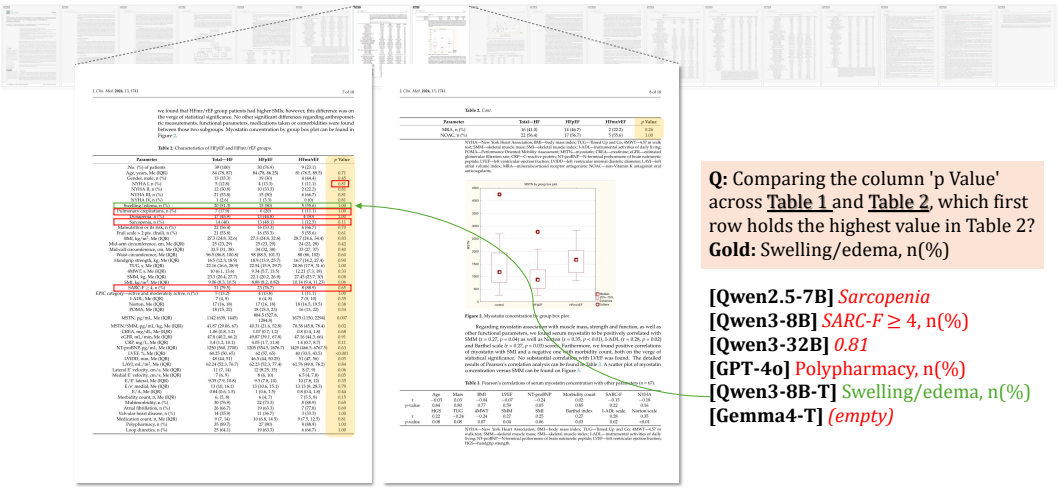


Figure E.3: **Model failure analysis in L_3 (Cell/column-level).** The question requires locating specific rows across two tables and summing their values. Non-Thinking models produce empty responses, while Thinking models hallucinate incorrect values. Only GPT-4o correctly identifies both values (15 + 7.5 = 22.5).



Q: Comparing the column 'p Value' across Table 1 and Table 2, which first row holds the highest value in Table 2?
Gold: Swelling/edema, n(%)

[Qwen2.5-7B] *Sarcopenia*
[Qwen3-8B] *SARC-F* ≥ 4, n(%)
[Qwen3-32B] *0.81*
[GPT-4o] Polypharmacy, n(%)
[Qwen3-8B-T] Swelling/edema, n(%)
[Gemma4-T] *(empty)*

Figure E.4: **Model failure analysis in L_3 (Cell/column-level).** The question requires comparing p-values across a table split over two pages, then retrieving the corresponding row label. All non-Thinking models return different incorrect rows, revealing that each model misinterprets the question or misreads the table differently. Only Qwen3-VL-8B (Thinking) identifies the correct row (Swelling/edema).