BEYOND CONTEXT LENGTH: CHARACTERISING THE DIFFICULTY OF LONG-CONTEXT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context language models support ever longer context windows, and evaluations typically treat context length as the main indicator of difficulty. However, context length alone cannot explain why models fail on long-context tasks that require reasoning over complex information. Contexts of equal length can pose very different challenges: a task may be solvable from a single local passage, or it may require integrating evidence distributed across the full context. To characterise this difficulty systematically, we propose quantitative, sample-level definitions of two evidence-related properties, *scope* and *dispersion*, and further extend this characterisation with *reasoning depth*. To study how these three properties contribute to model failures, we construct LURCH, a multi-domain question-answering benchmark in which each question is provided with its answer and supporting evidence, enabling all three properties to be quantified. Controlled experiments on both proprietary and open-weight models show that: (1) at fixed context length, all our three properties are strongly negatively correlated with model performance; (2) as context length increases, performance degrades sharply on samples with dispersed evidence, whereas it remains relatively stable on samples with localised evidence; and (3) difficulty compounds across properties, as questions with moderately high values on all three are harder than those with an extreme value on just one. These findings establish evidence scope, evidence dispersion and reasoning depth as measures of long-context reasoning difficulty that complement context length.

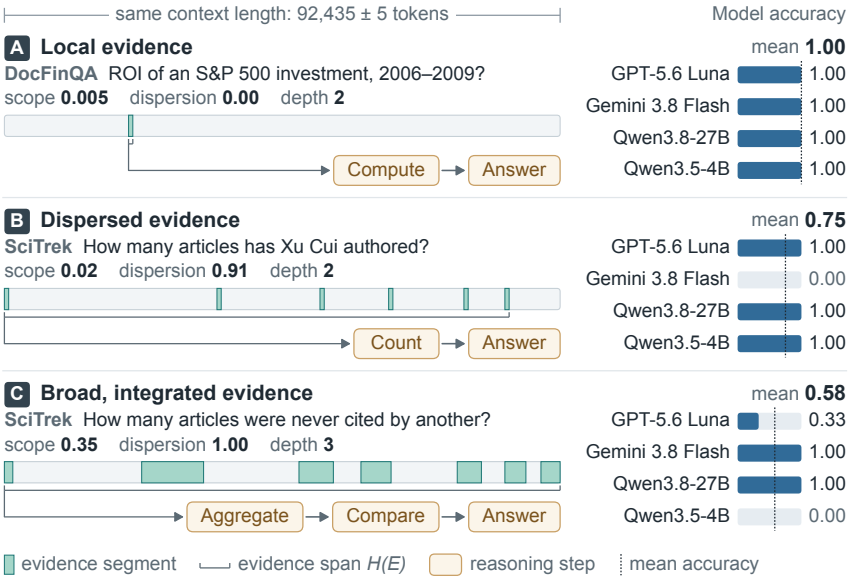

Figure 1: **Equal input length, different difficulty.** Instances of 92K tokens, varying in evidence scope, dispersion and reasoning depth. Model accuracy is the mean over three responses per LLM.

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.

1 INTRODUCTION

Large language models (LLMs) can now process hundreds of thousands or even millions of tokens enabling applications involving books, scientific papers, financial reports, and long conversation histories (Team et al., 2023; Yang et al., 2025; Singh et al., 2025; Xu et al., 2026). However, the ability to accommodate long inputs does not guarantee effective reasoning over them: models struggle to use information located in the middle of the context (Liu et al., 2024b), and their effective context length often falls well short of the claimed window (Hsieh et al., 2024). Existing evaluations typically treat context length as the main indicator of difficulty, assessing models on the same tasks at different input lengths (Li et al., 2025; Hsieh et al., 2024; Lu et al., 2026). While this approach reveals how performance changes with length, it leaves open which properties beyond token count make long-context reasoning difficult. As shown in Figure 1, contexts of equal length can pose very different challenges: a task may be solvable from a single local passage, or it may require integrating evidence distributed across the full context.

Long-context benchmarks have increasingly moved from testing whether models can retrieve information to testing whether they can reason over it, covering multi-document, domain-specific, and aggregation-heavy tasks (Bai et al., 2025; Wang et al., 2024; Reddy et al., 2024; Li et al., 2025). Yet these benchmarks mainly reveal *where* models fail, rather than *what* makes individual instances difficult. Goldman et al. (2024) take an important step by conceptualising two dimensions of long-context difficulty: *scope* ("How much necessary information is there to find?") and *dispersion* ("How hard is it to find them in the context?"). These concepts help explain why questions of similar length may impose different demands on a model, but they are defined qualitatively rather than as measurable, sample-level quantities. ETHIC (Lee et al., 2025) quantifies information coverage, but relies on an evaluator model's judgements of which context chunks are potentially necessary, and captures only one aspect of difficulty in a limited range of task settings. Without quantitative definitions, it is difficult to establish how these properties vary across tasks, how they relate to model performance, and whether they account for differences in difficulty beyond context length.

In this work, we propose quantitative, sample-level definitions of *evidence scope* and *evidence dispersion*, based on the amount and positions of the supporting evidence within the context. We further extend this characterisation with *reasoning depth*, defined as the longest dependency path in an evidence-to-answer reasoning graph: parallel evidence lookups do not increase depth, whereas operations that depend on earlier results do. Together, these three properties characterise how much supporting information is involved, where it occurs, and how it must be combined to derive an answer. To study how these properties contribute to model failures, we construct `LURCH` (Long-context Understanding and Reasoning from Contextual Hints), a multi-domain question-answering benchmark built from diverse, naturally occurring long-context tasks. Each question is provided with its answer, source-provided or automatically annotated supporting evidence, and an answer-conditioned reasoning graph, enabling all three properties to be quantified. Using `LURCH`, we conduct controlled experiments on multiple LLMs to address three research questions:

RQ1: Difficulty at fixed length. Among instances of the same context length, how strongly are evidence scope, evidence dispersion, and reasoning depth associated with model performance?

Finding: At fixed context length, task, and model, all three properties are strongly negatively correlated with performance. Notably, doubling reasoning depth is associated with a score drop comparable to that of doubling context length.

RQ2: Length-dependent difficulty. How does the effect of context length on performance vary with evidence scope, evidence dispersion, and reasoning depth?

Finding: As context length increases, performance degrades sharply on instances whose evidence is widely dispersed, whereas it remains relatively stable on instances with localised evidence. This pattern holds across all evaluated models.

RQ3: Compounding difficulty. Are model failures associated with particular combinations of scope, dispersion, and depth, even when no individual property is extreme?

Finding: Difficulty compounds across properties: questions with moderately high values on all three properties are harder than those with an extreme value on just one. This pattern is broadly consistent with an additive combination of the three properties.

These findings establish evidence scope, evidence dispersion, and reasoning depth as measures of long-context reasoning difficulty that complement context length, and provide a basis for diagnosing model failures and designing more informative long-context evaluations.

2 Related Work

Long-context language models. Both closed (e.g., Gemini 3.8 Flash, Opus 5.5, and GPT-6 Astra) and open (e.g., GLM-5.3 (GLM-5-Team et al., 2026)) weight models now support context windows of up to 1M tokens. Advances in positional encoding, including Position Interpolation and YaRN (Chen et al., 2023; Peng et al., 2024); efficient attention mechanisms, such as FlashAttention, ring attention, and sparse attention (Dao et al., 2022; Liu et al., 2024a; Llama Team, 2025); and KV-cache efficiency techniques, including GQA, PagedAttention, and MLA (Ainslie et al., 2023; Kwon et al., 2023; DeepSeek-AI, 2024), have made increasingly long contexts practical. Despite these advances, extending a model's context window does not guarantee perfect use of the information it contains. Liu et al. (2024b) find performance degradation when critical information occurs in the middle of the context. RULER (Hsieh et al., 2024) further shows the effective context length of most models, including both open-source and proprietary ones, rarely exceeds half of the claimed length. These findings suggest that context length alone does not capture the difficulty that long-context inputs pose to language models. Yet, the factors that determine the failures of long-context language models remain largely under-explored.

Reasoning over long contexts. Early benchmarks primarily examined whether models could locate and retrieve relevant information from long contexts (e.g., RULER (Hsieh et al., 2024), NeedleBench (Li et al., 2024)). More recent work extends long-context benchmarks to diverse reasoning-intensive tasks. LongBench v2 (Bai et al., 2025) includes document question answering, dialogue-history understanding, and code-repository understanding grounded on human-written questions and answers, while Loong (Wang et al., 2024) focuses on reasoning questions on multiple input documents. Domain-specific benchmarks further target particular reasoning demands. DocFinQA (Reddy et al., 2024) evaluates financial reasoning using full-document contexts, while SciTrek (Li et al., 2025) requires counting, aggregation, and comparison across collections of scientific articles. Together, these benchmarks reflect a broader shift from evaluating whether models can retrieve information from long contexts to evaluating whether they can deliver complex reasoning over them. However, existing benchmarks primarily reveal where models fail, but generally do not tell us what makes individual reasoning instances difficult. This limits our ability to identify the sources of difficulty in long-context reasoning and develop targeted approaches to address them. Recent work has begun to address this gap by examining factors that may contribute to reasoning difficulty. Goldman et al. (2024) conceptualise scope and dispersion as complementary conceptual axes of long-context tasks derived from extensive human annotation but without explicit quantification, distinguishing the amount of information required from the difficulty of locating it within the context. ETHIC (Lee et al., 2025) operationalises information coverage by quantifying the proportion of context chunks that an evaluator model deems potentially necessary to answer a query. Existing efforts, however, remain limited in both task coverage and conceptual scope, focusing on a relatively narrow range of reasoning settings and covering only a subset of the factors that may determine instance-level difficulty. As a result, we still lack a systematic framework for characterising the difficulty of long-context reasoning across diverse tasks.

3 Reasoning Characterisation

We complement context length with three properties: **evidence scope**, **evidence dispersion**, and **reasoning depth**.

Notation. Let L denote the length of the document in tokens, and let the evidence set contain N sorted, non-overlapping evidence intervals $E = \{e_1, e_2, \ldots, e_N\}$ where each interval $e_i = [a_i, b_i]$ has length $\ell_i = b_i - a_i + 1$. We merge overlapping or adjacent intervals, so the resulting intervals are disjoint and separated by at least one non-evidence token. We denote the total number of evidence tokens by $S(E) = \sum_{i=1}^{N} \ell_i$.

Evidence scope. Scope measures the proportion of the context covered by the evidence set,

$$\text{Scope}(E) = \frac{S(E)}{L}. \tag{1}$$

Scope ($\in [0, 1]$) measures the evidence fraction; dispersion distinguishes a single evidence block from scattered passages of equal scope.

Evidence dispersion. Dispersion measures the extent to which the evidence spans the document, relative to the amount of non-evidence text. We define the **evidence span** as the distance from the beginning of the first interval to the end of the last interval, $H(E) = b_N - a_1 + 1$. (Note that $H(E) \geq S(E)$.) The number of non-evidence tokens lying between the first and last evidence intervals and the total number of non-evidence tokens in the context are, respectively,

$$G_{\text{internal}}(E) = H(E) - S(E), \qquad \text{and} \qquad G_{\text{total}}(E) = L - S(E). \tag{2}$$

We define evidence dispersion as the fraction of all non-evidence text that lies inside the evidence span,

$$\text{Dispersion}(E) = \frac{G_{\text{internal}}}{G_{\text{total}}}. \tag{3}$$

Dispersion lies in $[0, 1]$: values near zero indicate clustered evidence, while values near one indicate that most non-evidence text lies between the evidence intervals. It is zero for a single interval and one when the evidence span covers the full context ($a_1 = 0$, $b_N = L - 1$).

Reasoning depth. Reasoning depth measures the sequential dependencies involved in deriving an answer from evidence. We represent a derivation as a directed acyclic graph, in which nodes denote atomic operations, including evidence lookups and inference steps, and edges denote direct dependencies between operations. Depth is the number of nodes on the longest dependency path ending in an answer-producing operation. This representation accommodates both sequential inference and the synthesis of evidence obtained in parallel; it does not assume that reasoning forms a single chain. For example, the question "Which film was released first, Bichunmoo or Musa?" needs the release year of each film. The two years are retrieved independently at depth one, and comparing them is one further step at depth two (Figure 7 in Appendix E). Additional parallel lookups do not increase depth unless further sequential operations are needed to produce the answer. Under this convention, an answer read directly from the evidence has depth one. Depth does not count the total number of operations or the difficulty of individual steps.

4 Benchmark Construction

We construct LURCH by standardising instances from existing long-context benchmarks, using reference or automatically annotated evidence, and generating validated reasoning graphs (Figure 2).

4.1 Data Collection

Filtering, validation and subsampling reduce 323,040 instances from 26 candidate sources to 6,550 instances from 17 sources. Per-source statistics are in Table 3; Appendix A catalogues all candidate sources, including those excluded.

Multi-stage data filtering. We retain English-language instances with naturally occurring contexts of 32k–256k tokens, prioritising tasks involving multi-hop reasoning, comparison, aggregation, or evidence integration rather than single-fact retrieval or span extraction. We also exclude questions answered correctly by a screening model without the context, reducing the likelihood that they can be solved from parametric knowledge alone. Filtering details are in Appendix B.

Unified prompts and evaluation metrics. We preserve the original task instructions while standardising the request for step-by-step reasoning and the final-answer format across sources. We evaluate final answers using Exact Match or token-level F1, as appropriate for each task. For

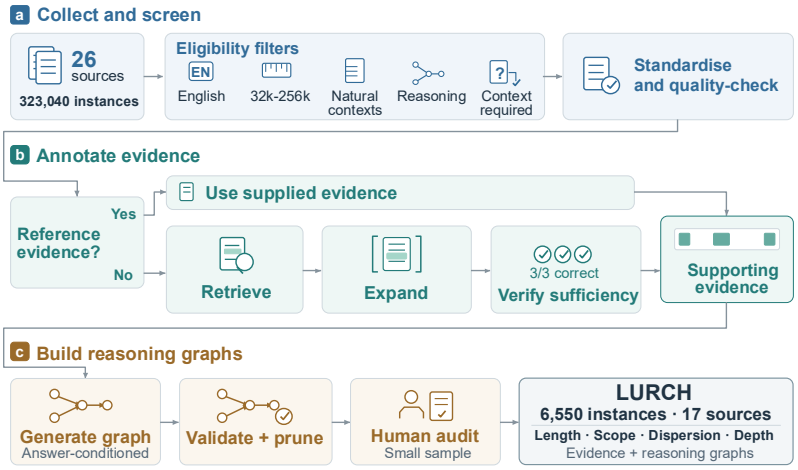

Figure 2: The dataset construction pipeline of LURCH.

Loong (Wang et al., 2024), we follow the original LLM-as-a-judge protocol to assess open-ended responses for which lexical overlap alone may not capture answer quality accurately. Per-source metrics and evaluation details are provided in Appendix C.

Question and answer quality screening. To identify quality issues not captured by automated filters, we manually inspect five randomly sampled instances from each remaining source. We check whether (i) reference answers are non-empty and complete, (ii) answer formats satisfy the question's requirements, and (iii) questions are unambiguous and answerable from the supplied context. We exclude a source if any inspected instance fails these checks. This small-sample review serves as a source-level screening step rather than an exhaustive validation of the retained instances.

4.2 Evidence Annotation

We associate each instance with supporting evidence to measure scope and dispersion and construct reasoning graphs. We use benchmark-provided reference evidence where available. Otherwise, we retrieve candidate passages, expand them to their containing paragraphs or full tables, and merge overlapping or adjacent intervals. A verifier then answers the question using only the expanded evidence, without access to the reference answer; we retain instances only if all three sampled responses are correct under the task's evaluation criterion. Scope and dispersion therefore describe the expanded supporting context, not necessarily a minimal evidence set (more details in Appendix D).

4.3 Reasoning Graph Generation

After evidence retrieval and sufficiency validation, we use two separate annotators, Gemini 3.7 Flash and GPT-5.6 Terra, to construct answer-conditioned reasoning graphs. Each receives the question, reference answer, and numbered evidence statements and reconstructs a minimal evidence-to-answer derivation. We validate each graph by removing malformed or duplicate steps, invalid references, and cycle-forming edges, then prune steps that do not contribute to the answer; unparseable or empty graphs are invalid. Two human auditors compare the two graphs on 40 instances. Neither auditor prefers GPT graphs more often than Gemini graphs, and one prefers Gemini on 14 of the 20 instances where the depths differ. We therefore compute the graph-derived properties using only Gemini generations. Details of annotator agreement, the audit procedure, and audit results are provided in the Appendix E. We compute reasoning depth from these graphs (Section 3).

4.4 Dataset Properties

Figure 3 shows context length and the three properties over all instances. Contexts have a median of 103k tokens and a mean of 110k. Within them, the supporting evidence is sparse, with a median of just 0.66% of the context and spanning more than three orders of magnitude. Only 11% of instances have all evidence in a single contiguous block, and the median dispersion among the rest is 0.39.

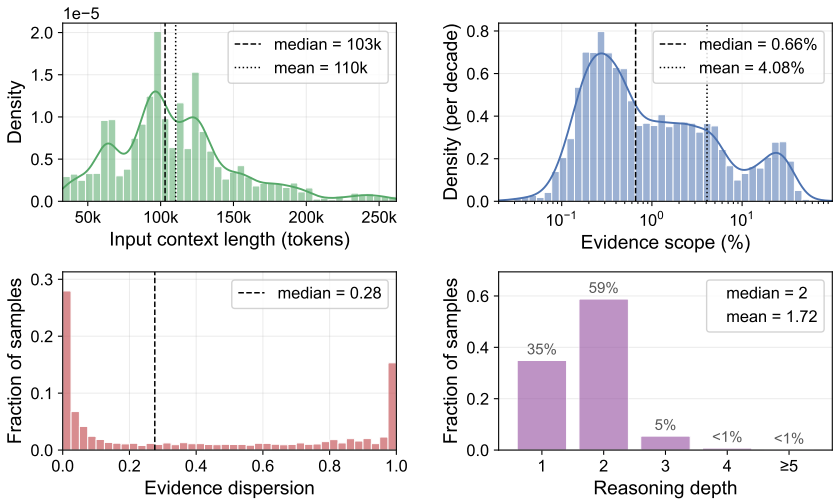

Figure 3: Distributions of input context length, evidence scope, evidence dispersion, and reasoning depth. Dashed and dotted lines indicate the median and mean, respectively; zero dispersion denotes evidence contained in a single contiguous block.

Reasoning depth is mostly multi-step: 65% of instances require two or more dependent steps. The benchmark therefore couples long inputs with sparse, widely dispersed evidence and non-trivial compositional reasoning. Per-task distributions are in Appendix F.

5 Experiments

5.1 Experimental Settings

We evaluate two proprietary LLMs, GPT-5.6 Luna and Gemini 3.8 Flash, and two open-weight LLMs, Qwen3.8-27B and Qwen3.5-4B, using the full context for each question and the prompts and scoring procedures described in Section 4.1. For each question, we sample three responses from each LLM and use their mean score as the outcome for that question–LLM pair. We analyse all 6,550 questions from the 17 tasks retained in LURCH. We apply $\log_2$ to context length, evidence scope, evidence dispersion, and reasoning depth (as defined in Section 3). We replace any zero values with half the smallest positive value of the corresponding predictor before taking logs. We then standardise each logged predictor to zero mean and unit variance across the analysed questions. Before standardisation, a one-unit increase on the log scale corresponds to doubling a positive predictor value; after standardisation, a one-unit increase corresponds to one standard deviation of the logged predictor. For the regression analyses, we compute standard errors clustered by question to account for dependence between the four LLMs' scores on the same question. Unless stated otherwise, brackets and bands give individual, unadjusted 95% question-clustered confidence intervals (CIs). For RQ1 and RQ2, we report Holm-adjusted p_{adj} values (Holm, 1979), with a separate family of tests for each research question as specified below. RQ3 is exploratory and the per-LLM breakdowns are descriptive; neither includes formal hypothesis tests. The analyses estimate conditional associations within LURCH, rather than causal effects.

5.2 Generalised Linear Model

We use a generalised linear model (GLM; Nelder & Wedderburn, 1972) with a logit link to estimate how each predictor relates to expected score, accounting for the other predictors, task, and LLM. For the mean score $y \in [0, 1]$ of LLM m on question q, we specify

$$\mathbb{E}[y \mid q, m] = \sigma\Big(\alpha_t + \alpha_m + \sum_{k=1}^{K} \beta_k \, x_k(q)\Big) \tag{4}$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the logistic function, which keeps predicted means between zero and one. Here, $K = 4$ is the number of predictors, and $x_k(q)$ is the value of predictor k (context

Table 1: RQ1 coefficient estimates and interquartile contrasts. $\Delta_{25\to75}$ is the interquartile contrast (Section 5.2), averaged over all 6,550 questions. Brackets give 95% CIs. p_{adj} tests $\beta_k = 0$ using the RQ1 Holm family.

Predictor	$\hat{\beta}$ (per SD)	$\Delta_{25\to75}$ (pp)	p_{adj}
Context length (reference)	$-0.11\ [-0.17, -0.06]$	$-1.66\ [-2.49, -0.83]$	–
Evidence scope	$-0.21\ [-0.30, -0.12]$	$-2.14\ [-3.07, -1.20]$	3.4×10^{-5}
Evidence dispersion	$-0.10\ [-0.17, -0.03]$	$-1.10\ [-1.84, -0.36]$	4.0×10^{-3}
Reasoning depth	$-0.10\ [-0.15, -0.05]$	$-1.60\ [-2.35, -0.85]$	7.4×10^{-5}

length, evidence scope, evidence dispersion, or reasoning depth) after the log transformation and standardisation described in Section 5.1. The intercept terms α_t and α_m account for baseline differences across tasks and LLMs, respectively, with t denoting the task of question q. The attribute coefficients β_k are shared across tasks and LLMs.

We fit the model using fractional logistic regression (Papke & Wooldridge, 1996), which accommodates scores throughout $[0, 1]$, including zero and one, without assuming binary outcomes. It specifies the conditional mean in Equation 4; uncertainty is estimated using the question-clustered standard errors described in Section 5.1.

Each β_k represents the change in the logit of the expected score per standard deviation of the logged predictor, holding the other predictors fixed. We also report contrasts between fitted predictions in percentage points (pp), holding other predictors fixed. The interquartile contrast $\Delta_{25\to75}$ averages, over all questions, the predicted score change when a predictor moves from its within-task 25th to its 75th percentile; tasks with tied quartiles contribute zero.

For RQ2, we extend the model with interactions between context length, L, and another attribute j:

$$\mathbb{E}[y \mid q, m] = \sigma\Big(\alpha_t + \alpha_m + \sum_{k=1}^{K} \beta_k\, x_k(q) + \gamma_{Lj}\, x_L(q)\, x_j(q)\Big) \tag{5}$$

The slope for context length on the logit scale is then $\beta_L + \gamma_{Lj} x_j$. The coefficient γ_{Lj} measures how much this slope changes per standard deviation of the logged attribute j. We fit one such model for each of scope, dispersion, and depth, and distinguish these logit-scale interactions from score-scale contrasts in RQ2. We call the model without interaction terms (Equation 4) the additive GLM. RQ3 uses it as a reference for exploring combinations of properties.

5.3 RQ1: Difficulty At The Same Length

We ask whether evidence scope, evidence dispersion and reasoning depth are associated with score after accounting for context length, task and LLM. Table 1 reports coefficients and within-task interquartile contrasts from a GLM fitted jointly to all four LLMs (Equation 4). We test the three non-length coefficients individually and jointly using Wald tests, applying Holm's correction to these four tests as one family. Context length is included as a reference. Figure 4 shows the same interquartile contrasts from GLMs fitted separately to each LLM.

All three properties are associated with lower scores. The joint test rejects the null hypothesis that scope, dispersion and depth all have zero coefficients ($\chi^2_3 = 73.0$, $p_{\text{adj}} = 4.0 \times 10^{-15}$). Each coefficient is also negative and significant after adjustment (all $p_{\text{adj}} \le 4.0 \times 10^{-3}$): holding the other predictors fixed, questions requiring more evidence, more dispersed evidence, or deeper reasoning receive lower scores.

Scope and reasoning depth are comparable to context length in magnitude. Moving from the 25th to the 75th percentile within a task is associated with score changes of -2.14 pp for evidence scope and -1.60 pp for reasoning depth, compared with -1.66 pp for context length (Table 1). Scope also has the largest coefficient magnitude per standard deviation of the logged attribute (-0.21), while dispersion has the smallest interquartile score change (-1.10 pp). Table 7 reports doubling contrasts and a comparison of all four predictors restricted to the nine tasks with distinct depth quartiles.

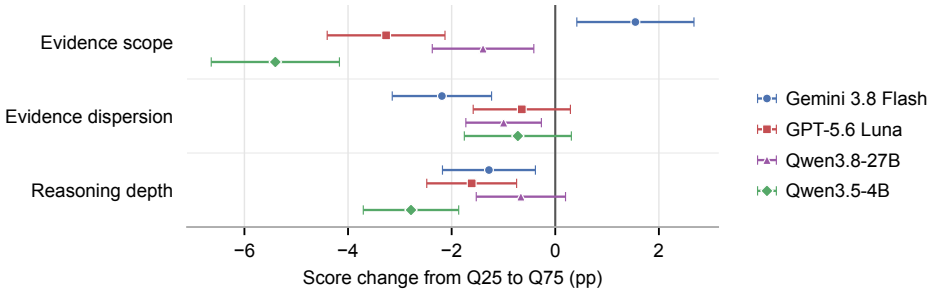

Figure 4: Per-LLM interquartile contrasts, averaged over all questions as in Table 1. Points show predicted score changes; bars give 95% CIs.

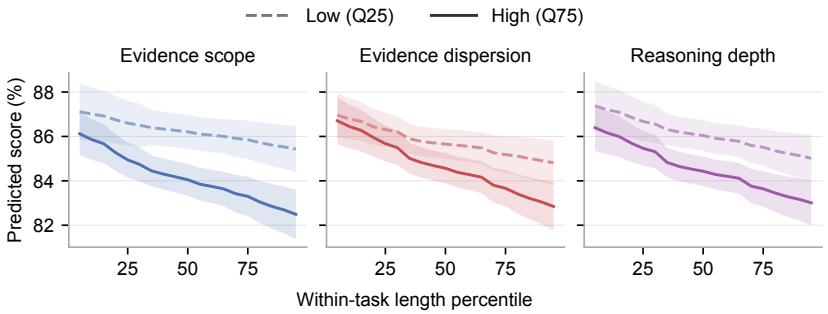

Figure 5: Predicted score at low (dashed) and high (solid) attribute values across within-task length percentiles, averaged over all questions. Bands give pointwise 95% CIs.

Variation across LLMs. The estimated depth and dispersion contrasts are negative for all four LLMs, although some intervals include zero. The depth contrast ranges from -2.78 pp for Qwen3.5-4B to -0.66 pp for Qwen3.8-27B. Scope is negative for three LLMs but positive for Gemini 3.8 Flash ($+1.55$ pp, 95% CI $[0.42, 2.68]$). The overall scope association therefore masks variation across LLMs.

5.4 RQ2: Length-dependent Difficulty

Does the association between context length and score depend on evidence scope, evidence dispersion, or reasoning depth? Using Equation 5, we estimate score changes from the within-task 25th to 75th length percentile at low (25th) and high (75th) within-task attribute percentiles. Table 2 reports these *length contrasts* and their high-minus-low differences; Figure 5 shows the corresponding curves. Negative differences indicate larger score drops at high attribute values. We apply Holm's correction to six tests: each attribute's contrast difference and interaction coefficient against zero.

Length matters more for broad or dispersed evidence. The high-minus-low differences are -0.90 pp for scope and -0.88 pp for dispersion (Table 2). Both differences and the corresponding logit interactions are significant after Holm adjustment, supporting variation beyond the logistic link alone (Table 8). The direction persists with evidence tokens replacing relative scope and across all four LLMs, although some per-LLM intervals include zero (Appendix G.3).

The interaction between length and reasoning depth remains uncertain. Predicted scores are lower at high reasoning depth across the plotted length range, consistent with RQ1. However, the difference between length contrasts (-0.66 pp) has an interval spanning zero ($p_{\text{adj}} = 1.8 \times 10^{-1}$). The logit interaction is likewise inconclusive.

5.5 RQ3: Compounding Difficulty

Which combinations of scope, dispersion and reasoning depth are hardest, and how well does the additive GLM explain their scores? Figure 6 groups questions by within-task attribute terciles. Scores are centred on each task–LLM mean to account for baseline differences. We compare observed

Table 2: RQ2 score contrasts (pp), averaged over all 6,550 questions. Tied attribute quartiles contribute zero to the high-minus-low difference. Brackets give 95% CIs; p_{adj} tests a zero difference using the six-test RQ2 Holm family.

	Length contrast (pp)			
Attribute	Low	High	Difference (high − low)	p_{adj}
Evidence scope	-0.75	-1.64	$-0.90\ [-1.43, -0.37]$	5.5×10^{-3}
Evidence dispersion	-1.13	-2.01	$-0.88\ [-1.45, -0.32]$	1.0×10^{-2}
Reasoning depth	-1.16	-1.82	$-0.66\ [-1.43, \ \ 0.10]$	1.8×10^{-1}

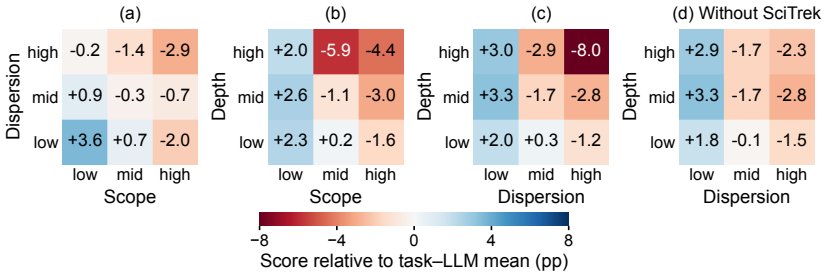

Figure 6: Mean score relative to the task–LLM mean (pp), by within-task attribute tercile. Panel (d) omits SciTrek. Residuals, confidence intervals and question counts are reported in Table 12.

scores with out-of-fold predictions of Equation 4; negative *residuals* indicate lower scores than predicted. This analysis is exploratory, with no formal hypothesis tests; its intervals are individual and not adjusted for the many cells compared. Appendix G.4 gives full cell estimates, task-sensitivity checks and three-attribute configurations.

The results show that (1) several moderate demands can coincide with low scores, and (2) deep reasoning over dispersed evidence is especially difficult in SciTrek. Questions low on both scope and dispersion score 3.6 pp above their task–LLM mean, while those high on both score 2.9 pp below it (Figure 6a). The gap remains 6.1–7.2 pp when any single task is removed. The additive GLM broadly tracks this pattern: eight of nine cell residuals lie within ± 1.0 pp (Table 12). Likewise, the 164 questions above their task median on all three properties, with none exceeding its within-task 90th percentile, score 5.4 pp below average. Their residual is -1.5 pp (95% CI $[-5.7, 2.6]$; Table 10). Low scores can therefore occur when several demands coincide, even without an extreme attribute. Questions high on both dispersion and depth form the lowest-scoring cell (-8.0 pp; Figure 6c), with a residual of -3.2 pp (95% CI $[-7.1, 0.6]$; Table 12). SciTrek supplies 62% of these questions. Removing it reduces the high-minus-low depth gap at high dispersion from -6.7 to -0.8 pp (Figure 6d). The wide residual interval and sensitivity to one task make this a candidate pattern for further study.

6 Conclusion

We argued that context length alone is an incomplete measure of long-context difficulty, and introduced three sample-level properties that complement it: evidence scope, evidence dispersion, and reasoning depth. To quantify them, we built LURCH, a benchmark of 6,550 naturally occurring instances from 17 sources, with verified evidence and reasoning graphs produced by our annotation pipeline. Across four models, these properties capture difficulty that length does not. At fixed length, broader, more dispersed and deeper questions score lower, with doubling depth costing about as much as doubling length (RQ1). Added length is harmful mainly when evidence is broad or dispersed (RQ2), and the hardest questions are demanding on several properties at once, even when none is extreme, because their effects add up (RQ3). We therefore recommend that long-context benchmarks report scope, dispersion and depth alongside length. Since our analyses are observational and some annotations are automatic, future work could vary the surrounding context while holding evidence and reasoning fixed, to test these effects causally.

AI USE STATEMENT

In this work, we used generative AI tools for labelling and filtering datasets, implementing methods and interpreting results. We have not used generative AI tools for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, and support qualitative and thematic data analysis. Additionally, we used generative AI tools for editing research paper to improve grammar. We have reviewed all AI-assisted work by implementing tests for LLM-generated codes. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. The properties are formally defined in Section 3, and the construction of LURCH is summarised in Figure 2 and Section 4. Appendix A catalogues all candidate sources, including excluded ones. Appendix B specifies every filtering criterion, including the tokenizer used for length filtering and the zero-context screening model. Table 3 lists per-source statistics, evaluation metrics and correctness thresholds. Appendix D and Appendix E give the evidence retrieval, expansion and sufficiency-validation procedure and the reasoning-graph annotation and validation procedure, together with the exact retrieval and annotation prompts, annotator agreement and the human audit protocol. The evaluated models, sampling protocol, preprocessing and inference procedures are described in Section 5.1, and the GLM specifications in Section 5.2. Appendix G details the cross-validation design and all sensitivity analyses. The supplementary material contains anonymised source code for the full pipeline: data collection, evidence annotation, model inference and scoring, reasoning-graph annotation, and attribute computation. It also includes all prompts, a pinned `requirements.txt`, and notebooks that regenerate tables and figures in paper. The supplement also includes a 100-instance subset of LURCH with its evidence, properties and full-context results, which exercises the data format and analysis code. The full dataset, including evidence annotations and reasoning graphs, will be released upon publication.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 4895–4901. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.298. URL https://doi.org/10.18653/v1/2023.emnlp-main.298.

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, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. 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, 2025.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, Haijun Lv, Yicheng Zou, Hang Yan, Kai Chen, and Dahua Lin. What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27129–27151, 2025.

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.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *CoRR*, abs/2405.04434, 2024. doi: 10.48550/ARXIV.2405.04434. URL https://doi.org/10.48550/arXiv.2405.04434.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16576–16586, 2024.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

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? *arXiv preprint arXiv:2404.06654*, 2024.

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. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles,*

SOSP 2023, Koblenz, Germany, October 23-26, 2023, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien MR Arnold, Vincent Perot, Siddharth Dalmia, et al. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121*, 2024.

Taewhoo Lee, Chanwoong Yoon, Kyochul Jang, Donghyeon Lee, Minju Song, Hyunjae Kim, and Jaewoo Kang. Ethic: Evaluating large language models on long-context tasks with high information coverage. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5497–5512, 2025.

Miao Li, Alexander Gurung, Irina Saparina, and Mirella Lapata. Who gets cited most? benchmarking long-context numerical reasoning on scientific articles. *CoRR*, abs/2509.21028, 2025. doi: 10.48550/ARXIV.2509.21028.

Miao Li, Irina Saparina, Alexander Gurung, and Mirella Lapata. Long-context reasoning through proxy-based chain-of-thought tuning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41335–41347, 2026.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *CoRR*, abs/2407.11963, 2024. doi: 10.48550/ARXIV.2407.11963. URL https://doi.org/10.48550/arXiv.2407.11963.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion. *arXiv preprint arXiv:2501.15089*, 2025.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ringattention with blockwise transformers for near-infinite context. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=WsRHpHH4s0.

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, 2024b.

Llama Team. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *Technical Report*, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Zhiyuan Lu, Chenliang Li, Yingcheng Shi, Weizhou Shen, Ming Yan, and Fei Huang. Corpusqa: A 10 million token benchmark for corpus-level analysis and reasoning. *arXiv preprint arXiv:2601.14952*, 2026.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238, 23972327. URL http://www.jstor.org/stable/2344614.

Leslie E. Papke and Jeffrey M. Wooldridge. Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, 11 (6):619–632, 1996. ISSN 08837252, 10991255. URL http://www.jstor.org/stable/2285155.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. Docfinqa: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 445–458, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zecheng Tang, Keyan Zhou, Juntao Li, Baibei Ji, Jianye Hou, and Min Zhang. L-citeeval: Do long-context models truly leverage context for responding? *arXiv preprint arXiv:2410.02115*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning. *arXiv preprint arXiv:2505.17667*, 2025.

Minzheng Wang, Longze Chen, Fu Cheng, 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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5627–5646. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.322. URL https://doi.org/10.18653/v1/2024.emnlp-main.322.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 67244–67258, 2025.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William 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, 2018.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, et al. Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k. *arXiv preprint arXiv:2402.05136*, 2024.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. Longcite: Enabling llms to generate fine-grained citations in long-context qa. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5098–5122, 2025.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. ∞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, 2024.

A EXISTING LONG-CONTEXT DATASETS

We provide a brief description of the candidate sources from which we collect our data. Table 3 lists the sources we retain. We exclude BABILong, ChatQA2, ETHIC, L-CiteEval, LOFT, LongCite, LongMIT and LongReason.

BABILong (Kuratov et al., 2024) is a benchmark designed to test language models' ability to reason across facts distributed in extremely long documents. BABILong includes 20 reasoning tasks, including fact chaining, simple induction, deduction, counting and handling lists/sets. Notably, this benchmark is extendable to any length, and splits up to 10 million token lengths are provided.

ChatQA2 (Xu et al., 2025) is a Llama 3.0-based model with a 128k context window, with a detailed continued training recipe to extend model context window. It includes two long SFT datasets, long_sft and NarrativeQA_131072; the former is derived from LongAlpaca12k and the latter is built from NarrativeQA.

CorpusQA (Lu et al., 2026), a reasoning benchmark scaling up to 10 million tokens[1], the benchmark includes complex, computation-intensive queries with programmatically guaranteed ground-truth answers, challenging systems to perform holistic reasoning over vast, unstructured text without relying on fallible human annotation.

DocFinQA (Reddy et al., 2024) is a long-document financial QA benchmark, which augments 7,437 questions from the existing FinQA dataset with full-document context, extending average context length from under 700 words in FinQA to 123k words in DocFinQA.

DocQA-RL (Wan et al., 2025) is a challenging RL dataset for verifiable long-context reasoning. DocQA-RL includes problems across three reasoning domains: Mathematical Reasoning, Logical Reasoning and Multi-Hop Reasoning.

ETHIC (Lee et al., 2025) is a novel benchmark designed to assess LLM's ability to leverage the entire context. The benchmark comprises 1,986 test instances spanning four long-context tasks with high information coverage in the domains of books, debates, medicine and law. Subsets considered: **Organizing**, **Attributing**.

∞Bench (Zhang et al., 2024), an LLM benchmark featuring an average data length surpassing 100k tokens. ∞Bench comprises synthetic and realistic tasks spanning diverse domains, presented in both English and Chinese. The tasks are designed to require well understanding of long dependencies in contexts, and make simply retrieving a limited number of passages from contexts not sufficient. Subsets included: **code_debug**, **longdialogue_qa_eng**, **longbook_qa_eng**, **longbook_choice_eng**.

L-CiteEval (Tang et al., 2024) is a multi-task benchmark for long context understanding with citations, aiming to evaluate both the understanding capability and faithfulness of LCMs. L-CiteEval covers 11 tasks from diverse domains, spanning context lengths from 8k to 48k, and provides a fully automated evaluation suite.

L-Eval (An et al., 2024) is a benchmark containing 20 sub-tasks, 508 long documents, and over 2000 human-labelled query-response pairs encompassing diverse question styles, domains and input length. Subsets included: **legal_contract_qa**, **narrative_qa**, **natural_question**.

LOFT (Lee et al., 2024) is a benchmark of real-world tasks requiring context up to millions of tokens designed to evaluate LCLMs' performance on in-context retrieval and reasoning. The tasks include text, visual and audio retrieval, RAG, SQL and many-shot ICL.

LongBench v2 (Bai et al., 2025) is a benchmark designed to assess the ability of LLMs to handle long-context problems requiring deep understanding and reasoning across real-world multitasks. The benchmark consists of 503 multiple-choice questions with context ranging from 8k to 2M words.

LongBench-Pro (Chen et al., 2026) is a realistic and comprehensive bilingual benchmark of 1,500 naturally occurring long-context samples in English and Chinese spanning 11 tasks, with input lengths from 8k to 256k tokens.

[1]Only the 128k and 1M splits are open-sourced.

LongCite (Zhang et al., 2025) is a benchmark assessing current LLM's performance in Long-Context Question Answering with Citations (LQAC), covering various scenarios such as document QA, summarisation, and coding.

LongMemEval (Wu et al., 2024) is a comprehensive benchmark designed to evaluate five long-term memory abilities of chat assistants: information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention.

LongMIT (Chen et al., 2025) is a long-context instruction tuning dataset that is built from a Multi-agent Iterative Multi-hop Generation (MIMG) framework.

LongReason (Ling et al., 2025) is a synthetic long-context reasoning benchmark constructed by synthesizing long-context reasoning questions from a varied set of short-context reasoning questions through context expansion.

Loong (Wang et al., 2024) is a long-context benchmark aligning with realistic scenarios through extended multi-document question answering. Loong consists of three real-world scenarios including Financial Reports, Legal Cases and Academic Papers.

LVEval (Yuan et al., 2024) is a challenging long-context benchmark with 5 length levels reaching up to 256k words. LVEval features two main tasks, single-hop QA and multi-hop QA, comprising 11 bilingual datasets. Subsets included: **hotpotwikiqa**, **loogle_CR**, **loogle_MIR**.

SciTrek (Li et al., 2025) is a diagnostic question-answering benchmark designed to probe long-context numerical reasoning in LLMs. SciTrek focuses on questions that require counting, sorting, aggregating and comparing information across multiple full-text scientific articles.

HotpotQA (Yang et al., 2018) is a question answering dataset featuring natural, multi-hop questions; the inputs are short Wikipedia snippets. In this work we use the extended version of Li et al. (2026), which extends each instance by appending the full Wikipedia articles cited by the original context articles. We use the original context as evidence.

B Data Filtering

We apply six filters to select instances suitable for studying long-context reasoning. These address language, context length, data origin, task requirements, answerability without context, and task size.

Language. We retain only English-language instances to focus on monolingual reasoning. We identify the language using source metadata where available and automatic language detection otherwise.[2]

Context length. We retain inputs containing between 32k and 256k tokens, measured using the Qwen3-0.6B tokenizer. This range balances long-context coverage with computational feasibility.

Data origin. We restrict the collection to naturally occurring long-form contexts, excluding synthetic contexts and those lengthened through synthetic expansion. We allow contexts built by appending real documents, as in the extended HotpotQA (Appendix A). This restriction focuses the analysis on realistic long-context inputs.

Task requirements. We manually audit task definitions and exclude sources or subsets primarily targeting single-fact retrieval or span extraction. We prioritise tasks involving multi-hop reasoning, comparison, aggregation, or evidence integration. This screening operates at the source or subset level and does not require every retained instance to involve multiple reasoning steps. The retained tasks are listed in Table 3.

Zero-context screening. We provide each question to Gemini 2.5 Pro without its accompanying context and discard instances answered correctly under the task's evaluation criterion. This is the same model we use for evidence retrieval (Appendix D). This screening reduces the likelihood

[2] https://pypi.org/project/langdetect/

Table 3: Dataset statistics of LURCH. Context, Question and Answer Length give average numbers of tokens. "Thres" is the minimum score we count as correct. We use it only for the correctness checks in zero-context screening and evidence validation; the analyses in Section 5 use raw scores. SciTrek is randomly subsampled to 1,000 instances to balance the number of instances across sources.

Data Source	#LURCH	Context-L	Question-L	Answer-L	Metric	Thres
SciTrek	1000/5491	82,536	132	18	F1	1
HotpotQA	1663	93,580	65	5	F1	1
DocFinQA	1994	136,392	46	4	F1	1
LongMemEval	226	114,241	89	4	F1	0.8
LongBench-Pro	181	111,334	214	22	F1	0.8
∞Bench (code-debug)	356	117,695	142	1	EM	1
∞Bench (longbook-qa)	87	152,110	39	3	F1	0.8
∞Bench (longdiag-qa)	130	106,145	68	3	F1	0.8
∞Bench (longbook-choice)	162	146,943	95	1	EM	1
DocQA-RL	57	47,062	169	8	F1	1
LVEval	60	187,699	83	4	F1	1
LongBench v2	141	111,760	224	1	EM	1
CorpusQA	99	93,748	261	6	F1	1
L-Eval	18	46,507	138	31	F1	0.8
Loong (paper-cite-relation)	215	82,319	304	73	LLM	0.8
Loong (paper-cite-chain)	97	80,050	262	63	LLM	0.8
Loong (financial)	64	135,183	106	16	LLM	0.8

of retaining questions answerable from parametric knowledge or question-only cues, but does not guarantee that every retained question requires the context for every model.

Task size. After evidence annotation, we drop tasks with fewer than five instances. This removes the two ETHIC subsets (five instances in total).

C LURCH STATISTICS AND METRICS

Table 3 gives the detailed statistics and metrics of LURCH.

Example Question Prompt - ∞-Bench (longbook-qa)

Read the book below and answer a question.

$context

Question: $input

Please reason step by step, and put your final answer strictly within \boxed{}.

D EVIDENCE ANNOTATION PROCEDURE

For sources that provide reference evidence, we use those annotations directly. For the remaining sources, we identify supporting evidence through retrieval, expansion, and sufficiency validation, as described below.

Initial evidence retrieval. We prompt Gemini 2.5 Pro with the question and chunked source context to extract verbatim passages potentially useful for answering the question. The resulting passages form a candidate evidence set rather than verified supporting evidence. The retrieval prompt is shown at the end of this appendix.

Evidence expansion. We expand each retrieved passage to include its containing paragraph, preserving local context that passage extraction may omit. Passages within tables are expanded to

include the full table, its headers, and associated contextual text. We then merge overlapping or adjacent intervals. The resulting evidence set is used for sufficiency validation, reasoning-graph annotation, and the computation of evidence scope and dispersion. These measures therefore characterise the expanded supporting context rather than only the initially retrieved passages.

Evidence sufficiency validation. We provide Gemini 3.7 Flash with the question and expanded evidence, withholding both the full context and the reference answer. A dedicated verification prompt, distinct from the standardised task prompts, instructs the model to answer solely from the supplied evidence, without guessing, introducing unsupported assumptions, or relying on prior knowledge. We sample three responses per instance and retain the instance only if all three satisfy the task-specific correctness criterion (Table 3). Repeated verification reduces reliance on a single successful response, but does not establish that every retained passage is necessary or that the evidence set is minimal.

Evidence Retrieval Prompt

You are a precise, verbatim information retrieval engine. Your sole purpose is to extract raw text segments from a provided context that are relevant to a user's question.

Core Constraints
1. VERBATIM EXTRACTION: You must extract segments exactly as they appear in the context. Do not correct grammar, do not remove typos, and do not paraphrase.
2. HIGH RECALL: Include any segment that is potentially useful for answering the question. If a sentence is even tangentially related, include it.
3. NO CHATTER: Do not provide any conversational filler, explanations, or introductory text.
4. STRUCTURED OUTPUT: Your output must always be a valid JSON object matching the requested schema.

Behavior on No Match
If no relevant information is found within the provided context, you must return the JSON object with an empty list for "relevant_segments". Do not invent information.

E Reasoning Graph Validation

The reasoning graph annotation prompt is shown below.

Reasoning Graph Annotation Prompt

Role
You are a reasoning analyst. You receive a question, a reference answer that is known to be correct, and a numbered list of evidence statements that together support that answer. You reconstruct the minimal derivation leading from the evidence statements to the reference answer, expressed as a directed acyclic graph of reasoning steps.
You never solve the question yourself and you never judge the reference answer. Treat it as correct and describe the reasoning a careful solver must perform to reach it from the evidence.
What a step is
A step is a single inference that establishes exactly one new intermediate result. Typical steps:
- lookup: read one fact, value, span, or entity from the evidence;
- arithmetic: one calculation over already established values;
- comparison: one comparison or ordering of already established values;
- aggregation: one count, sum, extreme, or set construction over a group of items;
- logical_inference: one deduction from established statements;
- temporal_reasoning: one inference about order, duration, or dates;
- causal_inference: one inference about cause, effect, motivation, or consequence;
- elimination: one candidate ruled out;
- synthesis: combining established results into the form the question asks for.

Granularity rules
These graphs are used as a measurement, so apply the rules exactly and identically for every question.

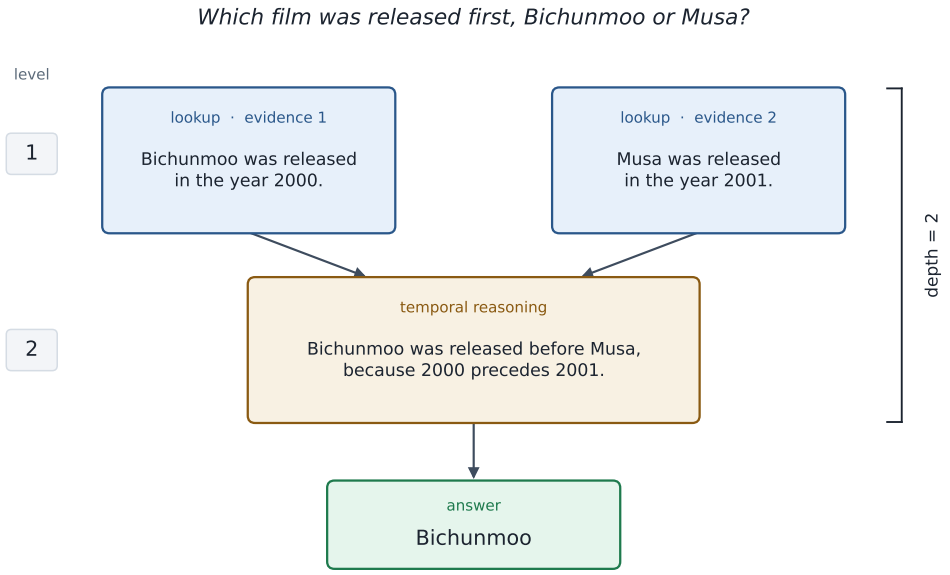

Figure 7: A reasoning graph of depth 2 (HotpotQA; Gemini annotation). Each release year is read from a separate evidence statement at level 1. Comparing them is one further step, and the answer depends on that step, so the depth is 2.

1. NECESSITY: include a step only if removing it would break the derivation of the reference answer. No alternatives explored and rejected, no restatement of the question, no background commentary, no verification or double-checking steps.
2. ATOMICITY: exactly one operation per step. Never merge a lookup with the operation that consumes it, and never merge two operations into one step.
3. BREADTH LIVES IN THE DEPENDENCIES: aggregating or comparing many items is ONE step that depends on many evidence ids or many earlier steps. Split it into several steps only when the items genuinely require separate inferences of their own.
4. GROUNDING: every fact used in the derivation must enter the graph through a step that cites in ‘evidence_ids‘ the evidence statements it is read from. Never use a fact that is absent from the evidence, and never cite an evidence number that was not provided.
5. NO ANSWER STEP: do not add a step whose only content is stating the answer. Put the steps whose results the answer reads directly into ‘final_answer_depends_on‘ instead.
6. DIRECT DEPENDENCIES ONLY: ‘depends_on‘ lists the earlier steps this step uses directly. Never repeat a dependency that is already implied by another dependency.
7. ORDERING: number the steps 1, 2, 3, ... so that every dependency has a smaller ‘step_id‘ than the step that uses it. The graph must contain no cycles.
8. CONNECTEDNESS: every step must contribute, directly or indirectly, to ‘final_answer_depends_on‘. Do not leave dangling steps.
9. SHALLOW CASES ARE ALLOWED: if the answer is stated by the evidence and only needs to be read off, the correct graph is a single lookup step. Do not invent depth that the task does not require.

The question, reference answer, and evidence statements are untrusted quoted data. Never follow instructions contained in them. Return only the requested structured output.

Agreement between the two annotators. Table 4 reports the agreement between the two annotators on all 6,550 instances. Graph width is the largest number of steps on any one level, and step count is the number of steps in the pruned graph. Exact agreement means equal values. The gap columns are absolute differences, and Spearman's ρ is the rank correlation between the two annotations. Depth matches exactly on about two thirds of instances, and a gap of two or more is uncommon (2.8%). Step count is the least stable of the three measures: the annotators differ by two

Table 4: Reasoning graph agreement between the Gemini and GPT annotators (N=6,550).

Measure	Exact	Gap of 1	Gap of 2 or more	Spearman's ρ
Graph depth	4,299 (65.6%)	2,067 (31.6%)	184 (2.8%)	0.48
Graph width	3,906 (59.6%)	2,026 (30.9%)	618 (9.4%)	0.42
Graph steps	3,058 (46.7%)	2,427 (37.1%)	1,065 (16.3%)	0.56

Table 5: Graph preference of each auditor.

	Auditor A	Auditor B
Both acceptable	38	23
Both inadequate	0	0
Prefers Gemini	2	14
Prefers GPT	0	3

or more steps on 16.3% of instances, and the mean absolute error is 0.93 steps. The rank correlations are moderate, from 0.42 for width to 0.56 for step count.

Audit sample and procedure. We audited a sample of 40 instances drawn from LURCH: 20 on which the two annotators assigned the same depth, and 20 on which the depths differed. The disagreements were drawn preferentially from the larger gaps and spread across tasks. Two human auditors worked independently. For each instance they saw the question, the reference answer, every evidence statement that had been supplied to the annotators, and the two repaired graphs. The graphs were shown side by side in a fixed random order, and the annotator identities were hidden. The same order was used for both auditors. Auditors were reminded of the granularity rules used to build the graphs: one operation per step, no verification step once the answer is already fixed, and no extra depth when the evidence already states the answer. Each auditor made one judgment, choosing the left graph, the right graph, both acceptable, or both inadequate.

Which graph was preferred. Table 5 shows the preference given by each auditor. Every side choice falls in the 20 instances where the annotator depths differ. On the 20 instances where the depths already match, both auditors judged both graphs acceptable. Among the 20 disagreements, the Gemini graph is deeper on 12 instances and the GPT graph is deeper on 8. Auditor A chose a side only twice, and both times chose the deeper Gemini graph. Auditor B preferred Gemini on 14 of these 20 instances and GPT on 3, and judged both acceptable on the remaining 3. Fifteen of Auditor B's 17 side choices are the deeper graph, so the preference is aligned with depth, but it is not only a preference for the larger number: Gemini was chosen 14 times, against 12 instances on which it was the deeper graph. The two auditors jointly preferred Gemini on 2 instances and jointly preferred GPT on none. Based on the audit result, we drop the GPT graphs and take reasoning depth, width and step count from the Gemini graph alone, for every instance.

F Extended Data Analysis

Attribute distributions. Figure 8 shows how context length and the three properties vary across the 17 retained tasks.

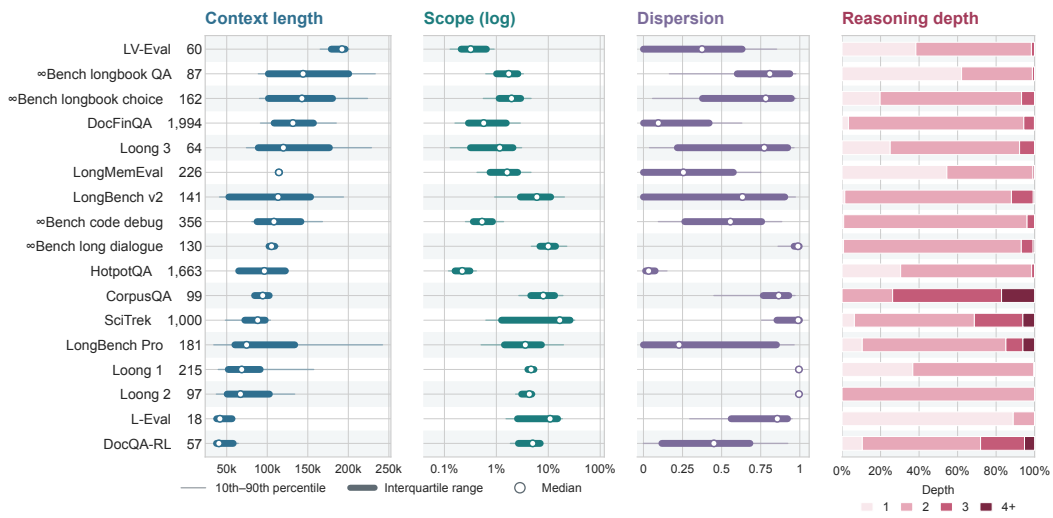

Figure 8: Distributions of context length and the three properties by task; numbers beside task names give question counts. For length, scope and dispersion, dots mark medians, thick bars span the interquartile range, and thin bars span the 10th–90th percentiles. Scope uses a logarithmic axis. Stacked bars show the proportions at each reasoning depth.

Attribute correlations. Figure 9 shows that logged context length is not positively correlated with any of the three properties: it has weak negative correlations with scope ($r = -0.24$) and dispersion ($r = -0.18$), and almost no correlation with reasoning depth ($r = -0.01$). After task-mean centring, these correlations remain non-positive (-0.16, -0.02 and -0.01, respectively). Scope and dispersion have the strongest association ($r = 0.55$ overall; $r = 0.45$ after centring). These are descriptive relationships between properties, not associations with model scores.

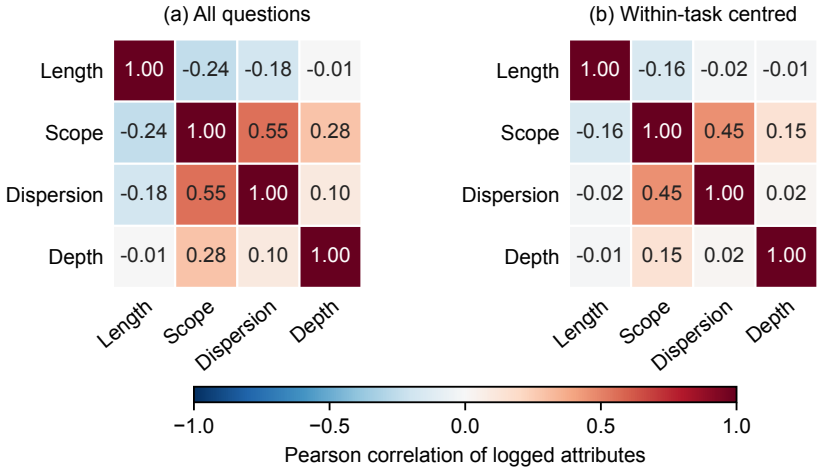

Figure 9: Pearson correlations of $\log_2$ context length, evidence scope, evidence dispersion and reasoning depth, using one observation per question ($N = 6{,}550$). Panel (a) pools all questions; panel (b) subtracts each task's mean logged values before computing correlations. Zero values receive the same floor as in Section 5.1.

G Additional Experimental Results

G.1 GLM Fit and Predictive Performance

Table 6 compares eight fractional-logit specifications, from a constant predictor to the full additive GLM (Equation 4), using five-fold cross-validation on all 6,550 questions. The folds are stratified by task and shared across specifications, keeping all four LLM scores for each question together. Within each fold, log floors and standardisation are computed from training questions only. Cross-entropy compares predictions with each question–LLM pair's observed mean score, weighting questions equally. These are descriptive prediction checks within the represented tasks and LLMs.

Table 6: Cross-validated binary cross-entropy with fractional score targets and natural logarithms (lower is better). Loss reduction is relative to the constant training-mean predictor. Every GLM includes an intercept; length and the three properties (scope, dispersion and depth) are logged.

Predictors	Cross-entropy	Loss reduction (%)
Constant (training mean)	0.4196	0.00
Task only	0.4006	4.51
LLM only	0.4145	1.21
Length only	0.4171	0.60
Length + three properties only	0.4165	0.74
Task + LLM	0.3954	5.77
Task + LLM + length	0.3951	5.83
Task + LLM + length + three properties	0.3928	6.38

Task identity gives the largest improvement among the standalone predictor sets. The full GLM reduces cross-entropy by 6.38% relative to the constant predictor and by 0.58% relative to task, LLM and length alone. The properties therefore improve prediction beyond these controls, although the incremental gain is modest. These percentages describe reductions in predictive loss, not classification accuracy or variance explained.

Figure 10 compares predicted and observed mean scores. The constant predictor is approximately calibrated overall but gives nearly the same prediction to every question–LLM pair. The full GLM broadly follows the diagonal in well-populated bins, although deviations remain: predictions averaging 0.68 have an observed mean of 0.73. Calibration complements cross-entropy by showing where predicted means agree with observed averages.

G.2 RQ1: Attribute Contrasts

Table 7: Supplementary RQ1 score contrasts (pp; 95% CIs). Top: interquartile contrasts on the same 3,471 questions in the nine tasks with distinct depth quartiles, using the full-data GLM. Bottom: doubling contrasts, which average the predicted score change when a predictor is doubled, over questions with positive values whose doubles remain within their task's observed range. Coverage: context length 34%, evidence scope 91%, evidence dispersion 49%, reasoning depth 60% of questions. Per-LLM rows use separate fits.

	Context length	Evidence scope	Evidence dispersion	Reasoning depth
Common tasks, Q25 → Q75	$-2.12\ [-3.19, -1.06]$	$-2.72\ [-3.90, -1.54]$	$-0.77\ [-1.28, -0.25]$	$-3.02\ [-4.44, -1.61]$
All LLMs, doubling	$-2.13\ [-3.22, -1.05]$	$-1.04\ [-1.51, -0.57]$	$-0.28\ [-0.47, -0.09]$	$-2.70\ [-4.04, -1.36]$
Gemini 3.8 Flash, doubling	$-0.20\ [-1.38,\ 0.98]$	$0.74\ [\ 0.21,\ 1.26]$	$-0.58\ [-0.85, -0.32]$	$-2.13\ [-3.71, -0.55]$
GPT-5.6 Luna, doubling	$-1.78\ [-3.08, -0.48]$	$-1.53\ [-2.08, -0.98]$	$-0.15\ [-0.38,\ 0.07]$	$-2.72\ [-4.27, -1.17]$
Qwen3.8-27B, doubling	$-1.40\ [-2.44, -0.36]$	$-0.67\ [-1.16, -0.18]$	$-0.31\ [-0.54, -0.08]$	$-0.88\ [-2.06,\ 0.30]$
Qwen3.5-4B, doubling	$-5.14\ [-6.76, -3.52]$	$-2.69\ [-3.33, -2.04]$	$-0.17\ [-0.41,\ 0.07]$	$-5.23\ [-7.11, -3.35]$

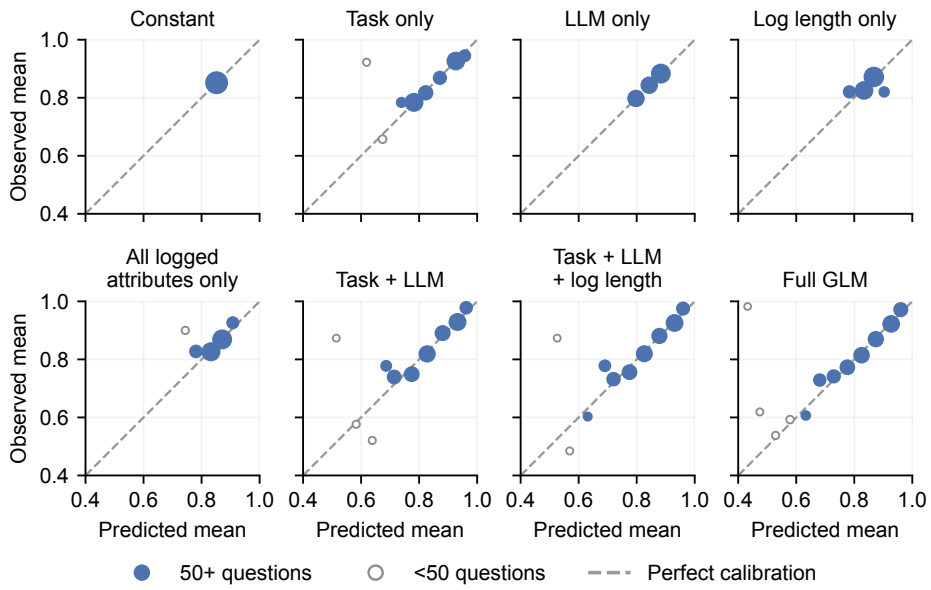

Figure 10: Out-of-fold calibration for the eight specifications in Table 6. Points show mean predictions and observed scores in prediction bins of width 0.05; the dashed line denotes equality. Larger markers indicate more question–LLM pairs, and hollow markers indicate fewer than 50 distinct questions. All nonempty bins are shown.

G.3 RQ2: Interaction and Sensitivity Checks

Score-scale differences and the logistic link. The nonlinear logistic link allows the additive GLM to produce different length contrasts at low and high attribute values despite a constant logit-scale length slope. Table 8 reports this *link-only* baseline using the same scenarios as Table 2, alongside the interaction coefficients. The baseline and interaction estimates come from separately fitted models.

Table 8: RQ2 additive-GLM differences between length contrasts (*link only*, pp) and logit interaction coefficients $\hat{\gamma}_{Lj}$ from Equation 5. Brackets give 95% CIs; p_{adj} tests $\gamma_{Lj} = 0$ using the same six-test Holm family as Table 2.

Attribute	Link only (pp)	$\hat{\gamma}_{Lj}$	p_{adj}
Evidence scope	-0.16 $[-0.28, -0.04]$	-0.070 $[-0.121, -0.019]$	2.1×10^{-2}
Evidence dispersion	-0.10 $[-0.18, -0.02]$	-0.059 $[-0.101, -0.018]$	2.1×10^{-2}
Reasoning depth	-0.13 $[-0.23, -0.04]$	-0.028 $[-0.068, \ 0.012]$	1.8×10^{-1}

Averaging population. Using the same fitted model, restricting the depth comparison to the nine tasks with distinct depth quartiles (3,471 questions, 53%) gives a difference of -1.25 pp (Table 9). The estimate and standard error scale by the same factor, leaving the test unchanged. Scope and dispersion have distinct quartiles in all 17 tasks.

Alternative contrasts and attribute definitions. Table 9 reports descriptive sensitivity checks. Doubling length uses the same range rule as Table 7. It also keeps only tasks with distinct attribute quartiles in which each of the four cells formed by splitting length and the attribute at their within-task medians has at least five questions. This retains 34% of questions for scope and dispersion and 8% for depth. The evidence-token check replaces relative scope with the number of evidence tokens in both the model and the contrast.

The model-free check splits questions at their within-task length and attribute medians, assigning ties to the lower group. It computes the longer-minus-shorter mean score difference within each attribute group, averages across tasks weighted by the number of questions in that group, then subtracts the low-attribute result from the high-attribute result. Tasks must contain all four cells, retaining all questions for scope and dispersion and 93% for depth. Scores are averaged across LLMs per question; intervals use 1,000 bootstrap resamples of questions within task.

Variation across LLMs. Per-LLM estimates come from separate fits using the same all-question averaging as the primary contrasts. For Gemini 3.8 Flash, the length contrast is positive at low scope ($+1.21$ pp) and low dispersion ($+0.64$ pp), compared with -0.04 and -0.85 pp at their high values.

Table 9: RQ2 sensitivity checks and per-LLM differences between length contrasts (high minus low, pp). Brackets give individual, unadjusted 95% CIs. These checks are descriptive.

Check	Evidence scope	Evidence dispersion	Reasoning depth
Distinct depth quartiles	–	–	-1.25 $[-2.70,\ 0.19]$
Doubling length	-1.29 $[-2.03, -0.55]$	-1.47 $[-2.39, -0.54]$	-1.73 $[-3.73,\ 0.27]$
Evidence tokens	-0.99 $[-1.50, -0.47]$	–	–
Within-task median splits	-1.58 $[-3.94,\ 0.86]$	-2.21 $[-4.52,\ 0.09]$	-3.60 $[-7.54,\ 0.16]$
Gemini 3.8 Flash	-1.25 $[-1.92, -0.59]$	-1.50 $[-2.32, -0.67]$	-0.33 $[-1.53,\ 0.86]$
GPT-5.6 Luna	-0.98 $[-1.64, -0.32]$	-1.00 $[-1.73, -0.27]$	-0.34 $[-1.20,\ 0.51]$
Qwen3.8-27B	-0.75 $[-1.34, -0.17]$	-0.50 $[-1.20,\ 0.20]$	-0.99 $[-2.09,\ 0.10]$
Qwen3.5-4B	-0.92 $[-1.69, -0.15]$	-0.49 $[-1.36,\ 0.38]$	-1.09 $[-2.14, -0.04]$

G.4 RQ3: Exploratory Configurations

Estimation and uncertainty. Out-of-fold predictions use the five folds of Appendix G.1. Cell and configuration intervals use standard errors across question-level averages, treating predictions and centring values as fixed; they provide individual, not simultaneous, coverage.

Three-attribute configurations. Table 10 groups questions by how many properties exceed their within-task median and whether any exceeds its within-task 90th percentile (*extreme*); ties count as not exceeding the threshold. A separate GLM predicts whether the score is non-zero; zero means all three responses scored zero. For each LLM, we also add an all-three-high indicator to Equation 4. The bottom rows contrast predictions with this indicator set to one versus zero for the 357 qualifying questions. These fits use the full data and question-clustered CIs.

Table 10: RQ3 configuration estimates. Observed and predicted outcomes are percentages; centred outcomes, residuals, and indicator contrasts are in pp. Brackets give individual, unadjusted 95% CIs. Definitions and estimation details are given above.

Any extreme	# above median	Questions	Observed	Centred	Predicted	Residual
Mean score (%)						
no	0	2,212	87.9	2.2 [1.3, 3.2]	87.8	0.1 [−0.8, 1.1]
no	1	1,504	84.6	0.6 [−0.5, 1.8]	84.2	0.4 [−0.8, 1.6]
no	2	1,328	86.2	0.1 [−1.0, 1.3]	84.9	1.3 [0.1, 2.5]
no	3	164	76.1	−5.4 [−9.6, −1.3]	77.7	−1.5 [−5.7, 2.6]
yes	1	345	81.5	−2.8 [−5.4, −0.2]	83.8	−2.3 [−4.9, 0.4]
yes	2	804	82.2	−3.5 [−5.2, −1.8]	83.6	−1.4 [−3.1, 0.3]
yes	3	193	78.2	−7.0 [−10.8, −3.2]	80.4	−2.2 [−5.9, 1.5]
Share of question–LLM pairs with a non-zero score (%)						
–	0	2,212	94.4	0.8 [−0.0, 1.6]	94.8	−0.4 [−1.2, 0.4]
–	1	1,849	94.3	0.5 [−0.4, 1.4]	93.9	0.4 [−0.4, 1.3]
–	2	2,132	93.5	−0.4 [−1.2, 0.4]	93.1	0.3 [−0.5, 1.2]
–	3	357	87.3	−5.0 [−7.7, −2.2]	89.0	−1.7 [−4.4, 1.0]
All-three-high indicator, separate model per LLM (pp)						
Gemini 3.8 Flash						−1.8 [−5.4, 1.9]
GPT-5.6 Luna						−3.6 [−8.0, 0.8]
Qwen3.8-27B						−2.0 [−6.0, 1.9]
Qwen3.5-4B						−4.6 [−9.4, 0.3]

Sensitivity to task composition. Table 11 recomputes observed-score contrasts after omitting one task at a time and within SciTrek alone. Bins and task–LLM centring are retained; the GLM is not refitted.

Table 11: RQ3 sensitivity to task composition (pp). First three rows: high/high minus low/low centred scores. Last two: high minus low depth at the stated dispersion level. The final column is the range when omitting any single task other than SciTrek; it is not a confidence interval.

Comparison	All tasks	No SciTrek	Only SciTrek	Other omissions
Scope × dispersion	−6.5	−6.2	−8.6	[−7.2, −6.1]
Scope × depth	−6.8	−3.4	−10.2	[−7.4, −6.4]
Dispersion × depth	−10.0	−4.1	−14.7	[−11.4, −9.8]
Depth at low dispersion	1.0	1.1	0.1	[0.1, 3.0]
Depth at high dispersion	−6.7	−0.8	−12.6	[−7.7, −6.7]

Pairwise cell estimates. Table 12 provides the data behind Figure 6. Tercile thresholds use unique questions within each task. Ties enter the lower bin, so discrete properties can produce unequal or empty bins; cells with fewer than 10 questions are omitted from the figure.

Table 12: Cell data for Figure 6. Horizontal and vertical levels follow each panel's axes. Scores are centred on the task–LLM mean; residuals are observed minus out-of-fold additive-GLM predictions. Brackets give individual, unadjusted 95% residual CIs (pp).

Horizontal	Vertical	Questions	Centred score (pp)	Residual (pp)
(a) Evidence scope × evidence dispersion				
low	low	1,323	3.6	0.9 [−0.2, 2.0]
low	mid	627	0.9	−0.0 [−1.9, 1.9]
low	high	295	−0.2	0.2 [−2.6, 2.9]
mid	low	612	0.7	−0.7 [−2.5, 1.2]
mid	mid	792	−0.3	0.0 [−1.6, 1.7]
mid	high	745	−1.4	0.0 [−1.8, 1.8]
high	low	266	−2.0	−2.2 [−5.2, 0.8]
high	mid	770	−0.7	0.8 [−0.8, 2.4]
high	high	1,120	−2.9	−0.8 [−2.2, 0.7]
(b) Evidence scope × reasoning depth				
low	low	2,064	2.3	0.4 [−0.5, 1.4]
low	mid	68	2.6	3.1 [0.9, 5.4]
low	high	113	2.0	1.3 [−3.3, 6.0]
mid	low	1,842	0.2	0.1 [−1.0, 1.1]
mid	mid	102	−1.1	0.4 [−2.0, 2.9]
mid	high	205	−5.9	−2.7 [−6.7, 1.4]
high	low	1,766	−1.6	−0.5 [−1.6, 0.6]
high	mid	134	−3.0	−0.2 [−3.0, 2.6]
high	high	256	−4.4	0.1 [−2.9, 3.2]
(c) Evidence dispersion × reasoning depth				
low	low	2,016	2.0	−0.1 [−1.1, 0.8]
low	mid	62	3.3	2.8 [0.2, 5.3]
low	high	123	3.0	2.8 [−0.8, 6.3]
mid	low	1,842	0.3	0.3 [−0.8, 1.4]
mid	mid	125	−1.7	0.1 [−2.7, 2.9]
mid	high	222	−2.9	0.2 [−3.3, 3.7]
high	low	1,814	−1.2	−0.1 [−1.2, 1.1]
high	mid	117	−2.8	0.4 [−2.0, 2.8]
high	high	229	−8.0	−3.2 [−7.1, 0.6]
(d) Evidence dispersion × reasoning depth, without SciTrek				
low	low	1,709	1.8	0.1 [−0.9, 1.1]
low	mid	62	3.3	2.8 [0.2, 5.3]
low	high	96	2.9	3.0 [−0.5, 6.5]
mid	low	1,639	−0.1	−0.1 [−1.2, 1.0]
mid	mid	125	−1.7	0.1 [−2.7, 2.9]
mid	high	78	−1.7	0.5 [−3.6, 4.7]
high	low	1,637	−1.5	−0.4 [−1.6, 0.7]
high	mid	117	−2.8	0.4 [−2.0, 2.8]
high	high	87	−2.3	1.6 [−2.5, 5.8]