CertiHop: Certified Necessary-Evidence Supervision for Long-Context Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Long-context question answering benchmarks often annotate supporting evidence, yet support is weaker than necessity: a model may answer correctly without combining every designated item. This gap grows when a compact proof is expanded into long documents, where redundancy, alternative paths, answer leakage, and parametric knowledge can reduce effective dependency order. We introduce CertiHop, a construction and certification framework for long-context necessary-evidence reasoning. Each example begins with an executable dependency tree, expands into a multi-document dossier, and is certified only on the final context. Full must uniquely determine the answer, whereas Zero and every required-leaf leave-one-out intervention must preserve the gold answer as possible but not uniquely determine it. The resulting family supplies paired sufficient/insufficient supervision and parent-level Joint Irreducibility (JI); non-critical interventions distinguish selective sensitivity from generic abstention. Across four Qwen-series backbones, CertiHop yields LongBench-3 point estimates 11.92–36.25 points above equal-size continued-normal training. The controlled A3 comparison yields positive LongBench-3 differences at all four scales. A 9B RULER decline exposes a checklist-completeness boundary. CertiHop turns intended graph dependency into an auditable property of the final text.

1 Introduction

Modern language models accept tens or hundreds of thousands of tokens, but window size is an interface property whereas reliable use of distant information is a behavioral capability. Performance can vary sharply with evidence position, context length, and lexical overlap (Liu et al., 2024; Bai et al., 2023; Hsieh et al., 2024; Modarressi et al., 2025). Multi-hop datasets address a related challenge by associating an answer with several supporting facts (Yang et al., 2018; Trivedi et al., 2022). Still, answer accuracy establishes only that some route to the answer was available. It does not identify which annotated facts the model had to combine.

This ambiguity is not hypothetical. A nominal four-hop question may have an effective dependency order of one when a paragraph directly reveals the answer, a marked fact is redundant, or a pretrained model answers from memory. Prior analyses show that compositional questions and HotpotQA examples admit single-hop and lexical shortcuts (Min et al., 2019; Jiang & Bansal, 2019). MuSiQue responds by composing single-hop questions and filtering edges whose successor can be answered without its predecessor (Trivedi et al., 2022). Counterfactual and pseudo-evidentiality methods further use evidence interventions to diagnose or reduce disconnected reasoning (Lee et al., 2021; Guo et al., 2023). These results establish the problem and make a broad claim of first counterfactual multi-hop data untenable.

Long-context synthesis creates a second failure boundary. Recent methods improve supervision through faithful reasoning, multi-agent generation, concept graphs, evidence graphs, trajectory compilation, or model-in-the-loop dependency screening (Yang et al., 2025a; Chen et al., 2025; Sun et al., 2026; Su et al., 2026; Jia et al., 2025; Yuan et al., 2026). These advances make generated tasks more grounded, diverse, or traceable. But a dependency that

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.

appears valid in a compact question graph can disappear when its facts are rewritten and embedded in a long dossier: the realization step can add semantic redundancy, an answer-bearing paraphrase, or an unenumerated alternative path. We call this the post-realization necessity gap. The decisive question is therefore counterfactual and must be asked on the model's actual input: would the answer cease to be uniquely determined if this evidence item were neutralized while everything else remained comparable?

We introduce CertiHop, a framework that makes post-realization necessity a release criterion. It first constructs an explicit dependency tree whose leaves are atomic evidence statements and whose internal nodes are executable composition operations. It then embeds the proof core in a long, topically coherent multi-document dossier. Only after this dossier is complete does CertiHopevaluate the complete context, a zero-evidence intervention, and every per-leaf leave-one-out intervention. A parent example is certified only when the complete context yields the gold answer and every designated-evidence intervention prevents a unique answer. In-place neutral replacements preserve document count, evidence position, and input length, limiting trivial condition cues.

Our contribution is not another taxonomy of increasingly complicated trees. The tree is an instrument for controlling a scientific variable; the intervention family is the unit of release, learning, and evaluation. CertiHop-Train provides matched sufficient and insufficient long-context supervision. CertiHop-Bench measures parent-level Joint Irreducibility, requiring correct behavior on the complete context and the entire associated intervention family. Matched non-critical interventions test whether a model merely abstains after any perturbation, and public long-context benchmarks test transfer beyond our controlled format.

Our contributions are:

- We identify the post-realization necessity gap and formalize the distinction among supporting evidence, intended dependency at construction time, and certified dependency on the final long-text surface.
- We introduce a scalable tree-controlled pipeline whose release unit is an exhaustive, surface-matched complete/zero/all-leaf-LOO intervention family certified after long-context realization.
- We reuse this family for paired sufficient/insufficient SFT and parent-level Joint Irreducibility evaluation, with non-critical interventions guarding against generic abstention.
- We propose matched controls that separate necessary-dependency supervision from long-context exposure, generation quality, and certification-based selection while holding output protocol fixed, and evaluate capacity and dense-to-MoE transfer across four Qwen-series backbones together with frozen external benchmarks.

2 Problem Formulation

Let q be a question, a its intended answer, and C a long context containing distractor content D and a set of designated atomic evidence leaves $E = \{e_1, \ldots, e_m\}$. A reasoning structure T specifies how the leaves compose into a. We use $\mathcal{A}(q, C)$ to denote the set of answers that remain compatible with the information explicitly available in C. The answer is uniquely determined when $\mathcal{A}(q, C) = \{a\}$.

Supporting versus necessary evidence. A set $S \subseteq E$ supports a if the answer can be derived from $D \cup S$. It is a minimal determining set if $D \cup S$ uniquely determines a but $D \cup (S \setminus \{e\})$ does not for every $e \in S$. This second condition is the scientific target of this work. It prevents a nominal m-hop label from being satisfied by a shorter route that ignores one or more annotated facts.

Intervention family. The Full context contains every required leaf; Zero neutralizes the proof core; and C^{-i} replaces only e_i with position- and form-matched non-entailing text.

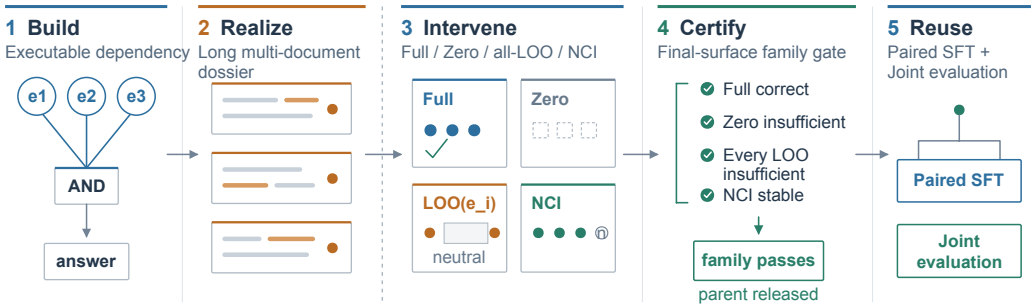

Figure 1: The CertiHop pipeline. A compact dependency specifies construction intent, but long-form realization may leak or duplicate an answer path. We therefore construct a matched intervention family on the finalized surface and release a parent only when Full, Zero, every required-leaf LOO, and non-critical interventions jointly satisfy the certificate. The same certified family supplies paired SFT and parent-level evaluation.

Certification requires

$$\mathcal{A}(q, C^{\text{full}}) = \{a\}, \qquad \substack{a\in\mathcal{A}(q,C^{\text{zero}}),\\ |\mathcal{A}(q,C^{\text{zero}})|>1}, \qquad \forall i, \ \substack{a\in\mathcal{A}(q,C^{-i}),\\ |\mathcal{A}(q,C^{-i})|>1}. \tag{1}$$

These conditions exclude the case in which an intervention uniquely implies a different answer. They operationalize necessity rather than claim a model-independent proof: answer determinacy is approximated by executable checks, a verifier separated from the generator, and targeted heterogeneous audits.

Valid interventions. An unsuccessful answer under C^{-i} is evidence of missing information only when the replacement is neutral. Valid neutralization is monotone with respect to the gold completion: it removes a constraint without contradicting a, so a remains compatible. It must entail neither the removed proposition nor its negation, add no alternative answer route, preserve non-target facts, hide protected literals, and match the original position and surface form. These properties are checked independently of the student's prediction; contradictory replacements are excluded from the certificate.

Alternative valid paths. When multiple minimal evidence sets derive the answer, we represent an AND/OR hyperedge family rather than declare every leaf globally necessary. The controlled split prioritizes single minimal families; multi-path cases are retained only when every intended path can be enumerated and tested.

Operational certificate. Let $v(q, C, a) \in \{0, 1\}$ indicate whether the declared audit stack judges a uniquely determined by C, and let $u(q, C, a) \in \{0, 1\}$ indicate that it judges a compatible but non-unique. The machine certificate for parent p is

$$\Gamma(p) = v(q, C^{\text{full}}, a)\, u(q, C^{\text{zero}}, a) \prod_{i=1}^{m} u(q, C^{-i}, a). \tag{2}$$

We admit p to the machine-certified release candidate only when $\Gamma(p) = 1$ and all structural, leakage, and tokenizer checks pass. Under valid interventions and exact v, u, this is equivalent to E being a minimal determining set relative to D (Appendix E). "Machine-certified" means certified under the declared protocol, not that no conceivable external reasoner could recover a; the final stratified human/heterogeneous audit is reported separately (Appendix H).

3 The CertiHop Framework

3.1 Atomic evidence and executable dependency trees

We obtain seed facts from publicly available datasets and controlled synthetic worlds. Each fact is normalized into an atomic statement with provenance, answer-bearing literals, and an explicit semantic role. A dependency tree connects these leaves through typed operations such as entity resolution, temporal ordering, comparison, arithmetic aggregation, set composition, and Boolean conjunction. Every internal node is executable or independently checkable. Trees that contain disconnected leaves, unknown references, circular dependencies, or direct answer leakage are rejected before language generation.

3.2 From proof cores to long dossiers

The compact proof core is expanded into a topically coherent dossier containing multiple documents. Evidence leaves are distributed across documents and positions, while distractors are selected to be locally plausible but non-entailing. The construction records absolute evidence positions, maximum pairwise span, document dispersion, tree depth, number of required leaves, and answer type. Training contexts cover multiple length buckets rather than padding every example to a single length.

Expansion is a high-risk stage: a fluent paragraph can restate the answer, duplicate a required fact, or introduce an alternative route. We therefore treat the compact tree as a construction specification, not as a certificate. Necessity is re-evaluated after the final dossier is produced.

3.3 Matched intervention family

For each parent problem, CertiHop constructs one Full surface, one Zero surface, and one LOO(e_i) surface per designated leaf. Missing evidence is replaced in place rather than simply deleted. Replacement sentences match discourse role and approximate length but do not entail the removed proposition. Document identifiers and explicit condition labels are hidden from the student model. We additionally test non-critical intervention invariance on audited subsets to ensure the model is not rewarded for refusing whenever any text changes.

Appendix F visualizes one six-leaf tree and its matched final-surface intervention.

3.4 Final-surface certification

Certification combines structural checks for tree connectivity and leaf use; textual checks for leakage, duplication, and template concentration; Full/Zero/all-LOO behavioral checks by a verifier distinct from the generator; and targeted heterogeneous audits that quarantine unresolved disagreements. A request, candidate, or partially verified parent is never called Gold: machine admission occurs only after its complete intervention family passes every applicable automated gate, while the final release label additionally requires the preregistered audit in Appendix H.

3.5 Type-specific determinacy and negative-space controls

Irreducibility is defined over task semantics rather than a universal deletion heuristic. Each answer type has an executable determinacy contract: when two completions consistent with the visible dossier yield different answers, the target is insufficient. Deterministic canonicalization handles Boolean aliases, dates, numbers, and ordered or unordered lists before any model-based adjudication. To reject the policy "refuse whenever text changes," two preregistered non-critical replacements preserve every required leaf and must retain the original answer. Required-leaf LOO measures sensitivity; NCI measures selectivity.

Table 1: Frozen experimental release candidate. SFT exports one Full and one deterministic representative Missing surface per parent; exhaustive evaluation retains Full, Zero, every leaf-level LOO, and two non-critical interventions.

Statistic	Train	Validation	Test
Machine-certified parents	8,984	1,123	1,123
Canonical SFT rows	17,968	2,246	2,246
Exhaustive-eval parents	–	200	200
Exhaustive-eval surfaces	–	2,420	2,434

3.6 End-to-end construction algorithm

The pipeline is intentionally asymmetric: high-throughput models may propose text, but no generator approves its own output. We normalize source facts, construct and execute the compact tree, screen its complete intervention family, naturalize and expand it into a long dossier, rebuild matched interventions, certify the final surface with a separate verifier, and only then group-split and export. Any later surface edit invalidates the certificate. The appendix records the certification flow, audit roles, model registry, and reproducibility fields available for review.

3.7 SFT construction

Every training parent has first passed the exhaustive Full/Zero/all-LOO certificate. The primary 1:1 SFT budget then exports one Full surface and exactly one representative LOO surface per parent; the latter is selected by the minimum SHA-256 digest of the fixed sampling seed, parent ID, and evidence ID. Selection never uses verifier confidence, student predictions, or benchmark scores, and the selected leaf role is retained in the source manifest for audit. All interventions remain available for certification, while the exhaustive family is evaluation-only. Sufficient targets contain the exact answer in the requested task format; insufficient targets state that the evidence does not uniquely determine an answer. Output protocol is treated as a controlled variable rather than part of the method: every causal comparison assigns it identically by parent. We never attribute a benchmark gain to necessity supervision when only the output format differs.

4 The CertiHop Data

We instantiate the framework as two parent-disjoint resources. CertiHop-Train contains paired sufficient and insufficient examples for SFT. CertiHop-Bench preserves every intervention for causal evaluation. The parent, rather than an individual surface, is the unit of splitting: all complete, zero, leave-one-out, and non-critical variants of one dependency structure remain in exactly one split.

The frozen experimental release candidate contains 11,230 machine-certified parents and 22,460 canonical Full/representative-Missing rows. Its parent-grouped split contains 8,984/1,123/1,123 parents for train/validation/test, with no parent overlap. Exact Qwen3.5-27B tokenization places all canonical training prompts below 32,768 tokens (7,497/10,212/27,457 minimum/median/maximum), and the maximum within-parent Full/Missing prompt-length difference is 0.2829%. Dependency order spans 4–12, proof depth 2–6, and dossiers contain 4–6 documents; Appendix G gives exact distributions. Questions and Full prompts are exact-unique; normalized-text and semantic-similarity audits find no overlap between the training split and the frozen external evaluation files. Table 1 distinguishes the canonical SFT projection from the exhaustive evaluation family.

The exhaustive validation and test exports contain 200 parents each and 4,854 surfaces in total. Every family is checked for the expected number of conditions before scoring; an incomplete family is marked incomplete rather than entering the JI denominator. Technical failures (missing, malformed, or truncated provider responses) remain distinct from complete

semantic failures. A complete parent that violates Full, Zero, or any required-leaf LOO condition is rejected and is never repeatedly sampled until it passes.

For A1, the formal training split yields two exposure-matched 19,968-row arms: 17,968 CertiHop rows plus the same 2,000 LongAlign replay rows at identical schedule positions. Full-only repeats each certified Full surface, whereas Paired uses one Full and one deterministically selected certified Missing surface per parent. Output protocol is assigned identically by parent, with 75% direct-answer and 25% structured targets. At global batch 16, each run uses 1,248 optimizer updates. Controlled worlds use fictionalized identifiers and values to limit parametric recall; source-derived examples preserve provenance and are screened against internal and external evaluation questions.

5 Experimental Design

5.1 Models

We evaluate pinned Qwen3.5-4B/9B/27B and Qwen3.6-35B-A3B snapshots (Qwen Team, 2026a;b) with identical examples, Stage-2 arms, seeds, and evaluation inputs. The dense models provide a within-family capacity curve; the 35B-total, approximately 3B-active sparse MoE tests architecture transfer. This matrix supports neither cross-vendor nor cross-family claims. All reported training runs use full-parameter SFT, with optimization settings held fixed within each comparison. Training ran on Huawei accelerators through the TorchTurbo/ms-swift stack; exact accelerator model/count, CANN and framework versions, and resolved optimizer fields are accepted only from returned run manifests rather than reconstructed from historical launch files (Appendix I).

5.2 Baselines and matched budgets

For each backbone and seed, we train on the same 4,000 LongAlign examples (Bai et al., 2024a) and use the resulting full-parameter Stage-1 checkpoint. Stage 2 branches from that checkpoint to 2,000 held-out LongAlign (continued-normal), 2,000 official LongFaith-SFT (Yang et al., 2025a), or 2,000 CertiHop examples; every arm also receives the same 1,500-example replay. Formal seeds are 42, 2026, and 3407. Base and Stage 1 are diagnostic anchors; the primary effect is CertiHop minus continued-normal. CertiHop–LongFaith is reported only after a shared generation-budget gate. Training and evaluation factors are fixed within each valid comparison.

The official LongAlign file contains 9,888 observed records despite its 10K name. We retain only examples below 32K for every target tokenizer and a Qwen2.5 compatibility tokenizer. The deterministic 2K LongAlign and CertiHop subsets differ by 0.198% in reference-tokenizer prompt tokens. The selected 2K official LongFaith prompts are 82.22% shorter, so LongFaith is sample-matched but not a token-matched causal control; continued-normal is the tighter exposure control.

5.3 Controlled causal ablations

Three ablations use all four backbones and seeds 42, 2026, and 3407. A1 compares Full-only with Full+Missing on 8,984 machine-certified parents (19,968 rows/arm; 1,248 updates). A2 compares final-surface machine-certified data with same-batch candidates that pass structural, textual, completeness, length, and Full-correctness gates but fail Zero or at least one LOO. Each arm has 1,552 parents: 704/408/240/200 have 6/8/10/12 leaves, so complete Full+Zero+all-LOO families yield $704 \times 8 + 408 \times 10 + 240 \times 12 + 200 \times 14 = 15{,}392$ core rows; 2,000 identical replay rows give 17,392 total (1,087 updates). A3 reuses each dossier byte-for-byte while changing whether the question requires one leaf or multiple leaves (6,144 parents, 14,288 rows/arm; 893 updates). Training factors are fixed within each ablation.

A2 arms preserve complete intervention families; no failed leaf is post-hoc selected. Parents are matched without replacement exactly on batch, requested length, answer type, leaf count, and depth, then by token count, document count, and surface-quality features. The

maximum continuous-covariate standardized mean difference is 0.0334. Technical failures and Full-incorrect candidates are excluded.

The equal-quality control reuses the dependency target's final dossier byte-for-byte and changes only the question–answer contract: a single payload leaf now determines the answer, and its in-place replacement creates the matched insufficient condition. It is independently recertified after derivation. Thus generator style, topic, length, evidence position, and background prose are held exactly fixed while effective dependency order changes. This same-context comparison follows the broader controlled-evaluation principle that task difficulty should vary without simultaneously replacing the surrounding context (Cao et al., 2025).

5.4 Evaluation

CertiHop-Bench directly evaluates evidence necessity. A parent passes Joint Irreducibility (JI) only when the model answers Full correctly and abstains under Zero and every required-leaf LOO intervention. We also report Full, Zero, LOO-macro, worst-leaf LOO, and non-critical intervention (NCI) accuracy.

For parent p with m_p required leaves, let F_p, Z_p, and $L_{p,i}$ indicate correct behavior under the complete, zero, and ith leave-one-out conditions. Joint Irreducibility is

$$\mathrm{JI} = \frac{1}{|P|} \sum_{p \in P} F_p Z_p \prod_{i=1}^{m_p} L_{p,i}. \tag{3}$$

The conjunction prevents success on only easy deletions. Causal comparisons use the same parents and families; cross-corpus results are leaf-count stratified. Two position-matched non-critical replacements per parent guard against generic refusal: $|\mathrm{FullAcc} - \mathrm{NCI}| \leq 2$ percentage points must hold, or higher JI is interpreted as over-abstention. Length, leaf-count, depth, position, and dispersion strata are diagnostic only.

External evaluation separates complementary forms of evidence. BABILong QA2/QA3 at 8K, 16K, and 32K test reasoning over facts distributed through a long haystack (Kuratov et al., 2024). RULER variable tracking at the same lengths is a structure-aligned diagnostic (Hsieh et al., 2024). LooGLE long-dependency QA and the MuSiQue, 2WikiMultiHopQA, and MultiFieldQA-en subsets of LongBench test realistic or natural multi-document transfer (Li et al., 2024; Bai et al., 2023; Trivedi et al., 2022); LongBench v2 is reported separately because its multiple-choice protocol is not commensurate with LongBench v1 F1 (Bai et al., 2024b). The frozen raw-output scorer remains the registered BABILong endpoint. Because complete matched raw-output aggregates are not reported in this draft, we make no primary BABILong claim; first/last-answer extraction is confined to a post-hoc appendix sensitivity. Public benchmarks demonstrate transfer, not item-level evidence necessity.

RULER-VT-3 and LongBench-3 are frozen unweighted aggregates over their declared task sets. Missing or failed task runs are not silently omitted from an aggregate.

For the controlled ablations, fixed tasks are aggregated within each training seed before inference. We report paired seed-level effects, sample standard deviations, and paired-t intervals over seeds 42, 2026, and 3407. Holm correction is applied within each ablation to the eight frozen RULER-VT-3/LongBench-3 comparisons (four backbones by two metrics); post-hoc BABILong extraction sensitivities are excluded from that family.

Complete internal families support JI; Pair Joint Accuracy (PJA) is limited to historical one-Missing diagnostics and never substitutes for JI.

6 Results

6.1 Cross-model comparison

Table 2 reports three-seed Stage-2 point estimates against the exposure-matched continued-normal control; failed tasks are never silently removed. Staged-main seed intervals remain pending.

Table 2: Three-seed frozen-scorer point estimates for the matched continued-normal control. Staged-main seed intervals remain pending.

Backbone	Stage-2 data	Seeds	RULER-VT-3	LongBench-3
Qwen3.5-4B	Continued-normal	3	99.77	16.79
	CertiHop	3	**99.88**	**51.74**
Qwen3.5-9B	Continued-normal	3	**99.96**	39.83
	CertiHop	3	85.12	**51.75**
Qwen3.5-27B	Continued-normal	3	99.97	25.40
	CertiHop	3	**100.00**	**61.65**
Qwen3.6-35B-A3B	Continued-normal	3	99.41	41.66
	CertiHop	3	99.24	**57.23**

Table 3: Seed-42 fixed-budget secondary results. "Hits" gives ceiling-hit counts for CertiHop/continued-normal/LongFaith; every example remains scored.

Backbone	Metric (output ceiling)	CertiHop	Δ vs. CN	Δ vs. LF	Hits (C/CN/LF)
Qwen3.5-4B	RULER-VT-8K (8192)	99.80	−0.16	+8.80	0/0/45
	LooGLE QA F1 (1024)	23.82	+17.01	+19.31	5/5/139
Qwen3.5-9B	RULER-VT-8K (8192)	84.40	−15.48	−14.20	0/0/7
	LooGLE QA F1 (1024)	24.91	+18.06	+19.38	1/5/140
Qwen3.5-27B	LooGLE QA F1 (1024)	20.77	+13.36	+14.39	5/1/62
Qwen3.6-35B-A3B	LooGLE QA F1 (1024)	21.61	+14.55	+14.42	4/2/28

Fixed-budget scores include outputs reaching the declared ceiling and are not zero-hit latent-capability estimates.

Relative to continued-normal, CertiHop raises LongBench-3 point estimates by 34.95, 11.92, 36.25, and 15.57 points at 4B, 9B, 27B, and sparse-MoE scale. RULER differences are +0.11, −14.84, +0.03, and −0.17; the 9B loss reflects incomplete variable lists (Appendix D).

Table 3 gives a fully enumerated seed-42 fixed-budget evaluation. CertiHop leads LooGLE QA-F1 for all four backbones by 13.36–18.06 points over continued-normal, while retaining the negative 9B RULER cell. Every example remains scored; ceiling hits are reported explicitly. A two-seed 16K extension (Appendix C) reproduces the positive 4B and negative 9B RULER contrasts; its BABILong QA2 analysis remains post-hoc.

6.2 Which design components matter?

Appendix Table 4 gives all three-seed panels. A1 is mixed. A2 LongBench-3 means are positive for all four backbones, but only sparse-MoE survives Holm correction ($p_{\text{Holm}} = 0.042$). A3 is clearest: all 12 seed-by-backbone differences are positive (means $+3.28, +7.18, +5.94, +3.56$), with 9B surviving Holm ($p_{\text{Holm}} = 0.032$). Figure 2 shows paired values and intervals; post-hoc BABILong is excluded from multiplicity testing.

Validation diagnostic. On a preregistered 200-parent validation family, one frozen 9B checkpoint achieved JI, Full, Zero rejection, worst-leaf/all-LOO rejection, and NCI of 1.00 across 2,420 rows, with zero technical failures. SHA-256 artifacts are recorded. Matched controls, held-out testing, and replication remain pending, so this is diagnostic.

7 Related Work

Long-context data uses rationales, evidence graphs, trajectories, and filtering (Yang et al., 2025a; Chen et al., 2025; Sun et al., 2026; Su et al., 2026; Yuan et al., 2026; Gao et al., 2025; Zhu et al., 2025); multi-hop QA admits shortcuts (Trivedi et al., 2022; Min et al., 2019; Jiang & Bansal, 2019). LongFaith and LongCrafter do not release exhaustive post-realization necessity families. Sufficiency and counterfactual work tests missing context or perturbed chains (Jain & Garimella, 2026; Wu et al., 2025; Guo et al., 2023; Lee et al., 2021); CertiHop certifies finalized Full/Zero/all-leaf LOO families for paired SFT and parent-level

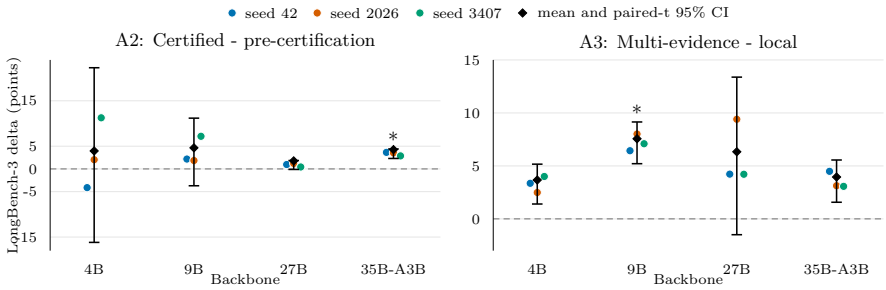

Figure 2: Paired seed-level LongBench-3 effects for A2/A3. Points are seeds; diamonds and whiskers give the mean and paired-t 95% interval ($n = 3$); stars mark Holm survivors among eight frozen tests per ablation.

scoring. Existing suites remain transfer diagnostics (Bai et al., 2023; Li et al., 2024; Hsieh et al., 2024; Yang et al., 2025b; Zhu et al., 2024).

8 Limitations

Certification depends on verifier coverage, templates, unenumerated paths, and parametric recall. JI is one seed-42 9B validation checkpoint without matched controls, test evaluation, or replication. Ablations have three seeds; staged-main intervals and the 120-parent audit are pending. The single-seed secondary evaluation has 251/8,994 8K and 397/13,212 LooGLE ceiling hits, mostly LongFaith, with all rows scored. A two-seed 16K extension covers only 4B/9B and has 351/17,988 ceiling hits (0/1/350 for CertiHop/continued-normal/LongFaith); seed-42 harmonization and the larger-backbone extension were unfinished and are not pooled. Backbones are Qwen-series; 9B RULER exposes list-completeness limits.

9 Conclusion

CertiHop certifies final text and reuses Full/Zero/all-LOO families for paired SFT and joint evaluation. It yields higher long-document QA point estimates, including seed-42 full-denominator LooGLE; controlled A3 gives the clearest evidence for multi-evidence supervision. The 9B list failure and pending controlled JI define the claim boundary.

AI use, ethics, and reproducibility statements

AI use. Generative AI assisted synthetic-data generation, code development, experiment diagnostics, literature discovery, research ideation and execution, drafting and language-level editing, and formalization of mathematical arguments. The authors independently reviewed and verified all AI-assisted material and take responsibility for the final text, claims, and artifacts. Ethics. We use public data and controlled synthetic worlds; provenance/licenses are recorded, unresolved records and internal metadata are withheld, and release requires review. Reproducibility. The text defines interventions, controls, metrics, and statistics; the anonymous supplement provides manifests, frozen protocols, audit materials, scoring scripts, and run templates with versioned SHA-256 fingerprints.

References

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. arXiv preprint arXiv:2308.14508, 2023.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. In Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 1376–1395, 2024a. doi: 10.18653/v1/2024.findings-emnlp.74. URL https://aclanthology.org/2024.findings-emnlp.74/.

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. arXiv preprint arXiv:2412.15204, 2024b.

Shuyang Cao, Kaijian Zou, and Lu Wang. SYNC: A synthetic long-context understanding benchmark for controlled comparisons of model capabilities. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 33627–33648, 2025. doi: 10.18653/v1/2025.emnlp-main.1707. URL https://aclanthology.org/2025.emnlp-main.1707/.

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, pp. 27129–27151, 2025.

Chaochen Gao, Xing Wu, Zijia Lin, Debing Zhang, and Songlin Hu. Longmagpie: A self-synthesis method for generating large-scale long-context instructions. arXiv preprint arXiv:2505.17134, 2025.

Wangzhen Guo, Qinkang Gong, Yanghui Rao, and Hanjiang Lai. Counterfactual multihop QA: A cause-effect approach for reducing disconnected reasoning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4214–4226, 2023. doi: 10.18653/v1/2023.acl-long.231. URL https://aclanthology.org/2023.acl-long.231/.

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.

Akriti Jain and Aparna Garimella. Knowing what's missing: Assessing information sufficiency in question answering. In Findings of the Association for Computational Linguistics: EACL 2026, pp. 4163–4174, 2026. doi: 10.18653/v1/2026.findings-eacl.217. URL https://aclanthology.org/2026.findings-eacl.217/.

Junlong Jia, Ziyang Chen, Xing Wu, Chaochen Gao, Zijia Lin, Debing Zhang, Songlin Hu, and Binghui Guo. Entropylong: Effective long-context training via predictive uncertainty. arXiv preprint arXiv:2510.02330, 2025.

Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2726–2736, 2019. doi: 10.18653/v1/P19-1262. URL https://aclanthology.org/P19-1262/.

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, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c0d62e70dbc659cc9bd44cbcf1cb652f-Paper-Datasets_and_Benchmarks_Track.pdf.

Kyungjae Lee, Seung-won Hwang, Sang-eun Han, and Dohyeon Lee. Robustifying multi-hop QA through pseudo-evidentiality training. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 6110–6119, 2021. doi: 10.18653/v1/2021.acl-long.476. URL https://aclanthology.org/2021.acl-long.476/.

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, pp. 16304–16333, 2024.

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/.

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, 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 Schütze. Nolima: Long-context evaluation beyond literal matching. In Proceedings of the 42nd International Conference on Machine Learning, 2025.

Qwen Team. Qwen3.5 model cards. https://huggingface.co/Qwen/Qwen3.5-4B, 2026a. Pinned 4B, 9B, and 27B repository revisions are listed in the appendix.

Qwen Team. Qwen3.6-35b-a3b model card. https://huggingface.co/Qwen/Qwen3.6-35B-A3B, 2026b. Pinned repository revision is listed in the appendix.

Qisheng Su, Zhen Fang, Shiting Huang, Yu Zeng, Yiming Zhao, Kou Shi, Ziao Zhang, Lin Chen, Zehui Chen, Lijun Wu, and Feng Zhao. ACC: Compiling agent trajectories for long-context training. arXiv preprint arXiv:2605.21850, 2026. URL https://arxiv.org/abs/2605.21850v2. version 2, 14 June 2026.

Zechen Sun, Zecheng Tang, Juntao Li, Wenpeng Hu, Wenliang Chen, Zhuchen Luo, and Qiaoming Zhu. Cgmis: Concept-graph based multi-hop instructions synthesis for enhancing long-context reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 33153–33161, 2026.

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.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. Cofca: A step-wise counterfactual multi-hop QA benchmark. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=q2DmkZ1wVe.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. Longfaith: Enhancing long-context reasoning in llms with faithful synthetic data. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 3236–3256, 2025a.

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, 2025b. 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, 2018.

Chenhao Yuan, Yinhao Xu, Shuwen Xu, Xizhi Yang, Jiaxiang Liu, Chenxi Zhou, Shaoping Huang, Haolin Ren, Pengfei Cao, Jun Zhao, and Kang Liu. Longcrafter: Towards diverse long-context understanding via evidence-graph-guided instruction synthesis. arXiv preprint arXiv:2607.06160, 2026. URL https://arxiv.org/abs/2607.06160v2. version 2, 3 August 2026.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 18–37, 2024. doi: 10.18653/v1/2024.acl-short.2.

Wenhao Zhu, Pinzhen Chen, Hanxu Hu, Shujian Huang, Fei Yuan, Jiajun Chen, and Alexandra Birch. Generalizing from short to long: Effective data synthesis for long-context instruction tuning. arXiv preprint arXiv:2502.15592, 2025. URL https://arxiv.org/abs/2502.15592.

A External Ablation Results

Table 4: Three-seed frozen-scorer diagnostic effects (treatment minus control); none substitutes for registered parent-level JI.

ID	Backbone	Treatment − control	Seeds	ΔRULER-VT-3	ΔLongBench-3
A1	Qwen3.5-4B	Full+Missing − Full-only	3/3	+4.93	+0.86
A1	Qwen3.5-9B	Full+Missing − Full-only	3/3	−0.92	−4.16
A1	Qwen3.5-27B	Full+Missing − Full-only	3/3	+0.05	−0.50
A1	Qwen3.6-35B-A3B	Full+Missing − Full-only	3/3	−0.03	−2.43
A2	Qwen3.5-4B	Certified − Pre-cert.	3/3	−0.37	+3.06
A2	Qwen3.5-9B	Certified − Pre-cert.	3/3	+0.34	+3.74
A2	Qwen3.5-27B	Certified − Pre-cert.	3/3	+0.00	+0.87
A2	Qwen3.6-35B-A3B	Certified − Pre-cert.	3/3	+0.02	+3.35
A3	Qwen3.5-4B	Multi-evidence − Local	3/3	+2.76	+3.28
A3	Qwen3.5-9B	Multi-evidence − Local	3/3	+2.51	+7.18
A3	Qwen3.5-27B	Multi-evidence − Local	3/3	+0.03	+5.94
A3	Qwen3.6-35B-A3B	Multi-evidence − Local	3/3	+0.06	+3.56

B Post-hoc BABILong Output-Extraction Sensitivity

Table 5: Post-hoc BABILong first/last-answer extraction sensitivities. These values are excluded from primary claims and multiplicity correction. Panel A reports three-seed staged-main scores; Panel B reports treatment-minus-control effects; Panel C reports seed-42 fixed-budget QA2-8K sensitivities.

Panel	Backbone	Comparison/arm	Seeds	Score/effect	Δ vs. CN	Δ vs. LF
A	Qwen3.5-4B	Continued-normal	3	27.42	–	–
		CertiHop	3	47.87	+20.45	–
A	Qwen3.5-9B	Continued-normal	3	44.17	–	–
		CertiHop	3	49.27	+5.10	–
A	Qwen3.5-27B	Continued-normal	3	53.66	–	–
		CertiHop	3	64.00	+10.34	–
A	Qwen3.6-35B-A3B	Continued-normal	3	45.84	–	–
		CertiHop	3	57.63	+11.79	–
B	Qwen3.5-4B	A1: Full+Missing − Full-only	3	+0.32	–	–
B	Qwen3.5-9B	A1: Full+Missing − Full-only	3	+1.33	–	–
B	Qwen3.5-27B	A1: Full+Missing − Full-only	3	-3.46	–	–
B	Qwen3.6-35B-A3B	A1: Full+Missing − Full-only	3	-4.66	–	–
B	Qwen3.5-4B	A2: Certified − Pre-cert.	3	-1.75	–	–
B	Qwen3.5-9B	A2: Certified − Pre-cert.	3	+1.31	–	–
B	Qwen3.5-27B	A2: Certified − Pre-cert.	3	-8.09	–	–
B	Qwen3.6-35B-A3B	A2: Certified − Pre-cert.	3	+3.44	–	–
B	Qwen3.5-4B	A3: Multi-evidence − Local	3	+7.71	–	–
B	Qwen3.5-9B	A3: Multi-evidence − Local	3	+12.15	–	–
B	Qwen3.5-27B	A3: Multi-evidence − Local	3	+18.58	–	–
B	Qwen3.6-35B-A3B	A3: Multi-evidence − Local	3	+13.51	–	–
C	Qwen3.5-4B	CertiHop, QA2-8K (8192)	1	50.15	+15.42	+38.44
C	Qwen3.5-9B	CertiHop, QA2-8K (8192)	1	53.05	+3.70	+5.20

Panel C retains every row, including 198 LongFaith and one continued-normal ceiling hits across the two QA2-8K cells; it is a fixed-budget output sensitivity, not a zero-hit latent-capability estimate.

C Two-Seed 16K Fixed-Budget Extension

For seeds 2026 and 3407, we reran the 4B/9B three-arm comparison on QA2-8K and RULER-VT-8K with a shared max_new_tokens=16,384, max_model_len=131,072, and greedy decoding. All 17,988 rows have non-empty predictions and zero input-limit hits. Output-ceiling hits were 0, 1, and 350 for CertiHop, continued-normal, and LongFaith, respectively; all rows remain scored. The seed-42 harmonization and larger-backbone extension were incomplete at the submission cutoff and are not pooled into Table 6.

Table 6: Two-seed 16K fixed-budget extension (seeds 2026/3407; mean ± sample SD). QA2 uses the post-hoc best_of_two output extractor; RULER-VT uses the frozen official string-match scorer.

Backbone	Metric	CertiHop	Continued-normal	LongFaith	Δ vs. CN	Δ vs. LF
Qwen3.5-4B	QA2-8K†	49.15±0.28	35.99±1.63	13.86±0.35	+13.17	+35.29
	RULER-VT-8K	99.90±0.03	99.96±0.00	85.72±1.53	−0.06	+14.18
Qwen3.5-9B	QA2-8K†	53.15±0.57	49.65±0.14	49.30±0.92	+3.50	+3.85
	RULER-VT-8K	83.26±3.14	99.96±0.00	98.60±0.00	−16.70	−15.34

†Post-hoc output-extraction sensitivity, not a frozen primary endpoint. The reported SD describes two training seeds and is not an inferential interval.

D Post-hoc Completeness Follow-up

The frozen 9B RULER failure motivated a narrowly scoped follow-up, CertiHop-Completeness. Two 3,500-row seed-42 candidates share the same Stage-1 parent and training configuration, differing only in whether 600 or 1,000 rows use synthetic Full/Missing closure supervision. Table 7 reports the complete validation result. Closure-600 raises synthetic closure parent-joint accuracy from 0.00 to 0.57, while closure-1000 reaches 0.20; however, both reduce internal JI relative to the frozen CertiHop checkpoint. The follow-up therefore identifies a sharp trade-off between exhaustive insufficiency behavior and complete-list production rather than replacing the primary model.

Table 7: Seed-42 completeness follow-up on the preregistered internal and synthetic closure validation sets.

Model	JI	Full	Zero reject	Closure PJA
Frozen CertiHop	**1.000**	**1.000**	**1.000**	0.00
Closure-600	0.015	**1.000**	0.145	**0.57**
Closure-1000	0.515	0.811	0.940	0.20

E Certificate–Minimality Equivalence

Under Section 2's assumptions, $v(q, C^{\text{full}}, a) = 1$ means that $D \cup E$ uniquely determines a. Valid neutralization is monotone, so each C^{-i} remains compatible with a while removing only e_i's constraint; $u(q, C^{-i}, a) = 1$ therefore means that $D \cup (E \setminus \{e_i\})$ admits at least two answers including a. These statements are exactly sufficiency of E and element-wise minimality. The zero term analogously establishes that D alone leaves a possible but non-unique. Their binary conjunction yields $\Gamma(p) = 1$. Conversely, if E is a minimal determining set, D is insufficient, interventions are valid and monotone, and v, u are exact, every factor in $\Gamma(p)$ equals one.

F Worked Dependency and Intervention Example

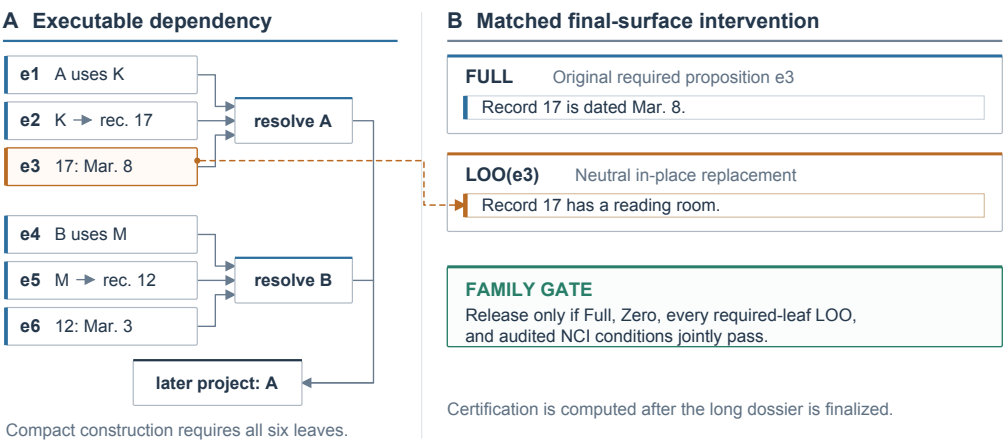

Figure 3: A six-leaf dependency and one member of its matched intervention family. Replacing e_3 preserves the dossier surface but removes the proposition needed to resolve project A's date. The parent is certified only after the analogous intervention for every other leaf also passes.

G Data Composition and Provenance

Table 8 is computed from one Full row per parent in the frozen 11,230-parent release-candidate manifest. The canonical SFT export stores five answer types and all listed structural fields, but its flattened rows do not materialize per-parent language, domain, or upstream-license columns. We therefore report no guessed language/domain proportions. Cohort-level provenance is 5,728 parents from v42, 2,627 from v43, and 2,875 from the controlled v44 cohort; record-level source and license metadata remain subject to release approval. The public-data supplement freezes the observed LongAlign and LongFaith upstream revisions and hashes used by the comparison construction.

Table 8: Exact composition of the 11,230 machine-certified parents. Percentages use parents, not the two-row SFT projection.

Field	Value: count (percentage of parents)
Answer type	boolean: 622 (5.54); date: 2,064 (18.38); entity/text: 3,586 (31.93); list: 2,123 (18.90); number: 2,835 (25.24)
Dependency order	4: 541 (4.82); 6: 3,851 (34.29); 8: 2,966 (26.41); 10: 2,260 (20.12); 12: 1,612 (14.35)
Proof depth	2: 541 (4.82); 3: 6,173 (54.97); 4: 3,080 (27.43); 5: 855 (7.61); 6: 581 (5.17)
Documents	4: 5,157 (45.92); 5: 3,377 (30.07); 6: 2,696 (24.01)

Prompt tokens under the pinned Qwen3.5-27B tokenizer: minimum 7,497, median 10,212, maximum 27,457. Exact-unique questions: 11,230/11,230; normalized unique templates: 11,000/11,230 (97.95%); largest normalized-template group: 7.

H Certification Flow and Pending Release Audit

Table 9 separates recorded machine-admission evidence from the unfinished release-validity audit. We do not reconstruct a single upstream denominator because v42–v44 used cohort-specific generation and retry paths, and the preserved manifests do not expose one comparable candidate count for every gate. The only explicit residual technical count in the merged evidence is 21 v43 parents, all quarantined. For v43, all 899 globally flagged semantic near-duplicate pairs received decisions and none was judged duplicate.

Table 9: Certification-flow evidence available in the supplement. "Pending" is not counted as a pass.

Cohort	Machine-admitted	Tech. quarantined	Near-dup. decisions	Final audit
v42	5,728	not unified	not unified	pending
v43	2,627	21	899/899; 0 duplicates	pending
v44	2,875	not unified	not unified	pending
Merged	11,230	$\geq$ 21 recorded	no cross-cohort overlap	pending

Before any model-score inspection, a deterministic SHA-256 selector froze 120 parents stratified across all five answer types (24 each), four requested-length buckets (31/31/31/27 at 8K/16K/24K/32K), and observed leaf-count and proof-depth strata. The rubric checks Full uniqueness; Zero and every LOO for compatible non-uniqueness; both NCIs for answer preservation; leakage, redundant paths, contradictions, and replacement naturalness. Each label is pass, fail, or uncertain; uncertainty is never a pass. The annotation file is currently empty, so disagreement, false-positive rate, and final audited release status are not reported. Two independent completions of the same 120-parent panel are required before inter-annotator agreement can be claimed.

I Frozen Model and Upstream-Data Registry

Table 10: Frozen model/tokenizer snapshots. Model and tokenizer use the same revision within each row.

Key	Repository	Revision
qwen35_4b	Qwen/Qwen3.5-4B	851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a
qwen35_9b	Qwen/Qwen3.5-9B	c202236235762e1c871ad0ccb60c8ee5ba337b9a
qwen35_27b	Qwen/Qwen3.5-27B	fc05daec18b0a78c049392ed2e771dde82bdf654
qwen36_a3b	Qwen/Qwen3.6-35B-A3B	995ad96eacd98c81ed38be0c5b274b04031597b0

LongAlign code is pinned at 450ecc8c9faeafd83a169610fad94d4b7fac1357; its data revision is 12f17c4baff1001f0d44c4f8feab09ee2ee8c6dc. The observed long.jsonl has 9,888 records and SHA-256 d7a1c39738e645ae1d0f8609a3cd62e50a3fa982e23cb752f64630bdbb7cee08. LongFaith code is pinned at e3df74a0b17124c05256379f9c6dfb67a3a662fb; the selected official GPT-4o-mini SFT file contains 2,048 records and SHA-256 bf4a339fb755b86bc96d521286a0641a5af2749e64713aea3e0bd3d3e44b20f8. Its repository README states MIT, but the pinned commit contains no license file; we disclose that distinction rather than normalizing it away.

Formal training used Huawei accelerators and full-parameter SFT. The supplementary run-manifest schema requires accelerator model and count, CANN, PyTorch, torch_npu, TorchTurbo, ms-swift, tokenizer/chat-template hashes, tuning method, optimizer, learning rate, scheduler, epochs or optimizer updates, packing, precision, cutoff, and effective batch size. Fields not yet returned by the remote job remain explicitly pending; historical CUDA or HPU YAML files are not used to fill them.

J Reproducibility Checklist for Data Certification

For every released parent, we preserve source provenance, the compact dependency structure, all final intervention surfaces, verifier outputs, programmatic audit results, tokenizer-specific length checks, split membership, and a release decision. Full and missing variants remain in the same split. API credentials, internal endpoints, and unapproved proprietary artifacts are never included in the public release.

K Generation and Audit Model Roles

The production protocol uses explicitly separated roles. GLM-5.2 batch inference proposes evidence-preserving naturalizations and long dossiers (temperature 0.35 for naturalization and 0.2 for long-surface expansion). GLM-5.1 batch inference is the primary Full/Zero/all-LOO behavioral verifier at temperature 0. GLM-5.3 online inference is designated for targeted independent votes and the frozen stratified audit; DeepSeek-V4-Pro is designated for recorded disagreement adjudication. The latter audit is pending as documented in Appendix H. Online models are not on the all-example admission path, and no generator is the sole verifier of its own output. Programmatic execution, schema, leakage, deduplication, and tokenizer checks remain mandatory regardless of model votes.

Output-budget retries repair only technical incompleteness. Initial long-surface requests used a 24,576-token completion ceiling; a documented subset exhausted that budget through provider-side hidden reasoning despite requesting disabled thinking. Exact failed document IDs are retried once with a 49,152-token ceiling, then rejoined with first-pass complete responses and re-audited. Complete semantic failures are quarantined rather than resampled into acceptance. The supplementary manifest records the available model snapshots, request parameters, retry evidence, and known cohort counts; unavailable cross-cohort rates remain marked not_unified.

L Construction Cost and Throughput Accounting

Our reporting schema records each stage's model and transport, entering and leaving parent counts, request and prompt hashes, token usage, retries, semantic rejections, wall time, and available cost. Normalized throughput is reported only when all denominators exist; queueing is separated from service time, missing totals remain unavailable, and heterogeneous units are never collapsed.