WILDTRACE: BENCHMARKING NATURAL EVIDENCE TRAILS IN LONG-CONTEXT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Answering complex questions over long documents frequently requires integrating evidence that the source itself disperses naturally across distant passages. In an incident report, the operating condition, design flaw, and missed safety check that jointly explain a disaster may appear dozens of sections apart; in a novel, a character's true motive may surface only through scenes far removed from the moment it becomes relevant. This source-internal evidence integration is central to real-world long-document analysis, yet existing benchmarks largely sidestep it. Needle probes, planted facts, and reverse-engineered multi-hop chains embed evidence that may differ from the host text in distribution, placement, or register, making it unclear whether strong performance reflects genuine source reasoning or distributional artifacts. We introduce WILDTRACE, a benchmark of 481 tasks over 214 naturally occurring long-form sources such as technical incident reports and lesser-known literary narratives, where all evidence trails arise from the document's own causal, temporal, and narrative logic. Drawing on Pearl's causal hierarchy and prior multi-hop reasoning typologies, we define seven source-internal evidence geometries that characterize the distinct relational demands of analytical reading in long documents. A source-first construction pipeline mines candidate trails from document structure before writing questions; each item then undergoes multi-stage validation covering clue necessity, answer groundedness, rubric fidelity, contamination resistance and answerability. Across 18 frontier systems evaluated under evidence-withheld conditions, the strongest reaches 75.3%, with pronounced weaknesses on reasoning-intensive geometries such as counterfactual branching and causal attribution. As models are increasingly entrusted with real-world high-stakes analytical tasks, this gap between accessing information and reasoning over naturally dispersed evidence emerges as a defining challenge for the next stage of long-context research.

1 INTRODUCTION

Long-document reasoning is not merely retrieval over more tokens. For a given analytic question, a model must decide which ordinary source fragments are load-bearing evidence, which same-register details are distractors, and which causal, temporal, comparative, abductive, or counterfactual dependency makes the selected fragments jointly warrant an answer. In literary analysis, a character's motive may not reside in the scene where the decisive action occurs; it may be warranted only by linking an early promise, a later contradiction, and a delayed consequence that changes how earlier scenes should be read. Structured technical documents pose the same burden in a more procedural register. Even when a report contains findings, headings, tables, and cross-references, a follow-up question may require linking an operating condition, a design constraint, and a review lapse whose dependency is not packaged as the answer to that query. The relevant passages may all be locatable, but the task is to recover why they jointly support the conclusion.

Many existing evaluations measure important adjacent capabilities, but often control the evidence environment in ways that reduce the burden of evidence discovery and relation recovery. Needle-style and controlled long-context probes measure access, position sensitivity, aggregation, and robustness, yet the relevant facts are inserted or otherwise benchmark-controlled (Liu et al., 2024; Hsieh et al., 2024; Kuratov et al., 2024; Li et al., 2025). Pre-selected-passage and retrieval-grounded settings test whether a model can use supplied support, but the task interface has already decided which

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.

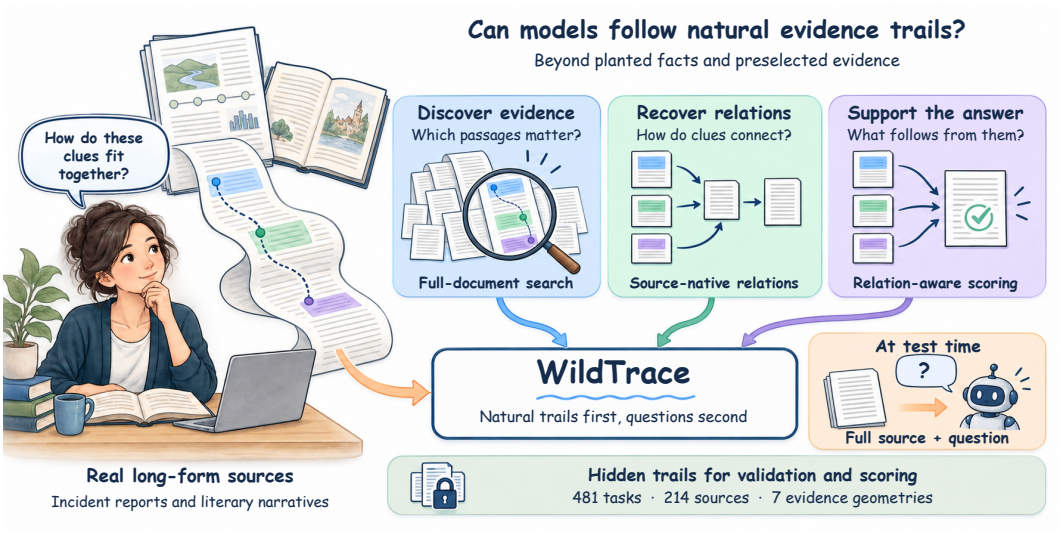


Figure 1: **WildTrace evaluates reasoning over natural evidence trails.** Real long-form sources disperse answer-critical evidence across distant passages. The benchmark tests whether models can discover this evidence, recover its source-internal relations, and preserve the dependency that supports an answer. Construction proceeds source-first: supported trails are identified before questions are written. At evaluation time, models receive only the full source and public question; evidence trails and rubrics remain hidden and support validation and scoring.

passages matter (Bai et al., 2024; Yen et al., 2025). Compositional multi-hop datasets test bridge following, comparison, and shortcut-resistant reasoning, but their chains are often assembled or reverse-engineered around target answers rather than recovered from a raw source's own narrative, procedural, and logical structure (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022).

This tradeoff is not merely philosophical. Prior analyses show that nominally multi-hop questions can be answerable from single-hop evidence or dataset artifacts, while long-context performance can depend strongly on where evidence appears and how literally the query matches the target span (Min et al., 2019; Chen & Durrett, 2019; Gururangan et al., 2018; McCoy et al., 2019; Liu et al., 2024; Modarressi et al., 2025). Inserted or localized evidence can therefore introduce construction cues—position, lexical overlap, evidence density, or stylistic mismatch with surrounding text—that need not be present when the relevant fragments are naturally authored within the source. In downstream analytical use, support is rarely pre-marked for the query: an analyst gives the model a source and asks for a conclusion that can be warranted from it. The missing evaluation boundary is therefore not long context or multi-hop reasoning in isolation, but full-document, evidence-withheld reasoning over source-induced evidence trails: deciding what counts as evidence, preserving why it counts, and reasoning over the dependency that makes it support the answer.

We introduce WILDTRACE, a benchmark for multi-hop reasoning over *natural evidence trails* in long contexts. A natural evidence trail is a set of source-internal clues and relations that jointly warrant an answer, with both the clue layout and the inferential dependency induced by the document itself. WILDTRACE is therefore built source-first rather than question-first. Starting from 214 public long-context sources, including technical incident reports and lesser-known literary narratives, we segment each source into overlapping windows, convert local context into evidence records—entities, events, attributes, temporal markers, quotations, local relations, and short summaries—and link these records into a typed source-internal graph. An edge is added only when the surrounding source text supports the relation between the linked facts. Geometry-specific queries then mine naturally occurring trails for seven evidence geometries: forward chains, intersections, comparisons, temporal reconstructions, causal fan-in, abductive explanations, and counterfactual branches (Pearl, 2019). Only after such a trail has been grounded do we write the public question, reference answer, key facts, and criterion-level rubric. Across construction, this source-first pipeline instantiated 3,506 validation-tracked candidate trail–question artifacts.

Candidate mining is high-recall by design: a trail becomes a benchmark item only after verification. Natural long-context questions can look multi-hop while still being answerable from a single local clue, from public familiarity, from an unsupported fact, or from a reasoning pattern different from the assigned geometry. We therefore organize promotion around four threats to benchmark fidelity: source grounding, multi-hop necessity, shortcut resistance, and answerability. Grounding checks verify that answers, key facts, evidence spans, and rubric criteria are supported by the source and reward answer-critical relations rather than decorative details. Necessity and shortcut checks use leave-one-out and single-clue ablations to reject items where a proper subset of the evidence already yields the unique answer, while geometry-specific review rejects collapsed relations. No-document probes filter contamination and public-knowledge shortcuts. Evidence-conditioned checks verify that the hidden trail is sufficient when supplied, and human-led plus external audits catch unsupported facts, redundant clues, ambiguous answers, and geometry mismatches. After these gates, the public release[1] retains 481 accepted tasks over 214 long-form sources, a 13.7% survival rate.

We evaluate 18 frontier systems under full-document, evidence-withheld prompts using three non-contestant judges. The strongest evaluated system reaches 75.3%, leaving substantial headroom even for frontier long-context models. The main result is not a top-1 ranking but a diagnostic account of where source-internal reasoning breaks. Models often produce locally plausible answers while failing the evidence relation: they name the right entity but drop a disambiguating constraint, recover one local story but miss the critical comparison, list plausible causal ingredients without the dependency that makes them explanatory, or follow the factual path while losing the counterfactual branch. Performance also varies by source family, context tier, and evidence geometry. These patterns suggest that the bottleneck is not only how much context a model can ingest, but whether it can turn natural evidence trails into warranted answers by selecting the right source fragments and preserving the relation among them. To summarize, our contributions are as follows:

- **A source-internal formulation of long-context reasoning.** We define natural evidence trails as source-induced configurations of dispersed evidence that jointly warrant an answer, and organize them into seven evidence geometries capturing distinct relational demands.
- **A source-first, validity-gated benchmark release.** WILDTRACE contains 481 tasks over 214 long-form source files. Evidence structures are identified before question writing, and each released item undergoes multi-stage checks of source support, evidence necessity, answerability, rubric fidelity, and geometry consistency.
- **A diagnostic evaluation of frontier systems.** We evaluate 18 systems under a full-document, evidence-withheld protocol with criterion-level multi-judge scoring. The results reveal substantial remaining headroom, systematic performance changes with document scale, and distinct reasoning profiles across evidence geometries.

2 WILDTRACE BENCHMARK AND EVALUATION DESIGN

2.1 A RUNNING EXAMPLE: LOCATING EVIDENCE IS NOT ENOUGH

We start with a task case as it shows the full capability that WILDTRACE evaluates: precise evidence localization, source-grounded support, and reasoning over the dependency that makes dispersed clues jointly answer the question. Figure 2 shows a locked L6 comparative item from a roughly 950K-token literary source. At test time, the model receives the full novel and a public question asking it to compare two distant episodes in which missing communication makes an innocent character appear morally guilty: Jack Tanerton's sailing on the *Rose of Delhi* and Mamie Lee's contested marriage to James West. The answer-critical evidence is not localized in a single passage. Two early windows at 9.2–9.3% establish Jack's apparent fault and Aunt Dean's suppression of the letters that would have explained him; three much later windows at 80.2–81.6% establish Mamie's apparent fault, the documentary reversal of that judgment, and the failed wartime contact path.

The model must do more than retrieve these windows or repeat their shared topic. The full-credit answer is a mechanism-level comparison. Jack's apparent disobedience is manufactured by deliberate suppression: Aunt Dean prevents the relevant letters from reaching their recipients, making Jack's departure appear willful. Mamie's apparent guilt is undone differently: documentary proof of

[1] Public dataset release: an anonymous repository.

marriage and James West's failed contact path explain why silence was not abandonment. A response that recovers only one episode, or says only that both cases involve missing communication, is relevant but incomplete. It identifies the shared surface frame, but misses the mechanism-level comparison: Jack's false guilt is created by deliberate suppression of letters, whereas Mamie's is resolved through documentary proof and failed wartime contact.

Case Card: Comparative Evidence Trail

TASK	Compare two distant episodes in which missing communication makes an innocent character look morally guilty: Jack Tanerton sailing on the *Rose of Delhi*, and Mamie Lee's contested marriage to James West.
GOLD	**Jack:** Aunt Dean burns both Jack's farewell and the Rector's no-sail letter, so Jack is wrongly judged to have defied a refusal he never received. **Mamie:** James's Bristol certificate proves the marriage, and his letters to Ireland plus battlefield injury explain why his silence was not abandonment. **Contrast:** both episodes reverse moral blame through communication evidence, but the mechanisms differ: deliberate suppression versus missing proof and failed wartime contact.
ANCHORS	C1 **@9.2%, hidden refusal:** Aunt Dean burns Jack's farewell and the Rector's no-sail letter. C2 **@9.3%, moral consequence:** Jack is read as disobedient and nearly disinherited. C3 **@80.2%, apparent fault:** Mamie returns with a child but no marriage lines, making her claim suspect. C4 **@81.5%, documentary proof:** James returns with the Bristol certificate proving the marriage. C5 **@81.6%, missing contact:** James's Ireland letters, wounding, and discharge explain the silence.
ROUTE TEST	**C1** gives deliberate suppression, **C2** gives its moral effect, **C3** creates Mamie's apparent guilt, **C4** overturns it with proof, and **C5** explains the failed contact; the answer must contrast these mechanisms, not just name missing communication.
PASS	Gemini-3.1-Pro 96.7 **Mechanism contrast:** Jack's false guilt is engineered by Aunt Dean's intentional interception, while Mamie's false guilt is overturned by James's certificate, misdirected letters, and wartime injury.
FAIL	GPT-5.1 45.7 **Missing one side:** recovers parts of the letter-suppression plot but loses the full Mamie reversal, so the comparison is under-supported. Qwen3.7-Plus 36.7 **Collapsed mechanism:** keeps the shared "missing communication" theme but merges suppression, marriage proof, and misdirected wartime letters into one generic misunderstanding. MiniMax-M2.7 0.0 **Context-window limit:** this L6 source is about 950K estimated tokens, exceeding MiniMax's supported context window, so the full prompt was not sent and the item receives no route credit.

Figure 2: **A WILDTRACE task design example.** The public task asks for a comparative evidence trail; the card displays construction artifacts that are hidden during evaluation. **C1–C5** are the five load-bearing clue windows, and @ marks each window's approximate source location. The item tests whether a model can recover both mechanisms across distant passages: Jack's apparent guilt comes from deliberate letter suppression, whereas Mamie's is resolved by documentary proof plus failed wartime contact. **PASS/FAIL** rows summarize representative route outcomes: high credit preserves this contrast; partial credit names nearby facts but drops or collapses a mechanism; the context-window row shows an endpoint that could not receive the full source and therefore receives no item credit.

The construction artifact turns this mechanism-level requirement into a checkable benchmark item. It stores the five load-bearing source windows, the grounded spans within them, the typed comparative relation, a reference answer, criterion-level rubric, and leave-one-out checks. Removing Jack's windows removes the deliberate-suppression mechanism; removing Mamie's windows removes the documentary reversal and failed-contact mechanism; removing the contrast leaves two local plot summaries rather than the requested comparison. The model sees only the full source and the public question, and earns full credit only if it locates the right evidence and preserves the cross-case relation that makes the answer warranted.

2.2 BENCHMARK OBJECT AND RELEASE SNAPSHOT

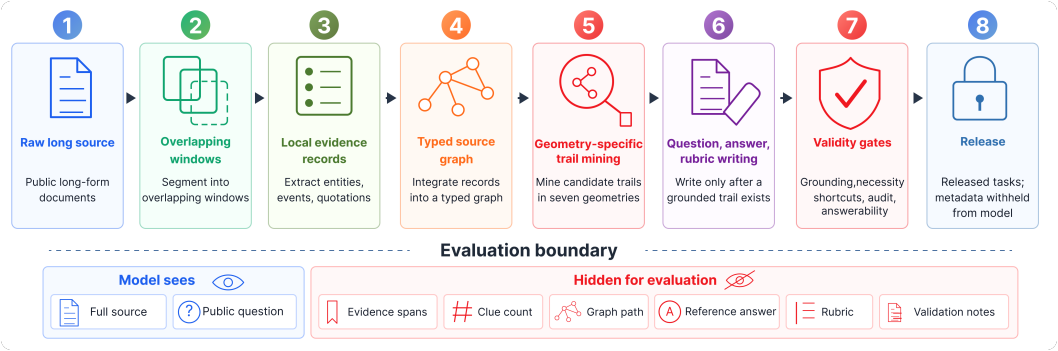


Figure 3: **WILDTRACE construction and evaluation boundary.** The pipeline is source-first: long sources are segmented into local records, linked into typed source-internal graphs, mined for geometry-specific evidence trails, and only then converted into questions, answers, and rubrics. At evaluation time, models see only the full source and public question; evidence spans, clue counts, graph paths, reference answers, rubrics, and validation notes remain hidden and are used only for audit and scoring.

The running example instantiates the benchmark object that WILDTRACE constructs at scale. Each item begins with one raw long-form source and a naturally occurring evidence trail: a set of source spans and typed relations that jointly warrant an answer. The public task exposes only the full source and a question. The construction record retains the hidden trail, primary geometry label, reference answer, criterion-level rubric, and validation outcomes for audit and scoring, but none of these artifacts are shown to the evaluated model.

Formally, a release item is a tuple $x = (S, q, T, a, R)$. Here S is the raw long-form source, q is the public question, $T = (\text{spans}, \text{edges}, g)$ is the hidden evidence trail with primary geometry g, a is the reference answer, and R is the criterion-level rubric. At test time, only (S, q) is visible. The model must therefore recover which passages are evidential and preserve the relation that makes them jointly support an answer; evidence spans, clue counts, graph paths, reference answers, rubrics, and validation notes are withheld.

The finalized release contains 481 locked tasks over 214 long-form sources, selected from 3,506 validation-tracked candidate trail–question artifacts. It covers seven evidence geometries, meets the powered L0–L6 geometry-by-tier sample floor, and includes three L7 stress probes. Figure 3 summarizes the construction and evaluation boundary; Appendix Figure 7 and Appendix A.2 give the release-composition snapshot.

2.3 SOURCE REGIMES AND EVIDENCE GEOMETRIES

The release combines three public source regimes because source-internal evidence takes different forms across document types. Technical incident reports stress causal, procedural, timeline, measurement, and corrective-action evidence. English literary narratives stress motive, delayed revelation, perspective shift, discourse order, and implicit explanation. Chinese literary sources provide an initial Chinese literature slice and diversify the release beyond English without treating WILDTRACE as a comprehensive multilingual benchmark.

Context length is stratified for balance and analysis, not treated as the sole definition of difficulty. We use half-octave full-document tiers: L0–L6 are the powered release tiers and every geometry-by-tier cell meets the final sample floor, while L7 contains only three >1M-token stress probes. Exact tier boundaries and counts are reported in Appendix Table 4.

The seven evidence geometries specify the answer-critical dependency, not a surface template: chains propagate constraints; intersections satisfy scattered conditions; comparisons align a shared dimension; temporal items recover event order; causal items aggregate contributors to an outcome; abductive items infer a latent explanation from observations; and counterfactual items preserve an

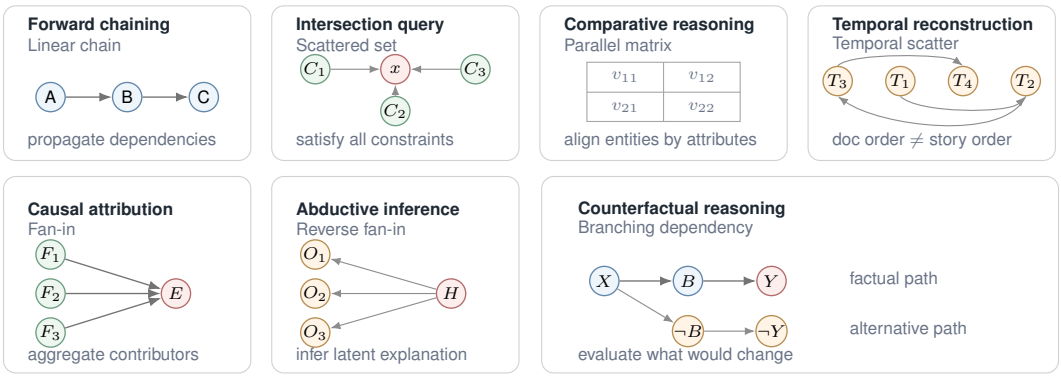


Figure 4: **Seven source-induced evidence geometries.** Each schematic node is a grounded source span and each edge is a source-internal relation discovered from the document rather than inserted by the benchmark. The geometry label names the primary dependency required for rubric credit: chains propagate constraints, intersections satisfy scattered conditions, comparisons align attributes, temporal items recover event order, causal items aggregate contributors, abductive items infer latent explanations, and counterfactual items preserve an actual path plus an alternative branch.

actual path plus an alternative branch. Labels are primary diagnostic labels at the item-and-rubric level, so ambiguous secondary relations are recorded but do not change the scored relation. Figure 4 gives the taxonomy; Appendix A.2 reports label reliability.

2.4 Source-first Trail Induction

WildTrace is constructed source-first rather than question-first. We do not write a desired multi-hop question and then search for or synthesize supporting passages. Candidate generation begins by asking whether the raw source already contains a supported evidence trail with a clear dependency structure. This order is central to the benchmark: clue placement, relation type, and distractors are properties of the document before they become properties of an evaluation item.

Mechanically, long documents are segmented into overlapping windows, converted into local evidence records, and linked into a typed source-internal graph. Rules and LLM-assisted extraction propose entities, events, attributes, temporal markers, quotations, and local relations; an edge is retained only when the endpoint facts are source-supported and the typed relation is explicit or reviewably entailed by the surrounding windows.

Geometry-specific graph queries then mine candidate trails. Only after a supported trail exists do we write the question, answer, key facts, and rubric. Questions expose enough information to be well-posed without revealing the hidden evidence relation, and rubrics score answer-critical relations rather than surface similarity. Appendix A.3 gives the artifact schema, induction pseudocode, and the rule/LLM/review boundary.

2.5 Validity Gates and Release Promotion

Natural evidence trails are easy to over-generate and hard to validate. A candidate may look multi-hop while one clue already determines the answer; a publicly familiar source may make the question answerable without the document; a proposed causal bridge may be plausible but unsupported; or a nominal geometry may collapse into a simpler local extraction task. The release-promotion checks therefore implement a threat model for benchmark validity rather than a generic quality score.

We organize promotion into four families of validity checks, with the individual checks numbered D1–D9 in the audit artifact and Appendix A.5. *Source grounding* requires every answer-critical fact and retained clue to be supported by and relevant to the source. *Multi-hop necessity and shortcut resistance* uses subset, locality, dispersion, and leave-one-out checks to reject local bypasses, shortcut edges, and single-clue-sufficient items. *Contamination resistance* probes whether the question can be answered from public familiarity or memorized summaries without the source. *Answerability and*

rubric fidelity verifies that strong readers can answer when the hidden evidence windows are supplied and that the rubric scores the answer core rather than decorative details.

The checks are binding rather than ceremonial. No-document probing rejects 38% of the validation-tracked pool; leave-one-out and single-clue audits reject items where one clue alone determines the unique answer. After all filters and audits, 3,506 tracked candidates yield 481 locked tasks, a 13.7% survival rate. Appendix A.5 gives the D1–D9 checks and grouped pass rules.

Automatic checks are followed by manual and external acceptance. Reviewers inspect the source windows, quoted support, evidence relation, geometry label, reference answer, and rubric; failures trigger rejection, evidence re-anchoring, rubric repair, or relabeling rather than silent score discounting. In a blinded relabeling study of 98 items (14 per geometry), two annotators achieved 93.9% agreement on primary labels (Cohen's $\kappa = 0.929$). Agreement with the original benchmark labels was 89.8–90.8% ($\kappa = 0.881$–0.893; Appendix A.2). Final merge hygiene removes duplicates and evidence-overlap conflicts before an item enters the locked release. Figure 6 details the validation checks and promotion funnel.

2.6 Full-Document Evaluation and Rubric Scoring

Evaluation follows the evidence-withheld boundary summarized in Appendix Table 12. The evaluated model receives only the public question and the complete source document; all construction and scoring artifacts remain hidden from the model.

Open-form responses are graded against criterion-level rubrics by three non-contestant judges—Claude-Sonnet-4.6, Qwen3.5-Plus, and Gemini-2.5-Flash—and the task score is their mean credit. The rubrics reward answer-critical facts and relations rather than exact-match or span overlap. Production judges see the question, rubric, and response, but not the source; rubric criteria are source-verified during item construction and audited through validation and external review. Appendices A.5 and A.7 report role separation, judge agreement, evidence-visible checks, and expert validation. The reported score panel uses the frozen release score matrix, and the public release provides the locked task set, validation metadata, and score matrices.

We further audited judge scores on 60 question–response pairs stratified by score range and source family and covering all seven geometries. Two domain experts verified and adjudicated model-assisted criterion-level draft scores against the rubric and verbatim source evidence. Their consensus scores correlated with the panel mean at $r \approx 0.85$, with a mean absolute difference of approximately 0.14 on the normalized 0–1 scale (Appendix A.5, Table 10).

3 Results

3.1 Frontier Performance Reveals Substantial Headroom

WildTrace remains meaningfully unsaturated at the frontier. Table 1 summarizes performance across the 18 evaluated systems. The strongest system reaches 75.3% mean rubric credit, preserving 24.7 points of headroom to full credit.

Criterion-level scoring makes this headroom informative. Each response earns credit for recovered answer-critical facts and relations, while full credit requires the complete evidence structure encoded by the rubric. The resulting score captures analytical completeness: it distinguishes answers that recover relevant content from answers that successfully assemble the requested comparison, chronology, explanation, or branch. The leaderboard therefore summarizes how fully each system converts a long source into a warranted answer.

3.2 Performance Changes Systematically with Document Scale

Full-document reasoning becomes increasingly demanding as the source grows. Figure 5A reports performance across the powered L0–L6 context tiers. Most systems achieve their strongest results on the shortest tier and lower scores on the longest powered tiers. The leading systems nevertheless retain substantial performance through L6, producing a graded decline rather than an abrupt context cliff. The shape of the curves adds an important second insight. Several systems

Table 1: **Evidence-withheld full-document score panel.** Scores are frozen three-judge rubric means from the benchmark-release score matrix; bold rows overlap the leader's 95% task-bootstrap interval.

Rank	Model	Score	CI	Rank	Model	Score	CI
1	**Gemini-3.1-Pro**	**75.3**	[72.7, 77.7]	10	Gemini-2.5-Pro	61.4	[58.5, 64.2]
2	**Claude Opus 4.8**	**72.6**	[69.7, 75.3]	11	DeepSeek-V4	60.4	[57.6, 63.2]
3	**GPT-5.5**	**71.0**	[68.3, 73.7]	12	GPT-4.1	60.4	[57.6, 63.2]
4	**GLM-5.2**	**70.9**	[68.2, 73.6]	13	GPT-5.1	42.1	[38.8, 45.2]
5	**Qwen3.7-Max**	**70.1**	[67.4, 72.8]	14	Qwen3-Max	35.2	[31.7, 38.6]
6	Claude Opus 4.6	69.2	[66.5, 71.9]	15	Kimi-K2.6	35.1	[31.4, 38.6]
7	Qwen3.6+	66.7	[63.8, 69.5]	16	DeepSeek-V3.2	33.0	[30.1, 35.9]
8	Qwen3.7-Plus	65.7	[63.0, 68.4]	17	Doubao-Seed-2.1	30.7	[26.9, 34.3]
9	GPT-5.4	62.2	[59.4, 65.1]	18	MiniMax-M2.7	28.9	[25.5, 32.4]

show local rebounds across adjacent tiers, indicating that source length and evidence organization jointly shape performance. Longer documents broaden the space of plausible passages and extend the span over which the question, partial findings, and cross-passage relations must be maintained. At the same time, coherent evidence trails can remain recoverable at long scales when their clues reinforce a clear relation. WILDTRACE therefore captures not only context reach, but the ability to sustain goal-directed evidence integration as the source grows. The three L7 items are reserved as $>$1M-token stress probes and are excluded from the powered tier trend.

3.3 EVIDENCE GEOMETRY SEPARATES DISTINCT REASONING DEMANDS

The relation encoded by an evidence trail is a major axis of performance. Figure 5B reveals a pronounced geometry profile. Averaged across the evaluated systems over L0–L6, intersection items reach 70% mean rubric credit and comparative items reach 61%, while counterfactual items average 49%. Forward, causal, temporal, and abductive items form an intermediate band at approximately 53–54%. The 21-point spread between intersection and counterfactual makes evidence geometry a central diagnostic dimension of WILDTRACE.

The ordering reflects the structural operation required by each geometry. Intersection items accumulate multiple compatible constraints toward a single entity, event, or conclusion. Comparative items align two cases along a shared dimension, providing a natural scaffold for organizing evidence. Forward chains require intermediate constraints to be carried across successive steps, while temporal items require event order to be reconstructed when it differs from document order. Causal and abductive items require the answer to preserve explanatory direction between observations, contributors, and outcomes. Counterfactual reasoning places the strongest branch-management demand on the model. A complete answer must maintain the factual and alternative states simultaneously, identify the point at which they diverge, and propagate only the consequences licensed by the alternative branch. This combination of state separation and dependency tracking helps explain why counterfactual items remain the most challenging geometry in the aggregate profile.

These geometry profiles turn the overall score into a capability signature. Systems with similar aggregate performance can exhibit different strengths in constraint convergence, cross-case alignment, temporal reconstruction, explanatory attribution, and branch-sensitive reasoning. Document scale defines how far the model must search and retain; evidence geometry defines how the recovered information must be composed.

4 DISCUSSION

The central object is a question-conditioned evidence state. The results support a shift from viewing long-context capability as token access to viewing it as the construction of a question-conditioned evidence state. A successful system must transform a long source into a compact state that preserves answer-critical facts, entity bindings, and relations. Document scale expands the material that must be filtered, while evidence geometry determines the structure that must survive this filtering. For an intersection item, this state is a set of converging constraints; for a comparison, a pair of cases aligned along a shared dimension; for temporal reconstruction, an event order distinct from discourse order; and for counterfactual reasoning, separated factual and alternative branches.

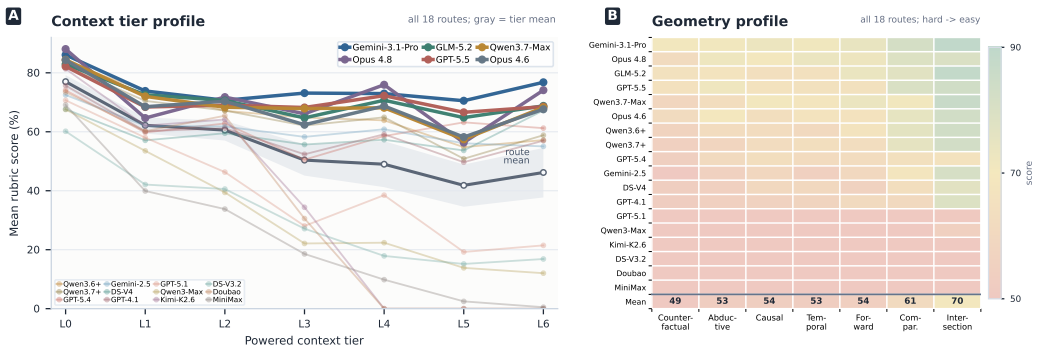


Figure 5: **Length and geometry provide complementary diagnostics.** Performance by context tier and evidence geometry. Panel A reports mean rubric credit across the powered L0–L6 context tiers. Panel B reports mean credit by primary evidence geometry. L7 is excluded from tier comparisons because it contains only three stress-test items.

This perspective connects the length and geometry results: effective long-context reasoning depends on selective retention with relational fidelity.

Different long-context architectures instantiate different forms of effective memory. Publicly documented systems already span interleaved local/global attention, latent KV compression, hybrid compressed sparse attention, and compressive memory (Gemma Team, 2025; Munkhdalai et al., 2024; DeepSeek-AI, 2026). These mechanisms allocate long-range computation differently: global access lets distant tokens interact directly, local/global hybrids concentrate that interaction in selected layers, and compressed states keep compact traces of earlier context. WILDTRACE is a natural behavioral probe for the resulting trade-offs. Intersection and comparative items test whether multiple anchors remain jointly addressable, temporal and causal items test preservation of order and direction, and counterfactual items test whether factual and alternative states remain distinct. Open-weight architecture and memory ablations, scored by geometry, could connect these mechanisms to reasoning behavior.

Internal and external memory should be studied jointly. Attention and compression determine what survives within a model call, while a harness determines what is revisited and restored across calls. Their interaction can be geometry-specific: a summary may preserve salient facts yet blur event order, causal direction, or branch identity, whereas structured evidence memory can keep these dependencies explicit. This makes model–harness co-adaptation a central future direction. Matched-budget comparisons of direct prompting, shared harnesses, and harness-aware post-training—and transfer across harnesses—would separate general long-context gains from policy-specific adaptation.

The key objective is relation-preserving compression. A useful memory need not reproduce the document; it must preserve the evidence state sufficient for the question. Training can therefore supervise these structures—entity bindings, comparison axes, temporal order, causal dependencies, and factual-versus-alternative branches—alongside final answers. WILDTRACE offers a natural testbed for measuring how they survive as context length, memory budget, and compression increase. The desired behavior is graceful forgetting: drop redundant surface detail before answer-critical dependencies. Pairing natural evidence trails with explicit relational structure gives a common substrate for architecture, memory policy, post-training, and orchestration.

5 LIMITATIONS

The main limitations of WILDTRACE concern coverage and experimental scope. The release focuses on public technical incident reports and literary narratives, mostly in English, with Chinese literature as an initial non-English slice; legal, medical, scientific, code, and enterprise settings will need additional collections. Each task has one primary evidence geometry, while real questions may combine several relations, so explicitly multi-geometry items remain future work. We evaluate direct full-document inference; retrieval-augmented, iterative, and harness-aware systems are a complementary setting.

AI USE STATEMENT

We used generative AI for manuscript polishing, methodological feedback, figure preparation, and benchmark construction and evaluation, including evidence extraction, drafting candidate questions, reference answers, and rubrics, validation probes, and response scoring. Source documents were not AI-generated. AI-assisted artifacts underwent the validation and human-review procedures described in the paper. The authors take responsibility for the final manuscript, results, and released artifacts.

ETHICS STATEMENT

WildTrace uses publicly available literary works and technical incident reports, with source provenance and licensing information documented in the release. These sources may describe harm and contain information about real individuals; we do not intentionally collect private user data. Benchmark scores reflect performance on the evaluated tasks and should not be interpreted as certification for high-stakes deployment.

REPRODUCIBILITY STATEMENT

The anonymous released dataset provide the task set, rubrics, evidence annotations, validation metadata, and score matrices. Section 2 and Appendix A document benchmark construction, validation, and evaluation. Exact reproduction of results from proprietary models may be affected by subsequent API changes.

REFERENCES

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.776. URL https://aclanthology.org/2024.acl-long.776/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://www.nature.com/articles/s41586-025-09962-4.

Jifan Chen and Greg Durrett. Understanding Dataset Design Choices for Multi-hop Reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4026–4032. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1405. URL https://aclanthology.org/N19-1405/.

Ziyang Chen, Xing Wu, Junlong Jia, Chaochen Gao, Qi Fu, Debing Zhang, and Songlin Hu. LongBench Pro: A More Realistic and Comprehensive Bilingual Long-Context Evaluation Benchmark. *arXiv preprint arXiv:2601.02872*, 2026. URL https://arxiv.org/abs/2601.02872.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf. Technical report.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. In *International Conference on Learning Representations*, pp. 42414–42448, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/465f22be10e07b301c6ed58f0472f704-Paper-Conference.pdf.

Xiang Feng, Jiawei Zhou, Zhangfeng Huang, Kewei Wang, Shanshan Ye, Jinxin Hu, Zulong Chen, Yong Luo, and Jing Zhang. DocScope: Benchmarking Verifiable Reasoning for Trustworthy Long-Document Understanding. *arXiv preprint arXiv:2605.08888*, 2026. URL https://arxiv.org/abs/2605.08888.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786. URL https://arxiv.org/abs/2503.19786.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Lawrence Keunho Jang, Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks. *arXiv preprint arXiv:2604.24964*, 2026. URL https://arxiv.org/abs/2604.24964.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106519–106554. Curran Associates, Inc., 2024. doi: 10.52202/079017-3381. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c0d62e70dbc659cc9bd44cbcf1cb652f-Paper-Datasets_and_Benchmarks_Track.pdf.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. LooGLE: Can Long-Context Language Models Understand Long Contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16304–16333. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.859. URL https://aclanthology.org/2024.acl-long.859/.

Mo Li, Songyang Zhang, Taolin Zhang, Haodong Duan, Yunxin Liu, and Kai Chen. NeedleBench: Evaluating LLM Retrieval and Reasoning Across Varying Information Densities. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=cEvmIKsRw0.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.467. URL https://aclanthology.org/2023.findings-emnlp.467/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, et al. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. In *International Conference on Learning Representations*, pp. 40903–40986, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/444a3737adaee10d86ad2ef5f74468e6-Paper-Conference.pdf.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. Compositional Questions Do Not Necessitate Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4249–4257. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1416. URL https://aclanthology.org/P19-1416/.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. NoLiMa: Long-Context Evaluation Beyond Literal Matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44554–44570. PMLR, 2025. URL https://proceedings.mlr.press/v267/modarressi25a.html.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention. *arXiv preprint arXiv:2404.07143*, 2024. doi: 10.48550/arXiv.2404.07143. URL https://arxiv.org/abs/2404.07143.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. QuALITY: Question Answering with Long Input Texts, Yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.391. URL https://aclanthology.org/2022.naacl-main.391/.

Judea Pearl. The Seven Tools of Causal Inference, with Reflections on Machine Learning. *Communications of the ACM*, 62(3):54–60, 2019. doi: 10.1145/3241036. URL https://doi.org/10.1145/3241036.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Xiangkun Hu, Zheng Zhang, Qian Wang, and Yue Zhang. NovelQA: Benchmarking Question Answering on Documents Exceeding 200K Tokens. In *International Conference on Learning Representations*, pp. 23444–23466, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3adb85a348a18cdd74ce99fbbab20301-Paper-Conference.pdf.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5627–5646. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.322. URL https://aclanthology.org/2024.emnlp-main.322/.

Renxi Wang, Honglin Mu, Liqun Ma, Lizhi Lin, Yunlong Feng, Timothy Baldwin, Xudong Han, and Haonan Li. SCALAR: Scientific Citation-based Live Assessment of Long-context Academic Reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7817–7830. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.366. URL https://aclanthology.org/2026.eacl-long.366/.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. *arXiv preprint arXiv:2504.12516*, 2025. URL https://arxiv.org/abs/2504.12516.

Biao Xiang, Soyeon Caren Han, and Yihao Ding. BRIDGE: Benchmark for multi-hop Reasoning In long multimodal Documents with Grounded Evidence. *arXiv preprint arXiv:2603.07931*, 2026. URL https://arxiv.org/abs/2603.07931.

Kai Yan, Zhan Ling, Kang Liu, Yifan Yang, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. MIR-Bench: Can Your LLM Recognize Complicated Patterns via Many-Shot In-Context Reasoning? In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-2823. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/796076672b00f54fb01d05a2e5fde363-Paper-Datasets_and_Benchmarks_Track.pdf.

Van Yang, Hongye Jin, Shaochen Zhong, Song Jiang, Qifan Wang, Vipin Chaudhary, and Xiaotian Han. 100-LongBench: Are de facto Long-Context Benchmarks Literally Evaluating Long-Context Ability? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17560–17576. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.903. URL https://aclanthology.org/2025.findings-acl.903/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question

Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to Evaluate Long-context Models Effectively and Thoroughly. In *International Conference on Learning Representations*, pp. 98914–98965, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f5332c8273d02729730a9c24dec2135e-Paper-Conference.pdf.

Haozhen Zhang, Tao Feng, Pengrui Han, and Jiaxuan You. AcademicEval: Live Long-Context LLM Benchmark. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=LjQ4voE5bs.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending Long Context Evaluation Beyond 100K Tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.814. URL https://aclanthology.org/2024.acl-long.814/.

A Additional Construction and Validation Details

This appendix begins with related work and benchmark positioning, then gives the operational details behind Section 2. Figure 4 in the main text gives the source-internal evidence-geometry taxonomy, and Figure 8 summarizes the release-promotion loop. We then spell out the artifact record, validation gates, manual-review procedure, role separation, and evaluation boundary used to audit the final benchmark.

A.1 Related Work and Benchmark Positioning

Table 2 is a boundary map, not a ranking of benchmark quality. Its columns ask whether a feature is a primary evaluation focus under the following operational definitions. "Raw-source evaluation" means that the evaluated unit is a naturally occurring long document rather than a benchmark-assembled passage set. "Same-source evidence" means the answer-critical evidence is distributed within one source. "Evidence hidden at test" means the model is not given supporting spans or clue windows at evaluation time. "Shortcut / necessity audit" means the benchmark checks whether the answer can be recovered from a single local clue or shortcut edge. "Relation taxonomy" means the dataset labels the primary evidence relation being tested, not merely the task domain.

Long-context evaluation. Long-context evaluation has moved from broad coverage toward sharper tests of realistic reasoning. Earlier suites such as LongBench, L-Eval, InfiniteBench, LooGLE, HELMET, and LOONG established diverse long-input tasks across QA, summarization, few-shot learning, code, and multitask reasoning (Bai et al., 2024; An et al., 2024; Zhang et al., 2024; Li et al., 2024; Yen et al., 2025; Wang et al., 2024). Newer suites emphasize adjacent axes: LongBench v2 and LongBench Pro stress harder realistic long-context tasks, 100-LongBench audits whether de facto long-context tests isolate long-context ability, and BRIDGE and DocScope add step-, page-, region-, and fact-level evidence supervision for long-document reasoning (Bai et al., 2025; Chen et al., 2026; Yang et al., 2025; Xiang et al., 2026; Feng et al., 2026). Controlled probes such as Lost-in-the-Middle, RULER, BABILong, NeedleBench, and NoLiMa further isolate position sensitivity, needle retrieval, synthetic fact tracing, aggregation, and lexical-overlap effects (Liu et al., 2024; Hsieh et al., 2024; Kuratov et al., 2024; Li et al., 2025; Modarressi et al., 2025). These benchmarks are crucial for measuring access and robustness. In many of these settings, however, key parts of the evidence environment are supplied, inserted, supervised, or otherwise controlled by the benchmark rather than discovered from a raw source.

Multi-hop and long-document QA. Multi-hop QA datasets such as HotpotQA, 2WikiMultiHopQA, and MuSiQue introduced bridge and comparison questions, supporting facts, reasoning paths, and filters against shortcut-solvable composition (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). Long-document QA benchmarks such as NarrativeQA, Qasper, QuALITY, NovelQA, MIR-Bench, SCALAR, and AcademicEval move toward longer books, papers, scripts, and domain-grounded tasks (Kočiský et al., 2018; Dasigi et al., 2021; Pang et al., 2022; Wang et al., 2025; Yan et al., 2025; Wang et al., 2026; Zhang et al., 2025). WildTrace inherits the need for supporting evidence and shortcut resistance, but changes the origin of the reasoning path: candidate paths are mined from the source's own narrative, temporal, causal, and entity-relation structure before a question is written.

Retrieval and citation-grounded generation. Retrieval-augmented and citation-grounded systems make provenance explicit: the model retrieves, cites, or verifies passages as part of the task interface (Lewis et al., 2020; Gao et al., 2023; Liu et al., 2023). WildTrace evaluates a different boundary. It withholds all evidence metadata and asks whether a model can infer which naturally occurring passages are evidential in the first place. This makes the task a test of source-internal evidence selection and relation preservation, not only of using retrieved support.

Frontier benchmark design. Recent frontier evaluations increasingly emphasize unsaturated tasks, expert or workflow-grounded construction, and verifiable or rubric-based scoring. Humanity's Last Exam uses expert-vetted closed-ended questions at the frontier of academic knowledge (Center for AI Safety et al., 2026); BrowseComp focuses on hard-to-find web facts with short verifiable answers (Wei et al., 2025); DeepResearch Bench evaluates long-form research reports and citation

Benchmark / resource	Unit	Raw-source eval.	Same-source evidence	Evidence hidden at test	Shortcut / necessity audit	Relation taxonomy
Broad and application-centric long-context suites						
LongBench (Bai et al., 2024)	Long-context task suite	part.	part.	part.	part.	–
L-Eval (An et al., 2024)	Long-document evaluation	P	part.	part.	part.	–
∞Bench (Zhang et al., 2024)	100K+ context suite	part.	part.	part.	part.	–
LooGLE (Li et al., 2024)	Long-dependency QA	P	part.	P	part.	–
HELMET (Yen et al., 2025)	Application-centric suite	part.	part.	part.	part.	–
LongBench v2 (Bai et al., 2025)	Realistic long-context multitask	P	part.	part.	part.	–
100-LongBench (Yang et al., 2025)	Length-controlled evaluation	part.	part.	part.	part.	–
Synthetic, haystack, and controlled long-context probes						
RULER (Hsieh et al., 2024)	Controlled context probe	–	–	P	part.	–
BABILong (Kuratov et al., 2024)	Reasoning-in-a-haystack	–	–	P	part.	–
NeedleBench (Li et al., 2025)	Needle tracing / retrieval	–	–	P	part.	–
NoLiMa (Modarressi et al., 2025)	Lexical-gap needle retrieval	–	–	P	–	–
Compositional multi-hop QA benchmarks						
HotpotQA (Yang et al., 2018)	Multi-hop QA	–	–	part.	part.	part.
2WikiMultiHopQA (Ho et al., 2020)	Multi-hop QA	–	–	part.	P	part.
MuSiQue (Trivedi et al., 2022)	Composed multi-hop QA	–	–	part.	P	part.
Long-document, narrative, and domain-QA benchmarks						
NarrativeQA (Kočiský et al., 2018)	Book / script QA	P	part.	P	part.	–
QASPER (Dasigi et al., 2021)	Scientific-paper QA	P	part.	part.	part.	–
QuALITY (Pang et al., 2022)	Long-passage QA	P	part.	P	part.	–
NovelQA (Wang et al., 2025)	Novel QA	P	part.	P	part.	–
Source-internal evidence benchmark						
WILDTRACE	Source-internal evidence item	P	P	P	P	P

Table 2: **Evaluation-boundary map for representative long-context, multi-hop, and document-QA benchmarks.** The table compares each benchmark only with respect to the boundary targeted by WILDTRACE, not overall benchmark quality or the value of other task designs. P marks a primary evaluation focus, part. marks partial, task-specific, or indirect coverage, and – means the axis is not a primary focus under the operational definitions in Appendix A.1. WILDTRACE combines these axes in one setting: naturally distributed source evidence, evidence-hidden full-document evaluation, necessity validation, and a relation-level taxonomy of reasoning demands.

quality (Du et al., 2026); Terminal-Bench 2.0 and Odysseys evaluate realistic long-horizon agent workflows with execution tests or graded rubrics (Merrill et al., 2026; Jang et al., 2026). WILDTRACE follows the same validity-first direction, but targets a different object of measurement: evidence relations that are internal to a single long source and hidden from the evaluated model.

Positioning. WILDTRACE is complementary to existing long-context, multi-hop, narrative-QA, and retrieval-grounded benchmarks. Prior suites cover important parts of the design space; WILDTRACE isolates a particular conjunction as its primary evaluation unit: a single raw long-form source whose clues, distractors, and dependencies arise internally, with evidence hidden from the evaluated model. Success requires preserving the evidence relation that warrants the answer, not only locating a fact or producing a plausible response.

A.2 SOURCE, GEOMETRY, AND CANDIDATE ARTIFACTS

Source selection and tiering. The benchmark release uses 214 long-form source files from three intentionally different source families: public English technical incident reports, English literary narratives, and Chinese literary sources. These families stress different evidence regimes. Technical reports make causal, procedural, and corrective-action relations explicit but require preserving operational dependencies. Literary sources make many relations implicit through delayed revelation, perspective, discourse order, and motive. Chinese sources provide an initial Chinese literature slice and diversify the source regime beyond English texts. Table 4 reports how the 481 locked items are distributed across full-document length tiers.

Benchmark property	Benchmark release
Locked tasks	481 questions in the benchmark release.
Source coverage	214 long-form source files: English technical reports, English literary narratives, and Chinese literary sources.
Geometry counts	Forward chain 73; intersection 71; temporal reconstruction 71; causal attribution 67; comparative 67; counterfactual reasoning 67; abductive inference 65.
Context tiers	L0 $\leq$128K: 90; L1 128K–181K: 59; L2 181K–256K: 61; L3 256K–362K: 58; L4 362K–512K: 59; L5 512K–724K: 58; L6 724K–1M: 93.
Frontier stress tier	L7 $>$1M: 3 items, reported as a stress probe rather than a powered comparison tier.
Sample floor	Fully met: L0/L6 have at least 12 items per geometry, and L1–L5 have at least 8 items per geometry.
Release rule	Only items passing release-promotion checks and external acceptance enter the locked release.

Table 3: **What is in the locked benchmark release.** The table summarizes the final dataset after release-promotion checks and external acceptance, including task count, source count, source families, geometry balance, context tiers, and release artifacts. Counts follow the frozen release and audit ledger, not earlier candidate pools.

Tier	Length	Items	Reporting role
L0	$\leq$128K	90	Short full-document tier; at least 12 items per geometry.
L1	128K–181K	59	Lower-mid tier used for degradation analysis; at least 8 items per geometry.
L2	181K–256K	61	Mid tier near common long-context routing boundaries; at least 8 items per geometry.
L3	256K–362K	58	Boundary tier where context handling and relation integration begin to diverge; at least 8 items per geometry.
L4	362K–512K	59	Long tier with hard abductive and branch-heavy cases; at least 8 items per geometry.
L5	512K–724K	58	Long tier; at least 8 items per geometry.
L6	724K–1M	93	Long-context stress tier; at least 12 items per geometry.
L7	$>$1M	3	Frontier stress probe, not a powered comparison tier.

Table 4: **Full-document length tiers used for analysis.** L0–L6 define the powered benchmark-release length strata and meet the release floor across evidence geometries; L7 contains only three $>$1M-token stress-probe items.

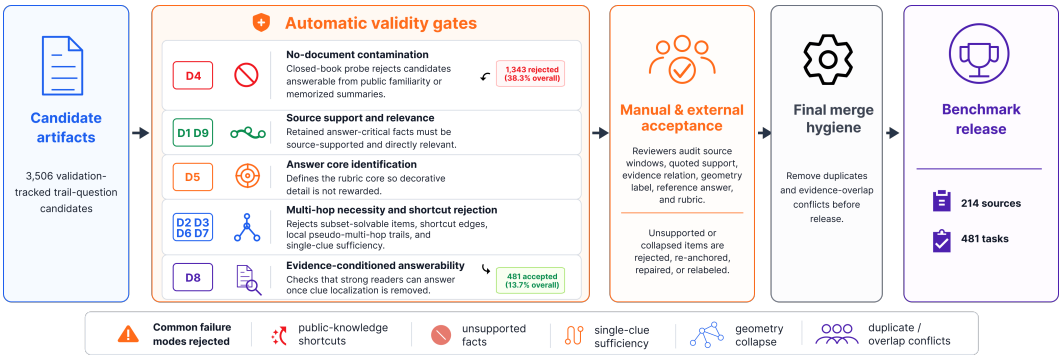


Figure 6: **Strict promotion and validation funnel.** Candidate artifacts enter the locked release only after no-document contamination probes, source-support and relevance checks, answer-core review, multi-hop necessity and shortcut rejection, evidence-conditioned answerability, manual/external acceptance, and final merge hygiene. The funnel separates construction-time validation artifacts from the evidence- withheld evaluation setting.

Primary geometry labels. Geometry labels are primary labels, not claims that no other relation appears in the source. Reviewers assign the label according to the dependency that must be preserved to earn the rubric core. For example, a counterfactual branch may contain causal steps, but it remains counterfactual when the answer-critical operation is preserving the alternative non-actual outcome. External acceptance audits reject geometry-collapse cases where the declared label is not the answer-critical relation.

Geometry relabeling reliability. We validated this primary-label scheme with a blinded relabeling study over 98 items, stratified 14 per geometry. Two annotators saw only the question, evidence spans, and rubric criteria, not the original label or construction notes. They assigned a primary label and could mark a secondary label when another relation was genuinely present but not answer-critical. The annotators disagreed on 6 of 98 primary labels, giving 93.9% agreement and Cohen's $\kappa = 0.929$. Agreement against the benchmark's original primary labels was 89.8% for annotator A ($\kappa = 0.881$) and 90.8% for annotator B ($\kappa = 0.893$); allowing secondary labels raised agreement into the mid-90s. Remaining disagreements mainly reflect adjacent explanatory relations where a secondary relation is present but not the rubric-critical one.

Reliability check	Agreement	κ	Read
Annotator A vs. B primary labels	93.9%	0.929	Six primary-label disagreements over 98 stratified items.
Annotator A vs. original primary	89.8%	0.881	Original primary labels are broadly reproducible under blinded relabeling.
Annotator B vs. original primary	90.8%	0.893	Primary-or-secondary agreement with the original label reaches the mid-90s.

Table 5: **Blinded evidence-geometry relabeling.** Two annotators assigned primary geometry labels from the question, evidence spans, and rubric, without seeing the original construction label. Agreement is high but not perfect after release filtering. Remaining disagreements mostly reflect primary-versus-secondary choices among adjacent explanatory labels, so geometry slices should be read as diagnostic groupings by the answer-critical relation rather than as mutually exclusive linguistic categories.

Candidate artifact record. Each promoted item is backed by an auditable construction record: source identifier, full-document tier, source-family assignment used for aggregate analysis, intended evidence geometry, selected source windows, grounded evidence spans, induced graph relation, question, reference answer, answer-critical key facts, criterion-level rubric, validation outcomes, and manual-review notes. The evaluated model sees none of these construction artifacts except the

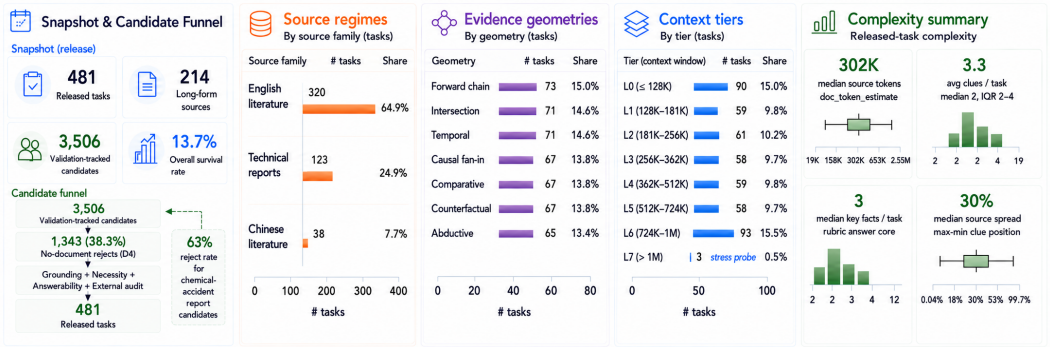


Figure 7: **WILDTRACE release statistics.** The dashboard summarizes the locked release: task and source counts, validation-tracked candidate funnel, source-family mix, evidence-geometry counts, context-tier distribution, and released-task complexity statistics computed from the strict481 artifact.

raw full source and the question. This separation is the core distinction between WILDTRACE and evaluations that provide retrieved passages or benchmark-designed clue layouts.

Artifact schema excerpt. The public release uses opaque item identifiers and a clean schema. A full task record contains model-facing fields plus hidden audit fields in the answer-bearing split; internal construction batches, repair notes, and legacy iteration identifiers are not part of the release artifact. The excerpt below illustrates the schema shape without relying on a legacy internal item id or construction log.

```
{
  "question_id": "wildtrace-0001",
  "source_id": "source_technical_001",
  "corpus_file": "corpus/source_technical_001.txt",
  "tier": "L0",
  "source_family": "technical_report",
  "geometry": "abductive",
  "question": "...",
  "evidence_spans": [
    {"span_id": "span_a", "role": "observed outcome", "quote": "..."},
    {"span_id": "span_b", "role": "latent explanation", "quote": "..."}
  ],
  "trail_edges": [
    {"from": "span_b", "to": "span_a", "type": "explains"}
  ],
  "reference_answer": "...",
  "rubric": [
    {"id": "P1", "points": 50, "criterion": "..."},
    {"id": "P2", "points": 50, "criterion": "..."}
  ],
  "validation_summary": "passed release promotion"
}
```

A.3 SOURCE-FIRST INDUCTION

Item induction proceeds from source structure to question text. First, long documents are segmented into overlapping windows and converted into local evidence records containing entities, events, attributes, temporal markers, relations, quotations, and local summaries. Second, these local records are linked into a document graph only when the surrounding source context supports the relation. Third, geometry-specific graph queries mine candidate paths: conjunctions for intersection, parallel attributes for comparison, text-order/story-order mismatches for temporal reconstruction, fan-in dependencies for causal attribution, reverse fan-in observations for abductive explanation, and factual/alternative branches for counterfactual reasoning. Only after this path exists do we write

the question and answer. This prevents the benchmark from becoming a prompt-first search for convenient passages.

Stage	Operation	Implementation and validation boundary
1	Segment source	Split raw source S into overlapping windows with saved offsets, relative positions, source id, and tier metadata. This step is deterministic.
2	Extract local records	For each window, create candidate records for entities, events, attributes, quotations, temporal markers, and local relations. This step is LLM-assisted but every retained record keeps source offsets or quoted support.
3	Propose typed edges	Add candidate edges for entity/coreference continuity, repeated attributes, temporal order, causal language, observation-to-explanation links, factual branches, and comparison dimensions. Rule filters propose high-precision edges; LLM assistance supplies paraphrase and implicit-relation candidates.
4	Verify edge support	Keep an edge only if the endpoint facts are source-supported and the typed relation is explicit or reviewably entailed by the surrounding windows. Unsupported paraphrases, ambiguous pronouns, and thematic rather than causal links are rejected or sent to manual review.
5	Query geometries	Run geometry-specific graph queries to produce candidate trails $T = (\text{spans}, \text{edges}, g)$: intersections, comparisons, temporal reconstructions, causal fan-in, abductive explanations, forward chains, and counterfactual branches.
6	Write item artifacts	Write q, a, key facts, and rubric R from the accepted trail, not before it exists. The public question exposes enough constraints for well-posedness while withholding evidence spans, clue count, graph path, and rubric.
7	Promote or reject	Apply the D1–D9 validation checks, leave-one-out necessity, no-document probing, evidence-conditioned answerability, manual review, and external acceptance. Items answerable from one clue, wrong geometry, or unsupported edges are dropped or repaired before release.

Table 6: **Pipeline overview for source-first trail induction.** The table expands the construction order used by WILDTRACE: source → local records → typed graph edges → hidden trail → question/rubric → release gate. LLMs assist extraction and candidate relation proposal, but promotion depends on source-grounded validation, leave-one-out necessity, and human/external review, so the public question is written only after a supported evidence trail exists.

Audit-level pseudocode. Table 6 is a pipeline overview rather than an executable algorithm. The audit procedure used for release promotion is captured more directly by the following pseudocode; implementation scripts may over-generate and repair candidates, but a candidate enters the locked release only through this acceptance boundary.

```
for source S in source_set:
  windows = segment(S, overlapping_windows, keep_offsets=True)
  records = extract_local_records(windows)
  graph = {}
  for candidate_edge in propose_edges(records):
    if source_supports(candidate_edge) and relation_type_is_valid(candidate_edge):
      graph.add(candidate_edge)
  for geometry g in GEOMETRIES:
    for trail T in query_graph(graph, geometry=g):
      item = write_question_answer_rubric(S, T, g)
      if passes_D1_to_D9(item) and passes_manual_external_review(item):
        release.add(lock(item))
```

Graph induction is therefore not a single model judgment. Rules and offsets provide the initial scaffold; LLM-assisted passes propose local records and candidate implicit relations; the source-support, single-clue-sufficiency, and relevance checks (D1, D7, and D9), together with manual review,

Component	Released/auditable specification
Segmentation	Nominal 25K-character overlapping induction windows with saved source id, character offsets, relative position, and tier metadata; retained clues additionally store precise offsets and quoted support.
Local record extraction	LLM-assisted JSON records over entities, events, attributes, temporal markers, quotations, local relations, and short local summaries; retained records must carry source offsets or quoted evidence.
Edge proposal and retention	Rule filters propose high-precision continuity, temporal, comparison, causal, abductive, and branch links; LLM assistance proposes paraphrase or implicit-relation candidates. An edge is retained only when the endpoint facts and relation are supported by the surrounding source windows.
Geometry query constraints	Intersection requires a unique entity satisfying all constraints; comparison requires a shared dimension; temporal requires story-time order rather than document order; causal and forward items require a supported dependency direction; abductive items require observations supporting a latent explanation; counterfactual items require both actual and alternative branches.
Promotion checks	No-document probing, evidence dispersion, leave-one-out necessity, and evidence-conditioned answerability are applied before release; final acceptance rejects items with unsupported facts, local shortcuts, single-clue sufficiency, or collapsed geometry.
Release artifact	Each locked item stores source id, tier, source family, geometry, question, evidence spans, trail edges, reference answer, rubric criteria, validation outcomes, and review notes. Model-facing evaluation hides evidence spans, clue count, graph path, reference answer, rubric, and validation notes.

Table 7: **Construction parameters and audit surface.** The table gives the stable parameters and fields needed to audit construction bias without treating the over-generation code path as a deterministic recipe for reproducing the exact same item set.

decide whether the relation is actually grounded in the source. For causal, forward-chain, abductive, and counterfactual items, reviewers also inspect the direction of the dependency: thematic parallels, narrator-stated explanations that collapse abduction, missing alternative outcomes, and single-clue sufficient shortcuts are rejection cases rather than low-confidence edges.

Construction prompt skeletons. The release package contains executable loading and scoring utilities; construction-time generation is not part of the public model-evaluation API. To make the annotation boundary auditable, we use short, role-separated prompt skeletons rather than a single end-to-end generator. The local-record prompt asks for source-grounded records only:

```
Given SOURCE_ID, WINDOW_OFFSET, and WINDOW_TEXT, return JSON records:
entities, events, attributes, temporal markers, quotations,
local relations, and a short summary. Every record must include either
a character offset or a quoted passage from WINDOW_TEXT. Do not infer
facts not supported by the window.
```

The item-writing prompt is invoked only after a supported trail exists:

```
Given SOURCE_ID, GEOMETRY, EVIDENCE_SPANS, and TRAIL_EDGES,
write a public question, reference answer, key facts, and a
criterion-level rubric. The question must not reveal evidence spans,
clue count, graph path, or the reference answer. Each rubric point must
map to an answer-critical fact or relation supported by the spans.
```

A.4 Question Writing and Rubric Artifacts

The question is written to expose enough constraints for the intended reasoning task without leaking the hidden evidence relation. For example, intersection questions may name the constraints that define uniqueness, while abductive and counterfactual questions must avoid naming the missing hypothesis or branch point because recovering that relation is the task. The reference answer and rubric are written from wide context around the evidence spans. Rubric criteria are then reduced

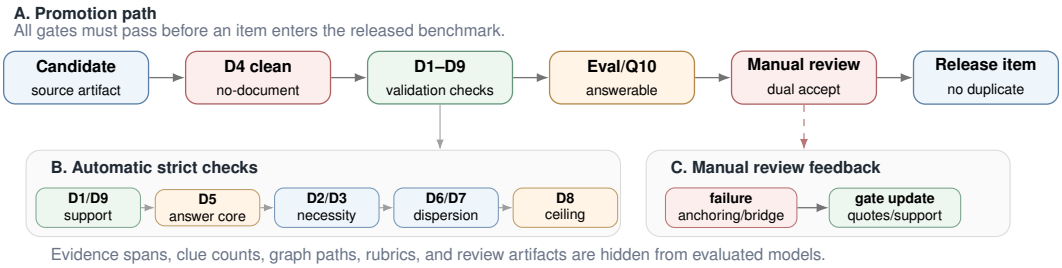


Figure 8: **Release-promotion protocol and feedback loop.** Candidates must pass contamination probes, D1–D9 strict checks, evidence-conditioned answerability, dual manual review, and final merge hygiene before entering the canonical set. The lower panels summarize review feedback: candidates are rejected or revised when the evidence trail collapses into local extraction, unsupported key facts, contamination, or a mismatched reasoning geometry.

to answer-critical facts and relations: facts that are merely vivid, stylistically useful, or locally true but not necessary are removed before scoring. This is why the benchmark gives partial credit for preserved source relations rather than for surface answer similarity.

A.5 Strict Promotion Gates

Gate order. Release promotion is dependency ordered. No-document probes run before release to reject contaminated or memorized items. D1–D9 then test whether the answer is grounded, the evidence path is irreducible, the rubric core is precise, and the item remains answerable when evidence localization is provided. The evidence-conditioned ceiling check verifies that strong readers can answer once evidence selection is removed. Dual manual review inspects the artifacts and original source windows before final merge hygiene removes duplicates or evidence-overlap conflicts.

D1–D9 grouping. The nine dimensions in Table 8 are not nine independent difficulty knobs. They form four validity groups. D1 and D9 enforce source support and relevance. D5 identifies the answer core so the rubric does not reward decorative detail. D2, D3, D6, and D7 enforce evidence dependence by rejecting shortcut edges, local bypasses, and single-clue sufficiency. D8 tests answerability under evidence-conditioned prompts, while D4/D4-alt handles no-document contamination. This grouping is what Figure 8 expands in its lower-left panel.

D4 no-document probe. D4 is an empirical filter for closed-book answerability and memorized shortcuts, not a claim of absolute source novelty. A candidate fails when a strong no-document probe can recover the answer core without the source. This removes candidates whose answer is recoverable from plot familiarity, public summaries, or canonical facts before they can enter the locked release. At the validation-tracked candidate-pool level, D4 rejects 38% of all candidates, which is why it is treated as a release gate rather than only as a post hoc diagnostic.

Gate-threshold sensitivity. The continuous gates are operating checks rather than claims of formal calibration. We use local sensitivity checks to ensure nearby operating points do not qualitatively change the accepted set, and external acceptance review remains binding for borderline cases. D7 is intentionally not summarized as a continuous threshold sweep here: the final release decision uses explicit leave-one-out acceptance, including rejection of any item where one clue alone yields the unique full answer.

Manual review. Manual review is a human-led audit of the candidate artifact, assisted where useful by model-generated traces but not delegated to an uninspected model verdict. Reviewers inspect the question, reference answer, key facts, rubric, evidence windows, quoted support, geometry label, and validation failures. A final accepted item must remain grounded, non-local, answerable, and faithful to its declared geometry under this review. The lower-right panel of Figure 8 represents feedback from this review: a failure does not simply lower confidence, it updates the evidence gate by forcing

Code	Threat to validity	Operationalization	Pass rule
D1	Unsupported reference answer	Verify each retained rubric fact against the cited source region under wide local context.	All answer-critical facts are source-supported.
D2	Subset-solvable item	Remove each required clue in turn and test whether the intended path still determines the answer.	Removing any required clue breaks the path or leaves the answer underdetermined.
D3	Structural bypass / shortcut edge	Check whether the intended reasoning chain can be collapsed by a direct edge or a single local statement.	No shortcut edge or local bypass remains.
D4	Memorization / contamination	Query strong models without the document and score against the same answer criteria.	No-document probes cannot recover the answer core; alternate probes are used for blocked or ambiguous cases.
D5	Ambiguity and rubric dilution	Use multi-model full-evidence answers to identify the shared answer structure or core facts. Detail-heavy geometries use fact containment; structure-heavy geometries use structural agreement.	Pairwise agreement passes, and non-core padding facts are removed from the answer criteria.
D6	Pseudo-multi-hop from local concentration	Compute evidence spread and minimum gap over clue offsets before expensive validation.	Retained clues must be sufficiently dispersed for the declared geometry; targeted miners may impose stricter spread gates.
D7	Single-clue sufficiency	Re-score each clue window in isolation *against the D5 core facts*, rather than against the original fact list.	No single clue window can recover the answer core or yield the unique full answer by itself.
D8	Unanswerable item or broken rubric	Give strong models all clue windows and ask whether the item is answerable once evidence localization is removed.	Evidence-conditioned readers must be able to recover the answer core from the supplied clue windows.
D9	Tangential or padded rubric facts	Judge whether each grounded fact is necessary or directly relevant to answering the question.	Only relevant facts remain; at least two grounded answer-critical facts survive.

Table 8: **Nine validation dimensions for release promotion.** Each D-check targets a concrete threat to benchmark fidelity: unsupported evidence, shortcut answerability, contamination, rubric dilution, local sufficiency, low ceiling, or mismatched relevance. The protocol is designed to preserve source-internal multi-hop necessity rather than to maximize apparent difficulty.

rejection, evidence re-anchoring, rubric repair, or geometry relabeling before a candidate can be reconsidered.

Judge-panel reliability. The final scoring panel uses Claude-Sonnet-4.6, Qwen3.5-Plus, and Gemini-2.5-Flash as non-contestant judges. Table 9 reports answer-cell and model-rank agreement, plus a judge-family sensitivity check because Qwen-family systems are also evaluated. Answer-cell Pearson correlations are at least 0.81, pairwise Kendall τ over model ranks is at least 0.91, and removing the Qwen judge shifts any model mean by at most 1.8 points.

Expert validation of judge scores. We additionally checked whether the rubric panel tracks expert judgment on a stratified 60-response sample. The sample is balanced across the three judge-score bins and the three source families, and covers all seven evidence geometries. Two domain experts scored each response against the same rubric with verbatim key-fact quotes visible; point-level model-assisted pre-scores were used only as drafts, and the experts verified and adjudicated every rubric decision against the evidence. The resulting consensus scores correlate with the three-judge panel at Pearson $r \approx 0.85$, with mean absolute human–judge difference ≈ 0.14 and mean signed difference ≈ 0. This supports using the panel mean as a practical proxy for expert rubric scoring, while not replacing larger-scale human rescoring.

Judge diagnostic	Value	Unit
Gemini-2.5-Flash grand mean	67.4	score points
Claude-Sonnet-4.6 grand mean	62.3	score points
Qwen3.5-Plus grand mean	60.9	score points
Sonnet vs. Qwen3.5 Pearson	0.886	answer cells
Qwen3.5 vs. Gemini-Flash Pearson	0.817	answer cells
Sonnet vs. Gemini-Flash Pearson	0.809	answer cells
Sonnet vs. Qwen3.5 Kendall τ	0.987	model ranks
Qwen3.5 vs. Gemini-Flash Kendall τ	0.922	model ranks
Sonnet vs. Gemini-Flash Kendall τ	0.908	model ranks
Qwen judge bias on Qwen-family outputs	-4.06	pp vs. Sonnet/Gemini mean
Qwen judge bias on non-Qwen outputs	-3.88	pp vs. Sonnet/Gemini mean
Qwen-family bias gap	-0.18	pp
Remove-Qwen-judge max model shift	1.80	score points
All-judge vs. no-Qwen rank Kendall τ	0.974	model ranks

Table 9: **Final-panel judge reliability on the benchmark release.** The table reports agreement among the three non-contestant judges at two levels: per-answer rubric scores and single-judge model rankings. The Qwen-family bias rows compare Qwen3.5-Plus scores with the mean of the Sonnet and Gemini judges on answer cells scored by all three judges; negative values mean the Qwen judge is stricter, and similar bias on Qwen versus non-Qwen outputs indicates no large observed family-specific preference in this panel.

Validation item	Result
Sample size	60 question–response pairs
Score-bin stratification	20 low, 20 medium, 20 high by three-judge score bin
Source-family stratification	20 technical-report, 20 English-literature, 20 CJK-literature pairs
Geometry coverage	All seven evidence geometries, with 7–11 pairs per geometry
Expert scoring protocol	Two domain experts verified and adjudicated model-assisted point-level draft labels against the same rubric, with verbatim key-fact quotes visible
Human–panel Pearson correlation	≈ 0.85 between expert consensus score and three-judge mean
Mean absolute difference	≈ 0.14 on the normalized 0–1 rubric scale
Mean signed difference	≈ 0.00, indicating no systematic over- or under-scoring in this sample

Table 10: **Expert validation of the LLM judge panel.** The stratified 60-response check validates the three-judge rubric mean against adjudicated expert scores. The result supports using the non-contestant judge panel as a practical scoring proxy, while the paper treats it as a sample validation rather than a full replacement for exhaustive human rescoring.

Rubric-only versus evidence-visible judging. Production scoring is rubric-conditioned grading with source-grounded rubrics: the judge sees the question, criterion-level rubric, and model response, not the full source or evidence windows. Source verification is front-loaded into item construction, D1/D7/D9 audits, and external acceptance review. As a spot check, we reran 124 stratified answer cells with Claude-Sonnet-4.6 under two conditions: the production rubric-only prompt and the same prompt with gold facts plus verbatim clue passages. The evidence-visible condition was slightly more generous on average ($+3.6$ percentage points), and 6% of cells moved down by 20 points or more. This check did not show broad rubric-only over-credit in the sampled cells.

Observed failure modes. The most common rejected candidates were not low-quality questions in a generic sense. They were specifically invalid for source-internal multi-hop reasoning: a single local clue already determined the answer, a key fact was paraphrased beyond source support, a causal or abductive bridge was plausible but not evidenced, the item label mismatched its geometry, or the answer could be recovered without the document. These observations inform the release gates above rather than serving as an alternate benchmark result.

Check	n	Rubric-only	Evidence-visible	Δ	Cells down ≥ 20 pp	Read
Sonnet spot check	124	65.8	69.4	+3.6 pp	6%	Adding gold facts and clue passages does not reveal systematic over-credit by the production rubric-only judge; the evidence-visible condition is slightly more generous on average.

Table 11: **Rubric-only versus evidence-visible judging.** Production scoring grades the question, criterion rubric, and model response; the spot-check condition additionally includes gold facts and verbatim clue passages. Small average shifts suggest that the source-grounded rubric carries most evidence constraints in this sample, while large downward moves flag responses that looked plausible under the rubric but contradicted or fabricated against the shown evidence.

A.6 Executable Reproducibility Protocol

The public release package at https://huggingface.co/datasets/nmrbench/nmr-bench contains the locked task JSON, questions-only JSONL, source files, schema, checksums, Croissant/Responsible-AI metadata, evaluation scripts, reference scores, multi-judge score matrices, and a full evaluation protocol. The model-under-test prompt is instantiated without a system message, few-shot example, chain-of-thought request, evidence span, clue count, reference answer, or rubric:

Model-under-test prompt template.

```
Answer using ONLY the document below.  Include every specific detail
from the text.

Question:
{question}

Document:
{full_document}
```

The reported leaderboard uses this question-before-document ordering (Q-before-D) because it makes the task goal explicit before very long inputs are serialized. This is the standardized released route, not a claim that prompt order is irrelevant. The instruction "Include every specific detail" was intended to reduce omission of answer-critical facts; it can also encourage verbosity. Rubric scoring therefore credits only answer-critical criteria.

For models that accept decoding parameters, generation uses low-temperature decoding and a fixed output budget under the released harness. Benchmark scores are computed from logged model responses rather than from construction artifacts. Because closed endpoints can change over time, new evaluations should archive the endpoint snapshot and decoding configuration used for that run.

The released evaluation script is the source of truth for prompt serialization, input budgeting, status logging, and score normalization. New model runs should report their configuration separately from the locked benchmark release rather than editing the task artifact.

The L0–L7 context tiers are dataset strata based on the release field `doc_token_estimate`; they are used to balance and analyze source lengths. Endpoint-specific input handling is implemented in the evaluation script and archived with each run.

At a protocol level, each non-contestant judge receives the public question, criterion-level rubric, and model response, then returns criterion points and a total score. Scores are normalized to $[0, 1]$, and the task score is the simple mean of the non-contestant judges for that logged response. The release includes both the averaged score matrix and per-judge matrices so that reviewers can inspect judge sensitivity without rerunning all model endpoints.

A.7 Evaluation Boundary and Role Separation

Evaluation boundary. The primary task is full-document and evidence-withheld. During evaluation, a model receives the full source and the question, but not evidence spans, clue counts, graph paths, reference answers, rubrics, validation notes, or manual-review artifacts. Clue-window prompts are reserved for construction diagnostics, answerability checks, and access diagnostics; they are not merged into the primary full-document leaderboard.

Component	WildTrace protocol
Model sees	Full source and public question.
Model does not see	Evidence spans, clue counts, source-window hints, graph path, reference answer, rubric, validation notes, or manual-review artifacts.
Output	Open-form answer.
Scoring	Criterion-level rubric scoring by three non-contestant judge models.
Main metric	Mean rubric credit from the frozen full-document score matrix.
Run metadata	Endpoint configuration is archived separately from the locked task artifact.

Table 12: **Full-document evaluation boundary.** Evaluated models receive only the full source and public question. Construction and scoring artifacts remain hidden during evaluation.

Benchmark release and evaluation logs. The final dataset contains 481 locked tasks. The main score panel reports the frozen three-judge rubric mean under the released full-document evaluation harness. Endpoint logs are retained as run provenance and do not redefine the locked task set or headline score panel.

Geometry	L0	L1	L2	L3	L4	L5	L6
Forward	**82.7 (12)**	70.8 (8)	72.3 (9)	74.7 (9)	62.9 (10)	51.0 (8)	72.0 (16)
Intersection	**87.5 (12)**	**80.4 (10)**	**85.9 (8)**	**92.5 (8)**	**92.7 (8)**	**86.4 (10)**	73.8 (13)
Comparative	**91.1 (12)**	62.2 (9)	**82.6 (9)**	**88.4 (9)**	**83.2 (8)**	53.9 (8)	71.3 (12)
Temporal	**82.2 (13)**	78.1 (8)	76.2 (10)	50.7 (8)	64.4 (9)	46.2 (8)	70.2 (15)
Abductive	**88.6 (12)**	56.7 (8)	63.6 (9)	59.7 (8)	65.1 (8)	69.0 (8)	70.8 (12)
Causal	79.9 (15)	72.8 (8)	44.6 (8)	54.7 (8)	72.7 (8)	61.0 (8)	75.7 (12)
Counterfactual	**82.4 (14)**	67.3 (8)	62.3 (8)	44.3 (8)	60.4 (8)	56.8 (8)	62.2 (13)

Entries are mean scores for the top-six headline systems, with item counts in parentheses. L7 is omitted because it has only three stress-probe tasks.

Table 13: **Evidence geometry and context tier interact.** Entries are top-six mean scores with item counts in parentheses, grouped by primary geometry and document-length tier. The pattern is not a pure length law: some geometries are brittle even at mid lengths, while some long-tier cells remain robust for frontier systems.

Source family	Gemini-3.1-Pro	Opus 4.8	GLM-5.2	GPT-5.5	Qwen3.7-Max	Opus 4.6	Gemini-2.5	MiniMax-M2.7
Technical reports	**77.7**	74.3	70.9	68.9	75.4	70.1	64.1	36.1
EN literature	**74.7**	71.6	71.5	71.5	68.4	69.5	60.2	25.2
CJK literature	72.5	**75.2**	69.7	74.4	68.2	63.1	62.5	36.9

Scores are diagnostic three-judge means within each source family under the same released scoring protocol.

Table 14: **Source-family slice.** Scores are diagnostic benchmark-release means within technical reports, English literature, and the smaller CJK literature slice. The columns should be read as domain diagnostics rather than a multilingual generalization claim: different source families stress different evidence regimes, and no model dominates every family.

B Additional Results and Diagnostic Ablations

The main text reports the benchmark-release leaderboard, context-tier curves, and evidence-geometry profile. This appendix keeps only supporting diagnostics that are needed to interpret those results: geometry-by-tier interaction, source-family slices, and an additive factor diagnostic. Unless otherwise noted, these slices are recomputed from the final benchmark-release multi-judge score matrix and should be read as diagnostics, not alternate leaderboards.

Factor group	Single-group R^2	Incremental R^2
Model identity	0.189	0.189
Context tier	0.130	0.032
Evidence geometry	0.022	0.028
Structural metadata	0.104	0.005
Source family	0.002	0.001
Full additive model	0.353	–

Table 15: **ANOVA-style diagnostic over benchmark-release answer cells.** The additive model includes model identity, tier, geometry, source family, and structural metadata; incremental R^2 is the drop from removing a group. This is descriptive rather than causal: length effects coexist with model, geometry, source-family, and structural effects.