SCENARIO COLLAPSE: MEASURING SCENARIO-LOCAL EVIDENCE BINDING IN LONG-CONTEXT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can retrieve the right evidence from a long context yet apply its value to the wrong case or document version. We call this failure *scenario-local binding*. We introduce SCENARIOCOLLAPSEBENCH, a paired-scenario benchmark with 10,132 scenario pairs and deterministic evidence, state, and decision traces, together with *found-but-not-bound* (FBNB), which measures neighboring-scenario value substitution after successful evidence localization. Across ten models, substantial binding errors remain even after the correct evidence is found. The same failure appears in 200 naturally occurring historical versions of UK legislation and paired adaptations of RULER and BABILong. A matched Qwen3-14B thinking ablation shows that additional reasoning reduces, but does not eliminate, the problem. Oracle experiments locate the dominant residual error in value/state construction rather than evidence access. Finally, making scenario-local state explicit through scenario-specific identifiers, state tables, and verification substantially reduces binding errors and collapse. These results identify scenario-local binding as a distinct reasoning problem and provide a practical design principle for systems that reason across similar cases, versions, or branches.

1 INTRODUCTION

Large language models are increasingly evaluated in long, structured, and evolving contexts, including policy interpretation, workflow authorization, audit support, agent planning, compliance review, and counterfactual decision support (Bai et al., 2024; Hsieh et al., 2024; Shaham et al., 2022; Bai et al., 2025; Kuratov et al., 2024; Yang et al., 2018; Trivedi et al., 2022; Dasigi et al., 2021; Yu et al., 2023; Zhang et al., 2024; An et al., 2024). In these settings, reliability depends on more than finding relevant evidence. Models may need to reason over multiple cases, document versions, branches, or jurisdictions that share most entities, labels, rules, and surface structure while differing in only a few decision-relevant updates. Correct reasoning therefore requires preserving not only what evidence says, but also which local world that evidence belongs to.

Our key observation is that finding the right evidence is not enough. A model can localize the correct evidence and still construct the wrong local world. Consider two vendor-payment records in which event E22 sets an approval limit of USD 70,000 in Scenario A and USD 100,000 in Scenario B. For a USD 90,000 invoice, a model may cite E22 correctly for both scenarios yet approve both by carrying Scenario B's value into Scenario A. An identifier- or span-localization metric can assign evidence credit even though the scenario-local state is wrong. This exposes a post-localization failure that evidence retrieval or citation-location scores alone do not directly measure (Gao et al., 2023; Buchmann et al., 2024; Petroni et al., 2021; Thorne et al., 2018; DeYoung et al., 2020; Rashkin et al., 2023).

We distinguish two levels of this behavior. A *scenario-local binding failure* is a field-level error in which a model correctly localizes evidence in the target scenario but assigns it a value belonging to another scenario. Such a failure can lead to *scenario collapse* when two distinct scenario-local states are no longer preserved separately, causing one scenario to be represented as the other or both predictions to converge to the same local world. This differs from retrieval failure, where the relevant evidence is not found; value-extraction error, where the retrieved value is wrong but does not come from the neighboring scenario; and rule-use error, where the local state is correct but the final decision is wrong. We use scenario collapse as a behavioral descriptor; its mechanism is left unspecified.

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.

Existing long-context and state-tracking evaluations generally do not annotate the complete association among scenario-scoped evidence, the value expressed by that evidence, and the local state in which the value is valid. To make this behavior directly measurable, we introduce SCENARIOCOLLAPSEBENCH, a factorized paired-scenario benchmark in which near-duplicate local worlds share rules, entities, local identifiers, distractors, and surface structure while differing in decision-relevant state updates. Deterministic latent traces expose evidence identifiers, evidence values, final states, rule paths, and decisions for direct scoring. We further define FBNB, a field-level conditional metric that asks: among fields for which the model has already found the required scenario-local evidence, how often does it nevertheless assign the neighboring scenario's value?

We use the benchmark as an instrument for studying scenario-local binding rather than treating the raw localization–binding difference as an independent quantity: binding accuracy includes successful localization by definition. The informative question is therefore conditional—given that the required evidence was localized, how often is the value bound to the correct local state, and how often is the neighboring scenario's value substituted? We test this question on the factorized benchmark suite, on 200 naturally occurring historical version pairs from UK legislation, and through paired adaptations of RULER and BABILong. Oracle and structural interventions then probe whether residual errors arise from evidence access, value recovery, local-state construction, or downstream rule use.

Our contributions are threefold:

- **Phenomenon.** We identify and operationalize a separation between evidence localization and scenario-local state construction, distinguishing field-level binding failures from pair-level scenario collapse.
- **Measurement.** We introduce SCENARIOCOLLAPSEBENCH and FBNB, which make this failure directly measurable through deterministic evidence–value–state annotations, together with a common-support variant for cross-model comparison and a conditional post-localization binding rate.
- **Evidence and diagnosis.** We show that post-localization binding errors persist across model families, including reasoning-oriented models, in naturally occurring document versions, and in paired adaptations of existing long-context benchmarks. Oracle interventions place the dominant residual error in the post-localization value/state-construction stage, while structural interventions show that explicit scenario separation reduces these errors.

2 RELATION TO EXISTING BENCHMARKS

Long-context evaluation has moved beyond single-span retrieval toward aggregation, state tracking, multi-document reasoning, and reasoning over alternatives (Bai et al., 2024; Hsieh et al., 2024; Shaham et al., 2022; Bai et al., 2025; Kuratov et al., 2024; Yen et al., 2025; Modarressi et al., 2025). LongBench, LongBench v2, SCROLLS, RULER, BABILong, and related benchmarks test aspects of long-context behavior, but they generally do not explicitly annotate the binding among a scenario-local evidence item, the value expressed by that item, and the local world in which that value is valid.

Multi-hop QA benchmarks such as HotpotQA and MuSiQue test reasoning across supporting facts, while QASPER includes supporting-evidence annotations (Yang et al., 2018; Trivedi et al., 2022; Dasigi et al., 2021). Dialogue-state tracking datasets such as MultiWOZ evaluate slot-value tracking over turns (Budzianowski et al., 2018; Eric et al., 2020). These tasks are closely related, but their supervision does not directly isolate the case where a model localizes the correct local evidence while attaching a neighboring scenario's value to the target state. Knowledge-conflict work studies a complementary failure regime: NQ-Swap and related analyses examine conflicts between contextual evidence and parametric knowledge (Longpre et al., 2021; Wang et al., 2025). Our setting instead studies conflicts within the supplied context: two local worlds contain valid but scenario-specific values, and the question is whether the model preserves the association between evidence, value, and local world.

Table 1 summarizes the diagnostic ingredients. SCENARIOCOLLAPSEBENCH complements existing benchmarks by isolating a distinct evaluation axis: the association among scenario-scoped evidence, its value, and the local world in which that value is valid.

Table 1: Relation to existing benchmarks. Columns describe whether the original task structure and released supervision expose the ingredients needed to score scenario-local evidence–value binding. "Partial" means that a state/value label is available but the complete scenario-scoped evidence–value association is not directly annotated. Paired adaptations of RULER and BABILong are evaluated separately in Section 4.8.

Benchmark	Main target	Multiple local worlds	Reused local IDs	Conflicting values	Evidence–value binding	Binding directly scorable?
Needle-in-a-haystack	Long-context retrieval	No	No	No	No	No
LongBench / v2	General long-context understanding	Sometimes	No	Sometimes	Usually no	No
RULER	Retrieval, tracing, aggregation	Limited	No	Sometimes	No	No
SCROLLS	Long-sequence QA / NLI / summarization	Limited	No	Sometimes	No	No
BABILong	Long-context state reasoning	Usually one world	No	Yes	State labels only	Partial
bAbI (Weston et al., 2016)	Synthetic state tracking	Usually one world	No	Yes	State labels	Partial
HotpotQA / MuSiQue	Multi-hop QA	Multiple documents/facts	No	Sometimes	Supporting facts / QA	No
QASPER	Scientific-document QA	One paper	No	Sometimes	Supporting evidence	No
MultiWOZ	Dialogue-state tracking	One dialogue state	No	Yes	Slot-value labels	Partial
Counterfactual QA (Yu et al., 2023)	Reasoning under alternatives	Sometimes	Rarely	Yes	Usually no	No
ScenarioCollapseBench	Scenario-local evidence–value binding	Yes	Yes; namespace ablations	Yes	Yes	Yes

3 Design and Metrics

3.1 Overview

Each ScenarioCollapseBench instance contains Scenario A and Scenario B. The two scenarios share task rules, entity names, local event identifiers, distractor structure, and surface form, but differ in one or more decision-relevant state updates. The model must answer each scenario separately and cite the evidence events that determine the final state. The benchmark separates four capabilities: (1) localizing relevant evidence, (2) extracting the value expressed by that evidence, (3) binding the value to the correct scenario-local state, and (4) applying a downstream decision rule.

3.2 Benchmark Construction

Each scenario is represented as an ordered timestamped event sequence

$$S = (e_1, e_2, \ldots, e_n), \qquad t_1 \leq t_2 \leq \cdots \leq t_n. \tag{1}$$

Each event contains an identifier, timestamp, event type, affected attribute, value, and authority flag. Only authoritative updates modify scenario state. Non-authoritative notes, summaries, reports, dashboards, archives, digests, and policy comments act as distractors and may contain stale or conflicting values. For each scenario, the generator records gold evidence identifiers, gold values, the final state, the decision, and the rule path.

Scenario A and Scenario B share rules, entities, local IDs, and distractor layout. A targeted perturbation changes one or more decision-relevant state updates. In the main paired condition, the perturbation is selected so that the correct downstream decisions differ. Additional ablations include different-state/same-decision and same-decision/conflicting-value conditions.

3.3 Task Families and Data Tiers

The benchmark contains 10,132 paired instances and 10,132 matched single-scenario instances across six task families: dataset-export authorization, vendor-payment approval, software-workflow authorization, access-control review, deployment approval, and data-export request audit. The paired set contains 6,117 factorized synthetic, 2,994 naturalized synthetic, and 1,021 human-written instances. The human-written tier reduces template regularity but remains grounded in deterministic latent traces.

Table 2 summarizes context scale using the Qwen2.5 tokenizer after benchmark prompt serialization and before provider-specific system wrappers. Paired benchmark inputs have a median length of 4,780 tokens and a maximum of 7,120 tokens; decisive evidence is distributed throughout the context, with the 90th-percentile relative evidence position at 0.91. The paired RULER and BABILong adaptations occupy a similar multi-thousand-token regime. These settings test state separation over structured multi-thousand-token contexts with decisive evidence distributed throughout the input.

Table 2: Context-scale statistics. Token counts use the Qwen2.5 tokenizer on serialized benchmark inputs. Relative evidence position is the token offset of the final decisive evidence divided by total input length.

Statistic	Median	P90	Max
Paired benchmark input tokens	4,780	6,420	7,120
Single-scenario input tokens	2,390	3,210	3,560
Events per paired instance	144	208	240
Decisive-evidence relative position	0.57	0.91	0.99
RULER paired-adaptation tokens	6,120	6,940	7,220
BABILong paired-adaptation tokens	6,480	7,080	7,360

3.4 EVIDENCE–VALUE BINDING

For paired instance i, scenario s, and required field k, let $v^*_{i,s,k}$ be the gold value and $E^*_{i,s,k}$ the decisive scenario-scoped evidence set. The model predicts $\hat{v}_{i,s,k}$ and $\hat{E}_{i,s,k}$. We use two primitive indicators:

$$L_{i,s,k} = \mathbf{1}\Big[\mathrm{loc}(\hat{E}_{i,s,k}, E^*_{i,s,k}) = 1\Big], \tag{2}$$

$$V_{i,s,k} = \mathbf{1}\big[\mathrm{norm}(\hat{v}_{i,s,k}) = \mathrm{norm}(v^*_{i,s,k})\big]. \tag{3}$$

Here loc is the deterministic scenario-scoped evidence scorer defined in Section 3.7 and Appendix F: all decisive evidence for the field must be localized to the target scenario, while wrong-scenario or contradictory same-attribute evidence is disqualifying. norm is the deterministic field-specific value canonicalization used by the released scorer. Unicode text is normalized to NFKC, leading/trailing whitespace is removed, and repeated internal whitespace is collapsed; case folding is applied only to schema-declared case-insensitive categorical fields; categorical aliases are mapped through a field-specific closed vocabulary; numeric values are parsed after removing formatting separators and compared in the schema-declared unit and precision; dates are converted to ISO-8601 form when the field schema is a date; boolean/status fields use an explicit per-field alias table; missing values remain distinct from explicit null/negative values; and scenario namespaces on evidence identifiers are never removed or merged.

Field-level binding is

$$\mathrm{Bind}_{i,s,k} = L_{i,s,k} V_{i,s,k}. \tag{4}$$

Joint localization-and-binding accuracy (BINDINGACC) averages this indicator over required fields. Because it includes successful localization by definition, BINDINGACC $\leq$ EVIDENCELOC. We therefore also report the post-localization binding rate

$$\textsc{CondBind} = \frac{\sum_{i,s,k} L_{i,s,k} V_{i,s,k}}{\sum_{i,s,k} L_{i,s,k}}, \tag{5}$$

which asks: among fields for which the required evidence was localized, how often was the value also bound correctly? The metric is undefined when the denominator is zero.

3.5 FOUND-BUT-NOT-BOUND

Let $\bar{s}$ denote the neighboring scenario in the pair. We define neighboring-scenario substitution as

$$N_{i,s,k} = \mathbf{1}\big[\mathrm{norm}(\hat{v}_{i,s,k}) \neq \mathrm{norm}(v^*_{i,s,k}) \ \wedge \ \mathrm{norm}(\hat{v}_{i,s,k}) = \mathrm{norm}(v^*_{i,\bar{s},k})\big]. \tag{6}$$

FBNB is the neighboring-value substitution rate conditional on successful target-scenario localization:

$$\mathrm{FBNB} = \frac{\sum_{i,s,k} L_{i,s,k} N_{i,s,k}}{\sum_{i,s,k} L_{i,s,k}}. \tag{7}$$

The conjunction inside $N_{i,s,k}$ makes the definition valid even when the two scenarios share a gold value for a field: such a field cannot be counted as a neighboring-value substitution unless the predicted value is simultaneously wrong for the target and equal to the neighbor. The metric is undefined when the localization denominator is zero.

3.6 COMMON-SUPPORT FBNB

For a compared model set $\mathcal{M}$, define the common localization support

$$\mathcal{S}_{\text{common}}(\mathcal{M}) = \left\{(i,s,k) : L_{i,s,k}^{(m)} = 1 \; \forall m \in \mathcal{M}\right\}. \tag{8}$$

For model m,

$$\text{CS-FBNB}_m = \frac{\sum_{(i,s,k)\in\mathcal{S}_{\text{common}}(\mathcal{M})} N_{i,s,k}^{(m)}}{|\mathcal{S}_{\text{common}}(\mathcal{M})|}. \tag{9}$$

The model set is included in the support definition because changing the comparison set can change the intersection. CS-FBNB is undefined when $|\mathcal{S}_{\text{common}}(\mathcal{M})| = 0$. FBNB remains the primary within-model diagnostic; CS-FBNB provides an identical denominator for direct cross-model comparison.

3.7 SCENARIO-SCOPED EVIDENCE IDENTIFIERS

Evidence identifiers are local to a scenario unless explicitly namespace-separated. An identifier such as E07 inside a Scenario A object is interpreted as A:E07. In namespace-separated conditions, A-E07 and B-E07 are literal. A wrong-scenario citation receives localization credit of zero even if the same surface identifier exists in both scenarios. Ambiguous global citations are not localized unless the output schema or surrounding text unambiguously assigns them to a scenario. Multi-evidence fields require all decisive authoritative events.

3.8 COLLAPSE AND DECISION METRICS

BINDINGACC is field-level, whereas collapse metrics classify pair-level state patterns. A collapse to A occurs when the predicted state for Scenario A matches A's gold state and the predicted state for Scenario B matches A's gold state; collapse to B is defined symmetrically. Let $\mathcal{P}_\Delta$ be the evaluated pairs whose normalized gold states differ. The narrow collapse rate is

$$\text{COLLAPSERATE} = \frac{|\text{CollapseToA}| + |\text{CollapseToB}|}{|\mathcal{P}_\Delta|}. \tag{10}$$

Pairs with missing or unparsable required state fields remain in $\mathcal{P}_\Delta$ and are assigned to the non-collapse error categories defined in Appendix G. A binding failure can occur without narrow collapse, and a pair can collapse even when some individual fields remain correct; we therefore report both levels.

For decisions, we report mean per-scenario decision accuracy within a pair,

$$\text{MeanDecisionAcc} = P(\text{both correct}) + \tfrac{1}{2}P(\text{one correct}), \tag{11}$$

and strict pair accuracy $P(\text{both correct})$ separately. This naming avoids interpreting the mean per-scenario score as exact pair correctness.

3.9 MODELS AND INFERENCE SETTINGS

The full-benchmark model suite contains GPT-4.1 (OpenAI, 2025a), DeepSeek-V3 (DeepSeek-AI, 2024), Gemini 2.5 Pro (Google DeepMind, 2025), Claude 3.7 Sonnet (Anthropic, 2025), Qwen2.5-32B/14B/7B (Qwen Team et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), Gemma-2-9B (Gemma Team et al., 2024), and Mistral-7B (Jiang et al., 2023). Additional coverage includes DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI et al., 2025), Qwen3-14B with thinking enabled (Yang et al., 2025), and OpenAI o3 (OpenAI, 2025b) on validated subsets, natural-version data, and transfer experiments. A matched Qwen3-14B on/off ablation isolates the effect of explicit thinking mode within one checkpoint in Section 4.6. All main experiments use the same default task prompt unless otherwise specified. Greedy or deterministic decoding is used where supported; open-weight models use temperature 0.0 and top-p 1.0. Exact model identifiers and run-level inference settings are reported in Appendix H.

Table 3: Full-benchmark results on the factorized SCENARIOCOLLAPSEBENCH suite with pair-cluster bootstrap 95% confidence intervals for the originally reported metrics. BINDINGACC requires both successful localization and a correct normalized value. CONDBIND is computed conditionally from localized fields; its bootstrap intervals are reported in Appendix M.

Model	Single Acc	Mean Dec. Acc	Ev.Loc	Bind.Acc	CondBind	FBNB	Collapse
GPT-4.1	96.3 ± 1.6	66.9 ± 1.1	93.4 ± 1.1	59.8 ± 3.2	64.0	24.6 ± 3.1	22.4 ± 1.1
DeepSeek-V3	97.9 ± 1.5	66.8 ± 2.5	85.7 ± 1.7	61.5 ± 2.3	71.8	20.9 ± 2.3	30.3 ± 2.4
Gemini 2.5 Pro	93.8 ± 1.9	62.8 ± 2.2	91.7 ± 1.6	57.3 ± 1.2	62.5	27.4 ± 1.3	25.1 ± 1.4
Claude 3.7 Sonnet	96.6 ± 2.1	68.2 ± 2.1	87.5 ± 1.6	63.6 ± 1.8	72.7	22.8 ± 2.2	31.0 ± 1.9
Qwen2.5-32B	88.2 ± 1.6	59.7 ± 1.4	86.8 ± 2.7	53.4 ± 1.7	61.5	27.9 ± 1.5	18.6 ± 3.2
Llama-3.1-8B	86.1 ± 1.2	49.6 ± 3.2	74.1 ± 1.0	46.3 ± 1.7	62.5	33.9 ± 2.8	36.2 ± 1.5
Qwen2.5-14B	91.7 ± 1.2	52.9 ± 2.1	80.6 ± 1.9	54.1 ± 2.3	67.1	27.6 ± 1.7	38.5 ± 3.0
Gemma-2-9B	79.3 ± 2.0	44.8 ± 2.4	77.2 ± 2.2	33.1 ± 2.7	42.9	31.5 ± 2.2	49.3 ± 2.9
Qwen2.5-7B	84.8 ± 1.8	41.9 ± 2.4	69.4 ± 2.3	42.0 ± 2.6	60.5	36.7 ± 2.8	41.7 ± 1.8
Mistral-7B	76.2 ± 2.3	42.7 ± 2.5	72.9 ± 2.2	31.6 ± 2.8	43.3	38.2 ± 3.1	43.6 ± 1.1

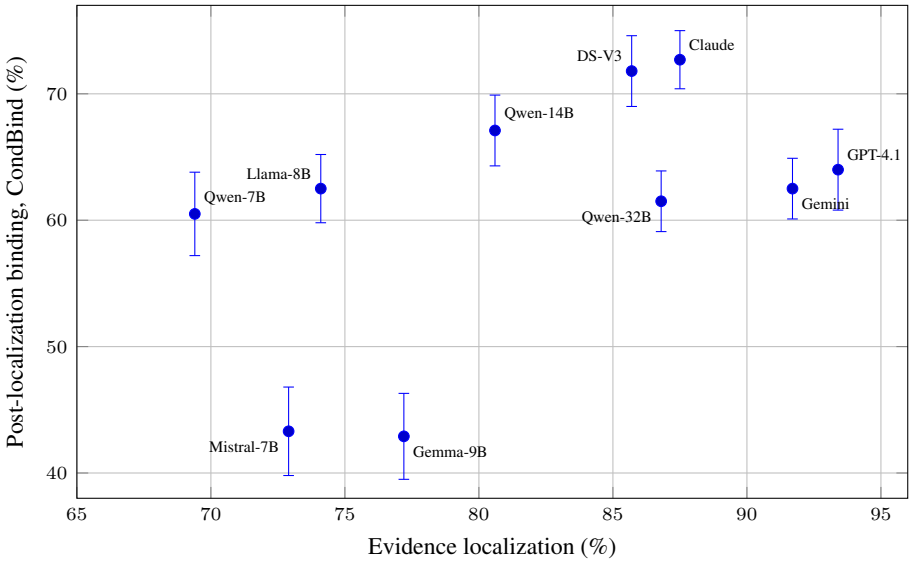


Figure 1: Evidence localization versus post-localization binding on the full benchmark. Vertical bars are pair-cluster bootstrap 95% confidence intervals for CONDBIND. The conditional vertical metric is not mechanically constrained to lie below evidence localization.

4 RESULTS

All values are percentages unless otherwise specified. The full-benchmark results cover ten models. Additional model coverage, the matched thinking ablation, natural-version validation, and paired benchmark adaptations are reported separately to preserve the distinction between evaluation settings.

4.1 POST-LOCALIZATION BINDING ERRORS

Models perform well on single scenarios but degrade when two near-duplicate scenarios are presented together. Evidence localization remains high, yet a substantial fraction of localized fields still fail value binding. Because BINDINGACC includes localization by definition, the raw difference between EVIDENCELOC and BINDINGACC is descriptive rather than an independent test of dissociation. We therefore interpret the results conditionally. For GPT-4.1, $59.8/93.4 = 64.0\%$ of localized fields are also value-correct. Across the ten full-benchmark models, this post-localization binding rate ranges from 42.9% to 72.7%. FBNB further isolates post-localization errors in which the model substitutes the neighboring scenario's gold value.

4.2 Common-Support FBNB

The conditional denominator of FBNB differs across models because each model localizes a different subset of required fields. We therefore recompute FBNB on the intersection localized successfully by every model in the factorized benchmark suite. This common support contains 58.4% of target fields. Appendix Table 7 reports the shared-denominator analysis; neighboring-scenario substitution remains substantial.

4.3 Decision Correctness at Scenario and Pair Levels

Mean per-scenario decision accuracy gives half credit when exactly one scenario is correct, whereas strict pair accuracy requires both decisions to be correct. Appendix Table 8 gives the full breakdown. For GPT-4.1, mean decision accuracy is 66.9%, while strict pair accuracy is 50.7%.

4.4 Diagnostic Instance

A representative vendor-payment instance illustrates a found-but-not-bound error. Scenario A states in local event E07 that Nora's approval limit is EUR 15,000, while Scenario B uses the same local identifier E07 for EUR 25,000. The invoice amount is EUR 18,000 and vendor risk is medium in both scenarios, so the correct decisions are reject for A and approve for B. An illustrative output cites E07 for both scenarios but assigns A the EUR 25,000 value from B. Evidence localization for A is correct while value binding is wrong; because the spending limit is the only differing state field, the predicted A state becomes identical to B's gold state, yielding a collapse to B. Appendix J gives the full instance and scoring.

4.5 Oracle Decomposition

Appendix Table 4 progressively removes uncertainty in evidence access and state construction. Oracle evidence IDs raise localization to 100% but binding only to 71.4%; oracle spans reach 75.6%, while supplying the evidence value raises binding to 93.2%. Gold state plus model rule use yields 97.8% decision accuracy. These results localize most residual difficulty after evidence access without identifying a unique internal mechanism.

4.6 Additional Model Coverage and Matched Thinking Ablation

DeepSeek-R1-Distill-Qwen-14B, Qwen3-14B Thinking, and OpenAI o3 extend coverage beyond the ten-model factorized suite. On the 1,021-pair human-written tier, DeepSeek-R1-Distill-Qwen-14B obtains 85.3% evidence localization, 62.1% binding accuracy, 18.9% FBNB, and 24.6% collapse. In the matched Qwen3-14B experiment (Appendix Table 5), thinking increases conditional binding by 7.7 points and strict pair accuracy by 12.9 points, while reducing FBNB by 7.8 points and collapse by 9.9 points. The matched ablation isolates an inference-mode benefit within this checkpoint.

4.7 Natural-Version External Validation

We evaluate 200 independently screened historical version pairs from the Version-Aware RAG UK Legal Corpus (Mathitharan, 2026), where the same Act/Part contains a substantive decision-relevant change. The model answers a version-specific question and cites the decisive provision. Appendix Table 6 reports the results. All thirteen models exhibit post-localization binding errors; the common support contains 246 of 400 target fields (61.5%).

4.8 Transfer to Existing Benchmarks

We construct matched paired adaptations of 200 RULER (Hsieh et al., 2024) and 200 BABILong (Kuratov et al., 2024) instances, preserving the original retrieval or state-tracking task while requiring two conflicting local worlds to be maintained simultaneously. Appendix Tables 9 and 10 compare original and paired conditions. Across both benchmarks, the paired condition exposes post-localization binding errors not measured by original single-world accuracy alone.

4.9 Interventions for Scenario Separation

Appendix Table 11 evaluates retrieval support, boundaries, namespaces, explicit state construction, verification, and scenario splitting over the full benchmark. Generic chain-of-thought prompting is inconsistent; per-scenario identifiers, explicit state tables, and parser-based verification give larger binding gains and collapse reductions. Scenario splitting is retained as an ablation because it removes simultaneous state maintenance from the paired condition.

4.10 Data Tiers

The failure persists across factorized synthetic, naturalized synthetic, and human-written tiers (Appendix Table 12). The human-written tier is hardest overall, with 85.9% evidence localization and 44.3% binding accuracy, reducing the likelihood that the effect is solely a rigid-template artifact.

5 Discussion and Limitations

Our results extend a growing body of work showing that long-context capability is not equivalent to reliable long-context use. Prior studies have mainly emphasized whether models can find, retain, or combine information in large inputs. *Lost in the Middle* showed that access to relevant information depends strongly on where it appears in context (Liu et al., 2024), while RULER and BABILong showed that performance degrades as retrieval is combined with tracing, aggregation, or multi-step reasoning (Hsieh et al., 2024; Kuratov et al., 2024). LongBench v2 similarly shows that realistic long-context reasoning remains difficult even for strong models (Bai et al., 2025; Yen et al., 2025; Modarressi et al., 2025). Our results add a different failure point. In ScenarioCollapseBench, a model can already have the right evidence and still attach the wrong value to the target scenario. For GPT-4.1, evidence localization reaches 93.4%, yet only 64.0% of localized fields are also value-correct; across the ten-model factorized suite, conditional binding ranges from 42.9% to 72.7%. The oracle study points in the same direction: giving the correct evidence identifier raises localization to 100% but binding reaches only 71.4%, whereas supplying the evidence value raises binding to 93.2%. The problem therefore extends beyond "can the model find the right passage?" to "can it keep the right fact attached to the right local world?" It also complements work on knowledge conflict, which has largely studied competition between contextual evidence and parametric memory (Longpre et al., 2021; Wang et al., 2025). Here, both values are valid in the supplied context; the challenge is preserving which version, case, or branch each value belongs to.

This distinction has direct implications for real systems because many workflows contain several similar but simultaneously valid states. A legal assistant may compare historical versions of a statute; a compliance system may review similar cases under different rules; and an enterprise or software agent may reason over parallel records that share most fields but differ in one decision-relevant value. In such settings, a plausible citation is not enough. Work on citation and attribution has shown that evidence can improve verifiability without guaranteeing answer quality, especially for complex claims (Gao et al., 2023; Buchmann et al., 2024). Our results make one possible failure concrete: a system may cite the correct local evidence while using a value from a neighboring case. The natural-version experiment shows that this behavior is not confined to synthetic templates. Across 200 historical UK legal-version pairs, all thirteen evaluated models exhibit post-localization binding errors. The natural-version evaluation identifies a failure mode that retrieval accuracy, citation accuracy, or single-document QA can miss. Systems operating over versions, cases, jurisdictions, branches, or parallel records may therefore need to evaluate not only whether evidence was retrieved, but whether each extracted value remains tied to the correct scope throughout reasoning.

The intervention results suggest a practical direction. More reasoning helps, but does not solve the problem by itself: enabling thinking in the matched Qwen3-14B experiment increases conditional binding by 7.7 points and reduces collapse by 9.9 points, while residual errors remain. Generic "think step by step" prompting is also inconsistent. By contrast, structural separation produces larger and more stable gains. Per-scenario identifiers raise binding from baseline model-macro levels of 60.6% to 78.2% for API models and from 43.4% to 76.4% for open-weight models; explicit state tables also improve binding, and parser-based verification reaches 85.7% and 84.9%, respectively. Scenario splitting performs better still and serves as an ablation of simultaneous state maintenance.

Together, these results suggest that reliable long-context systems should make scenario boundaries explicit instead of expecting models to maintain them implicitly. Retrieved passages can carry version or case identifiers, intermediate outputs can maintain separate state tables, and a verifier can check evidence–value–scenario consistency before downstream decisions. For legal review, compliance, authorization, audit, and agentic workflows, this shifts system design from retrieval alone toward retrieval plus scoped state construction and verification. The present evaluation leaves several questions open. SCENARIOCOLLAPSEBENCH is a mechanism-isolating benchmark; deployment prevalence of scenario-local binding failures remains outside its scope. Its factorized schemas make evidence, values, states, and decisions auditable, but omit ambiguity, incomplete records, heterogeneous formats, institutional procedures, and interactions among more than two evolving cases. The human-written tier reduces template regularity but still derives from deterministic latent traces, while the natural-version study is limited to 200 pairs of UK primary legislation. Future work should therefore test naturally occurring multi-case and multi-version workflows across jurisdictions and domains, including full retrieval and review pipelines.

Our experiments localize the residual error stage more precisely than the underlying mechanism. Oracle interventions place the dominant residual difficulty after evidence localization, in value/state construction, and namespace ablations show that reused identifiers contribute without fully explaining the failure. We use scenario collapse as a behavioral descriptor; distinguishing internal mechanisms requires additional analysis. Prior work shows that entity-state tracking is itself a non-trivial capability and that attention architecture can affect it (Kim & Schuster, 2023; Fagnou et al., 2024). Future work can combine causal interventions with representation analyses to test how models encode scenario identity, state updates, and scope boundaries, and whether neighboring values become entangled during attention, representation, or decoding. Finally, the evaluated contexts cover roughly 2.4k–7.1k reference tokens, with transfer adaptations extending to about 7.4k. The present evaluation targets multi-thousand-token state separation; context-window scaling remains open. LongBench v2 and BABILong suggest that behavior changes as inputs and reasoning chains grow (Bai et al., 2025; Kuratov et al., 2024). Future evaluations should vary context length, the number of simultaneous scenarios, conflict density, update frequency, and the distance between evidence and decision to determine whether binding errors scale differently from retrieval errors. The mitigation results motivate broader validation: per-scenario identifiers, explicit state tables, and verification reduce errors within this benchmark, while the matched Qwen3-14B ablation establishes a reasoning benefit for one checkpoint. Future work should test whether scope-aware training, structured state representations, and verifier-guided inference generalize across architectures and real workflows. Results also depend on prompt templates, parsers, normalization rules, model versions, and inference settings, so exact replication requires run-specific artifacts and dated model configurations.

6 CONCLUSION

This work identifies scenario-local evidence binding as a distinct reliability problem in long-context reasoning: finding the right evidence does not ensure that its value remains attached to the right local world. SCENARIOCOLLAPSEBENCH makes this failure directly measurable through scenario-scoped evidence–value–state traces, FBNB, and pair-level collapse categories. Across factorized, human-written, naturally versioned legal, and adapted long-context settings, the failure persists across model families, while additional reasoning reduces but does not eliminate it. Oracle and intervention results place the dominant residual difficulty in constructing and preserving scenario-local state rather than in evidence access alone. The central implication is that systems reasoning over versions, cases, branches, or parallel records should treat scope preservation as a first-class requirement: reliable retrieval should be followed by explicit state construction and verification, rather than assumed to guarantee correct reasoning.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure. Additionally, we used generative AI tools solely to improve the readability of the manuscript, specifically for grammar correction, sentence-level restructuring, and wording refinement. We have reviewed all AI-assisted edits and verified that they did not alter the scientific content, claims, or conclusions of the paper. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The benchmark uses deterministic latent state traces, and Sections 3.4–3.8 define the primary scoring rules. The appendix documents human-written validation (Appendix D), natural legal-version selection (Appendix E), model and inference settings (Appendix H), parser handling (Appendix I), paired RULER/BABILong construction (Appendix L), and bootstrap procedures (Appendix M). The anonymous supplementary artifact contains the generator, split manifests, prompts and output schemas, scoring code, raw outputs, bootstrap scripts, and run metadata required for exact replication.

REFERENCES

An, C.; Gong, S.; Zhong, M.; Zhao, X.; Li, M.; Zhang, J.; Kong, L.; and Qiu, X. 2024. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. ACL 2024.

Anthropic. 2025. Claude 3.7 Sonnet System Card. Anthropic technical report.

Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2024. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. ACL 2024.

Bai, Y.; Tu, S.; Zhang, J.; Peng, H.; Wang, X.; Lv, X.; Cao, S.; Xu, J.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2025. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. ACL 2025.

Buchmann, J.; Liu, X.; and Gurevych, I. 2024. Attribute or Abstain: Large Language Models as Long Document Assistants. EMNLP 2024.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ: A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. EMNLP 2018.

Dasigi, P.; Lo, K.; Beltagy, I.; Cohan, A.; Smith, N. A.; and Gardner, M. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. NAACL 2021.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. ACL 2020.

Efron, B.; and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; Kumar, A.; Goyal, A.; Ku, P.; and Hakkani-Tur, D. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. LREC 2020.

Fagnou, E.; Caillon, P.; Delattre, B.; and Allauzen, A. 2024. Chain and Causal Attention for Efficient Entity Tracking. EMNLP 2024.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling Large Language Models to Generate Text with Citations. EMNLP 2023.

Gemma Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.

Google DeepMind. 2025. Gemini 2.5 Pro Model Card. Google DeepMind.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Hsieh, C.-P.; Sun, S.; Kriman, S.; Acharya, S.; Rekesh, D.; Jia, F.; Zhang, Y.; and Ginsburg, B. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? First Conference on Language Modeling (COLM 2024).

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Renard Lavaud, L.; Lachaux, M.-A.; Stock, P.; Le Scao, T.; Lavril, T.; Wang, T.; Lacroix, T.; and El Sayed, W. 2023. Mistral 7B. arXiv:2310.06825.

Kim, N.; and Schuster, S. 2023. Entity Tracking in Language Models. ACL 2023.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models Are Zero-Shot Reasoners. NeurIPS 2022.

Kuratov, Y.; Bulatov, A.; Anokhin, P.; Rodkin, I.; Sorokin, D.; Sorokin, A.; and Burtsev, M. 2024. BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack. NeurIPS 2024.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12:157–173.

Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2021. Entity-Based Knowledge Conflicts in Question Answering. EMNLP 2021.

Mathitharan, K. 2026. Version-Aware RAG – UK Legal Corpus. Hugging Face dataset.

Modarressi, A.; Deilamsalehy, H.; Dernoncourt, F.; Bui, T.; Rossi, R. A.; Yoon, S.; and Schuetze, H. 2025. NoLiMa: Long-Context Evaluation Beyond Literal Matching. ICML 2025.

OpenAI. 2025a. Introducing GPT-4.1 in the API. OpenAI technical documentation.

OpenAI. 2025b. o3 Model Documentation. OpenAI technical documentation.

Petroni, F.; Piktus, A.; Fan, A.; Lewis, P.; Yazdani, M.; De Cao, N.; Thorne, J.; Jernite, Y.; Karpukhin, V.; Maillard, J.; Plachouras, V.; Rocktäschel, T.; and Riedel, S. 2021. KILT: A Benchmark for Knowledge Intensive Language Tasks. NAACL 2021.

Qwen Team. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Rashkin, H.; Nikolaev, V.; Lamm, M.; Aroyo, L.; Collins, M.; Das, D.; Petrov, S.; Tomar, G. S.; Turc, I.; and Reitter, D. 2023. Measuring Attribution in Natural Language Generation Models. Computational Linguistics, 49(4):777–840.

Shaham, U.; Segal, E.; Ivgi, M.; Efrat, A.; Yoran, O.; Haviv, A.; Gupta, A.; Xiong, W.; Geva, M.; Berant, J.; and Levy, O. 2022. SCROLLS: Standardized CompaRison Over Long Language Sequences. EMNLP 2022.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: A Large-scale Dataset for Fact Extraction and VERification. NAACL 2018.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. Transactions of the Association for Computational Linguistics, 10:539–554.

Wang, F.; Wan, X.; Sun, R.; Chen, J.; and Arik, S. O. 2025. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. ACL 2025.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2016. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. ICLR 2016.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. EMNLP 2018.

Yen, H.; Gao, T.; Hou, M.; Ding, K.; Fleischer, D.; Izsak, P.; Wasserblat, M.; and Chen, D. 2025. HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly. ICLR 2025.

Yu, W.; Jiang, M.; Clark, P.; and Sabharwal, A. 2023. IfQA: A Dataset for Open-domain Question Answering under Counterfactual Presuppositions. EMNLP 2023.

Zhang, X.; Chen, Y.; Hu, S.; Xu, Z.; Chen, J.; Hao, M.; Han, X.; Thai, Z.; Wang, S.; Liu, Z.; and Sun, M. 2024. InfinityBench: Extending Long Context Evaluation Beyond 100K Tokens. ACL 2024.

A Additional Results

The following tables provide the detailed diagnostic, transfer, intervention, and data-tier results referenced in the main text.

Table 4: Oracle decomposition on a fixed stratified 1,000-pair diagnostic subset, macro-averaged across GPT-4.1, DeepSeek-V3, Gemini 2.5 Pro, and Claude 3.7 Sonnet. Every condition uses the same instance IDs. Progressive oracle information probes where errors remain after evidence access.

Condition	Ev.Loc	Bind.Acc	Dec.Acc	Coll.
Standard prompting	89.3	57.6	66.8	31.1
Oracle evidence ID	100.0	71.4	68.9	18.6
Oracle evidence span	100.0	75.6	83.4	15.7
Oracle evidence value	100.0	93.2	86.1	6.9
Gold state + model rule use	–	–	97.8	0.0
Model state + symbolic rule	89.3	57.6	75.4	31.1
Symbolic oracle parser	–	100.0	100.0	0.0

Table 5: Matched Qwen3-14B thinking ablation on the 1,021-pair human-written tier. The checkpoint, instance IDs, prompt, context, output schema, maximum output length, and decoding parameters are held fixed; only thinking mode is changed. Confidence intervals are paired pair-cluster bootstrap intervals.

Qwen3-14B mode	Ev.Loc	Bind.Acc	CondBind	FBNB	Collapse	Strict Pair Acc
Thinking OFF	87.9	64.8	73.7	23.4	27.1	43.9
Thinking ON	90.4	73.6	81.4	15.6	17.2	56.8
ON–OFF	+2.5	+8.8	+7.7	-7.8	-9.9	+12.9
95% paired CI	[0.7,4.3]	[5.5,12.1]	[4.9,10.6]	[-10.6,-5.1]	[-13.0,-6.7]	[9.4,16.4]

Table 6: External validation on 200 naturally occurring historical UK legal-version pairs. CondBind conditions value correctness on successful localization; CS-FBNB uses the same 246-field common support (61.5% of 400 target fields) for every model. Pair-cluster bootstrap intervals for CondBind, FBNB, and collapse are reported in Appendix M. The evaluation measures observability under natural document revision.

Model	Ev.Loc	Bind.Acc	CondBind	FBNB	CS-FBNB	Collapse
GPT-4.1	92.6	76.4	82.5	13.5	12.6	14.8
DeepSeek-V3	86.9	60.5	69.6	18.6	20.3	27.1
Gemini 2.5 Pro	90.8	73.7	81.2	15.9	14.2	17.5
Claude 3.7 Sonnet	89.4	78.8	88.1	11.2	9.8	20.7
Qwen2.5-32B	88.2	69.6	78.9	16.7	15.4	20.4
Qwen2.5-14B	81.7	61.8	75.6	24.1	25.6	22.9
Llama-3.1-8B	76.8	58.1	75.7	21.3	22.0	27.4
Gemma-2-9B	79.0	52.7	66.7	28.2	30.1	34.8
Qwen2.5-7B	72.5	44.0	60.7	32.9	29.7	38.6
Mistral-7B	71.6	49.2	68.7	31.2	33.3	39.8
DeepSeek-R1-Distill-Qwen-14B	88.9	76.5	86.1	10.8	9.3	11.7
Qwen3-14B Thinking	91.2	81.1	88.9	11.1	10.6	8.9
OpenAI o3	95.0	88.1	92.7	7.4	6.1	5.8

Table 7: FBNB and common-support FBNB on the full benchmark. The common support comprises 58.4% of target fields and is identical across the ten models.

Model	FBNB	CS-FBNB
GPT-4.1	24.6	23.1
DeepSeek-V3	20.9	22.4
Gemini 2.5 Pro	27.4	25.8
Claude 3.7	22.8	21.0
Qwen2.5-32B	27.9	26.7
Llama-3.1-8B	33.9	32.4
Qwen2.5-14B	27.6	29.0
Gemma-2-9B	31.5	33.1
Qwen2.5-7B	36.7	34.8
Mistral-7B	38.2	36.9

Table 8: Pair-level decision correctness. Both, One, and Neither count pairs with two, one, or zero correct scenario decisions. "Both" is strict pair accuracy; the final column is mean per-scenario accuracy within each pair.

Model	Both	One	Neither	Mean Dec. Acc
GPT-4.1	50.7	32.4	16.9	66.9
DeepSeek-V3	52.0	29.6	18.4	66.8
Gemini 2.5 Pro	48.9	27.8	23.3	62.8
Claude 3.7	54.0	28.4	17.6	68.2
Qwen2.5-32B	45.1	29.2	25.7	59.7
Llama-3.1-8B	32.8	33.6	33.6	49.6
Qwen2.5-14B	36.2	33.4	30.4	52.9
Gemma-2-9B	27.4	34.8	37.8	44.8
Qwen2.5-7B	24.3	35.2	40.5	41.9
Mistral-7B	24.2	37.0	38.8	42.7

Table 9: RULER original vs. paired-local-world adaptation. CONDBIND is value correctness conditional on successful localization. Pair-cluster bootstrap intervals for CONDBIND and FBNB are reported in Appendix M.

Model	Original Acc	Mean Dec. Acc	Ev.Loc	Bind.Acc	CondBind	FBNB
GPT-4.1	94.8	82.7	93.6	78.1	83.4	12.4
DeepSeek-V3	90.9	78.4	89.4	73.5	82.2	14.7
Gemini 2.5 Pro	93.7	81.2	92.5	76.2	82.4	13.6
Claude 3.7	95.1	85.8	94.0	82.4	87.7	9.8
Qwen2.5-32B	89.6	75.6	88.1	69.8	79.2	16.8
Qwen2.5-14B	85.4	70.9	84.2	65.7	78.0	18.9
Qwen2.5-7B	77.2	61.8	76.9	55.9	72.7	24.1
Llama-3.1-8B	79.5	64.3	79.8	59.4	74.4	22.7
Gemma-2-9B	75.8	57.9	74.3	51.6	69.4	27.9
Mistral-7B	73.9	58.7	72.7	52.8	72.6	27.4
DeepSeek-R1-Distill-Qwen-14B	92.3	84.1	91.5	81.0	88.5	8.9
Qwen3-14B Thinking	93.4	86.7	92.7	83.7	90.3	9.7
OpenAI o3	97.1	92.8	96.5	90.6	93.9	5.9

Table 10: BABILong original vs. paired-local-world adaptation. CONDBIND is value correctness conditional on successful localization. Pair-cluster bootstrap intervals for CONDBIND and FBNB are reported in Appendix M.

Model	Original Acc	Mean Dec. Acc	Ev.Loc	Bind.Acc	CondBind	FBNB
GPT-4.1	88.9	75.2	87.4	68.9	78.8	15.9
DeepSeek-V3	84.3	68.5	82.7	61.8	74.7	20.5
Gemini 2.5 Pro	86.6	72.9	85.9	66.5	77.4	18.7
Claude 3.7	89.5	78.4	88.2	72.6	82.3	13.4
Qwen2.5-32B	82.2	66.1	80.1	59.7	74.5	21.8
Qwen2.5-14B	77.8	61.7	76.4	55.2	72.3	24.1
Qwen2.5-7B	68.4	49.8	67.9	42.8	63.0	32.3
Llama-3.1-8B	71.9	53.6	71.5	47.1	65.9	29.7
Gemma-2-9B	66.7	46.9	65.6	39.4	60.1	34.6
Mistral-7B	63.8	45.7	63.4	40.7	64.2	33.1
DeepSeek-R1-Distill-Qwen-14B	87.6	79.1	86.9	75.5	86.9	10.4
Qwen3-14B Thinking	89.2	81.8	88.1	78.9	89.6	10.4
OpenAI o3	94.8	89.4	93.7	86.8	92.6	6.8

Table 11: Ablations and interventions for scenario separation, reported as model-macro means ± pair-cluster bootstrap 95% half-widths. API = GPT-4.1, DeepSeek-V3, Gemini 2.5 Pro, Claude 3.7 Sonnet; OW = Qwen2.5-32B/14B/7B, Llama-3.1-8B, Gemma-2-9B, Mistral-7B. Each model is evaluated on all 10,132 benchmark pairs. Scenario splitting is an ablation rather than a mitigation under the original paired condition.

Category	Method	API Bind	OW Bind	API Coll	OW Coll
Baseline	Standard	60.6 ± 1.8	43.4 ± 1.6	27.2 ± 1.5	38.0 ± 1.7
Generic reasoning	Think step by step	57.1 ± 1.9	47.8 ± 1.7	30.9 ± 1.6	36.4 ± 1.7
Retrieval support	Retrieval buffer	72.6 ± 1.6	48.9 ± 1.7	24.0 ± 1.4	29.7 ± 1.6
Boundary	Strong delimiters	68.4 ± 1.7	48.3 ± 1.7	19.7 ± 1.3	35.2 ± 1.7
Namespace	Per-scenario IDs	78.2 ± 1.4	76.4 ± 1.5	12.8 ± 1.1	16.1 ± 1.2
Explicit state construction	State table	70.9 ± 1.6	66.2 ± 1.6	14.3 ± 1.2	26.8 ± 1.5
Explicit state verification	Parser + verifier	85.7 ± 1.2	84.9 ± 1.3	7.6 ± 0.9	6.2 ± 0.8
Ablation	Scenario split	90.8 ± 1.0	78.5 ± 1.4	4.3 ± 0.7	11.6 ± 1.0

Table 12: Data-tier breakdown, macro-averaged across the ten-model full-benchmark suite after computing each model's score over all instances in the tier. Decision Acc. is mean per-scenario decision accuracy.

Tier	Pairs	Decision Acc.	Ev.Loc	Bind.Acc	FBNB
Factorized syn.	6,117	65.2	91.7	60.4	22.9
Naturalized syn.	2,994	63.4	78.2	58.6	23.1
Human-written	1,021	51.6	85.9	44.3	35.7

B Limitations and Scope

Limitations, scope, and future directions are integrated into Section 5.

C Task Families and Data Tiers

Table 13: The six structured decision task families in ScenarioCollapseBench.

Task family	State variables	Decision type
Dataset export authorization	role, clearance, destination, dataset sensitivity	access allowed / denied
Vendor payment approval	role, spending limit, amount, vendor risk, region	payment approved / rejected
Software workflow authorization	role, environment clearance, release status, deployment target	permitted / blocked
Access-control review	user role, group membership, exception status	access granted / revoked
Deployment approval	release status, reviewer approval, environment, incident flag	deployment allowed / paused
Data-export request audit	requester status, legal basis, region, data class	compliant / non-compliant

Table 14: Benchmark data tiers, consistent with Section 3.3 and Table 12.

Tier	Paired	Single
Factorized synthetic	6,117	6,117
Naturalized synthetic	2,994	2,994
Human-written	1,021	1,021
Total	10,132	10,132

D Human-Written Tier Protocol

The human-written tier contains 1,021 retained paired instances. Each instance begins from a latent state trace generated by the deterministic state machine. Rewrite annotators convert the trace into naturalistic audit-log, workflow-note, compliance-review, deployment, or access-control prose while preserving local evidence IDs, authoritative status, final values, and rule-relevant facts. Rewrite annotators see the paired traces because their task is to preserve the targeted scenario difference

while varying surface form. Separate validation annotators, disjoint from the rewrite annotators, independently recover final state, decisive evidence, and decision without seeing gold labels. Each instance is labeled by three validators and retained only when the final state, evidence set, and decision are recoverable after adjudication. The retained set contains 170 pairs for each of the first five task families and 171 data-export audit pairs, for a total of 1,021. Initial writing produced 1,294 pairs; 273 were excluded (21.1%). Final-decision kappa is 0.82, final-state exact agreement is 0.78, evidence-set exact agreement is 0.74, and evidence-ID micro-F1 is 0.86.

E Natural Legal Version Validation Protocol

The external-validation corpus is the Version-Aware RAG UK Legal Corpus (Mathitharan, 2026), which stores point-in-time snapshots of five UK Public General Acts from legislation.gov.uk. The corpus contains 96 Parts and 355 historical Part-level versions. Candidate pairs are drawn from versions of the same Act and Part and screened for a substantive change affecting a version-specific definition, eligibility criterion, exception, threshold, or obligation. Two screeners independently judge substantive relevance; disagreements are adjudicated before annotation. Of 248 pairs passing the initial substantive-change screen, 200 are retained after gold annotation and validation, with the retained set stratified across the five Acts and change types.

For each retained pair, two annotators independently identify the decisive provision and record the version-specific value/state and answer without rewriting the source text. Disagreements are adjudicated by a third annotator. Before adjudication, exact agreement is 0.89 for decisive-provision selection, 0.91 for normalized value/state, and 0.94 for the final answer. The scenarios are the two original legal snapshots; their text is not rewritten to create near-duplicates. Each released item records the source/version date, decisive provision, gold value/state, answer, and annotation provenance. This evaluation tests whether right-evidence/wrong-version binding failures occur under natural document revision rather than estimating real legal error rates. As reported in Section 4.7 and Table 6, the common support contains 246 of 400 target fields (61.5%). Every compared model successfully localizes every field in this intersection, so CS-FBNB uses an identical denominator across models.

F Extra Citation Scoring

Extra citations are ignored only when unrelated to the evaluated field and non-conflicting. They are penalized when they support a contradictory same-attribute value, are stale non-authoritative records for that attribute, or belong to the neighboring scenario namespace. Correct value without required scenario-local evidence does not receive binding credit. Wrong-scenario evidence does not receive localization credit even when the surface identifier matches.

G Collapse Taxonomy

Collapse is evaluated on pairs whose gold states differ. The normalized predicted states are compared with the two gold state objects. Categories are assigned in order: correct separation; collapse to A; collapse to B; swap; distractor binding; shared invalid; and other error. Narrow CollapseRate counts only collapse to A or B. These state-based categories are distinct from the decision-correctness categories in Table 8.

H Models and Inference Settings

Ten models are evaluated on the full factorized benchmark. DeepSeek-R1-Distill-Qwen-14B, Qwen3-14B Thinking, and OpenAI o3 extend model coverage on validated subsets, natural-version data, and transfer experiments. Section 4.6 also reports the matched Qwen3-14B thinking on/off ablation on the 1,021-pair human-written tier. Tables 6–10 report all thirteen configurations where coverage is shared. The default task prompt is shared across models unless otherwise specified. Greedy or deterministic decoding is used where supported; open-weight models use temperature 0.0 and top-p 1.0.

Table 15: API model identifiers.

Model	Provider	Identifier
GPT-4.1	OpenAI	`gpt-4.1`
DeepSeek-V3	DeepSeek	`deepseek-chat`
Gemini 2.5 Pro	Google	`gemini-2.5-pro`
Claude 3.7 Sonnet	Anthropic	`claude-3-7-sonnet-20250219`
OpenAI o3	OpenAI	`o3-2025-04-16`

Table 16: Open-weight model identifiers and inference modes.

Model	Checkpoint / family	Inference mode
Qwen2.5-32B	`Qwen/Qwen2.5-32B-Instruct-AWQ`	AWQ
Qwen2.5-14B	`Qwen/Qwen2.5-14B-Instruct`	4-bit
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	4-bit
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	4-bit
Gemma-2-9B	`google/gemma-2-9b-it`	4-bit
Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.3`	4-bit
DeepSeek-R1-Distill-Qwen-14B	`deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`	reasoning
Qwen3-14B Thinking	`Qwen/Qwen3-14B`	thinking enabled

Confidence intervals use 10,000 pair-cluster bootstrap resamples (Efron and Tibshirani, 1993). Paired instance IDs are sampled with replacement, retaining both scenarios and all required fields from each sampled pair; percentile 95% intervals are computed from the empirical bootstrap distribution using random seed 20250317. For matched on/off and original-versus-paired comparisons, the same sampled IDs are used in both conditions before taking the difference.

Table 17: Run-level reproducibility settings. API dates identify the evaluation window. Local open-weight runs use vLLM 0.8.5 with weight-only 4-bit quantization where shown; Qwen2.5-32B uses AWQ and two A100-80GB GPUs, while 7B–14B models use one A100-80GB GPU. Model-specific checkpoint revisions and complete environment manifests are included in the supplementary artifact.

Model	Exact identifier / checkpoint	Run date	Max ctx used	Max output	Decoding	Backend / reasoning
GPT-4.1	`gpt-4.1`	2026-06-15	32768	2048	$T = 0, p = 1$	OpenAI API
DeepSeek-V3	`deepseek-chat`	2026-06-15	32768	2048	$T = 0, p = 1$	DeepSeek API
Gemini 2.5 Pro	`gemini-2.5-pro`	2026-06-15	32768	2048	deterministic	Google API
Claude 3.7 Sonnet	`claude-3-7-sonnet-20250219`	2026-06-15	32768	2048	$T = 0$	Anthropic API
OpenAI o3	`o3-2025-04-16`	2026-06-15	32768	4096	provider deterministic	reasoning effort=medium
Qwen2.5-32B	`Qwen/Qwen2.5-32B-Instruct-AWQ`	2026-06-20	32768	2048	$T = 0, p = 1$	vLLM 0.8.5; AWQ; 2×A100-80GB
Qwen2.5-14B	`Qwen/Qwen2.5-14B-Instruct`	2026-06-20	32768	2048	$T = 0, p = 1$	vLLM 0.8.5; 4-bit; 1×A100-80GB
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	2026-06-20	32768	2048	$T = 0, p = 1$	vLLM 0.8.5; 4-bit; 1×A100-80GB
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	2026-06-20	32768	2048	$T = 0, p = 1$	vLLM 0.8.5; 4-bit; 1×A100-80GB
Gemma-2-9B	`google/gemma-2-9b-it`	2026-06-20	8192	768	$T = 0, p = 1$	vLLM 0.8.5; 4-bit; 1×A100-80GB
Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.3`	2026-06-20	32768	2048	$T = 0, p = 1$	vLLM 0.8.5; 4-bit; 1×A100-80GB
DeepSeek-R1-Distill-Qwen-14B	`deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`	2026-06-20	32768	8192	$T = 0, p = 1$	vLLM 0.8.5; reasoning
Qwen3-14B Thinking	`Qwen/Qwen3-14B`	2026-06-20	32768	8192	$T = 0, p = 1$	vLLM 0.8.5; thinking budget 6144

I Parser Failure Handling

Parser failures are counted as incorrect for metrics requiring missing fields. Their handling is part of the evaluation protocol; a missing required value or evidence field does not receive binding credit. Table 18 reports parser-failure rates separately so schema compliance can be distinguished from semantic errors.

Table 18: Parser-failure rate (%). "Primary set" denotes the model's main reported benchmark/subset; "Transfer avg." is the mean parser-failure rate across the paired RULER and BABILong runs.

Model	Primary set	Legal versions	Transfer avg.
GPT-4.1	0.2	0.0	0.1
DeepSeek-V3	0.6	0.5	0.7
Gemini 2.5 Pro	0.3	0.0	0.2
Claude 3.7 Sonnet	0.2	0.0	0.1
Qwen2.5-32B	0.8	0.5	0.9
Qwen2.5-14B	1.1	1.0	1.4
Qwen2.5-7B	1.8	1.5	2.0
Llama-3.1-8B	1.5	1.0	1.7
Gemma-2-9B	2.6	2.0	2.8
Mistral-7B	3.2	2.5	3.5
DeepSeek-R1-Distill-Qwen-14B	1.0	0.5	0.8
Qwen3-14B Thinking	0.7	0.0	0.6
OpenAI o3	0.1	0.0	0.0

J Illustrative Diagnostic Instance and Scoring

Task. Vendor payment approval. Approve if the requester's current spending limit covers the invoice and vendor risk is not high. Authoritative records update the state; non-authoritative records are distractors.

Scenario A. Invoice: EUR 18,000. E07, an authoritative limit update, sets Nora's limit to EUR 15,000. E11, a non-authoritative memo, describes an earlier EUR 25,000 limit. E14, an authoritative risk update, states that Northbay Analytics is medium risk.

Scenario B. Invoice: EUR 18,000. E07 sets Nora's limit to EUR 25,000. E11 describes an earlier EUR 15,000 limit. E14 states that Northbay Analytics is medium risk.

Gold and illustrative error. The gold decisions are reject for A and approve for B. An illustrative failure cites E07 in each scenario but assigns EUR 25,000 to both spending-limit fields, with vendor risk otherwise correct. The A spending-limit field has localization $= 1$, binding $= 0$, and FBNB $= 1$. The pair collapses to B. This is a constructed diagnostic illustration, not a verbatim model-output transcript.

K Scenario-Separation Controls

K.1 State Differences and Decision Agreement

The benchmark distinguishes different-state/different-decision pairs from different-state/same-decision and same-decision/conflicting-value conditions. Decision agreement does not imply identical local states: different values can fall on the same side of a downstream threshold. State binding and final decision accuracy must therefore be scored separately. Table 8 reports decision correctness, whereas binding and narrow collapse measure state-related behavior.

K.2 Boundaries and Namespace-Separated Identifiers

Shared local IDs are interpreted within their scenario objects. Explicit namespace separation uses distinct identifiers such as A-E07 and B-E07. Table 19 reproduces the corresponding results from Table 11, retaining separate API and open-weight (OW) summaries. Boundary and namespace interventions improve binding and reduce collapse in both groups; the reported effects do not eliminate all errors.

Table 19: Boundary and namespace ablations, reproduced from Table 11. Values are percentages.

Condition	API Bind	OW Bind	API Coll	OW Coll
Standard	60.6	43.4	27.2	38.0
Strong delimiters	68.4	48.3	19.7	35.2
Per-scenario IDs	78.2	76.4	12.8	16.1

L Paired Adaptation Protocol for Existing Benchmarks

For RULER and BABILong, we sample 200 source instances from each benchmark, as described in Section 4.8. The comparison uses the same underlying tasks in original and paired conditions. Original instances retain their single-local-world form. Paired items combine two source instances with the same task schema and different target values as Scenario A and Scenario B. Scenario order is randomized, explicit boundaries are retained, and both scenarios use the same output schema.

Pair construction. Candidate pairs must satisfy three constraints: (i) the underlying task type is identical, (ii) the gold target values differ, and (iii) the source instances are matched as closely as possible in context length and distractor structure so that pairing does not introduce a new task family. For RULER, the decisive support unit is the source-local key, target span, or trace item used by the original task. For BABILong, the decisive support unit is the state-changing fact or update required for the final answer. These source-local support units are namespace-scoped after pairing and are used to compute EvidenceLoc and binding metrics.

Matched evaluation. Every paired item is evaluated together with its two constituent original items. This produces a within-source comparison: the underlying retrieval or state-tracking problem remains the same, while the paired condition adds the requirement to preserve two conflicting local worlds simultaneously. We therefore interpret the change from original accuracy to mean paired-scenario decision accuracy together with EvidenceLoc, BindingAcc, CondBind, and FBNB in the paired condition.

Illustration. A RULER retrieval task with target value v_A is paired with a matched task whose target v_B differs. The model returns each value under its scenario namespace. A BABILong state-tracking task is paired with a matched world whose decisive update differs, producing conflicting final states under the same reasoning schema.

Table 20: Post-localization binding on the full benchmark. Intervals use the pair-cluster bootstrap described in Appendix H.

Model	CondBind	95% CI
GPT-4.1	64.0	[60.8,67.2]
DeepSeek-V3	71.8	[69.0,74.6]
Gemini 2.5 Pro	62.5	[60.1,64.9]
Claude 3.7 Sonnet	72.7	[70.4,75.0]
Qwen2.5-32B	61.5	[59.1,63.9]
Llama-3.1-8B	62.5	[59.8,65.2]
Qwen2.5-14B	67.1	[64.3,69.9]
Gemma-2-9B	42.9	[39.5,46.3]
Qwen2.5-7B	60.5	[57.2,63.8]
Mistral-7B	43.3	[39.8,46.8]

Table 21: Length-stratified results on the full benchmark, macro-averaged across the ten-model full-benchmark suite. Bins use the same reference tokenizer as Table 2. The degradation with length is descriptive; context length is correlated with event count and distractor density in this diagnostic suite.

Reference-token bin	Pairs	CondBind	FBNB	Collapse
$\leq$ 4k	2,960	69.1	19.7	24.1
4–6k	4,734	62.8	25.6	31.4
$>$ 6k	2,438	56.7	31.9	36.8

M Additional Statistical Details

M.1 Full-Benchmark Conditional Binding Intervals

M.2 Length-Stratified Robustness

M.3 External-Validation Confidence Intervals

Table 22: Pair-cluster bootstrap 95% confidence intervals for the natural legal-version evaluation.

Model	CondBind 95% CI	FBNB 95% CI	Collapse 95% CI
GPT-4.1	[78.6,86.4]	[10.1,16.9]	[9.9,19.7]
DeepSeek-V3	[64.7,74.5]	[14.4,22.8]	[20.9,33.3]
Gemini 2.5 Pro	[77.1,85.3]	[12.1,19.7]	[12.2,22.8]
Claude 3.7 Sonnet	[84.6,91.6]	[7.8,14.6]	[15.0,26.4]
Qwen2.5-32B	[74.4,83.4]	[12.8,20.6]	[14.8,26.0]
Qwen2.5-14B	[70.6,80.6]	[19.4,28.8]	[17.1,28.7]
Llama-3.1-8B	[70.4,81.0]	[16.5,26.1]	[21.2,33.6]
Gemma-2-9B	[61.1,72.3]	[22.9,33.5]	[28.2,41.4]
Qwen2.5-7B	[54.5,66.9]	[27.2,38.6]	[31.8,45.4]
Mistral-7B	[62.7,74.7]	[25.5,36.9]	[32.9,46.7]
DeepSeek-R1-Distill-Qwen-14B	[82.2,90.0]	[7.4,14.2]	[7.2,16.2]
Qwen3-14B Thinking	[85.2,92.6]	[7.6,14.6]	[4.9,12.9]
OpenAI o3	[89.8,95.6]	[4.4,10.4]	[2.6,9.0]

M.4 Transfer Confidence Intervals

Table 23: Pair-cluster bootstrap 95% confidence intervals for the paired RULER and BABILong adaptations.

Model	RULER CondBind CI	RULER FBNB CI	BABILong CondBind CI	BABILong FBNB CI
GPT-4.1	[79.6,87.2]	[9.0,15.8]	[74.3,83.3]	[11.9,19.9]
DeepSeek-V3	[78.1,86.3]	[10.9,18.5]	[69.9,79.5]	[15.9,25.1]
Gemini 2.5 Pro	[78.4,86.4]	[9.9,17.3]	[72.8,82.0]	[14.2,23.2]
Claude 3.7	[84.2,91.2]	[6.6,13.0]	[78.0,86.6]	[9.4,17.4]
Qwen2.5-32B	[74.8,83.6]	[12.7,20.9]	[69.7,79.3]	[17.0,26.6]
Qwen2.5-14B	[73.3,82.7]	[14.4,23.4]	[67.1,77.5]	[19.0,29.2]
Qwen2.5-7B	[67.1,78.3]	[18.9,29.3]	[56.9,69.1]	[26.5,38.1]
Llama-3.1-8B	[69.1,79.7]	[17.7,27.7]	[60.0,71.8]	[24.1,35.3]
Gemma-2-9B	[63.5,75.3]	[22.4,33.4]	[53.8,66.4]	[28.7,40.5]
Mistral-7B	[66.8,78.4]	[22.0,32.8]	[58.0,70.4]	[27.3,38.9]
DeepSeek-R1-Distill-Qwen-14B	[84.8,92.2]	[5.8,12.0]	[82.8,91.0]	[7.0,13.8]
Qwen3-14B Thinking	[86.8,93.8]	[6.4,13.0]	[85.6,93.6]	[7.0,13.8]
OpenAI o3	[91.0,96.8]	[3.2,8.6]	[89.3,95.9]	[3.8,9.8]