STRUCTSCOPE: CAN LLMS MAINTAIN STRUCTURAL SCOPE IN LONG CONTEXTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid growth of structured information across long contexts, such as scientific documents, software artifacts, and harness-managed contexts, creates a need for models that can determine not only which information is relevant but also which structural scope makes it admissible. However, existing long-context benchmarks primarily evaluate models' capability to process text information with explicit structured information as fixed background. To bridge this gap, we introduce **StructScope**, a novel benchmark designed to evaluate models' capability in *structural scope maintenance*, i.e., applying query-specified scope constraints when selecting and attributing information. Although models show strong structural reasoning when evaluated on compact structural parse trees, our experiments reveal that their performance degrades sharply as the same structures are embedded within longer contexts. Further fine-grained experiments identify scope-constrained selection, particularly within nested structures, as a key bottleneck. Our findings highlight *structural scope maintenance* as an important dimension of long-context capability, motivating dedicated evaluation, training, and scope-aware context management in harness systems.

1 INTRODUCTION

As LLMs expand context windows to millions of tokens, their applications are rapidly extending from simple conversational assistants to complex document analysis, software engineering, and long-horizon agentic workflows. Across these demanding domains, model inputs are rarely plain text; instead, they take the form of system-assembled contexts that interleave natural language with explicit structural markers (Xia et al., 2024; Yang et al., 2024b; Deng et al., 2023; Zhou et al., 2023). Crucially, agent harnesses (Yao et al., 2023; Wu et al., 2023) organize instructions, resources, and tool observations through explicit scopes and references (Yang et al., 2024b; Lu et al., 2024). Together with the user request, these structures dictate how a model must navigate its long context to determine which instructions apply and which resources should be consulted. Figure 1 illustrates these requirements through a real-world workflow for updating a plugin. For example, even after correctly identifying *Update a Plugin* as the applicable scope, a model may fail to maintain that scope, allowing instructions under *Create a New Plugin* to influence its subsequent decisions. The four stages abstracted from this workflow motivate our evaluation of *structural scope maintenance*.

Existing long-context benchmarks evaluate retrieval, aggregation, and reasoning across diverse tasks (Li et al., 2024a; Bai et al., 2025; Yen et al., 2025; Hsieh et al., 2024; Kuratov et al., 2024; Modarressi et al., 2025; Maekawa et al., 2025), with some incorporating document hierarchies and structure-guided retrieval (Cui et al., 2025; Wang et al., 2026; Saad-Falcon et al., 2024). While Michelangelo (Vodrahalli et al., 2024) shifts the focus towards structural evaluation, it specifically targets latent contextual structures that must be resolved through semantic reasoning. However, because task success in these settings relies heavily on reasoning, it is hard to disentangle a model's inability to navigate structural relations from its inability to reason over the accessed content. To bridge this gap, we introduce **StructScope** to directly and systematically evaluate models' ability to apply query-specified scope constraints over explicit document structures and reference relations. To bypass the highly fragmented and heterogeneous nature of real-world agent data (e.g., skills, tool responses), which makes it challenging to organize and formalize for systematic benchmarking, we utilize arXiv LaTeX sources as a well-structured alternative. Specifically, by formalizing explicit LaTeX elements into deterministic structure trees, we ensure that correct answers depend solely on structural

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

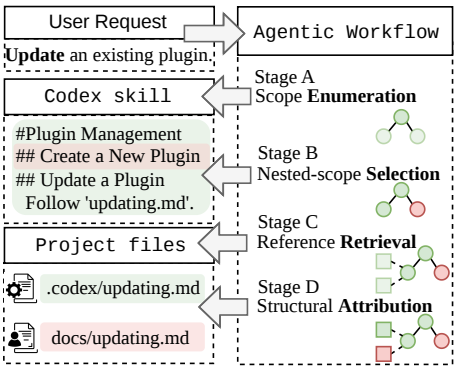


Figure 1: A four-stage abstraction inspired by an agentic workflow. Green highlights denote admissible scopes or targets, whereas red highlights indicate elements requiring structural exclusion.

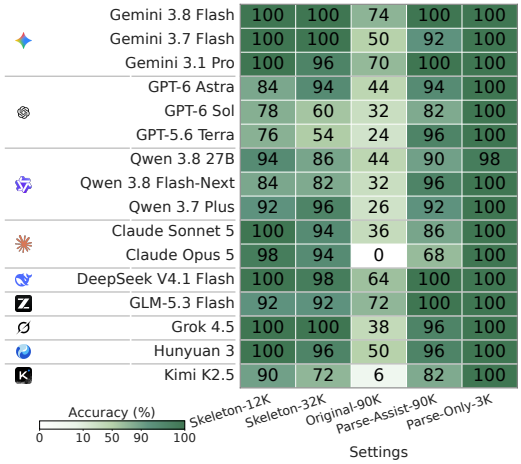


Figure 2: Full workflow query accuracy across five context configurations on a 50-query subset.

relations. Anchoring our evaluation against a simple, structure-only baseline, our benchmark isolates the evaluation of structural ability from complex semantic reasoning, enabling a precise diagnosis of how models select and attribute information under scope constraints.

StructScope is an automated benchmark pipeline capable of continuously generating evaluation instances from newly published arXiv LaTeX sources. In this study, our evaluation bank is constructed from 64 papers (spanning 70K–102K tokens), naturally preserving text alongside heading hierarchies, object definitions, and cross-references. We extract structural parse trees and reference links to generate 470 queries. This design enables automatic generation of the gold-standard and output verification, completely independent of the surrounding semantic claims. We design five tasks to evaluate the atomic and composite structural operations shown in Figure 1. By fixing the query and structural relations while manipulating context length, background text, and structural presentation, our design offers three key advantages: (i) **controlled context configurations**, (ii) **fine-grained capability diagnosis**, and (iii) **scalable, automated evaluation**.

Our evaluation reveals substantial length-dependent degradation (Figure 2). While most models excel on compact structural parse trees (Parse-Only-3K), even state-of-the-art models drop to 26–74% accuracy at original document length (Original-90K). Fine-grained analysis shows that scope-constrained selection, particularly under nested exclusions, emerges as a key bottleneck, whereas path attribution remains relatively robust. Furthermore, combining multiple structural operations increases failure rates. At comparable lengths, replacing background text with neutral filler produces marginal impact, whereas shortening background text yields substantial gains, suggesting that performance is more sensitive to the amount of text across which structural constraints must be tracked than to the semantic informativeness of the background text. While models demonstrate strong performance on standalone compressed representations, the utilization of the same information is generally limited and highly model-dependent when integrated into lengthy documents, motivating the design of agent harness systems that facilitate more robust tracking and utilization of structural information over extended context windows (e.g., through structure-aware context compaction).

In summary, our main contributions are as follows.

- We introduce StructScope, a scalable automated pipeline for evaluating *structural scope maintenance* in long contexts with explicit structure.
- We provide complementary diagnostics through task composition and context configuration.
- We instantiate a 470-query bank and reveal substantial length-dependent failures in *structural scope maintenance* through an evaluation of 35 models spanning diverse scales and architectures.

All resources (including the construction pipeline, the full bank, and the raw model responses) will be released with the paper.

2 Related Work

Long-Context Evaluation. The rapid expansion of context windows in large language models has driven a paradigm shift in evaluation, evolving from basic retrieval tasks to complex, multi-hop reasoning. Foundational diagnostics like the "needle-in-a-haystack" test and observations of the U-shaped performance curve (Liu et al., 2024; Roberts et al., 2025; Kuratov et al., 2024) initially highlighted models' failure to robustly access information across extended contexts. To provide more comprehensive assessments, recent benchmarks such as LooGLE (Li et al., 2024a), RULER (Hsieh et al., 2024), and HELMET (Yen et al., 2025) systematically scale context lengths to evaluate retrieval and aggregation. Furthermore, benchmarks like LongBench v2 (Bai et al., 2025), NoLiMa (Modarressi et al., 2025), and others (Goldman et al., 2024; Kaiser et al., 2025) push the boundary towards genuinely difficult long-context reasoning with synthetic distractors or tunable difficulties. Other specialized evaluations test interactive long-term memory (Wu et al., 2024), step-wise counterfactual reasoning (Wu et al., 2025), and massive database operations (Maekawa et al., 2025). While Michelangelo (Vodrahalli et al., 2024) introduces latent structure queries to move beyond literal matching, its evaluation still relies on reasoning over the underlying contextual content. In contrast to these works, which fundamentally entangle semantic reasoning with context utilization, StructScope explicitly isolates a model's structural ability.

Structure-Aware Document Processing. Real-world documents rely on rich explicit structures, driving the development of datasets incorporating hierarchies and layouts (Xia et al., 2024; Deng et al., 2025; Jang et al., 2026). For scientific literature, benchmarks broadly assess semantic understanding and citation generation (Cui et al., 2025; Wang et al., 2026; Zhang et al., 2025; Feng et al., 2026), while advanced methods increasingly leverage structure-guided retrieval and LaTeX parsing (Saad-Falcon et al., 2024; Huang et al., 2025; Jin et al., 2025; Verhoeff, 2026). However, these existing works typically emphasize either the multimodal layout-parsing capabilities of vision models or complex semantic reasoning over structured text. Crucially, they do not isolate structural navigation itself as a fundamental long-context challenge. StructScope strips away visual and semantic complexities to isolate core structural operations.

Structured Contexts in Agent Systems. While modern agentic frameworks heavily rely on navigating explicit structural boundaries—such as file hierarchies, HTML elements, and tool states (Yang et al., 2024b; Deng et al., 2023; Xie et al., 2024; Debenedetti et al., 2024)—current end-to-end benchmarks suffer from a critical diagnostic flaw. They fundamentally entangle high-level planning with this low-level structural grounding (Aghzal et al., 2026; Letta, 2025). This entanglement obscures whether task failures stem from flawed reasoning or lost structural context, frequently enabling models to bypass authentic navigation via semantic shortcuts (He & Roy, 2026). To resolve this, StructScope deliberately strips away agentic reasoning confounds, explicitly isolating a model's capacity for *structural scope maintenance*.

3 The StructScope Benchmark

StructScope evaluates whether a model can maintain explicit structural constraints while processing a long document. The benchmark separates two sources of difficulty: the gold answer is computed only from explicit structural relations, while the input condition determines whether those relations are presented compactly, embedded in a source document, or mixed with neutral or unrelated text. Thus, semantic interpretation of the claims in the document is not needed to determine the answer, although the original document conditions still test the recovery and maintenance of structural relations over long contexts.

3.1 Explicit Structure Representation

StructScope instantiates explicit document structure using LaTeX sources. Sectioning commands define heading boundaries, labeled environments define objects, and reference commands such as `\ref` identify their targets through matching `\label` keys.

We represent these elements as a directed graph with two node sets and their ownership and reference relations. Let $\mathcal{H}$ denote the set of headings, including the document root, and $\mathcal{O}$ the set of labeled objects, such as figures, tables, equations and theorems. Individual headings and objects are denoted

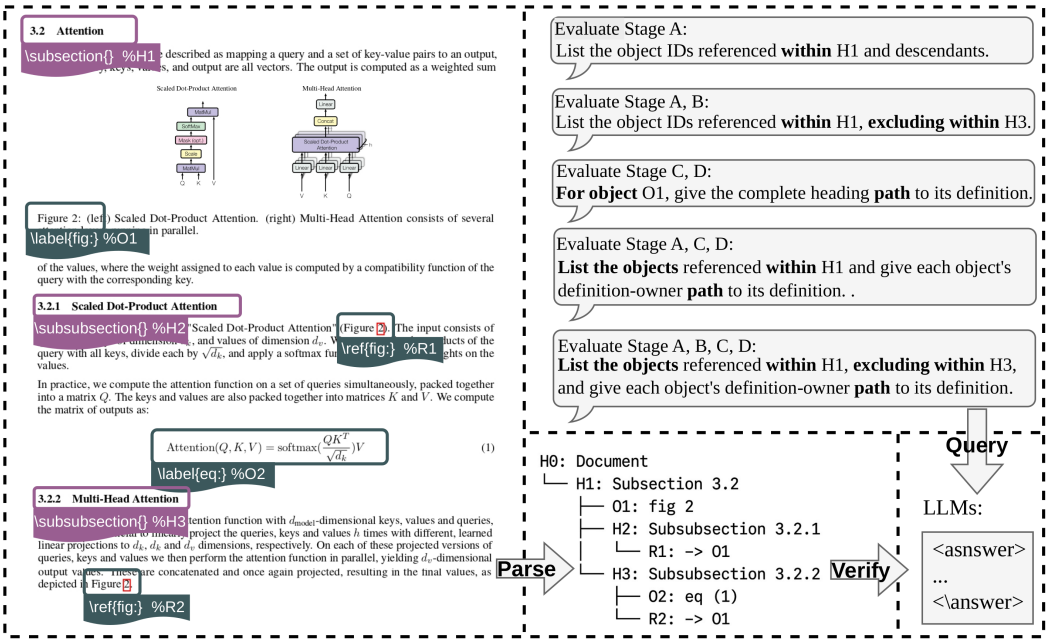


Figure 3: StructScope task construction and verification overview. The document snippet is adapted from (Vaswani et al., 2017) to illustrate our structural annotations, where purple markers denote headings and green markers denote object labels and references. Although depicted using a visual layout for clarity, actual model inputs and structural parsing only rely on LaTeX sources.

by $h \in \mathcal{H}$ and $o \in \mathcal{O}$. This representation captures structural relations independently of their surface syntax.

The *ownership* relation $R_{\text{own}} \subseteq \mathcal{H} \times (\mathcal{H} \cup \mathcal{O})$ links each heading to the elements it directly owns. An edge (h, h') indicates that h' is an immediate child heading of h, while (h, o) indicates that the definition of o begins in the source region directly owned by h. For LaTeX objects, this location is the beginning of the defining environment. The *reference* relation $R_{\text{ref}} \subseteq \mathcal{H} \times \mathcal{O}$ contains (h, o) when at least one reference to o occurs in the source text directly owned by h. Multiple references to the same object within a heading contribute a single pair. Ownership and reference therefore distinguish where an object is defined from where it is cited.

A deterministic parser extracts these elements and relations from LaTeX sources, as illustrated in Figure 3. Headings and objects receive flat, shuffled IDs that reveal neither document order nor hierarchy depth. These IDs remain fixed across context configurations for each paper; the ordered IDs in the figure are used only for illustration.

3.2 Task Design and Evaluation

Inspired by the workflow in Figure 1, we define five tasks by designating document elements as query input. Scope queries specify a positive heading $P \in \mathcal{H}$ and, when exclusion is required, a heading $N \in \mathcal{H}$ that is a descendant of P. The Supplied-target queries instead provide an object-ID set $L \subseteq \mathcal{O}$.

The scope of a heading includes itself and all descendant headings under R_{own}, together with their contents. We use $\text{Ref}(P)$ to denote the set of distinct objects referenced anywhere in the scope of P, collected through R_{ref}. **Enumeration (EN)** returns this reference set. **Selection (SE)** additionally excludes objects referenced within the scope of N, giving

$$\text{Sel}(P, N) = \text{Ref}(P) \setminus \text{Ref}(N). \quad (1)$$

Exclusion applies to object identities. An object referenced under N is therefore removed even if it is also referenced elsewhere under P.

Table 1: Input and output formulations for the five StructScope tasks and their corresponding stages in the workflow(Figure 1). The document context is given in every task.

Task	Stages	Input	Output
EN	A	P	Object IDs referenced within the scope of P
SE	A, B	P, N	Object IDs in the scope of P after excluding references under N
RA	C, D	L	Each object ID in L with its complete definition path
EN-RA	A, C, D	P	Each object referenced within P with its definition path
SE-RA	A, B, C, D	P, N	Each object selected from P after excluding N, with its definition path

Retrieval and Attribution (RA) receives the object-ID set L directly in the query. *Retrieval* means locating each object's definition in the document by its ID; definition locations and paths are not supplied. *Attribution* recovers the complete heading path from the root to the definition owner by tracing R_{own} backward, including every intervening heading and excluding the object itself. Table 1 summarizes the five tasks by their inputs, required outputs, and their corresponding stages in the abstracted workflow. EN and SE return object IDs, while **EN-RA** and **SE-RA** additionally require a definition path for each enumerated or selected object. These composite forms therefore combine target IDs within the specified scopes with definition lookup and structural attribution.

We report *query-level exact-match accuracy*, the percentage of queries whose required output is entirely correct. Specifically, EN and SE require the exact gold object set, ignoring the order of object entries. RA requires exactly the supplied objects alongside their complete definition paths. EN-RA and SE-RA require both the exact object set and every definition path to be perfectly recovered. Missing or extra objects, duplicate or unknown IDs, and incorrect or incomplete paths receive no query-level credit. Responses are parsed with deterministic normalization; unavailable, truncated, or unrecoverable answers count as incorrect. We further evaluate more granular intermediate diagnostics, including object-set accuracy (evaluated only on object IDs regardless of path errors) and object-level recall (measured both by ID-only and joint ID-and-path recovery). Complete results for these metrics, alongside a detailed breakdown of the failure modes, are provided in Appendix D.

3.3 Controlled Context Configurations

Within each task form, we vary the context while retaining the query, IDs, structural relations, and gold answer. Figure 4 illustrates the configurations. For readability, **90K** denotes original document lengths of **70–102K tokens** and their corresponding long-context variants; **3K** denotes compact parses of **2–4K tokens**. These are shorthand labels, not identical lengths across papers.



Figure 4: Context configurations over the same document structure.

Original retains the original document. **Skeleton** retains the structure-bearing markup but replaces the background text with neutral filler, repeating `The quick brown fox jumps over the lazy dog.` (Hsieh et al., 2024; Mohtashami & Jaggi, 2023). Adjusting the filler yields the 12K, 32K, and original-length variants. Structural elements retain their source order, and filler is distributed across heading regions to approximately preserve their relative positions. **Parse-Only** presents the complete hierarchy, ownership, and reference relations as a compact serialization without prose. **Parse-Assist** inserts the corresponding parse relations into their associated sections of the original source. We additionally evaluate **Parse-Appended**, which places the complete parse serialization as a single block after the document. We evaluate SE-RA across these primary context configurations, while the remaining task forms are evaluated under Skeleton-32K and Original. Detailed length matching and parse placement construction, along with the evaluation results for the **Parse-Appended** configuration, are provided in Appendix B and Appendix B.3.

3.4 Dataset Construction

The evaluation bank contains 470 queries from 64 papers. We generate eligible heading pairs and object sets from the parsed relations. During construction, target objects are restricted to definitions outside the positive scope to avoid trivial local attribution. We retain nonempty targets of two to four objects and sample across reference-to-definition distances. Gold sets and paths are computed programmatically and reused across the corresponding context variants. Source statistics, parsing and sampling rules, and template examples are provided in Appendix A.

4 Experiments

We evaluate the complete 470-query bank across diverse locally deployed open-weight models with 8B–122B parameters, covering both dense and mixture-of-experts (MoE) architectures (Jiang et al., 2024). We additionally evaluate proprietary models on a 50-query subset (Figure 2). The evaluated open-weight families include Qwen (Bai et al., 2023; Yang et al., 2024a), Gemma (Gemma Team, 2024), Seed-OSS (ByteDance Seed Team, 2024), Mistral/Devstral (Jiang et al., 2023), Nemotron (Parmar et al., 2024), and Muse (Meta AI, 2024). Main experiments use greedy decoding ($T = 0$) with thinking mode disabled to ensure comparable reasoning budgets. The evaluation prompts provide Chain-of-Thought (CoT) guidance and restrict outputs to a 6K-token limit, operating under a 16K maximum generation length (extended to 32K for supplementary thinking-enabled runs). Detailed generation statistics and failure analyses are provided in Appendix E, with supplementary thinking-mode budgets reported in Appendix F.5.

Table 2: Comprehensive evaluation of model capabilities on structural tasks. The left panel breaks down performance by task composition; the right panel contrasts SE-RA across different text configurations. Cell shading corresponds to accuracy, with darker green denoting higher performance.

	Task Composition*								**Context Configuration** (SE-RA)					
	32K				90K				3K	12K	32K	90K		
Model	EN	SE	RA	EN-RA	EN	SE	RA	EN-RA	Parse	Skeleton	Skeleton	Parse	Skeleton	**Original**
Qwen3.8 27B	99.0	94.1	92.2	95.1	79.4	65.7	93.1	71.6	99.6	86.6	83.6	92.6	62.3	59.8
Qwen3.8 Flash-Next	100.0	99.0	95.1	91.2	75.5	52.9	81.4	54.9	100.0	89.6	85.5	89.4	49.1	47.4
Qwen3.6 27B	99.0	100.0	94.1	88.2	57.8	44.1	86.3	47.1	100.0	84.7	84.3	90.9	42.6	40.0
Qwen3.5 27B	95.1	88.2	93.1	87.3	57.8	25.5	73.5	41.2	99.8	82.8	78.9	74.3	25.5	22.3
Muse Glimmer 30B	95.1	91.2	94.1	91.2	34.3	16.7	76.5	25.5	99.8	93.0	86.8	78.9	19.8	18.9
Gemma-4 31B	94.1	95.1	96.1	71.6	52.0	21.6	75.5	29.4	99.1	91.5	74.3	83.6	24.9	16.6
Qwen3.6 35B-A3B	91.2	80.4	40.2	52.0	42.2	24.5	52.0	16.7	94.7	51.3	49.8	67.0	18.5	13.6
Qwen3.5 122B-A10B	98.0	93.1	67.6	57.8	26.5	10.8	67.6	20.6	97.0	60.4	58.7	61.9	14.7	10.9
Qwen-AgentWorld	96.1	85.3	36.3	49.0	50.0	12.7	30.4	10.8	96.6	31.3	19.8	25.1	11.9	10.0
Qwen3-Next 80B-A3B	88.2	49.0	46.1	37.3	20.6	2.9	30.4	7.8	93.6	53.8	43.2	33.2	3.6	3.2
Qwen3.5 35B-A3B	77.5	68.6	53.9	50.0	17.6	5.9	44.1	10.8	99.8	34.3	26.0	47.7	4.5	3.0
Seed-OSS 36B	100.0	86.3	76.5	60.8	43.1	6.9	60.8	18.6	99.8	62.8	50.0	23.6	5.7	2.6
GPT-OSS 120B	97.1	62.7	64.7	46.1	23.5	6.9	38.2	6.9	98.7	57.2	27.0	36.8	1.7	1.7
Gemma-4 26B-A4B	52.0	17.6	36.3	13.7	12.7	2.0	47.1	2.9	92.6	10.0	2.3	17.9	1.5	1.5
Gemma-4 E4B	38.2	2.9	0.0	2.9	9.8	1.0	6.9	2.0	82.8	2.8	0.6	8.7	1.1	1.1
Nemotron-3 Super*	77.5	30.4	70.6	25.5	30.4	4.9	41.2	5.9	97.1	22.5	10.8	54.9	0.0	1.0
Qwen3 30B-A3B	49.0	13.7	15.7	12.7	1.0	0.0	25.5	1.0	94.3	7.8	6.9	10.6	1.1	0.2
Qwen3.5 9B	47.1	39.2	17.6	12.7	14.7	3.9	40.2	5.9	97.4	11.7	8.5	13.7	0.9	0.0
Devstral-Small-2 24B	78.4	14.7	42.2	19.6	0.0	0.0	4.9	0.0	96.1	12.1	6.6	0.6	0.0	0.0
GPT-OSS 20B*	41.2	2.9	8.8	1.0	2.9	0.0	9.8	0.0	80.6	6.9	0.0	3.8	0.0	0.0
Mistral-Small-3.2 24B*	31.4	8.8	34.3	5.9	0.0	0.0	5.9	1.0	82.7	5.7	2.3	1.0	0.0	0.0

*: Evaluated under a fixed 102 queries subset randomly selected.

4.1 Main Results

4.1.1 Compact-Structure Baseline and Overall Performance

Table 2 establishes a compact-structure baseline for the full composite task. Most models with available Parse-Only results achieve over 90% accuracy, with several approaching 100%. These results

show that the query rules and required structural operations are within these models' capabilities when the relevant relations are presented compactly. However, this strong baseline does not consistently transfer to longer contexts, where the same queries and gold answers yield substantially lower accuracy. We establish Parse-Only as the execution baseline for subsequent analyses of context length, task composition, and parsing assistance.

4.1.2 Context Length and Structural Presentation

As shown in Figure 2 and Table 2, performance degradation at extended context lengths is a shared limitation, though the critical thresholds vary. We observe two distinct degradation patterns. The first is **late degradation**, where models remain stable up to 32K but drop sharply at their native document lengths (e.g., Gemini 3.8 Flash drops from 100% at 32K to 74% at 90K, and Claude Opus 5 collapses from 94% to 0%). The second is **early degradation**, where models steadily decline as length increases; for instance, Muse Glimmer 30B degrades from 87% at 12K to 62% at 32K, falling to 17% at the original length.

Introducing Parse-Assist mitigates this degradation, but the recovery is highly model-dependent. Strong models like Gemini 3.8 Flash achieve full recovery, whereas Claude Opus 5 recovers to 68%. Together with the high Parse-Only scores, this contrast indicates that models can use the relations when they are presented compactly but have difficulty recovering and maintaining them when they are distributed across a long document. This motivates more effective structure-aware context organization.

4.1.3 Basic Structural Operations and Composition

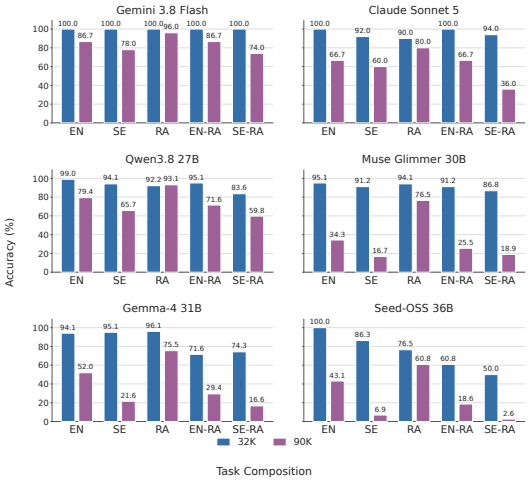


Figure 5: Accuracy across task compositions under 32K and 90K text configurations.

Figure 5 compares Enumeration (EN), Selection (SE), Retrieval and Attribution (RA), and their composite forms for six models. Structural operations are **highly length-sensitive**, especially Enumeration (EN) and Selection (SE). While Retrieval and Attribution (RA) is less sensitive, fluctuating mostly within 20 percentage points from Skeleton-32K to Original-90K, whereas EN and SE decline substantially. SE is the most vulnerable operation, decreasing by 22.0–79.4 points; at the original length, its accuracy falls to 16.7% for Muse Glimmer 30B and 6.9% for Seed-OSS 36B. The broader results in Table 2 follow the same pattern, identifying nested-scope exclusion as the key bottleneck.

Table 3: Intermediate-result analysis on Original-90K. Downstream RA is evaluated as the object-level ID and path accuracy for the extracted items, conditioned on exact set-level correctness ($+/-$) of the upstream EN or SE. Subscripts denote the number of queries.

		EN-RA		SE-RA	
Model	RA	EN$^+$	EN$^-$	SE$^+$	SE$^-$
Qwen3.8 27B	93.1_{102}	90.1_{81}	66.7_{21}	85.1_{67}	74.3_{35}
Muse Glimmer 30B	76.5_{102}	68.4_{38}	59.4_{64}	82.4_{17}	61.2_{85}
Gemma-4 31B	75.5_{102}	62.5_{48}	74.1_{54}	93.3_{15}	65.5_{87}
Seed-OSS 36B	60.8_{102}	59.4_{32}	42.9_{70}	66.7_{3}	33.3_{99}

Combining operations introduces additional difficulty by **propagating upstream errors**. As shown in Table 3, an incorrect upstream enumeration or selection propagates errors into the subsequent context and significantly degrades downstream performance. For instance, in SE-RA, Gemma-4 31B's downstream accuracy drops from 93.3% to 65.5% when the upstream selection fails. Driven by these cascading errors, the overall composite accuracy suffers a massive decline. From Skeleton-32K to Original-90K, EN-RA and SE-RA decrease by 13.3–65.7 and 23.8–67.9 points relative to their single-step baselines. Even at the shorter Skeleton-32K length, Gemma-4 31B's accuracy drops from 94.1% (EN) to 71.6% (EN-RA). Consequently, the difficulty of composite tasks is multiplicative rather than additive.

4.1.4 Background Content and Structural Interference

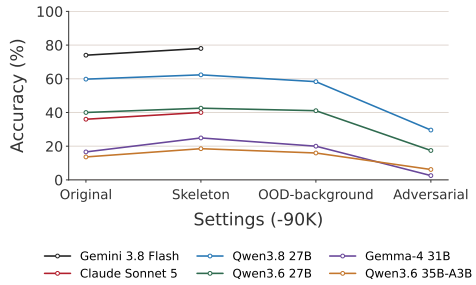


Figure 6: Full-workflow accuracy under background text manipulation.

Figure 6 shows comparatively modest changes under background manipulation. Skeleton replaces the background prose with neutral filler, while OOD-background uses biology text stripped of structural markers. Among models evaluated under all four conditions, accuracy differs from Original by at most 8.3 percentage points for Skeleton and 4.7 points for OOD-background.

In contrast, the Adversarial condition injects approximately 5K tokens of plain text into the original document, serving as interference by making misleading claims about the explicit structure. This interference reduces accuracy for every illustrated model, including Qwen3.8-27B from 59.8% to 29.6%. Models therefore remain vulnerable when plain text describes false structural attributions that contradict the actual explicit hierarchy.

4.1.5 Structural Factors and Task Difficulty

We provide a breakdown of SE-RA accuracy under Original-90K for six models, grouping queries by target-set size, definition-path depth, citation-to-definition distance, and document heading count. Target-set size shows the most consistent pattern, with lower accuracy on four-object queries than on two-object queries across these models. The remaining factors show less uniform patterns. Deeper paths and longer reference distances are associated with lower accuracy for some models, but neither factor, nor heading count, exhibits a consistent monotonic trend across models. Detailed accuracies and query counts for each group are provided in Appendix F.2(Table 13).

4.2 Discussion

4.2.1 Can Retrieval Substitute for Structural Scope Maintenance?

To test whether retrieve-then-read pipelines(Jin et al., 2025) can bypass structural index construction, we evaluate a BM25 baseline (Li et al., 2024b). Under a matched 32K context budget, retrieval yields only 4.5–10.4% accuracy (Table 4), lagging far behind the contiguous Skeleton-32K baseline (49.8–85.5%). Standard similarity retrieval fails because naive chunking fundamentally disrupts the global hierarchical topology required for nested scope exclusion and exhaustive set recall.

Table 4: SE-RA accuracy (%) on matched queries per model. RAG uses BM25 with 1,024-token chunks, top-32 retrieval, and a 32K token budget.

Model	N	Original-90K	RAG-32K	Parse-Assist-90K
Qwen3.8 27B	470	59.8	10.4	92.6
Qwen3.8 Flash-Next	470	47.4	7.2	89.4
Qwen3.6 27B	378	37.6	7.1	89.9
Gemma-4 31B	470	16.6	5.3	83.6
Qwen3.6 35B-A3B	470	13.6	4.5	67.0

4.2.2 How Do Architecture and Post-Training Impact Structural Performance?

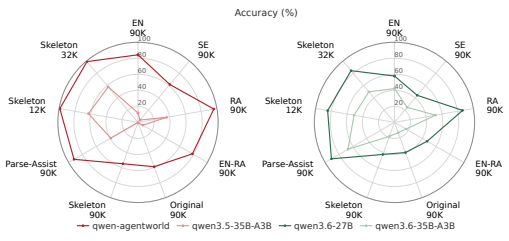


Figure 7: Model architecture and post-training comparisons.

Figure 7 reveals clear performance trends regarding post-training alignment and model architecture when strictly controlling for foundation bases within the Qwen family. Regarding post-training (evaluated with "think mode" enabled), the agentic-aligned Qwen-AgentWorld exhibits a comprehensive advantage over its instruction-tuned counterpart (Qwen3.5-35B-A3B) derived from the exact same pre-trained base, suggesting a strong connection between *structural scope maintenance* and agentic capabilities. Regarding architecture, dense designs demonstrate a decisive

advantage: the dense Qwen3.6-27B substantially outperforms its version-matched Mixture-of-Experts (MoE) sibling (Qwen3.6-35B-A3B) across all structural operations and context configurations.

4.2.3 Can Reasoning Compensate for Context Length?

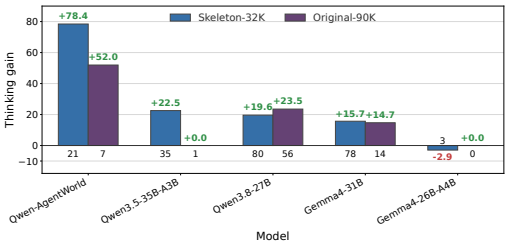


Figure 8: Impact of enabling the thinking mode on SE-RA accuracy. Black numbers on the x-axis denote the thinking-disabled baseline.

Figure 8 shows that enabling reasoning ("thinking") generally improves structural performance, but the magnitude of these gains does not exhibit a direct correlation with the models' capabilities. For the majority of models, the performance increases remain relatively marginal. Qwen-AgentWorld stands out as a singular exception, demonstrating a highly significant improvement in accuracy. This confirms that while reasoning acts as a structural catalyst, it cannot fundamentally bypass a model's intrinsic long-context limitations. Further detailed breakdowns of these performance shifts are provided in Appendix F.5.

4.2.4 Does Semantic Long-Context Ability Predict Structural Scope?

To contextualize our benchmark, we compare it with AA-LCR (Artificial Analysis, 2025), a standard for semantic long-context evaluation (Appendix F.1). While they exhibit a moderate rank correlation (Spearman's $\rho = 0.679$), the weaker linear correlation (Pearson's $r = 0.538$) suggests that gains in semantic reasoning do not necessarily guarantee proportional improvements in structural ability. This indicates that StructScope can offer a complementary perspective for long-context evaluation.

5 Limitations

Substrate and generalization. StructScope is evaluated on LaTeX scientific papers. Although our modeled structural relations generalize to other substrates, differences in markup conventions, document quality, and pretraining familiarity may affect both performance and failure patterns. Furthermore, our evaluation is limited to single-turn queries over static documents, whereas real agents operate over dynamic, heterogeneous, multi-turn contexts. Extending StructScope to such settings is a challenging direction for future work.

Operation coverage. We evaluate scope expansion, nested-scope selection, definition retrieval, upward path attribution, and set difference. This is a deliberate subset of possible structural operations. We leave cross-relation traversals, set operations such as intersection and union, positional and ordering relations, and structural mutation outside the benchmark.

Context length control. Relying on real LaTeX documents inherently yields approximate context lengths. Consequently, some of our context configurations are constrained within length bands rather than exact token targets. While our pipeline supports highly granular length stratification, we prioritize these broader intervals to establish the presence of general length-dependent degradation before identifying exact transition points.

6 Conclusion

We introduced StructScope to evaluate structural scope maintenance: applying query-specified scope constraints when selecting and attributing information. Built from real LaTeX sources, the benchmark provides complementary diagnostics through task composition and controlled context configurations. While models excel on compact structural parses, they degrade substantially in longer contexts, bottlenecked primarily by scope-constrained selection under nested exclusions. Background replacement has comparatively limited effects, while parsing assistance yields model-dependent recovery. These findings highlight structural scope maintenance as an important dimension of long-context capability, motivating dedicated evaluation, training, and scope-aware context management in agent harnesses.

REFERENCES

Mohamed Aghzal, Gregory J. Stein, and Ziyu Yao. Why do llm-based web agents fail? a hierarchical planning perspective, 2026. URL https://arxiv.org/abs/2603.14248.

Artificial Analysis. Announcing artificial analysis long context reasoning (aa-lcr), 2025. URL https://artificialanalysis.ai/articles/announcing-aa-lcr.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yushi Bai et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *ACL*, 2025.

ByteDance Seed Team. Seed-oss: Open-weight models for advanced reasoning. Hugging Face Repository, 2024. URL https://huggingface.co/seed.

Hao Cui et al. Curie: Evaluating llms on multitask scientific long context understanding and reasoning. In *ICLR*, 2025.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, et al. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Chao Deng et al. Longdocurl: A comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *ACL*, 2025.

Xiang Deng et al. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2023.

Xiang Feng et al. Docscope: Benchmarking verifiable reasoning for trustworthy long-document understanding. *arXiv preprint arXiv:2605.08888*, 2026.

Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Omer Goldman et al. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. In *EMNLP*, 2024.

Kang He and Kaushik Roy. Swe-adept: An llm-based agentic framework for deep codebase analysis and structured issue resolution, 2026. URL https://arxiv.org/abs/2603.01327.

Cheng-Ping Hsieh et al. Ruler: What's the real context size of your long-context language models? In *COLM*, 2024.

Xinhao Huang et al. Seal: Structure and element aware learning to improve long structured document retrieval. In *EMNLP*, 2025.

Haeun Jang, Hwan Chang, and Hwanhee Lee. Doc-pp: Document policy preservation benchmark for large vision-language models. In *Findings of ACL*, 2026.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Jiajie Jin et al. Hierarchical document refinement for long-context retrieval-augmented generation. In *ACL*, 2025.

Daniel Kaiser, Arnoldo Frigessi, Ali Ramezani-Kebrya, and Benjamin Ricaud. Cogniload: A synthetic natural language reasoning benchmark with tunable length, intrinsic difficulty, and distractor density. *arXiv preprint arXiv:2509.18458*, 2025.

Yuri Kuratov et al. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Letta. Context-bench: Benchmarking llms on agentic context engineering. *Letta Blog*, October 2025. URL https://www.letta.com/blog/context-bench/.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? In *ACL*, 2024a.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 881–893, 2024b.

Nelson F. Liu et al. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 2024.

Jiarui Lu et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. *arXiv preprint arXiv:2408.04682*, 2024.

Seiji Maekawa, Hayate Iso, and Nikita Bhutani. Holistic reasoning with long-context lms: A benchmark for database operations on massive textual data. In *ICLR*, 2025.

Meta AI. Muse: Open agentic models for edge devices. Meta Research Releases, 2024. URL https://research.meta.ai/.

Ali Modarressi et al. Nolima: Long-context evaluation beyond literal matching. In *ICML*, 2025.

Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite-context length for transformers. *arXiv preprint arXiv:2305.16300*, 2023.

Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, et al. Nemotron-4 15b technical report. *arXiv preprint arXiv:2402.16819*, 2024.

Jonathan Roberts, Kai Han, and Samuel Albanie. Needle threading: Can llms follow threads through near-million-scale haystacks? In *ICLR*, 2025.

Jon Saad-Falcon et al. Pdftriage: Question answering over long, structured documents. In *EMNLP Industry Track*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tom Verhoeff. Ai-friendly latex: Using latex code as a knowledge source for retrieval-augmented generation. *arXiv preprint arXiv:2605.22923*, 2026.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. Michelangelo: Long context evaluations beyond haystacks via latent structure queries, 2024. URL https://arxiv.org/abs/2409.12640.

Renxi Wang et al. Scalar: Scientific citation-based live assessment of long-context academic reasoning. In *EACL*, 2026.

Di Wu et al. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. Cofca: A step-wise counterfactual multi-hop qa benchmark. In *ICLR*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Renqiu Xia et al. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024.

Tianbao Xie et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

An Yang, Baosong Yang, Zheng Biao, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

John Yang, Carlos E. Jimenez, et al. Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Howard Yen et al. Helmet: How to evaluate long-context language models effectively and thoroughly. In *ICLR*, 2025.

Jiajie Zhang et al. Longcite: Enabling llms to generate fine-grained citations in long-context qa. In *Findings of ACL*, 2025.

Shuyan Zhou et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

A Source Papers and Benchmark Construction

A.1 Corpus and Per-Paper Metadata

Table 5 summarizes the fixed bank, and Table 6 lists every paper with its title, arXiv posting month, source length, structural counts, and SE-RA query count. The source directories identify collection snapshots rather than publication-date filters. Posting months span April 2023–June 2026. Exact arXiv version dates and author metadata are not stored in this input bank and are not inferred from directory names.

Table 5: Bank statistics recomputed from the question records. Distances are per-query means and path lengths include the root.

Property	Value
Papers / SE-RA queries	64 / 470
Source tokens (min / median / max)	70,457 / 84,092 / 102,120
arXiv posting months	2023-04 to 2026-06
Gold objects per query	2 objects: 225; 3 objects: 141; 4 objects: 104
Gold object occurrences	1289
Minimum / maximum gold path length	3 / 5
Reference-to-definition distance in characters (min / median / max)	1,524 / 57,251 / 204,038

Table 6: Source-paper metadata for the 64-paper bank. Month is decoded from the arXiv identifier, not a model cutoff date. Tokens are stored source-token counts. H counts headings excluding the document root, O counts labeled objects, C counts retained reference occurrences, and Q counts SE-RA queries.

arXiv ID	Title	Month	Tokens	H	O	C	Q
2304.14419	Unsupervised Learning of Robust Spectral Shape Matching	2023-04	71,458	32	28	39	12
2305.06426	Planning a Community Approach to Diabetes Care in Low- and Middle-Income Countries Using Optimization	2023-05	74,467	67	76	134	20
2406.14194	VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model	2024-06	76,698	63	42	69	8
2411.09355	Prices, Bids, Values: One ML-Powered Combinatorial Auction to Rule Them All	2024-11	74,289	71	28	120	10
2412.19363	Large Language Models for Market Research: A Data-augmentation Approach	2024-12	84,453	44	46	73	8
2502.06096	Post-detection inference for sequential changepoint localization	2025-02	77,521	41	70	155	2
2503.12181	Action-Gradient Monte Carlo Tree Search for Non-Parametric Continuous (PO)MDPs	2025-03	71,941	83	52	60	5
2503.20607	A decision-theoretic approach to dealing with uncertainty in quantum mechanics	2025-03	94,057	30	63	138	4
2505.16831	Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs	2025-05	95,791	39	33	48	1
2506.03610	Orak: A Foundational Benchmark for Training and Evaluating LLM Agents on Diverse Video Games	2025-06	86,822	115	39	60	3
2506.10054	Uni-DPO: A Unified Paradigm for Dynamic Preference Optimization of LLMs	2025-06	73,277	108	37	75	12
2507.01833	Refining Gelfond Rationality Principle: Towards More Comprehensive Foundational Principles for Answer Set Semantics	2025-07	94,910	51	68	245	11
2507.04441	A Category-Theoretic Analysis of Conformal Prediction	2025-07	101,130	34	62	129	1
2507.07969	Reinforcement Learning with Action Chunking	2025-07	79,128	35	32	53	5
2509.23597	Characteristic Root Analysis and Regularization for Linear Time Series Forecasting	2025-09	95,332	95	59	94	13
2510.08045	Verifying Quantized GNNs With Readout Is Decidable But Highly Intractable	2025-10	78,350	49	73	158	2
2512.01152	Open-Set Domain Adaptation Under Background Distribution Shift: Challenges and A Provably Efficient Solution	2025-12	85,994	36	44	99	5
2512.23770	SB-TRPO: Towards Safe Reinforcement Learning with Hard Constraints	2025-12	101,631	84	25	54	13

Table 6 continued

arXiv ID	Title	Month	Tokens	H	O	C	Q
2601.11016	Contextual Distributionally Robust Optimization with Causal and Continuous Structure: An Interpretable and Tractable Approach	2026-01	91,062	42	93	220	2
2601.14053	LLMOrbit: A Circular Taxonomy of Large Language Models -From Scaling Walls to Agentic AI Systems	2026-01	75,279	133	22	31	1
2602.02112	Unifying Masked Diffusion Models with Various Generation Orders and Beyond	2026-02	85,913	41	48	63	5
2602.12224	Two-Sided Time-Independent Regret for Matching Markets with Limited Interviews	2026-02	89,156	77	50	200	5
2602.14872	The Implicit Curriculum: Learning Dynamics in RL with Verifiable Rewards	2026-02	72,695	92	90	178	8
2602.16813	Flow Map Language Models: One-step Language Modeling via Continuous Denoising	2026-02	80,905	82	78	156	10
2602.17315	Flickering Multi-Armed Bandits	2026-02	84,174	112	54	146	16
2602.23163	A Decision-Theoretic Formalisation of Steganography With Applications to LLM Monitoring	2026-02	83,909	124	26	46	8
2603.01589	SafeSci: Safety Evaluation of Large Language Models in Science Domains and Beyond	2026-03	94,672	49	12	19	13
2603.03099	Why Adam Can Beat SGD: Second-Moment Normalization Yields Sharper Tails	2026-03	82,154	44	101	167	6
2603.07514	A Unified View of Score-Based and Drifting Models	2026-03	70,587	100	43	91	20
2603.18563	Reasonably reasoning AI agents can avoid game-theoretic failures in zero-shot, provably	2026-03	86,796	99	68	169	4
2603.28590	MonitorBench: A Comprehensive Benchmark for Chain-of-Thought Monitorability in Large Language Models	2026-03	90,188	75	12	17	2
2604.00987	Bridging Structured Knowledge and Data: A Unified Framework with Finance Applications	2026-04	94,000	78	71	107	4
2604.07413	FORGE: Fine-grained Multimodal Evaluation for Manufacturing Scenarios	2026-04	91,850	46	22	35	2
2604.09417	Do We Really Need to Approach the Entire Pareto Front in Many-Objective Bayesian Optimisation?	2026-04	72,057	57	48	78	3
2604.10727	Tail-Aware Information-Theoretic Generalization for RLHF and SGLD	2026-04	72,559	87	77	221	9
2604.10958	Continuous-time Online Learning via Mean-Field Neural Networks: Regret Analysis in Diffusion Environments	2026-04	76,701	30	83	231	6
2604.15174	MambaSL: Exploring Single-Layer Mamba for Time Series Classification	2026-04	70,457	45	23	42	9
2604.15558	Preregistered Belief Revision Contracts	2026-04	72,531	133	81	190	19
2604.18882	Formally Verified Patent Analysis via Dependent Type Theory: Machine-Checkable Certificates from a Hybrid AI + Lean 4 Pipeline	2026-04	84,859	51	65	495	4
2604.20744	AAC: Admissible-by-Architecture Differentiable Landmark Compression for ALT	2026-04	73,073	140	42	238	16
2604.24737	Learning to Think from Multiple Thinkers	2026-04	92,500	106	82	271	11
2604.26976	Fitting Horn DL Ontologies to ABox and Query Examples: A Tale of Simulation Quantifiers and Finite Models	2026-04	86,761	36	27	81	3
2605.00733	EASE: Federated Multimodal Unlearning via Entanglement-Aware Anchor Closure	2026-05	75,565	78	32	71	20
2605.02814	IConFace: Identity-Structure Asymmetric Conditioning for Unified Reference-Aware Face Restoration	2026-05	77,852	30	13	14	2
2605.09296	Micro-Defects Expose Macro-Fakes: Detecting AI-Generated Images via Local Distributional Shifts	2026-05	73,505	60	79	160	14
2605.14323	Dynamic Latent Routing	2026-05	88,076	56	52	98	1
2605.15053	TFGN: Task-Free, Replay-Free Continual Pre-Training Without Catastrophic Forgetting at LLM Scale	2026-05	75,931	92	26	49	5
2605.15285	Universal Approximation of Nonlinear Operators and Their Derivatives	2026-05	82,598	74	105	271	15
2605.15959	When and Why Adversarial Training Improves PINNs: A Neural Tangent Kernel Perspective	2026-05	84,515	166	109	161	20
2605.16679	CHI-Bench: Can AI Agents Automate End-to-End, Long-Horizon, Policy-Rich Healthcare Workflows?	2026-05	82,976	142	32	48	11
2605.17160	When Bits Break Recourse: Counterfactual-Faithful Quantization	2026-05	76,495	167	85	140	3
2605.18661	AI for Auto-Research: Roadmap & User Guide	2026-05	77,117	100	10	11	4
2605.18905	Stability and Discretization Error of State Space Model Neural Operators	2026-05	84,303	75	74	161	2
2605.21489	Variance Reduction for Expectations with Diffusion Teachers	2026-05	92,361	79	41	130	14

Table 6 continued

arXiv ID	Title	Month	Tokens	H	O	C	Q
2605.23491	CoSPlay: Cooperative Self-Play at Test-Time with Self-Generated Code and Unit Test	2026-05	88,303	65	48	73	1
2605.24042	Hidden-State Privacy Has an Empty Middle	2026-05	89,802	154	95	189	4
2605.24484	SPACE: Unifying Symmetric and Asymmetric Routing Problems for Generalist Neural Solver	2026-05	74,367	86	34	48	5
2605.24900	ProActor: Timing-Aware Reinforcement Learning for Proactive Task Scheduling Agents	2026-05	80,092	188	48	62	7
2605.25789	On the Benefits of Free Exploration for Regret Minimization in Multi-Armed Bandits	2026-05	84,132	55	59	144	4
2605.26654	Bilevel Optimization over Saddle Points of Zero-Sum Markov Games	2026-05	97,770	33	63	124	4
2605.27593	Voluntary Collusion with Secret Tools in Competing LLM Agents	2026-05	85,476	81	26	57	3
2605.29267	When and How Human Curation Backfires: Preference Alignment under Multi-Model Self-Consuming Loop	2026-05	84,053	40	83	225	1
2606.00440	SDR: Set-Distance Rewards for Radiology Report Generation	2026-06	102,120	50	22	35	1
2606.00570	Revisiting Parameter-Based Knowledge Editing in Large Language Models: Theoretical Limits and Empirical Evidence	2026-06	86,055	35	54	86	8

A.2 Parser Support and Query Sampling

The parser recognizes `section`, `subsection`, `subsubsection`, and `paragraph`, including starred forms, after masking comments and literal regions. Object labels must be unique, referenced by a supported command, and enclosed in a supported object environment. Supported environments are `figure`, `table`, `equation`, `align`, `gather`, `multline`, `eqnarray`, `algorithm`, `lstlisting`, `theorem`, `lemma`, `proposition`, `definition`, `corollary`, `remark`, and `example`; literal masking takes precedence. Supported reference commands are `\ref`, `\autoref`, `\cref`, `\Cref`, and `\eqref`. Multi-target commands such as `\cref{a,b}`, unresolved labels, and labels outside supported environments are excluded. Gold follows this declared parser support rather than all references a full LaTeX engine could resolve.

Ownership uses the beginning of the defining environment, not rendered float placement. Identifier permutations are seeded from the source hash; `H00` denotes the root in actual prompts and outputs. SE-RA candidates use nested heading pairs, two to four eligible objects, and a minimum gold path length of three headings including the root. Query-bank construction restricts eligible targets to definitions outside the positive scope, while the displayed task rules focus on the citation scope and nested exclusion. This construction condition is applied before prompt rendering and is not repeated in the task-facing examples.

Selection cycles through five distance bands bounded by 35K, 70K, 105K, and 150K characters, with at most 20 SE-RA queries per paper. Within a band, candidates are ordered by minimum path depth, mean path depth, target-set size, and distance, with anchor indices breaking ties. The stored bank contains 470 SE-RA queries. The positive-only construction yields 405 EN/EN-RA queries; evaluated subsets are reported separately in Appendix D.1.

A.3 Actual Source and Serialization Fragments

The following contiguous excerpt comes from the annotated source of arXiv:2603.28590. The snippet ends before the section does.

```
\section{Experiments}
\arxivbenchheadingid{H30}


In this section, we conduct experiments to answer the following
    research questions:
\vspace{-5pt}
\begin{itemize}[leftmargin=*]
  \vspace{-2pt}
```

```
  \item \textbf{RQ1:} When can CoT be reliably used to monitor the
    decision drivers of existing LLMs? Which types of decision-critical
     factors and model families exhibit lower monitorability?
  \vspace{-2pt}
  \item \textbf{RQ2:} To what extent does CoT remain monitorable under
    stress-test conditions (i.e., with direct concealment and monitor-
    aware evasion instructions)?
  \vspace{-2pt}
  \item \textbf{RQ3:} Why do LLMs fail to conceal decision-relevant
    information in their CoTs, and what strategies could enable
    successful concealment?
\end{itemize}
\vspace{-2pt}

\textbf{Evaluated LLMs.} We evaluate three open-source models (Qwen3-14
    B, Qwen3-235B-A22B~\citep{yang2025qwen3}, and GPT-OSS-120B~\citep{
    agarwal2025gpt}), and four closed-source models (GPT-5-mini, GPT-5~\
    citep{singh2025openai}, Claude Haiku 4.5, Gemini-3-Flash), which
    cover a broad spectrum of LLMs with different levels of
    capabilities. We run multiple trajectories per prompt. For the
    tasks that need LLM as a judge for the verification, and those
    monitoring tasks, we use Qwen3-32B. The monitor model is given
    exact decision-critical factors to look for, which ensures
    monitorability reflects the properties of the evaluated LLM a
```

The following XML excerpt retains complete heading records from that paper's actual Parse-Only serialization. The context header and code fence are omitted, and the XML comment marks the omitted records. Both the full serialization and this excerpt are well-formed XML. Heading IDs begin with `H`, object IDs with uppercase `O`, and the root ID `H00` ends in two zeros. The unabridged context is retained in the accompanying records.

```
<document>
  <root id="H00">
  </root>
  <heading id="H48" kind="section" parent="H00" title="Introduction">
    <defines target="O09" type="figure"/>
    <defines target="O10" type="figure"/>
    <cites target="O09"/>
    <cites target="O12"/>
  </heading>
  <heading id="H34" kind="section" parent="H00" title="Related Works">
  </heading>
  <heading id="H12" kind="section" parent="H00" title="">
  </heading>
  <heading id="H26" kind="subsection" parent="H12" title="Evaluation
    Design">
    <defines target="O12" type="table"/>
  </heading>
  <heading id="H07" kind="subsection" parent="H12" title="Evaluation
    Tasks">
  </heading>
  <heading id="H40" kind="subsubsection" parent="H07" title="Input
    Intervention">
  </heading>
  <heading id="H73" kind="subsubsection" parent="H07" title="Outcome
    Justification">
  </heading>
  <heading id="H15" kind="subsubsection" parent="H07" title="Solution
    Process">
  </heading>
  <heading id="H11" kind="subsection" parent="H12" title="Evaluation
    Metrics">
  </heading>
  <heading id="H30" kind="section" parent="H00" title="Experiments">
```

```
    <defines target="O02" type="figure"/>
  </heading>
  <heading id="H36" kind="subsection" parent="H30" title="
    Monitorability Across Factors and Model Families (RQ1)">
    <defines target="O07" type="table"/>
    <cites target="O10"/>
    <cites target="O03"/>
    <cites target="O08"/>
    <cites target="O02"/>
  </heading>
  <!-- Remaining headings omitted from this excerpt. -->
</document>
```

B Context Construction and Measured Lengths

B.1 Token Accounting

Table 7 separates context-block and full-prompt lengths under the Qwen2.5-14B-Instruct-1M reference tokenizer, without a model-specific chat template. Full prompts additionally contain instructions, a worked example, the query, and the output contract. The nominal 90K and 3K labels are shorthand, not fixed lengths. Source documents occupy approximately 70–102K tokens; Parse-Only contexts span 1,229–8,268 tokens and their full prompts span 3,014–10,119 tokens. Complete per-query and per-paper counts, together with prompt hashes, are retained in the accompanying CSV records.

Table 7: Actual token counts for SE-RA inputs under the fixed Qwen2.5-14B-Instruct-1M reference tokenizer. Context includes its textual wrappers; prompt additionally includes task instructions, query, worked example, and output contract. Statistics are over 470 query instances per configuration; no chat template is counted.

Configuration	Context min / median / max			Prompt min / median / max		
Parse-Only	1,229	4,319	8,268	3,014	6,115	10,119
Skeleton-12K	12,064	12,068	12,073	13,841	13,867	13,924
Skeleton-32K	32,064	32,068	32,073	33,841	33,867	33,924
Skeleton-90K	71,439	83,083	103,164	73,227	84,881	104,963
Original	71,277	82,881	102,986	73,063	84,679	104,775
Parse-Assist	74,309	89,480	108,034	76,095	91,272	109,837
Parse-Appended	73,351	87,911	106,360	75,137	89,718	108,163
OOD-background	71,345	82,891	103,030	73,133	84,689	104,833
Adversarial	76,325	88,058	108,115	78,127	89,854	109,903
BM25-32K	31,830	32,356	32,938	33,613	34,150	34,775

B.2 Construction Records and Background Examples

In Figure 4, blue indicates source structural markers, purple serialized relations, black original prose, and gray filler. Skeleton retains the structural spans and their source order while distributing filler within intervening regions. Length reduction thins filler toward the requested 12,000- or 32,000-token source budget. This approximately preserves relative structural positions, without asserting identical token offsets. Below is an actual Skeleton-12K excerpt from arXiv:2603.28590.

```
\section{Structural heading}
\arxivbenchheadingid{H30}
\arxivbenchobjectid{O02}
\begin{figure}
\label{fig:scaling_all_message}
\end{figure}
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
```

```
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
\subsection{Structural heading}
\arxivbenchheadingid{H36}
\ref{fig:bar_radar}
\ref{fig:scaling_action_only}
\ref{fig:scaling_cot_only}
\ref{fig:scaling_all_message}
\ref{fig:bar_radar}
\arxivbenchobjectid{O07}
\begin{table}
\label{tab:stress_test_tpr_cot_only}
\end{table}
The quick brown fox jumps over the lazy dog.
The quick br
```

OOD-background draws prose from eight biology source papers, removes their structural markers, and inserts it into the replacement regions. Adversarial inserts plain-text membership and ownership claims while preserving the markup-derived gold. The following records give the donor IDs, an actual replacement, and a misleading insertion concerning `fig:teaser`, whose reference is outside the positive scope.

```
OOD-background donor arXiv IDs:
2603.24626, 2604.11451, 2604.17786, 2604.14259, 2604.23773, 2604.16896,
     2211.11346, 2512.24427

Verbatim replacement unit (inserted into arXiv:2603.28590):
Our analytic approximation enables us to identify T cell
    characteristics that benefit most from Dendritic cell clustering,
    to result in an enhanced stimulation distribution.
Donor arXiv ID: 2604.17786

Verbatim adversarial insertion (same target paper):
In a similar vein, Experiments makes use of the figure labeled fig:
    teaser.

Insertion offset in the source: 7143
Object ID: O09
```

The adversarial builder requests about 5K added tokens; Table 7 reports the realized input lengths. This condition combines misleading content with added text and has no neutral-insertion control at the same added length.

B.3 Parse-Assist versus Parse-Appended

Parse-Assist places each heading's parsed relations beside the corresponding source region. **Parse-Appended** places the complete serialization after the original source. Table 8 uses the intersection of available queries across all four columns within each model. Both assisted conditions expose the same relations but differ in position, grouping, and formatting overhead; the comparison does not isolate traversal cost from presentation.

Table 8: SE-RA query accuracy (%) on the intersection of available queries across all four configurations for each model. Q explicitly reports the matched denominator.

Model	Q	Original	Parse-Assist	Parse-Appended	Parse-Only
Qwen3.8 27B	470	59.8	92.6	96.2	99.6
Qwen3.8 Flash-Next	102	42.2	85.3	96.1	100.0
Qwen3.6 27B	470	40.0	90.9	99.4	100.0
Qwen3.5 27B	398	20.1	73.4	97.0	99.7
Gemma-4 31B	470	16.6	83.6	93.0	99.1
Qwen3.6 35B-A3B	470	13.6	67.0	77.4	94.7
Qwen3.5 122B-A10B	470	10.9	61.9	79.8	97.0
Qwen-AgentWorld	470	10.0	25.1	31.9	96.6
Qwen3-Next 80B-A3B	470	3.2	33.2	54.3	93.6
Qwen3.5 35B-A3B	470	3.0	47.7	69.8	99.8
Seed-OSS 36B	470	2.6	23.6	43.2	99.8
GPT-OSS 120B	470	1.7	36.8	65.3	98.7
Gemma-4 26B-A4B	470	1.5	17.9	46.6	92.6
Gemma-4 E4B	470	1.1	8.7	7.0	82.8
Qwen3 30B-A3B	470	0.2	10.6	14.7	94.3
Qwen3.5 9B	418	1.0	18.9	43.5	97.4
Devstral-Small-2 24B	102	0.0	0.0	0.0	96.1
GPT-OSS 20B	470	0.0	3.8	17.9	80.6
Gemini 3.7 Flash	22	54.5	81.8	95.5	100.0
GPT-5.6 Terra	22	22.7	100.0	100.0	100.0
Qwen3.7 Plus	22	31.8	86.4	90.9	100.0
Claude Sonnet 5	22	31.8	95.5	100.0	100.0
Grok 4.5	22	50.0	90.9	100.0	100.0

C Actual Task Prompts and Model Responses

The five examples use arXiv:2305.06426 and the saved prompts for EN, SE, RA, EN-RA, and SE-RA. Shared instructions are printed once, followed by each task's specific instruction, query, output contract, answer, and gold output. Answer blocks are quoted verbatim from Qwen3.8 27B Original runs.

C.1 Shared Prompt Scaffold

The five task prompts share the following base instructions. They are shown once to avoid repeating the same context, heading, identifier, and reasoning instructions.

```
Every block of the context is introduced by a line starting with "
    Context - ". Everything you are given describes this one paper;
    none of it states the answer, so it has to be worked out from what
    you are given.

Headings come at four levels, and all four count: \section, \subsection,
     \subsubsection and \paragraph. A \paragraph is a heading here
    exactly as a \section is, so a path may be several levels deep and
    no level may be skipped when reading one off.

Definitions, used exactly as given:
  subtree(X)      - heading X together with all of its descendant
    headings, at every level, and all prose owned by any of them.
  cited_in(X)     - the set of objects targeted by a ref-like command
    (\ref, \autoref, \cref, \Cref, \eqref) appearing anywhere in
    subtree(X). Where the object is defined does not matter here; only
    where it is cited from.
  definition      - an object's definition is the LaTeX environment
    that contains its \label - the \begin{figure}, \begin{table}, \
    begin{equation}, \begin{align}, \begin{algorithm}, \begin{theorem},
     \begin{lemma}, \begin{definition} and so on. The definition is
    located where that environment BEGINS. Prose that merely mentions
```

```
    or discusses the object is not its definition, and neither is a \
    ref to it.
  defined_in(X)   - the objects whose own definition, as just defined,
    sits under a heading in subtree(X).
  heading path    - the chain of heading identifiers from the document
    root down to and including the heading that owns a definition, with
     EVERY intervening heading present. If a definition sits under a \
    paragraph whose parent is a \subsection whose parent is a \section,
     the path has four entries: the root, the section, the subsection
    and the paragraph. Do not stop at the nearest section or subsection.
     The root is H00. The object itself is never part of its own path.

Identifiers (H.. for headings, O.. for objects) are arbitrary labels.
    They do not encode nesting level and they are not in document order,
     so neither an object's own identifier nor its neighbours' tell you
     where it sits.


Work out the answer before writing it. Your working should not exceed
    about 6,000 tokens.
```

C.2 Task-Specific Prompt Blocks and Recorded Answers

C.2.1 EN

Query `2305.06426:H16:positive_only` on arXiv:2305.06426. The following gives the cleaned task-specific blocks; shared blocks are shown once above.

```
Task: for one paper, work out which objects satisfy a stated condition.
     Return only their object identifiers; no heading paths are
    required.

[Shared instructions are shown above.]

Paper: "Planning a Community Approach to Diabetes Care in Low- and
    Middle-Income Countries Using Optimization" (arXiv:2305.06426).

Anchor P = H16 (subsection), title path "CHW Provider Problem > Multi-
    Patient Problem".
There is no negative anchor in this control query.

The objects you are asked about are exactly these:
    cited_in(H16)

Give only the identifiers of those objects. Do not return heading paths.


Put your answer in a single <answer>...</answer> block with nothing
    after it. Inside it, list only the selected object identifiers,
    separated by " ; ". Do not include heading paths.
<answer>O01 ; O08 ; O05</answer>
List each object exactly once; order is not scored. If no objects
    qualify, write <answer></answer>.
Your working should not exceed about 6,000 tokens.
```

Recorded model answer. Qwen3.8 27B, Original. The answer block is quoted verbatim; the complete response is retained in the accompanying text file.

```
<answer>O30 ; O16 ; O38</answer>
```

Gold output.

```
<answer>O16 ; O38 ; O30</answer>
```

The final answer block is parseable. The order-insensitive scorer marks the required output as exact: Set exact = yes and ID recall = 3/3. Entry order is ignored by the scorer; reasoning outside the final answer block is not scored.

C.2.2 SE

Query `2305.06426:H16:H67` on arXiv:2305.06426. The following gives the cleaned task-specific blocks; shared blocks are shown once above.

```
Task: for one paper, work out which objects satisfy a stated condition.
     Return only their object identifiers; no heading paths are
    required.

[Shared instructions are shown above.]

Paper: "Planning a Community Approach to Diabetes Care in Low- and
    Middle-Income Countries Using Optimization" (arXiv:2305.06426).

Anchor P = H16 (subsection), title path "CHW Provider Problem > Multi-
    Patient Problem".
Anchor N = H67 (subsubsection), title path "CHW Provider Problem >
    Multi-Patient Problem > Lagrangian Relaxation.".
H67 lies inside subtree(H16).

The objects you are asked about are exactly these:
    cited_in(H16)  MINUS  cited_in(H67)

Give only the identifiers of those objects. Do not return heading paths.


Put your answer in a single <answer>...</answer> block with nothing
    after it. Inside it, list only the selected object identifiers,
    separated by " ; ". Do not include heading paths.
<answer>O01 ; O08 ; O05</answer>
List each object exactly once; order is not scored. If no objects
    qualify, write <answer></answer>.
Your working should not exceed about 6,000 tokens.
```

Recorded model answer. Qwen3.8 27B, Original. The answer block is quoted verbatim; the complete response is retained in the accompanying text file.

```
<answer>O30 ; O16 ; O38</answer>
```

Gold output.

```
<answer>O16 ; O38 ; O30</answer>
```

The final answer block is parseable. The order-insensitive scorer marks the required output as exact: Set exact = yes and ID recall = 3/3. Entry order is ignored by the scorer; reasoning outside the final answer block is not scored.

C.2.3 RA

Query `2305.06426:H16:H67` on arXiv:2305.06426. The following gives the cleaned task-specific blocks; shared blocks are shown once above.

```
Task: for one paper, give the chain of headings leading to the
    definition of each supplied object identifier. The target objects
    are already given; no citation-based selection is required.

[Shared instructions are shown above.]

Paper: "Planning a Community Approach to Diabetes Care in Low- and
    Middle-Income Countries Using Optimization" (arXiv:2305.06426).

Requested object identifiers: O16, O38, O30.

For each supplied object, find its definition and return the complete
    heading path from the document root to the heading owning that
    definition. Return exactly the supplied objects. No P/N filtering
    or reference-based selection is part of this task.

Put your answer in a single <answer>...</answer> block with nothing
    after it. Inside it, write one entry per object: the object
    identifier, a colon, then its heading path with " > " between
    headings. Separate entries with " ; ".
<answer>O01: H00 > H05 ; O08: H00 > H05 > H09</answer>
Each path runs root-first, starts at H00, and ends at the heading that
    owns that object's definition. List each object exactly once; the
    order of the entries is not scored.
Your working should not exceed about 6,000 tokens.
```

Recorded model answer. Qwen3.8 27B, Original. The answer block is quoted verbatim; the complete response is retained in the accompanying text file.

```
<answer>O16: H00 > H59 > H20 > H48 ; O38: H00 > H59 > H20 > H48 ; O30:
    H00 > H59 > H20 > H21</answer>
```

Gold output.

```
<answer>O16: H00 > H59 > H20 > H48 ; O38: H00 > H59 > H20 > H48 ; O30:
    H00 > H59 > H20 > H21</answer>
```

The final answer block is parseable. The order-insensitive scorer marks the required output as exact: Set exact = yes and ID recall = 3/3; ID+Path recall = 3/3. Entry order is ignored by the scorer; reasoning outside the final answer block is not scored.

C.2.4 EN-RA

Query `2305.06426:H16:positive_only` on arXiv:2305.06426. The following gives the cleaned task-specific blocks; shared blocks are shown once above.

```
Task: for one paper, work out which objects satisfy a stated condition,
     and for each of them give the chain of headings leading to where
    it is defined.

[Shared instructions are shown above.]

Paper: "Planning a Community Approach to Diabetes Care in Low- and
    Middle-Income Countries Using Optimization" (arXiv:2305.06426).

Anchor P = H16 (subsection), title path "CHW Provider Problem > Multi-
    Patient Problem".
There is no negative anchor in this control query.

The objects you are asked about are exactly these:
    cited_in(H16)
```

```
For each of those objects, give the heading path of the heading that
    owns its DEFINITION: find the environment that contains its \label
    anywhere in the paper, note which heading that environment sits
    under, and read off the chain of headings from the root down to it.

Put your answer in a single <answer>...</answer> block with nothing
    after it. Inside it, write one entry per object: the object
    identifier, a colon, then its heading path with " > " between
    headings. Separate entries with " ; ".
<answer>O01: H00 > H05 ; O08: H00 > H05 > H09</answer>
Each path runs root-first, starts at H00, and ends at the heading that
    owns that object's definition. List each object exactly once; the
    order of the entries is not scored.
Your working should not exceed about 6,000 tokens.
```

Recorded model answer. Qwen3.8 27B, Original. The answer block is quoted verbatim; the complete response is retained in the accompanying text file.

```
<answer>O30: H00 > H59 > H20 > H21 ; O16: H00 > H59 > H20 > H48 ; O38:
    H00 > H59 > H20 > H48</answer>
```

Gold output.

```
<answer>O16: H00 > H59 > H20 > H48 ; O38: H00 > H59 > H20 > H48 ; O30:
    H00 > H59 > H20 > H21</answer>
```

The final answer block is parseable. The order-insensitive scorer marks the required output as exact: Set exact = yes and ID recall = 3/3; ID+Path recall = 3/3. Entry order is ignored by the scorer; reasoning outside the final answer block is not scored.

C.2.5 SE-RA

Query `2305.06426:H16:H67` on arXiv:2305.06426. The following gives the cleaned task-specific blocks; shared blocks are shown once above.

```
Task: for one paper, work out which objects satisfy a stated condition,
     and for each of them give the chain of headings leading to where
    it is defined.

[Shared instructions are shown above.]

Paper: "Planning a Community Approach to Diabetes Care in Low- and
    Middle-Income Countries Using Optimization" (arXiv:2305.06426).

Anchor P = H16 (subsection), title path "CHW Provider Problem > Multi-
    Patient Problem".
Anchor N = H67 (subsubsection), title path "CHW Provider Problem >
    Multi-Patient Problem > Lagrangian Relaxation.".
H67 lies inside subtree(H16).

The objects you are asked about are exactly these:
    cited_in(H16)  MINUS  cited_in(H67)

For each of those objects, give the heading path of the heading that
    owns its DEFINITION: find the environment that contains its \label
    anywhere in the paper, note which heading that environment sits
    under, and read off the chain of headings from the root down to it.

Put your answer in a single <answer>...</answer> block with nothing
    after it. Inside it, write one entry per object: the object
    identifier, a colon, then its heading path with " > " between
    headings. Separate entries with " ; ".
<answer>O01: H00 > H05 ; O08: H00 > H05 > H09</answer>
```

```
Each path runs root-first, starts at H00, and ends at the heading that
    owns that object's definition. List each object exactly once; the
    order of the entries is not scored.
Your working should not exceed about 6,000 tokens.
```

Recorded model answer. Qwen3.8 27B, Original. The answer block is quoted verbatim; the complete response is retained in the accompanying text file.

```
<answer>O30: H00 > H59 > H20 > H21 ; O16: H00 > H59 > H20 > H48 ; O38:
    H00 > H59 > H20 > H48</answer>
```

Gold output.

```
<answer>O16: H00 > H59 > H20 > H48 ; O38: H00 > H59 > H20 > H48 ; O30:
    H00 > H59 > H20 > H21</answer>
```

The final answer block is parseable. The order-insensitive scorer marks the required output as exact: Set exact = yes and ID recall = 3/3; ID+Path recall = 3/3. Entry order is ignored by the scorer; reasoning outside the final answer block is not scored.

D Evaluation Coverage and Supplementary Metrics

D.1 Record Selection

The appendix is restricted to the 35 models in the main paper, comprising 21 locally deployed models and 14 API models. Thinking-enabled runs are additional settings of these same models. All rows are matched to question IDs in the fixed 64-paper manifest. Most local runs cover 470 queries; the API heatmap uses 50 queries from five papers. Absent records are not fabricated, while recorded unsuccessful requests remain in the denominator. The appendix uses the saved output snapshot rather than claiming to reconstruct first generation attempts. Per-model coverage and every matched query ID remain in the accompanying score records.

Some main-table SE-RA entries use the 102-query subset, while the supplementary tables below report all currently available bank-matched records. Their scores may therefore differ when coverage has expanded.

The 102-query composition analyses use the first 102 manifest-ordered eligible queries of each form. EN/EN-RA and SE/RA/SE-RA have different candidate banks. Matched within-form context contrasts should therefore be distinguished from comparisons between forms.

D.2 Metric Denominators and Normalization

Set accuracy is the proportion of exact gold object sets, ignoring entry order and path correctness. **ID recall** divides recovered gold-object occurrences by all gold-object occurrences across queries. **ID+Path recall** uses the same denominator but requires both the ID and its complete root-to-owner path. Both recalls are micro-averaged. Query accuracy additionally requires the exact set and every correct path for RA, EN-RA, and SE-RA; for EN and SE it equals Set accuracy.

All metrics use the same validity gate. Unavailable, empty, truncated, and unrecoverable outputs contribute zero numerator credit, retaining their queries and gold objects in the denominators. Duplicate or unknown object IDs invalidate an answer. Deterministic normalization handles supported ID formatting, separator variants, escaped formatting, and legacy JSON records; it does not supply missing objects or repair wrong paths. Strict and relaxed parsing counts are in Table 11, and failure proportions are in Appendix E.

D.3 Full-Workflow Results under Three Supplementary Metrics

For compactness, we retain the Original SE-RA breakdown here. It places Set accuracy, ID recall, and ID+Path recall beside query accuracy, with their query and gold-object denominators. The same per-query metrics for other configurations remain in the accompanying records.

Table 9: Original SE-RA results (%). Set scores exact object sets; ID recall and ID+Path recall are micro-averaged over gold-object occurrences. All recorded queries, including invalid answers, remain in the denominators. Q denotes queries and G gold-object occurrences.

Model	Q	G	Query	Set	ID recall	ID+Path
Qwen3.8 27B	470	1289	59.8	69.1	90.8	84.3
Qwen3.8 Flash-Next	470	1289	47.4	64.3	87.7	73.0
Qwen3.6 27B	470	1289	40.0	54.0	84.1	69.9
Qwen3.5 27B	470	1289	22.3	31.1	51.3	41.4
Muse Glimmer 30B	470	1289	18.9	23.8	48.1	40.4
Gemma-4 31B	470	1289	16.6	26.6	63.0	50.5
Qwen3.6 35B-A3B	470	1289	13.6	24.5	66.2	45.2
Qwen3.5 122B-A10B	470	1289	10.9	14.9	44.7	29.2
Qwen-AgentWorld	470	1289	10.0	12.6	51.8	31.5
Qwen3-Next 80B-A3B	470	1289	3.2	9.6	33.2	17.7
Qwen3.5 35B-A3B	470	1289	3.0	6.4	17.2	11.5
Seed-OSS 36B	470	1289	2.6	6.0	29.1	19.3
GPT-OSS 120B	470	1289	1.7	5.1	15.4	7.2
Gemma-4 26B-A4B	470	1289	1.5	2.8	16.4	9.8
Gemma-4 E4B	470	1289	1.1	2.1	20.5	5.5
Nemotron-3 Super	102	300	1.0	3.9	4.0	1.7
Qwen3 30B-A3B	470	1289	0.2	1.5	8.8	2.0
Qwen3.5 9B	418	1151	1.0	3.1	10.9	5.7
Devstral-Small-2 24B	470	1289	0.0	0.0	11.1	1.5
GPT-OSS 20B	470	1289	0.0	0.2	0.6	0.2
Mistral-Small-3.2 24B	115	334	0.0	0.0	28.4	1.5
Gemini 3.8 Flash	50	159	74.0	78.0	93.7	89.9
Gemini 3.7 Flash	50	159	50.0	68.0	87.4	77.4
Gemini 3.1 Pro	50	159	70.0	72.0	93.1	91.8
GPT-6 Astra	50	159	44.0	54.0	83.6	67.3
GPT-6 Sol	50	159	32.0	44.0	79.9	66.0
GPT-5.6 Terra	50	159	24.0	40.0	82.4	61.6
Qwen3.7 Plus	50	159	26.0	40.0	68.6	57.9
Claude Sonnet 5	50	159	36.0	42.0	72.3	58.5
Claude Opus 5	50	159	0.0	4.0	39.6	23.3
DeepSeek V4.1 Flash	50	159	64.0	74.0	93.1	86.8
GLM-5.3 Flash	50	159	72.0	76.0	95.0	92.5
Grok 4.5	50	159	38.0	52.0	77.4	62.9
Hunyuan 3	50	159	50.0	60.0	84.9	78.0
Kimi K2.5	50	159	6.0	8.0	53.5	37.7

D.4 Basic and Composite Task Results

Table 10 contains all available EN, SE, RA, and EN-RA records under Original and Skeleton-32K. Coverage is shown separately for each cell rather than assuming every run contains 102 queries. Path metrics are inapplicable for EN and SE.

Table 10: Supplementary task metrics (%) using available saved queries within the bank. Q reports coverage separately for each task and configuration; comparisons with the 102-query main-table subset must use that subset, not all rows pooled here. ID and ID+Path are micro-averaged recalls. Paths are not requested for EN/SE.

Model	Task	Context	Q	Query	Set	ID	ID+Path
Qwen3.8 27B	EN	Original	103	79.6	79.6	91.0	–
Qwen3.8 27B	EN	Skeleton-32K	103	99.0	99.0	100.0	–
Qwen3.8 27B	SE	Original	102	65.7	65.7	91.7	–
Qwen3.8 27B	SE	Skeleton-32K	102	94.1	94.1	99.3	–
Qwen3.8 27B	RA	Original	102	93.1	100.0	100.0	96.3
Qwen3.8 27B	RA	Skeleton-32K	102	92.2	100.0	100.0	96.7
Qwen3.8 27B	EN-RA	Original	103	71.8	79.6	91.7	87.0
Qwen3.8 27B	EN-RA	Skeleton-32K	405	94.1	98.8	99.9	96.7
Qwen3.8 Flash-Next	EN	Original	103	74.8	74.8	89.5	–
Qwen3.8 Flash-Next	EN	Skeleton-32K	103	100.0	100.0	100.0	–
Qwen3.8 Flash-Next	SE	Original	102	52.9	52.9	83.0	–

Table 10 continued

Model	Task	Context	Q	Query	Set	ID	ID+Path
Qwen3.8 Flash-Next	SE	Skeleton-32K	102	99.0	99.0	99.7	–
Qwen3.8 Flash-Next	RA	Original	102	81.4	100.0	100.0	91.3
Qwen3.8 Flash-Next	RA	Skeleton-32K	102	95.1	100.0	100.0	97.0
Qwen3.8 Flash-Next	EN-RA	Original	103	54.4	68.0	89.5	80.5
Qwen3.8 Flash-Next	EN-RA	Skeleton-32K	103	91.3	99.0	100.0	94.2
Qwen3.6 27B	EN	Original	103	57.3	57.3	78.7	–
Qwen3.6 27B	EN	Skeleton-32K	103	99.0	99.0	99.6	–
Qwen3.6 27B	SE	Original	102	44.1	44.1	77.3	–
Qwen3.6 27B	SE	Skeleton-32K	102	100.0	100.0	100.0	–
Qwen3.6 27B	RA	Original	102	86.3	100.0	100.0	91.3
Qwen3.6 27B	RA	Skeleton-32K	102	94.1	100.0	100.0	96.0
Qwen3.6 27B	EN-RA	Original	103	46.6	62.1	85.2	75.5
Qwen3.6 27B	EN-RA	Skeleton-32K	103	87.4	99.0	99.6	93.1
Qwen3.5 27B	EN	Original	103	57.3	57.3	76.5	–
Qwen3.5 27B	EN	Skeleton-32K	103	95.1	95.1	99.3	–
Qwen3.5 27B	SE	Original	102	25.5	25.5	49.3	–
Qwen3.5 27B	SE	Skeleton-32K	102	88.2	88.2	97.3	–
Qwen3.5 27B	RA	Original	102	73.5	100.0	100.0	87.3
Qwen3.5 27B	RA	Skeleton-32K	102	93.1	99.0	99.3	96.0
Qwen3.5 27B	EN-RA	Original	103	40.8	52.4	75.1	66.1
Qwen3.5 27B	EN-RA	Skeleton-32K	103	86.4	95.1	98.2	93.9
Muse Glimmer 30B	EN	Original	103	34.0	34.0	52.7	–
Muse Glimmer 30B	EN	Skeleton-32K	103	95.1	95.1	97.1	–
Muse Glimmer 30B	SE	Original	102	16.7	16.7	33.0	–
Muse Glimmer 30B	SE	Skeleton-32K	102	91.2	91.2	97.7	–
Muse Glimmer 30B	RA	Original	102	76.5	100.0	100.0	89.3
Muse Glimmer 30B	RA	Skeleton-32K	102	94.1	100.0	100.0	98.0
Muse Glimmer 30B	EN-RA	Original	103	25.2	36.9	58.5	49.8
Muse Glimmer 30B	EN-RA	Skeleton-32K	103	90.3	97.1	99.3	95.7
Gemma-4 31B	EN	Original	103	51.5	51.5	78.0	–
Gemma-4 31B	EN	Skeleton-32K	103	94.2	94.2	98.2	–
Gemma-4 31B	SE	Original	102	21.6	21.6	58.7	–
Gemma-4 31B	SE	Skeleton-32K	102	95.1	95.1	99.3	–
Gemma-4 31B	RA	Original	102	75.5	100.0	100.0	87.7
Gemma-4 31B	RA	Skeleton-32K	102	96.1	100.0	100.0	98.3
Gemma-4 31B	EN-RA	Original	103	29.1	46.6	78.7	63.5
Gemma-4 31B	EN-RA	Skeleton-32K	103	71.8	93.2	96.0	81.9
Qwen3.6 35B-A3B	EN	Original	103	41.7	41.7	72.9	–
Qwen3.6 35B-A3B	EN	Skeleton-32K	103	91.3	91.3	97.8	–
Qwen3.6 35B-A3B	SE	Original	102	24.5	24.5	60.7	–
Qwen3.6 35B-A3B	SE	Skeleton-32K	102	80.4	80.4	93.7	–
Qwen3.6 35B-A3B	RA	Original	102	52.0	100.0	100.0	70.0
Qwen3.6 35B-A3B	RA	Skeleton-32K	102	40.2	100.0	100.0	62.3
Qwen3.6 35B-A3B	EN-RA	Original	103	16.5	30.1	69.3	46.6
Qwen3.6 35B-A3B	EN-RA	Skeleton-32K	103	52.4	88.3	95.3	67.5
Qwen3.5 122B-A10B	EN	Original	103	26.2	26.2	53.1	–
Qwen3.5 122B-A10B	EN	Skeleton-32K	103	98.1	98.1	98.9	–
Qwen3.5 122B-A10B	SE	Original	102	10.8	10.8	28.0	–
Qwen3.5 122B-A10B	SE	Skeleton-32K	102	93.1	93.1	97.3	–
Qwen3.5 122B-A10B	RA	Original	102	67.6	98.0	98.3	80.7
Qwen3.5 122B-A10B	RA	Skeleton-32K	102	67.6	99.0	99.0	83.0
Qwen3.5 122B-A10B	EN-RA	Original	103	20.4	31.1	70.0	51.6
Qwen3.5 122B-A10B	EN-RA	Skeleton-32K	103	58.3	98.1	100.0	73.6
Qwen-AgentWorld	EN	Original	103	49.5	49.5	56.7	–
Qwen-AgentWorld	EN	Skeleton-32K	103	96.1	96.1	96.4	–
Qwen-AgentWorld	SE	Original	102	12.7	12.7	21.0	–
Qwen-AgentWorld	SE	Skeleton-32K	102	85.3	85.3	85.7	–
Qwen-AgentWorld	RA	Original	102	30.4	95.1	98.0	42.0
Qwen-AgentWorld	RA	Skeleton-32K	102	36.3	100.0	100.0	48.0
Qwen-AgentWorld	EN-RA	Original	103	10.7	13.6	71.8	38.6
Qwen-AgentWorld	EN-RA	Skeleton-32K	233	44.2	44.6	73.8	49.8
Qwen3-Next 80B-A3B	EN	Original	103	20.4	20.4	41.2	–
Qwen3-Next 80B-A3B	EN	Skeleton-32K	103	88.3	88.3	97.5	–
Qwen3-Next 80B-A3B	SE	Original	102	2.9	2.9	21.3	–
Qwen3-Next 80B-A3B	SE	Skeleton-32K	102	49.0	49.0	79.7	–
Qwen3-Next 80B-A3B	RA	Original	102	30.4	97.1	96.7	57.3
Qwen3-Next 80B-A3B	RA	Skeleton-32K	102	46.1	100.0	100.0	68.0

Table 10 continued

Model	Task	Context	Q	Query	Set	ID	ID+Path
Qwen3-Next 80B-A3B	EN-RA	Original	103	7.8	24.3	50.5	23.5
Qwen3-Next 80B-A3B	EN-RA	Skeleton-32K	103	36.9	84.5	91.7	53.4
Qwen3.5 35B-A3B	EN	Original	103	17.5	17.5	30.3	–
Qwen3.5 35B-A3B	EN	Skeleton-32K	103	77.7	77.7	91.7	–
Qwen3.5 35B-A3B	SE	Original	102	5.9	5.9	19.0	–
Qwen3.5 35B-A3B	SE	Skeleton-32K	102	68.6	68.6	83.7	–
Qwen3.5 35B-A3B	RA	Original	102	44.1	96.1	96.0	61.7
Qwen3.5 35B-A3B	RA	Skeleton-32K	102	53.9	99.0	98.7	66.3
Qwen3.5 35B-A3B	EN-RA	Original	103	10.7	15.5	29.2	19.9
Qwen3.5 35B-A3B	EN-RA	Skeleton-32K	103	50.5	72.8	91.3	67.1
Seed-OSS 36B	EN	Original	103	42.7	42.7	58.8	–
Seed-OSS 36B	EN	Skeleton-32K	103	100.0	100.0	100.0	–
Seed-OSS 36B	SE	Original	102	6.9	6.9	25.0	–
Seed-OSS 36B	SE	Skeleton-32K	102	86.3	86.3	93.3	–
Seed-OSS 36B	RA	Original	102	60.8	97.1	98.7	75.0
Seed-OSS 36B	RA	Skeleton-32K	102	76.5	100.0	100.0	90.0
Seed-OSS 36B	EN-RA	Original	103	18.4	31.1	50.9	34.3
Seed-OSS 36B	EN-RA	Skeleton-32K	103	60.2	95.1	98.6	79.1
GPT-OSS 120B	EN	Original	103	23.3	23.3	37.9	–
GPT-OSS 120B	EN	Skeleton-32K	103	97.1	97.1	98.6	–
GPT-OSS 120B	SE	Original	102	6.9	6.9	22.7	–
GPT-OSS 120B	SE	Skeleton-32K	102	62.7	62.7	85.0	–
GPT-OSS 120B	RA	Original	102	38.2	94.1	94.0	63.0
GPT-OSS 120B	RA	Skeleton-32K	102	64.7	92.2	93.0	79.7
GPT-OSS 120B	EN-RA	Original	103	6.8	19.4	29.2	15.2
GPT-OSS 120B	EN-RA	Skeleton-32K	103	45.6	81.6	88.4	63.5
Gemma-4 26B-A4B	EN	Original	103	12.6	12.6	28.9	–
Gemma-4 26B-A4B	EN	Skeleton-32K	103	52.4	52.4	69.0	–
Gemma-4 26B-A4B	SE	Original	102	2.0	2.0	11.3	–
Gemma-4 26B-A4B	SE	Skeleton-32K	102	17.6	17.6	44.0	–
Gemma-4 26B-A4B	RA	Original	102	47.1	99.0	99.3	70.7
Gemma-4 26B-A4B	RA	Skeleton-32K	102	36.3	100.0	100.0	56.0
Gemma-4 26B-A4B	EN-RA	Original	103	2.9	9.7	33.9	14.8
Gemma-4 26B-A4B	EN-RA	Skeleton-32K	103	13.6	42.7	72.9	32.9
Gemma-4 E4B	EN	Original	103	9.7	9.7	36.1	–
Gemma-4 E4B	EN	Skeleton-32K	103	37.9	37.9	58.5	–
Gemma-4 E4B	SE	Original	102	1.0	1.0	15.3	–
Gemma-4 E4B	SE	Skeleton-32K	102	2.9	2.9	23.0	–
Gemma-4 E4B	RA	Original	102	6.9	98.0	98.7	23.7
Gemma-4 E4B	RA	Skeleton-32K	102	0.0	99.0	99.3	7.0
Gemma-4 E4B	EN-RA	Original	103	1.9	9.7	37.9	10.1
Gemma-4 E4B	EN-RA	Skeleton-32K	103	2.9	43.7	64.6	7.9
Nemotron-3 Super	EN	Original	103	30.1	30.1	42.6	–
Nemotron-3 Super	EN	Skeleton-32K	103	77.7	77.7	78.0	–
Nemotron-3 Super	SE	Original	102	4.9	4.9	5.3	–
Nemotron-3 Super	SE	Skeleton-32K	102	30.4	30.4	42.7	–
Nemotron-3 Super	RA	Original	102	41.2	92.2	92.7	60.7
Nemotron-3 Super	RA	Skeleton-32K	102	70.6	97.1	97.0	83.7
Nemotron-3 Super	EN-RA	Original	103	5.8	18.4	20.9	11.6
Nemotron-3 Super	EN-RA	Skeleton-32K	103	25.2	40.8	39.0	27.8
Qwen3 30B-A3B	EN	Original	103	1.0	1.0	12.3	–
Qwen3 30B-A3B	EN	Skeleton-32K	103	48.5	48.5	65.7	–
Qwen3 30B-A3B	SE	Original	102	0.0	0.0	2.0	–
Qwen3 30B-A3B	SE	Skeleton-32K	102	13.7	13.7	27.3	–
Qwen3 30B-A3B	RA	Original	102	25.5	99.0	99.0	40.7
Qwen3 30B-A3B	RA	Skeleton-32K	102	15.7	92.2	95.3	31.7
Qwen3 30B-A3B	EN-RA	Original	103	1.0	4.9	23.8	6.9
Qwen3 30B-A3B	EN-RA	Skeleton-32K	103	12.6	55.3	71.8	24.5
Qwen3.5 9B	EN	Original	103	14.6	14.6	31.4	–
Qwen3.5 9B	EN	Skeleton-32K	103	47.6	47.6	77.6	–
Qwen3.5 9B	SE	Original	102	3.9	3.9	9.3	–
Qwen3.5 9B	SE	Skeleton-32K	102	39.2	39.2	69.7	–
Qwen3.5 9B	RA	Original	102	40.2	96.1	96.7	57.7
Qwen3.5 9B	RA	Skeleton-32K	102	17.6	93.1	93.3	33.0
Qwen3.5 9B	EN-RA	Original	103	5.8	14.6	30.3	16.6
Qwen3.5 9B	EN-RA	Skeleton-32K	103	12.6	55.3	77.6	37.5
Devstral-Small-2 24B	EN	Original	103	0.0	0.0	16.6	–

Table 10 continued

Model	Task	Context	Q	Query	Set	ID	ID+Path
Devstral-Small-2 24B	EN	Skeleton-32K	103	78.6	78.6	86.3	–
Devstral-Small-2 24B	SE	Original	102	0.0	0.0	12.3	–
Devstral-Small-2 24B	SE	Skeleton-32K	102	14.7	14.7	32.0	–
Devstral-Small-2 24B	RA	Original	102	4.9	95.1	98.0	17.7
Devstral-Small-2 24B	RA	Skeleton-32K	102	42.2	100.0	100.0	64.7
Devstral-Small-2 24B	EN-RA	Original	103	0.0	0.0	13.4	2.9
Devstral-Small-2 24B	EN-RA	Skeleton-32K	103	19.4	55.3	70.8	31.8
GPT-OSS 20B	EN	Original	103	2.9	2.9	4.0	–
GPT-OSS 20B	EN	Skeleton-32K	103	40.8	40.8	40.4	–
GPT-OSS 20B	SE	Original	102	0.0	0.0	0.0	–
GPT-OSS 20B	SE	Skeleton-32K	102	2.9	2.9	4.7	–
GPT-OSS 20B	RA	Original	102	9.8	33.3	28.7	12.3
GPT-OSS 20B	RA	Skeleton-32K	102	8.8	35.3	33.0	14.0
GPT-OSS 20B	EN-RA	Original	103	0.0	1.9	3.6	1.4
GPT-OSS 20B	EN-RA	Skeleton-32K	103	1.0	17.5	17.3	2.9
Mistral-Small-3.2 24B	EN	Original	103	0.0	0.0	26.0	–
Mistral-Small-3.2 24B	EN	Skeleton-32K	103	31.1	31.1	43.0	–
Mistral-Small-3.2 24B	SE	Original	102	0.0	0.0	20.3	–
Mistral-Small-3.2 24B	SE	Skeleton-32K	102	8.8	8.8	15.7	–
Mistral-Small-3.2 24B	RA	Original	102	5.9	67.6	83.3	13.7
Mistral-Small-3.2 24B	RA	Skeleton-32K	102	34.3	99.0	100.0	57.0
Mistral-Small-3.2 24B	EN-RA	Original	103	1.0	1.9	37.2	3.2
Mistral-Small-3.2 24B	EN-RA	Skeleton-32K	103	5.8	40.8	72.2	21.7
Gemini 3.8 Flash	EN	Original	30	86.7	86.7	98.8	–
Gemini 3.8 Flash	EN	Skeleton-32K	30	100.0	100.0	100.0	–
Gemini 3.8 Flash	SE	Original	50	78.0	78.0	93.7	–
Gemini 3.8 Flash	SE	Skeleton-32K	50	100.0	100.0	100.0	–
Gemini 3.8 Flash	RA	Original	50	96.0	98.0	97.5	96.2
Gemini 3.8 Flash	RA	Skeleton-32K	50	100.0	100.0	100.0	100.0
Gemini 3.8 Flash	EN-RA	Original	30	86.7	86.7	98.8	98.8
Gemini 3.8 Flash	EN-RA	Skeleton-32K	30	100.0	100.0	100.0	100.0
Claude Sonnet 5	EN	Original	30	66.7	66.7	94.1	–
Claude Sonnet 5	EN	Skeleton-32K	30	100.0	100.0	100.0	–
Claude Sonnet 5	SE	Original	50	60.0	60.0	85.5	–
Claude Sonnet 5	SE	Skeleton-32K	50	92.0	92.0	97.5	–
Claude Sonnet 5	RA	Original	50	80.0	100.0	100.0	88.1
Claude Sonnet 5	RA	Skeleton-32K	50	90.0	100.0	100.0	94.3
Claude Sonnet 5	EN-RA	Original	30	66.7	73.3	94.1	85.9
Claude Sonnet 5	EN-RA	Skeleton-32K	30	100.0	100.0	100.0	100.0

E Generation Statistics and Observed Failure Modes

E.1 Generation Lengths and Output Validity

Table 11 summarizes Original SE-RA generation statistics using a nominal 16K budget. Server-reported tokens use each model's tokenizer and differ from the reference input counts. Measured medians and actual output-limit overrides are retained in `generation_statistics.csv`.

Table 11: Original SE-RA generation statistics for selected models with a nominal 16K budget. n counts available server usage records. Actual medians and request limits are retained in the accompanying data. Strict and relaxed counts use complete, scorable saved responses.

Model	n	Median	Nominal limit	Strict	Relaxed
Qwen3.8 27B	470	3,464	16K	460	0
Qwen3.8 Flash-Next	469	3,028	16K	451	0
Qwen3.6 27B	470	4,223	16K	460	0
Qwen3.5 27B	470	9,596	16K	244	34
Muse Glimmer 30B	470	2,044	16K	382	0
Gemma-4 31B	470	1,286	16K	441	0
Qwen3.6 35B-A3B	470	5,536	16K	411	0
Qwen3.5 122B-A10B	470	9,953	16K	258	7
Qwen-AgentWorld	470	981	16K	370	29

Table 11 continued

Model	n	Median	Nominal limit	Strict	Relaxed
Qwen3-Next 80B-A3B	461	10,017	16K	274	84
Seed-OSS 36B	470	8,132	16K	417	0
GPT-OSS 120B	470	12,080	16K	152	1
Gemma-4 E4B	470	2,910	16K	327	4
Devstral-Small-2 24B	470	626	16K	413	2
Mistral-Small-3.2 24B	115	1,310	16K	82	2
Gemini 3.8 Flash	50	13,390	16K	50	0
GPT-6 Astra	50	431	16K	50	0
GPT-6 Sol	50	800	16K	47	0
GPT-5.6 Terra	50	1,357	16K	50	0
Qwen3.7 Plus	50	5,901	16K	49	0
Claude Sonnet 5	50	5,518	16K	46	1
Claude Opus 5	50	1,858	16K	43	0
DeepSeek V4.1 Flash	50	14,068	16K	50	0
GLM-5.3 Flash	50	8,096	16K	50	0
Grok 4.5	47	6,437	16K	46	0
Hunyuan 3	50	9,473	16K	40	10
Kimi K2.5	45	5,736	16K	42	2

Table 12 partitions recorded queries into six disjoint outcomes. “Wrong” is a complete, parseable answer failing the task metric; “unavailable” is a failed request; “truncated” follows the length finish reason; “empty” has no response text; and “unparseable” covers remaining invalid answers. Truncated responses receive no credit even if a partial answer block is recoverable.

Table 12: Original SE-RA outcome distribution. Cells are percentages with raw counts as subscripts; the six categories are disjoint and sum to the query count Q. Truncation is determined by the recorded finish reason, even when a partial answer is recoverable.

Model	Q	Correct	Wrong	Unavailable	Truncated	Empty	Unparseable
Qwen3.8 27B	470	59.8_{281}	38.1_{179}	0.0_{0}	0.0_{0}	0.0_{0}	2.1_{10}
Qwen3.8 Flash-Next	470	47.4_{223}	48.5_{228}	0.2_{1}	0.0_{0}	0.0_{0}	3.8_{18}
Qwen3.6 27B	470	40.0_{188}	57.9_{272}	0.0_{0}	0.0_{0}	0.0_{0}	2.1_{10}
Qwen3.5 27B	470	22.3_{105}	36.8_{173}	0.0_{0}	40.4_{190}	0.0_{0}	0.4_{2}
Muse Glimmer 30B	470	18.9_{89}	62.3_{293}	0.0_{0}	0.0_{0}	0.0_{0}	18.7_{88}
Gemma-4 31B	470	16.6_{78}	77.2_{363}	0.0_{0}	0.4_{2}	0.0_{0}	5.7_{27}
Qwen3.6 35B-A3B	470	13.6_{64}	73.8_{347}	0.0_{0}	0.2_{1}	0.0_{0}	12.3_{58}
Qwen3.5 122B-A10B	470	10.9_{51}	45.5_{214}	0.0_{0}	40.6_{191}	0.0_{0}	3.0_{14}
Qwen-AgentWorld	470	10.0_{47}	74.9_{352}	0.0_{0}	14.7_{69}	0.0_{0}	0.4_{2}
Qwen3-Next 80B-A3B	470	3.2_{15}	73.0_{343}	1.9_{9}	6.6_{31}	0.0_{0}	15.3_{72}
Qwen3.5 35B-A3B	470	3.0_{14}	20.9_{98}	0.0_{0}	75.5_{355}	0.0_{0}	0.6_{3}
Seed-OSS 36B	470	2.6_{12}	86.2_{405}	0.0_{0}	1.5_{7}	0.0_{0}	9.8_{46}
GPT-OSS 120B	470	1.7_{8}	30.9_{145}	0.0_{0}	40.2_{189}	0.0_{0}	27.2_{128}
Gemma-4 26B-A4B	470	1.5_{7}	32.8_{154}	0.0_{0}	61.9_{291}	0.0_{0}	3.8_{18}
Gemma-4 E4B	470	1.1_{5}	69.4_{326}	0.0_{0}	2.8_{13}	0.0_{0}	26.8_{126}
Nemotron-3 Super	102	1.0_{1}	5.9_{6}	0.0_{0}	92.2_{94}	0.0_{0}	1.0_{1}
Qwen3 30B-A3B	470	0.2_{1}	31.9_{150}	0.0_{0}	62.3_{293}	0.0_{0}	5.5_{26}
Qwen3.5 9B	418	1.0_{4}	18.9_{79}	0.0_{0}	78.0_{326}	0.0_{0}	2.2_{9}
Devstral-Small-2 24B	470	0.0_{0}	88.3_{415}	0.0_{0}	1.5_{7}	0.0_{0}	10.2_{48}
Mistral-Small-3.2 24B	115	0.0_{0}	73.0_{84}	0.0_{0}	11.3_{13}	0.0_{0}	15.7_{18}
Gemini 3.8 Flash	50	74.0_{37}	26.0_{13}	0.0_{0}	0.0_{0}	0.0_{0}	0.0_{0}
Gemini 3.7 Flash	50	50.0_{25}	50.0_{25}	0.0_{0}	0.0_{0}	0.0_{0}	0.0_{0}
Gemini 3.1 Pro	50	70.0_{35}	30.0_{15}	0.0_{0}	0.0_{0}	0.0_{0}	0.0_{0}
GPT-6 Astra	50	44.0_{22}	56.0_{28}	0.0_{0}	0.0_{0}	0.0_{0}	0.0_{0}
GPT-6 Sol	50	32.0_{16}	62.0_{31}	0.0_{0}	0.0_{0}	0.0_{0}	6.0_{3}
GPT-5.6 Terra	50	24.0_{12}	76.0_{38}	0.0_{0}	0.0_{0}	0.0_{0}	0.0_{0}
Qwen3.7 Plus	50	26.0_{13}	72.0_{36}	0.0_{0}	0.0_{0}	0.0_{0}	2.0_{1}
Claude Sonnet 5	50	36.0_{18}	58.0_{29}	0.0_{0}	0.0_{0}	0.0_{0}	6.0_{3}
Claude Opus 5	50	0.0_{0}	86.0_{43}	0.0_{0}	0.0_{0}	4.0_{2}	10.0_{5}
DeepSeek V4.1 Flash	50	64.0_{32}	36.0_{18}	0.0_{0}	0.0_{0}	0.0_{0}	0.0_{0}
GLM-5.3 Flash	50	72.0_{36}	28.0_{14}	0.0_{0}	0.0_{0}	0.0_{0}	0.0_{0}
Grok 4.5	50	38.0_{19}	54.0_{27}	6.0_{3}	0.0_{0}	0.0_{0}	2.0_{1}
Hunyuan 3	50	50.0_{25}	50.0_{25}	0.0_{0}	0.0_{0}	0.0_{0}	0.0_{0}
Kimi K2.5	50	6.0_{3}	82.0_{41}	10.0_{5}	0.0_{0}	0.0_{0}	2.0_{1}
GPT-OSS 20B	470	0.0_{0}	1.5_{7}	1.5_{7}	93.6_{440}	0.0_{0}	3.4_{16}

E.2 Verbatim Failure Examples and Scorer Records

The following cases show actual answer blocks alongside concise scorer summaries, identified by model, paper, and query. Complete responses and machine-readable records are retained in the example files.

E.2.1 Extra object despite correct paths for gold objects

Qwen3.8 27B, Original, arXiv:2411.09355, query `task7:2411.09355:H59:H32`.

```
<answer>O25: H00 > H27 > H13 ; O26: H00 > H39 > H55 > H60 ; O15: H00 >
    H39 > H12</answer>
```

Scorer analysis. The summary below reports observable errors and metric outcomes; it does not describe the model's internal reasoning.

```
Omissions: none.
Intrusions: O25.
Set exact: no; query exact: no.
ID recall: 2/2; ID+Path recall: 2/2.
```

E.2.2 Correct object set with an incorrect ancestor

Qwen3.8 27B, Original, arXiv:2603.28590, query `task7:2603.28590:H30:H56`.

```
<answer>O01: H00 > H04 > H19 > H42 ; O03: H00 > H04 > H19 > H42 ; O04:
    H00 > H04 > H19 > H42 ; O08: H00 > H04 > H19 > H42</answer>
```

Scorer analysis. The summary below reports observable errors and metric outcomes; it does not describe the model's internal reasoning.

```
Omissions: none.
Intrusions: none.
Set exact: yes; query exact: no.
ID recall: 4/4; ID+Path recall: 0/4.
```

E.2.3 Missing object with a correct returned path

Gemma-4 31B, Original, arXiv:2605.24484, query `task7:2605.24484:H10:H71`.

```
<answer>O24: H00 > H85 > H55 > H03</answer>
```

Scorer analysis. The summary below reports observable errors and metric outcomes; it does not describe the model's internal reasoning.

```
Omissions: O22.
Intrusions: none.
Set exact: no; query exact: no.
ID recall: 1/2; ID+Path recall: 1/2.
```

F Additional Analyses

F.1 Cross-Benchmark Score Pairs

StructScope scores use Original SE-RA on the same 50-query subset as the heatmap. Among the matched models, 27 cover all 50 queries; Nemotron-3 Super covers 32 queries from three of the five papers. Using these available records gives Spearman's $\rho = 0.699$ and Pearson's

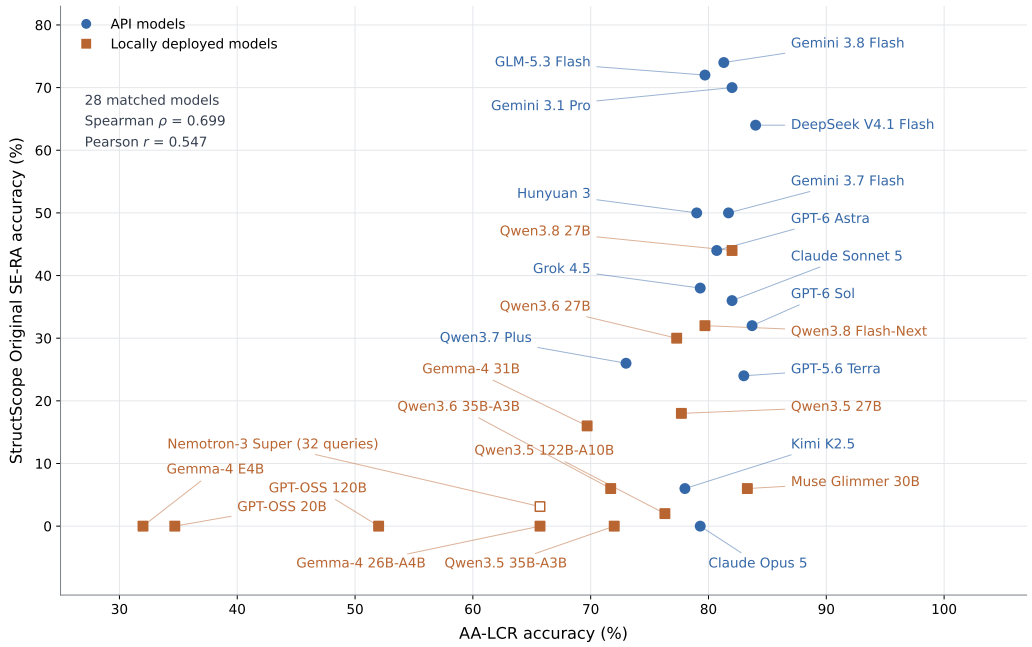


Figure 9: StructScope Original SE-RA accuracy and AA-LCR scores for the 28 matched main-paper models. All StructScope points use the 50-query heatmap subset except Nemotron-3 Super, whose hollow marker indicates 32 available queries.

$r = 0.547$. Spearman measures rank association, averaging tied ranks, whereas Pearson measures linear association between percentage scores.

F.2 Structural Factors and Task Difficulty

Table 13 groups recorded Original SE-RA results for the six models discussed in the main text. API models use 50 queries and local models 470, with bin sizes shown explicitly. These descriptive breakdowns do not control for other structural factors or paper identity.

Table 13: Original SE-RA accuracy (%) grouped by observed structural factors. Subscripts give queries per bin. Gemini 3.8 Flash and Claude Sonnet 5 use 50 queries each; Qwen3.8 27B, Muse Glimmer 30B, Gemma-4 31B, and Seed-OSS 36B use 470. Distance is the query mean of absolute character offsets between the first reference to each gold object in the positive scope and its definition; K means 1,000 characters. Path depth includes the root. These are descriptive bins, not controlled factor interventions.

Factor	Bin	Gemini	Claude	Qwen	Muse	Gemma	Seed
Reference-to-definition distance	$<$ 35K	66.7_{24}	33.3_{24}	61.4_{127}	28.3_{127}	28.3_{127}	3.9_{127}
Reference-to-definition distance	35–70K	54.5_{11}	27.3_{11}	67.3_{147}	23.8_{147}	13.6_{147}	4.1_{147}
Reference-to-definition distance	70K+	100.0_{15}	46.7_{15}	53.1_{196}	9.2_{196}	11.2_{196}	0.5_{196}
Mean gold path depth	3	9.1_{11}	0.0_{11}	65.1_{232}	25.4_{232}	19.8_{232}	2.6_{232}
Mean gold path depth	(3, 4)	100.0_{17}	52.9_{17}	55.6_{135}	15.6_{135}	11.9_{135}	2.2_{135}
Mean gold path depth	4+	86.4_{22}	40.9_{22}	53.4_{103}	8.7_{103}	15.5_{103}	2.9_{103}
Headings per paper	$<$ 50	25.0_{4}	0.0_{4}	77.2_{101}	29.7_{101}	27.7_{101}	2.0_{101}
Headings per paper	50–99	70.0_{30}	36.7_{30}	47.7_{214}	19.6_{214}	12.1_{214}	3.3_{214}
Headings per paper	100+	93.8_{16}	43.8_{16}	65.2_{155}	11.0_{155}	15.5_{155}	1.9_{155}
Gold objects per query	2	91.7_{12}	41.7_{12}	69.8_{225}	21.8_{225}	19.1_{225}	3.6_{225}
Gold objects per query	3	82.4_{17}	47.1_{17}	51.1_{141}	18.4_{141}	17.7_{141}	2.8_{141}
Gold objects per query	4	57.1_{21}	23.8_{21}	50.0_{104}	13.5_{104}	9.6_{104}	0.0_{104}

F.3 PARSE-ONLY-CONDITIONED PERFORMANCE

Table 14 groups the same queries by Parse-Only correctness and reports their Skeleton-32K and Original accuracy.

Table 14: SE-RA accuracy (%) conditioned on the same query Parse-Only correctness. Each model uses its 470-query intersection; Q is the conditional group size.

Model	Parse-Only	Q	Skeleton-32K	Original
Qwen3.6 35B-A3B	Correct	445	51.0	14.2
Qwen3.6 35B-A3B	Incorrect	25	28.0	4.0
Qwen3.5 122B-A10B	Correct	456	59.2	11.0
Qwen3.5 122B-A10B	Incorrect	14	42.9	7.1
Qwen3-Next 80B-A3B	Correct	440	44.5	3.2
Qwen3-Next 80B-A3B	Incorrect	30	23.3	3.3
Qwen-AgentWorld	Correct	454	20.5	9.7
Qwen-AgentWorld	Incorrect	16	0.0	18.8

F.4 INTERMEDIATE RESULTS IN COMPOSITE ANSWERS

Table 3 groups composite responses by object-set correctness. A conditional path success requires a complete, scorable, nonempty answer with the exact path for *every returned object*, including out-of-scope objects. Omissions affect set correctness rather than this conditional score. This query-level rate thus differs from ID+Path recall, which penalizes missing gold objects. The intermediate-result records and analysis manifest accompany the appendix; the main-text table is not repeated.

F.5 THINKING-MODE RESULTS AND NOMINAL BUDGETS

Table 15 pairs default and thinking-enabled outputs by query ID. Table 16 presents budgets of 16K and 32K, respectively, alongside recorded truncation counts.

Table 15: Matched SE-RA accuracy (%) with default and thinking-enabled decoding. Matching is by query ID separately for each model and context.

Model	Context	Q	Default	Thinking
Qwen-AgentWorld	Skeleton-32K	102	20.6	99.0
Qwen-AgentWorld	Original	102	6.9	58.8
Qwen-AgentWorld	Parse-Assist	102	14.7	92.2
Qwen3.8 27B	Skeleton-32K	102	80.4	100.0
Qwen3.8 27B	Original	102	55.9	79.4
Qwen3.8 Flash-Next	Skeleton-32K	102	95.1	98.0
Qwen3.8 Flash-Next	Original	102	42.2	77.5
Qwen3.6 27B	Skeleton-32K	102	87.3	98.0
Qwen3.6 27B	Original	102	35.3	50.0
Gemma-4 31B	Skeleton-32K	102	78.4	94.1
Gemma-4 31B	Original	102	13.7	28.4
Qwen3.5 35B-A3B	Skeleton-32K	102	35.3	57.8
Qwen3.5 35B-A3B	Original	102	1.0	1.0
Qwen3.5 35B-A3B	Parse-Assist	102	46.1	39.2
Gemma-4 26B-A4B	Skeleton-32K	102	2.9	0.0
Gemma-4 26B-A4B	Original	102	0.0	0.0

Table 16: Output budgets for the matched thinking comparison, 16K for default decoding and 32K for thinking. Parentheses give matched query counts. Truncated counts saved responses with the length finish reason.

Model	Context	Mode	limit (Q)	Truncated
Qwen-AgentWorld	Skeleton-32K	Default	16K (102)	4
Qwen-AgentWorld	Skeleton-32K	Thinking	32K (102)	0

Table 16 continued

Model	Context	Mode	limit (Q)	Truncated
Qwen-AgentWorld	Original	Default	16K (102)	14
Qwen-AgentWorld	Original	Thinking	32K (102)	0
Qwen-AgentWorld	Parse-Assist	Default	16K (102)	0
Qwen-AgentWorld	Parse-Assist	Thinking	32K (102)	0
Qwen3.8 27B	Skeleton-32K	Default	16K (102)	0
Qwen3.8 27B	Skeleton-32K	Thinking	32K (102)	0
Qwen3.8 27B	Original	Default	16K (102)	0
Qwen3.8 27B	Original	Thinking	32K (102)	2
Qwen3.8 Flash-Next	Skeleton-32K	Default	16K (102)	0
Qwen3.8 Flash-Next	Skeleton-32K	Thinking	32K (102)	2
Qwen3.8 Flash-Next	Original	Default	16K (102)	0
Qwen3.8 Flash-Next	Original	Thinking	32K (102)	1
Qwen3.6 27B	Skeleton-32K	Default	16K (102)	0
Qwen3.6 27B	Skeleton-32K	Thinking	32K (102)	0
Qwen3.6 27B	Original	Default	16K (102)	0
Qwen3.6 27B	Original	Thinking	32K (102)	1
Gemma-4 31B	Skeleton-32K	Default	16K (102)	0
Gemma-4 31B	Skeleton-32K	Thinking	32K (102)	0
Gemma-4 31B	Original	Default	16K (102)	0
Gemma-4 31B	Original	Thinking	32K (102)	2
Qwen3.5 35B-A3B	Skeleton-32K	Default	16K (102)	26
Qwen3.5 35B-A3B	Skeleton-32K	Thinking	32K (102)	33
Qwen3.5 35B-A3B	Original	Default	16K (102)	80
Qwen3.5 35B-A3B	Original	Thinking	32K (102)	68
Qwen3.5 35B-A3B	Parse-Assist	Default	16K (102)	29
Qwen3.5 35B-A3B	Parse-Assist	Thinking	32K (102)	56
Gemma-4 26B-A4B	Skeleton-32K	Default	16K (102)	60
Gemma-4 26B-A4B	Skeleton-32K	Thinking	32K (102)	102
Gemma-4 26B-A4B	Original	Default	16K (102)	50
Gemma-4 26B-A4B	Original	Thinking	32K (102)	102