DocHop-QA: Towards Multi-Hop Reasoning over Multimodal Document Collections

Anonymous ACL submission

Abstract

Despite rapid progress in large language models (LLMs), current QA benchmarks still overlook the core challenge of real-world scientific information seeking: synthesizing multimodal evidence scattered across multiple documents and structural formats. Existing QAs remain narrow in scope, relying on unimodal text and short-span reasoning that fail to capture the complexity of real information-seeking. We introduce **DocHop-QA**, a benchmark of 11,379 instances for evaluating multimodal, multi-document, multi-hop scientific QA. Built from publicly available PubMed articles, DocHop-QA incorporates textual passages, tables, and layout cues, enabling cross-document inference without explicit hyperlinks. To scale realistic QA construction, we develop an LLM-driven generation pipeline grounded in 11 scientific reasoning concepts, producing diverse and coherent question–answer pairs. To highlight the utility and versatility of the dataset, we propose a task-driven evaluation framework spanning four settings, including generative answering, multimodal evidence integration and structured index prediction. Experiments show that current models struggle with DocHop-QA's long-context, multi-evidence demands, establishing it as a rigorous testbed for advancing next-generation scientific QA systems.

1 Introduction

Recent advances in Large Language Models (LLMs) have substantially improved Question Answering (QA) performance, particularly in single-document and short-context settings. However, real-world information-seeking rarely conforms to such simplified conditions. In practice, answering complex questions often requires integrating fragmented evidence across multiple documents and heterogeneous formats, including narrative text, tables, and structured layouts (Yoon et al., 2022). This setting naturally gives rise to *multi-hop QA*, where models must retrieve and reason over multiple pieces of evidence rather than rely on localized textual cues. In scientific domains, for example, answering questions such as "What proteins are involved in pathway X, and what experiments validate their roles?" typically demands synthesizing various evidence across several publications.

Despite recent progress, deploying LLMs for multi-hop QA in realistic document-centric environments remains challenging, reflecting fundamental NLP failure modes including sensitivity to retrieval quality, difficulty grounding semi-structured tables, and error accumulation across long reasoning chains. Specifically, omitting a single critical evidence node can break the reasoning chain, while retrieving loosely related passages introduces distraction and confusion (Zhu et al., 2025; Bai et al., 2025b). Yet existing benchmarks insufficiently capture these failure modes, remaining largely confined to Wikipedia-sourced, unimodal settings with short extractive answers.

To fill this gap, we introduce DocHop-QA, a large-scale benchmark for complex, multi-hop QA over multimodal, multi-document corpora, constructed from scientific articles. Its generation methodology is domain-agnostic and scalable, making it applicable to a wide range of scientific domains. Constructed from PubMed articles, it includes unstructured text, structured tables, and layout features, without relying on predefined hyperlinks or annotated reasoning chains. To efficiently generate diverse and challenging QA pairs, we present a scalable LLM-driven pipeline based on 11 real-world scientific reasoning concepts. To evaluate the applicability and flexibility of DocHop-QA, we benchmark its performance across four representative QA tasks: (1) Generative Text Extraction, (2) Structured Generative Answering, (3) BBox Entity Index Extraction, and (4) XML Entity Index Extraction. DocHop-QA intentionally isolates the multi-document evidence integration

stage in a controlled setting, rather than modeling the full end-to-end retrieval pipeline. Each instance requires identifying multiple relevant evidence spans across documents and synthesising a coherent natural-language answer, targeting the core capability of cross-document reading and reasoning. This design enables rigorous evaluation of evidence aggregation and cross-document synthesis, a fundamental yet underexplored challenge in scientific document understanding.

Our key contributions are as follows: 1) We introduce **DocHop-QA**, a large-scale benchmark for multi-hop QA over multimodal, multi-document scientific corpora, incorporating text, tables, and layout features. 2) We develop a **scalable LLM-driven generation pipeline** guided by 11 reasoning concepts inspired by real-world scientific inquiry, enabling the automated creation of diverse and challenging multi-hop QA instances without gold chains or hyperlinks. 3) We **evaluate DocHop-QA across four QA tasks**, ranging from structured index prediction to generative answering, demonstrating its broad applicability for assessing both discriminative and generative reasoning over textual, structural, and visual information.

2 Related Work

Multi-hop QA Multi-hop QA tasks span multiple documents, pages, and modalities, reflecting a realistic information seeking process. Early benchmarks such as QAngaroo (Welbl et al., 2018) and HotpotQA (Yang et al., 2018) pioneered multi-hop reasoning over text, but rely on predefined knowledge bases or relatively short contexts. Later, HybridQA (Chen et al., 2020) and WebQA (Chang et al., 2022) expanded evidence modalities by incorporating semi-structured data such as tables, yet still lean toward web-derived resources and exhibit evidence that is relatively locally concentrated. More recent datasets, including FanOutQA (Zhu et al., 2024), MEQA (Li et al., 2024b), LLeQA (Louis et al., 2024) and Loong (Wang et al., 2024) increase difficulty along axes such as hop depth or context length to probe model reasoning under greater complexity. Nonetheless, many remain text-centric and predominantly feature short answers, leaving a gap to the paragraph-level synthesis demanded in scientific information seeking. Moreover, many still depend on explicitly structured representations such as relational DBs (Zhu et al., 2025) and knowledge graphs (Kim et al., 2025) to support complex reasoning. More recently,

Dataset	M Hop	VRD	#QA	Q Gen	Source	M Doc	Modality
HotpotQA	O	X	113K	H	Wiki	X	Tx
MEQA	O	X	2.2K	LLM + H	WikiEvents	X	Tx
NOVELHOPQA	O	X	3.8K	LLM	Gutenberg	X	Tx
HybridQA	O	X	70K	H	Wiki	O	Tx+Tab
Loong	O	X	1.6K	LLM + H	Industry	O	Tx
FanOutQA	O	X	1K	H	Wiki	O	Tx
M3SciQA	O	X	1.4K	H	EMNLP Papers	O	Tx+Tab+Img
DocVQA	X	O	50K	H	Industry	X	Tx+Img
SlideVQA	X	O	14K	H	Slideshare	X	Tx+Img
ECG-QA	X	O	414K	H	PTB-XL	X	Tx+Img
SPIQA	X	O	270K	LLM + H	arXiv	X	Tx+Tab+Img
MMVQA	X	O	263K	LLM + H	PubMed	X	Tx+Tab+Img
VisDoMBench	X	O	2.2K	LLM	Wikip.+arXiv	O	Tx+Tab+Img
Ours	O	O	11K	LLM + H	PubMed	O	Tx+Tab+Img

Table 1: **Comparison of QA dataset benchmarks.** M Hop: Multi Hop, VRD: Visually Rich Document, Q Gen.: Question Generation method (H: Human, LLM: Large Language Model), M Doc: Multiple Document, Modality (Tx: Text, Tab: Table, Img: Image). Further details about Multi-hop and reasoning are in Table 4

NOVELHOP (Gupta et al., 2025) targets multi-hop reasoning over long narrative contexts but is constrained to unimodal text. MEBench (Lin et al., 2025) extends this to entity-dense scenarios but remains text-centric, while M3SciQA (Li et al., 2024a) introduces multimodal multi-document QA yet relies on citation-based anchor links. In contrast, DocHop-QA constructs semantically related pairs of real scientific documents, requiring models to aggregate and synthesize paragraph- and table-level evidence across documents without relying on explicit structural anchors.

Visually Rich Document QA Visually rich document QA extends question answering beyond flat text to real-world documents with complex layouts, tables, and figures. Early datasets such as DocVQA (Mathew et al., 2021) and InfographicVQA (Mathew et al., 2022) focus on single page document images, often framed as span extraction or retrieval tasks. SlideVQA (Tanaka et al., 2023) and BLIVA (Hu et al., 2024) introduce multi-hop reasoning, but are primarily studied in single-document settings. More recent benchmarks such as ECG-QA (Oh et al., 2023), SPIQA (Pramanick et al., 2024) , MMVQA (Ding et al., 2024), and VisDoMBench (Suri et al., 2025) further extend VRD QA toward multimodal and multi-document settings. However, they largely emphasize retrieval-oriented evaluation, offer limited open-ended multi-hop reasoning, and mostly manually constructed for specific document types, limiting scalability. In contrast, DocHop-QA supports scientific document reasoning, incorporating text, tables, and layout features without manual construction constraints. Table 1 summarizes the comparison of prior datasets with ours across key dimensions.

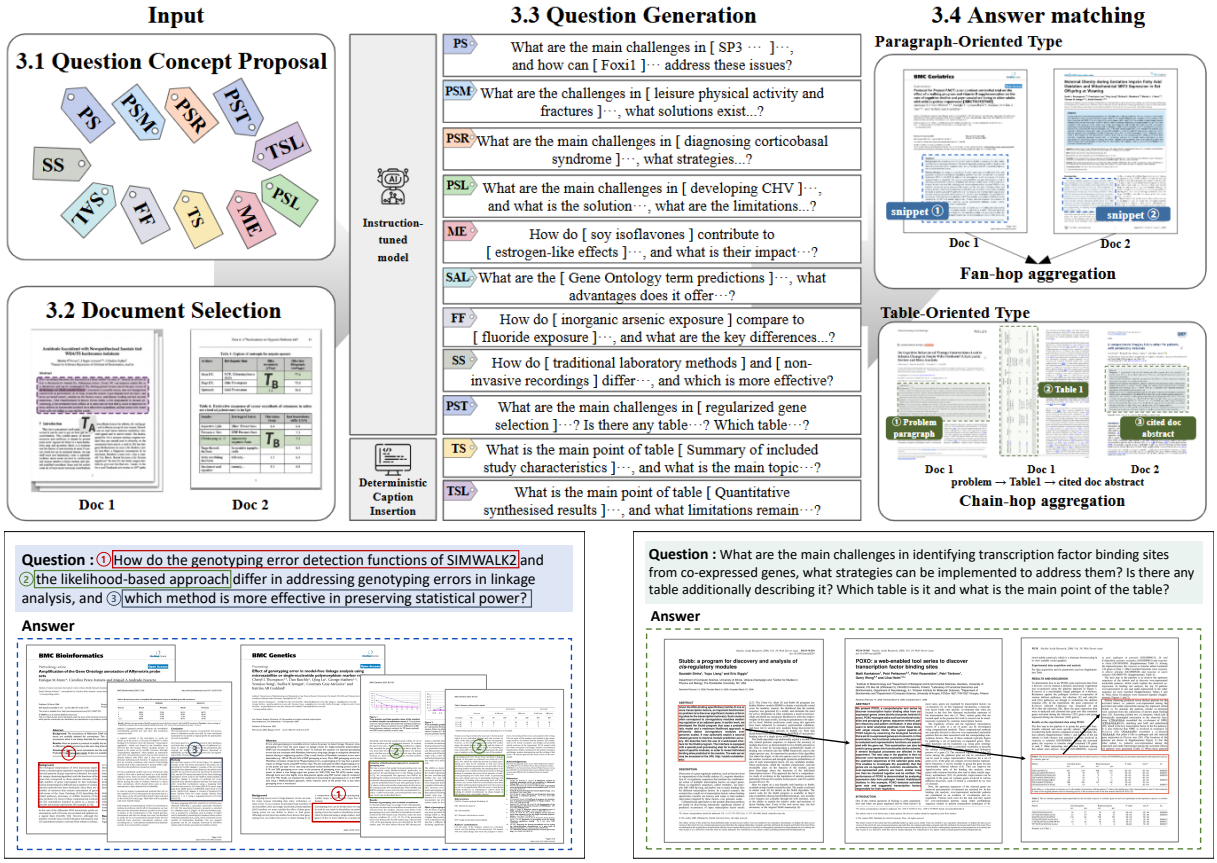


Figure 1: **Overview of the DocHop-QA construction pipeline across four stages** (1) Question Concept Proposal, (2) Document Selection, (3) Question Generation, and (4) Answer Matching. Bottom row shows representative Comparison-type (left) and PST-type (right) QA instances.

3 Dataset Construction

DocHop-QA is constructed through a four-stage pipeline illustrated in Figure 1: (1) proposing diverse question concepts, (2) selecting document pairs suitable for multi-hop reasoning, (3) generating questions and (4) matching answers. We build on the PubMed Central (PMC) open access repository which provides structured XML and PDF versions of biomedical and life science literature. From roughly 2M full text scientific articles we obtained approximately 20K document pairs covering 9,250 unique documents forming a controlled testbed for cross-document evidence integration over heterogeneous content.

3.1 Question Concept Proposal

To enable systematic and scalable multi-hop question generation we define 11 *question concepts* grounded in reasoning patterns frequently observed in PubMed articles. The concepts were derived through a human-in-the-loop process. We first used NotebookLM to propose cross-document question candidates grounded in uploaded source documents which authors then reviewed and refined into 90 seed QA pairs. Qualitative analysis identified core semantic keywords such as `Problem`, `Solution` and `Effect` capturing fundamental components of scientific reasoning. Using these as scaffolds we generated an additional 1,000 multi-hop QA pairs with ChatGPT through prompts tailored to elicit multi-step reasoning. We then consolidated structurally overlapping or infrequent scaffolds into 11 representative question concepts summarised in Table 2. Paragraph-Oriented concepts follow a fan-hop aggregation pattern while Table-Oriented concepts follow a chain-hop sequential inference pattern as illustrated in Figure 1.

3.2 Document Selection

We selected document pairs by ensuring topical alignment without predefined explicit links between documents. The selection strategy differs by question type to reflect the distinct reasoning structures each type demands.

[1] See Fig. 22 for concept-wise distributions of section and paragraph level answer snippet positions.

Type	Subtype	Keyword	Q Type	Question Concept	Question Example
Paragraph-Oriented (75.3%)	Non-Comparison	<**P**roblem><sep><**S**olution>	PS(1215)	What are the main challenges in [ ], and how can [ ] address these issues?	What are the main challenges in understanding the dual role of the transcription factor Sp3 in tumor progression, and how can the forkhead transcription factor Foxi1 address these issues by regulating the expression of key subunits of the vacuolar H+ ATPase?
		<**P**roblem><sep><**S**olution><sep><**M**echanism>	PSM (720)	What are the challenges in [ ], what solutions exist to address these challenges, and how do these solutions help resolve the issues?	What are the challenges in the relationship between leisure physical activity and the risk of osteoporotic fractures in men, what solutions exist to address these challenges, and how do these solutions help resolve the issues?
		<**P**roblem><sep><**S**olution><sep><**R**esult/effect>	PSR (977)	What are the main challenges in [ ], what strategies can be implemented to address them, and what are the observed effects of these strategies?	What are the main challenges in accurately diagnosing corticobasal syndrome, what strategies can be implemented to address them, and what are the observed effects of these strategies?
		<**P**roblem><sep><**S**olution><sep><**L**imitation>	PSL (1,269)	What are the main challenges in [ ], and how can [ ] address these issues? Additionally, what are the limitations of this approach?	What are the main challenges in developing consumer health vocabulary (CHV) due to the heterogeneity and ambiguity of consumer expressions, and what is the solution to this challenge? Additionally, what are the limitations of this approach?
		<**M**echanism><sep><**E**ffect>	ME (1,155)	How does [ ] contribute to [ ], and what is its impact on [ ]?	How do soy isoflavones contribute to estrogen-like effects, and what is their impact on male breast cancer incidence?
		<**S**tudytopic><sep><**A**dvantage><sep><**L**imitation>	SAL (850)	What are the [ ], what advantages does it offer, and what limitations exist?	What are the Gene Ontology term predictions, what advantages does it offer, and what limitations exist?
	Comparison	<**F**eature><sep><**F**eature>	FF (1,229)	How does [ ] compare to [ ], and what are the key similarities and differences?	How do the effects of inorganic arsenic exposure on children's intellectual function compare to those of fluoride exposure, and what are the key similarities and differences?
		<**S**olution><sep><**S**olution>	SS (1,152)	How do [ ] and [ ] differ in addressing [ ], and which approach is more effective?	How do traditional laboratory and clinical research methods and non-invasive abdominal recordings differ in addressing the prediction of preterm labor, and which approach is more effective?
Table-Oriented (24.7%)	Non-Refer Table	<**P**roblem><sep><**S**olution><sep><**T**able>	PST (874)	What are the main challenges in [ ], what strategies can be implemented to address them? Is there any table additionally describing it? Which table is it and what is the main point of the table?	What are the main challenges in regularized gene selection in cancer microarray meta-analysis, what strategies can be implemented to address them? Is there any table additionally describing it? Which table is it and what is the main point of the table?
	Refer Table	<**T**able><sep><**S**tudytopic><sep><**L**imitation>	TSL (573)	What is the main point of Table [ ] in document [ ], and what is the main topic of the referenced paper? Also, what limitations remain in the referenced paper?	What is the main point of table Quantitative synthesised results in given document, and what is the main topic of the referenced paper? Also, what limitations remain in the referenced paper?
		<**T**able><sep><**S**tudytopic>	TS (1,365)	What is the main point of Table [ ] in document [ ], and what is the main topic of the referenced paper?	What is the main point of table Summary of included study characteristics in given document, and what is the main topic of the referenced paper?

Table 2: **Illustrative examples of the proposed question concepts**, categorized by (1) **Type**, denoting the primary answer modality; (2) **Subtype**, indicating the reasoning structure; (3) **Keyword**, representing the core reasoning elements and semantic placeholders used in question generation; (4) **Q Type**, a shorthand for the placeholder composition; (5) **Question Concept**, the abstract question structure using placeholders; and (6) **Question Example**, a concrete instantiation of each question concept. This design enables controlled, diverse, and semantically grounded multi-hop QA generation across heterogeneous document modalities.[1]

Paragraph-Oriented Type Document pairs are selected by keyword cooccurrence and lightweight semantic similarity over titles and abstracts without predefined citation links between documents. This encourages open ended evidence aggregation across semantically related but structurally unlinked documents. The detailed process is provided in Appendix D.1.

Table-Oriented Type Unlike paragraph-oriented settings, Table-Oriented types intentionally use citation-linked tables to model structured chain-hop reasoning. For the PST type we retained only pairs in which at least one table appears in a results section as these tables typically summarise intervention outcomes. For TS and TSL types which model reasoning over tables citing external studies we targeted systematic review articles containing citation linked tables. These citation links serve as intra document structural anchors for sequential evidence transitions rather than as inter-document pairing criteria.

3.3 Question Generation

Paragraph-Oriented Type We tuned an instruction following model on instances pairing question concept templates with document abstracts and gold question examples (Appendix D.2). We instruction-tuned the Qwen2-Instruct model using Unsloth's 4-bit LoRA framework(Hu et al., 2021)[2], chosen for its open-source availability, enabling reproducible and cost-efficient large-scale question generation. To preserve controllability and semantic consistency, we trained separate models for each question category (comparison, non-comparison, and PST). Joint training proved suboptimal, frequently omitting contrastive cues in comparison questions and producing fragmented reasoning in non-comparison and PST cases. During inference, we supplied a few-shot prompt that includes the concept, 2 abstracts, and 2 example questions using a step by step reasoning skeleton.

Table-Oriented Type For the PST type, questions were generated using the same model based approach as the Paragraph-Oriented type, but differed in the composition of the input to better reflect the table content. For TS and TSL types we designed a deterministic procedure in which table captions are inserted directly into predefined question placeholders. This design is appropriate because the reasoning transitions for these types are

[2] https://huggingface.co/unsloth/Qwen2-7B-Instruct-bnb-4bit

explicitly defined by table references making LLM-based generation unnecessary while preserving full structural controllability. Post processing filters applied to all types are described in Appendix D.2.

3.4 Answer Matching

In DocHop-QA, answer retrieval is guided by two distinct reasoning patterns[3]: *fan-hop* for Paragraph-Oriented types, where semantically related content is aggregated from documents (Zhu et al., 2024), and *chain-hop* for Table-Oriented types, which require explicit links between structured and unstructured content (Xu et al., 2021; Chen et al., 2019). Following the reasoning patterns illustrated in Figures 1 our matching pipeline adopts unified retrieval for Paragraph-Oriented types and tailored strategies for Table-Oriented ones.

Paragraph-Oriented Type For Paragraph-Oriented types, evidence is dispersed across documents. Semantically complementary five snippets are retrieved independently from each source document and aggregated without explicit cross-referencing between sources. Candidate segments are ranked by hybrid lexical and semantic similarity to the question and only instances with sufficiently coherent evidence are retained.

Table-Oriented Type For Table-Oriented types, evidence is sequential. We constructed up to four snippets per instance, connecting problem and solution paragraphs, tables, and referenced content across documents. Retrieval combines semantic similarity with keyword filtering and applies logical chaining when anchor references are present. Instances with low similarity were filtered to preserve dataset integrity. Detailed retrieval logic is provided in Appendix D.3.

4 Dataset Quality Assurance

To evaluate the quality of the DocHop-QA dataset, we define three criteria assessed via a 4-point rubric: ***Reality/Fluency*** ensures that questions are natural, plausible, and meaningful within the context. ***Accuracy*** assesses whether each snippet meaningfully addresses the question and contributes to forming a complete answer. ***Completeness*** evaluates whether the retrieved snippets collectively form a sufficient answer.

Human-LLM Quality assurance To validate dataset quality, we conduct a human evaluation on 50 carefully selected QA instances, covering all 11 question concepts, both paragraph- and table-oriented types, and diverse hop patterns(annotator demographics, IRB statement, and evaluation interface details are provided in Appendix E.2). We report Gwet's AC2 as the primary agreement metric, as it is robust to marginal distribution skewness commonly observed in quality annotation tasks. Human annotations are compared against three instruction-tuned LLM judges: GPT-4o[4], Gemini2.5 Flash[5], and Qwen3.5-9B, using the same rubric (prompt details in Appendix E.1), all of which show substantial agreement with human judgments across all criteria (Table 3). Additional metrics including Spearman's ρ and Krippendorff's α are provided in Appendix E. The evaluation is then extended to the full dataset using Qwen3.5-9B, an open-source model that achieves competitive agreement while remaining locally deployable. Across all 11,379 instances, approximately **92%**, **87%**, and **76%** of QA pairs receive a score of 3 or above on Reality/Fluency, Accuracy, and Completeness, respectively, suggesting that the quality trends confirmed by human evaluation generalize to the entire DocHop-QA corpus. Detailed results are provided in Appendix E.

Criterion	Model	AC2 (ordinal)
Completeness	GPT-4o	0.555
	Gemini2.5	0.331
	Qwen3.5-9B	0.342
Accuracy	GPT-4o	0.632
	Gemini2.5	0.424
	Qwen3.5-9B	0.741
Reality	GPT-4o	0.874
	Gemini2.5	0.911
	Qwen3.5-9B	0.935

Table 3: Human–LLM agreement measured by Gwet's AC2 agreement coefficients on 50 selected QA instances

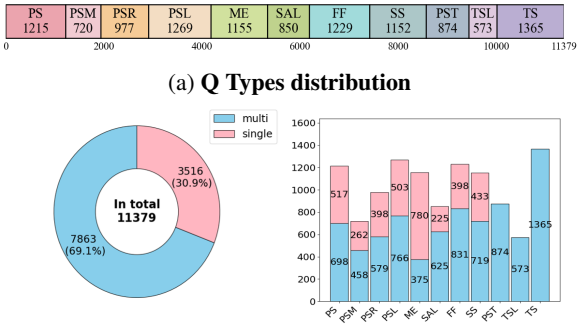
5 Dataset Analysis

DocHop-QA comprised 11,379 instances spanning 11 evenly distributed question concepts, each representing approximately 10% of dataset, with both single- and multi-document instances across Question concept types shown in Figure 2. The following analyses characterize DocHop-QA from three perspectives: (1) Document-level complexity, (2)

[3]Further details are provided in Appendix D.3

[4]https://openai.com/index/hello-gpt-4o/

[5]https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash



(a) **Q Types distribution**

(b) **Used Document Type** (c) **QA Distribution**

Figure 2: **Overview of document usage across question types.** Multi-document reasoning dominates.

Answer Dispersion, and (3) Reasoning complexity.

5.1 Dataset Coverage

As shown in Figure 3, most documents span ~10 pages, contain ~20 subsections and ~60 paragraphs totaling 10,000 to 15,000 tokens, closely reflecting the scale and complexity of real-world scientific literature. The questions constructed from these documents inherit their semantic richness, reflecting the depth and complexity of biomedical literature. The questions constructed from these documents inherit their semantic richness, reflecting the depth and complexity of biomedical literature. Furthermore, keyword analysis across 11 question concepts (see Appendix F.1) reveals both topical diversity and domain specificity, with biomedical terms dominating, as expected from PubMed-sourced data.

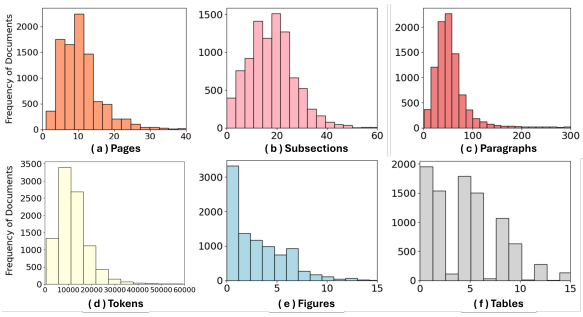


Figure 3: **Distribution of document component types.** DocHop-QA includes multi-page, multi-section VRDs, supporting complex multi-hop and multimodal QA.

5.2 Answer Dispersion

DocHop-QA encodes structurally diverse and non-position-biased reasoning behaviors across its 11 question concepts. To support section level analysis, diverse section titles are mapped into 8 unified Super-Section tags described in Appendix F.2. As shown in Figure 4, Paragraph-Oriented types predominantly retrieve from Abstract while Table-Oriented types draw from Results and Discussion, consistent with their respective fan-hop and chain-hop strategies (see Figure 21 and Table 15). This section-level divergence is further supported by structural distance analysis in Figure 5 and further explained in Appendix F.2. Fan-hop dispersion was measured by absolute positional distance and normalized distance ratio within documents, and chain-hop using both element-wise and their average using inter-snippet distances between consecutive steps. Paragraph-Oriented instances using fan-hop strategy exhibit broadly distributed snippet distances without strong positional bias, while Table-Oriented instances of chain-hop strategy show a heavy-tailed distribution where most transitions are local but a non-negligible fraction spans long ranges, reflecting a mix of short and long-range dependencies. Further analysis of chain-hop transition in Figure 23 reveals that tables frequently serve not only as terminal evidence but also as intermediate reasoning bridges. These findings confirm that DocHop-QA systematically encodes diverse retrieval behaviors across document structure, posing genuine challenges for both evidence localization and multi-hop aggregation.

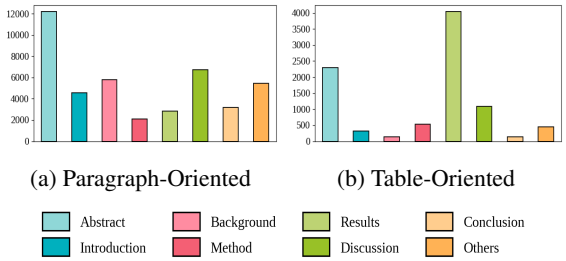
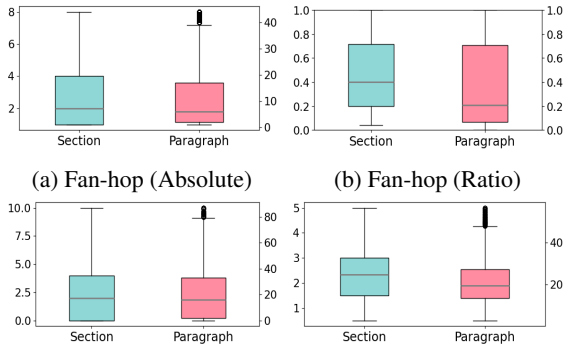


Figure 4: **Answer snippet distribution for each question types**; **Abstract** is the most frequent for Paragraph-Oriented whereas **Result** dominates Table-Oriented).

5.3 Reasoning Complexity

Table 4 reports context source, answer form, hop count and reasoning depth across existing benchmarks, where reasoning depth is defined as the average number of decomposable sub-questions per instance. While DocHop-QA shows comparable hop counts to prior datasets (avg. 4.63 hops), the nature of evidence integration is fundamentally different. Although Loong (Wang et al., 2024) exhibits higher hop counts and reasoning depth, its reasoning traversal relies on explicit citation links and predefined structural connections between doc-



(a) Fan-hop (Absolute) (b) Fan-hop (Ratio)

(c) Chain-hop (Element) (d) Chain-hop (Average)

Figure 5: **Section and paragraph level structural distance of fan-hop and chain-hop questions.**

Dataset	Context	Answer	Num Hops	Reasoning Depth
HotpotQA	Wiki	short span and supporting facts	2.0	1.42
HybridQA	Wiki	short span	3.03	2.05
Loong	Industry	short spans or NL sentence	10.18	3.25
FanOutQA	Wiki	collections of short span	6.40	2.96
MEQA	Wiki	short span	2.53	2.53
NovelHopQA	Gutenberg	1 -2 sentences	2.50	2.50
Ours	**Pubmed**	**collection of paragraphs**	**4.63**	**2.46**

Table 4: **Comparison of QA benchmark dataset** described in Table 1 in terms of context granularity, answer form, and reasoning complexity

uments, reducing the open-ended discovery burden on the model. Similarly, most existing benchmarks are text-centric, explicitly linked, and rely on short-span answers. In contrast, DocHop-QA draws from real scientific literature and requires synthesizing information across multiple paragraphs and tables, with cross-document reasoning constructed through semantic similarity rather than predefined structural connections. These properties establish DocHop-QA as a benchmark that demands richer semantic integration and higher reasoning complexity than hop statistics alone would suggest.

6 Experiments and Results

To evaluate the utility of DocHop-QA, we conduct experiments across four QA tasks covering different reasoning paradigms and input modalities, including generative and discriminative settings. These tasks assess models' ability to reason over complex, document-rich inputs involving structured index prediction, free-form generation, and multimodal integration. Discriminative tasks involving entity index prediction, namely Task 3 (BBox Entity Index Extraction) and Task 4 (XML Entity Index Extraction) are discussed in detail in Appendix G, along with detailed experiment setups, metrics, and ablations. Appendix H further discusses different ablation studies.

Model	Input	BLEU	R-1	R-2	R-L	R-LS
InternVL2	T	14.53	32.11	11.99	19.56	20.91
InternVL2	I	6.39	28.74	6.99	16.12	18.50
InternVL2	T+I	7.48	31.23	8.55	16.46	20.26
Qwen2.5-VL	T	7.32	27.66	8.30	15.45	19.44
Qwen2.5-VL	I	12.40	34.48	11.49	19.50	22.53
Qwen2.5-VL	T+I	9.45	31.35	9.96	17.94	21.40
Qwen3-VL 8B	T	26.00	–	–	28.80	–
Qwen3-VL 8B	I	18.20	–	–	21.00	–
Qwen3-VL 8B	T+I	29.70	–	–	32.10	–
GPT-4o	I	28.92	**45.79**	**23.81**	31.26	**34.12**
Gemini 2.5	T	11.94	24.55	12.85	16.81	19.41
Gemini 2.5	I	11.96	25.47	10.96	15.33	18.68
Gemini 2.5	T+I	11.71	24.57	12.61	16.30	18.95
Gemini 3 Pro	T+I	**31.8**	–	–	**34.5**	–

Table 5: **Generative text extraction results under one-shot prompting across models and input configurations.** BLEU and ROUGE-1/2/L/LSum are reported. T: Text, I: Image.

Task 1: Generative Reasoning. This task tests models' ability to generate natural-language answers given different combinations of text and image inputs. We assessed three prompting styles (question-only, zero-shot, and one-shot) and three input modalities (text-only, image-only, text+image). We further compare different input modes with InternVL2-4B, Qwen2.5-VL-7B, and Qwen3-VL-8B for preprocessed inputs and GPT-4o, Gemini2.5 Flash, and Gemini3 Pro[6] for full-document ingestion (XML, PDF). Evaluation is based on BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores against gold answers. The generative reasoning task assesses the ability of large vision-language models to produce free-form answers from document-level context, without relying on predefined answer indices. The results of one-shot prompting setting is described in Table 5. Gemini3 Pro demonstrates superior performance (BLEU: 31.8) with full-document ingestion, as opposed to input-processed architectures, where texts and images are incorporated into prompts, consuming valuable input context window. Detailed results are provided in Appendix G.

Task 2: Structured Generative Answering. For this task, models generate lists of answer entity indices in response to questions, using labeled XML content and, optionally, associated page images. We compared InternVL2-4B (Chen et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025a), and GPT-

[6] https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-pro?hl=ko

Model	S	Img	Sample F1	Sample Re	Sample Pr	Accuracy Be	Accuracy Co	Accuracy Ov
InternVL2	Z	–	7.03	9.77	7.39	0.88	0.88	28.30
InternVL2	Z	✓	4.24	6.72	4.03	0.35	0.75	17.44
InternVL2	O	✓	4.34	5.77	4.68	0.44	0.62	17.31
Qwen2.5-VL	Z	–	14.06	51.62	13.69	4.04	34.05	75.48
Qwen2.5-VL	Z	✓	7.92	58.53	7.06	1.54	50.04	69.73
Qwen2.5-VL	O	✓	8.12	18.55	7.41	0.04	9.89	41.08
GPT-4o	Z	-	**16.21**	54.12	15.85	5.18	36.86	**79.38**
GPT-4o	Z	✓	9.37	60.27	8.91	2.12	52.46	73.15
GPT-4o	O	✓	9.15	21.43	8.75	0.97	12.21	45.62

Table 6: **Structured Generative Answering results.** S: Setup (Z: Zero-shot, O: One-shot), Img: Image input.

4o in zero-shot and one-shot settings. Evaluation spans multi-label spatial accuracy (belong/overlap/-contain)[7] and sample-level precision, recall, and F1. This task assesses whether generative models can produce structured answer indices from labeled contextual inputs. GPT-4o achieves the highest F1 score (16.21) followed by Qwen2.5-VL (14.06) in a zero-shot, text-only setting (Table 6), showing strong performance without fine-tuning. InternVL2 performs poorly across sets and struggles with consistent output formatting, underscoring the need for task-specific pretraining. Adding document images generally hurts performance due to alignment challenges between text and visual layouts. The structured generative task highlights both the promise and limitations of using instruction-following LLMs for entity-index generation. DocHop-QA provides a challenging testbed for structured reasoning in generative settings, enabling deeper studies of prompt design, modality alignment, and decoding strategies.

Qualitative Analysis: Cross-Document Evidence Integration. We present a representative qualitative example to illustrate the reasoning challenges in DocHop-QA. Figure 6 shows an example from Task 1 (Generative Reasoning), where answering the question requires integrating complementary findings across two independent documents. Document A reports tentative linkage-analysis evidence suggesting possible QTLs associated with smoking behavior, while Document B identifies susceptibility chromosomal regions for nicotine dependence using extended pedigree analysis. Correct answering requires models to synthesize the shared conclusion that smoking-related traits have genetic influences while distinguishing differences in confidence, phenotype focus, and specificity of the reported loci. Stronger models such as Gemini 3 Pro successfully aggregate evidence across documents and produce coherent comparative summaries. However, even strong models such as GPT-4o may hallucinate unsupported chromosomal regions not explicitly stated in the evidence. Weaker models frequently fail to integrate the two studies correctly and instead misinterpret the relationship between smoking behavior and nicotine dependence. This example demonstrates that DocHop-QA requires more than local evidence retrieval: models must perform cross-document reasoning, comparative interpretation, and coherent answer synthesis from distributed scientific evidence.

[7] https://meta-pytorch.org/torcheval/stable/generated/torcheval.metrics.MultilabelAccuracy.html

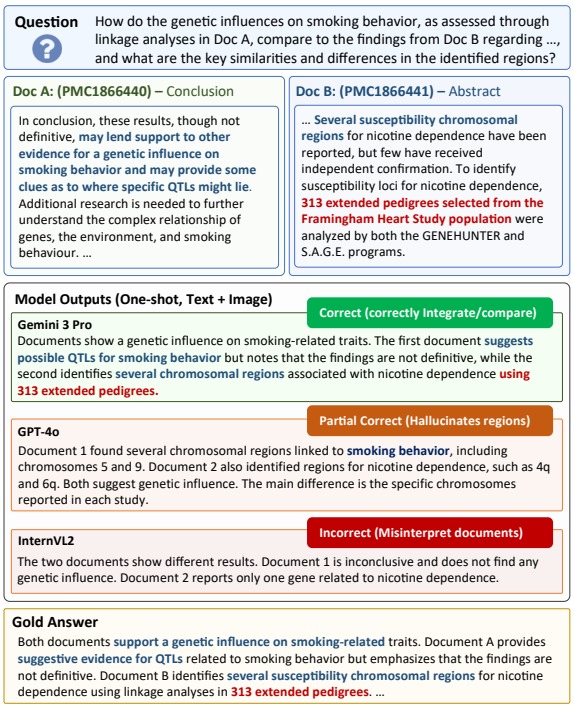


Figure 6: Representative qualitative sample for cross-document evidence integration.

7 Conclusion

We presented DocHop-QA, a new benchmark designed to evaluate multi-hop reasoning over multimodal, multi-document scientific corpora. Unlike prior datasets, DocHop-QA captures realistic information seeking scenarios by integrating unstructured text, structured tables, and visual layout features, without relying on explicit links. Using 11 concepts, we generated diverse QAs and evaluated performance on 4 representative tasks. Results highlight critical insights into model behavior, modality integration challenges, and prompt effects, underscoring the need for further research in multi-hop, multimodal reasoning. We hope DocHop-QA facilitates future research toward more robust QA systems for real world documents.

Limitations

DocHop-QA has a few limitations worth noting. First, parts of the dataset are generated using LLM-assisted pipelines, which may introduce stylistic regularities in question formulation despite concept-guided prompting and human validation. Second, we evaluate end-to-end reasoning over long, multi-document contexts without requiring access to explicit intermediate-reasoning annotations. Third, while table-oriented questions reflect common scientific reading behavior, they primarily emphasize high-level semantic interpretation rather than fine-grained numerical or cell-level operations. Finally, our evaluation targets general-purpose models in zero-shot or light prompting settings, leaving task-specific adaptation for future work.

Ethics Statement

DocHop-QA is constructed from publicly available scientific articles in the PubMed Central Open Access corpus. The dataset does not contain personal, sensitive, or private information, and all source documents are used in accordance with their original licenses. The benchmark is designed for research purposes to study multimodal, multi-document reasoning and is not intended for clinical decision-making or medical advice. While the dataset is automatically generated with assistance from large language models, we apply concept-guided generation, filtering, and human validation to mitigate hallucinations and low-quality instances.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2025b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504.

Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. 2020. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zhoujun Cheng, Haoyu Dong, Ran Jia, Pengfei Wu, Shi Han, Fan Cheng, and Dongmei Zhang. 2022. Fortap: Using formulas for numerical-reasoning-aware table pretraining. *Preprint*, arXiv:2109.07323.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

Abhay Gupta, Kevin Zhu, Vasu Sharma, Sean O'Brien, and Michael Lu. 2025. Novelhopqa: Diagnosing multi-hop reasoning failures in long narrative contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26145–26162.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, page 2256–2264.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4083–4091, New York, NY, USA. Association for Computing Machinery.

Yunsoo Kim, Yusuf Abdulle, and Honghan Wu. 2025. Biohopr: A benchmark for multi-hop, multi-answer reasoning in biomedical domain. *arXiv preprint arXiv:2505.22240*.

Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. 2024a. M3sciqa: A multi-modal multi-document scientific qa benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15419–15446.

Ruosen Li, Zimu Wang, Son Tran, Lei Xia, and Xinya Du. 2024b. Meqa: A benchmark for multi-hop event-centric question answering with explanations. *Advances in Neural Information Processing Systems*, 37:126835–126862.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, page 74–81, Barcelona, Spain. Association for Computational Linguistics.

Teng Lin, Yuyu Luo, Honglin Zhang, Jicheng Zhang, Chunlin Liu, Kaishun Wu, and Nan Tang. 2025. Mebench: Benchmarking large language models for cross-document multi-entity question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1494.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

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

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myoung Kwon, and Edward Choi. 2023. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *Advances in Neural Information Processing Systems*, 36:66277–66288.

Vaishali Pal, Evangelos Kanoulas, Andrew Yates, and Maarten de Rijke. 2024. Table question answering for low-resourced Indic languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 75–92, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *arXiv preprint arXiv:2407.09413*.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A Rossi, and Dinesh Manocha. 2025. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6088–6109.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, and 1 others. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Jian Wu, Linyi Yang, Dongyuan Li, Yuliang Ji, Manabu Okumura, and Yue Zhang. 2025. Mmqa: Evaluating llms with multi-table multi-hop complex questions. In *The Thirteenth International Conference on Learning Representations*, page 1.

Weiwen Xu, Yang Deng, Huihui Zhang, Deng Cai, and Wai Lam. 2021. Exploiting reasoning chains for multi-hop science question answering. *arXiv preprint arXiv:2109.02905*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1192–1200. ACM.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Wonjin Yoon, Richard Jackson, Aron Lagerberg, and Jaewoo Kang. 2022. Sequence tagging for biomedical extractive question answering. *Bioinformatics*, 38(15):3794–3801.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. *arXiv preprint arXiv:2402.14116*.

Rongzhi Zhu, Xiangyu Liu, Zequn Sun, Yiwei Wang, and Wei Hu. 2025. Mitigating lost-in-retrieval problems in retrieval augmented multi-hop question answering. *arXiv preprint arXiv:2502.14245*.

A Table of Contents

Category	Subcategory	Location	Key Items (Purpose)
	How to prove the dataset construction pipeline?		
Pipeline & Construction	Question Concept Proposal	Sec. 3.1	Table 2 (Question Concept Type Examples)
	Document Pairing Strategy	Sec. 3.2, App. D.1	Alg. 1 (Document Selection Algorithm)
	Question Generation	Sec. 3.3, App. D.2	Fig. 1 (Pipeline Overview) Tables 8–10 (Question Generation Prompt)
	Answer Matching	Sec. 3.4, App. D.3	Table 11 (Filtering Keyword List)
	Is this dataset valid?		
Quality & Validation	Quality Assurance	Sec. 4, App. E, I	Tables 3, 13 (Inter-annotator Agreement Result) Figs. 7–8 (LLM-as-judge Result) Tables 28, 29 (LLM Evaluation Prompt) Fig. 26 (Human Evaluation Interface)
	What does the dataset cover?		
Diversity & Coverage	Dataset Coverage	Sec. 5.1, App. F.1	Fig. 2 (Dataset Statistics) Fig. 3 (Document Analysis) Figs. 9–19 (Question Keyword)
	Answer Dispersion & Reasoning Scope	Sec. 5.2, App. F.2	Table 14, Fig. 20 (Super Section Mapping Result) Figs. 4, 21, Table 15 (Section-level Distribution) Figs. 5, 22 (Answer Snippet Distance) Fig. 23 (Chain-hop Transition)
	Why is this dataset hard?		
Reasoning Analysis	Reasoning Complexity	Sec. 5.3	Table 4 (Reasoning Depth Comparison)
	What actually makes the task difficult?		
Controlled Experiment	Task 1: Generative Text Extraction	Sec. 6, App. G	Table 5 (Main Result) Table 16 (Full Result) Fig. 25 (Question Type Embedding Analysis)
	Task2: Structured Generative Answering	Sec. 6, App. G	Table 6 (Main Result)
	Task3, 4: BBox & XML Entity Index Extraction	App. G	Fig. 24 (Result Comparisons) Tables 17, 18 (Main Result)
	Ablation Analysis	App. H	Table 23 (Modality Ablation) Table 24 (Shortcut Ablation) Table 25 (Controlled Evidence Ablation) Table 26 (Human Answerability and Realism Analysis) Table 27 (Retrieval Ablation)
	Can this work be reliably reproduced and extended?		
Reproducibility & Implementation	Dataset and Code	App. B	https://shorturl.at/rqfcM
	License of Artifacts	App. B	–
	Reproducibility Cost	App. D.4	Table 12 (Computational Cost)
	Implementation Details	App. G.1, G.4	Tables 19– 22 (Hyperparameters),

Table 7: Structural overview of DocHop-QA.

B Data and Code Release

Link to Code and Dataset Full implementation and reproduction scripts for Dataset Construction (Section 3), Dataset Quality Assurance (Section 4), and Experiments and Results (Section 6) are at the following link. Access it here: https://shorturl.at/rqfcM

License for Artifacts All experiments are conducted using publicly available datasets and pretrained multimodal foundation models subject to their respective licenses and terms of use. We release the DocHop-QA dataset and the code necessary to reproduce the overall experimental procedures. Users are responsible for ensuring compliance with the licenses and usage restrictions associated with the underlying datasets and pretrained backbones.

C Related Work

Table-Oriented QA Several prior works explore table-centric QA, but they differ fundamentally from our setting. Pal et al. (2024) generate table QA for low-resource languages, yet their method relies on structured web pages and does not evaluate multi-hop or path-level reasoning. Likewise, FORTAP(Cheng et al., 2022) focuses on numerical reasoning over tabular data, but the framework lacks support for textual inference or multimodal evidence integration. More recently, Wu et al. (2025) propose MMQA, which evaluates multi-hop reasoning at the schema level over relational tables, rather than document-centric evidence integration. Our DocHop-QA targets document-centric multi-hop reasoning, where evidence spans heterogeneous elements such as paragraphs, tables, and visual layouts across multiple documents.

D Dataset Construction Details

D.1 Document Selection Details

As mentioned in Section 3.2, we formalized the semantic similarity-based document pairing strategy into a concrete algorithmic procedure, as illustrated in Algorithm 1. Candidate pairs are first identified using the following two filters: (1) at least one overlapping keyword, and (2) a TF-IDF cosine similarity score of at least 0.3[8]. After this first-stage filtering, we compute a total matching score for each remaining pair as a weighted combination of normalized keyword overlap and TF-IDF similarity. For each source document, we then kept only the top three scoring partners for downstream question generation, limiting repeated reuse of the same documents and promoting document diversity.

[8]After testing multiple thresholds (0.1–0.9) and manually inspecting retrieved documents, 0.3 was selected for consistently yielding the most relevant results.

Algorithm 1: **Document Selection**

Input: Source corpus $\mathcal{C} = \{d_1, \ldots, d_N\}$ (PubMed XMLs)
Parameter: Keyword extractor (YAKE) Φ ; TF–IDF vectorizer ψ ; similarity threshold τ ; weight α ; top-L
Output: Paired set $\mathcal{P} \subseteq \{(i,j) \mid 1 \le i < j \le N\}$

$\mathcal{P} \leftarrow \varnothing$; initialize $S_{ij} \leftarrow -\infty$ for all $i < j$
for $i \leftarrow 1$ **to** N **do**
  extract *pmc id*$_i$, *title*$_i$, *abstract*$_i$ from d_i
  fulltext$_i \leftarrow$ *title*$_i \parallel$ *abstract*$_i$
  $\mathcal{K}_i \leftarrow \Phi(\textit{fulltext}_i)$; remove stopwords
$\mathbf{X} \leftarrow \psi(\textit{fulltext}_1, \ldots, \textit{fulltext}_N)$
$\mathbf{M} \leftarrow \mathrm{cosine}(\mathbf{X}\mathbf{X}^\top)$
for $i \leftarrow 1$ **to** N **do**
  for $j \leftarrow i+1$ **to** N **do**
    $C_{ij} \leftarrow \mathcal{K}_i \cap \mathcal{K}_j$
    if $|C_{ij}| > 0$ **and** $\mathbf{M}_{ij} \ge \tau$ **then**
      $s_{\mathrm{kw}} \leftarrow \dfrac{\log(1+|C_{ij}|)}{\log(1+\max_k |\mathcal{K}_k|)}$
      $s_{\mathrm{sim}} \leftarrow \mathbf{M}_{ij}$
      $S_{ij} \leftarrow \alpha\, s_{\mathrm{kw}} + (1-\alpha)\, s_{\mathrm{sim}}$
  $\mathcal{N}_i \leftarrow$ Top-L indices j by S_{ij}
  for all $j \in \mathcal{N}_i$ **do**
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{(i,j)\}$
return $\mathcal{P}$

D.2 Question Generation Details

Instruction Tuning We presented few shot instances for three types of question generation: Non-Comparison, Comparison, and Non-Refer-Table (PST). Each instance follows a unified schema comprising a prompt, an instruction, two abstracts, and the resulting question. The prompt encapsulates the underlying question concept, the instruction clarifies how the concept should be instantiated, and Contexts 1 and 2 provide the input abstracts. Although the schema is identical across types, we train a distinct model for each category because our preliminary experiments revealed that type-specific models generate a higher proportion of valid questions than a single unified model. Representative examples of each question type are shown in Tables 8, 9, and 10.

Question Post-processing To ensure that generated questions strictly follow the intended question concepts and remain usable for downstream QA, we apply a lightweight post-processing step to filter invalid generations. We use the following criteria: (1) Generation failures: We remove instances

Prompt

Instruction: How does [ ] compare to [ ], and what are the key similarities and differences?

Chain-of-Thought (internal use only): Generate a question by replacing only the placeholder [ ] in the question concept below using document-specific information derived from the provided contexts. You must use two documents.

Instruction

How does [ ] compare to [ ], and what are the key similarities and differences?

Input : {

Context 1

`Regionalization of pediatric emergency care in the US. The 2006 IOM Report highlighted uneven care. Pediatric emergency services are concentrated in a limited number of children's hospitals and trauma centers.`

Context 2

`Current state of pediatric emergency care in Korea. A 2010 survey showed widespread unpreparedness: many EDs lacked essential pediatric equipment, and consultations were often handled by inexperienced residents.` }

Output

How do the characteristics of the healthcare systems in the United States compare to Korea, and what are the key similarities and differences?

Table 8: **An example of our dataset instance based on comparison-type multi-hop question generation.** Fields such as prompt, instruction, and context are shown in full.

Prompt

Instruction: What are the main challenges in [ ], and how can [ ] address these issues?

Generate a question by replacing only the placeholder [ ] in the question concept below using document-specific information derived from the provided contexts.

Chain-of-Thought (internal use only): Generate a question by replacing only the placeholder [ ] in the question concept below using document-specific information derived from the provided contexts. You must use two documents.

Instruction

What are the main challenges in [ ], and how can [ ] address these issues?

Input : {

Context 1

`Anaemia is a common haematological disorder in chronic disease, often complicating diagnosis and worsening quality of life and survival. This paper reviews its pathogenesis and potential therapeutic targets.`

Context 2

`Cloning the human EPO gene (1984) enabled ESA therapies for cancer-related anaemia. Trials show ESAs reduce transfusion needs, provide sustained correction, lower transfusion-related risks, and improve blood flow and quality of life.` }

Output

What are the main challenges in iron availability for red cell precursors in cancer-related anaemia, and how can erythropoiesis-stimulating agents address this issue?

Table 9: **An example of our dataset instance based on Non-comparison type multi-hop question generation.** Fields such as prompt, instruction, and context are shown in full.

where the model produces empty outputs or outputs that do not conform to the expected format (e.g., malformed JSON fields or missing required components). (2) Question–concept mismatch: We filter out questions that fail to instantiate the target concept, operationalized by missing the required concept keywords/placeholders specified in the corresponding prompt. This step prevents concept drift and enforces consistency between the intended reasoning structure and the generated question. After applying these filters, we retain only question–context pairs that pass all checks, forming the final dataset used in our experiments.

D.3 Answer Matching Details

In DocHop-QA, multi-hop reasoning follows two paradigms. Fan-hop aggregates semantically complementary information from multiple, independently retrieved documents. Each snippet contributes to the final answer without explicit cross-referencing. On the other hand, Chain-hop performs step-by-step inference where information in one document directly supports content in another, often forming a logical link-such as when a table summarizes results from an external study.

Paragraph-Oriented Type To ground each QA instance with relevant context, we parsed the full-text files of both documents. The subsection-level content was extracted, and segments with fewer than two sentences were discarded to avoid sparse inputs. For each segment, we computed seman-

Prompt

Instruction: What are the main challenges in [ ], what strategies can be implemented to address them? Is there any table additionally describe it? Which table is it and what is the main point of the table?

Chain-of-Thought (internal use only): Generate a question by replacing only the placeholder [ ] in the question concept below using document-specific information derived from the provided contexts. You must use two documents.

Instruction

What are the main challenges in [ ], what strategies can be implemented to address them? Is there any table additionally describe it? Which table is it and what is the main point of the table?

Input : {

Context 1

`Inflammatory diseases, including rheumatic disorders, are a major global health burden. While multiple treatments exist, long-term steroidal or non-steroidal therapy can cause serious adverse effects.`

Context 2

`A randomized, double-blind, placebo-controlled trial enrolled 80 UC patients in remission (Sep 2009-Oct 2010). Participants received SM (140 mg/day) or placebo alongside standard therapy for six months.` }

Output

What are the main challenges in traditional anti-inflammatory in treating inflammatory diseases, what strategies can be implemented to address them? Is there any table additionally describe it? Which table is it and what is the main point of the table?

Table 10: **An example of our dataset instance based on PST type multi-hop question generation.** Fields such as prompt, instruction, and context are shown in full.

tic similarity to the question using a hybrid scoring method: a weighted combination of TF-IDF (weight 0.2) and BERT-based(Zhang et al., 2019) sentence embeddings (weight 0.8). This configuration prioritized deeper semantic understanding while preserving term-level sensitivity. Similarity scores were aggregated across both documents, and the top 5 segments were retrieved. To ensure quality, only answer instances with all retrieved snippets above a similarity score of 0.4 were retained. This threshold was empirically selected based on preliminary validation, which showed that higher similarity was strongly associated with answer plausibility. All filtered QA-context pairs were saved in structured JSON.

Table-Oriented Type We employed two strategies depending on the snippet's role: some snippets were retrieved via semantic-similarity scoring augmented with keyword-based filtering, whereas others were extracted using predefined logic. To support the semantic-similarity route, we curated three keyword lists targeting problem-, solution-, and limitation-related paragraphs (Table 11); these lists served as an initial paragraph filter in each context and were then combined with semantic-similarity scoring to select answer snippets. If no paragraphs are filtered we computed similarity scores between questions and all paragraphs. We designed logical answer chains tailored to each QA type (PST Type and TS/TSL Type), which are explained as follows. Green-colored snippets used the semantic similarity strategy, while others were retrieved via predefined logic.

Category	Keywords
Problem-related	burden, challenge, gap, unmet need, lack of, insufficient, controversy, uncertainty, poor outcomes, high prevalence, incidence, rising rates, increasing trend, public health concern, problem, barrier, limitation, disparity, inequity, inaccessibility, misdiagnosis, delay, complication, risk factor, vulnerability, deficiency, inconsistency, overuse, underuse, shortage, low adherence, noncompliance, failure, inadequacy, ineffectiveness, complexity, fragmentation, obstacle, hazard, disease burden
Solution-related	approach, strategy, intervention, treatment, therapy, model, framework, tool, instrument, protocol, program, solution, innovation, novel, effective, efficacy, improvement, enhancement, implementation, guideline, practice, plan, policy, recommendation, management, design, structure, evaluation, algorithm, process, assessment, mechanism, procedure, initiative, rollout, deployment, trial, interdisciplinary, collaborative, multimodal, integrated, systematic, scalable, feasible, evidence-based, optimization
Limitation-related	limitation, limit, lack, weakness, shortcoming, bias

Table 11: Keyword categories used for paragraph filtering during semantic retrieval.

(1) PST Type

- **Snippets 1 & 2**: Problem and Solution paragraphs were keyword filtered. The table is assumed to be in the solution document; the problem document provides motivation. We then selected the most relevant paragraph per document using 0.2 * TF-IDF + 0.8 * BERTScore.

- **Snippet 3**: The most relevant table content, selected based on similarity to Snippet 2.
- **Snippet 4**: Paragraph in the document referring the table.

(2) TS & TSL Type

- **Snippet 1**: Table metadata from the source document.
- **Snippet 2**: Paragraph referring to the table.
- **Snippet 3**: Abstract of the referenced document, linked via citation markers in the table.
- **Snippet 4**: Paragraph describing the cited document's limitations. Snippet 4 was retrieved via a query-driven similarity search using the prompt "What is the limitation of this paper?" and subsequently keyword-filtered. If the similarity score for snippet 4 fell below 0.4, the snippet was discarded, and the QA instance was reclassified as TS type, containing only the first three snippets.

D.4 Computational Cost Details

We measured generation and matching costs to assess the computational scalability of our pipeline. As summarized in Table 12, the measured question/answer generation time and GPU usage indicate that once prompts are prepared, scaling up the QA set can be done efficiently with moderate resource requirements.This demonstrates that the proposed LLM-driven pipeline is computationally scalable for large-scale QA dataset construction.

Question Type	Time per Q gen	GPU peak usage (Q gen)	Time per A matching	GPU peak usage (A matching)
Comparison	6.84s	11.82GB	5.34s	0.99GB
Non-comparison	2.22s	12.10GB	—	—
PST	12.56s	14.89GB	0.05s	15.38GB
TSL	24.03s	—	4.45s	—

Table 12: **Computation efficiency of QA generation and matching.** Time and GPU usage measured per question (Q gen) and answer matching (A matching) step for each question type. The results show that once prompts are prepared, QA scaling can be efficiently achieved with moderate GPU memory usage.

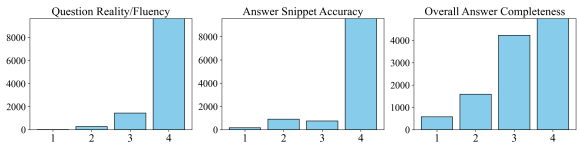


Figure 7: **Qwen evaluation across the full dataset**

E Dataset Quality Assurance

E.1 LLM-based Dataset Quality Assurance

As shown in Figure 7, the full dataset achieves consistently high scores across all criteria. Among question types, Paragraph-Oriented QA performs particularly well, with most scores rated 3 or 4 (Figure 8). Some instances had lower accuracy, likely due to snippets covering only parts of the question, reducing alignment with its overall intent. Table-Oriented QA showed slightly lower Reality/Fluency due to its format, but still maintained high overall quality. These results confirm that DocHopQA exhibits robust quality and is suitable for complex QA tasks. All three LLM judges (Qwen3.5-9B, GPT-4o, and Gemini2.5 Flash) were prompted with the same evaluation criteria, output format, and few-shot examples (Table 28, Table 29).

E.2 Human-based Dataset Quality Assurance

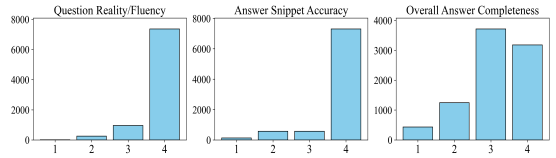
Annotator Demographics Fifteen annotators participated in the human evaluation.The sample was gender-balanced (7F/8M) and included 6 undergraduates, 6 master's, and 3 doctoral students. The participants had academic backgrounds in business (n=3), industrial and management engineering (n=4), computer engineering unrelated to AI (n=2), and AI-related computer science (n=6). The age distribution was as follows: 20 years (n = 8), 30 years (n = 5) and older than 50 years (n = 2).

Institutional Review Board Statement The study was approved by the Institutional Review Board (Exemption Number: PIRB-2025-E020, Date: June 9, 2025).

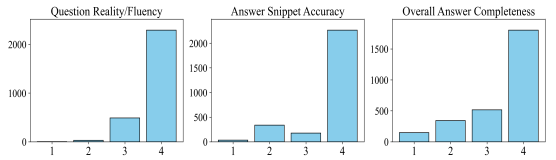
Evaluation Interface Annotators rated each QA pair via a Google Form interface (Figure 26). The form included Likert-scale items for three evaluation criteria and mandatory comment fields for low ratings, ensuring consistent evaluation and facilitating both quantitative and qualitative feedback.

E.3 Human-LLM Agreement Analysis

To further assess the alignment between human judgments and automated evaluators, we report Spearman's ρ and Krippendorff's α on the 50-instance subset. Table 13 summarizes the agreement across three evaluation dimensions. We observe moderate agreement between human annotators and LLM evaluators, with variation across



(a) Paragraph-Oriented



(b) Table-Oriented

Figure 8: **LLM based Quality Assurance result**, (1) Question Reality/Fluency (2) Answer Snippet Accuracy (3) Overall Answer Completeness. Scores are based on a 4-point scale (1:poor, 4: excellent).

Evaluation	Model	Spearman's ρ	Krippendorff's α
Completeness	GPT-4o	0.354	0.199
	Gemini-2.5	0.606	0.382
	Qwen3.5-9B	0.037	0.015
Accuracy	GPT-4o	0.240	0.144
	Gemini-2.5	0.395	0.285
	Qwen3.5-9B	0.027	-0.441
Reality	GPT-4o	-0.163	-0.160
	Gemini-2.5	0.445	0.000
	Qwen3.5-9B	0.099	0.113

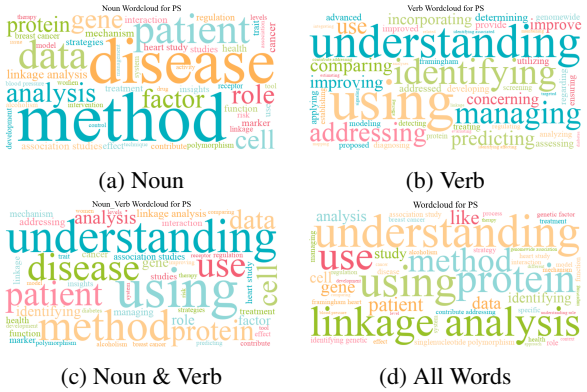
Table 13: Human–LLM agreement measured by Spearman's ρ and Krippendorff's α on a 50-instance subset.

models and criteria. For the *Reality* dimension, agreement scores are low due to a ceiling effect, where most instances receive high scores, resulting in low variance and unstable statistics. Beyond evaluation scores, we manually verified whether automated snippet matching correctly identifies human-identified supporting evidence. In most cases, retrieved snippets align well with human annotations, indicating limited noise from automated matching. These results suggest that automated evaluation can approximate human judgment when appropriately calibrated, though limitations persist under low variance conditions.

F Dataset Analysis Details

F.1 Dataset Coverage

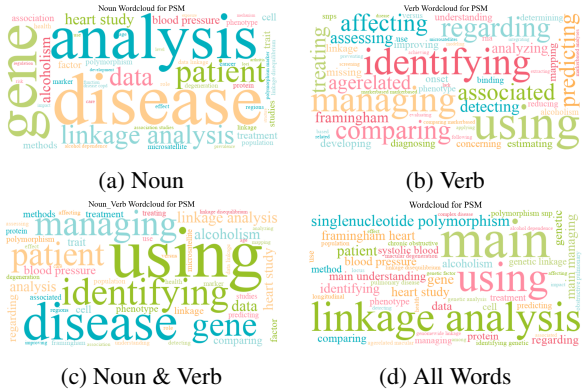
Keyword Analysis To validate the topical diversity and domain coverage of DocHop-QA, we conducted word cloud analysis on question texts across all 11 question concepts. After removing stopwords and terms specific to each question concept, we extracted nouns, verbs, and combinations of nouns and verbs from each question. Word clouds were generated per concept (Figures 9 to 19). As expected from data sourced from PubMed, biomedical terms such as *patient*, *gene*, and *cell* dominate across concepts, confirming domain specificity. Importantly, each concept exhibits distinct keyword distributions. Concepts without comparison focus on mechanistic terms, while comparison concepts surface contrastive terms such as *compare* and *associated*. Together, these patterns confirm that DocHop-QA achieves both broad topical coverage across biomedical subfields and structural diversity across reasoning types.



(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

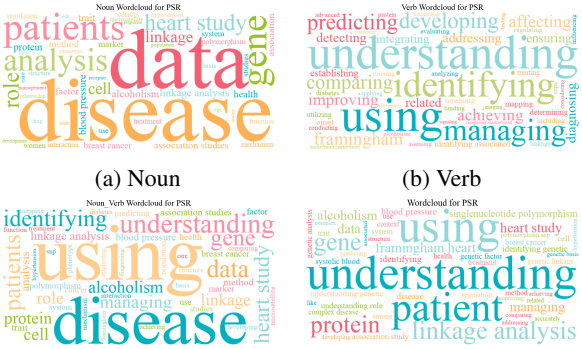
Figure 9: **Frequent words in PS-type questions:** Showing terms like *"method"* and *"disease,"* indicating conceptual understanding in biomedical topics.



(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

Figure 10: **Frequent words in PSM-type questions:** Showing terms like *analysis*, *affecting*, and *identifying*, highlighting analytical reasoning about genetic mechanisms and disease traits.



(a) Noun (b) Verb

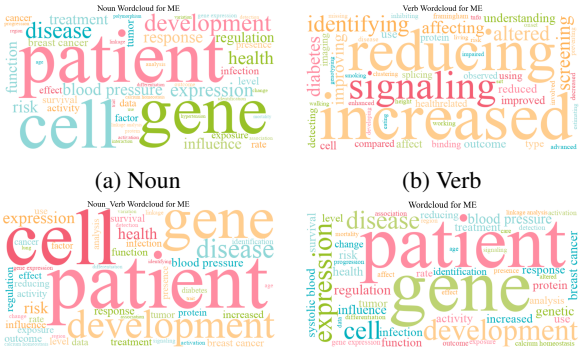
(c) Noun & Verb (d) All Words

Figure 11: **Frequent words in PSR-type questions:** Suggesting a focus on result-oriented reasoning about biomedical solutions or interventions.

(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

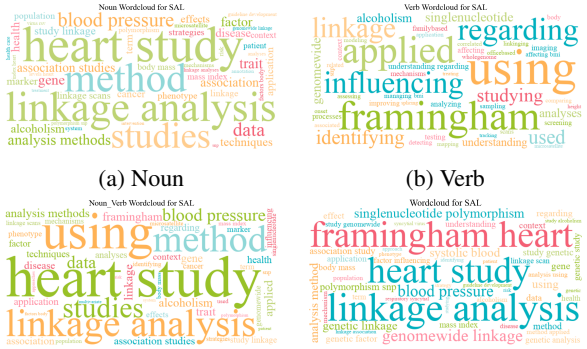
Figure 12: **Frequent words in PSL-type questions:** Showing terms like *conducting*, *addressing*, and *diagnosing*, highlighting evaluation limitations and follow up aspects of solutions.



(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

Figure 13: **Frequent words in ME-type questions:** Showing terms like *increased*, *reducing*, and *signaling*, indicating a focus on causal mechanisms and biological effects.



(a) Noun (b) Verb

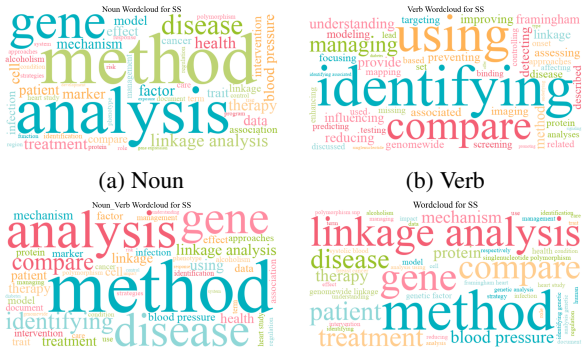
(c) Noun & Verb (d) All Words

Figure 14: **Frequent words in SAL-type questions:** Showing terms like *heart* and *blood*, indicating strengths and limitations of biological studies.

(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

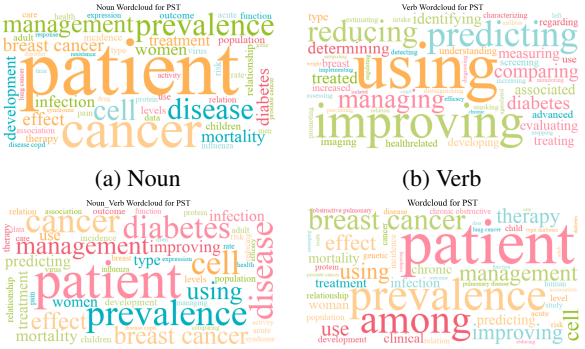
Figure 15: **Frequent words in FF-type questions:** Showing terms like *associated*, *effect*, and *mechanism*, indicating functional comparisons between two entities.



(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

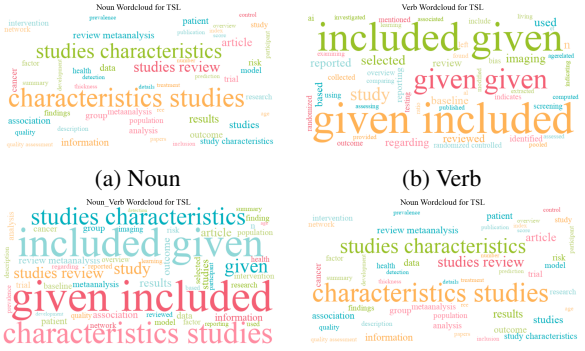
Figure 16: **Frequent words in SS-type questions:** Showing terms like *compare*, *method*, and *analysis*, reflecting contrast of scientific approaches to biomedical problems.



(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

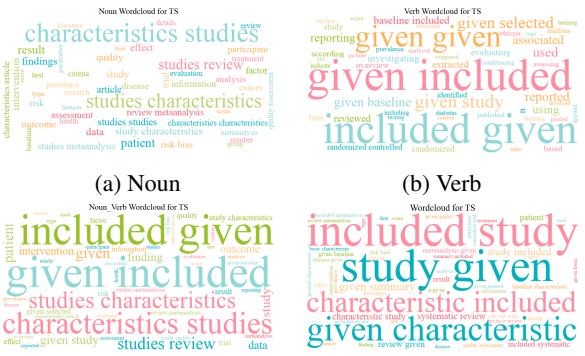
Figure 17: **Frequent words in PST-type questions:** Showing terms like *predicting*, *patient*, and *infection*, highlighting challenges and strategies based on tabular data.



(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

Figure 18: **Frequent words in TSL-type questions:** Showing terms like *given*, *included*, and *characteristics*, reflecting their deterministic, caption-based generation with limited lexical variety.



(a) Noun (b) Verb

(c) Noun & Verb (d) All Words

Figure 19: **Frequent words in TS-type questions:** Showing terms like *given*, *included*, and *study*, reflecting their deterministic, table-driven design with limited lexical diversity.

Super-Section	Detail Section Title
Abstract	**abstract**, etc.
Introduction	**introduction**, etc.
Background	**background**, **background and recent developments**, background and overview, etc.
Method	**method(s)**, material(s) and method(s), method and implementation, etc.
Results	**result(s)**, **results and discussion**, implementation and results, result analysis, etc.
Discussion	**discussion**(s), **discussion and conclusions**, limitations, summary and discussion, etc
Conclusion	**conclusion(s)**, conclusions and future directions, conclusion and perspectives, etc.
Others	case report, **review**, dataset and evaluation strategy, **abbreviations**, etc.

Table 14: **Detailed results of Super-Section Mapping**, with frequent section titles bolded.

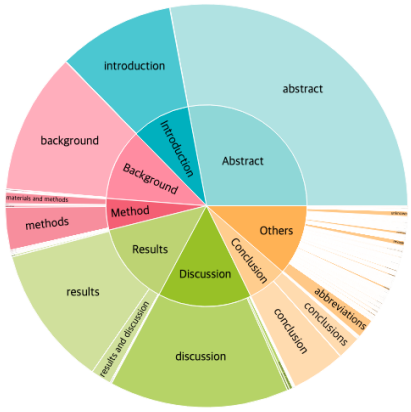


Figure 20: **Section Mapping Results.**

F.2 Answer Dispersion

Super-Section Mapping Analysis To facilitate section-level analysis, we standardized a diverse set of original section titles into 8 unified Super-Section tags. This mapping aimed to reduce the variation caused by inconsistent section titles (e.g., “Materials and Methods", “Subjects and Methods", “Methodology") and to improve the interpretability of section-based statistics. The results of standardization are described in Table 14 and Figure 20.

Section-level Answer Dispersion Figure 21 presents the distribution of answer snippet across QA types. Paragraph-Oriented types predominantly retrieved answers from the Abstract, aligning with their fan-hop retrieval strategy. In contrast, Table-Oriented types mostly extracted snippets from the

Results and Discussion sections, reflecting the presence of structured, interpretive content found in them. These patterns highlight that each Q type draws from distinct sections of the document, underscoring the need for models to understand section-level semantics when reasoning over scientific papers.

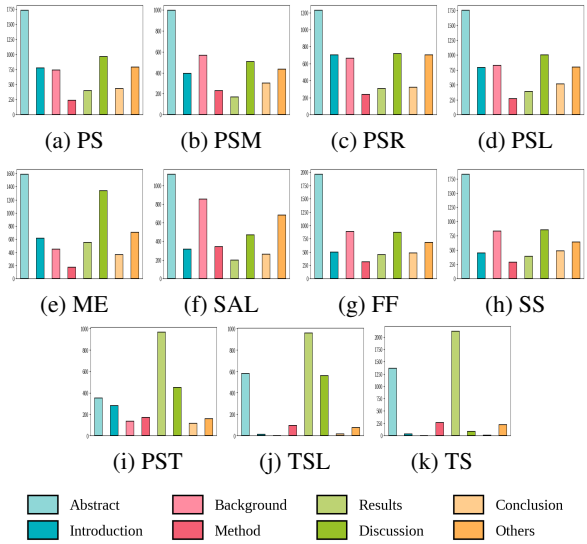


Figure 21: **Section-level distribution of answer spans across the 11 Q types.**

Table Oriented Section Analysis We analyzed the origin of tables (Table 15) and found that most were extracted from the *Results* section. This supports our decision to include only documents with at least one table in result section for PST. TSL, TS tables also mainly came from Results, reflecting their common role in summarizing outcomes across studies. These patterns show that the dataset aligns with typical scientific reporting, where key findings are often presented in results tables.

Section	PST	TS	TSL
Abstract	5	3	3
Introduction	1	16	5
Background	1	2	0
Method	33	132	47
Results	**830**	**1057**	**478**
Discussion	1	41	4
Conclusion	0	2	1
Others	3	112	35

Table 15: **Distribution of Table Origins Across Table-Oriented types** (PST, TS, TSL) with Section Mapping

Structural Distance Analysis We quantify how answer snippets are distributed across the document structure to verify that our 11 question concepts yield diverse, non-position-biased snippet layouts (Figure 22. While fan-hop and chain-hop characterize two complementary reasoning patterns based on answer snippet aggregation and dependency structure, the underlying 11 question concept types span these patterns and induce distinct structural retrieval and traversal behaviors. For **chain-hop**, we concatenate the two documents into a single sequence to account for cross-document distances, and compute the snippet-to-snippet *distance* between consecutive steps based on snippet's position. We report distances at the element-wise and as an instance-wise average. All box plots retain only values within the interquartile range (IQR; 25th–75th percentile).

Chain-Hop Transition Analysis To analyze the structural diversity of chain-hop reasoning in DocHop-QA, we visualize ground-truth hop transitions across documents and document sections using a Sankey-style representation shown in Figure 23. This analysis reveals that chain-hop questions exhibit diverse hop patterns, spanning document sections such as abstracts, background, methods, results, and conclusions, often across different documents. Notably, table entities appear not only as terminal answer evidence but also as intermediate nodes within reasoning chains, indicating that tables frequently serve as both sources and bridges of information.

These observations highlight that DocHop-QA chain-hop questions require models to perform structured, dependency-sensitive reasoning across heterogeneous document components, rather than relying on linear or section-local inference. The diversity of hop transitions underscores the need for multi-document understanding that integrates both textual and tabular evidence in a coherent reasoning process.

G Additional Experiments and Results

G.1 General Preprocessing

All documents are converted from PDF to XML and image formats. OCR and table detection are applied to extract layout-aware content for entity extraction tasks. We use a 70/10/20 train/dev/test split, ensuring that each document appears in only one split to prevent information leakage.

At the document level, we first extracted structured entities, including titles, abstracts, paragraphs, figures, and tables, from the XML version. For visual grounding, we align these XML entities with

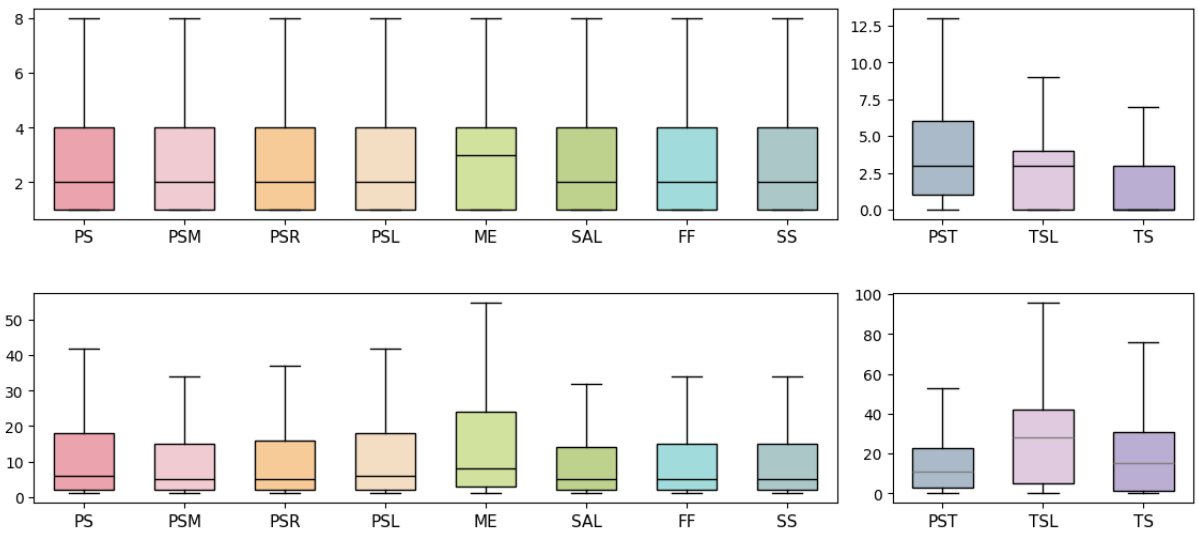


Figure 22: **Structural evidence distribution across question concept types.** We show section-level (top) and paragraph-level (bottom) positions of answer snippets for each concept. Fan-hop questions exhibit broadly distributed positions within documents, indicating global aggregation, whereas chain-hop questions show progressive structural between snippets, reflecting ordered multi-step reason-chaining processes. Notably, PSM, ME tend to retrieve evidence from relatively localized yet deeper structural regions, while PST, TSL, TS span wider structural ranges, often bridging descriptive and analytical sections. Overall, the diverse structural ranges covered across question concept types highlight the heterogeneity of reasoning difficulty and retrieval behaviors encoded in our taxonomy.

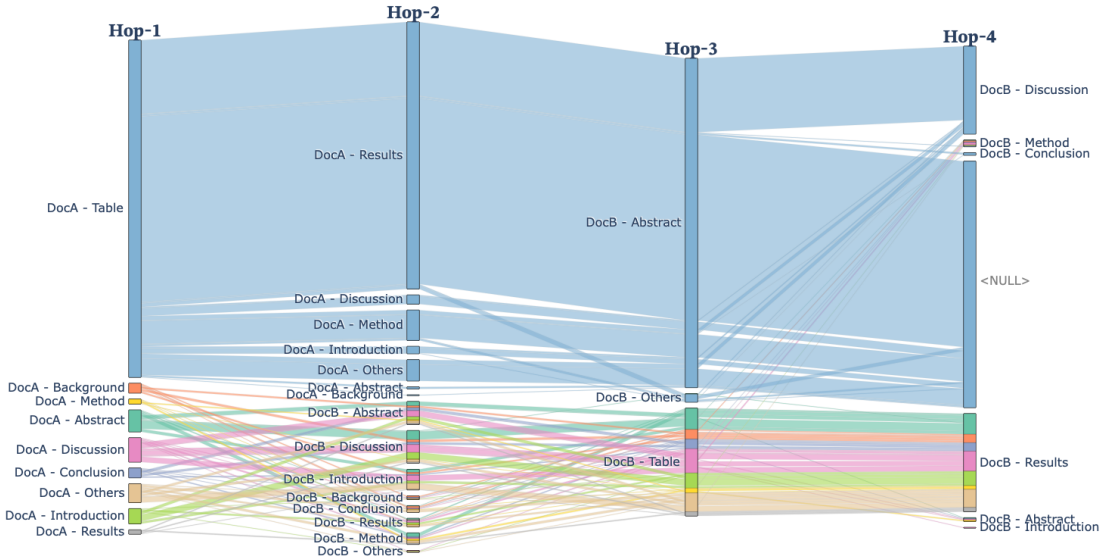


Figure 23: **Ground-truth chain-hop transitions across documents and sections (Sankey Diagram).** Each node denotes the document-section location of an answer snippet at each hop, and links represent hop-to-hop transitions in the ground-truth chain. Tables appear not only as endpoints but also as intermediate bridges. This diversity highlights the structural complexity of multi-document reasoning required by DocHop-QA

their corresponding positions in the PDF using the PaddleX OCR toolkit, and save each page of the PDF as an image. This enables us to generate both text-based and image-based inputs across tasks.

For entity-based tasks, at the sample level, we combined all entities from the relevant context documents into a unified list, assigning global indices. For instance, if Document 1 has 10 entities and Document 2 has 32, entities are indexed sequentially from [Entity 0] (e.g., <doc1_title>) to [Entity 41] (e.g., <doc2_paragraph20>). Answer labels are represented as index lists referencing this combined

sequence. For models that are limited to one input image, such as LayoutLMv3, all pages from the context documents are merged into a single image and resized to comply with the model's input constraints.

For fine-tuning, all classification-based models are trained with binary cross-entropy loss with the AdamW optimizer for up to 50 epochs. Early stopping is applied with a patience of either 5 or 10 epochs, based on validation performance measured by sample-level recall. Training of open-sourced models is conducted locally using a single NVIDIA A100 GPU. Close-sourced model inference are executed via API calls.

G.2 Additional Task Descriptions

Task 1: Generative Reasoning We test generative LLMs under three prompting strategies (question-only, zero-shot, one-shot) and three input types (text-only, image-only, text+image). Models capable of full-document ingestion, such as Gemini-2.5-Flash, Gemini-3-Pro, and GPT-4o, are given full XML and PDF files. Generated answers are evaluated with BLEU and ROUGE (1, 2, L, Lsum) against synthesized gold answers from reference context snippets using GPT-4.1.

Task 2: Structured Generative Answering This task uses generative models InternVL2-4b, Qwen2.5-VL-7B-Instruct, and GPT-4o to produce lists of entity indices. Inputs include labeled XML entities and document images. Prompts follow zero-shot or one-shot formats. To avoid memory issues, we cap the number of entities and pages. Evaluation includes multi-label accuracy (Belong, Contain, and Overlap) and sample-level F1 based on spatial containment.

Task 3: BBox Entity Index Extraction Formulated as a discriminative, multi-label classification task, this evaluates how well models can identify the relevant answer-contained bounding boxes (e.g., paragraphs, figures, tables) in OCR document layouts. We first map each XML entity to one or more OCR-detected bounding boxes, which are compiled into a single 1000x1000 image per document set. We use two layout-aware vision-language models, LayoutLMv3 (Huang et al., 2022) and LayoutXLM (Xu et al., 2020), that receive bounding box coordinates, associated OCR texts, and document images, and predict relevant bounding box indices. Entities that cannot be matched through OCR extraction are marked with index 0. The evaluation mirrors the metrics used in Task 2.

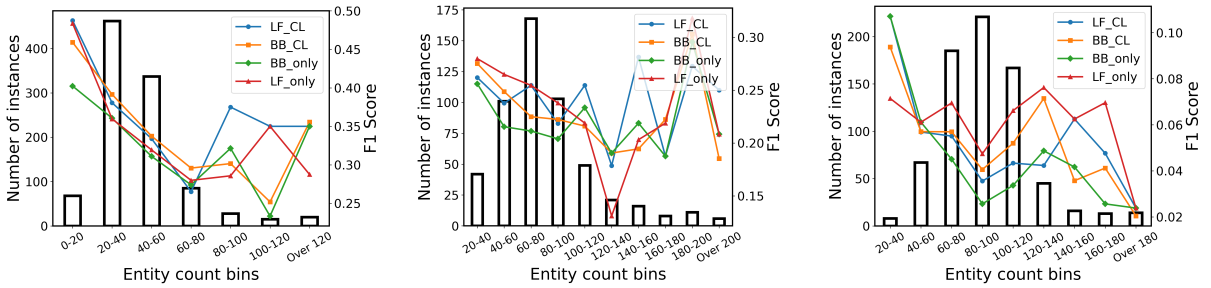
Task 4: XML Entity Index Extraction This task further investigates multi-label entity classification without dependencies on bounding box extractions but instead, predicts indices directly from XML-defined entities, assessing how well models can reason over text-structured representations of documents. We examine two setups: 1) Concatenated inputs, where the question and all XML entities are fed as one sequence, and 2) Per-entity inputs, where text and optional CLIP-based image embeddings are pooled for each entity. Models tested include BigBird (Zaheer et al., 2020) and Longformer (Beltagy et al., 2020). Metrics used are identical to Task 2.

G.3 Additional Result Analysis

Results of Task 1: Generative Reasoning As shown in Table 16, the best performance is from Gemini3-Pro in a one-shot setting using text and image inputs with full-document ingestion, which achieves a BLEU score of 31.8 and strong performance across ROUGE-L (34.5). This suggests that modern large vision-language models are more capable of locating relevant information and reasoning across multiple and complex document images and textual content.

Comparing input-processed architectures, Qwen2.5 and Qwen3 consistently outperforms InternVL2 in setups that utilize image input. This advantage likely stems from Qwen's greater input capacity, which allows it to process up to 100 document pages, while InternVL2 is limited to 10 with the simplest configuration. As a result, the Qwen models are better suited for large-scale document understanding, particularly when full visual context is required. However, InternVL2 demonstrates better performance when using text-only inputs compared to Qwen2.5, achieving a BLEU of 12.65 and ROUGE-2 of 11.11, which slightly surpasses Qwen2.5's text-only setup (BLEU: 9.66, ROUGE-2: 10.93). This implies that InternVL2's language modeling capabilities are competitive when visual grounding is not required, and reaffirms the importance of architecture-modality alignment.

Interestingly, Qwen2.5 performs best when using image input alone, and worst when using only text. This indicates that Qwen2.5 is likely optimized for visual reasoning and may struggle to integrate text-only signals in the absence of layout structure. However, adding both image and text did not significantly boost performance, suggesting



(a) Paragraph & Single (b) Paragraph & Multi (c) Table & Multi

Figure 24: **Task 3: Entity Count vs. F1.** Mean F1 per entity-count bin (lines) with bin sizes (bars). The top row reports results over all instances, while the bottom row excludes zero-F1 cases. Performance generally degrades as entity counts increase, with a more pronounced effect for paragraph-oriented settings.

Model	Setup	Text	Image	BLEU	Rouge1	Rouge2	RougeL	RougeLSum
InternVL	Question	-	✓	6.22	30.02	7.24	15.95	19.04
InternVL	Zero-shot	-	✓	5.54	28.67	7.03	14.62	19.46
InternVL	One-shot	-	✓	6.39	28.74	6.99	16.12	18.50
InternVL	Question	✓	-	11.20	32.69	10.02	18.49	20.32
InternVL	Zero-shot	✓	-	12.65	33.36	11.11	18.93	21.40
InternVL	One-shot	✓	-	14.53	32.11	11.99	19.56	20.91
InternVL	Question	✓	✓	6.91	30.09	8.05	15.85	19.79
InternVL	Zero-shot	✓	✓	6.91	30.06	8.48	15.55	20.60
InternVL	One-shot	✓	✓	7.48	31.23	8.55	16.46	20.26
Qwen2.5	Question	-	✓	15.06	36.64	13.43	20.79	23.89
Qwen2.5	Zero-shot	-	✓	21.56	38.45	16.92	23.90	25.84
Qwen2.5	One-shot	-	✓	12.40	34.48	11.49	19.50	22.53
Qwen2.5	Question	✓	-	7.77	28.62	9.33	15.71	20.44
Qwen2.5	Zero-shot	✓	-	9.66	30.81	10.93	17.28	21.96
Qwen2.5	One-shot	✓	-	7.32	27.66	8.30	15.45	19.44
Qwen2.5	Question	✓	✓	9.80	31.89	10.70	17.82	22.12
Qwen2.5	Zero-shot	✓	✓	12.78	33.69	12.86	19.80	23.66
Qwen2.5	One-shot	✓	✓	9.45	31.35	9.96	17.94	21.40
Qwen3	One-shot	✓	-	26.00	-	-	28.80	-
Qwen3	One-shot	-	✓	18.20	-	-	21.00	-
Qwen3	One-shot	✓	✓	29.7	-	-	32.1	-
Gemini2.5	Question	✓	-	7.34	28.26	10.91	15.28	20.44
Gemini2.5	Zero-shot	✓	-	9.22	29.24	11.14	15.89	20.93
Gemini2.5	One-shot	✓	-	11.94	24.55	12.85	16.81	19.41
Gemini2.5	Question	-	✓	7.57	28.76	10.71	15.41	20.92
Gemini2.5	Zero-shot	-	✓	9.03	29.62	12.08	16.20	21.88
Gemini2.5	One-shot	-	✓	11.96	25.47	10.96	15.33	18.68
Gemini2.5	Question	✓	✓	7.38	28.85	10.55	15.15	20.46
Gemini2.5	Zero-shot	✓	✓	8.89	30.70	12.24	17.07	22.63
Gemini2.5	One-shot	✓	✓	11.71	24.57	12.61	16.30	18.95
Gemini3	One-shot	✓	✓	**31.8**	-	-	**34.5**	-
GPT4o	Question	-	✓	24.80	41.25	19.46	26.85	29.92
GPT4o	Zero-shot	-	✓	26.45	43.16	21.05	28.63	31.55
GPT4o	One-shot	-	✓	28.92	**45.79**	**23.81**	31.26	**34.12**

Table 16: **Comprehensive results for generative text extraction task** using InternVL2, Qwen 2.5, Qwen 3, Gemini 2.5, Gemini 3, and GPT-4o across different prompt setups and input configuration. We measure performance through BLEU and ROUGE metrics.

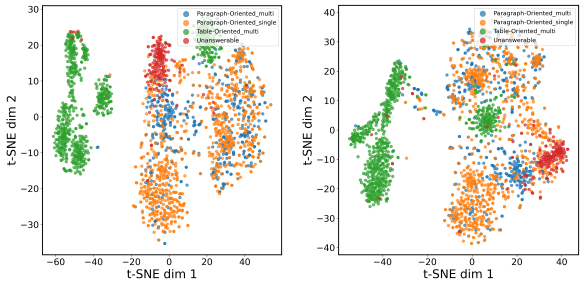
that naive concatenation of modalities may introduce redundancy or noise, and that more sophisticated cross-modal fusion strategies are needed for optimal performance on DocHop-QA. The more recent Qwen3, however, outperforms it's predecessor with better visual and textual alignment performing best (BLEU: 29.7, ROUGEL: 32.1) with both text and image inputs.

Prompt design plays a critical role for input processed generative QA performance. For Qwen2.5, zero-shot prompting consistently outperforms one-shot setups, likely due to the substantial length of context documents. Including a one-shot example in the prompt consumes valuable input space, which forces truncation of relevant information, especially under token limits. This finding highlights an important trade-off in prompting long-context models: example-rich prompts can backfire in high-volume settings like full-document QA.

We also evaluated Gemini 2.5 and Gemini 3

across XML and PDF inputs and GPT-4o with PDF inputs. Unlike InternVL2, Qwen2.5, and Qwen3, these models were accessed via API and thus were not bound by local memory constraints, allowing full document inputs without severe truncation. As a result, one-shot prompting performed best across most configurations, with top BLEU score from Gemini3 of 31.8 using whole XML & PDF input. Gemini3 outperforms any other models significantly, demonstrating its superior ability in document understanding and multimodal QA-based text extraction. Comparing to Qwen3-VL, which is one of the latest vision-language models Gemini3 also outperforms its results (BLEU: 29.7). These results illustrate the potential of cloud-based LLMs for processing full-document QA tasks without aggressive input filtering, further expanding the applicability of DocHop-QA in real-world settings.

In summary, this task underscores how DocHop-QA enables robust benchmarking of free-form QA grounded in multimodal, multi-page and cross-document setups. Users can evaluate generative models under various prompting and input modality combinations, uncovering limitations in memory, alignment, and format-following behavior.



(a) Image-only input (b) Text-only (XML) input

Figure 25: **Task 1: t-SNE of Qwen last layer hidden representations** Visualization under image-only and text-only (XML) inputs, analyzing how LLMs and VLMs differentiate question types and contextual complexity. Table and paragraph-oriented instances form separable clusters in both settings, and unanswerable cases show a distinct grouping, suggesting modality-consistent internal structuring of question types.

Embedding Analysis for Question Types We analyze the last-layer hidden states of the Qwen model using t-SNE in Figure 25 to understand how LLMs and LVLMs internally differentiate question types and context complexities. Clear separation emerges between table-oriented and paragraph-oriented questions for both text-only and image-only inputs, suggesting that the model can distinguish question types at a coarse level, likely driven by surface-level cues in question formulation rather than explicit tabular reasoning. This indicates an awareness of question format and required evidence type, but does not imply effective reasoning over tabular content, which is consistent with the substantially lower performance observed on table-oriented tasks. Additionally, we identify a distinct cluster corresponding to unanswerable questions, demonstrating that recent LLMs can partially detect answerability without fine-tuning. However, this separation is less pronounced for paragraph-oriented and fan-hop questions, where partial answers may be available but full reasoning chains cannot be completed. This observation reflects the inherent difficulty of DocHop-QA unanswerable cases and highlights the tendency of dataset-generation LLMs to introduce hallucinated intermediate hops. Overall, the similarity of clustering patterns across text-only and image-only inputs suggests consistent internal representations across modalities, while reinforcing the need for more robust reasoning and answerability-aware modeling.

Results of Task 2: Structured Generative Answering We prompt generative vision-language models (InternVL2, Qwen2.5-VL, GPT-4o) to output structured entity indices. Inputs include labeled entity texts and optionally, document page images. Prompts follow zero-shot or one-shot formats. Table 6 shows that GPT-4o in the zero-shot, text-only configuration yields the highest F1 (16.21). Visual input degrades performance due to layout misalignment. One-shot prompts reduce input space, harming performance on long documents. This task highlights the potential of LLMs for structured reasoning, but also the limits of current multimodal alignment.

Results of Task3: BBox Entity Index Extraction This task features multi-label classification over OCR-extracted bounding boxes. As shown in Table 17, LayoutLMv3 significantly outperforms LayoutXLM in overlap-based accuracy (95.52%) and F1 (12.35), showing its robustness in spatially grounding questions. However, full containment accuracy is near zero, suggesting difficulties in aligning complex layouts. These findings show that DocHop-QA is a valuable benchmark for pushing the limits of document layout understanding in QA settings. It reveals critical challenges in handling long, multi-page, multi-document inputs, while of-

Model	Sample			Accuracy		
	F1	Re	Pr	Be	Co	Ov
LayoutLMv3	**12.35**	34.97	8.47	0.00	0.04	95.52
LayoutXLM	7.13	4.49	24.06	15.64	0.00	33.96

Table 17: **Results on BBox Entity Index Extraction** reporting sample-level F1, Recall (Re), and Precision (Pr), along with spatial accuracy: Belong (Be), Contain (Co), and Overlap (Ov), for LayoutLMv3 and LayoutXLM.

Setup	Text	Img	Sample			Accuracy		
			F1	Re	Pr	Be	Co	Ov
Concat	BB	–	23.18	18.81	30.45	4.22	0.09	62.21
Concat	LF	–	21.83	20.09	24.12	0.22	0.00	66.17
PerEnt	BB	–	21.25	17.22	27.94	2.07	0.09	58.96
PerEnt	LF	–	22.99	18.71	30.08	3.25	0.00	63.80
PerEnt	BB	CL	**23.24**	21.29	25.77	0.48	0.09	64.98
PerEnt	LF	CL	23.18	18.81	30.45	4.22	0.09	62.21

Table 18: **Results for XML Entity Index Extraction.** We compared concatenated (Concat) vs. per-entity (PerEnt) setups across text-only and text+image CLIP (CL) inputs with BigBird (BB) and Longformer (LF).

Model	Setup	Text	Image	Max Entities	Max Pages	Max Tokens
InternVL	Question	–	✓	–	10	1024
InternVL	Zero-shot	–	✓	–	10	1024
InternVL	One-shot	–	✓	–	5	1024
InternVL	Question	✓	–	25	–	1024
InternVL	Zero-shot	✓	–	25	–	1024
InternVL	One-shot	✓	–	25	–	1024
InternVL	Question	✓	✓	25	5	1024
InternVL	Zero-shot	✓	✓	25	5	1024
InternVL	One-shot	✓	✓	25	5	1024
Qwen2.5	Question	–	✓	–	10	1024
Qwen2.5	Zero-shot	–	✓	–	10	1024
Qwen2.5	One-shot	–	✓	–	10	1024
Qwen2.5	Question	✓	–	200	–	1024
Qwen2.5	Zero-shot	✓	–	200	–	1024
Qwen2.5	One-shot	✓	–	50	–	1024
Qwen2.5	Question	✓	✓	100	10	1024
Qwen2.5	Zero-shot	✓	✓	100	10	1024
Qwen2.5	One-shot	✓	✓	50	10	1024

Table 19: **Generative text extraction best model hyperparameters.**

fering fine-grained bounding box supervision to support model development. DocHop-QA may be used to design new training strategies (e.g., multi-image stitching, hierarchical encoding) or pretraining tasks that improve entity-level reasoning under spatial and modality constraints.

Results of Task4: XML Entity Index Extraction We further investigate multi-label entity classification using XML-based entity indices. The best F1 score (23.24) was achieved using BigBird with both text and image inputs in the per-entity setup (Table 18), indicating that visual context improves model understanding and highlighting the value of layout-grounded visual features. Model comparisons reveal complementary strengths: BigBird excels in capturing long-range context in the text-concatenated setup, while Longformer performs better on segmented inputs. Visual inputs further boost per-entity performance, showing the benefit of aligning text with document layout. DocHop-QA exposes label bias and modality interaction challenges, providing a benchmark for developing models that integrate textual, visual, and positional cues in entity-level reasoning.

Impact of Context Entity Cardinality on Model Performance To investigate how contextual complexity affects model performance, we analyze Task 4 performance as a function of the number of context entities involved in each question instance described in Figure 24. Across both text-only and text-plus-image settings, we observe a general degradation in F1 score as the number of entities increases, particularly for paragraph-oriented instances. For single-document questions, performance declines noticeably as entity counts approach 80, while for multi-document questions, a similar trend appears around 100 entities. This degradation likely reflects increased noise from repetitive or overly detailed contextual information, such as experimental settings or hyperparameters that are irrelevant to the question. Interestingly, a modest performance increase is observed beyond these thresholds, which we attribute to the small number of instances in higher entity-count bins. For table-oriented, multi-document instances, overall scores remain lower than paragraph-oriented counterparts. This gap underscores the difficulty of contextualizing tabular information within large entity sets and suggests that improved mechanisms for entity selection and table-aware reasoning are crucial for addressing complex DocHop-QA instances.

G.4 Hyperparameters

We provide the best hyperparameter setups in Tables 19 to 22. For the generative models in Task 1 and Task 2, we controlled the number of input entities and page images for input processed setups to manage memory issues. Maximum tokens are set to 1024 for all setups. Being a smaller model, InternVL2 only ran successfully with a maximum of 25 entities and 10 pages. Qwen2.5 could handle up to 200 entities and 100 page images but, interestingly, performed best with only 10 pages. We mirror the best setup from Qwen2.5 for experiments done for Qwen3. For fine-tuned models in Task 3 and Task 4, we explore different values for learning rate {2e-01, 2e-02, 2e-03, 2e-04, 2e-05},

early stop patience {None, 5, 10}, and warm up {None, 10}. For learning rates, 2e-02 generally produces a better F1 score, however, predicted indices tend to skew to the more frequent entity indices, raising recall but sacrificing precision. For closed-source models, we follow the hyperparameter setup of the best performing open-source model when possible.

Model	Setup	Text	Image	Max Entities	Max Pages	Max Tokens
InternVL	Zero-shot	✓	–	25	–	1024
InternVL	Zero-shot	✓	✓	25	5	1024
InternVL	One-shot	✓	–	25	–	1024
InternVL	One-shot	✓	✓	25	5	1024
Qwen2.5	Zero-shot	✓	–	100	–	1024
Qwen2.5	Zero-shot	✓	✓	50	10	1024
Qwen2.5	One-shot	✓	–	50	–	1024
Qwen2.5	One-shot	✓	✓	50	10	1024

Table 20: **Structured generative answering best model hyperparameters.**

Model	Learning Rate	Batch Size	Max Len	Early Stop	Warm Up
LayoutLMv3	2e-02	32	512	10	10
LayoutXLM	2e-05	64	512	10	10

Table 21: **BBox Entity Index Extraction best model hyperparameters.**

Setup	Text Emb	Img Emb	Max Len	Learning Rate	Early Stop	Batch Size	Warm Up
Concat	BB	–	2048	2e-02	5	16	10
Concat	LF	–	2048	2e-02	5	8	10
PerEnt	BB	–	4096	2e-02	0	64	0
PerEnt	LF	–	4096	2e-02	0	64	0
PerEnt	BB	CLIP	4096	2e-02	5	64	10
PerEnt	LF	CLIP	4096	2e-02	5	64	10

Table 22: **XML Entity Index Extraction best model hyperparameters.** Concat: Concatenated, PerEnt: Per Entity, BB: BigBird, LF: Longformer

H Ablation Studies

Modality Contribution Analysis To examine the necessity of multimodal reasoning, we conduct a controlled ablation using Qwen3-VL-8B under four input configurations. As shown in Table 23, removing visual input leads to a measurable performance drop relative to the full multimodal setting, while image-only input is substantially weaker, confirming that textual content remains the dominant modality. Nonetheless, joint multimodal input consistently outperforms text-only, indicating that visual cues contribute non-trivially to answer synthesis. The oracle-text upper bound further reveals that evidence integration itself remains a key bottleneck, even when relevant snippets are directly provided.

Setting	Input	BLEU	ROUGE-L
Full (Text+Image)	OCR text + page images	29.7	32.1
Text-only	OCR text only	26.0	28.8
Image-only	Page images only	18.2	21.0
Oracle-text	Gold snippets only	35.4	38.2

Table 23: Modality ablation on DocHop-QA using Qwen3-VL-8B.

Shortcut Learning Analysis To investigate whether models can exploit parametric knowledge or generic scientific phrasing rather than performing genuine document-grounded reasoning, we evaluate Qwen3-VL-8B in two settings: with full document access and with question input alone. As reported in Table 24, question-only performance is extremely low (BLEU 6.1, ROUGE-L 9.6), confirming that answers cannot be derived from prior knowledge alone. Combined with the document-grounded, multi-hop template constraints used during question generation, these results indicate that DocHop-QA requires genuine evidence aggregation across multiple documents rather than superficial semantic matching.

Setting	BLEU	ROUGE-L
With Documents	29.7	32.1
Question Only	6.1	9.6

Table 24: Shortcut learning analysis on DocHop-QA

Controlled Evidence-Access Analysis We evaluate Qwen3-VL-8B (one-shot) under different evidence-access settings on the same 50-instance subset used for human evaluation to further illustrate the task difficulty (Table 25). We find that question-only performance is very low (10.2), indicating that answers cannot be generated from

Setting	ROUGE-L
Question-only	10.2
Retrieved (GTR top-5)	28.2
Oracle (gold snippets only)	48.5
Human (full document)	58.6

Table 25: Controlled evidence-access performance analysis comparing input setups question-only, GTR retrieved context access, and oracle access to human performance with full document access.

parametric knowledge or general scientific phrasing alone. Oracle, having access to the gold snippets, improves over Retrieved, suggesting that retrieval contributes meaningfully to overall task difficulty. However, even with gold evidence provided, model performance (48.5) remains below human performance (58.6). This suggests that integrating evidence across documents and synthesizing answers remains challenging, even when correct snippets are available. The consistent performance gaps across conditions indicate that DocHop-QA requires non-trivial evidence aggregation rather than superficial summarization or localization. These results provide empirical evidence that DocHop-QA questions are evidence-dependent and reflects reasoning complexity.

Human Answerability and Realism Analysis To further examine whether DocHop-QA reflects realistic and answerable scientific information-seeking scenarios, we conducted a human evaluation on 50 QA instances covering diverse reasoning concepts and hop types. Three domain experts (1 CS PhD, 1 Biomedical Engineering PhD, 1 Medical Science PhD) independently answered the questions and evaluated answerability confidence. As shown in Table 26, 84% of the questions were judged answerable with an average confidence score of 4.1/5. Human-generated answers also achieved substantially higher overlap with reference answers than current model outputs. These results suggest that DocHop-QA does not primarily consist of ill-posed or artificially constructed queries, but instead requires evidence-grounded multi-document reasoning.

Metric	Result
Answerable (%)	84
Confidence (1–5)	4.1
Human vs. Gold (ROUGE-L)	58.6

Table 26: Human evaluation on 50 instances.

Retrieval Quality and Multi-Hop Recall Table 27 reports snippet-level and complete multi-hop recall for three retrieval systems. While Top-5 snippet recall is relatively high across all retrievers, complete multi-hop recall, which requires all supporting evidence pieces to be jointly retrieved, remains substantially lower (e.g., 0.39 at Top-5 for GTR-large). This recall gap directly reflects the core fan-hop aggregation challenge in DocHop-QA: retrieving any single relevant snippet is comparatively tractable, whereas locating the full set of evidence required for a multi-hop answer remains an open challenge.

Retriever	Top-5 Recall	Top-10 Recall	Complete R@5	Complete R@10
TF-IDF + BERT	0.68	0.77	0.32	0.41
BM25	0.71	0.79	0.35	0.42
GTR-large	0.76	0.84	0.39	0.48

Table 27: Retrieval quality analysis on DocHop-QA, evaluated with Recall(Top-K and Complete Recall)

I Supplementary Materials

System Prompt: You are an evaluator for a multi-hop QA dataset called DocHop-QA. Evaluate each QA instance on the following three criteria using a 4-point scale.

1. Reality/Fluency: 4=realistic and coherent; 3=slightly awkward but interpretable; 2=unnatural or hard to interpret; 1=incomprehensible.
2. Accuracy: 4=all snippets directly relevant; 3=mostly relevant; 2=tangentially related; 1=entirely unrelated.
3. Completeness: 4=all parts addressed; 3=mostly covered; 2=partial coverage only; 1=no useful information.

Input: **Question:** {question} **Snippets:** {snippets}

Output: `{"reality_fluency": {"score": <1-4>, "reason": "..."}, "accuracy": {"score": <1-4>, "reason": "..."}, "completeness": {"score": <1-4>, "reason": "..."}}`

Table 28: **LLM evaluator prompt for dataset quality assurance.** Three criteria are assessed on a 4-point scale with structured JSON output.

Few-shot Example A (high scores)
Question: How does L-carnitine differ from black tea in protecting LDL and liver lipids from oxidation?
Snippets: Snippet 1 describes L-carnitine protecting LDL lipids and proteins. Snippet 2 describes black tea reducing liver lipid peroxidation and maintaining glutathione levels.

Output:
```
{"reality_fluency": {"score": 4,
"reason": "Clear, scientifically
meaningful comparative question."},
"accuracy": {"score": 4, "reason":
"Both snippets directly correspond to
the two subjects of the question."},
"completeness": {"score": 3, "reason":
"Mechanisms are described but
direct comparative efficacy data is
missing."}}
```

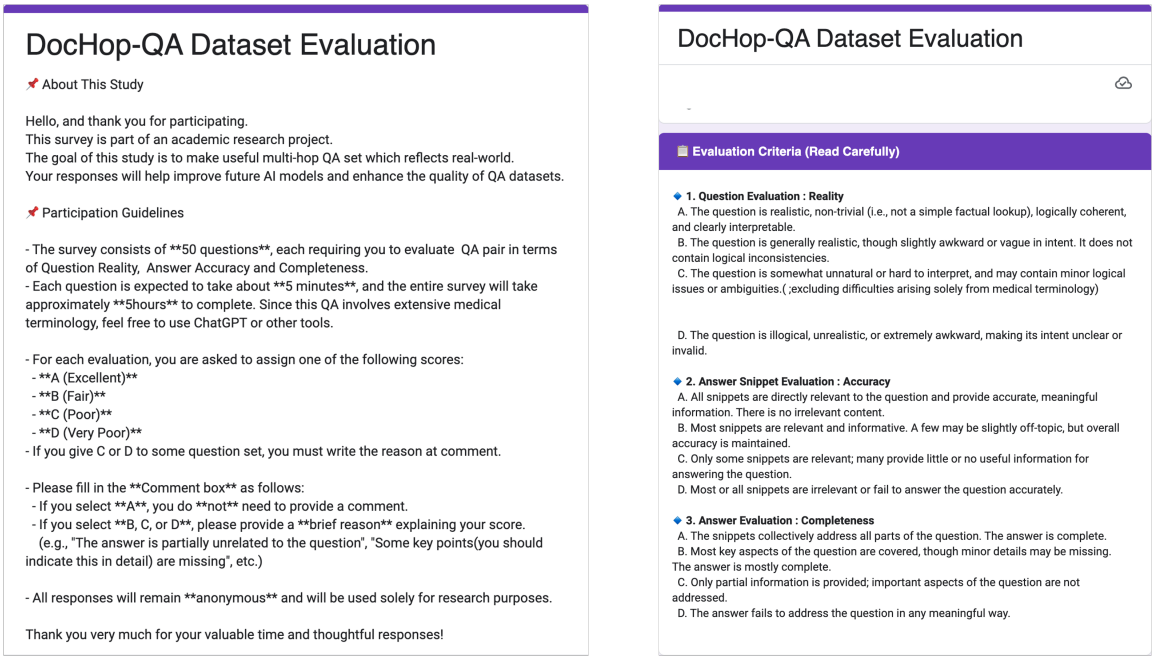
Few-shot Example B (mid scores)
Question: How do DNMT epigenetic mechanisms in lung cancer and Arab-American health disparities differ in disease prevention?
Snippets: All 5 snippets focus on DNMT and lung tumorigenesis; Arab-American health disparities are not mentioned.

Output:
```
{"reality_fluency": {"score": 3,
"reason": "Complex but interpretable
as a comparative research question."},
"accuracy": {"score": 2, "reason":
"Snippets only cover one of the two
required topics."},
"completeness": {"score": 2, "reason":
"Only half the question can be
addressed; the other topic has no
coverage."}}
```

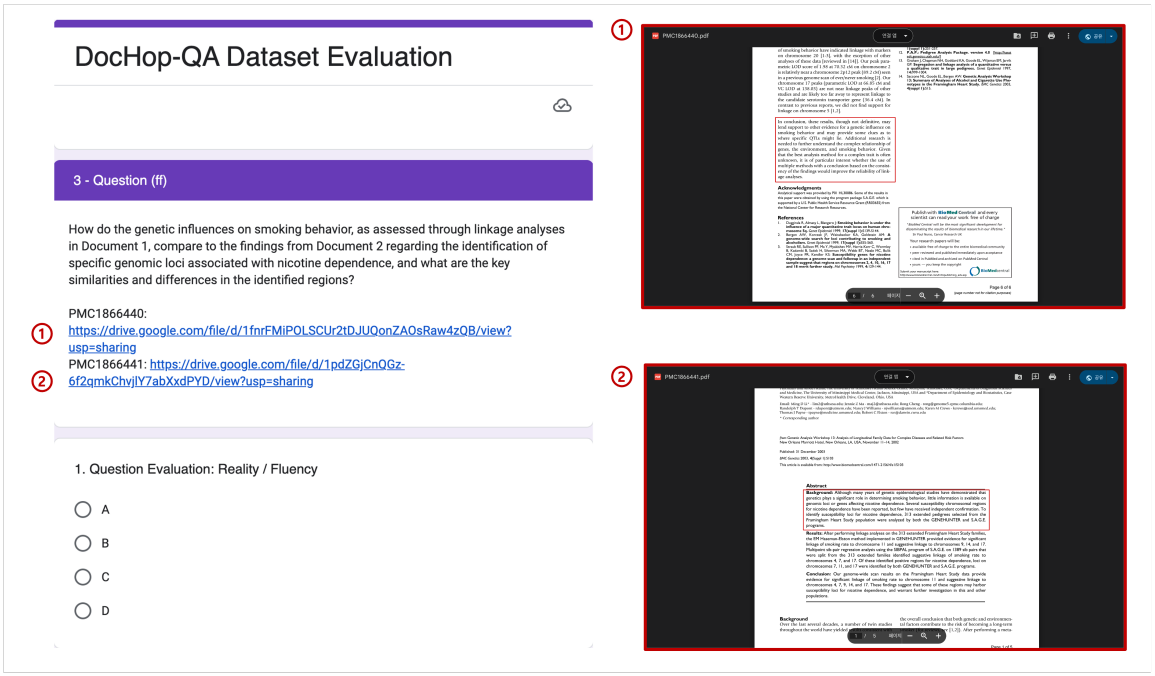
Few-shot Example C (mid-high scores)
Question: What surgical approaches for Moyamoya disease and pulmonary artery stenosis show better outcomes?
Snippets: Snippet 1 details revascularization for Moyamoya. Snippet 2 covers balloon dilation for pulmonary stenosis. Snippet 3 discusses general pediatric cardiac outcomes.

Output:
```
{"reality_fluency": {"score": 3,
"reason": "Somewhat complex but
clearly understandable as a surgical
outcomes comparison."},
"accuracy": {"score": 3, "reason":
"Two snippets directly address the
two conditions; the third is loosely
related."},
"completeness": {"score": 3, "reason":
"Relevant surgical detail provided
for both, though direct outcome
comparisons are not explicit."}}
```

Table 29: **Few-shot examples provided to the LLM evaluator.** Examples A, B, and C illustrate high, mid, and mid-high scoring cases respectively.



(a) Overview page providing the study purpose and outlining the evaluation criteria.



(b) Example QA instance with hyperlinks to PDF documents displaying the answer snippet with a bounding box.

Figure 26: **Screenshots of the structured Google Form interface used for human evaluation.**