Deep-Reporter: Deep Research for Grounded Multimodal Long-Form Generation

Fangda Ye[1†], Kuicai Dong[2†], Zhifei Xie[2†], Yuxin Hu[1†], Yihang Yin[3], Shurui Huang[1], Shikai Dong[4], Chen Zhang[1], Jianzhu Bao[2], Shuicheng Yan[1]
[1] National University of Singapore [2] Nanyang Technological University
[3] University of Edinburgh [4] Beijing Institute of Technology

Abstract

Recent agentic search frameworks enable deep research via iterative planning and retrieval, reducing hallucinations and enhancing factual grounding. However, they remain text-centric, overlooking the multimodal evidence that characterizes real-world expert reports. We introduce a pressing task: multimodal long-form generation. Accordingly, we propose DEEP-REPORTER, a unified agentic framework for grounded multimodal long-form generation. It orchestrates: (i) *Agentic Multimodal Search and Filtering* to retrieve and filter textual passages and information-dense visuals; (ii) *Checklist-Guided Incremental Synthesis* to ensure coherent image-text integration and optimal citation placement; and (iii) *Recurrent Context Management* to balance long-range coherence with local fluency. We develop a rigorous curation pipeline producing 8K high-quality agentic traces for model optimization. We further introduce M^{2}LONGBENCH, a comprehensive testbed comprising 247 research tasks across 9 domains and a stable multimodal sandbox. It enables unified multimodal assessment, fair comparison, and accessible evaluation without commercial APIs. Extensive experiments demonstrate that long-form multimodal generation is a challenging task, especially in multimodal selection and integration, and effective post-training can bridge the gap. Our code is available at `https://github.com/fangda-ye/Deep-Report`.

1 Introduction

LLMs have driven remarkable progress in question answering (Rajpurkar et al., 2016; Kwiatkowski et al., 2019; Rein et al., 2024). Building upon this foundation, recent research has increasingly focused on complex long-form generation, including Wiki article (Shao et al., 2024; Yang et al., 2025c), academic paper (Lu et al., 2024; Ghafarollahi and Buehler, 2025), and industry reports (Tian et al., 2025). These systems typically tackle multifaceted queries, such as "*Generate an analysis of the renewable energy market*", which demand extensive factual grounding across multiple sources. Existing approaches employ agentic search (Zhang et al., 2025; Jin et al., 2025; Li et al., 2025b; Team et al., 2026), where autonomous agents iteratively plan queries, retrieve information from external sources, and synthesize findings. This paradigm mitigates hallucinations and enhances answer factuality.

However, existing long-form generation systems remain text-centric (Figure 1a), overlooking multimodal generation (Xi et al., 2025; Gu et al., 2025; Bai et al., 2024a; Bao et al., 2022b; Ye et al., 2026). While recent works (Yang et al., 2025b; Xia et al., 2024; Zhang et al., 2026c) reveal that real-world expert reports rely heavily on visual evidence including: tables, charts, infographics, complex layouts, etc. Incorporating such multimodal elements fundamentally enhances both interpretability and engagement of generated reports. Yet this integration introduces three critical challenges: (1) ***Agentic Multimodal Retrieval***: How can we effectively adapt text-based agentic search frameworks to retrieve, filter, and integrate multimodal information? (2) ***Coherent Multimodal Long-Form Generation***: How can we maintain *global sectional coherence*, achieve *image-text consistency*, and manage extensive *multimodal context*? (3) ***Evaluation of Multimodal Generation***: Given open-ended nature of agentic search and long-form generation, how can we evaluate in a *rigorous and reproducible setting*?

To address these challenges, we introduce DEEP-REPORTER (§2), a unified agentic framework that adapts existing agentic search and long-form generation paradigms to multimodal setting. DEEP-REPORTER supports comprehensive multimodal agentic search, achieved by collaborative interactions to decompose complex queries, iteratively

[†]These authors contributed equally. Correspondence to: Fangda Ye <fangda.ye@u.nus.edu> and Kuicai Dong <kuicai001@e.ntu.edu.sg>.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41137–41177

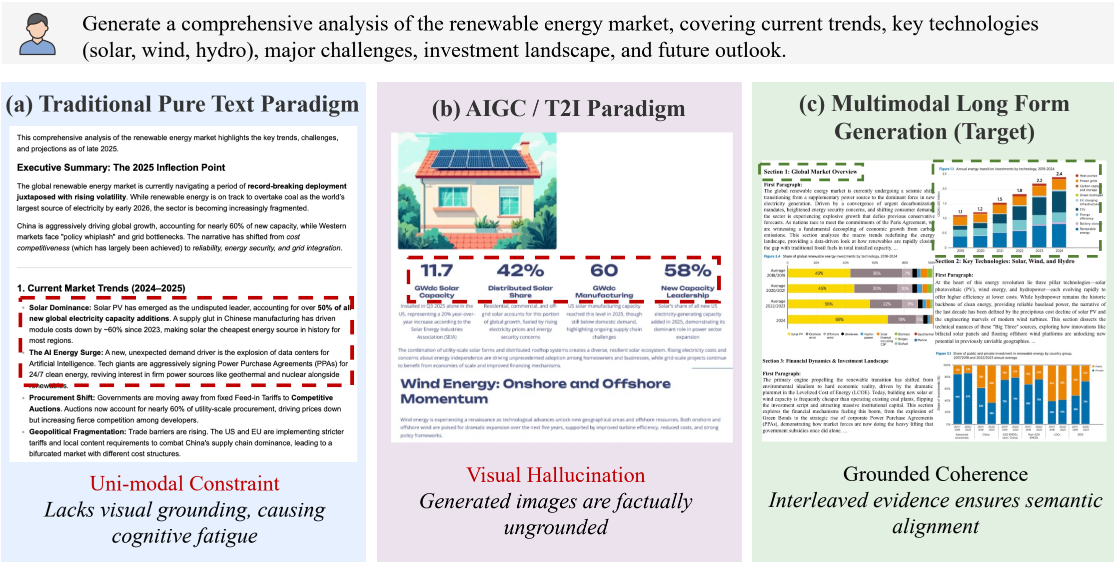


Figure 1: **Comparison of paradigms in long-form report generation.** While traditional systems (a) lack visual engagement and pure T2I approaches (b) suffer from factual hallucinations and narrative fragmentation, our method (c) achieves high coherence and factuality by retrieving and integrating real-world visual evidence.

retrieves both textual and visual passages, refine/filter relevant evidence. Subsequently, DEEP-REPORTER employs a "Checklist-Guided Incremental Synthesis" mechanism, formalizing the report structure into "Semantic Anchors" and utilizing recurrent context update strategy, to ensure coherent interleaving of text and images throughout the report (as shown in Figure 1c). To equip open-source models with these capabilities, we propose a pipeline to curate high-quality agentic trace (§2.2). We specifically target two critical competencies: (1) ***Agentic Multimodal Search***: teaching agents to strategically plan queries, retrieve heterogeneous content, and filter high-value visual evidence, and (2) ***Coherent Multimodal Synthesis***: demonstrating when and where to cite visual elements while maintaining global coherence via checklist-driven generation. Through rigorous post-processing with LLM-based scoring and human verification on visual hallucinations, we curate 8k expert trajectories to enhance model multimodal competencies.

To address the lack of standardized multimodal testbeds, we construct M^{2}LONGBENCH (§3), a benchmark comprising 247 complex research tasks across 9 domains. Unlike existing benchmarks (Wu et al., 2025a; Du et al., 2025b) that are either text-only or rely on dynamic web APIs, M^{2}LONGBENCH establishes a stable multimodal sandbox containing 95K images and 108M text chunks (aggregated from 45K webpages and 6.4K PDFs), with annotations averaging 102 images and 168 text chunks as ground-truth evidence per report. This enable rigorous evaluation of visual retrieval and image-text coherence in long-form generation. M^{2}LONGBENCH addresses three critical gaps: (1) *Unified Multimodal Assessment*: simultaneous evaluation of textual and visual retrieval precision, along with image-text integration quality in generated reports; (2) *Fairness & Transparency*: researchers can isolate retrieval algorithm contributions within a controlled corpus (Chen et al., 2025), eliminating temporal drift from evolving web content and opacity of black-box APIs; (3) *Accessibility*: the self-contained sandbox removes dependencies on expensive commercial search APIs, lowering barriers while maintaining evaluation rigor. Experiments on M^{2}LONGBENCH confirm that DEEP-REPORTER substantially outperforms RAG baselines, and our curated agentic traces significantly boost multimodal evidence selection and generation quality. In summary, we present a holistic solution for grounded multimodal long-form generation, with three contributions:

- We propose DEEP-REPORTER that orchestrates multimodal retrieval and checklist-guided synthesis, achieving coherent multimodal generation and visual evidence grounding.
- We develop a rigorous trajectory curation pipeline that constructs 8K high-quality expert traces, enabling post-training of open-sourced LLMs to achieve precise multimodal selection and coherent multimodal synthesis.
- We design M^{2}LONGBENCH for comprehensive and reproducible multimodal long-form evaluation, revealing the pressing need for advanced multimodal agentic framework and post-training.

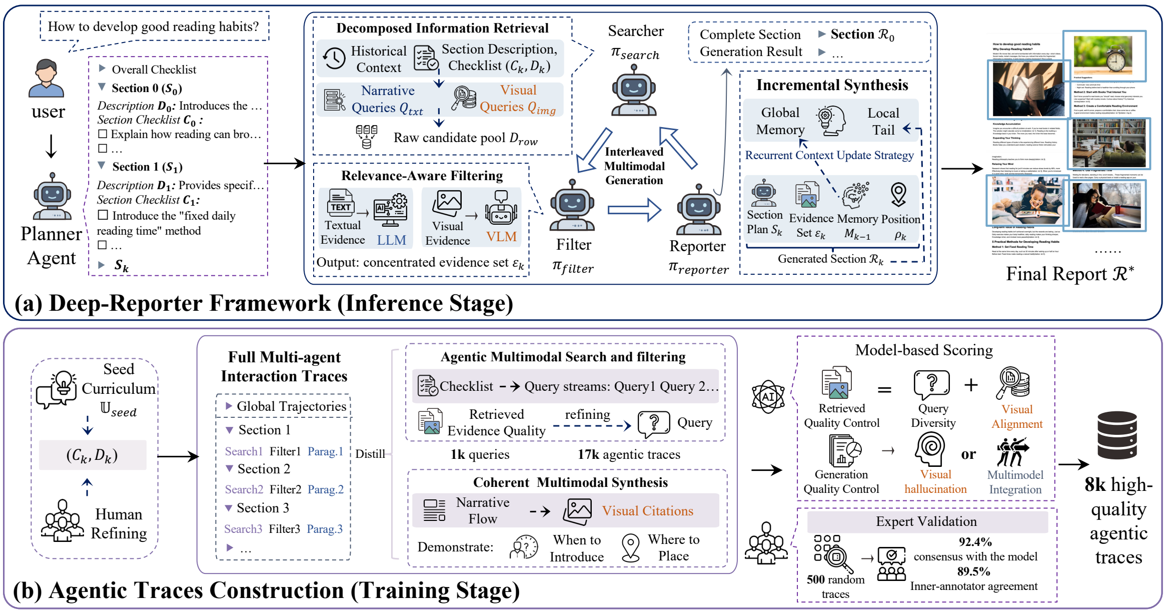


Figure 2: **Deep-Reporter Architecture.** (a) The multi-agent framework orchestrates planning, multimodal information seeking, and incremental writing to generate professional reports. (b) The data synthesis pipeline constructs high-quality expert trajectories to equip open-weight models with deep research capabilities.

2 Method: Deep-Reporter

2.1 The Deep-Reporter Framework

Deep-Reporter (Figure 2a) transforms a user query into a comprehensive multimodal report $\mathcal{R}^*$ via three specialized agents: a *Planner*, a collaborative *Searcher-Filter*, and a *Reporter*.

Sectional Planning with Dual-granularity Checklist. Following prior work (Shao et al., 2024; Prasad et al., 2024), we use hierarchical planning to maintain global coherence. The Planner Agent π_{plan} decomposes the user input $\mathcal{U}$ into a structured global blueprint: $\pi_{plan}(\mathcal{U}) = \{S_1, \dots, S_N\}$, where each section $S_k = (D_k, \mathcal{C}_k)$ is formalized as a semantic checklist to enforce content rigor. Specifically, D_k denotes the coarse-grained content description, defining the scope and core topics of current section. Meanwhile, $\mathcal{C}_k = \{c_{k,1}, \dots, c_{k,m}\}$ provides fine-grained semantic anchors, specifying facts, and arguments (Bao et al., 2022a) to be addressed in the current section. This dual-granularity checklist ensures both coherence and precision. Furthermore, it allows optional human-in-the-loop refinement for better user intent alignment.

Agentic Multimodal Search and Filtering. Grounding long-form reports in verifiable evidence (Yao et al., 2022; Jin et al., 2025) requires retrieving both textual passages for factual content and visual elements (charts, diagrams, infographics). To achieve this, we implement a targeted agentic multimodal search mechanism that orchestrates dual-stream retrieval and relevance-aware filtering. For each target section S_k, the searcher π_{search} performs Decomposed Information Retrieval (Trivedi et al., 2023), formulating two complementary query streams conditioned on the sectional checklist $(D_k, \mathcal{C}_k)$ and historical context: (i) *Narrative Queries* $\mathcal{Q}_{txt}$ to retrieve factual passages and statistics, and (ii) *Visual Queries* $\mathcal{Q}_{img}$ to retrieve charts, diagrams, and infographics. These queries are executed against multimodal search tools, yielding a raw candidate pool $\mathcal{D}_{raw}$.

To ensure high-quality evidence, the Filter π_{filter} applies Relevance-Aware Filtering (Asai et al., 2024), selectively retaining passages and images that satisfy sectional constraints:

$$\mathcal{E}_k = \{d \in \mathcal{D}_{raw} \mid \mathbb{I}_{ver}(d, D_k, \mathcal{C}_k) = 1\} \quad (1)$$

where $\mathbb{I}_{ver}$ denotes a multimodal verification indicator. For textual evidence, we use LLMs to assess semantic entailment *w.r.t.* $(D_k, \mathcal{C}_k)$. For visual evidence, a VLM evaluates the image's *informativeness*, filtering out non-informative decorative images while prioritizing information-dense figures (e.g., statistical charts, technical diagrams). This yields a concentrated evidence set $\mathcal{E}_k$.

Multimodal Incremental Synthesis with Recurrent Context Management. Generating coherent long-form multimodal reports introduces two critical context challenges: (i) maintaining narrative continuity across previously generated sections, and (ii) grounding the current section in concentrated multimodal evidence (text passages + images). Naively concatenating full history and all retrieved evidence into a single prompt leads to context overflow and degraded comprehension, particularly problematic given that images can consume thousands of tokens each. To address this, the Reporter π_{report} employs a Recurrent Context Update strategy (Packer et al., 2023), compressing historical context into $\mathcal{M}_{k-1} = (m_{global}, m_{local})$, where m_{global} maintains a recursive summary of the narrative arc (preserving global coherence), while m_{local} retains the verbatim tail of the preceding section (ensuring smooth transitions). Such a memory mechanism balances long-range dependencies with local fluency. For the k-th section, synthesis is formulated as a state-conditioned generation process:

$$\mathcal{R}_k \sim P_\theta(\cdot \mid \underbrace{S_k}_{\text{Plan}}, \underbrace{\mathcal{E}_k}_{\text{Evidence}}, \underbrace{\mathcal{M}_{k-1}}_{\text{Memory}}, \underbrace{\rho_k}_{\text{Position}}) \quad (2)$$

where ρ_k denotes the section's position within the global structure. After generation, the memory undergoes a update: $\mathcal{M}_k \leftarrow \Phi(\mathcal{M}_{k-1}, \mathcal{R}_k)$, propagating contextual understanding forward.

Interleaved Multimodal Generation. A key capability of DEEP-REPORTER is seamlessly integrating visual evidence into narrative flow. To enable text-based LLMs to reason about image placement, we adopt caption-based transcription (Dong et al., 2025b), converting images in $\mathcal{E}_k$ into detailed textual descriptions. During generation, the model learns to insert citations (e.g., `![](cite:img1)`) at contextually appropriate positions.

2.2 Agentic Traces Construction

Current open-source LLMs remain text-centric, lacking the multimodal agentic reasoning capabilities required for our task. To bridge this gap, we construct a high-quality training corpus (Figure 2b) using DEEP-REPORTER (§2.1), combined with expert-in-the-loop curation and validation.

Stage 1: Expert-in-the-Loop Planning. We begin by curating a diverse seed curriculum $\mathbb{U}_{seed}$ spanning 9 domains. For each query, domain experts refine the sectional outlines and checklists $(D_k, \mathcal{C}_k)$ through iterative discussion. This expert involvement ensures that: (i) the research structure exhibits logical complexity and appropriate granularity, avoiding trivial decompositions; (ii) the checklists $\mathcal{C}_k$ contain concrete semantic anchors that require multimodal evidence; and (iii) the plans maintain diversity in both topical coverage and structural patterns. This process results in 1K queries along with outlines and checklists.

Stage 2: Agentic Trace Distillation. Using expert-refined plans as input, we execute DEEP-REPORTER with a frontier proprietary model to record the full multi-agent interaction traces, to effectively enhance grounded generation capabilities (Bao et al., 2023). We specifically focus on distilling two competencies: **1. Agentic Multimodal Search and Filtering.** We capture the model's strategy for: (i) decomposing checklist requirements into precise and executable query streams; (ii) iteratively refining queries based on retrieved evidence quality; and (iii) distinguishing information-dense visuals from decorative or low-value images. **2. Coherent Multimodal Synthesis**. We record the decision logic for visual citation placement within narrative flow. Specifically, the traces demonstrate: (i) *when* to introduce visual evidence; (ii) *where* to position citations at contextually appropriate positions to maximize image-text consistency. Building on 1K queries, we obtain 17K agentic traces in stage 2.

Stage 3: Rigorous Quality Assurance. Raw traces undergo strict filtering at scale using an advanced proprietary model. **1. Retrieval Quality Control.** The verification model assesses search queries via: (i) *Query Diversity*: whether the queries explore multiple facets of the checklist requirements rather than redundantly rephrasing similar searches; and (ii) *Visual Alignment*: whether visual queries appropriately target the visual modalities needed for the section. **2. Generation Quality Control.** The model aggressively flags traces exhibiting visual hallucinations (Huang et al., 2024; Bai et al., 2024b) or poor multimodal integration. Specifically, it identifies: (i) *Visual Selection Errors*: citing images that are factually incorrect or irrelevant; (ii) *Positioning Defects*: inserting visual citations at inappropriate locations that disrupt narrative flow; (iii) *Image-Text Inconsistency*: cases where textual descriptions contradict or misrepresent the cited visual content; and (iv) *Image Reuse Issues*: redundantly citing the same image across sections without justification. **3. Expert Validation.** Two domain experts independently annotated a random sample of 500 traces. The proprietary model achieved 92.4% accuracy against expert consensus, with an inter-annotator agreement of 89.5%. Trajectories flagged by the automated system are discarded, distilling the raw data into 8K high-quality agentic traces.

Benchmark	Retrieval Environment			Evi. Annotation		Report Annotation				Task Info		
	Source	Scale†	Modal	Text Chk	Image	Outl.	Checkl.	Modal	#Word	Domain	#Task	Expert
HELLOBENCH (Shao et al., 2024)	-	-	-	✗	✗	✗	✗	Text	739	5	647	✓
LONGEVAL (Wu et al., 2025a)	-	-	-	✗	✗	✓	✗	Text	3.5k	3	166	✓
WRITINGBENCH (Wu et al., 2025b)	-	-	-	✗	✗	✗	✓	Text	2.7k	6	1k	✗
LONGGENBENCH (Wu et al., 2024)	-	-	-	✗	✗	✗	✓	Text	15k	4	6.4k	✗
INTERLEAVEDBENCH (Liu et al., 2024)	-	-	-	✗	✗	✗	✗	MM	19	10	815	✓
LONGLAMP (Kumar et al., 2024)	Static	100	Text	✓	✗	✗	✗	Text	200	4	19.3k	✗
M-LONGDOC (Chia et al., 2025)	Static	180	MM	✓	✓	✗	✗	Text	180	3	851	✓
MMDOCIR (Dong et al., 2025a)	Static	313	MM	✓	✗	✗	✗	Text	21	10	1.6k	✓
DOCBENCH (Zou et al., 2025)	Static	229	Text	✗	✗	✗	✗	Text	15	5	1.1k	✓
MMDOCRAG (Dong et al., 2025b)	Static	222	MM	✓	✓	✗	✗	Text	24	10	4k	✓
M4DOCBENCH (Dong et al., 2025c)	Sandbox	304	MM	✓	✓	✗	✗	MM	60	4	158	✓
DEEPRESEARCH (Du et al., 2025b)	Web	∞	Text	✓	✗	✗	✓	Text	3.3k	22	100	✓
M²LONGBENCH (Ours)	**Sandbox**	**61.4k**	**MM**	✓	✓	✓	✓	**MM**	**5.5k**	9	247	✓

Table 1: **M²LONGBENCH versus existing benchmarks.** *Evi.* denotes candidate sets for text/image retrieval. *Modality* describe the expected input/output format (MM=Multimodal).

3 Benchmark: M²LONGBENCH

As reflected in Table 1, current benchmarks exhibit critical limitations: short multimodal responses, unimodal textual bias, and small-scale or uncontrolled retrieval environment. To address these gaps, we introduce M²LONGBENCH, comprising 247 tasks across 9 domains, supported by a stable multimodal sandbox containing 95K images and 108M text chunks. As mentioned in Section 1, M²LONGBENCH enables unified multimodal assessment, fair and transparent comparison, and accessibility. Specifically, we develop a rigorous expert-in-the-loop pipeline to ensure task authenticity, sandbox realism, and annotation reliability.

Task Collection and Curation. We curated a dataset of 247 high-quality tasks through expert sourcing, hybrid filtering, and structured extraction. **1. Source Selection:** Domain experts identified 100+ authoritative repositories across 9 domains, allowing us to crawl over 4,000 candidate reports. **2. Filtering:** A hybrid approach combining rule-based heuristics (e.g., token count >3k, visual richness) and LLM scoring narrowed candidates to $\sim$1,000. A subsequent manual review removed time-sensitive content to ensure high quality and domain diversity, yielding 247 final reports. **3. Processing:** We converted these reports to clean markdown using Crawl4AI (UncleCode, 2024) (web) and MinerU (Niu et al., 2025) (PDFs), followed by manual cleaning. Finally, we used LLM-assisted extraction with human verification to transform them into structured tasks (queries, outlines, checklists). These reports serve as task *prototypes*, reflecting the open-ended and high-complexity nature of real-world multimodal research problems.

Multimodal Sandbox Construction. We constructed a robust, large-scale multimodal sandbox through evidence aggregation, data ingestion, and multimodal indexing. **1. Evidence Aggregation:** We extracted citation URLs from source reports and expanded the corpus via the Google Custom Search API. For each section, we utilized an LLM to generate search queries based on task metadata, section descriptions, and checklists, retrieving the top-30 URLs per section to supplement original citations. **2. Data Ingestion:** Following URL deduplication, we conducted batch crawling and applied the previously described cleaning pipeline (Crawl4AI for HTML, MinerU for PDFs) to produce clean markdown and locally stored images. **3. Multimodal Indexing:** We performed hierarchical text chunking on the markdown content and embedded both text chunks and images using Jina Embeddings-v4 (Günther et al., 2025).

Silver-Standard Annotation. We constructed a dual-layer annotation set evaluating both the research process and generation outcome. **1. Process-Level Evidence:** We employed a retrieval-and-verification (Jin et al., 2025) approach to build evidence pools. For each section, we utilized LLM-formulated queries to fetch candidates (Top-40 images/Top-100 text chunks per section) from the sandbox, followed by model-based relevance scoring and strict expert adjudication to eliminate false positives. This yielded verified pools averaging 168 text chunks and 103 images per task. **2. Outcome-Level Reference:** Addressing the open-ended nature of long-form generation (Celikyilmaz et al., 2020; Becker et al., 2024), we constructed Expert-Refined Silver Reports ($\mathcal{R}_{ref}$). We first utilized GPT-4.1 to synthesize drafts strictly conditioned on the verified outlines and original reports. Subsequently, domain experts performed post-editing to rectify logical inconsistencies and ensure optimal instruction adherence, establishing a rigorous baseline for comparative evaluation. See §A for complete benchmark details.

Method	Backbone	Search			Filter			Selection		
		Text	Image	Overall	Text	Image	Overall	Text	Image	Overall
Naïve RAG	Qwen3-8B	34.9 (54.6)	32.7 (27.3)	33.8 (81.9)	-	-	-	46.3 (18.7)	7.8 (0.9)	27.0 (19.6)
Naïve RAG	Qwen3-32B	**35.0** (54.7)	32.8 (27.4)	33.9 (82.0)	-	-	-	45.7 (21.0)	10.4 (1.0)	28.1 (22.0)
Storm-MM	Qwen3-32B	29.5 (105.2)	38.4 (53.8)	34.0 (159.0)	-	-	-	48.3 (26.3)	12.7 (1.4)	30.5 (27.7)
DEEP-REPORTER	Qwen3-8B	29.3 (103.8)	37.9 (52.4)	33.6 (156.3)	45.5 (50.4)	50.2 (30.5)	47.9 (81.0)	**49.6** (24.5)	8.3 (0.7)	28.9 (25.2)
w/o Filter	Qwen3-8B	29.5 (104.0)	37.2 (52.5)	33.4 (156.5)	-	-	-	43.6 (31.9)	3.8 (0.3)	23.7 (32.2)
DEEP-REPORTER	Qwen3-8B♠	30.9 (103.8)	**39.8** (52.3)	35.3 (156.1)	46.5 (51.5)	51.5 (31.0)	49.0 (82.5)	46.5 (36.9)	45.0 (9.8)	45.7 (46.7)
w/o Filter	Qwen3-8B♠	30.5 (104.2)	39.6 (52.2)	35.1 (156.4)	-	-	-	39.2 (56.9)	33.6 (6.3)	36.4 (63.3)
DEEP-REPORTER	Qwen3-8B♣	30.3 (105.6)	38.6 (53.3)	34.5 (158.9)	46.2 (52.3)	50.1 (31.3)	48.2 (83.6)	48.5 (22.5)	9.8 (0.8)	29.1 (23.4)
w/o Filter	Qwen3-8B♣	31.7 (103.4)	39.4 (52.7)	**35.6** (156.1)	-	-	-	44.2 (34.5)	6.9 (0.5)	25.6 (35.1)
DEEP-REPORTER	Qwen3-32B	29.2 (106.0)	38.1 (54.4)	33.6 (160.4)	45.4 (51.5)	50.7 (30.4)	48.0 (81.9)	48.7 (27.6)	19.5 (2.0)	34.1 (29.6)
w/o Filter	Qwen3-32B	29.2 (105.8)	38.9 (54.4)	34.1 (160.2)	-	-	-	43.3 (32.1)	12.1 (1.0)	27.7 (33.1)
DEEP-REPORTER	Qwen3-32B♠	30.9 (104.5)	39.2 (52.9)	35.1 (157.5)	**46.9** (52.1)	51.6 (31.0)	**49.3** (83.2)	48.6 (40.3)	41.4 (7.6)	45.0 (47.8)
w/o Filter	Qwen3-32B♠	30.7 (104.3)	39.4 (52.5)	35.0 (156.8)	-	-	-	40.2 (55.2)	33.4 (5.3)	36.8 (60.5)
w/o Recur.	Qwen3-32B♠	30.4 (103.2)	38.7 (52.1)	34.5 (155.3)	46.2 (51.4)	51.0 (30.6)	48.7 (82.0)	47.3 (39.1)	39.6 (7.2)	43.5 (46.3)
w/o Both	Qwen3-32B♠	30.3 (103.0)	38.6 (52.0)	34.3 (155.0)	-	-	-	38.8 (52.7)	30.1 (5.0)	34.5 (57.7)
DEEP-REPORTER	Qwen3-32B♣	29.7 (105.5)	39.4 (54.2)	34.6 (159.7)	45.6 (51.6)	51.6 (31.3)	48.6 (82.8)	49.0 (25.9)	22.1 (2.5)	35.5 (28.4)
DEEP-REPORTER	Llama3.3-70B	24.6 (107.7)	33.3 (53.4)	29.0 (161.1)	40.2 (52.0)	46.2 (27.8)	43.2 (79.8)	42.4 (38.0)	14.3 (2.1)	28.3 (40.1)
DEEP-REPORTER	Llama3.3-70B♠	30.8 (104.6)	39.6 (52.5)	35.2 (157.1)	45.9 (54.3)	**52.0** (30.3)	48.9 (84.6)	47.4 (41.5)	41.2 (6.2)	44.3 (47.7)
DEEP-REPORTER	Llama3.3-70B♣	24.8 (107.5)	33.0 (53.5)	28.9 (161.0)	40.7 (52.1)	45.9 (29.0)	43.3 (81.0)	42.3 (37.4)	14.7 (1.9)	28.5 (39.3)
DEEP-REPORTER	GPT-4.1	30.0 (105.3)	39.5 (52.8)	34.8 (158.1)	45.8 (53.1)	51.5 (31.2)	48.6 (84.3)	47.8 (42.8)	**45.4** (7.3)	**46.6** (50.0)

Table 2: **Evidence quality on M^2LONGBENCH benchmark.** Each cell shows precision (%) and quantity of evidence that are either retrieved/retained/cited. The best score is in **purple boldface** and second best is underlined. ♠ and ♣ denote SFT and DPO respectively. *w/o Filter* removes the filter module; *w/o Recur.* removes recurrent context management; *w/o Both* removes both.

4 Experiments

4.1 Evaluation Metrics

Multimodal Evidence. We evaluate the process-level evidence precision of the agentic pipeline: (i) raw candidates retrieved by searcher (**Search**), (ii) concentrated candidates filtered by filter (**Filter**), and (iii) candidates selected as part of narrative flow by reporter (**Selection**). Specifically, we report Precision@Silver measuring the proportion of candidates that matches silver-standard annotations: $|\mathcal{R}^{(t)} \cap \mathcal{S}_{\text{silver}}|/|\mathcal{R}^{(t)}|$.

Multimodal Generation. We assess the long-form multimodal generation quality across three hierarchical levels: (i) structural adherence to planning constraints (**Section Anchor**), i.e., description and checklist, (ii) section-level multimodal grounding quality (**Section Content**), which measures the richness and clarity of generated section, and image-text coherence and citation format, and (iii) holistic evaluation on entire report (**Full Report**), favoring coherent and fluent writing while penalizing repetition and early termination. Specifically, we report the *Relative Quality Score* (Dubois et al., 2024; Lin et al., 2024), using LLM&VLM to score the generated content against expert-refined silver reports $\mathcal{R}_{\text{ref}}$ and then normalize the scores (refer to §D.2 for more details). The overall performance is computed as $(S_{\text{anchor}} + S_{\text{content}} + S_{\text{full}})/3$.

4.2 Experimental Setup

Retrieval Settings. All experiments utilize the same search tool, the multimodal sandbox (§3) to ensure reproducibility. For each search query, the tool performs semantic match and returns top-20 text chunks and top-10 images. These budgets are chosen to balance evidence coverage against context overhead: a top-20/10 budget provides sufficient candidate diversity for the subsequent filtering stage while keeping the raw evidence volume manageable (on average ∼156 candidates per task, reduced to ∼81 after filtering; see Table 2). End-to-end filter runtime scales roughly linearly with the retrieval budget, allowing users to trade off quality and latency by adjusting these parameters.

DEEP-REPORTER. We implement our framework using Qwen3-8B/32B (Yang et al., 2025a) and Llama-3.3-70B-Instruct (Dubey et al., 2024):

- **Base** utilizes the agentic workflow via in-context learning without weight updates.
- **SFT** performs fine-tuning on 8K agentic traces (§2.2) for agentic search and report generation.
- **DPO** applies preference learning (Rafailov et al., 2023) using 8K agentic trace (details in §C.2).
- **w/o Filter** is an ablation variant that removes filter to quantify the impact of noise reduction.
- **w/o Recur.** removes recurrent context management while keeping filter, isolating the contribution of cross-section memory.
- **w/o Both** removes both filter and recurrent context, quantifying their combined effect.

Naïve RAG Baselines. Rather than using agentic search, we compare DEEP-REPORTER against naïve RAG baselines using a non-iterative retrieval strategy. Given the sectional anchors, RAG baselines retrieve top-10 text chunks and top-5 images

Method	Backbone	Section Anchor Desc	Section Anchor Check	Section Anchor Avg	Section Content Rich	Section Content Coh	Section Content Place	Section Content Clar	Section Content Avg	Full Report Coh	Full Report Flu	Full Report Rep	Full Report Term	Full Report Avg	Overall	Report Length
Naïve RAG	Qwen3-8B	0.0	2.0	1.0	6.1	7.9	7.1	8.5	7.4	0.0	0.0	6.9	23.1	7.5	5.3	1.9k
Naïve RAG	Qwen3-32B	0.0	1.2	0.6	6.9	5.7	6.3	6.7	6.4	0.0	0.0	10.1	24.3	8.6	5.2	2.0k
Storm-MM	Qwen3-32B	22.3	25.3	23.8	7.7	5.7	7.3	6.5	6.8	10.1	20.7	19.4	14.6	16.2	15.6	3.2k
DEEP-REPORTER	Qwen3-8B	14.2	19.4	16.8	8.1	3.6	6.9	3.6	5.6	10.5	23.5	21.9	18.6	18.6	13.7	3.9k
w/o Filter	Qwen3-8B	13.0	14.6	13.8	8.1	2.0	6.1	3.2	4.9	8.1	21.9	17.8	20.7	17.1	11.9	3.8k
DEEP-REPORTER	Qwen3-8B♠	23.9	26.3	25.1	<u>44.1</u>	<u>36.4</u>	**40.1**	**43.7**	<u>41.1</u>	17.8	23.9	22.7	19.0	20.9	29.0	5.1k
w/o Filter	Qwen3-8B♠	24.7	27.9	26.3	30.0	21.5	27.1	28.3	26.7	15.0	25.9	23.9	21.1	21.5	24.8	5.2k
DEEP-REPORTER	Qwen3-8B♣	16.6	16.6	16.6	9.3	5.3	7.3	6.5	7.1	12.6	25.5	21.1	19.8	19.7	14.5	4.2k
w/o Filter	Qwen3-8B♣	17.4	23.5	20.4	8.1	3.6	6.5	2.4	5.2	10.9	22.7	16.6	21.9	18.0	14.5	4.1k
DEEP-REPORTER	Qwen3-32B	<u>38.1</u>	<u>38.9</u>	<u>38.5</u>	12.2	10.5	15.4	10.5	12.2	24.3	<u>35.2</u>	<u>34.8</u>	<u>29.2</u>	30.9	27.2	3.9k
w/o Filter	Qwen3-32B	31.6	36.8	34.2	9.7	4.9	7.7	5.7	7.0	19.8	30.8	34.0	**30.4**	28.7	23.3	4.0k
DEEP-REPORTER	Qwen3-32B♠	**39.7**	**39.7**	**39.7**	**47.8**	**37.2**	<u>38.5</u>	<u>43.3</u>	**41.7**	<u>27.1</u>	**36.4**	**36.8**	28.7	**32.3**	**37.9**	4.7k
w/o Filter	Qwen3-32B♠	34.8	38.1	36.4	30.4	24.3	28.7	27.1	27.6	23.5	25.1	23.9	25.9	24.6	29.6	4.9k
w/o Recur.	Qwen3-32B♠	36.4	38.8	37.6	39.2	30.4	32.8	34.4	34.2	11.3	24.3	25.0	18.6	19.8	30.5	4.5k
w/o Both	Qwen3-32B♠	36.0	37.8	36.9	27.5	20.2	25.2	25.1	24.5	9.7	21.5	22.6	15.8	17.4	26.3	4.3k
DEEP-REPORTER	Qwen3-32B♣	19.4	20.2	19.8	15.8	8.5	12.2	10.1	11.6	14.2	28.3	27.9	23.9	23.6	18.4	3.9k
DEEP-REPORTER	Llama3.3-70B	4.1	6.5	5.3	11.3	6.1	9.3	6.5	8.3	4.5	8.1	10.5	5.7	7.2	6.9	3.4k
DEEP-REPORTER	Llama3.3-70B♠	35.6	37.7	36.6	36.4	<u>36.4</u>	37.7	36.4	36.7	**33.2**	32.8	30.4	27.5	<u>31.0</u>	<u>34.8</u>	4.5k
DEEP-REPORTER	Llama3.3-70B♣	4.9	7.3	6.1	11.7	6.5	10.5	8.9	9.4	4.9	13.8	9.3	7.3	8.8	8.1	3.5k

Table 3: **Main results of generation quality. Section Anchor**: Desc=Description, Check=Checklist. **Section Content**: Rich=Richness, Coh=Image-text coherence, Place=Placement, Clar=Clarity. **Full Report**: Coh=Coherence, Flu=Fluency, Rep=Repetition, Term=Termination. ♠ and ♣ denote SFT and DPO respectively. *w/o Recur.* removes recurrent context management; *w/o Both* removes both filtering and recurrent context.

per section in parallel. The retrieved candidates are subsequently fed to backbone models for answer synthesis, without any relevance check or noise filtering.

Storm-MM Baseline. To provide a stronger agentic baseline, we construct Storm-MM, a minimally adapted STORM-style (Shao et al., 2024) agent under the same controlled environment as DEEP-REPORTER (same backbone LLM, same retrieval tool and sandbox, same evaluation metrics). We retain STORM's core multi-turn research pipeline and make three necessary changes to enable multimodal output: (i) the agent issues both text and image retrieval queries during multi-turn research, (ii) retrieved text chunks and image captions are merged into a unified multimodal evidence pool, and (iii) the section-level writer is prompted to generate output with both text and image citations.

4.3 Evidence Quality Analysis

Table 2 presents the multimodal evidence quality at three stages of the agentic pipeline.

Raw multimodal search yields comparable evidence quality across methods. Under identical retrieval settings, raw search performance remains largely consistent across both textual and visual modalities. For example, with the Qwen3-32B backbone, naïve RAG achieves a search precision of 33.9 in overall, while DEEP-REPORTER (Base) attains a comparable score of 33.6. This consistency confirms that all methods operate under a stable and fair retrieval environment, removing the impact of raw retrieval on downstream modules.

Agentic multimodal search and filtering substantially concentrates evidence. Despite similar raw retrieval quality, DEEP-REPORTER achieves significant gain on evidence precision after the filter stage, across all backbones. With Qwen3-8B, overall precision increases from 33.6 to 47.9 after filtering, while retaining a substantial volume of evidence (81.0), which markedly reduces the scale of multimodal context compared to raw retrieval. These results show that the agentic *Searcher–Filter* design effectively transforms noisy multimodal retrieval outputs into a concentrated evidence set optimized for downstream usage.

Training activates reliable multimodal evidence selection. The training variants consistently outperform their base counterparts for evidence selection, especially for images. For Qwen3-8B, training increases image Selection precision from 8.3 to 45.0, while expanding the number of cited images from 0.7 to 9.8 per report (13x improvement). This suggests that training does not merely refine retrieval behavior, but actively teaches the model to identify and integrate information-dense visual evidence into the narrative flow, enabling effective multimodal grounding in long-form generation.

4.4 Main Results: Multimodal Generation

Table 3 validates our core contributions via systematic comparisons.

The agentic workflow enables coherent multimodal long-form generation. Naïve RAG collapses to short and shallow outputs: with Qwen3-32B, it produces reports of only 2.0k tokens and achieves a low overall score of 5.2. In contrast, DEEP-REPORTER (Base) generates substantially longer reports (3.9k tokens) and reaches an overall score of 27.2, corresponding to a 4.2× improvement in generation quality. This gain reflects a consistent improvement across structural, multimodal, and holistic dimensions. DEEP-REPORTER (Base) exhibits strong adherence to section anchors (Section Anchor Avg: 0.6 → 38.5), substantially richer and more coherent multimodal content (Section Content Avg: 6.4 → 12.2), and markedly improved full-report coherence and fluency (Full Report Avg: 8.6 → 30.9). To further test whether the gains come from framework design rather than simply from iterative search, we compare against Storm-MM (§4.2), a stronger STORM-style iterative agentic baseline. Although Storm-MM already surpasses Naïve RAG (overall: 5.2 → 15.6), DEEP-REPORTER (Base) still leads by a large margin (overall: 27.2 vs. 15.6; Sel. Overall: 34.1 vs. 30.5). Together, these results show that the agentic workflow of DEEP-REPORTER can successfully transform multimodal retrieval results into structured, visually grounded long-form reports.

Post-training can boost multimodal generation in all dimensions. Building on a strong agentic base, post-training yields further substantial improvements in section anchor and content, and overall report. **1. For backbones that perform poorly in the base setting:** these gains are broad and pronounced: Llama3.3-70B improves from an overall score of 6.9 to 34.8 (over 5×), reflecting large and consistent improvements across structural adherence (6.9x), multimodal content quality (4.4x), and full-report coherence (4.3x) rather than isolated metric gains. **2. For stronger backbones:** post-training further strengthens performance across all dimensions while preserving generation stability. On Qwen3-32B, training increases the overall score from 27.2 to 37.9 (+39%) and extends report length from 3.9k to 4.7k tokens, with the most pronounced improvement observed in multimodal content, where Section Content Avg rises from 12.2 to 41.7. Together, these results show that post-training consistently enhances multimodal grounding and long-form synthesis quality by activating missing capabilities in weaker models and boosting existing strengths in stronger ones.

4.5 Ablation Analysis

We conduct ablations to isolate the effects of filtering, recurrent context management, model scale, and training objectives.

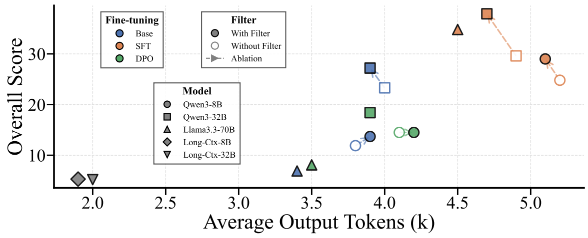


Figure 3: Overall performance with output tokens.

Filtering stabilizes multimodal generation by reducing context noise. Removing the Filter module consistently degrades generation quality, with the largest drops in section-level multimodal content. For Qwen3-32B♠, overall performance decreases from 37.9 to 29.6, while Section Content Avg drops sharply from 41.7 to 27.6. Specifically, raw retrieval yields over 150 candidates on average, whereas filtering compresses the evidence set to around 80 items, concentrating relevance and alleviating context overload. These results indicate that filtering prevents context pollution that disrupts image–text integration and citation decisions, rather than merely improving retrieval metrics. Notably, filtering dominates runtime (∼70% of total latency, Appendix E.1); post-training, in contrast, adds no inference overhead, making all quality gains from SFT essentially free.

Recurrent context management is essential and complementary to filtering. To isolate the impact of recurrent context management (§2.1), we evaluate two additional ablations on Qwen3-32B♠: *w/o recurrent ctx* removes the recurrent context update while keeping filtering, and *w/o both* removes both modules simultaneously. Removing recurrent context management causes the largest drop in Full Report quality (32.3 → 19.8), indicating that the long-context LLM alone cannot maintain cross-section coherence without explicit context accumulation. In contrast, removing the filter produces the largest drop in Section Content (41.7 → 27.6) and selection precision (Sel. Overall: 45.0 → 36.8 in Table 2). Removing both modules compounds the degradation (overall: 37.9 → 26.3; Sel. Overall: 45.0 → 34.5), confirming that these two components address complementary challenges: recurrent context for long-range coherence vs. filtering for evidence quality.

Model scale raises the ceiling, but agentic orchestration drives the qualitative shift. Larger backbones yield higher performance once the agentic workflow is in place, but scale alone is insufficient. Under Naïve RAG, Qwen3-32B underperforms Qwen3-8B (5.2 vs 5.3), whereas with

DEEP-REPORTER, Qwen3-32B substantially outperforms Qwen3-8B both before (27.2 vs 13.7) and after training (37.9 vs 29.0). Figure 3 shows that, at comparable lengths, larger models occupy higher-quality regions. Overall, scale amplifies performance, while agentic orchestration enables the transition from short summaries to grounded long-form reports.

Supervision granularity determines the effectiveness of training objectives. SFT consistently outperforms DPO for long-form multimodal generation. For Qwen3-32B, SFT achieves 37.9 overall versus 18.4 for DPO, with a large gap in Section Content (41.7 vs 11.6); similar trends hold for Qwen3-8B. This difference is consistent with supervision granularity: image selection and placement are sparse decisions in long trajectories, for which token-level SFT provides direct learning signals, while trajectory-level preference optimization offers weaker supervision (more details in § E).

Generalization to open web. To validate that findings on our sandbox transfer to an open environment, we conduct a small-scale experiment: for 100 tasks (Qwen3-32B backbone), we replace the sandbox retrieval tool with the Google Search API and re-run DEEP-REPORTER on the augmented evidence pool. Results show that 83.6% of URLs returned by live Google Search are already present in our sandbox, confirming that the corpus construction effectively covers what the open web provides for these tasks. Moreover, DEEP-REPORTER maintains comparable generation quality on live web evidence (Gen. Overall: 26.2 vs. 27.2 on sandbox), with no significant degradation in Section Anchor (40.3 vs. 38.5) or Full Report quality (28.6 vs. 30.9). These results demonstrate that the retrieval component is modular and readily replaceable, and that the factual grounding is preserved even in an open environment.

5 Related Work

From RAG to Autonomous Research Agents. RAG (Lewis et al., 2020) and ReAct (Yao et al., 2022) underpin modern agentic systems. Recent work extends them to autonomous, multi-step research agents, where explicit planning improves search trajectories (Zheng et al., 2025; Xue et al., 2025; Hu et al., 2025; Shi et al., 2025; Zhang et al., 2026a), search decision marking (Zhang et al., 2026b). However, existing agents remain predominantly *text-centric* (Geng et al., 2025; Lin et al., 2025): visual artifacts are linearized or treated as weak retrieval cues, losing their structured semantics (Yu et al., 2024), and rarely serve as first-class evidence in the reasoning loop (Wasserman et al., 2025; Abootorabi et al., 2025). While multimodal RAG pipelines such as MMDocRAG (Dong et al., 2025b) and M-LongDoc (Chia et al., 2025) introduce cross-modal retrieval, they target single-round document-level QA with short outputs, which differs fundamentally from the iterative, multi-section long-form generation addressed by our work.

Multimodal Long-Form Generation. Current systems face a fundamental trade-off between multimodality and long-range generation. Large Multimodal Models perform well on short interleaved reasoning (Li et al.; Du et al., 2025a) but degrade at document scale, while long-form text generators achieve stable 10k+ outputs (Bai et al., 2024a; Gu et al., 2025; He et al., 2025; Guo et al., 2025) without grounded visual evidence. Although recent agentic approaches synthesize charts or multimodal reports (Li et al., 2025a; Kaur et al., 2025; Yang et al., 2025b), they favor synthetic presentation over evidence fidelity, leaving authentic visual artifacts unsupported as first-class evidence.

6 Conclusion

We present DEEP-REPORTER, a unified agentic framework that advances long-form generation from text-centric to truly multimodal research capabilities. Through agentic multimodal search and filtering, checklist-guided incremental synthesis, and recurrent context management, our framework enables coherent integration of textual and visual evidence in long-form generation. We construct 8K high-quality agentic traces via expert-in-the-loop curation. Our comprehensive benchmark, M^2LONGBENCH, establishes a rigorous and reproducible testbed with 247 research tasks and a stable multimodal sandbox. Extensive experiments reveal that relevance-aware filtering and recurrent context management are complementary components critical for evidence quality and long-range coherence, and that training with curated agentic trajectories activates multimodal selection and integration capabilities. DEEP-REPORTER also generalizes to the open web. These findings validate the effectiveness of DEEP-REPORTER and the value of our data curation pipeline. We hope DEEP-REPORTER serves as a foundation for future research in grounded, multimodal content creation.

Limitations

Despite strong performance, our framework has several limitations:

Static Multimodal Sandbox. While our sandbox ensures fairness, transparency, and reproducibility through its controlled environment, it remains static by design. This presents challenges when users need to explore topics outside the predefined domains. However, our sandbox is inherently scalable. We encourage users to leverage our curation pipeline to expand the corpus using seed topics before performing deep research on brand-new out-of-domain subjects.

Offline Optimization Only. Our current optimization is limited to supervised fine-tuning (SFT) and direct preference optimization (DPO), without exploring online reinforcement learning methods. We conducted preliminary experiments with GRPO but found limited success, which we attribute to the inherent difficulty of open-ended RL in settings where rewards are sparse and hard to determine. We encourage future work to investigate more sophisticated online RL approaches tailored for multimodal long-form generation.

English-Only Focus. Our framework and benchmark currently support English only. Extending DEEP-REPORTER to multilingual settings remains an important direction for future work, particularly given the growing demand for research capabilities across diverse languages and cultural contexts.

Ethical Considerations

Intended use and usage constraints. The benchmark and datasets introduced in this work are intended solely for academic research, including the evaluation and analysis of multimodal agentic retrieval systems. They are not designed for real-world deployment, commercial use, or high-stakes applications. All benchmark components are constructed in accordance with the intended use and license conditions of the underlying resources. When source artifacts are restricted to research-only or non-commercial use, all derived data inherit the same constraints.

Bias and representation. Retrieved multimodal evidence may reflect biases present in underlying data sources, including imbalanced geographic, cultural, or demographic representation.

Privacy and sensitive content. Our data curation pipeline avoids private or personally identifiable information and relies exclusively on publicly available, non-personal research-oriented materials. The benchmark tasks focus on analytical and explanatory capabilities rather than personal profiling or surveillance-related use cases.

Use of AI assistants. Large language models were used as auxiliary tools during benchmark construction, including query formulation and draft synthesis for silver annotations. Most AI-generated outputs were reviewed and refined by human experts, who remained fully responsible for task design, annotation, and the final conclusions. ChatGPT was additionally used to improve writing quality and presentation.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive feedback, which significantly improved this work. We also thank the domain experts who contributed to task curation, outline refinement, and silver annotation verification.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024a. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024b. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu.

2022a. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10437–10449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022b. Aeg: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5134–5148.

Jonas Becker, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2024. Text generation: A systematic literature review of tasks, evaluation, and challenges. *arXiv preprint arXiv:2405.15604*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

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

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2025. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9244–9261.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025a. Mmdocir: Benchmarking multi-modal retrieval for long documents. *arXiv preprint arXiv:2501.08828*.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025b. Benchmarking retrieval-augmented multimomal generation for document question answering. *arXiv preprint arXiv:2505.16470*.

Kuicai Dong, Shurui Huang, Fangda Ye, Wei Han, Zhi Zhang, Dexun Li, Wenjun Li, Qu Yang, Gang Wang, Yichao Wang, and 1 others. 2025c. Doc-researcher: A unified system for multimodal document parsing and deep research. *arXiv preprint arXiv:2510.21603*.

Hang Du, Jiayang Zhang, Guoshun Nan, Wendi Deng, Zhenyan Chen, Chenyang Zhang, Wang Xiao, Shan Huang, Yuqi Pan, Tao Qi, and 1 others. 2025a. From easy to hard: The mir benchmark for progressive interleaved multi-image reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 859–869.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025b. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, and 1 others. 2025. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*.

Alireza Ghafarollahi and Markus J Buehler. 2025. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.

Hongchao Gu, Dexun Li, Kuicai Dong, Hao Zhang, Hang Lv, Hao Wang, Defu Lian, Yong Liu, and Enhong Chen. 2025. Rapid: Efficient retrieval-augmented long text generation with writing planning and information discovery. *arXiv preprint arXiv:2503.00751*.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and 1 others. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550.

Zhihan Guo, Jiele Wu, Wenqian Cui, Yifei Zhang, Minda Hu, Yufei Wang, and Irwin King. 2025. From general reward to targeted reward: Improving open-ended long-context generation models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5151–5166.

Jacqueline He, Howard Yen, Margaret Li, Shuyue Stella Li, Zhiyuan Zeng, Weijia Shi, Yulia Tsvetkov, Danqi Chen, Pang Wei Koh, and Luke Zettlemoyer. 2025. Precise information control in long-form text generation. *arXiv preprint arXiv:2506.06589*.

Yusong Hu, Runmin Ma, Yue Fan, Jinxin Shi, Zongsheng Cao, Yuhao Zhou, Jiakang Yuan, Xiangchao Yan, Wenlong Zhang, Lei Bai, and 1 others. 2025. Flowsearch: Advancing deep research with dynamic structured knowledge flow. *arXiv preprint arXiv:2510.08521*.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Rachneet Kaur, Nishan Srishankar, Zhen Zeng, Sumitra Ganesh, and Manuela Veloso. 2025. Chartagent: A multimodal agent for visually grounded reasoning in complex chart question answering. *arXiv preprint arXiv:2510.04514*.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, and 1 others. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Bingxuan Li, Yiwei Wang, Jiuxiang Gu, Kai-Wei Chang, and Nanyun Peng. 2025a. Metal: A multi-agent framework for chart generation with test-time scaling. *arXiv preprint arXiv:2502.17651*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Wenjun Li, Zhi Chen, Jingru Lin, Hannan Cao, Wei Han, Sheng Liang, Zhi Zhang, Kuicai Dong, Dexun Li, Chen Zhang, and Yong Liu. 2025b. Reinforcement learning foundations for deep research systems: A survey. *arXiv preprint arXiv: 2509.06733*.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

Jingru Lin, Chen Zhang, Stephen Y. Liu, and Haizhou Li. 2025. Ragcap-bench: Benchmarking capabilities of llms in agentic retrieval augmented generation systems. *Preprint*, arXiv:2510.13910.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2024. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, and 1 others. 2025. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. 2024. Adapt: As-needed decomposition and planning with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4226–4252.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.

Wenxuan Shi, Haochen Tan, Chuqiao Kuang, Xiaoguang Li, Xiaozhe Ren, Chen Zhang, Hanting Chen, Yasheng Wang, Lu Hou, and Lifeng Shang. 2025. DeepDiver: Adaptive search intensity scaling via open-web reinforcement learning. *arXiv preprint arXiv: 2505.24332*.

MiroMind Team, S Bai, L Bing, L Lei, R Li, X Li, X Lin, E Min, L Su, B Wang, and 1 others. 2026. Mirothinker-1.7 & h1: Towards heavy-duty research agents via verification. *arXiv preprint arXiv:2603.15726*.

Yong-En Tian, Yu-Chien Tang, Kuang-Da Wang, An-Zi Yen, and Wen-Chih Peng. 2025. Template-based financial report generation in agentic and decomposed information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2706–2710.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. 2025. Pairwise or pointwise? evaluating feedback protocols for bias in llm-based evaluation. *arXiv preprint arXiv:2504.14716*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

UncleCode. 2024. Crawl4ai: Open-source llm friendly web crawler & scraper. https://github.com/unclecode/crawl4ai.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. Real-mm-rag: A real-world multi-modal retrieval benchmark. *arXiv preprint arXiv:2502.12342*.

Siwei Wu, Yizhi Li, Xingwei Qu, Rishi Ravikumar, Yucheng Li, Tyler Loakman, Shanghaoran Quan, Xiaoyong Wei, Riza Batista-Navarro, and Chenghua Lin. 2025a. Longeval: A comprehensive analysis of long-text generation through a plan-based paradigm. *arXiv preprint arXiv:2502.19103*.

Yuhao Wu, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. 2024. Longgenbench: Benchmarking long-form generation in long context llms. *arXiv preprint arXiv:2409.02076*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and 1 others. 2025b. Writing-bench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*.

Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu, and Weinan Zhang. 2025. A survey of llm-based deep search agents: Paradigm, optimization, evaluation, and challenges. *arXiv preprint arXiv:2508.05668*.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. 2025. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. *arXiv preprint arXiv:2509.02479*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Luoxuan Weng, Yingchaojie Feng, Haozhe Feng, Minfeng Zhu, Bo Zhang, and 1 others. 2025b. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework. *arXiv preprint arXiv:2506.02454*.

Zhongyu Yang, Jun Chen, Dannong Xu, Junjie Fei, Xiaoqian Shen, Liangbing Zhao, Chun-Mei Feng, and Mohamed Elhoseiny. 2025c. Wikiautogen: Towards multi-modal wikipedia-style article generation. *arXiv preprint arXiv:2503.19065*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Fangda Ye, Yuxin Hu, Pengxiang Zhu, Yibo Li, Ziqi Jin, Yao Xiao, Yibo Wang, Lei Wang, Zhen Zhang, Lu Wang, and 1 others. 2026. Miroeval: Benchmarking multimodal deep research agents in process and outcome. *arXiv preprint arXiv:2603.28407*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Chen Zhang, Kuicai Dong, Dexun Li, Wenjun Li, Qu Yang, Wei Han, and Yong Liu. 2026a. SRR-Judge: Step-level rating and refinement for enhancing search-integrated reasoning in search agents. *arXiv preprint arXiv: 2602.07773*.

Wenlin Zhang, Kuicai Dong, Junyi Li, Yingyi Zhang, Xiaopeng Li, Pengyue Jia, Yi Wen, Derong Xu, Maolin Wang, Yichao Wang, Yong Liu, and Xiangyu Zhao. 2026b. To search or not to search: Aligning the decision boundary of deep search agents via causal intervention. *Preprint*, arXiv:2602.03304.

Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. 2025. Process vs. outcome reward: Which is better for agentic rag reinforcement learning. *Preprint*, arXiv:2505.14069.

Wenlin Zhang, Xiangyang Li, Qiyuan Ge, Kuicai Dong, Pengyue Jia, Xiaopeng Li, Zijian Zhang, Maolin Wang, Yichao Wang, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. 2026c. Exploring recommender system evaluation: A multi-modal user agent framework for a/b testing. *Preprint*, arXiv:2601.04554.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2025. Docbench: A benchmark for evaluating llm-based document reading systems. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 359–373.

Appendix Overview

This appendix provides detailed supplementary materials supporting the contributions of DEEP-REPORTER. The content is strictly organized as follows:

- **Appendix A: M^{2}LONGBENCH Construction Details.**
  - §A.1: Sandbox Construction and Implementation Details.
  - §A.2: Benchmark Statistics.
  - §A.3: Analysis of human labor, time investment, and computational resources.
  - §A.4: The specific LLM and VLM prompts used for data curation and filtering.
  - §A.5: Detailed visual breakdown of the annotation workflow.
- **Appendix B: Experimental Setup Details.**
  - §B.1:Specific model versions, revision hashes and licenses.
  - §B.2: Detailed system configuration and more environment settings.
  - §B.3: Full prompts for Planner, Searcher, Filter, and Reporter agents.
- **Appendix C: Training Implementation.**
  - §C.1: SFT hyperparameters and compute resource usage.
  - §C.2: DPO dataset construction methodology and training parameters.
- **Appendix D: Evaluation Details.**
  - §D.1: Detailed scoring rubrics.
  - §D.2: Final Score calculation logic.
  - §D.3: Human-LLM alignment study and discussion.
  - §D.4: Breakdown of evaluation costs.
- **Appendix E: Additional Experimental Results.**
  - §E.1: Analysis of generation latency and inference costs.
  - §E.2: Capability Radar Charts and fine-grained Domain Performance tables.
  - §E.3: Analysis of DPO Underperformance.
- **Appendix F: Qualitative Case Studies.**
  - Display of 3 generated report samples.

A M^{2}LONGBENCH Construction Details

A.1 Sandbox Construction and Implementation Details

More Details. We implement the sandbox as a multimodal retrieval and grounding pipeline (Figure 4), with a modular design similar to (Dong et al., 2025c).

We start from curated analytical reports and their reference URLs, and run a lightweight ingestion pipeline (URL deduplication, concurrent fetching, extraction/cleaning) to normalize heterogeneous sources. For web content, we convert pages into Markdown to preserve structural cues (e.g., headings, lists, tables) while keeping links to the original sources.

For PDF reports, we convert them into Markdown using MinerU v2.0.6, and apply the same processing pipeline used for web-extracted Markdown content. Textual content is segmented at subheading boundaries and each paragraph is retained to maintain semantic integrity. Short paragraphs are merged until the length approaches but not exceeds 350 words. Images smaller than 10KB are removed and each visual chunk retains its page index and layout bounding box for localization.

We embed all chunks and images using jina-embeddings-v4 (Jin et al., 2025). To improve text-to-visual access and cross-modal alignment, we caption every visual chunk with InternVL3.5-38B (Wang et al., 2025). We index embeddings in Milvus for dense top-K retrieval, and store metadata and structured artifact in MySQL. During retrieval, image candidates are returned with their original file paths, captions, as well as base64-encoded visual contents, allowing downstream models to directly ingest images for further cross-modal reasoning and generation.

Sandbox Extensibility. Beyond the curated sources used in this release, the M^{2}LONGBENCH sandbox is designed to be *incrementally extensible*. Given a user-specified *seed topic* (optionally with a small set of seed URLs), we provide an automated discovery script that performs topic-guided search, URL harvesting, and controlled web crawling to collect candidate evidence pages and reports for future sandbox expansion. The script applies the same normalization and deduplication rules as our ingestion pipeline, preserves source provenance (original URLs and link graph metadata), and supports incremental indexing into our storage (MySQL) and dense retrieval backend (Milvus).

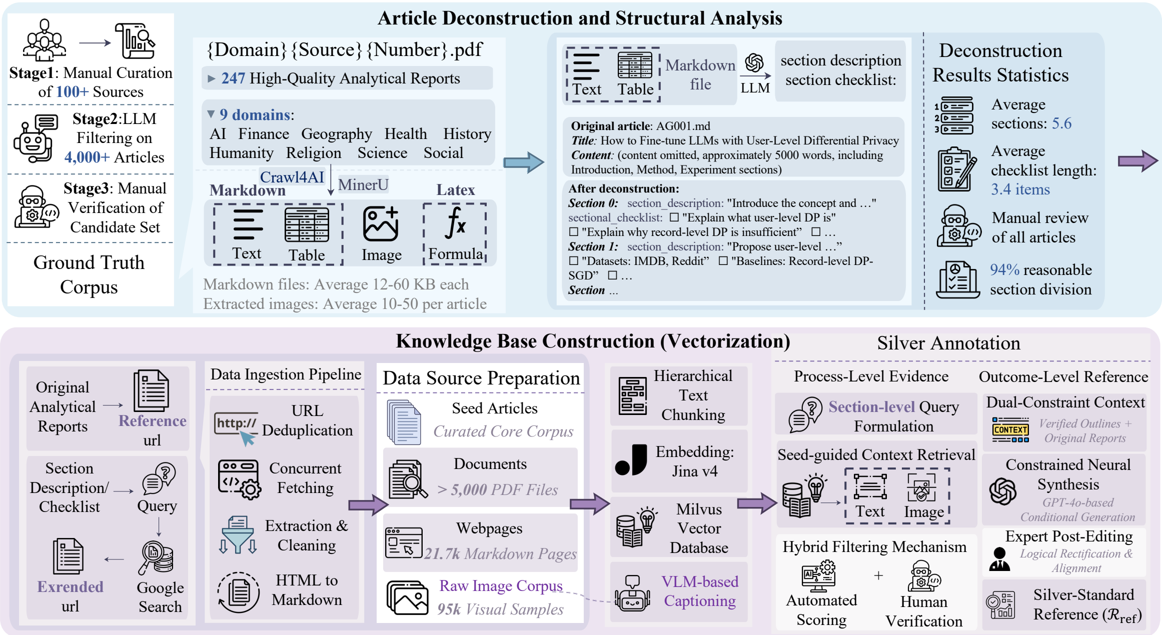


Figure 4: Demonstration of the sandbox construction.

This tooling makes it straightforward to grow the sandbox over time with new domains and emerging topics while maintaining a stable, transparent, and reproducible retrieval environment.

Personally Identifiable and Offensive Content. All data sources are publicly accessible authoritative repositories intended for open information dissemination. During data collection and processing, we explicitly filter out personally identifiable information (PII), sensitive personal data, and offensive or harmful content. The curated tasks and sandbox materials focus exclusively on technical, scientific, and analytical content and do not involve private individuals or user-generated data.

A.2 Benchmark Statistics

Table 4 provides a comprehensive statistical breakdown of M^{2}LongBench. The data quantitatively validates our design choices in addressing the limitations of prior benchmarks:

Sandbox Sufficiency: Replacing the Live Web. A core contribution of M^{2}LongBench is decoupling long-form research from the instability and opacity of live search engines without sacrificing information density. As shown in the *Sandbox Scale* columns of Table 4, we construct a retrieval environment that is informationally complete. This scale ensures that the sandbox contains sufficient "distractor" noise and diverse evidence to mimic real-world discovery, effectively serving as a stable, transparent, and accessible alternative to commercial web search APIs.

Multimodal Complexity. To bridge the gap in multimodal long-form generation, M^{2}LongBench presents tasks of significant depth. The *Source Report Metadata* in Table 4 indicates that the ground-truth blueprints average 5.5k tokens and 9.5 images. This complexity compels agents to move beyond simple factoid retrieval, requiring them to synthesize narrative structures and integrate visual evidence coherently across lengthy documents.

Structural Granularity for Rigorous Evaluation. Finally, to support our *Unified Multimodal Assessment*, the dataset features high structural specificity. The *Tasks* columns reveal that each report is decomposed into an average of 5.6 sections, guided by 3.4 fine-grained checklist constraints per section. This density of explicit instructions (totaling ∼19 constraints per task) transforms the vague "open-ended generation" problem into a measurable process, enabling precise adherence checking against the silver-standard annotations (167.9 chunks and 102.6 images per report).

A.3 Resource and Cost Analysis

The construction of M^{2}LongBench required substantial investment across human expertise, computational infrastructure, and API services. The resource breakdown is detailed as follows:

- **Human Labor:** Human participants were recruited via open and transparent channels (e.g.,

Domain	Source Report Metadata			Tasks		Sandbox Scale (Original / Expanded)				Silver Annot.	
	Count	Image	Token	Sec.	Check.	Web Pages	PDF Docs	Images	Chunks	Images	Chunks
AI	43	12.2	7.3k	5.7	3.6	28.5 / 59.8	16.8/ 27.9	39.7 / 289.9	103.0k / 179.3k	98.7	170.2
Finance	37	10.8	7.5k	5.9	3.4	16.4 / 79.5	1.7/ 13.3	20.8 / 285.9	59.3k / 238.2k	110.5	176.0
Geography	12	8.8	4.1k	5.6	3.0	8.3 / 80.1	2.0 / 9.5	45.6 / 548.3	29.9k / 240.0k	99.7	165.3
Health	28	7.0	4.2k	5.8	3.3	19.2 / 77.0	5.5 / 6.7	45.3 / 287.5	70.0k / 230.6k	108.5	172.5
History	22	7.4	3.3k	5.1	3.1	3.3 / 59.0	1.5 / 11.1	8.3 / 592.6	23.0k / 324.1k	99.1	154.0
Humanity	34	6.5	4.1k	5.1	3.4	11.3 / 62.6	3.6 / 7.3	26.0 / 263.3	199.1k / 911.2k	92.3	151.8
Religion	15	10.5	5.5k	6.1	3.5	17.4 / 74.0	1.8/ 5.8	66.4 / 342.5	63.0k / 221.7k	116.3	183.9
Science	41	10.3	5.6k	5.4	3.2	17.7 / 77.0	2.6 / 10.9	58.8 / 387.9	66.4k / 230.7k	96.7	162.4
Social	15	9.6	5.1k	6.4	3.7	20.0 / 78.4	2.4 / 9.2	73.3 / 263.9	72.4k / 234.9k	115.8	190.0
Overall	247	9.5	5.5k	5.6	3.4	17.1 / 70.9	7.6 / 12.8	41.8 / 342.8	82.9k / 313.2k	102.6	167.9

Table 4: **Statistical overview of M^{2}LongBench. Source Report Metadata** reflects the complexity of the ground-truth blueprints. **Sandbox Scale** highlights the difficulty of the retrieval task by contrasting original citations with our significantly expanded search space (Expanded), demonstrating the "needle-in-a-haystack" nature of the benchmark.

poster advertisements), primarily consisting of senior undergraduate students, graduate students, and Ph.D. candidates with relevant academic backgrounds. We devoted approximately 420 person-hours to benchmark construction and verification. This includes expert-driven task curation and topic selection (approximately 120 hours), manual review and refinement of section outlines and checklists (approximately 120 hours), and silver annotation with relevance verification across candidate textual and visual evidence, as well as manual review and refinement of the silver reports (approximately 180 hours). All participants were compensated at a rate of approximately $20 per hour. Human supervision was critical for ensuring task diversity, outline correctness, and high-quality multimodal relevance judgments.

- **Data Processing:** Computational preprocessing was conducted in an offline and fully reproducible manner. Specifically, we processed 45K webpages using Crawl4AI (approximately 36 CPU-hours), converted 6.4K PDFs using MinerU (approximately 220 GPU-hours), generated captions for 95K images (approximately 310 GPU-hours). All processing steps were performed once and reused across experiments.
- **API Usage (GPT-4.1).** We utilized GPT-4.1 primarily for query decomposition, outline validation, and relevance scoring during silver annotation. In total, we consumed approximately 18M tokens for query and outline generation and 26M tokens for relevance assessment and verification.

The total duration of the construction pipeline, from initial crawling to final indexing, was approximately 3 months.

A.4 Prompts for Data Curation

We provide the core prompts used in the data pipeline below.

A.5 Annotation Workflow Visualization

As described below, the annotation workflow proceeds through four stages: information extraction, article deconstruction, augmented search with embedding, and retrieval-driven silver annotation.

Prompt Template: Structured Table & Figure Decomposition

```
Given an image containing a table or figure, please provide a structured and detailed description with two levels of granularity:

Coarse-grained Description:
- Summarize the overall content and purpose of the image.
- Briefly state what type of data or information is presented (e.g., comparison, trend, distribution).
- Mention the main topic or message conveyed by the table or figure.

Fine-grained Description:
- Describe the specific details present in the image.
- For tables: List the column and row headers, units, and any notable values, patterns, or anomalies.
- For figures (e.g., plots, charts): Explain the axes, data series, legends, and any significant trends, outliers, or data points.
- Note any labels, captions, or annotations included in the image.
- Highlight specific examples or noteworthy details.

Deliver the description in a clear, organized, and reader-friendly manner, using bullet points or paragraphs as appropriate.
```

Prompt Template: Semantic Article Deconstruction & JSON Structuring

```
<system_role>
You are a meticulous and precise Technical Content Analyst. Your expertise lies in faithfully deconstructing written articles into a structured format that accurately preserves the original author's intent, structure, and flow of topics. Your output must be a clean, precise representation of the source material.
</system_role>

<user_prompt>
**Primary Goal**
Your task is to analyze the provided article in Markdown format and convert it into a structured JSON object. It is crucial that the generated outline and checklists **faithfully reflect the structure, sequence, and key points of the original article**.

**Input Article**
<article_content>
{article_markdown_content}
</article_content>

**Your Cognitive Process (Instructions)**
You MUST follow this structured thinking process:

1.  **De-noise Content**: First, mentally filter out and ignore all non-essential elements: advertisements, self-promotion, conversational filler ("Hey everyone," "Thanks for reading"), website navigation elements, and any metadata. Focus only on the substantive content that advances the main argument or narrative.

2.  **Infer Core Intent**: Based on the core content, determine the central topic to formulate the `query` and identify the key promises or questions answered to create the `overall_checklist`.

3.  **Identify Natural Content Divisions**:
   - **CRITICAL**: Completely ignore markdown heading hierarchy (#, ##, ###) as it may be corrupted or inconsistent due to web scraping issues.
   - Instead, read through the entire article and identify natural logical breaks where the discussion shifts to a substantially different aspect of the main topic.
   - Look for transition phrases, topic changes, shifts in perspective, or movement from one major concept to another.
   - **Target 3-8 sections total** - aim for meaningful, substantial sections rather than many small ones.
   - Each section should represent a complete thought or major component of the overall argument.

4.  **Section Boundaries Guidelines**:
   - A new section should only begin when there's a clear shift in focus, not just a new paragraph or minor sub-point.
```

```
   - Consider combining related content that might appear under separate headings if they serve
     the same logical purpose.
   - Look for natural narrative flow: Introduction -> Main Arguments/Evidence -> Analysis ->
     Conclusions, or Problem -> Methods -> Results -> Implications, etc.
   - Avoid creating sections that are too short (less than 2-3 substantial paragraphs) or too
     numerous (more than 8 sections).

5.  **Section Description and Sub-topics**: For each identified section:
   - Write a comprehensive `section_description` that captures the main thrust and purpose of
     that section.
   - Only include sub-topics if the section genuinely covers multiple distinct concepts that
     would benefit from enumeration.
   - Format sub-topics as: "Key sub-topics include: 1. Topic A details, 2. Topic B details, 3.
     Topic C details" - but only when this truly adds clarity.

6.  **Checklist Allocation**: Distribute the `overall_checklist` items across sections based on
     where those concepts are actually discussed in the article, ensuring comprehensive coverage
     without redundancy.

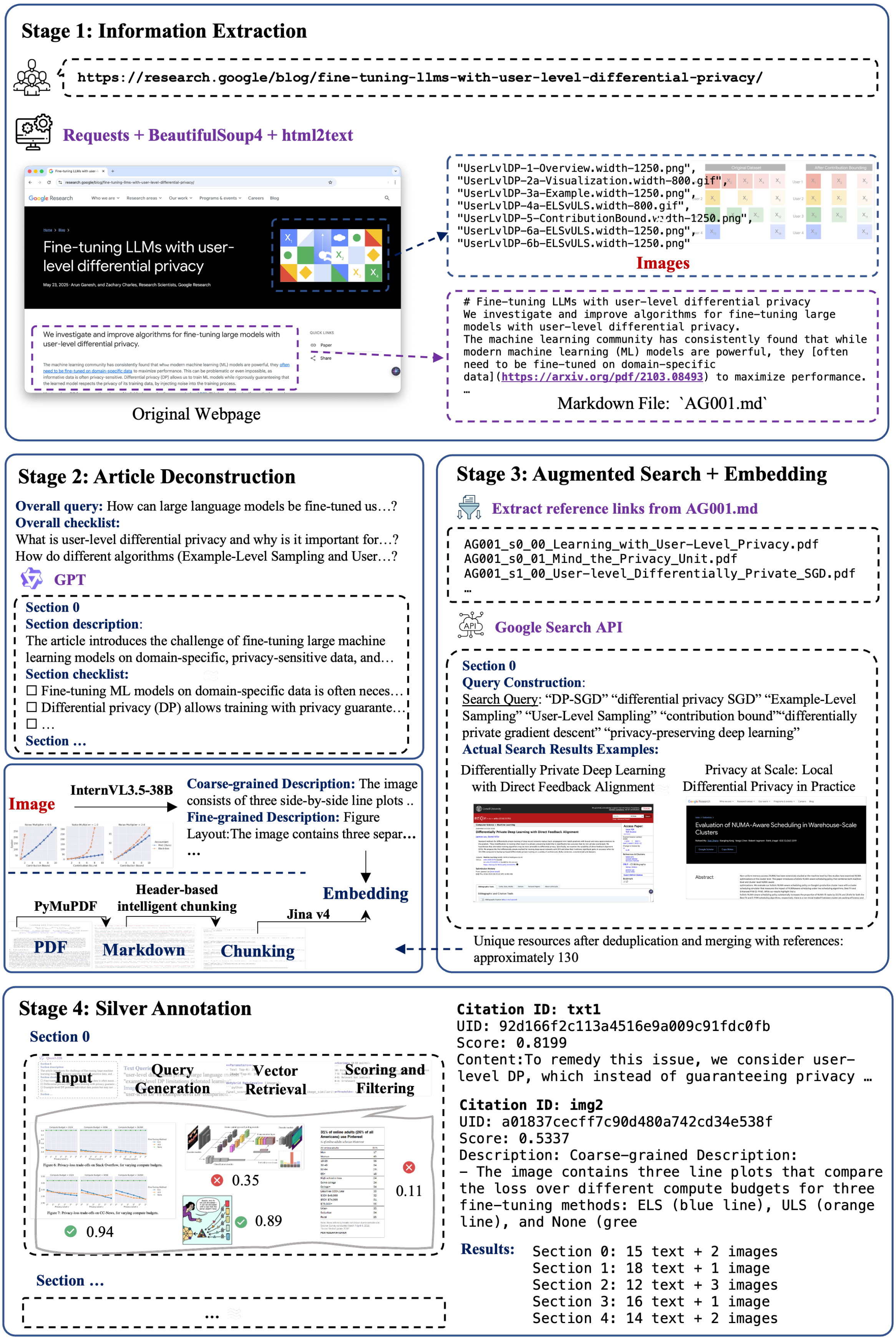
**Strict Output Specification**
Your output MUST be a single, valid JSON object and nothing else. The JSON object must follow
     this exact structure:

<output_format>
{{
  "query": "A concise query that accurately reflects the central topic of the article.",
  "overall_checklist": [
    "A key question the original article answers.",
    "A core concept the original article explains.",
    "A main conclusion or takeaway from the original article."
  ],
  "detailed_outline": [
    {{
      "section_description": "A detailed summary of what this section covers. **If, and only if,
     it is beneficial for clarity,** this field should also contain a list of key sub-topics. You
     can **include, create, or completely omit** sub-topics based on whether they genuinely help
     in structuring the section's content. For example, a simple section may only require a
     summary paragraph. A more complex section might look like: 'This section details the
     experiment results. Key sub-topics include: 1. Result A ..., 2. Result B...'.",
      "sectional_checklist": [
        "A specific point or claim made within this section of the original article.",
        "A definition provided in this section."
      ]
    }},
    {{
      "section_description": "A detailed summary for the next section, following the same rules.",

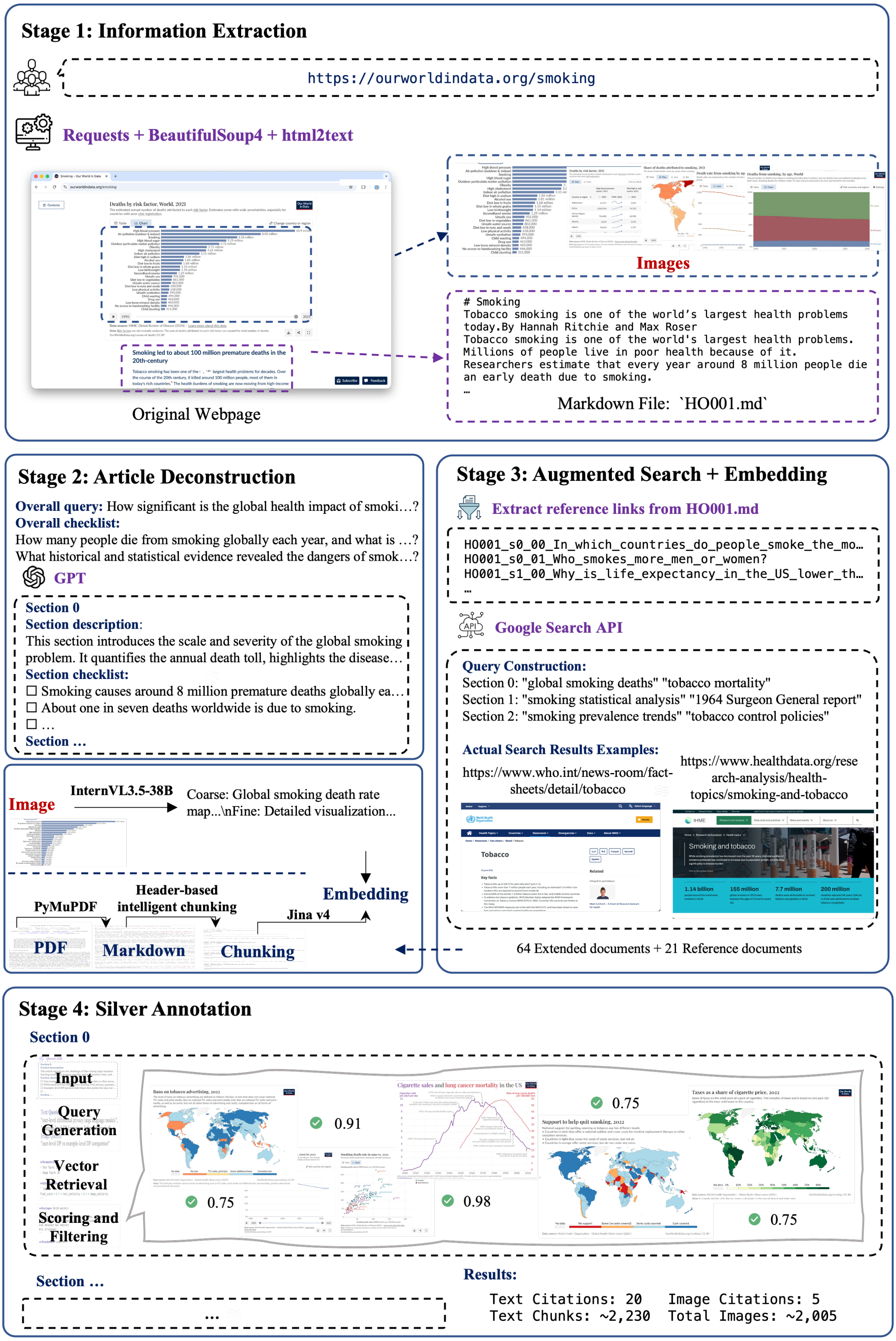
      "sectional_checklist": [
        "Another specific point covered in this part of the original article."
      ]
    }}
  ]
}}
</output_format>

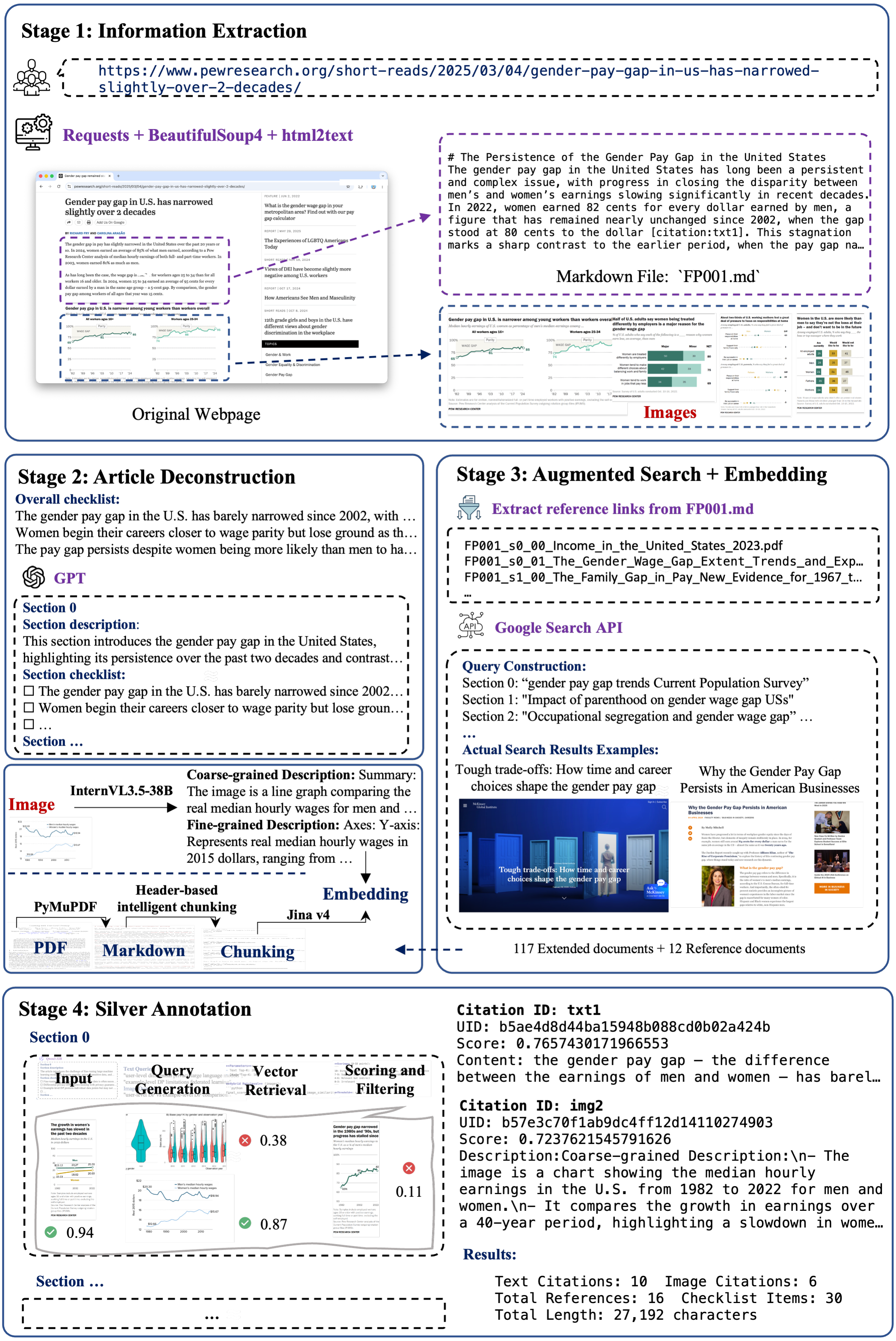
**Critical Constraints**
- Generate between 3-8 sections maximum
- Focus on logical content flow rather than markdown structure
- Each section must be substantial and meaningful
- Ensure sections follow the natural progression of the author's argument
- Do not create micro-sections for minor points

</user_prompt>
```



Stage 1: Information Extraction
https://research.google/blog/fine-tuning-llms-with-user-level-differential-privacy/
Requests + BeautifulSoup4 + html2text
Fine-tuning LLMs with user-level differential privacy
We investigate and improve algorithms for fine-tuning large models with user-level differential privacy.
"UserLvlDP-1-Overview.width-1250.png",
"UserLvlDP-2a-Visualization.width-800.gif",
"UserLvlDP-3a-Example.width-1250.png",
"UserLvlDP-4a-ELSvULS.width-800.gif",
"UserLvlDP-5-ContributionBound.width-1250.png",
"UserLvlDP-6a-ELSvULS.width-1250.png",
"UserLvlDP-6b-ELSvULS.width-1250.png"
Images
Fine-tuning LLMs with user-level differential privacy
We investigate and improve algorithms for fine-tuning large models with user-level differential privacy.
The machine learning community has consistently found that while modern machine learning (ML) models are powerful, they [often need to be fine-tuned on domain-specific data](https://arxiv.org/pdf/2103.08493) to maximize performance.
…
Markdown File: `AG001.md`
Original Webpage
Stage 2: Article Deconstruction
Overall query: How can large language models be fine-tuned us…?
Overall checklist:
What is user-level differential privacy and why is it important for…?
How do different algorithms (Example-Level Sampling and User…?
GPT
Section 0
Section description:
The article introduces the challenge of fine-tuning large machine learning models on domain-specific, privacy-sensitive data, and…
Section checklist:
☐ Fine-tuning ML models on domain-specific data is often neces…
☐ Differential privacy (DP) allows training with privacy guarante…
☐ …
Section …
Image
InternVL3.5-38B
Coarse-grained Description: The image consists of three side-by-side line plots ..
Fine-grained Description: Figure Layout:The image contains three separ…
…
Embedding
Header-based intelligent chunking
PyMuPDF
Jina v4
PDF
Markdown
Chunking
Stage 3: Augmented Search + Embedding
Extract reference links from AG001.md
AG001_s0_00_Learning_with_User-Level_Privacy.pdf
AG001_s0_01_Mind_the_Privacy_Unit.pdf
AG001_s1_00_User-level_Differentially_Private_SGD.pdf
…
Google Search API
Section 0
Query Construction:
Search Query: "DP-SGD" "differential privacy SGD" "Example-Level Sampling" "User-Level Sampling" "contribution bound" "differentially private gradient descent" "privacy-preserving deep learning"
Actual Search Results Examples:
Differentially Private Deep Learning with Direct Feedback Alignment
Privacy at Scale: Local Differential Privacy in Practice
Evaluation of NUMA-Aware Scheduling in Warehouse-Scale Clusters
Unique resources after deduplication and merging with references: approximately 130
Stage 4: Silver Annotation
Section 0
Input
Query Generation
Vector Retrieval
Scoring and Filtering
0.35
0.89
0.94
0.11
Section …
…
Citation ID: txt1
UID: 92d166f2c113a4516e9a009c91fdc0fb
Score: 0.8199
Content:To remedy this issue, we consider user-level DP, which instead of guaranteeing privacy …
Citation ID: img2
UID: a01837cecff7c90d480a742cd34e538f
Score: 0.5337
Description: Coarse-grained Description:
- The image contains three line plots that compare the loss over different compute budgets for three fine-tuning methods: ELS (blue line), ULS (orange line), and None (gree
Results:
Section 0: 15 text + 2 images
Section 1: 18 text + 1 image
Section 2: 12 text + 3 images
Section 3: 16 text + 1 image
Section 4: 14 text + 2 images


Stage 1: Information Extraction
https://ourworldindata.org/smoking
Requests + BeautifulSoup4 + html2text
Original Webpage
Images
Smoking
Tobacco smoking is one of the world's largest health problems
today.By Hannah Ritchie and Max Roser
Tobacco smoking is one of the world's largest health problems.
Millions of people live in poor health because of it.
Researchers estimate that every year around 8 million people die
an early death due to smoking.
…
Markdown File: `HO001.md`
Stage 2: Article Deconstruction
Overall query: How significant is the global health impact of smoki…?
Overall checklist:
How many people die from smoking globally each year, and what is …?
What historical and statistical evidence revealed the dangers of smok…?
GPT
Section 0
Section description:
This section introduces the scale and severity of the global smoking problem. It quantifies the annual death toll, highlights the disease…
Section checklist:
☐ Smoking causes around 8 million premature deaths globally ea…
☐ About one in seven deaths worldwide is due to smoking.
☐ …
Section …
Image
InternVL3.5-38B
Coarse: Global smoking death rate map...\nFine: Detailed visualization...
Embedding
PyMuPDF
Header-based intelligent chunking
Jina v4
PDF
Markdown
Chunking
Stage 3: Augmented Search + Embedding
Extract reference links from HO001.md
HO001_s0_00_In_which_countries_do_people_smoke_the_mo…
HO001_s0_01_Who_smokes_more_men_or_women?
HO001_s1_00_Why_is_life_expectancy_in_the_US_lower_th…
…
Google Search API
Query Construction:
Section 0: "global smoking deaths" "tobacco mortality"
Section 1: "smoking statistical analysis" "1964 Surgeon General report"
Section 2: "smoking prevalence trends" "tobacco control policies"
Actual Search Results Examples:
https://www.who.int/news-room/fact-sheets/detail/tobacco
https://www.healthdata.org/research-analysis/health-topics/smoking-and-tobacco
Tobacco
Smoking and tobacco
64 Extended documents + 21 Reference documents
Stage 4: Silver Annotation
Section 0
Input
Query Generation
Vector Retrieval
Scoring and Filtering
0.91
0.75
0.98
0.75
0.75
Section …
…
Results:
Text Citations: 20   Image Citations: 5
Text Chunks: ~2,230   Total Images: ~2,005


Stage 1: Information Extraction
https://www.pewresearch.org/short-reads/2025/03/04/gender-pay-gap-in-us-has-narrowed-slightly-over-2-decades/
Requests + BeautifulSoup4 + html2text
The Persistence of the Gender Pay Gap in the United States
The gender pay gap in the United States has long been a persistent and complex issue, with progress in closing the disparity between men's and women's earnings slowing significantly in recent decades. In 2022, women earned 82 cents for every dollar earned by men, a figure that has remained nearly unchanged since 2002, when the gap stood at 80 cents to the dollar [citation:txt1]. This stagnation marks a sharp contrast to the earlier period, when the pay gap na…
Markdown File: `FP001.md`
Original Webpage
Images
Stage 2: Article Deconstruction
Overall checklist:
The gender pay gap in the U.S. has barely narrowed since 2002, with …
Women begin their careers closer to wage parity but lose ground as th…
The pay gap persists despite women being more likely than men to ha…
GPT
Section 0
Section description:
This section introduces the gender pay gap in the United States, highlighting its persistence over the past two decades and contrast…
Section checklist:
☐ The gender pay gap in the U.S. has barely narrowed since 2002…
☐ Women begin their careers closer to wage parity but lose groun…
☐ …
Section …
Image
InternVL3.5-38B
Coarse-grained Description: Summary: The image is a line graph comparing the real median hourly wages for men and …
Fine-grained Description: Axes: Y-axis: Represents real median hourly wages in 2015 dollars, ranging from …
PyMuPDF
Header-based intelligent chunking
Embedding
Jina v4
PDF
Markdown
Chunking
Stage 3: Augmented Search + Embedding
Extract reference links from FP001.md
FP001_s0_00_Income_in_the_United_States_2023.pdf
FP001_s0_01_The_Gender_Wage_Gap_Extent_Trends_and_Exp…
FP001_s1_00_The_Family_Gap_in_Pay_New_Evidence_for_1967_t…
…
Google Search API
Query Construction:
Section 0: "gender pay gap trends Current Population Survey"
Section 1: "Impact of parenthood on gender wage gap USs"
Section 2: "Occupational segregation and gender wage gap" …
…
Actual Search Results Examples:
Tough trade-offs: How time and career choices shape the gender pay gap
Why the Gender Pay Gap Persists in American Businesses
117 Extended documents + 12 Reference documents
Stage 4: Silver Annotation
Section 0
Input
Query Generation
Vector Retrieval
Scoring and Filtering
0.38
0.11
0.87
0.94
Section …
…
Citation ID: txt1
UID: b5ae4d8d44ba15948b088cd0b02a424b
Score: 0.7657430171966553
Content: the gender pay gap — the difference between the earnings of men and women — has barel…
Citation ID: img2
UID: b57e3c70f1ab9dc4ff12d14110274903
Score: 0.7237621545791626
Description:Coarse-grained Description:\n- The image is a chart showing the median hourly earnings in the U.S. from 1982 to 2022 for men and women.\n- It compares the growth in earnings over a 40-year period, highlighting a slowdown in wome…
Results:
Text Citations: 10 Image Citations: 6
Total References: 16 Checklist Items: 30
Total Length: 27,192 characters

B Experimental Setup Details

B.1 Model Versions and Licenses

Our experiments employ the following models: Qwen3-8B, Qwen3-32B, Llama-3.3-70B-Instruct, InternVL3-5-38B, and Jina-Embeddings-v4.

Qwen3-8B, Qwen3-32B and InternVL3-5-38B are released under the Apache License, Version 2.0. Jina-Embeddings-v4 are released under the Qwen Research License Agreement. Llama-3.3-70B-Instruct is released under the Llama 3.3 Community License Agreement, which is based on the Apache License, Version 2.0.

All models are used in compliance with their respective licenses as released by the original authors, and no restrictions imposed by these licenses are violated in our experimental setting.

The exact model versions, and revision identifiers are summarized in Table 5.

Model	Revision (Git commit SHA)
Qwen3-8B	`b968826d9c46dd6066d109eabc6255188de91218`
Qwen3-32B	`9216db5781bf21249d130ec9da846c4624c16137`
Llama-3.3-70B-Instruct	`6f6073b423013f6a7d4d9f39144961bfbfbc386b`
InternVL3-5-38B	`de99855be3642cd44fe97c9b72d70e5ce2c07f69`
Jina-Embeddings-v4	`5f4b9cbb80cc95ba44fe6667dfd75710f7db2947`

Table 5: Model versions and Hugging Face Hub revisions used in our experiments.

B.2 System Configuration Details

Checklist-Guided Experimental Control. To ensure controlled and reproducible experiments for open-ended multimodal long-form generation, all reported results condition on a fixed, pre-defined checklist for each task. These checklists are derived during benchmark construction and specify section-level requirements. Accordingly, the planner component in DEEP-REPORTER is disabled in all experiments, and models are compared solely on their ability to retrieve, filter, and synthesize multimodal evidence under identical structural constraints. This design choice isolates the effects of agentic orchestration and evidence utilization from variability introduced by planning.

***System Configuration of* DEEP-REPORTER.** The Filter module is fixed and kept non-trainable across all experiments to ensure consistent relevance assessment. Specifically, the Filter employs `qwen-plus-latest` via API for textual relevance scoring and a locally deployed InternVL3.5-38B model for visual relevance estimation. The Searcher and Reporter agents share the same backbone model within each experiment (e.g., Qwen3-8B or Qwen3-32B) and are jointly trained when post-training is applied. This configuration ensures that performance differences arise from agentic coordination and training effects, rather than changes in filtering criteria.

Training Scope and Consistency. Post-training is applied jointly to the Searcher and Reporter agents using curated agentic trajectories. The Filter module remains unchanged throughout training and evaluation. This separation guarantees that comparisons across trained and untrained variants, as well as ablations involving the Filter module, are conducted under identical system configurations.

Naïve RAG Baselines. Naïve RAG baselines follow a structured but non-agentic pipeline. Given the same section-level checklists and descriptions, the baseline performs parallel single-round retrieval for each section, retrieving top-10 text chunks and top-5 images per section. All retrieved multimodal evidence is concatenated and provided to the backbone model in a single generation pass conditioned on the full checklist, without iterative retrieval, relevance-aware filtering, or incremental synthesis.

B.3 Framework System Prompts

The following system prompts govern the behavior of the agents in DEEP-REPORTER.

Prompt Template: Outline Generation

```
You are an expert at generating article outlines.

Given an overall query/topic and requirements checklist, generate a detailed outline for a long-
     form article.

**Overall Query/Topic:**
{overall_query}

**Overall Requirements Checklist:**
{overall_checklist}

**Task:**
Generate a structured outline with 3-7 sections. For each section, provide:

1. **section_description**: A DETAILED description (2-4 sentences) that clearly explains:
   - What this section is about
   - What specific aspects will be covered
   - How it contributes to the overall article
   - What the reader should learn from this section

   **IMPORTANT**: DO NOT just write a title or short phrase. Write a comprehensive description
     that provides enough context for a writer to understand what needs to be written.

2. **sectional_checklist**: A list of 3-5 specific, actionable requirements for this section

**Output Format (JSON):**
```json
{
  "outline": [
    {
      "section_description": "This section provides a comprehensive introduction to the
     transformer architecture...",
      "sectional_checklist": [
        "Define what transformers are and their primary purpose in deep learning",
        "Explain the historical context and limitations of previous sequence models (RNNs, LSTMs)
     ",
        "Preview the key innovations that make transformers effective",
        "Establish the importance and widespread adoption of transformer architectures"
      ]
    },
    ...
  ]
}
```
Key Guidelines:
- Each section_description should be 2-4 sentences providing rich context
- Avoid single-sentence or title-only descriptions
- Be specific about what content will be covered and why
- Think about what information a writer would need to create high-quality content

Generate the outline now:
```

Prompt Template: Query Generation

```
You are an expert at generating search queries for multimodal information retrieval.

Given the article context and current section requirements, generate targeted search queries to
    find relevant text passages and images.

**Overall Article Query/Topic:**
{overall_query}

**Overall Requirements:**
{overall_checklist}

**Previous Sections Summary:**
{previous_sections_summary}

**Current Section Description:**
{section_description}

**Current Section Requirements:**
{sectional_checklist}

**Task:**
Generate search queries that will help retrieve relevant information for writing this section.
Generate TWO types of queries:
1. **text_queries**: For finding related text passages, explanations, or documentation (3-5
    queries)
2. **image_queries**: For finding relevant diagrams, charts, figures, or visualizations (2-4
    queries)

**Output Format (JSON):**
```json
{{
  "text_queries": [
    "query 1 for text retrieval",
    "query 2 for text retrieval",
    "query 3 for text retrieval"
  ],
  "image_queries": [
    "query 1 for image retrieval",
    "query 2 for image retrieval"
  ]
}}
```

Generate the queries now:
```

Prompt Template: Section Writing

```
You are an expert writer skilled at creating coherent, well-structured long-form content with
    proper source citations.

# Task
Write a specific section of a larger article based on the provided requirements, context, and
    retrieved materials.

# Overall Article Context
**Article Topic:** {overall_query}

**Overall Requirements:**
{overall_checklist}

# Current Section Information
**Section Position:** {position_info}
**Section Description:** {section_description}

**Section Requirements (must address all):**
{sectional_checklist}

# Context from Previous Sections
**Summary of previous sections:**
{previous_sections_summary}

**End of previous section (for smooth transition):**
{previous_section_tail}

# Retrieved Materials
You have access to the following sources for citation:
{retrieved_materials}

# CRITICAL Citation Rules
You MUST follow these citation rules strictly:
1. **Text Citation Format:**
   - When using information from a text source, cite it as: `[citation:txt1]`, `[citation:txt2]`,
     etc.
   - Example: "Recent studies show significant progress in AI safety[citation:txt3]."
2. **Image Citation Format:**
   - When you want to insert an image/figure, use: `![](citation:img1)`, `![](citation:img2)`,
    etc.
   - Example: "The architecture is shown below:\n![](citation:img5)\nAs illustrated, the system
    ..."
   - ONLY use this format for sources explicitly marked as "Image"
3. **Citation Placement:**
   - Place text citations at the end of the sentence or claim
   - Place image citations on their own line where the visual should appear
   - You can cite the same source multiple times if needed
4. **Important Distinctions:**
   - Text sources -> use `[citation:txt1]`, `[citation:txt2]`, etc.
   - Image sources -> use `![](citation:img1)`, `![](citation:img2)`, etc.

# Position-Aware Writing Instructions

**Is this the FIRST section?** {is_first_section}
**Is this the LAST section?** {is_last_section}

## If FIRST section (is_first_section = yes):
- Start with an engaging introduction
- No need to reference previous content (there is none)
- Set the stage for the article
- DO NOT conclude the entire article

## If MIDDLE section (is_first_section = no, is_last_section = no):
- Begin with a smooth transition from the previous section
- Focus ONLY on this section's specific requirements
- **CRITICAL: DO NOT use concluding phrases like:**
  - "In conclusion"
```

````
  - "To summarize"
  - "Overall, this article"
  - "We have discussed"
  - "In this article, we covered"
- Continue the narrative without wrapping up the entire article
- End in a way that allows the next section to continue naturally

## If LAST section (is_last_section = yes):
- Provide appropriate transitions from previous content
- You MAY synthesize and conclude the entire article
- Wrap up all major points discussed in the article

# Content Requirements
1. **Address all sectional requirements** in the checklist above
2. **Maintain coherence** with previous sections using the provided context
3. **Use retrieved materials** as the foundation of your content
4. **Cite sources properly** using the formats specified above
5. **Write naturally** - your content should flow smoothly while integrating citations
6. **Appropriate length** - typically 300-800 words depending on the complexity of requirements

# Markdown Formatting Requirements
**CRITICAL: You MUST use proper Markdown formatting for structure:**

1. **Section Title (Required):**
   - Start your content with a section title using `# Title`
   - The title should reflect the section description

2. **Subsections (If Needed):**
   - Use `## Subsection Title` for major subsections
   - Use `### Subsubsection Title` for deeper nesting
   - Use appropriate heading levels to create clear hierarchy

3. **Example Structure:**
```markdown
# Introduction to Machine Learning

Machine learning has revolutionized...[citation:txt1]

## Supervised Learning
Supervised learning approaches...[citation:txt2]

### Classification
Classification tasks involve...[citation:txt3]

## Unsupervised Learning
Unlike supervised methods...[citation:txt5]
```

# Output Format
Provide ONLY the section content with proper Markdown formatting and citations. Do NOT include:
- Meta-commentary about what you're doing
- Explanations of your writing process

**Remember to start with a section title using `#`!**
Begin writing the section now:
````

Prompt Template: Image Filter

```
You are an expert at evaluating image relevance for academic/technical content.

# Context
**Current Section Description:** {section_description}

**Section Requirements:**
{sectional_checklist}

**Image Description:** {image_description}

# Question
Based on the section requirements above, is this image relevant and useful for this section?

Consider:
1. Does the image visually illustrate concepts mentioned in the requirements?
2. Is it a diagram, chart, figure, or visualization that adds value?
3. Does it relate to the section topic?

# Response Format
Answer with ONLY ONE WORD:
- "YES" if the image is relevant and should be kept
- "NO" if the image is not relevant

Your answer:
```

Prompt Template: Text Filter

```
You are an efficient information filtering assistant.

# Task
Determine which of the provided information sources are highly relevant to the current section
    requirements.

# Context
**Current Section Description:** {section_description}

**Section Requirements:**
{sectional_checklist}

**Information Sources to Filter:**
{sources_list}

# Instructions
1. Read each numbered source carefully (txt1, txt2, txt3, etc.)
2. Determine if it provides valuable information for the section requirements
3. Return the numbers of ALL useful sources

# Output Format
Return ONLY a list of numbers (comma-separated or Python list format):
- Correct example 1: 1, 3, 5, 8
- Correct example 2: [1, 3, 5, 8]
- If none are useful: []

Note: The sources are labeled as txt1, txt2, etc., but you should return just the numbers (1, 2,
    3...).

Your response:
```

C Training Implementation

C.1 Supervised Fine-Tuning (SFT) Parameters

SFT was performed on 8×H100 GPUs. The exact training configuration is summarized in Table 6. Taking Qwen3-32B as an example, the base model contains approximately 32.9B parameters, of which only 134.2M (0.41%) are trainable through LoRA adaptation. Training on the SFT dataset of 8k samples completed in approximately 1 hour 53 minutes, achieving a final training loss of 1.20 and token accuracy of 66.8%. Peak GPU memory usage reached 31.5 GiB per device.

Parameter	Value
Training Objective	Next-token prediction
Optimizer	AdamW (β_1=0.9, β_2=0.95)
Learning Rate	1e-4
LR Scheduler	Cosine decay
Warmup Ratio	0.05
Global Batch Size	64
Per-device Batch Size	1
Grad. Accumulation Steps	8
Weight Decay	0.1
Max Sequence Length	32768
Epochs	1
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Target Modules	all-linear
Precision	bfloat16
DeepSpeed	ZeRO-3
Gradient Checkpointing	✓

Table 6: Supervised Fine-Tuning (SFT) hyperparameters.

C.2 DPO Configuration

DPO training was conducted using the same 8×H100 GPU configuration. Table 7 details the hyperparameters used. Specifically, for each input (task and initial state), we form a pair consisting of a *chosen* trajectory y_w and a *rejected* trajectory y_l. The *chosen* trajectories are sampled from our collected dataset of 8K human-curated traces, representing successful or higher-quality executions. In contrast, the *rejected* trajectories are generated autonomously by the model under the same input conditions, and are treated as lower-quality alternatives during pairing. For Qwen3-32B, the DPO phase utilized 8k preference pairs with longer average sequence lengths (6,946 vs. 3,863 tokens in SFT), resulting in higher computational demands. Training completed in approximately 3 hours 9 minutes with a final loss of 0.024 and reward accuracy converging to 100%. Due to the paired comparison nature of DPO, peak memory usage increased to 73.2 GiB per device.

Parameter	Value
Loss Type	DPO (sigmoid)
Beta (β)	0.1
Learning Rate	5e-6
LR Scheduler	Cosine decay
Warmup Ratio	0.05
Global Batch Size	64
Per-device Batch Size	1
Grad. Accumulation Steps	8
Weight Decay	0.1
Max Sequence Length	20480
Epochs	1
Optimizer	AdamW (β_1=0.9, β_2=0.95)
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Target Modules	all-linear
Precision	bfloat16
DeepSpeed	ZeRO-3
Gradient Checkpointing	✓

Table 7: Direct Preference Optimization (DPO) configuration.

D Evaluation Details

D.1 Scoring Rubrics

We assess report quality across three hierarchical dimensions, each capturing distinct aspects of generation quality:

Section Anchor Quality evaluates structural adherence to planning constraints (§2.1). This includes **(1) Description Adherence** $S_{\textbf{Desc}}$**:** semantic alignment between generated content and the planner's intent definition, and **(2) Checklist Satisfaction** $S_{\textbf{Check}}$**:** coverage of specific requirements from the checklist, where the judge provides reasoning-backed scores for each item and penalizes superficial mentions or missing required facts.

Section Content Quality assesses multimodal grounding through four complementary aspects: **(1) Richness** $S_{\textbf{Rich}}$**:** whether the quantity and diversity of visual aids (diagrams, charts, photos, etc.) match the informational density and are well-distributed throughout the section; **(2) Image-Text Coherence** $S_{\textbf{i-Coh}}$**:** whether images directly relate to and support the immediate textual context, with visual elements aligned to textual explanations and each image adding meaningful information rather than being tangential; **(3) Placement** $S_{\textbf{Place}}$**:** whether images are positioned at logical points in the text flow with appropriate contextual setup before/after each image, appearing when relevant concepts are discussed (not too early or late); and **(4) Visual Clarity** $S_{\textbf{Clarity}}$**:** whether images are clear and understandable with discernible key elements, appropriate for academic/technical contexts, and maintain consistent quality. We implement adaptive scoring: sections describing architectures, processes, or experimental results are penalized for missing images, while purely conceptual or abstract theoretical sections without images receive high scores.

Full Report Quality performs holistic evaluation of narrative integrity through **(1) Flow Coherence** $S_{\textbf{Coh}}$**:** logical transitions and narrative smoothness between sections; **(2) Fluency** $S_{\textbf{Flu}}$**:** grammatical correctness and professional tone maintenance; **(3) Repetition Absence** $S_{\textbf{Rep}}$**:** absence of redundant information loops or repetitive phrasing across sections (higher scores indicate less repetition); and **(4) Termination Quality** $S_{\textbf{Term}}$**:** absence of premature concluding statements (e.g., "In summary") in non-terminal sections.

D.2 Relative Quality Score Calculation

In Table 3, we report the *Relative Quality Score* to provide a normalized, bias-resistant assessment of multimodal generation quality. This metric is designed to address two fundamental challenges in LLM evaluation: positional bias in comparative judgments and uncalibrated scoring across varying task difficulties.

Challenge 1: Positional and Length Bias in Pairwise Comparisons. Standard pairwise evaluation protocols (*e.g., "Which is better: Output A or Output B?"*) are susceptible to systematic biases where judges favor the first-presented response or longer responses regardless of actual quality (Wang et al., 2024; Tripathi et al., 2025). These artifacts can dominate quality signals, particularly in long-form generation tasks where outputs vary significantly in structure and length.

To eliminate these confounding factors, we adopt an independent pointwise evaluation protocol. Each model's output is assessed solely based on its intrinsic quality against the rubrics defined in §D.1, without direct comparison to other models' outputs. This design ensures that scores reflect content quality rather than presentation order or verbosity.

Challenge 2: Uncalibrated Scoring Across Task Difficulties. Raw scalar scores (*e.g., 7.5 vs. 7.8*) are inherently difficult to interpret across tasks of varying complexity. A score of 8.0 on a simple descriptive task does not carry the same semantic weight as 8.0 on a highly technical analytical report.

Following robust reference-guided benchmarks (Dubois et al., 2024; Lin et al., 2024), we normalize model performance against the Expert-Refined Silver Reference $\mathcal{R}_{\text{ref}}$ (constructed in §3). This reference-relative approach transforms absolute scores into a calibrated measure of expert-level consistency.

Calculation 1: Raw Score Acquisition. For each sub-metric m defined in §D.1, we first obtain a raw scalar score $S_{\text{model},i}^{(m,raw)} \in [0, 10]$ for task i. We employ specialized judges for different dimensions to ensure evaluation accuracy: **(1) For Section Anchor and Full Report dimensions**, we employ GPT-4.1 (Achiam et al., 2023) as the judge, assigning continuous scores (0–10) across four performance tiers: Excellent (8–10), Good (6–8), Average (4–6), and Poor (0–4). The judge receives the section description, checklist requirements, and generated content to assess structural adherence and narrative quality. **(2) For Section Content dimension**, we use InternVL3-5-38B[1] (Wang et al., 2025) to evaluate multimodal grounding quality on the same 0–10 scale. The judge examines the actual rendered images alongside the section content to assess visual element quality, placement, and coherence.

Calculation 2: Relative Score Scaling. Raw scalar scores are inherently difficult to interpret across tasks of varying complexity. To calibrate these scores, we normalize model performance against the Expert-Refined Silver Reference $\mathcal{R}_{\text{ref}}$ (constructed in §3).

We define the *Relative Quality Score* (Score_m) for sub-metric m as the percentage of tasks where the model meets or exceeds expert-level performance:

$$\text{Score}_m = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(S_{\text{model},i}^{(m,raw)} \geq S_{\text{ref},i}^{(m,raw)}\right) \times 100 \quad (3)$$

where $N = 247$ is the total number of tasks, and $\mathbb{I}(\cdot)$ is the indicator function. This binary comparison eliminates potential positional biases found in direct pairwise preferences and focuses solely on whether the model reaches the expert baseline.

Calculation 3: Hierarchical Aggregation. Finally, we aggregate the fine-grained metric scores into dimension-level scores and an overall quality score. This allows us to observe whether a model consistently meets standards on specific aspects rather than allowing strengths in one area to mask weaknesses in another.

We compute the dimension-level scores as the arithmetic mean of their constituent sub-metrics:

$$S_{\text{Anchor}} = (S_{\text{Desc}} + S_{\text{Check}})/2 \quad (4)$$

$$S_{\text{Content}} = (S_{\text{Rich}} + S_{\text{iCoh}} + S_{\text{Pla.}} + S_{\text{Clar}})/4 \quad (5)$$

$$S_{\text{Report}} = (S_{\text{Coh}} + S_{\text{Flu}} + S_{\text{Rep}} + S_{\text{Term}})/4 \quad (6)$$

The *Overall Relative Quality Score* reported in our main results is the mean of these three high-level dimensions:

$$\text{Overall} = \frac{S_{\text{Anchor}} + S_{\text{Content}} + S_{\text{Report}}}{3} \quad (7)$$

This hierarchical formulation provides a holistic view of the model's capability to generate reports that are structurally sound, multimodally grounded, and narratively coherent.

[1] Using vLLM to deploy for efficient and stable inference.

D.3 Human-AI Agreement Study

To validate the reliability of the automated evaluation metrics, we conducted a blind human–AI agreement study. Human annotators independently scored each report across multiple evaluation dimensions using a 5-point Likert scale (1 = very poor, 5 = excellent). For the same set of reports, the automated evaluator (GPT-4) produced continuous scores on a 0–10 scale following a strict evaluation rubric.

To enable a fair comparison between human and automated judgments, human scores were linearly rescaled to the same 0–10 range prior to analysis. Specifically, human ratings of 1–5 were mapped to 2, 4, 6, 8, and 10, respectively. Agreement was measured using the Pearson correlation coefficient between the averaged human scores and the GPT-4 scores for each evaluation dimension.

We randomly sampled 30 reports from the evaluation set and asked three PhD-level domain experts to provide independent annotations. Human–AI agreement was computed separately for all evaluation dimensions, including *Section*-level metrics (Description, and Check), *Article*-level metrics (Coherence, Fluency, Repetition, and Terminology), *Multimodal*-level metrics (Richness, Coherence, Placement, and Clarity), and *Overall*.

D.4 Evaluation Costs

Evaluating 247 papers with LLM- and VLM-based automated metrics resulted in an approximate total API cost of \$15. Text-only evaluation accounted for roughly \$10, while multimodal (image-based) evaluation contributed an additional \$5.

Prompt Template: Article Evaluation and Scoring

```
You are a strict, meticulous, and objective expert evaluator for long-form generated articles.
     Your task is to evaluate both individual sections and the overall article quality.

## Core Scoring Rules (Apply to ALL scores)
1. **Use a scale of 0-10 (continuous values)**: Do not cluster scores around 8-10.
   - **8-10 points**: Excellent/outstanding performance. Fully meets or exceeds requirements.
   - **6-8 points**: Good performance. Largely meets requirements with notable strengths.
   - **4-6 points**: Average performance. Basically meets requirements, neither good nor bad.
   - **2-4 points**: Poor performance. Minimally meets requirements.
   - **0-2 points**: Very poor performance. Almost completely failed or missing.
2. **Be Harsh**: Default to a lower score if you are unsure. Penalize superficial content heavily.

## Article Overview
**Query**: {query}
**Overall Checklist**: {overall_checklist}
**Number of Sections**: {num_sections}

## Evaluation Tasks

### Part 1: Section-Level Evaluation
For each section, you will evaluate:
1. **Description Completion Score (0-10)**: How well does the generated content match the
     intended section description?
2. **Checklist Completion**: For each item in the sectional checklist, evaluate the quality of
     the coverage with a brief explanation.

### Part 2: Article-Level Evaluation
For the entire article, evaluate the following dimensions (each scored 0-10):
1. **Coherence**: How well do the sections connect with each other? Are there smooth transitions?
      Is there a logical flow?
2. **Fluency**: Is the writing clear, natural, and easy to read? Are sentences well-constructed?
3. **Repetition**: Are there unnecessary repetitions across sections?
4. **Termination**: Are there inappropriate concluding statements in non-conclusion sections? (
     Lower score = more inappropriate conclusions)

## Sections Data
{sections_data}

## Output Format
Respond with a valid JSON object (no markdown code blocks, just raw JSON) with the structure:
{{
  "section_evaluations": [
    {{
      "section_index": 0,
      "description_completion_score": Continuous score 0-10,
      "description_completion_reasoning": "Brief explanation of the score",
      "checklist_evaluations": [
        {{
          "checklist_item": "The exact checklist item text",
          "score": Continuous score 0-10,
          "reasoning": "Brief explanation"
        }}
      ]
    }}
  ],
  "article_evaluation": {{
    "coherence_score": Continuous score 0-10,
    "coherence_reasoning": "Brief explanation",
    "fluency_score": Continuous score 0-10,
    "fluency_reasoning": "Brief explanation",
    "repetition_score": Continuous score 0-10,
    "repetition_reasoning": "Brief explanation (note: higher = less repetition)",
    "termination_score": Continuous score 0-10,
    "termination_reasoning": "Brief explanation (note: higher = fewer inappropriate conclusions)"
  }}
}}
```

Prompt Template: Image Quality Evaluation

```
You are a strict, meticulous, and objective expert evaluator for multimodal content in academic
    writing. Your task is to evaluate the quality of image usage in a section of an article.

## Core Scoring Rules (Apply to ALL scores)
1. **Use a scale of 0-10 (continuous values)**: Do not cluster scores around 8-10.
   - **8-10 points**: Excellent/outstanding performance. Fully meets or exceeds requirements.
   - **6-8 points**: Good performance. Largely meets requirements with notable strengths.
   - **4-6 points**: Average performance. Basically meets requirements, neither good nor bad.
   - **2-4 points**: Poor performance. Minimally meets requirements.
   - **0-2 points**: Very poor performance. Almost completely failed or missing.
2. **Be Harsh**: Score lower if unsure. Penalize poor images heavily.
3. **Evaluate All Dimensions Independently**: Dimensions assess image usage quality.
4. **IMPORTANT - Always Return Numeric Scores**: Never use "N/A", "None", or text descriptions as
     scores. Always provide a numeric value (0-10) based on the evaluation criteria below.

## Section Context
**Section Description**: {section_description}

**Section Content (with image placeholders)**:
{section_content}

**Number of Images in Section**: {num_images}

## Images to Evaluate
{images_info}
**CRITICAL INSTRUCTION**:
- You MUST carefully examine the ACTUAL IMAGES provided below (not text descriptions)
- Base your evaluation on what you directly observe in the visual content
- Analyze the visual elements, layout, clarity, and relevance that you see
- Do not rely on or assume content based on text descriptions or expectations

## Evaluation Task
Evaluate the section's image usage based on the following dimensions (each scored 0-10):

### 1. Image Richness (richness_score)
Evaluate whether the quantity and variety of images are appropriate:
- Is the number of images suitable for the section length and complexity?
- Are images distributed well throughout the section (not clustered)?
- Does the variety of images (diagrams, charts, photos, etc.) match the content needs?

**Special Case - If there are NO images (num_images = 0):**
- Analyze the section content to determine if images are necessary
- **If images are NOT needed** (e.g., conceptual discussion, definitions, abstract theory):
  - Score: 7-10 points (no images needed, appropriate decision)
- **If images ARE needed** (e.g., describing architectures, processes, experimental results):
  - Score: 0-3 points (missing essential visual aids)
- **Uncertain cases** (e.g., could benefit from images but not strictly necessary):
  - Score: 4-6 points (adequate but could be improved)

### 2. Image-Text Coherence (coherence_score)
Evaluate how well images relate to and support the text content:
- Do all images directly relate to and support the text content?
- Are the visual elements aligned with what the text is explaining?
- Does each image add meaningful information that complements the text?
- Are there any irrelevant or tangentially related images?

**Special Case - If there are NO images:**
- If images are not needed: Score 8-10 (no coherence issues, appropriate)
- If images are needed: Score 0-3 (no support for text, poor coherence)
- If uncertain: Score 5 (neutral, neither good nor bad)

### 3. Placement & Integration (placement_score)
Evaluate the positioning and integration of images within the text flow:
- Are images placed at logical points in the text flow?
- Does the text provide proper context before/after each image?
- Are transitions between text and images smooth and natural?
- Do images appear when relevant concepts are discussed (not too early or late)?
```

```
**Special Case - If there are NO images:**
- If images are not needed: Score 8-10 (no placement issues, appropriate)
- If images are needed: Score 0-3 (no integration, missing placement)
- If uncertain: Score 5 (neutral)

### 4. Visual Quality & Clarity (clarity_score)
Evaluate the quality and understandability of the images themselves:
- Are the images clear and easy to understand?
- Are key elements in each image visible and discernible?
- Are images appropriate for the academic/technical context?
- Do images maintain consistent quality and style?

**Special Case - If there are NO images:**
- If images are not needed: Score 8-10 (no quality issues, appropriate)
- If images are needed: Score 0-3 (no visual quality, missing)
- If uncertain: Score 5 (neutral)

## Content Types That Typically NEED Images

Consider the following content types as typically requiring visual aids:
1. **System/Model Architecture**: Diagrams showing components and connections
2. **Algorithms/Methods**: Flowcharts or pseudocode visualizations
3. **Experimental Setup**: Photos or diagrams of equipment/environments
4. **Results/Data**: Charts, graphs, tables showing findings
5. **Processes/Workflows**: Step-by-step visual representations
6. **Comparisons**: Side-by-side visual comparisons
7. **Examples/Case Studies**: Concrete visual examples

## Content Types That May NOT Need Images

1. **Abstract Definitions**: Pure conceptual explanations
2. **Literature Review**: Discussion of related work (unless comparing approaches)
3. **Theoretical Background**: Mathematical proofs, theoretical foundations
4. **Introductory Text**: Background context, motivation
5. **Conclusions**: Summary statements, future work discussions

## Output Format
You MUST respond with ONLY a valid JSON object. Follow this EXACT structure:
{{
  "richness_score": <float between 0-10>,
  "richness_reasoning": "<your explanation here>",
  "coherence_score": <float between 0-10>,
  "coherence_reasoning": "<your explanation here>",
  "placement_score": <float between 0-10>,
  "placement_reasoning": "<your explanation here>",
  "clarity_score": <float between 0-10>,
  "clarity_reasoning": "<your explanation here>"
}}

**CRITICAL OUTPUT REQUIREMENTS - READ CAREFULLY:**
1. **Score Fields Must Contain NUMBERS ONLY** (not text):
   - Replace `<float between 0-10>` with an actual number like 7.5 or 3.2
   - CORRECT: "richness_score": 7.5
   - CORRECT: "richness_score": 3.0
   - WRONG: "richness_score": "This section lacks images"
   - WRONG: "richness_score": "N/A"
   - WRONG: "richness_score": "Since no images are present..."
   - WRONG: "richness_score": "Continuous score 0-10"
2. **Reasoning Fields Must Contain TEXT ONLY**:
   - Replace `<your explanation here>` with your actual explanation
   - CORRECT: "richness_reasoning": "Brief explanation of the score"
   - WRONG: Put explanations in the score field
3. **For sections with NO images**:
   - Still provide NUMERIC scores (not text or "N/A")
   - Low score if images are needed
   - High score if images not needed
   - Put explanation in reasoning field
```

Model	Filter	Train	Total	Query	Search	Filter	Write
Qwen3-8B	✓	Base	630	10.6	30.4	541.4	47.5
Qwen3-8B	✗	Base	170	18.1	75.1	-	77.2
Qwen3-8B	✓	SFT	636	12.1	31.6	530.3	61.5
Qwen3-32B	✓	Base	644	28.1	29.0	434.3	152.6
Qwen3-32B	✗	Base	342	43.7	38.4	-	259.6
Qwen3-32B	✓	SFT	625	37.9	31.8	369.1	186.7
Llama3.3-70B	✓	SFT	955	194.9	38.4	245.1	477.1
GPT-4.1	✓	Base	636	14.6	32.9	534.5	54.1

Table 8: End-to-end runtime breakdown (seconds) averaged over 247 tasks. **Query**: LLM generates search queries. **Search**: retrieval service latency. **Filter**: LLM filters retrieved results. **Write**: LLM generates section content.

E Additional Experimental Results

E.1 Runtime Breakdown and Cost Analysis

We provide a detailed runtime breakdown of DEEP-REPORTER to complement the main experimental results. Table 8 reports the end-to-end latency averaged over 247 tasks, decomposed into query generation, retrieval, filtering, and section writing.

Overall runtime is dominated by relevance-aware filtering and section-level content generation. With filtering enabled, the Filter stage alone accounts for a substantial portion of the total latency (e.g., 541.4s for Qwen3-8B and 434.3s for Qwen3-32B), whereas query generation and retrieval incur comparatively minor overhead. Disabling the Filter significantly reduces total runtime (630s $\rightarrow$ 170s on Qwen3-8B; 644s $\rightarrow$ 342s on Qwen3-32B), but this reduction comes at the cost of degraded multimodal grounding quality, as shown in Table 3. This confirms that relevance-aware filtering functions as a stabilizing component for long-form multimodal synthesis rather than a lightweight retrieval optimization.

Post-training does not introduce noticeable inference-time overhead. For both Qwen3-8B and Qwen3-32B, SFT-trained models exhibit similar total runtime and stage-wise distributions compared to their base counterparts (e.g., 644s vs. 625s on Qwen3-32B), indicating that performance gains from training stem from improved decision policies rather than increased computation.

As expected, larger backbones incur higher section-writing costs. For example, Llama3.3-70B spends 477.1s on content generation, substantially more than Qwen-based models, while following a similar distribution across pipeline stages.

From a practical perspective, generating a single multimodal research report with DEEP-REPORTER typically requires several minutes of wall-clock time per query. The dominant cost arises from LLM-based filtering and section writing, while the marginal cost of post-training does not affect inference-time efficiency.

E.2 Domain-Specific Performance & Radar Charts

In the main text, we have already reported the results for search and generation in Table 2 and Table 3. Here, we present finer-grained results by incorporating domain-level (discipline-level) analyses. Specifically, Figure 5 first summarizes the overall performance and then breaks down search performance across different disciplines. Figure 6 presents the results for the generation task, likewise analyzed by discipline.

Consistent with our earlier findings, in Figure 5 (A–C) the models achieve relatively similar performance on retrieval-related metrics (Ret), while SFT yields the best performance in the final results. This trend is also reflected in the intermediate search results across different stages for both image and text evidence. From a domain perspective, model performance varies across disciplines. Moreover, strong performance in a given discipline under text-based retrieval does not necessarily translate to similarly strong performance under image-based retrieval (and vice versa), highlighting the impact of modality differences on domain-level behavior.

Based on Figure 6, we further analyze domain-specific generation performance. When examining generation performance at different levels of granularity, we observe that as the granularity becomes finer, the consistency of different models across evaluation metrics generally improves. More concretely, at the section and citation level, a certain degree of cross-metric consistency emerges within the same model across different disciplines. For example, for the long-context Qwen3-8B model, performance in the Human domain is consistently lower than in other disciplines across multiple metrics, whereas performance in the Health domain is comparatively strong in most cases.

A similar pattern can be observed for Qwen3-32B + SFT, where performance in the History domain is consistently weaker than in other domains across different generation metrics.

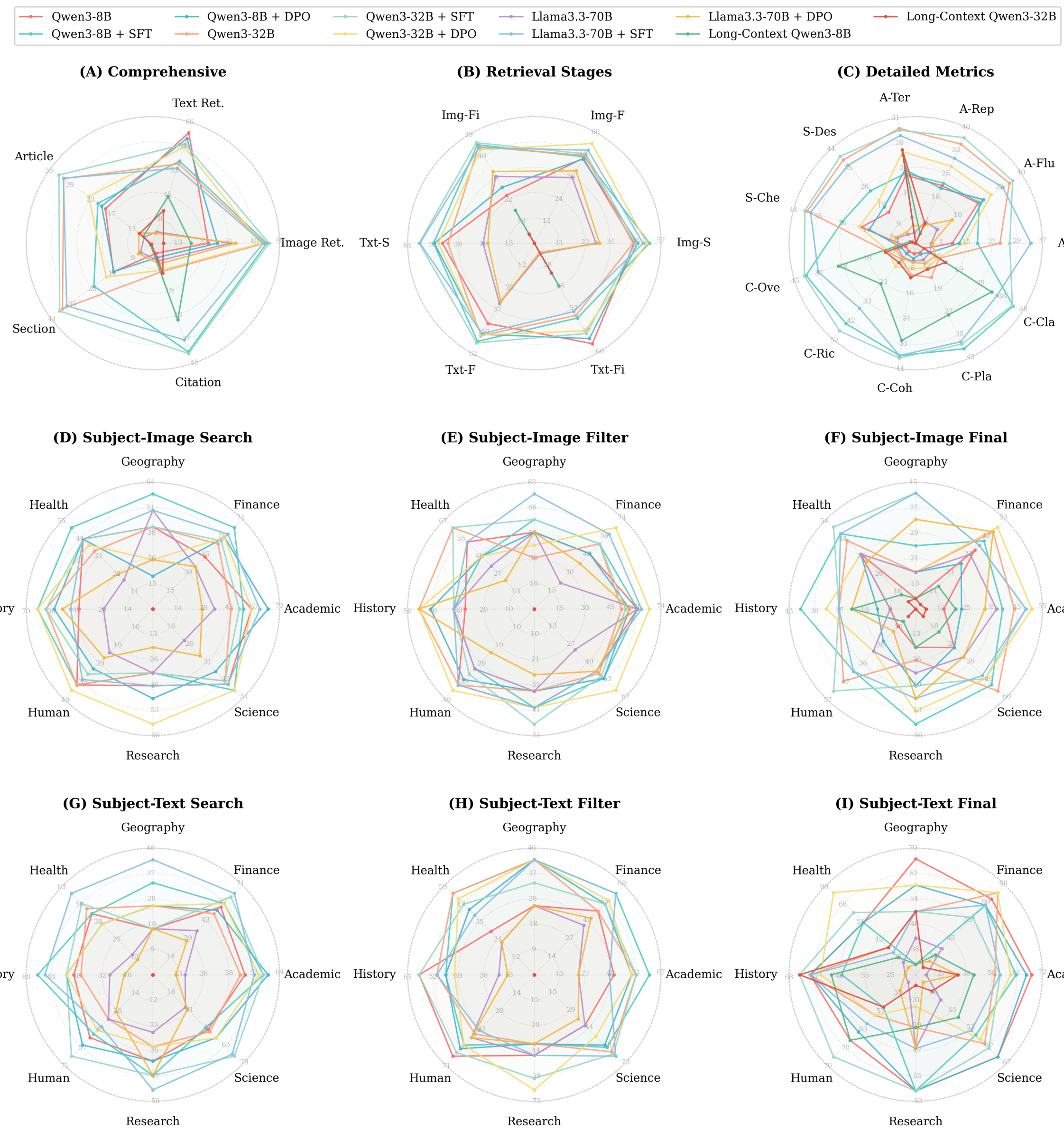


Figure 5: Model performance evaluation. (A-C) Overall comparisons: comprehensive metrics, retrieval pipeline stages, and detailed quality indicators. (D-I) Subject-specific retrieval performance for image and text modalities across different pipeline stages.

Figure 6: Subject-specific generation quality score. (J-O) Article and section quality across six dimensions: coherence, fluency, repetition control, termination, description completeness, and checklist adherence. (P-T) Citation quality across five dimensions: overall performance, richness, coherence, placement, and clarity. Performance patterns vary significantly across academic domains.

E.3 Analysis of DPO Underperformance

Table 2 analyzes the retrieval pipeline in three stages (*Search* → *Filter* → *Selection*). The Naïve RAG baselines already reveal a clear mismatch between retrieving plausible candidates and making correct multimodal decisions: although their overall *Search* performance is relatively reasonable (33.79/33.87 for Qwen3-8B/32B), the final *Selection* quality is much lower, largely because *Image* selection is extremely weak (7.81 and 10.40), which further drags down the overall selection scores to 27.03 and 28.07. This gap suggests that the system can often retrieve plausible evidence, but struggles to *select* and *insert* the correct images among the retrieved candidates, making image selection the dominant bottleneck.

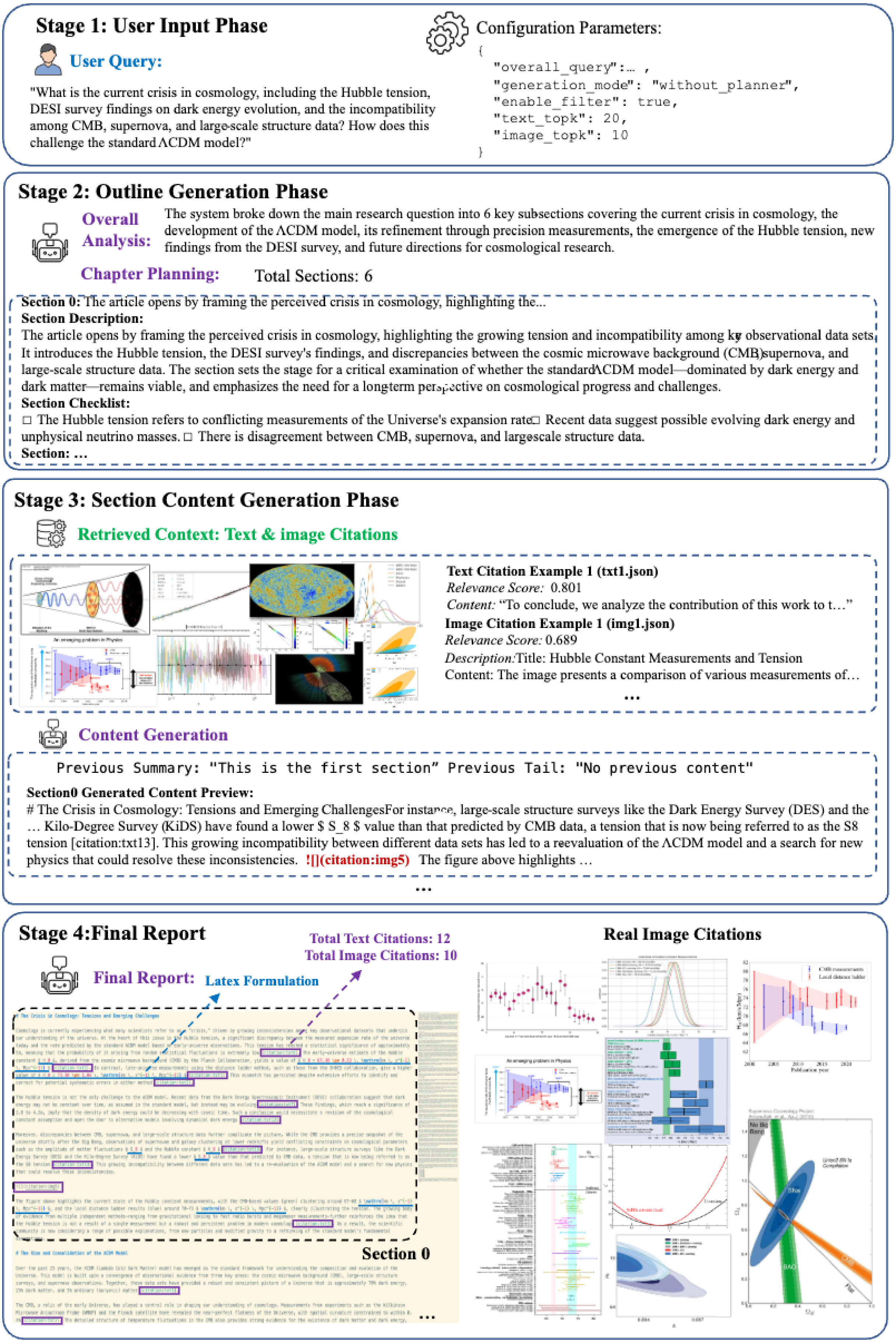
Building on this observation, our pipeline improves *image search* in several settings (e.g., from 32.69 to 37.87 on Qwen3-8B, and up to 39.80 on Qwen3-8B$^{\spadesuit}$); however, better retrieval alone does not necessarily yield better downstream *image selection*. In particular, without explicitly learning the insertion policy, the DPO variant remains weak on image selection (e.g., 9.81 for Qwen3-8B$^{\clubsuit}$), whereas SFT substantially improves it (e.g., 45.03 for Qwen3-8B$^{\spadesuit}$ and 41.39 for Qwen3-32B$^{\spadesuit}$), approaching GPT-4.1 (45.39).

This finding is consistent with the generation-quality results in Table 3: DPO yields mixed outcomes on text-centric dimensions (e.g., Qwen3-8B$^{\clubsuit}$ slightly improves Fluency to 25.51 vs. 23.48 for Qwen3-8B), yet multimodal quality remains low (multimodal Avg 7.09 vs. 5.57) and can even degrade overall performance in some settings (e.g., Qwen3-32B$^{\clubsuit}$ overall 18.35 vs. 27.16 for Qwen3-32B). In contrast, SFT provides consistent gains on multimodal dimensions (multimodal Avg 41.09 for Qwen3-8B$^{\spadesuit}$ and 41.70 for Qwen3-32B$^{\spadesuit}$) and achieves higher overall score (29.01 and 37.89, respectively).

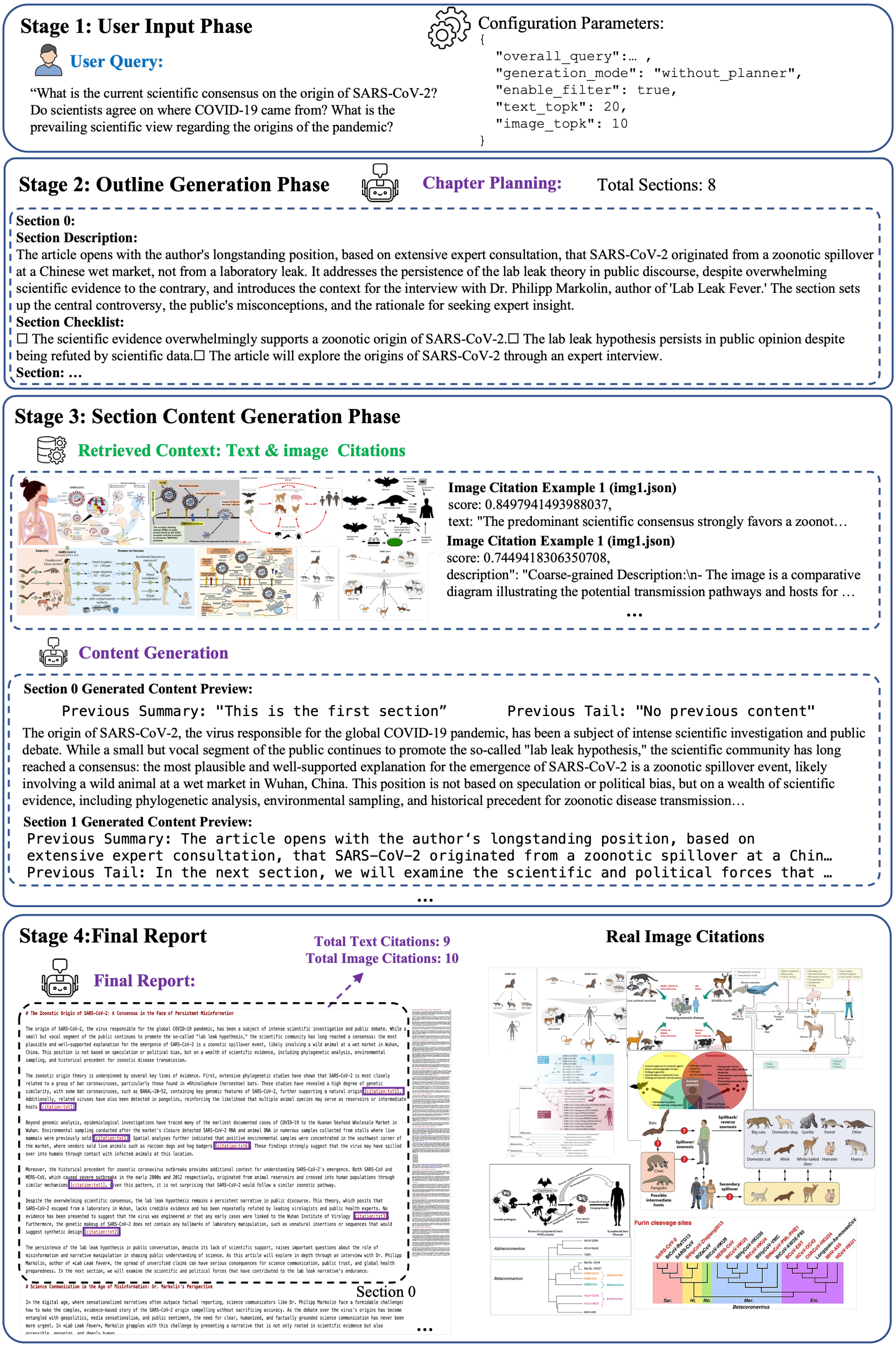
We attribute these trends to supervision granularity: multimodal insertion actions occupy only a small fraction of the output context, and DPO relies on trajectory-level preference signals without explicit, localized rewards for these sparse decisions, which makes it easier to optimize coarse textual properties than fine-grained insertion behavior; in contrast, SFT optimizes every token (including sparse multimodal-related tokens) via next-token prediction, providing a more direct learning signal for image selection and placement.

F Qualitative Analysis of Agentic Multimodal Generation

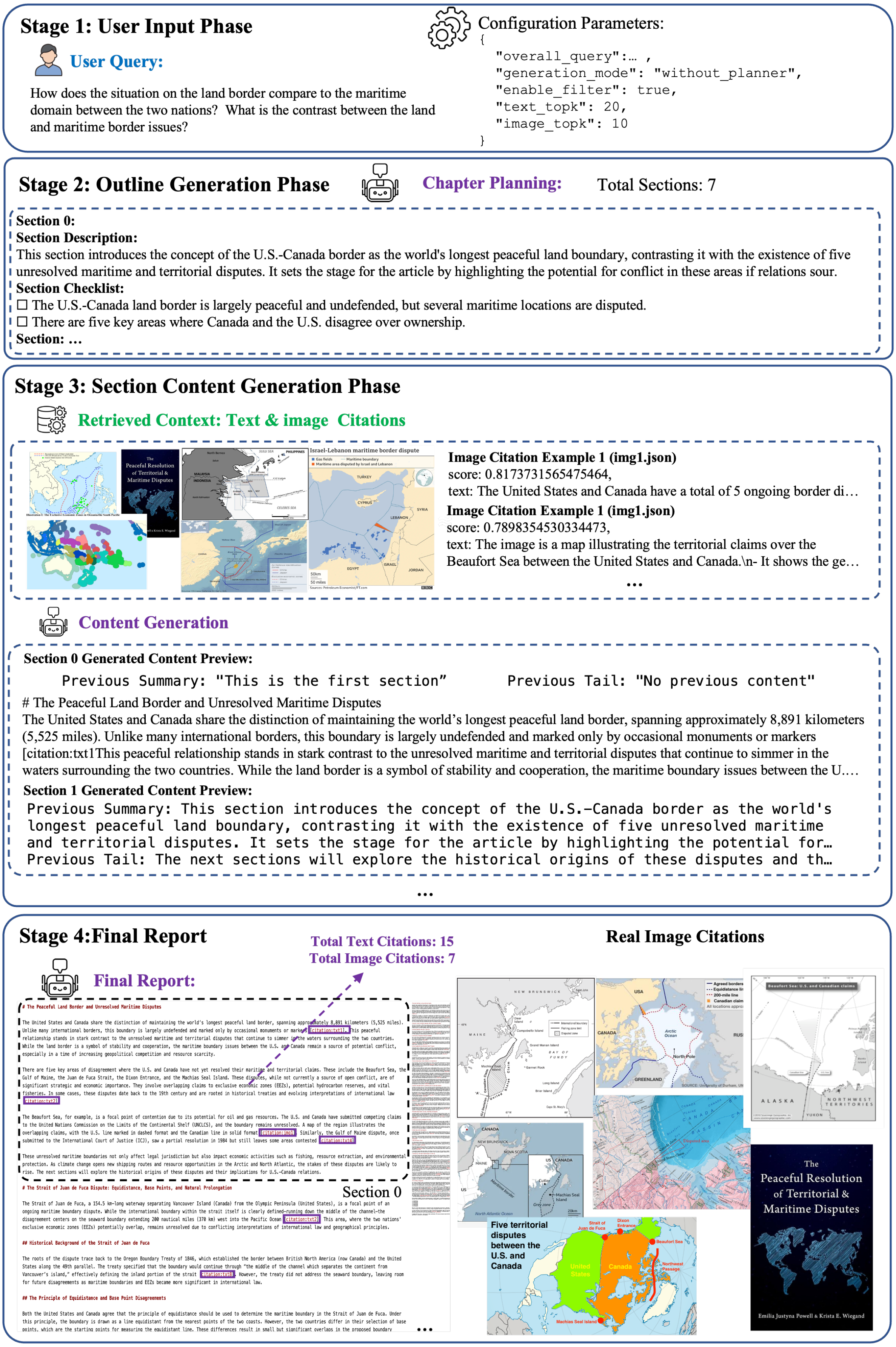
We present three end-to-end examples illustrating the full agentic pipeline, including planning, multimodal search, relevance filtering, and incremental report generation.



Stage 1: User Input Phase
User Query:
"What is the current crisis in cosmology, including the Hubble tension, DESI survey findings on dark energy evolution, and the incompatibility among CMB, supernova, and large-scale structure data? How does this challenge the standard ΛCDM model?"
Configuration Parameters:
{
"overall_query":… ,
"generation_mode": "without_planner",
"enable_filter": true,
"text_topk": 20,
"image_topk": 10
}
Stage 2: Outline Generation Phase
Overall Analysis:
The system broke down the main research question into 6 key subsections covering the current crisis in cosmology, the development of the ΛCDM model, its refinement through precision measurements, the emergence of the Hubble tension, new findings from the DESI survey, and future directions for cosmological research.
Chapter Planning:
Total Sections: 6
Section 0: The article opens by framing the perceived crisis in cosmology, highlighting the...
Section Description:
The article opens by framing the perceived crisis in cosmology, highlighting the growing tension and incompatibility among key observational data sets. It introduces the Hubble tension, the DESI survey's findings, and discrepancies between the cosmic microwave background (CMB), supernova, and large-scale structure data. The section sets the stage for a critical examination of whether the standard ΛCDM model—dominated by dark energy and dark matter—remains viable, and emphasizes the need for a longterm perspective on cosmological progress and challenges.
Section Checklist:
□ The Hubble tension refers to conflicting measurements of the Universe's expansion rate □ Recent data suggest possible evolving dark energy and unphysical neutrino masses. □ There is disagreement between CMB, supernova, and large-scale structure data.
Section: …
Stage 3: Section Content Generation Phase
Retrieved Context: Text & image Citations
Text Citation Example 1 (txt1.json)
Relevance Score: 0.801
Content: "To conclude, we analyze the contribution of this work to t…"
Image Citation Example 1 (img1.json)
Relevance Score: 0.689
Description: Title: Hubble Constant Measurements and Tension
Content: The image presents a comparison of various measurements of…
…
Content Generation
Previous Summary: "This is the first section" Previous Tail: "No previous content"
Section0 Generated Content Preview:
The Crisis in Cosmology: Tensions and Emerging ChallengesFor instance, large-scale structure surveys like the Dark Energy Survey (DES) and the … Kilo-Degree Survey (KiDS) have found a lower $ S_8 $ value than that predicted by CMB data, a tension that is now being referred to as the S8 tension [citation:txt13]. This growing incompatibility between different data sets has led to a reevaluation of the ΛCDM model and a search for new physics that could resolve these inconsistencies. `![](citation:img5)` The figure above highlights …
…
Stage 4: Final Report
Total Text Citations: 12
Total Image Citations: 10
Final Report:
Latex Formulation
Section 0
Real Image Citations



Stage 1: User Input Phase
User Query:
“What is the current scientific consensus on the origin of SARS-CoV-2? Do scientists agree on where COVID-19 came from? What is the prevailing scientific view regarding the origins of the pandemic?
Configuration Parameters:
{
"overall_query":… ,
"generation_mode": "without_planner",
"enable_filter": true,
"text_topk": 20,
"image_topk": 10
}
Stage 2: Outline Generation Phase
Chapter Planning:
Total Sections: 8
Section 0:
Section Description:
The article opens with the author's longstanding position, based on extensive expert consultation, that SARS-CoV-2 originated from a zoonotic spillover at a Chinese wet market, not from a laboratory leak. It addresses the persistence of the lab leak theory in public discourse, despite overwhelming scientific evidence to the contrary, and introduces the context for the interview with Dr. Philipp Markolin, author of 'Lab Leak Fever.' The section sets up the central controversy, the public's misconceptions, and the rationale for seeking expert insight.
Section Checklist:
☐ The scientific evidence overwhelmingly supports a zoonotic origin of SARS-CoV-2.☐ The lab leak hypothesis persists in public opinion despite being refuted by scientific data.☐ The article will explore the origins of SARS-CoV-2 through an expert interview.
Section: …
Stage 3: Section Content Generation Phase
Retrieved Context: Text & image Citations
Image Citation Example 1 (img1.json)
score: 0.8497941493988037,
text: "The predominant scientific consensus strongly favors a zoonot…
Image Citation Example 1 (img1.json)
score: 0.7449418306350708,
description": "Coarse-grained Description:\n- The image is a comparative diagram illustrating the potential transmission pathways and hosts for …
…
Content Generation
Section 0 Generated Content Preview:
Previous Summary: "This is the first section"
Previous Tail: "No previous content"
The origin of SARS-CoV-2, the virus responsible for the global COVID-19 pandemic, has been a subject of intense scientific investigation and public debate. While a small but vocal segment of the public continues to promote the so-called "lab leak hypothesis," the scientific community has long reached a consensus: the most plausible and well-supported explanation for the emergence of SARS-CoV-2 is a zoonotic spillover event, likely involving a wild animal at a wet market in Wuhan, China. This position is not based on speculation or political bias, but on a wealth of scientific evidence, including phylogenetic analysis, environmental sampling, and historical precedent for zoonotic disease transmission…
Section 1 Generated Content Preview:
Previous Summary: The article opens with the author's longstanding position, based on extensive expert consultation, that SARS-CoV-2 originated from a zoonotic spillover at a Chin…
Previous Tail: In the next section, we will examine the scientific and political forces that …
…
Stage 4:Final Report
Total Text Citations: 9
Total Image Citations: 10
Final Report:
Real Image Citations
The Zoonotic Origin of SARS-CoV-2: A Consensus in the Face of Persistent Misinformation
Section 0
Science Communication in the Age of Misinformation: Dr. Markolin's Perspective
…



Stage 1: User Input Phase
User Query:
How does the situation on the land border compare to the maritime domain between the two nations? What is the contrast between the land and maritime border issues?
Configuration Parameters:
{
"overall_query":… ,
"generation_mode": "without_planner",
"enable_filter": true,
"text_topk": 20,
"image_topk": 10
}
Stage 2: Outline Generation Phase
Chapter Planning:
Total Sections: 7
Section 0:
Section Description:
This section introduces the concept of the U.S.-Canada border as the world's longest peaceful land boundary, contrasting it with the existence of five unresolved maritime and territorial disputes. It sets the stage for the article by highlighting the potential for conflict in these areas if relations sour.
Section Checklist:
☐ The U.S.-Canada land border is largely peaceful and undefended, but several maritime locations are disputed.
☐ There are five key areas where Canada and the U.S. disagree over ownership.
Section: …
Stage 3: Section Content Generation Phase
Retrieved Context: Text & image Citations
Israel-Lebanon maritime border dispute
Image Citation Example 1 (img1.json)
score: 0.8173731565475464,
text: The United States and Canada have a total of 5 ongoing border di…
Image Citation Example 1 (img1.json)
score: 0.7898354530334473,
text: The image is a map illustrating the territorial claims over the Beaufort Sea between the United States and Canada.\n- It shows the ge…
…
Content Generation
Section 0 Generated Content Preview:
Previous Summary: "This is the first section"
Previous Tail: "No previous content"
The Peaceful Land Border and Unresolved Maritime Disputes
The United States and Canada share the distinction of maintaining the world's longest peaceful land border, spanning approximately 8,891 kilometers (5,525 miles). Unlike many international borders, this boundary is largely undefended and marked only by occasional monuments or markers [citation:txt1This peaceful relationship stands in stark contrast to the unresolved maritime and territorial disputes that continue to simmer in the waters surrounding the two countries. While the land border is a symbol of stability and cooperation, the maritime boundary issues between the U.…
Section 1 Generated Content Preview:
Previous Summary: This section introduces the concept of the U.S.-Canada border as the world's longest peaceful land boundary, contrasting it with the existence of five unresolved maritime and territorial disputes. It sets the stage for the article by highlighting the potential for…
Previous Tail: The next sections will explore the historical origins of these disputes and th…
…
Stage 4:Final Report
Total Text Citations: 15
Total Image Citations: 7
Real Image Citations
Final Report:
The Peaceful Land Border and Unresolved Maritime Disputes
Section 0
The Strait of Juan de Fuca Dispute: Equidistance, Base Points, and Natural Prolongation
Historical Background of the Strait of Juan de Fuca
The Principle of Equidistance and Base Point Disagreements
Five territorial disputes between the U.S. and Canada
The Peaceful Resolution of Territorial & Maritime Disputes
Emilia Justyna Powell & Krista E. Wiegand