CONSISTENCY CHECKING IN INTERACTIVE NARRATIVES AT MILLION-TOKEN SCALE: A BENCHMARK AND A DUAL-ROUTE EVALUATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are used in text-adventure games and role-playing applications, where narrative histories can span hundreds or thousands of turns. Maintaining factual consistency over such long histories remains challenging; failures to do so can introduce contradictions that disrupt gameplay and reduce immersion. Reliable evaluation is therefore essential, yet the effectiveness of consistency checkers at this scale remains underexplored. To address this gap, we introduce **LINCONBENCH** (**L**ong **I**nteractive **N**arrative **Con**sistency **Bench**mark), to our knowledge the first benchmark for contradiction detection and turn-level localization in interactive narratives extending to thousands of turns. It comprises 2,000 error-injected narrative variants derived from 50 human-authored Chinese novels across five source-length buckets, ranging from below 64K to 512K–1M tokens. Constructing such a benchmark requires reliable error annotations, yet exhaustively identifying naturally occurring contradictions is labor-intensive and prone to omissions or inconsistent judgments. We therefore adopt a two-stage pipeline that constructs multi-turn interaction histories while preserving their narration, then injects controlled contradictions with annotated error locations and supporting historical evidence. Each variant contains a single injected contradiction expressed in one or more turns, with contradiction validity and error-turn annotations verified by human reviewers. We further propose **DRPV**, which combines dual-route candidate discovery with proposition consolidation and full-prefix verification. Our experiments with nine baselines show that evaluators struggle to localize all annotated error turns, particularly in longer narratives and when historical evidence is more distant, whereas DRPV performs better across multiple metrics.

1 INTRODUCTION

In text-adventure games and role-playing applications, large language models can serve as narrative engines or character agents (Wang et al., 2024; Yu et al., 2025). Repeated interactions can produce narrative histories spanning hundreds or thousands of turns. Throughout these interactions, these systems must track world facts, including character identities, survival status, item ownership, abilities, and locations (Ran et al., 2025; He et al., 2026). Such facts may constrain later player actions even when absent from recent dialogue. Maintaining consistency therefore requires both recalling earlier facts and tracking subsequent state changes. For example, a companion explicitly confirmed dead should not later reappear alive without an explanation supported by the narrative. Such contradictions undermine players' trust in the story world, disrupting gameplay and reducing immersion. We study *narrative self-consistency*: whether later statements remain consistent with previously established facts, accounting for state changes supported by the narrative.

Reliable consistency evaluation is essential for comparing and improving these narrative engines. Evaluators may overestimate consistency by missing contradictions or underestimate it by mistaking valid state changes for errors. As histories grow, LLM-based evaluators must identify relevant earlier facts scattered across multiple turns and assess later statements against them while accounting for intervening state changes. Prior work shows that long-context performance is sensitive to information position (Liu et al., 2024; Yi et al., 2026), input length (Hsieh et al., 2024; Du et al., 2025), and reasoning complexity (Modarressi et al., 2025; Gupta et al., 2025). We therefore inves-

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.

tigate whether existing evaluators can reliably detect semantic contradictions and localize the turns that explicitly express them across increasingly long interaction histories.

Existing benchmarks study different aspects of long-context understanding and narrative consistency. General long-context benchmarks such as LongBench (Bai et al., 2024), LongBench v2 (Bai et al., 2025), and BABILong (Kuratov et al., 2024) evaluate understanding and reasoning over extended inputs, with the latter two reaching the million-token scale. ConStory-Bench (Li et al., 2026) evaluates LLM-generated narratives, with model-wise mean lengths ranging from 1.2K to 13.4K words, and assesses its checker on a separate error-injected diagnostic set. FlawedFictions (Ahuja et al., 2025) evaluates plot-hole detection and localization in counterfactually edited standalone stories, with its longer split averaging 2.7K words. In a related interactive narrative setting, NCP-Bench (Ma et al., 2026) primarily evaluates whether narrator agents maintain factual consistency and adhere to narrative commitments over up to 100 turns. These benchmarks advance the evaluation of long-context understanding and narrative consistency. However, interactive narrative consistency benchmarks at million-token scale remain limited. It therefore remains unclear how reliably consistency evaluators can detect contradictions and localize all erroneous turns in such histories.

To address this gap, we introduce **LINCONBENCH**, to our knowledge the first benchmark combining **(1) Consistency evaluation** through contradiction detection and turn-level localization, **(2) Long contexts** at the million-token scale, and **(3) Interactive histories** structured as player queries and narrative responses. It comprises 2,000 error-injected narrative variants derived from 50 human-authored Chinese novels across five source-length buckets, ranging from below 64K to 512K–1M tokens. Reliable evaluation requires accurate annotations of contradictions and their locations. Obtaining these annotations from naturally occurring errors in LLM-generated histories would require reviewers to search long texts, trace supporting historical evidence, and distinguish contradictions from legitimate state changes. Such manual review can be time-consuming (Bonomo et al., 2025; Yang & Jin, 2025), and expert reviewers may overlook inconsistencies or reach conflicting judgments (Subbiah et al., 2024; 2025). We therefore develop a two-stage pipeline comprising *multi-turn narrative construction* and *controlled error injection*. First, we segment each novel around narrative events, refine boundaries to place necessary context before interactions, and reverse-generate player queries while preserving the original narration as responses. Second, we extract persistent facts from early turns and edit one or more later responses to contradict a selected fact while retaining the original historical evidence. Each case is constructed separately from its unmodified interaction sequence, with error locations and supporting evidence recorded for targeted verification. A team of 20 human reviewers participated in full-text review of the 50 pre-injection interaction histories and validation of the 2,000 injected variants.

We further introduce **DRPV** (Dual-Route Proposition Verification), which combines dual-route candidate discovery with proposition consolidation and verification. Selected conflict propositions are checked against the complete interaction history up to their target turns, accounting for intervening state changes supported by the narrative. We evaluate representative baselines and DRPV under a common scoring protocol for contradiction detection and error-turn localization. DRPV leads evaluated methods in overall Turn-R, Turn-F1, and Complete Hit, although complete error-turn recovery remains difficult, particularly in the longest source-length bucket. Our contributions are threefold:

- **A Human-Verified Benchmark.** We introduce LINCONBENCH, a benchmark of long interactive narratives with controlled contradictions and human-verified annotations of erroneous turns and supporting historical evidence.
- **A Consistency Evaluator.** We propose DRPV, which integrates complementary candidate discovery routes with proposition consolidation and verification against complete narrative prefixes for contradiction detection and turn-level localization.
- **A Systematic Empirical Analysis.** Under a common evaluation protocol, we characterize evaluator limitations across source lengths and evidence distances and examine the contributions of DRPV's components through ablation studies.

2 RELATED WORK

Consistency Evaluation Methods. We organize related approaches into four categories. (1) *Fact-ledger-guided auditing*: NCP (Ma et al., 2026) checks narrator responses against an evolving fact

Table 1: Comparison of existing benchmarks with LINCONBENCH across key dimensions.

Benchmark	Primary evaluation target	Context source	Interactive narrative	Million-token scale
LongBench v2 (Bai et al., 2025)	Long-context reasoning	Mixed	✗	✓
FlawedFictions (Ahuja et al., 2025)	Story consistency	Human-authored	✗	✗
ConStory-Bench (Li et al., 2026)	Story consistency	LLM-generated	✗	✗
NCP-Bench (Ma et al., 2026)	Narrator consistency	LLM-generated	✓	✗
LINCONBENCH	**Story Consistency**	**Human-authored**	✓	✓

ledger and narrative commitments. (2) *Context selection and retrieval*: ArcANE (Song et al., 2026) uses character arcs to provide context appropriate to a character's stage in the narrative, while NarrativeFactScore (Jeong et al., 2025) retrieves scenes and character relationships to verify atomic facts in narrative summaries. (3) *Turn-level auditing*: WSE (Chen et al., 2026) evaluates each step using multiple judges that examine prior facts, conflicting claims, and possible state transitions. (4) *Candidate generation and verification*: FlawedFictions (Ahuja et al., 2025) explores a detector–verifier pipeline for plot holes, while ConStory-Checker (Li et al., 2026) extracts candidate spans, identifies contradictory pairs, and records supporting evidence. DRPV combines dual-route candidate discovery with proposition consolidation and verification against complete narrative prefixes.

Evaluation Benchmarks. General multi-turn benchmarks assess response quality, instruction following, and coherence across dialogue turns (Zheng et al., 2023; He et al., 2024; Deshpande et al., 2025). Long-context benchmarks assess recall, information integration, and long-range reasoning over million-token-scale texts (Bai et al., 2025; Wu et al., 2025a; Tavakoli et al., 2026). Writing and interactive narrative benchmarks assess text quality, game-rule adherence, and state tracking (Wu et al., 2025b; Lin et al., 2026; Yu et al., 2025). More closely related to our work, ConStory-Bench (Li et al., 2026) examines consistency errors in standalone narratives. FlawedFictions (Ahuja et al., 2025) evaluates plot-hole detection and localization in counterfactually edited standalone stories, while NCP-Bench (Ma et al., 2026) primarily evaluates the consistency of narrator agents in multi-turn interactions. Building on these efforts, LINCONBENCH evaluates consistency checkers over million-token-scale interactive histories, with explicit annotations of erroneous turns and supporting historical evidence. These annotations support systematic evaluation of contradiction detection and error-turn localization across narrative lengths and evidence distances.

3 LINCONBENCH

Assessing consistency evaluators over long interaction histories requires narratives with known contradictions. We select 50 Chinese novels from CNNovel125K (Ronsor, 2023)[1], with 10 in each of five source-length buckets. These buckets are based on token counts of the preprocessed novels. As illustrated in Figure 1, we construct LINCONBENCH from these novels in two stages: multi-turn narrative construction and controlled error injection. Human review comprises two stages. Before injection, reviewers examine 50 complete interaction histories, including generated player queries and narrative responses, for pre-existing semantic contradictions. Each history yields 40 separately constructed variants, followed by validation of each injected contradiction and its error-turn annotations. Twenty reviewers participate across both stages. Appendix A.3 details both review stages.

3.1 MULTI-TURN NARRATIVE CONSTRUCTION

We convert each preprocessed novel X into a multi-turn interaction sequence by preserving its narration as responses and generating player queries. The process comprises event-based segmentation, boundary refinement, and reverse query generation. To guide event attribution and query generation, we construct a protagonist profile π. Appendices A.2 and A.1 provide a worked construction example and details on protagonist identification, respectively.

Event-Based Segmentation. Drawing on the story grammar of Mandler & Johnson (1977), we use the setup, development, and outcome of episodes to guide segmentation. Conditioned on π,

[1] https://huggingface.co/datasets/RyokoAI/CNNovel125K

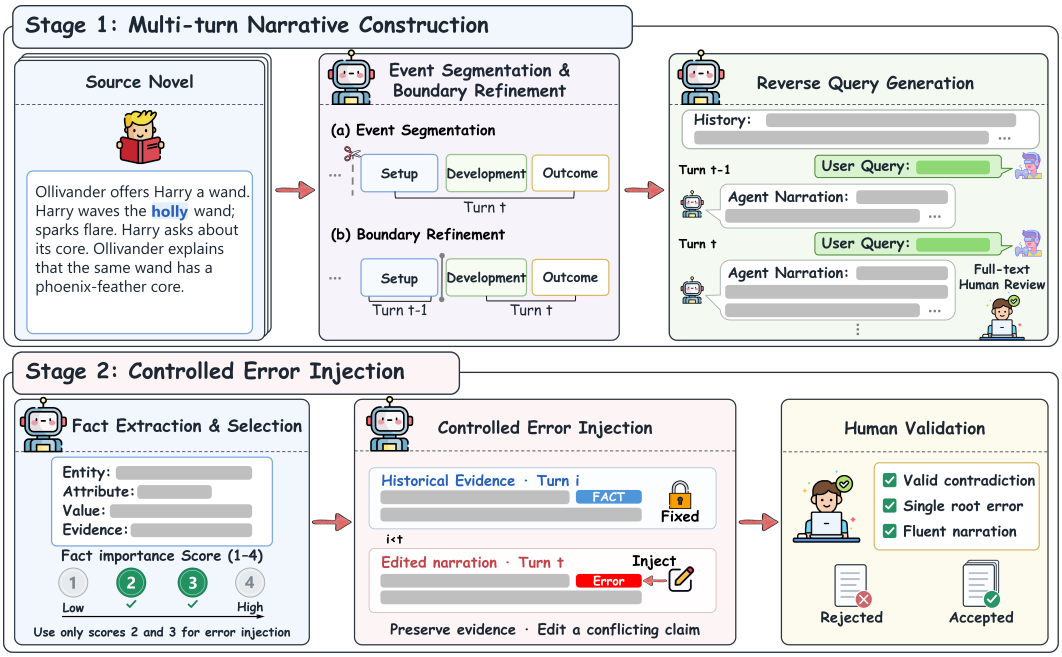

Figure 1: Overview of the LINCONBENCH construction pipeline. All interaction histories undergo full-text human review before injection; variants undergo contradiction and error-turn validation.

an LLM identifies episodes within each chapter, including protagonist-initiated actions and external events affecting the protagonist. We place initial boundaries where episode development begins, so that the preceding setup can inform a player query. This yields N initial passages $(\tilde{a}_1, \ldots, \tilde{a}_N)$.

Interaction-Aware Boundary Refinement. An initial passage may still begin with context needed to motivate the next player query. We refine these boundaries using narrative beats, small units of dramatic action within scenes (Mochocki & Koskimaa, 2021). Inspired by the LLM-based beat segmentation in Narrix (Zhang et al., 2026), we extract ordered, source-anchored beats for each passage. Active beats describe actions initiated by the protagonist; passive beats describe events not initiated by them. Each beat receives an agency label and a summary, with the model prompted to omit outcomes and describe active beats as high-level intentions.

We process the boundaries in narrative order, using the current passage texts at each step. For each passage after the first, an LLM compares its leading passive beats with the tail of the preceding passage. It selects a contiguous prefix introducing context not yet available there, stopping at the first active beat or a passive beat that merely continues an established situation. The selected source prefix p_i is moved to the preceding passage:

$$(\tilde{a}_{i-1}, \tilde{a}_i) \longrightarrow (\tilde{a}_{i-1} \oplus p_i, r_i), \qquad \tilde{a}_i = p_i \oplus r_i, \tag{1}$$

where r_i is the remaining text and $\oplus$ denotes text concatenation. If no prefix is selected, the boundary remains unchanged; if the entire passage is transferred, r_i is empty. After refinement, we remove empty passages and reindex the remaining ones as $(a_1, \ldots, a_T)$, where $T \leq N$. We update the beat annotations to reflect the refined passage boundaries. This procedure preserves content and order, apart from separator whitespace introduced during concatenation.

Reverse Query Generation. For each refined passage a_t, $t \geq 2$, we generate a player query q_t while keeping a_t as its fixed narrative response. The historical context H_t combines the tail of a_{t-1} with a rolling summary of $a_1, \ldots, a_{t-2}$ that tracks goals, relationships, and other narrative states. We reuse the beat summaries and agency labels from interaction-aware boundary refinement to construct a restricted outline o_t for each refined passage. To limit exposure to target-passage details, the outline contains only the summaries of the leading passive beats and the first active beat, or all available summaries if no active beat is present. The query is generated as:

$$q_t = \mathrm{QueryGen}_{\mathrm{LLM}}(\pi, H_t, o_t), \qquad t = 2, \ldots, T, \tag{2}$$

Table 2: Interactive corpus statistics by source-token range. Character counts exclude whitespace.

Token range	Books	Turns	Turns/book	Characters	Chars/turn
$(0, 64\text{K})$	10	1049	105	0.8M	815
$[64\text{K}, 128\text{K})$	10	1725	173	1.4M	816
$[128\text{K}, 256\text{K})$	10	2725	273	2.3M	866
$[256\text{K}, 512\text{K})$	10	5864	586	5.2M	884
$[512\text{K}, 1\text{M})$	10	10718	1072	9.8M	919
Total	50	22081	442	19.7M	890

where $\mathrm{QueryGen}_{\mathrm{LLM}}$ denotes an LLM prompted to express an immediate intention or response in the protagonist's voice without revealing subsequent outcomes. The resulting pre-injection sequence is $\mathcal{D}_{\mathrm{clean}} = (a_1, (q_2, a_2), \ldots, (q_T, a_T))$, with a prologue a_1 and $T - 1$ query–response pairs. Here, D_{clean} denotes the reconstructed interaction history after full-text human review and before controlled error injection. Table 2 summarizes the distribution across source-length buckets.

3.2 Controlled Error Injection

We construct each variant $\mathcal{D}$ separately from $\mathcal{D}_{\mathrm{clean}}$ by modifying one or more responses to contradict a single underlying fact while preserving its supporting historical evidence. The process comprises fact selection, contradiction injection, and human validation.

Fact Extraction & Selection. Inspired by FlawedFictionsMaker (Ahuja et al., 2025), we extract persistent facts from the first 20% of turns and pair each fact with a supporting source quotation. We score each fact's importance from 1 to 4, with higher scores indicating greater importance to the narrative. We retain only facts scored 2 or 3: minor details offer few suitable opportunities for contradiction, whereas contradicting facts central to the plot may require extensive narrative revisions. For each retained fact, we identify later passages that restate or depend on it, excluding candidates where intervening events have changed the relevant fact. For distance analysis, we use the valid supporting occurrence nearest to the injection target, which may differ from the quotation used for initial fact extraction.

Controlled Error Injection. For each candidate, an LLM makes localized edits that replace the fact's value with a mutually exclusive alternative while preserving its subject, attribute, and temporal scope. The same erroneous fact is introduced into one or more responses, while player queries and text outside the edited spans remain unchanged. For preliminary screening, an LLM reviews the supporting historical evidence, edited passages, and surrounding narrative context to classify each edited candidate as *contradictory*, *uncertain*, or *non-contradictory*. Non-contradictory candidates are discarded, while uncertain cases are revised or discarded before human review.

The benchmark comprises 2,000 corrupted variants from 50 novels, with 40 separately constructed variants per novel and 400 per source-length bucket. Each contains one injected contradiction, potentially expressed across multiple turns, with supporting historical evidence. The annotated contradictions include both query–narration and narration–narration conflicts. Human reviewers validated the contradiction and its annotations, which cover only turns explicitly expressing the error and exclude downstream-only consequences (Appendix A.3).

4 DRPV: Dual-Route Proposition Verification

Given an interaction sequence $\mathcal{D}$, DRPV identifies turns whose responses explicitly contradict previously established narrative facts. As illustrated in Figure 2, the process comprises dual-route candidate discovery, proposition consolidation and selection, and complete-prefix verification. The evaluator receives no gold annotations and is not told whether the input contains an error.

Dual-Route Candidate Discovery. We use two complementary routes: a *block route* for conflicts within contiguous narrative blocks and a *retrieval route* for conflicts involving earlier evidence beyond the target block. The block route partitions $\mathcal{D}$ into consecutive, non-overlapping blocks and

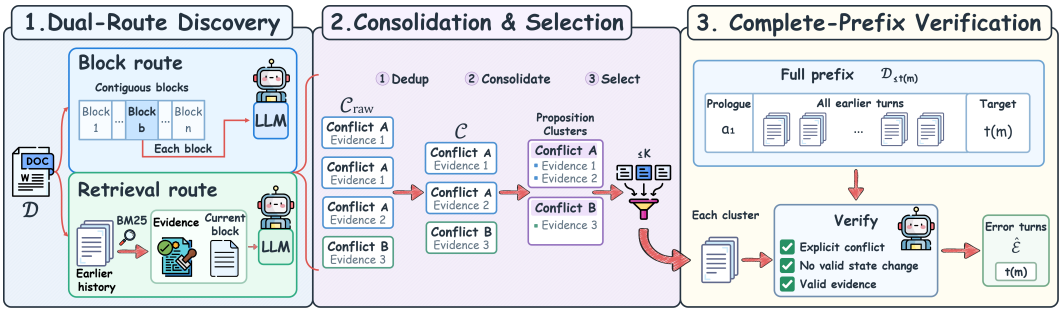

Figure 2: Overview of DRPV. Candidate conflicts from two discovery routes are consolidated into propositions, selected, and verified against complete narrative prefixes to localize erroneous turns.

prompts an LLM to identify conflicts between narrative responses within blocks. The retrieval route augments target blocks with earlier player queries and narrative responses using BM25-style lexical retrieval, then prompts an LLM to identify conflicts between the retrieved evidence and the target narration. Together, both routes produce the raw candidate pool $\mathcal{C}_{\text{raw}}$. Each candidate links a target statement to potentially conflicting earlier evidence and records quotations and turn IDs.

Proposition Consolidation and Selection. Multiple candidate records may describe the same conflict. We first deduplicate $\mathcal{C}_{\text{raw}}$ to obtain $\mathcal{C}$. An LLM then groups semantically equivalent candidates at the same target turn into clusters, each summarized by a *conflict proposition*. Candidates at different target turns remain separate, allowing repeated expressions of the same error to be localized individually. The LLM prioritizes clusters with explicit conflicting evidence and clear entity references, with support from both discovery routes as an additional cue. It selects a set $\mathcal{M}_K$ of up to K clusters for verification, with $K = 5$ by default.

Complete-Prefix Verification. For each selected cluster $m \in \mathcal{M}_K$ with target turn $t(m)$, an LLM verifier evaluates its proposition using the member candidates and the complete interaction prefix $\mathcal{D}_{\leq t(m)}$. The prefix includes the prologue and all query–response pairs through the target turn, excluding later turns. The verifier can therefore examine evidence beyond that identified during discovery, as well as intervening events that may resolve the proposed conflict. The verifier judges a conflict valid only if the target quotation explicitly contradicts an earlier narrative fact about the same entity and attribute under matching temporal conditions, with no narrative-supported state change reconciling the two. We denote the LLM-based verification procedure by $\text{Verify}_{\text{LLM}}$. It returns ACCEPT only when the LLM judges the conflict valid and the supporting candidate and evidence pass programmatic checks. The predicted error-turn set is:

$$\widehat{\mathcal{E}} = \{t(m) \mid m \in \mathcal{M}_K,\ \text{Verify}_{\text{LLM}}(m, \mathcal{D}_{\leq t(m)}) = \text{ACCEPT}\}. \tag{3}$$

A narrative is classified as inconsistent if $\widehat{\mathcal{E}} \neq \varnothing$. Appendix B provides further details on the discovery settings, selection constraints, and verification checks.

5 Experiments

LinconBench evaluates how accurately and completely methods localize the annotated injected errors in long narrative histories. The main evaluation on error-injected variants measures localization performance. Appendix I separately evaluates false alarms on reviewed pre-injection histories. Our experiments address four questions: (1) how accurately and completely do representative evaluators localize annotated errors in long interactive narratives; (2) how does localization vary with evidence distance; (3) how does localization vary with source length; and (4) how do DRPV's components and verification cap affect localization? Qualitative case studies are provided in Appendix J.

5.1 Experimental Setup

Baselines. We compare DRPV with nine baselines grouped into five categories, with table abbreviations given in parentheses: (1) *basic baselines*, including Direct Full Text (Direct) and

Table 3: Overall error-turn localization results. All scores are percentages; higher is better.

Metric	Direct	Sliding	NCP	ArcANE	NFS	WSE-S	WSE-V	Flawed	ConStory	**DRPV**
Turn-P	21.57	17.79	<u>27.99</u>	24.41	22.11	10.44	16.58	**30.23**	18.84	18.48
Turn-R	12.27	13.19	8.38	8.91	8.36	15.80	13.08	7.70	<u>16.19</u>	**28.91**
Turn-F1	15.65	15.15	12.90	13.05	12.13	12.57	14.63	12.28	<u>17.42</u>	**22.55**
CH	8.20	8.75	5.35	5.95	3.95	<u>9.00</u>	7.85	5.45	7.90	**19.90**

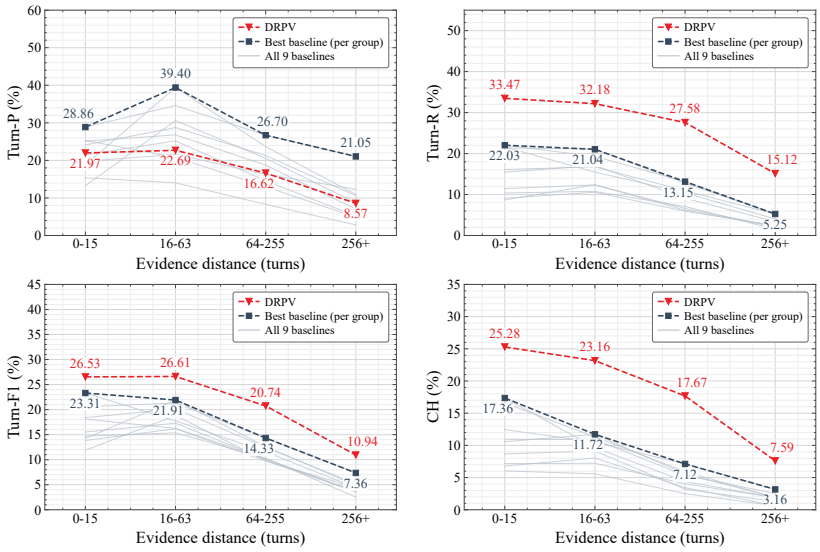

Figure 3: Localization by distance to the evidence. Gray lines show nine baselines.

Sliding Window (Sliding); (2) *fact-ledger-guided auditing*, represented by NCP (Ma et al., 2026); (3) *context selection and retrieval*, including ArcANE (Song et al., 2026) and NarrativeFactScore RAG (NFS) (Jeong et al., 2025); (4) *turn-level auditing*, including WSE Single (WSE-S) and WSE 3Vote (WSE-V) (Chen et al., 2026); and (5) *candidate generation and verification*, including FlawedFictions (Flawed) (Ahuja et al., 2025) and ConStory (Li et al., 2026). We use `deepseek-v4-flash` (DeepSeek-AI, 2026) for multi-turn narrative construction and controlled error injection. Token counts of the preprocessed source novels are computed using the `deepseek-ai/DeepSeek-V4-Flash` tokenizer. DRPV and all baselines use `deepseek-v4-pro` as the judge model. WSE-S uses the first judge's predictions, whereas WSE-V uses turn-level majority voting across three separate judge streams with the same LLM. We follow the reported baseline procedures and evaluate all methods under a shared error-turn localization protocol. Appendix C details each method's inputs, processing, output mapping, and settings.

Evaluation metrics. We report four primary localization metrics computed against the human-verified injected-error annotations. Turn-P measures the fraction of predicted error turns that match the annotations, while Turn-R measures the fraction of annotated error turns recovered. Turn-F1 summarizes their trade-off. These three metrics are micro-averaged by pooling turn-level counts across all evaluated narratives. CH measures the proportion of narratives for which every annotated error turn is recovered, allowing additional predictions. We interpret CH jointly with Turn-P and Turn-F1 because complete gold coverage does not require an exact prediction set. All four metrics use the full evaluation set, including empty predictions and execution failures scored as empty sets, and are reported as percentages. Higher values are better; the best and second-best results are bold and underlined, respectively. Appendix D defines these metrics and seven supplementary measures covering narrative flagging, partial and exact error-set recovery, and first-error localization. Evidence distance is measured in turns between the original injection target and the nearest valid supporting evidence for the contradicted fact.

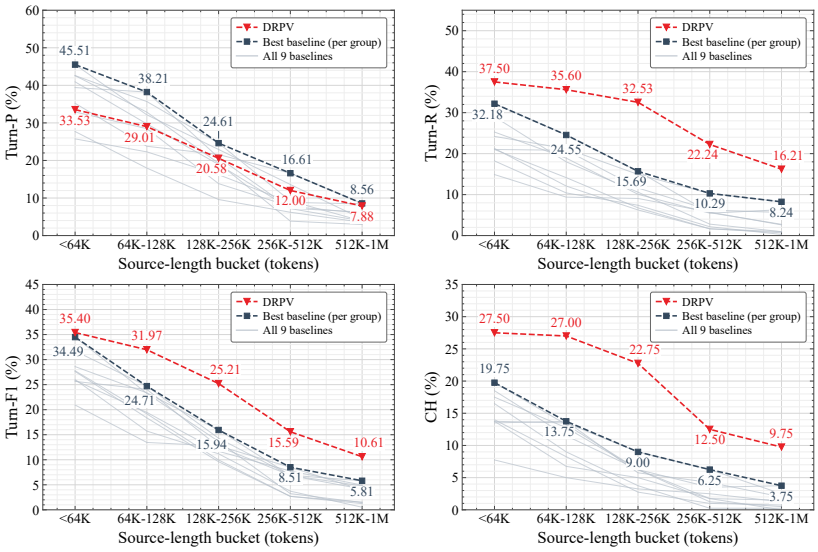

Figure 4: Performance across source-length buckets. Gray lines show nine baselines.

5.2 Main Results

Q1. Overall error-turn localization. Table 3 shows that DRPV achieves the highest Turn-R, Turn-F1, and CH among the evaluated methods. Compared with ConStory, the baseline with the highest Turn-F1, DRPV improves Turn-R by 12.72 percentage points, Turn-F1 by 5.13 points, and CH by 12.00 points, while Turn-P is 0.36 points lower. These results indicate greater recovery of annotated errors at similar prediction precision. The route ablations in Q4 support the contribution of complementary evidence views: using either block-based or retrieval-based discovery alone reduces both CH and Turn-F1. Nevertheless, DRPV's CH of 19.90% indicates that recovering all annotated error turns remains challenging. Appendix E provides supplementary results and further analysis of prediction behavior and error-set recovery. Appendix I complements these results with an analysis of false alarms on reviewed clean histories.

Q2. How does evidence distance relate to localization? Figure 3 shows generally lower localization scores at larger evidence distances, with CH and Turn-F1 lowest in the 256+ turn group for all methods. DRPV achieves the highest Turn-R, Turn-F1, and CH in every pooled distance group. Nevertheless, its CH decreases from 25.28% at 0–15 turns to 7.59% at 256+ turns, while Turn-F1 falls from 26.53% to 10.94%. Between the shortest- and longest-distance groups, Turn-P decreases from 21.97% to 8.57%, and Turn-R from 33.47% to 15.12%, indicating declines in both prediction precision and annotated-error recovery. Greater separation may make supporting facts harder to recover and require examining more intervening events for state changes that could resolve an apparent contradiction. However, the pooled decline is not consistently observed within source-length buckets, and these comparisons do not isolate the independent effect of evidence distance. Results for the length-stratified analyses and all 11 evaluation metrics are provided in Appendix F.

Q3. How does localization scale with narrative length? Figure 4 shows declining localization performance as source length increases. DRPV achieves the highest Turn-F1 in every source-length bucket, but its score decreases from 35.40% below 64K tokens to 10.61% at 512K–1M. Over the same range, Turn-P decreases from 33.53% to 7.88%, Turn-R from 37.50% to 16.21%, and CH from 27.50% to 9.75%. The decline therefore involves both reduced recovery of annotated errors and a smaller fraction of predictions matching the annotations. Longer histories may make it harder to distinguish relevant facts from competing narrative details and to track state updates. Missing supporting evidence can leave annotated errors undetected, while overlooking narrative-supported state changes can make compatible statements appear contradictory. These comparisons describe associations with source length rather than its independent causal effect. Appendix G reports all 11 evaluation measures and examines how narrative flagging differs from successful localization.

Table 4: Ablation results for DRPV. All scores are reported as percentages.

Configuration	Turn-P	Turn-R	Turn-F1	CH
DRPV	**18.48**	28.91	**22.55**	19.90
Block-only	15.65	19.64	17.42	13.00
Retrieval-only	17.87	16.06	16.92	9.15
w/o Proposition Consolidation	17.79	27.63	21.64	18.70
w/o Verification	15.23	**30.27**	20.27	**21.20**

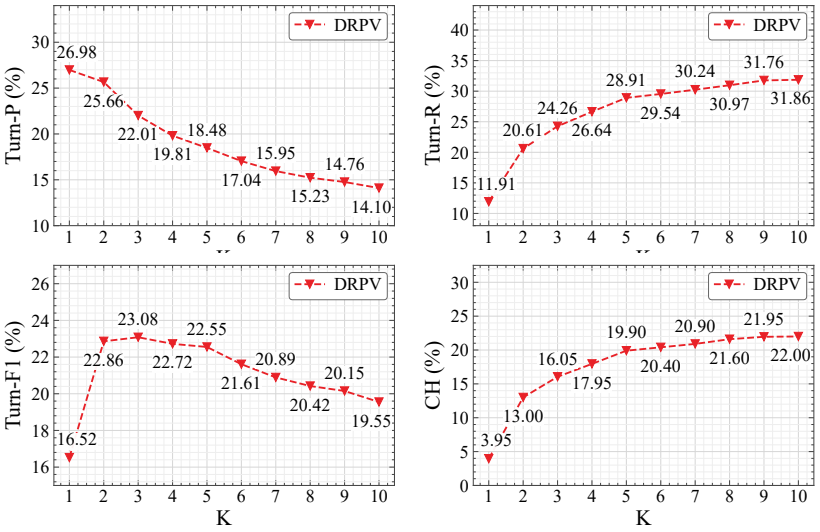

Figure 5: Sensitivity to the verification cap K. All metrics are reported as percentages.

Q4. Component analysis and budget sensitivity. Table 4 reports ablations of DRPV's discovery routes, proposition consolidation, and complete-prefix verification. Combining both discovery routes improves all four metrics over either route alone. Block-only and Retrieval-only reduce Turn-F1 from 22.55% to 17.42% and 16.92%, and CH from 19.90% to 13.00% and 9.15%, respectively. These results are consistent with complementary evidence views from local blocks and retrieved history. Removing proposition consolidation lowers all four metrics. Consolidation may reduce candidate redundancy, allowing limited verification slots to cover more distinct propositions. Holding selected clusters fixed, verification improves Turn-P by 3.25 percentage points and Turn-F1 by 2.28 points, while reducing Turn-R by 1.36 points and CH by 1.30 points. It filters non-gold predictions but also rejects some correctly localized errors. Figure 5 shows sensitivity to the verification cap. We use $K = 5$ by default to balance error coverage, localization quality, and verification workload. Turn-F1 peaks at $K = 3$. Increasing K beyond the default value of 5 yields modest gains in complete error coverage, while reducing Turn-F1 and increasing the number of logical verifications. Appendix H provides additional metrics and verification statistics.

6 Conclusion

We introduced LinConBench, a human-verified benchmark for semantic contradiction detection and error-turn localization in long interactive narratives. Built from human-authored Chinese novels, it provides controlled contradictions with error-turn annotations and historical evidence. We also propose DRPV, which combines dual-route candidate discovery with proposition consolidation and complete-prefix verification. Across nine baselines, DRPV achieves the highest overall Turn-R, Turn-F1, and CH. Its gains primarily reflect better recovery of annotated errors, but with lower precision than several baselines. Recovering all annotated error turns remains challenging. Across the evaluated LLM-based judges, longer narratives and greater evidence distances are generally associated with poorer localization performance. These findings highlight the need for more reliable evidence linking and state tracking over long interaction histories.

AI USE STATEMENT

We used generative AI tools to polish the manuscript for clarity and readability and to help construct LINCONBENCH. Specifically, we used `deepseek-v4-flash` for narrative segmentation and boundary refinement, player-query generation, fact extraction, controlled contradiction injection, and preliminary screening, as described in Section 3 and Appendix A. Narrative responses preserve the human-authored source text except for localized contradiction edits. A team of 20 human reviewers conducted full-text review of the 50 pre-injection interaction histories and validation of the 2,000 injected variants. Each variant was assessed by one reviewer to verify the injected contradiction and error-turn annotations, with uncertain or disputed cases resolved through discussion among reviewers. We used `deepseek-v4-pro` for the main evaluation and `Qwen3.8-Flash` for an additional evaluation (Appendix E.5). This additional evaluation examines dependence on the judge model but does not eliminate potential biases in benchmark construction. The authors take responsibility for the final manuscript and research artifacts, including all AI-assisted content.

REFERENCES

Kabir Ahuja, Melanie Sclar, and Yulia Tsvetkov. Finding flawed fictions: Evaluating complex reasoning in language models via plot hole detection. *CoRR*, 2025.

Alibaba Cloud. Qwen3.8-Flash, 2026.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proc. of ACL*, pp. 3119–3137, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proc. of ACL*, pp. 3639–3664, 2025.

Tommaso Bonomo, Luca Gioffré, and Roberto Navigli. Literaryqa: Towards effective evaluation of long-document narrative QA. In *Proc. of EMNLP*, pp. 34086–34107, 2025.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Proc. of ACL Findings*, pp. 2318–2335, 2024.

Yuqi Chen, Sixuan Li, Yunfeng Cai, Xueai Li, Ka Man Yan, and Ying Li. When stories evolve: Benchmarking LLM storytelling across agent architectures in open-ended world simulations. *CoRR*, 2026.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, 2026.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Proc. of ACL Findings*, pp. 18632–18702, 2025.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. In *Proc. of EMNLP Findings*, pp. 23281–23298, 2025.

Abhay Gupta, Kevin Zhu, Vasu Sharma, Sean O'Brien, and Michael Lu. NovelHopQA: Diagnosing multi-hop reasoning failures in long narrative contexts. In *Proc. of EMNLP*, pp. 26134–26151, 2025.

Shiyu He, Minchi Kuang, Mengxian Wang, Bin Hu, and Tingxiang Gu. EvoSpark: Endogenous interactive agent societies for unified long-horizon narrative evolution. In *Proc. of ACL*, pp. 32088–32107, 2026.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *CoRR*, 2024.

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? *CoRR*, 2024.

Yeonseok Jeong, Minsoo Kim, Seung-won Hwang, and Byung-Hak Kim. Agent-as-judge for factual summarization of long narratives. In *Proc. of EMNLP*, 2025.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Y. Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. In *Proc. of NeurIPS*, 2024.

Junjie Li, Xinrui Guo, Yuhao Wu, Roy Ka-Wei Lee, Hongzhi Li, and Yutao Xie. Lost in stories: Consistency bugs in long story generation by LLMs. In *Proc. of ACL Findings*, pp. 8400–8428, 2026.

Liangtao Lin, Jun Zheng, and Haidong Wang. Webnovelbench: Placing LLM novelists on the web novel distribution. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1828–1847, 2026.

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. *Trans. Assoc. Comput. Linguistics*, 2024.

Yingpeng Ma, Jianhao Yan, Bei Shi, Ka Hou Kam, Runnan Wang, Xuebo Liu, Yulong Chen, Yue Zhang, and Derek F Wong. Can llm agents stick to the script? a benchmark for long-horizon consistency in interactive narratives. *arXiv preprint arXiv:2608.08160*, 2026.

Jean M Mandler and Nancy S Johnson. Remembrance of things parsed: Story structure and recall. *Cognitive psychology*, pp. 111–151, 1977.

Michał Mochocki and Raine Koskimaa. Story beats in videogames as value-driven choice-based unit operations. *Images. The International Journal of European Film, Performing Arts and Audiovisual Communication*, pp. 5–31, 2021.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. Nolima: Long-context evaluation beyond literal matching. In *Proc. of ICML*, 2025.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. BOOKWORLD: From novels to interactive agent societies for story creation. In *Proc. of ACL*, pp. 15898–15912, 2025.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, pp. 333–389, 2009.

Ronsor. Bigknow2022: Bringing language models up to speed. `https://github.com/RyokoAI/BigKnow2022`, 2023.

Woojung Song, Nalim Kim, Sangjun Song, Chaewon Heo, Jongwon Lim, and Yohan Jo. Arcane: Do role-playing language agents stay in character at the right time? *CoRR*, 2026.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Adams, Lydia B. Chilton, and Kathleen R. McKeown. STORYSUMM: evaluating faithfulness in story summarization. In *Proc. of EMNLP*, pp. 9988–10005, 2024.

Melanie Subbiah, Akankshya Mishra, Grace Kim, Liyan Tang, Greg Durrett, and Kathleen McKeown. Is the top still spinning? evaluating subjectivity in narrative understanding. In *Proc. of EMNLP*, pp. 185–203, 2025.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. In *Proc. of ICLR*, pp. 133343–133429, 2026.

Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Proc. of ACL Findings*, pp. 14743–14777, 2024.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *Proc. of ICLR*, 2025a.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing. In *Proc. of NeurIPS*, 2025b.

Dingyi Yang and Qin Jin. What matters in evaluating book-length stories? a systematic study of long story evaluation. In *Proc. of ACL*, pp. 16375–16398, 2025.

Zihao Yi, Zhenqing Ling, Delong Zeng, Haohao Luo, Zhe Xu, Wei Liu, Jian Luan, Wanxia Cao, and Ying Shen. Attention basin: Why contextual position matters in large language models. In *Proc. of ACL*, pp. 26086–26106, 2026.

Pengfei Yu, Dongming Shen, Silin Meng, Jaewon Lee, Weisu Yin, Andrea Yaoyun Cui, Zhenlin Xu, Yi Zhu, Xingjian Shi, Mu Li, and Alex Smola. RPGBENCH: evaluating large language models as role-playing game engines. *CoRR*, 2025.

Chao Zhang, Shunan Guo, Abe Davis, and Eunyee Koh. Narrix: Remixing narrative strategies from examples for story writing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI 2026, Barcelona, Spain, April 13-17, 2026*, pp. 926:1–926:24, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proc. of NeurIPS*, 2023.

A LinconBench Construction Details

A.1 Protagonist Identification and Profiling

An LLM identifies the protagonist from the opening portion of the source text, expanding the inspected span if needed. It considers narrative focus, internal perspective, and plot-driving actions, using only the supplied text. Unresolved cases are handled manually.

The profile π records the protagonist's primary name, aliases, identity, personality, and speaking style. The name guides event attribution and beat analysis, while aliases and style support query generation in the protagonist's voice. The profile is reused throughout narrative construction.

A.2 A Worked Example of Benchmark Construction

A.2.1 Step 1: Protagonist Identification and Profiling

Source Passage S_0

On the first night of term, thirteen-year-old Rowan Bell waited until Alderwick Academy's west dormitory fell quiet, then slipped into the corridor with a brass star-key hidden in his palm. Professor Wren had lent it to him after supper, explaining that it belonged to the astronomy tower and warning him never to let another student carry it. Rowan crossed the moonlit quadrangle alone and reached the tower, where a blue chain sealed the spiral stair.

When he pressed the brass star-key into the moon-shaped lock, blue light ran through every link, and the chain fell loose. Wanting both hands free for the narrow climb, **he left the key in the tower lock.** Before he reached the first landing, a bell sounded once from the sealed observatory, and three silver moths streamed from a portrait toward the upper stair.

Rowan followed the moths to a landing above the tower clock, where they settled around a tall glass cabinet. Inside lay an atlas bound in green leather. A note fixed to its clasp said the book would open only for the student carrying Professor Wren's star-key. Rowan reached into his robe pocket, then remembered that the key remained in the lock below. He glanced toward the stair, intending to retrieve it before attempting to open the cabinet.

Before Rowan could turn back, the cabinet clicked open by itself. A strip of black parchment slid between the atlas pages and drifted toward the stairwell. Rowan caught it before it crossed the rail, unfolded it carefully, and read a warning written in pale ink: the atlas must not leave the cabinet before moonrise. He folded the parchment but did not touch the book. The bell did not sound again, and no footsteps came from the stair beneath him.

Mina Vale reached the landing while Rowan held the folded parchment. She had followed the moths' frost trail after noticing that his bed was empty. Rowan showed her the warning and explained that he had left the key in the lock after opening the stair. Mina asked whether anyone might have taken it while he climbed, but Rowan had heard no one below and had never returned for it. They left the atlas untouched and started back down together to retrieve the key. As they descended, **the brass star-key remained in the tower lock below.**

Input. The continuous source passage S_0.

Intermediate Result: Protagonist Profile π

Field	Extracted result
Primary name	Rowan Bell
Aliases	Rowan
Brief identity	Thirteen-year-old Alderwick Academy student
Personality	Curious, cautious around unfamiliar magic, and responsible for entrusted objects
Speaking / inner-monologue style	Brief first-person intentions and questions grounded in visible information

Output. Source passage S_0 and protagonist profile π.

A.2.2 STEP 2: EVENT-BASED SEGMENTATION

Input. S_0 and π.

Intermediate Result: Initial Event Segments

Segment $\tilde{a}_1$

Setup	Development	Outcome
Wren entrusts Rowan with the key; the tower stair is chained.	Rowan presses the brass star-key into the moon-shaped lock.	The stair opens; Rowan leaves the key in the lock.

Boundary anchor. Ends after "he left the key in the tower lock."

Complete source span. On the first night of term, thirteen-year-old Rowan Bell waited until Alderwick Academy's west dormitory fell quiet, then slipped into the corridor with a brass star-key hidden in his palm. Professor Wren had lent it to him after supper, explaining that it belonged to the astronomy tower and warning him never to let another student carry it. Rowan crossed the moonlit quadrangle alone and reached the tower, where a blue chain sealed the spiral stair.
When he pressed the brass star-key into the moon-shaped lock, blue light ran through every link, and the chain fell loose. Wanting both hands free for the narrow climb, he left the key in the tower lock.

Segment $\tilde{a}_2$

Setup	Development	Outcome
(in $\tilde{a}_1$) Rowan opens the tower stair, leaves the star-key in its lock, and intends to continue upward.	Bell and moths redirect him upward; he follows them to the atlas.	The note makes the key relevant; Rowan plans to retrieve it.

Boundary anchor. Begins with "Before he reached the first landing," immediately before the bell-and-moths trigger, and ends after "attempting to open the cabinet." The bell and moths develop the new episode by redirecting Rowan from the stairs to the upper landing; they also become context for the next query after refinement.

Complete source span. Before he reached the first landing, a bell sounded once from the sealed observatory, and three silver moths streamed from a portrait toward the upper stair.
Rowan followed the moths to a landing above the tower clock, where they settled around a tall glass cabinet. Inside lay an atlas bound in green leather. A note fixed to its clasp said the book would open only for the student carrying Professor Wren's star-key. Rowan reached into his robe pocket, then remembered that the key remained in the lock below. He glanced toward the stair, intending to retrieve it before attempting to open the cabinet.

Segment $\tilde{a}_3$

Setup	Development	Outcome
(in $\tilde{a}_2$) Rowan has found the atlas and intends to go back for its key.	The cabinet opens and parchment escapes, interrupting that plan; Rowan catches and reads it.	Rowan and Mina leave the atlas and descend for the key.

Boundary anchor. Begins with "Before Rowan could turn back," immediately before the cabinet's opening, and ends at the end of S_0. The opening changes Rowan's intended action and releases the warning; it is development, not merely scene setup.

Complete source span. Before Rowan could turn back, the cabinet clicked open by itself. A strip of black parchment slid between the atlas pages and drifted toward the stairwell. Rowan caught it before it crossed the rail, unfolded it carefully, and read a warning written in pale ink: the atlas must not leave the cabinet before moonrise. He folded the parchment but did not touch the book. The bell did not sound again, and no footsteps came from the stair beneath him.
Mina Vale reached the landing while Rowan held the folded parchment. She had followed the moths' frost trail after noticing that his bed was empty. Rowan showed her the warning and explained that he had left the key in the lock after opening the stair. Mina asked whether anyone might have taken it while he climbed, but Rowan had heard no one below and had never returned for it. They left the atlas untouched and started back down together to retrieve the key. As they descended, the brass star-key remained in the tower lock below.

Output. Initial passages $(\tilde{a}_1, \tilde{a}_2, \tilde{a}_3)$.

A.2.3 Step 3: Interaction-Aware Boundary Refinement

Input. The three initial passages and π.

Intermediate Result: Boundary-Relevant Narrative Beats

Beat	Seg.	Source anchor	Outcome-free summary	Agency
B1.1	1	"left the key in the tower lock"	Rowan intends to continue upward with both hands free.	Active
B2.1	2	"a bell sounded once"	A sound from the observatory draws attention upward.	Passive
B2.2	2	"three silver moths streamed"	A visible magical cue moves toward the upper stair.	Passive
B2.3	2	"Rowan followed the moths"	Rowan intends to follow the magical cue.	Active
B2.4	2	"He glanced toward the stair, intending to retrieve it"	Rowan intends to retrieve the key needed for the atlas.	Active
B3.1	3	"Before Rowan could turn back ... drifted toward the stairwell"	The cabinet and parchment redirect Rowan without his initiating them.	Passive
B3.2	3	"Rowan caught it"	Rowan intends to secure the drifting parchment.	Active
B3.3	3	"the atlas must not leave"	The parchment presents a restriction concerning the atlas.	Passive

Candidate prefixes: B2.1–B2.2 and B3.1; other non-boundary beats are omitted for brevity.

Output. Ordered beats, agency labels, and the two candidate passive prefixes.

Input. Initial boundaries, ordered beats, agency labels, and preceding-passage tails.

Intermediate Result: Boundary Refinement

Boundary	Before	Moved context	Stopping beat	After
$\tilde{a}_1 \mid \tilde{a}_2$	key left | bell, moths, Rowan follows	**complete bell-and-moths sentence** (B2.1–B2.2)	B2.3: Rowan follows	moths stream | Rowan follows
$\tilde{a}_2 \mid \tilde{a}_3$	plans to retrieve key | cabinet opens, parchment drifts, Rowan catches	**complete opening-and-parchment sentences** (B3.1)	B3.2: Rowan catches	parchment drifts | Rowan catches

No earlier passive beat in either moved prefix merely continues an established situation.

Refined Narrative Responses

Response	Exact source span	Boundary-local result
a_1	"On the first night" → "toward the upper stair"	Ends with the cues that motivate the next query.
a_2	"Rowan followed the moths" → "drifted toward the stairwell"	Begins at the first active beat and receives the next passive prefix.
a_3	"Rowan caught it" → "tower lock below"	Begins at the first active beat after the second boundary adjustment and preserves the remaining source narration.

Only passage membership changes; concatenating a_1, a_2, a_3 reconstructs S_0 in order, apart from separator whitespace.

Output. Refined passages (a_1, a_2, a_3) as fixed narrator responses.

A.2.4 Step 4: Reverse User Query Generation

Input. π, (a_1, a_2, a_3), H_t, and o_t.

Intermediate Result: Query Generation

Turn	Historical context H_t	Restricted outline o_t	Generated query and fixed response
T1	Opening context	–	Prologue a_1; no generated query.
T2	Tail of a_1: bell and moths; no earlier rolling summary.	First active beat B2.3; no leading passive beat remains.	q_2: *The bell and moths are drawing me upward. I follow them. What do I find?* Fixed a_2: "Rowan followed" → "drifted toward the stairwell."
T3	Tail of a_2: cabinet opens, and parchment drifts; rolling summary from a_1 tracks Wren's key left in the tower lock.	First active beat B3.2; no leading passive beat remains.	q_3: *The parchment is drifting away. I catch it. What does it say, and what happens next?* Fixed a_3: "Rowan caught it" → "tower lock below."

$\mathcal{D}_{\text{clean}} = (a_1, (q_2, a_2), (q_3, a_3))$.

Output. Clean interaction sequence $\mathcal{D}_{\text{clean}}$.

A.2.5 Step 5: Fact Extraction

Result: Candidate Fact Set $\mathcal{F}$

Fact	Entity / attribute	Value	Supporting evidence
F1	Rowan / age	Thirteen	T1: "thirteen-year-old Rowan Bell"
F2	Star-key / material	Brass	T1: "a brass star-key"
F3	Star-key / lender	Professor Wren	T1: "Professor Wren had lent it"
F4	Wren / instruction to Rowan	Do not let another student carry the key	T1: "warning him never to let another student carry it"
F5	Star-key / demonstrated function	Releases the tower chain	T1: "pressed the brass star-key into the moon-shaped lock, blue light ran through every link, and the chain fell loose"
F6	Star-key / location after a_1	Tower lock	T1: "left the key in the tower lock"
F7	Observatory / state	Sealed	T1: "the sealed observatory"

Production uses the first 20% of turns.

A.2.6 Step 6: Fact Selection

Result: Scores and Target Fact

Fact	Score (1–4)	Eligibility
F1	1	Not retained
F2	1	Not retained
F3	2	Retained
F4	2	Retained
F5	2	Retained
F6	**3**	**Retained; target in this example**
F7	1	Not retained

Retained: F3–F6. **Displayed target:** F6 (star-key location) in a_3. **State change:** None.

A.2.7 STEP 7: CONTROLLED ERROR INJECTION

Historical Evidence, Localized Edit, and Annotation

Item	Content
Historical evidence	T1 / a_1: "he left the key in the tower lock"
Before injection	T3 / a_3: "As they descended, the brass star-key ~~remained in the tower lock below~~."
After injection	T3 / a_3: "As they descended, the brass star-key **rested in Rowan's robe pocket** ."
Unchanged material	Queries q_2, q_3, and all text outside the edit
Contradicted fact	F6: star-key location; no intervening movement is narrated

A.2.8 STEP 8: HUMAN VALIDATION

Human Validation of the Worked Example

Contradiction: The edited passage places the star-key in Rowan's robe pocket, conflicting with its established location in the tower lock, with no intervening movement narrated.
First erroneous turn: T3.
Explicit error-turn set: {T3}.
Downstream-only turns: None.
Assessment: Valid contradiction; the edit remains fluent.

A.3 HUMAN REVIEW

A team of 20 human reviewers participated in two stages of benchmark validation: reviewing the complete pre-injection interaction histories for pre-existing semantic contradictions, and validating the injected contradictions and error-turn annotations. Uncertain or disputed cases in either stage were resolved through discussion among reviewers.

Pre-Injection Review. Reviewers examined all 50 complete interaction histories, covering both generated player queries and narrative responses. The review followed the semantic-contradiction criterion used throughout this paper: conflicting statements must concern established narrative facts under compatible temporal conditions, without a narrative-supported explanation reconciling them. Reviewers distinguished character beliefs, speculation, player intentions, and hypothetical events from facts established by the narration. Only histories with no identified semantic contradictions remaining unresolved were used for error injection. Each reviewed history served as the common base for 40 separately constructed variants.

Validation of Injected Contradictions. Following preliminary screening, each of the 2,000 variants was assessed by one reviewer. The reviewer examined the edited passages together with the supporting historical evidence and relevant narrative context, verifying that the edits contradicted a single underlying fact and that the conflict could not be explained by a valid state change or a difference in wording. The reviewer also assessed textual fluency and checked the annotations for the first erroneous turn and all turns explicitly expressing the injected erroneous fact. Turns describing only downstream consequences were excluded from the error annotations.

B DRPV IMPLEMENTATION DETAILS

B.1 DUAL-ROUTE CANDIDATE DISCOVERY

The block route scans consecutive, non-overlapping blocks of at most 128 turns. The retrieval route uses target blocks of at most 64 turns and 64,000 characters, with a shared 48,000-character history budget. Player queries and agent narrations are split into role-labeled passages of at most 1,000 characters. Queries combine visible target-turn text with sentence-aware chunks of up to 300 characters and approximately 50 characters of overlap. Packing budgets include source markers and separators; passage and chunk limits count source-text characters.

Retrieval is restricted to turns preceding the target block. Binary-term BM25-style scoring (Robertson & Zaragoza, 2009) uses Chinese bigrams and alphanumeric terms, with $k_1 = 1.2$, $b = 0.75$, and document statistics computed over that historical prefix. Low-frequency query terms and chronologically sampled postings produce candidate passages. Selection balances early, middle, and recent history across target turns under the budget, then presents the selected passages in source order.

Each candidate records a target narration turn and quotation, an evidence turn and source role, an evidence quotation, a conflict claim, and route provenance. The target quotation must match agent narration exactly, while the evidence quotation must match the referenced player query or agent narration exactly; retrieval candidates must additionally match the supplied target and historical passages, respectively. The retrieval route returns at most 12 candidates per block. Invalid entries are discarded. Discovery may expose later turns within a block, whereas final verification uses only the prefix through the individual target turn.

B.2 Proposition Consolidation and Selection

Validated candidates are deduplicated by target/evidence IDs and quotation pairs, after removing whitespace and applying Unicode case folding; duplicate records combine their route provenance. The complete resulting pool $\mathcal{C}$ is supplied to one LLM request, without truncation or batching. Across the benchmark, the largest candidate pool contains 57 records, and the largest consolidation input contains 6,989 tokens. The selector groups equivalent conflicts sharing the same target turn, entity, attribute, and temporal conditions, while keeping different target turns separate. It prioritizes explicit conflicting evidence and clear entity references; cross-route support is an additional ranking cue. Let $d(\mathcal{C})$ count distinct candidate target turns. The selected set satisfies:

$$|\mathcal{M}_K| \leq \min\{K, d(\mathcal{C})\}, \qquad K = 5. \tag{4}$$

This bounds the total cluster count without requiring different targets for different clusters: distinct propositions at one turn can consume separate slots. Each cluster must contain valid, unused candidate IDs with a common target. Invalid entries are discarded. Empty pools or selection with no valid clusters bypass verification and yield empty predictions.

B.3 Complete-Prefix Verification

Each selected cluster is verified independently using its proposition, member candidates, and the complete prefix $\mathcal{D}_{\leq t(m)}$, without later turns or silent truncation. The verifier applies the contradiction criterion in Section 4, checking for narrative-supported state changes and compatible interpretations. Acceptance requires a supporting cluster member and an exact evidence quotation from a player query or agent narration. Evidence may come from the current or an earlier turn, provided that it precedes the target conflicting statement in source order. This evidence may differ from the discovery evidence. Rejection requires null supporting-member and evidence fields.

An unresolved required verification is recorded as an execution failure, and no partial prediction is returned. The four such failures in the default run remain in the evaluation set and are scored with empty predicted sets, while retaining their failure status.

For completed narratives, accepted target IDs are deduplicated according to Eq. 3. Rejecting every selected cluster yields a completed empty prediction. The output contains at most K distinct error turns; supporting evidence and downstream consequences are not automatically included, and predictions are never expanded into continuous intervals.

C Baseline Implementation Details

C.1 Shared Evaluation Setting

We evaluate nine baseline implementations on LinconBench for error-turn localization. The following sections describe each implementation's inputs, processing steps, and prediction constraints. All judges use `Deepseek-v4-pro`, with temperature set to 0 and a maximum output budget of 65,536 tokens per judging request. These settings apply to individual requests; the number of requests and the narrative context provided depend on the method.

Table 5: Input configurations and output rules for the baseline implementations. "No fixed cap" means that no additional numerical limit is imposed on the number of reported error turns; context and output-token limits still apply.

Method	Input	Output
Direct	Full history	No fixed cap
Sliding	512-turn windows, stride 100	Union of window predictions
NCP	Current turn and incrementally updated fact ledger	Turn IDs of confirmed semantic contradictions
ArcANE	Character-arc context for the target narrative phase	Turn IDs of findings satisfying the shared contradiction criterion
NFS	Up to 32 evidence packets; 360,000-character evidence budget	Predictions restricted to visible turn IDs
WSE-S	Prefix through each 20-turn batch	All turns marked inconsistent by judge 1
WSE-V	Same batches, three separate judge streams	At least two inconsistent votes per turn
Flawed	Full history	Error turns in the first accepted proposal
ConStory	Full history	Up to 8 candidate items per subtype before verification

Evaluators receive narrative text with the original turn identifiers. Gold error locations, annotation labels, injection patches, and evidence-distance labels are excluded from judging prompts. Under the shared evaluation criterion, judges are instructed to distinguish established narrative facts from player intentions, dialogue, speculation, and hypothetical events, and to allow state changes supported by the narrative. Purely stylistic defects and spelling variants that preserve entity identity are not considered contradictions. The prompts for Direct, Sliding, NCP, ArcANE, NFS, Flawed, and ConStory describe the possibility of a consistent narrative or a single root contradiction expressed in one or more turns. The WSE prompt instead assigns a consistency label to each target turn without specifying the number of root contradictions. No evaluator is told whether the evaluated instance contains an injected error or where that error occurs.

Every method is mapped to a set of distinct predicted error-turn IDs. A nonempty set gives a positive narrative-level prediction, and its minimum is the predicted first error. Earlier supporting evidence and downstream consequences are not automatically included, and no prediction is expanded into a continuous turn interval. We retain each method's output constraints rather than imposing a common top-k rule. Table 5 summarizes these constraints. Cases unresolved after bounded validation retries are scored as failures and remain in the evaluation set.

C.2 Direct Auditing and Fact-Ledger Guidance

Direct Full Text (Direct). Direct submits the complete materialized narrative to a single judging request and asks for contradictions and their exact error-turn IDs. It uses neither a separate retrieval stage nor an external verifier. The prompt requires auditing the complete visible history and reporting only turns that explicitly express the conflicting fact.

Sliding Window (Sliding). Sliding partitions the narrative into contiguous windows of 512 turns with a stride of 100 turns. A narrative shorter than one window is evaluated once; otherwise, an additional end-anchored window is included when necessary to cover the final turn. Each window is judged independently using only its visible history. The final prediction is the union of the window-level error-turn sets, with duplicate turn IDs removed. Aggregation is deterministic and introduces no additional judging or verification call.

NCP. We follow the sequential fact-ledger auditing procedure of NCP (Ma et al., 2026). Narrative turns are processed in chronological order. At each turn, the evaluator checks the current narration against the active fact ledger and examines candidate state updates extracted from the response. Potential contradictions undergo a secondary confirmation step. When no conflict is confirmed, newly established facts are added to the ledger and superseded facts are marked as negated, preserving the history of state changes while maintaining the active narrative state. For our shared localization

protocol, confirmed semantic contradictions are mapped to the narration turns that explicitly express the conflicting statements.

C.3 Context Selection and Retrieval

ArcANE. We follow the Character Arc representation of ArcANE (Song et al., 2026) to organize characters' development across the narrative. Key events and evolving character states are represented as ordered phases along psychological or relational axes. Each phase records the relevant narrative span, the character's state, and the events supporting that state. Evaluation uses the arc context corresponding to the target narrative phase, with later phases excluded. Under our shared semantic-contradiction criterion, findings count as errors only when they establish incompatible narrative facts; changes in character behavior that are supported by the narrative are permitted.

NarrativeFactScore RAG (NFS). We follow the fact-extraction and evidence-retrieval procedure of NarrativeFactScore (Jeong et al., 2025), using earlier narrative scenes as evidence for evaluating the consistency of later statements. The narrative is divided into 12-turn scenes with a 2-turn overlap. An extraction call produces at most 3 atomic facts per turn, 36 facts per scene, and 16 alias groups per scene. Alias-normalized triples observed in at least two distinct evidence occurrences form a recurring-relation graph; duplicate occurrences from overlapping scenes are removed, and at most 96 graph edges are retained.

We concatenate each scene's extracted facts, including their subjects, predicates, objects, and qualifiers, and encode the resulting text with `BAAI/bge-m3` (Chen et al., 2024). We use $1{,}024$-dimensional dense embeddings, FP16 inference, ℓ_2 normalization, and a maximum encoder input length of $8{,}192$ tokens. For each scene with eligible earlier context, the retriever selects the highest-ranked nonoverlapping earlier scene; scenes without such context produce standalone packets. The ranking score combines embedding cosine similarity, entity-set Jaccard similarity, relation-set Jaccard similarity, and an attribute-conflict heuristic with weights 0.45, 0.20, 0.15, and 0.20, respectively. For fact pairs sharing a subject but having different objects, the heuristic assigns 1 to an exact predicate match and 0.55 to predicates sharing a character; it uses the maximum pair score, or 0 if no pair qualifies. Selection reserves one candidate packet per temporal phase, using up to 12 phases, and then fills the remaining slots by score, up to 32 packets.

Each packet contains extracted facts and up to 4 original evidence turns per scene. The judge receives the selected packets and the recurring-relation graph under a $360{,}000$-character budget. If a packet would exceed the budget, its raw-text excerpts are omitted first; a packet that still exceeds the budget is omitted entirely. To reuse extraction across variants of the same narrative, unchanged scene representations are cached, while every scene overlapping an edited turn is re-extracted from the materialized variant. Patch locations are used only to invalidate the affected cache entries. Retrieval scores and graph construction are recomputed from the resulting variant-specific representations, without using patch locations to rank or select evidence.

C.4 Turn-Level Auditing

WSE Single (WSE-S) and WSE 3Vote (WSE-V). Following the turn-level consistency procedure of When Stories Evolve (Chen et al., 2026), each judge assigns every target turn one of three labels: `consistent`, `inconsistent`, or `uncertain`. Targets are processed in batches of 20 turns. Each request includes the narrative prefix through the last turn in its batch, so earlier targets can access later turns within the same batch. This offline batched audit retains a separate decision for each target and requires supporting evidence to precede that target.

WSE-V uses three separate judge streams with the same LLM; the judges cannot inspect one another's outputs. A turn enters the predicted error set only if at least two judges label it inconsistent. Uncertain decisions are retained as uncertainty and do not count as detected contradictions. WSE-S uses the first judge's cached decisions for every narrative, with the judge index fixed across the evaluation. It does not select a judge separately for each sample. Excluding retries, a narrative with T evaluated turns requires $\lceil T/20 \rceil$ batch judgments per stream.

Positive judgments include quotations supporting the earlier fact and the later conflicting claim. These quotations are checked against the referenced turns for textual grounding. The label definitions and majority-voting rule are the same across all source-length buckets.

C.5 Candidate Generation and Verification

FlawedFictions (Flawed). We implement the generation-and-verification procedure of FlawedFictions (Ahuja et al., 2025), with separate generation and verification calls conditioned on the full narrative. The generator proposes an erroneous statement together with earlier contradictory evidence and the associated turn IDs. The verifier checks a positive proposal. If it is rejected, the rejection feedback is supplied to the next generator call, with at most 3 generator proposals in total. The process stops at the first accepted proposal or when the generator returns no contradiction. If all proposals are rejected, the method returns an empty error-turn set. The final output consists of the error-turn IDs in the accepted proposal, rather than the union of rejected and accepted proposals. This permits at most 3 generator and 3 verifier calls, excluding validation retries.

ConStory. We use the released ConStory category prompts (Li et al., 2026) to identify candidate contradictions and follow its candidate-generation and verification procedure. Findings are evaluated under the shared semantic-contradiction criterion. Five full-history scans cover characterization, factual detail, narrative style, timeline and plot, and world building, comprising 19 subtypes. Pure style or quality defects are excluded unless the quoted passages also establish incompatible narrative facts. Each subtype may return at most 8 candidate items. An item must identify an earlier supporting quotation and a later conflicting quotation from agent narration, together with their turn IDs.

Programmatic validation locates the quotations in the narrative and repairs recoverable location or ordering errors before semantic verification. Each nonempty category is then submitted to one additional full-history verification request that assesses all of its valid candidate pairs. The final prediction is the union of accepted later-turn IDs across categories. The pipeline therefore uses five category scans and up to five category-level verification calls, excluding retries. Its output is governed by the subtype-level candidate limit and verification decisions, without an additional global top-k turn filter.

D Evaluation Metrics

This appendix defines 11 evaluation measures and explains their aggregation rules. The four primary metrics are Turn-P, Turn-R, Turn-F1, and Complete Hit (CH). Seven supplementary measures characterize prediction behavior and additional aspects of localization: Flag Rate, Any-turn Hit (Hit), Exact Set (ES), First Exact (FE), first-error accuracy within one and three turns (FE$\pm$1 and FE$\pm$3), and conditional mean absolute error (MAE). Flag Rate describes the frequency of valid, nonempty predictions and is not assigned a preferred direction.

D.1 Evaluation Scope and Notation

LinconBench contains 2,000 error-injected narrative variants derived from 50 source novels across five source-length buckets, ranging from below 64K to 512K–1M tokens. Each variant contains a single injected contradiction expressed in one or more turns, with annotations validated through the human review procedure described in Appendix A.3. Let N denote the number of benchmark narratives, indexed by $i = 1, \ldots, N$. Overall evaluation uses all 2,000 narratives; each source-length bucket contains 400. The same definitions apply within each nonempty subgroup, and groups containing no benchmark instances are left unscored.

For instance i, let E_i denote the nonempty set of annotated error-turn IDs and $\widehat{E}_i$ the set of predicted error-turn IDs, with duplicate IDs counted once. The gold set includes only turns that explicitly state the injected erroneous fact, excluding earlier supporting evidence and turns that merely describe downstream consequences. We evaluate localization by exact turn-ID matching: any predicted turn outside E_i is counted as a false positive.

D.2 NARRATIVE FLAGGING

Flag Rate measures the proportion of evaluated narratives for which a valid final output contains a nonempty predicted error-turn set. Let I_{pred} be the index set of these narratives and $N_{\text{pred}} = |I_{\text{pred}}|$. With N denoting the total number of evaluated narratives,

$$\text{Flag Rate} = \frac{N_{\text{pred}}}{N}. \tag{5}$$

Valid empty predictions and execution failures contribute zero to the numerator and remain in the denominator. On this all-positive benchmark, Flag Rate equals narrative-level detection recall, but does not require any predicted turn to match an annotated error turn. We report it as a behavioral diagnostic, expressed as a percentage and without a preferred direction. Flag Rate on the error-injected variants does not estimate specificity or false-alarm rates on consistent narratives; Appendix I separately reports false-alarm rates on reviewed pre-injection histories.

D.3 TURN-LEVEL PREDICTION QUALITY

We assess turn-level localization using Turn-P, Turn-R, and Turn-F1. Turn-P measures the fraction of predicted error turns that match the annotations, while Turn-R measures the fraction of annotated error turns recovered. Turn-F1 is their harmonic mean.

We count correctly identified annotated error turns as true positives (TP), predicted turns outside the gold set as false positives (FP), and missed annotated error turns as false negatives (FN). These counts are relative to the injected-error annotations. True negatives are not used. Pooling the counts across narratives gives:

$$TP_{\text{turn}} = \sum_i |E_i \cap \hat{E}_i|, \quad FP_{\text{turn}} = \sum_i |\hat{E}_i \setminus E_i|, \quad FN_{\text{turn}} = \sum_i |E_i \setminus \hat{E}_i|. \tag{6}$$

We compute micro-averaged Turn-P, Turn-R, and Turn-F1 from these pooled counts:

$$\text{Turn-P} = \frac{TP_{\text{turn}}}{TP_{\text{turn}} + FP_{\text{turn}}}, \qquad \text{Turn-R} = \frac{TP_{\text{turn}}}{TP_{\text{turn}} + FN_{\text{turn}}}. \tag{7}$$

$$\text{Turn-F1} = \frac{2TP_{\text{turn}}}{2TP_{\text{turn}} + FP_{\text{turn}} + FN_{\text{turn}}}. \tag{8}$$

All evaluated narratives are included. An empty prediction ($\hat{E}_i = \varnothing$), including an execution failure scored as an empty set, contributes $|E_i|$ to FN_{turn} and zero to TP_{turn} and FP_{turn}.

D.4 SAMPLE-LEVEL LOCALIZATION

We assess sample-level localization using three increasingly stringent criteria: Any-turn Hit (Hit), Complete Hit (CH), and Exact Set (ES). Hit requires recovery of at least one gold error turn. CH extends this requirement to all gold error turns, while allowing additional predictions. ES further excludes non-gold predictions, requiring exact recovery of the gold set. Each score is the proportion of narratives satisfying its criterion:

$$\text{Hit} = \frac{1}{N} \sum_i \mathbf{1}[E_i \cap \widehat{E}_i \neq \varnothing], \tag{9}$$

$$\text{CH} = \frac{1}{N} \sum_i \mathbf{1}[E_i \subseteq \widehat{E}_i], \qquad \text{ES} = \frac{1}{N} \sum_i \mathbf{1}[E_i = \widehat{E}_i]. \tag{10}$$

Empty predictions fail all three criteria because each gold set is nonempty. The scores satisfy $\text{ES} \leq \text{CH} \leq \text{Hit}$, and Hit and CH coincide for instances with a single gold error turn. For example, $E_i = \{10, 20\}$ and $\widehat{E}_i = \{10, 20, 30\}$ satisfy Hit and CH but not ES. More generally, $\text{CH} - \text{ES}$ is the fraction of narratives with complete gold coverage and at least one additional non-gold prediction.

Table 6: Overall localization results and prediction diagnostics. Scores are percentages except MAE. MAE is computed over narratives with valid, nonempty predictions; Flag Rate reports the proportion of such narratives and has no preferred direction. For the remaining measures, arrows indicate the preferred direction; the best and second-best numerical results are bold and underlined, respectively.

	Prediction behavior	Turn-level			Sample-level			First-error			
Method	Flag Rate	Turn-P↑	Turn-R↑	Turn-F1↑	Hit↑	CH↑	ES↑	FE↑	FE±1 ↑	FE±3 ↑	MAE↓
Direct	54.70	21.57	12.27	15.65	15.05	8.20	<u>7.85</u>	11.60	12.30	13.40	105.82
Sliding	70.30	17.79	13.19	15.15	16.85	8.75	7.60	12.60	13.45	15.15	93.15
NCP	31.90	<u>27.99</u>	8.38	12.90	10.25	5.35	5.05	7.60	8.20	8.85	86.64
ArcANE	44.10	24.41	8.91	13.05	11.25	5.95	5.75	8.60	9.15	10.25	89.28
NFS	51.90	22.11	8.36	12.13	12.85	3.95	3.75	8.50	9.25	10.30	87.53
WSE-S	83.95	10.44	15.80	12.57	23.35	<u>9.00</u>	2.00	8.05	8.95	10.55	105.81
WSE-V	65.65	16.58	13.08	14.63	19.25	7.85	2.65	9.70	10.40	11.50	99.52
Flawed	34.40	**30.23**	7.70	12.28	9.55	5.45	5.20	7.60	8.35	9.20	**75.13**
ConStory	75.65	18.84	<u>16.19</u>	<u>17.42</u>	<u>27.05</u>	7.90	3.50	<u>12.90</u>	<u>13.95</u>	<u>15.20</u>	88.18
DRPV	93.80	18.48	**28.91**	**22.55**	**36.95**	**19.90**	**9.95**	**19.35**	**20.15**	**21.90**	<u>83.18</u>

D.5 First-Error Localization

We evaluate how accurately an evaluator identifies the first erroneous turn. First Exact (FE) requires an exact match, the ± 1 and ± 3 scores allow deviations of up to one and three turns, and mean absolute error (MAE) measures the average absolute deviation among narratives with a predicted location. Let $f_i = \min E_i$ be the first annotated error turn and $\widehat{f}_i = \min \widehat{E}_i$ the earliest predicted error turn for a nonempty prediction. Using the prediction subset I_{pred} defined in Section D.2, accuracy with a tolerance of k turns is:

$$\text{Acc}_{\pm k} = \frac{1}{N} \sum_{i \in \mathcal{I}_{\text{pred}}} \mathbf{1}[|\widehat{f}_i - f_i| \leq k], \qquad k \in \{0, 1, 3\}.$$

Here, $\text{FE} = \text{Acc}_{\pm 0}$, while FE$\pm$1 and FE$\pm$3 denote $\text{Acc}_{\pm 1}$ and $\text{Acc}_{\pm 3}$, respectively. All three accuracy scores use N as the denominator, so both valid empty predictions and execution failures contribute zero to the numerators. MAE is computed only over the N_{pred} narratives with a predicted location:

$$\text{MAE} = \frac{1}{N_{\text{pred}}} \sum_{i \in \mathcal{I}_{\text{pred}}} |\widehat{f}_i - f_i|.$$

MAE is measured in turns and is undefined when $N_{\text{pred}} = 0$. It uses the same valid, nonempty prediction subset counted in the numerator of Flag Rate. Thus, $N_{\text{pred}} = N \times \text{Flag Rate}$, with Flag Rate expressed as a fraction. Empty predictions and execution failures are excluded from MAE. Because this subset can differ across methods and settings, MAE should be interpreted jointly with Flag Rate and first-error accuracy. Comparisons use the earliest predicted turn, even if a later prediction matches f_i exactly.

E Precision–Recall Trade-offs and Error-Set Recovery

Table 6 reports all 11 evaluation measures defined in Appendix D. We use these results to characterize the precision–recall trade-offs of the evaluated methods, examine the effect of majority voting in WSE, and distinguish partial error recovery, complete coverage, and exact error-set recovery.

E.1 Precision–Recall Trade-offs across Evaluators

The evaluated methods exhibit different precision–recall trade-offs. FlawedFictions achieves the highest turn-level precision at 30.23%, but its recall is only 7.70%, yielding a Turn-F1 of 12.28%. DRPV achieves the highest recall at 28.91% and the highest Turn-F1 at 22.55%, with a precision of 18.48%. Compared with ConStory, DRPV increases recall by 12.72 percentage points while reducing precision by 0.36 points, yielding a Turn-F1 gain of 5.13 points. These differences also

reflect distinct aggregate prediction volumes. Under micro-aggregation, the ratio of total predicted error turns to total annotated error turns equals R/P. This ratio is approximately 1.56 for DRPV and 0.86 for ConStory. On the same evaluation set, DRPV therefore produces approximately 1.82 times as many predicted error turns, while recovering approximately 1.79 times as many annotated error turns. Its greater coverage thus accompanies a larger prediction volume, with correct matches increasing less proportionally than total predictions. Nevertheless, DRPV still misses approximately 71.09% of annotated error turns, and only 18.48% of its predictions match the annotations. Both incomplete recovery and non-gold predictions therefore remain substantial limitations.

E.2 Precision–Recall Trade-offs under Majority Voting

Under the WSE voting protocol described in Appendix C.4, majority voting increases precision from 10.44% for WSE-S to 16.58% for WSE-V, while reducing recall from 15.80% to 13.08% and Any-turn Hit from 23.35% to 19.25%. Turn-F1 nevertheless increases from 12.57% to 14.63%, indicating improved precision at the expense of error coverage. One possible explanation is that requiring agreement from at least two judge calls excludes some non-gold predictions that are not consistently reproduced. The same requirement may also exclude correct localizations supported by only one call. In long narratives, calls may retrieve different evidence, interpret intervening state changes differently, or identify different turns expressing the same underlying contradiction. Agreement on the existence of a conflict therefore need not produce agreement on its exact location, potentially reducing both turn-level recall and narrative-level localization coverage. Meanwhile, calls using the same backbone may share systematic errors, allowing some non-gold predictions to survive voting. These mechanisms could explain the observed trade-off, but the aggregate results do not distinguish their contributions.

E.3 Complete Coverage versus Exact Error-Set Recovery

DRPV achieves the highest Any-turn Hit, Complete Hit, and Exact Set scores at 36.95%, 19.90%, and 9.95%, respectively. The difference between Hit and CH indicates that 17.05% of narratives have at least one annotated error turn recovered but incomplete gold coverage. Another 9.95% have complete gold coverage accompanied by additional non-gold predictions, while 9.95% have their gold error sets recovered exactly. Thus, half of the narratives with complete gold coverage still contain additional non-gold predictions. DRPV exceeds Direct and Sliding in ES by 2.10 and 2.35 percentage points, respectively. Nevertheless, incomplete coverage and additional predictions remain substantial barriers to exact error-set recovery. Throughout this analysis, non-gold predictions denote predicted turns outside the human-verified error-turn sets.

Table 7: Distribution of 2,000 benchmark instances by source length and evidence distance.

Source length	0–15	16–63	64–255	256+	Total
$<$ 64K	83	255	62	0	400
64K–128K	76	193	131	0	400
128K–256K	38	120	226	16	400
256K–512K	33	77	237	53	400
512K–1M	35	89	187	89	400
Total	265	734	843	158	2000

E.4 Localization by Contradiction Type

We further compare DRPV on 470 Query–Narration cases and 1,530 Narration–Narration cases. Metrics are recomputed within each group following Appendix D. Figure 6 shows that Narration–Narration cases yield higher Turn-P, Turn-R, and Turn-F1 than Query–Narration cases. In particular, Turn-F1 is 24.88% versus 12.14%, whereas CH is lower at 18.89% versus 23.19%. This contrast shows that better aggregate turn-level localization does not necessarily translate into more frequent recovery of all annotated error turns.

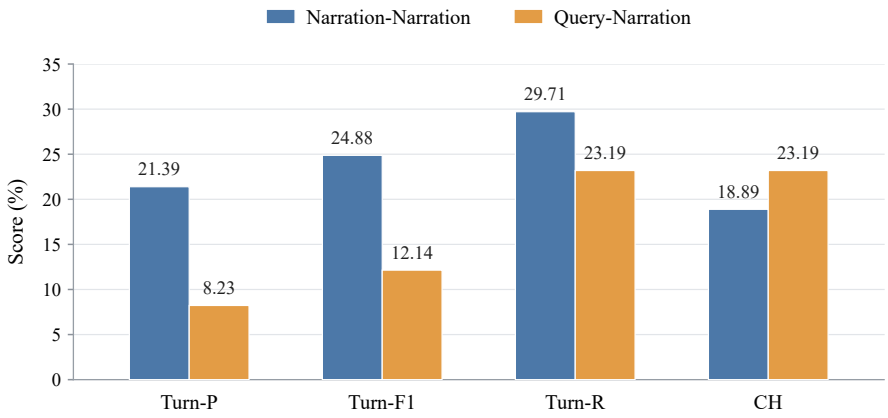

Figure 6: DRPV on Narration–Narration and Query–Narration cases. All scores are percentages; higher is better.

Table 8: Results using Qwen3.8-Flash as the judge model. Scores are percentages except MAE, which is measured in turns and computed only over narratives with valid, nonempty predictions. Flag Rate has no preferred direction. For the remaining metrics, the best and second-best results are bold and underlined, respectively.

	Prediction behavior	Turn-level			Sample-level			First-error			
Method	Flag Rate	Turn-P↑	Turn-R↑	Turn-F1↑	Hit↑	CH↑	ES↑	FE↑	FE±1 ↑	FE±3 ↑	MAE↓
Direct	55.20	22.01	12.59	16.02	15.60	8.60	<u>8.10</u>	12.00	12.70	13.85	103.60
Sliding	70.95	18.33	13.45	15.51	17.30	9.00	7.80	13.00	13.85	15.50	91.80
NCP	32.65	<u>28.46</u>	8.70	13.32	10.65	5.75	5.40	8.00	8.65	9.25	85.90
ArcANE	44.85	24.89	9.04	13.26	11.75	6.10	5.90	9.00	9.55	10.75	88.40
NFS	52.40	22.49	8.49	12.32	13.30	4.20	3.95	8.80	9.60	10.65	86.70
WSE-S	84.30	10.70	16.19	12.88	23.80	<u>9.40</u>	2.15	8.40	9.30	10.95	104.50
WSE-V	66.10	16.90	13.19	14.82	19.70	8.15	2.80	10.05	10.80	11.95	98.30
Flawed	35.00	**30.85**	8.10	12.83	10.00	5.70	5.50	7.90	8.70	9.60	**74.60**
ConStory	76.20	19.18	<u>16.45</u>	<u>17.71</u>	<u>27.60</u>	8.30	3.70	<u>13.25</u>	<u>14.30</u>	<u>15.60</u>	87.40
DRPV	94.10	18.90	**29.22**	**22.96**	**37.50**	**20.45**	**10.40**	**19.80**	**20.60**	**22.45**	<u>82.10</u>

E.5 Evaluation with an Alternative Judge Model

Benchmark construction and the main evaluation use models from the DeepSeek family, raising concerns about shared model-family biases. To examine whether our findings depend on this choice, we additionally evaluate DRPV and all nine baselines using `Qwen3.8-Flash` (Alibaba Cloud, 2026) as the judge model.

Table 8 shows that DRPV retains the highest Turn-R, Turn-F1, and CH. Its Turn-F1 exceeds ConStory by 5.25 percentage points, with comparable precision (18.90% vs. 19.18%). Its CH exceeds the strongest baseline, WSE-S, by 11.05 points. DRPV also leads in ES and all three first-error accuracy measures. Nevertheless, exact error-set recovery remains limited at 10.40%, and DRPV does not achieve the highest precision or lowest conditional MAE. These results support the stability of DRPV's main advantages under this alternative judge, without ruling out benchmark-construction biases.

F Detailed Results by Evidence Distance

We analyze localization by evidence distance across all 2,000 error-bearing narratives and within five source-length buckets, using the nearest valid supporting evidence as defined in Section 5.1.

Table 7 reports the sample distribution, and Table 9 provides all 11 evaluation measures for the pooled distance groups. Figures 7–11 show Turn-P, Turn-R, Turn-F1, and CH within each source-length bucket; Tables 10–14 report the seven supplementary measures: Flag Rate, Hit, ES, FE, FE$\pm$1, FE$\pm$3, and MAE. Metrics are recomputed within each group following Appendix D. Gray curves show all nine baselines, while the best-baseline curve connects their pointwise maxima; the winning method may change between groups.

F.1 Pooled Results

Table 9 reports the pooled results across four evidence-distance groups. DRPV achieves the highest Turn-R, Turn-F1, and CH in every pooled distance group, but does not achieve the highest Turn-P. Its CH decreases from 25.28% at 0–15 turns to 7.59% at 256+ turns, while FE decreases from 25.66% to 9.49%. Turn-F1 is similar in the first two groups (26.53% and 26.61%), before falling to 20.74% at 64–255 turns and 10.94% at 256+ turns. Flag Rate varies between 90.87% and 97.47%, reaching its highest value in the longest-distance group. Among the supplementary localization metrics, DRPV achieves the highest Hit, FE, FE$\pm$1, and FE$\pm$3 in every pooled distance group. DRPV also achieves the highest ES in the first three distance groups and ties with Direct, Flawed, and ConStory at 1.90% in the 256+ group. However, it does not achieve the highest Turn-P in any group. At 256+ turns, its CH, Turn-F1, and FE exceed the respective best-baseline scores of 3.16%, 7.36%, and 4.43%. Its conditional MAE in this group is 164.69 turns, with a Flag Rate of 97.47%. This MAE is computed over DRPV's valid, nonempty predictions; other methods can be evaluated on different subsets.

F.2 Length-Stratified Results

The main-text distance analysis pools narratives across source lengths. To examine whether the aggregate trend holds among narratives of similar length, we compare distance groups within each source-length bucket (Figures 7–11). The pooled decline is not consistently observed within these buckets. Below 64K tokens, DRPV's Turn-F1 rises from 29.79% at 0–15 turns to 44.02% at 64–255 turns. Within 256K–512K, Turn-F1 falls from 19.67% to 7.69% between the shortest and longest distance groups, although Flag Rate is 100% in both groups: returning a nonempty prediction does not ensure that an annotated error turn is identified. These results show that the pooled decline is not uniform across source-length buckets. The distance groups differ in their source-length composition: neither of the two shortest length buckets contains 256+ cases. Differences in narrative length may therefore contribute to the aggregate trend. Stratification allows comparisons within narrower length ranges, but does not hold length constant or establish an independent effect of evidence distance.

DRPV strictly exceeds every baseline in CH in 16 of the 18 populated length–distance groups and ties for the highest CH in the remaining two. It also strictly exceeds every baseline in Turn-R in all 18 groups and in Turn-F1 in 15 groups. A possible contributor is its combination of local and retrieval-based candidate discovery: retrieval can expose earlier supporting facts, while prefix verification checks whether intervening events resolve the proposed conflict. This interpretation is consistent with the design, but the grouped results do not establish which component produces the gains. The advantage is not universal: WSE-V leads Turn-F1 in the < 64K/0–15 group, while Flawed leads in the 128K–256K/256+ group. The latter contains only 16 cases, so its ranking warrants caution.

Table 9: Complete results by evidence distance, pooled across all source lengths. Scores are percentages except MAE (turns). Bold and underlined values indicate the best and second-best scores, respectively, excluding Flag Rate.

Metric	Distance	Direct	Sliding	NCP	ArcANE	NFS	WSE-S	WSE-V	Flawed	ConStory	DRPV
Flag Rate	0–15	60.38	70.19	36.98	50.57	53.21	81.89	67.55	40.00	81.51	94.72
	16–63	62.26	72.21	37.60	51.63	54.50	81.06	63.22	39.10	82.70	90.87
	64–255	49.82	67.97	27.16	38.91	52.31	86.36	67.62	31.32	71.65	95.37
	256+	36.08	74.05	22.15	25.95	35.44	87.97	63.29	19.62	54.43	97.47
Turn-P↑	0–15	13.22	21.47	18.81	24.11	25.13	15.36	25.31	**28.86**	19.77	21.97
	16–63	30.61	25.19	**39.40**	28.74	26.80	14.03	20.80	34.60	21.60	22.69
	64–255	20.55	14.56	23.67	21.29	18.72	8.30	13.01	**26.70**	16.29	16.62
	256+	8.73	5.18	10.94	10.61	7.04	2.81	4.85	**21.05**	12.32	8.57
Turn-R↑	0–15	15.47	16.10	8.69	11.44	10.38	22.03	21.61	9.11	21.40	**33.47**
	16–63	17.06	16.78	12.36	12.36	10.77	19.31	15.47	10.56	21.04	**32.18**
	64–255	9.09	10.91	6.18	6.61	7.09	13.15	10.42	5.94	12.79	**27.58**
	256+	3.40	5.25	2.16	2.16	1.54	5.25	4.01	2.47	5.25	**15.12**
Turn-F1↑	0–15	14.26	18.40	11.88	15.52	14.69	18.10	23.31	13.85	20.55	**26.53**
	16–63	21.91	20.14	18.82	17.29	15.37	16.25	17.74	16.18	21.32	**26.61**
	64–255	12.61	12.47	9.80	10.08	10.29	10.18	11.57	9.72	14.33	**20.74**
	256+	4.89	5.21	3.61	3.59	2.53	3.66	4.39	4.42	7.36	**10.94**
Hit↑	0–15	19.62	20.00	10.94	13.58	15.85	26.04	27.55	10.94	34.72	**38.11**
	16–63	19.48	20.84	14.71	14.71	15.67	26.57	20.98	12.53	33.92	**39.92**
	64–255	11.39	14.00	7.47	9.02	11.39	22.30	17.44	7.71	22.06	**36.30**
	256+	6.33	8.23	3.16	3.16	2.53	9.49	6.96	3.16	8.86	**24.68**
CH↑	0–15	10.57	12.45	6.79	8.68	6.04	16.60	17.36	7.17	10.94	**25.28**
	16–63	11.72	10.49	8.04	9.13	5.59	11.31	9.67	7.22	10.90	**23.16**
	64–255	5.58	7.12	3.44	3.20	2.49	5.93	4.39	4.03	5.34	**17.67**
	256+	1.90	3.16	0.63	1.27	0.63	1.90	1.90	1.90	2.53	**7.59**
ES↑	0–15	9.81	11.32	6.42	8.30	5.66	3.77	6.04	6.79	3.02	**13.58**
	16–63	11.31	10.08	7.49	9.13	5.31	3.00	3.68	6.95	5.18	**11.58**
	64–255	5.34	5.69	3.32	2.85	2.37	0.95	1.19	3.80	2.49	**8.90**
	256+	**1.90**	0.00	0.63	1.27	0.63	0.00	0.00	**1.90**	**1.90**	**1.90**
FE↑	0–15	14.34	15.09	7.17	11.70	10.94	12.45	16.98	8.68	17.74	**25.66**
	16–63	15.40	16.35	11.58	11.04	10.90	9.81	10.35	9.81	16.62	**22.21**
	64–255	8.78	10.08	5.46	6.76	6.88	6.17	8.07	6.29	9.73	**16.73**
	256+	4.43	4.43	1.27	1.90	1.90	2.53	3.16	2.53	4.43	**9.49**
FE±1↑	0–15	17.36	17.36	9.06	12.83	13.21	14.72	18.49	10.19	20.00	**26.79**
	16–63	15.80	17.17	11.72	11.58	11.72	10.49	11.31	10.22	17.98	**23.30**
	64–255	9.02	10.68	6.05	7.24	7.24	7.00	8.42	7.24	10.32	**17.32**
	256+	5.06	4.43	1.90	1.90	1.90	2.53	3.16	2.53	4.43	**9.49**
FE±3↑	0–15	19.62	20.38	9.81	13.58	14.34	18.11	21.13	10.94	23.40	**30.19**
	16–63	17.17	19.21	12.94	12.94	12.94	12.67	12.81	11.58	19.48	**26.02**
	64–255	9.73	11.86	6.29	8.42	8.30	7.83	8.90	7.83	10.91	**18.03**
	256+	5.06	5.06	1.90	1.90	1.90	2.53	3.16	2.53	4.43	**9.49**
MAE↓	0–15	57.44	64.67	67.13	59.30	65.62	82.23	69.30	59.92	**56.47**	61.19
	16–63	69.99	62.20	50.94	62.11	63.39	71.46	70.24	**49.68**	61.71	61.07
	64–255	143.87	113.40	116.17	114.24	104.49	117.96	111.30	**86.20**	107.22	92.78
	256+	248.51	179.39	229.51	238.66	181.55	225.96	222.34	268.48	221.02	**164.69**

Table 10: Supplementary metrics by evidence distance within the < 64K source-length bucket.

Metric	Distance	Direct	Sliding	NCP	ArcANE	NFS	WSE-S	WSE-V	Flawed	ConStory	DRPV
Flag Rate	0–15	83.13	79.52	53.01	62.65	60.24	68.67	57.83	57.83	97.59	90.36
	16–63	78.04	75.29	51.37	63.92	60.78	70.98	58.43	47.06	91.76	86.67
	64–255	69.35	74.19	53.23	59.68	51.61	66.13	59.68	53.23	90.32	88.71
	256+	–	–	–	–	–	–	–	–	–	–
Hit↑	0–15	33.73	25.30	18.07	18.07	18.07	32.53	32.53	20.48	**44.58**	37.35
	16–63	34.12	30.59	25.88	26.67	22.35	32.16	26.67	18.43	**50.20**	43.14
	64–255	29.03	30.65	27.42	24.19	20.97	35.48	29.03	29.03	53.23	**56.45**
	256+	–	–	–	–	–	–	–	–	–	–
ES↑	0–15	15.66	**18.07**	8.43	12.05	6.02	8.43	8.43	14.46	6.02	13.25
	16–63	**20.00**	16.86	14.90	18.04	8.63	5.49	5.49	12.55	8.63	14.90
	64–255	16.13	19.35	12.90	14.52	1.61	0.00	1.61	16.13	8.06	**22.58**
	256+	–	–	–	–	–	–	–	–	–	–
FE↑	0–15	19.28	20.48	8.43	14.46	10.84	20.48	21.69	15.66	25.30	**26.51**
	16–63	28.24	23.92	20.78	22.35	16.47	12.16	14.12	14.90	24.71	**28.63**
	64–255	22.58	25.81	22.58	20.97	6.45	11.29	19.35	22.58	24.19	**32.26**
	256+	–	–	–	–	–	–	–	–	–	–
FE±1 ↑	0–15	26.51	20.48	10.84	16.87	16.87	25.30	25.30	18.07	**30.12**	28.92
	16–63	28.63	24.71	21.18	23.53	17.65	13.33	15.69	15.29	26.27	**29.41**
	64–255	22.58	25.81	22.58	24.19	8.06	12.90	19.35	24.19	27.42	**33.87**
	256+	–	–	–	–	–	–	–	–	–	–
FE±3 ↑	0–15	30.12	25.30	13.25	18.07	19.28	32.53	30.12	20.48	**39.76**	33.73
	16–63	31.76	27.84	23.53	25.88	20.39	16.47	17.25	17.25	29.02	**32.94**
	64–255	25.81	27.42	24.19	29.03	8.06	12.90	19.35	24.19	27.42	**33.87**
	256+	–	–	–	–	–	–	–	–	–	–
MAE↓	0–15	24.29	26.95	26.75	26.65	20.38	**15.09**	16.33	21.62	17.69	21.21
	16–63	19.35	19.81	**17.92**	18.89	21.66	19.23	18.60	20.15	18.11	19.32
	64–255	24.23	22.35	22.27	**16.51**	30.97	32.44	20.81	17.85	27.41	20.64
	256+	–	–	–	–	–	–	–	–	–	–

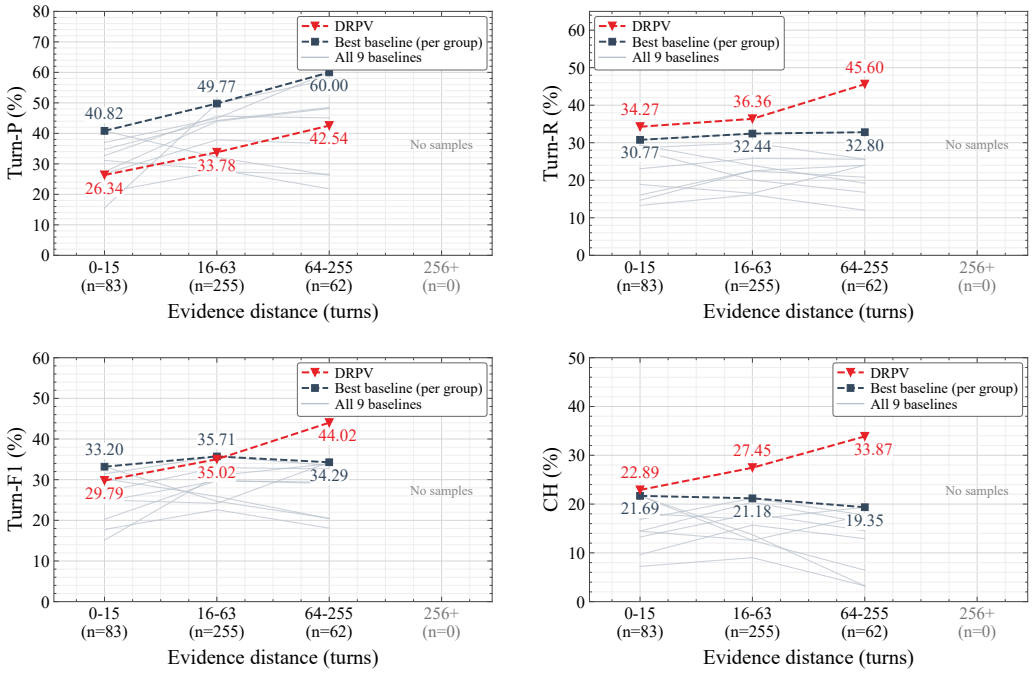

Figure 7: Performance across evidence-distance groups within the < 64K source-length bucket.

Table 11: Supplementary metrics by evidence distance within the 64K–128K source-length bucket.

Metric	Distance	Direct	Sliding	NCP	ArcANE	NFS	WSE-S	WSE-V	Flawed	ConStory	DRPV
Flag Rate	0–15	63.16	73.68	39.47	60.53	52.63	80.26	81.58	44.74	89.47	96.05
	16–63	65.28	74.09	44.04	52.33	55.96	85.49	67.36	43.01	90.16	89.12
	64–255	61.07	65.65	38.17	54.20	56.49	83.21	70.23	49.62	88.55	91.60
	256+	–	–	–	–	–	–	–	–	–	–
Hit↑	0–15	22.37	30.26	14.47	25.00	19.74	32.89	42.11	13.16	**48.68**	46.05
	16–63	20.21	22.80	16.06	12.95	12.95	30.57	26.94	14.51	39.38	**41.45**
	64–255	24.43	22.14	14.50	21.37	16.03	27.48	25.19	13.74	41.98	**50.38**
	256+	–	–	–	–	–	–	–	–	–	–
ES↑	0–15	13.16	17.11	11.84	15.79	7.89	3.95	7.89	7.89	3.95	**19.74**
	16–63	**11.92**	10.88	6.74	7.77	4.15	2.59	4.66	6.22	7.77	10.36
	64–255	13.74	11.45	6.87	6.11	4.58	1.53	2.29	5.34	3.82	**15.27**
	256+	–	–	–	–	–	–	–	–	–	–
FE↑	0–15	19.74	25.00	13.16	22.37	14.47	17.11	25.00	11.84	25.00	**36.84**
	16–63	15.54	18.65	11.92	8.81	7.77	13.99	13.47	10.88	22.80	**23.83**
	64–255	18.32	17.56	9.92	15.27	10.69	8.40	12.21	10.69	19.85	**25.19**
	256+	–	–	–	–	–	–	–	–	–	–
FE±1 ↑	0–15	21.05	30.26	14.47	23.68	15.79	18.42	26.32	14.47	26.32	**36.84**
	16–63	15.54	19.69	11.92	9.33	7.77	15.03	14.51	11.40	24.35	**26.42**
	64–255	19.08	18.32	11.45	16.79	11.45	10.69	13.74	15.27	22.14	**25.95**
	256+	–	–	–	–	–	–	–	–	–	–
FE±3 ↑	0–15	23.68	32.89	14.47	23.68	17.11	19.74	30.26	14.47	27.63	**42.11**
	16–63	16.06	22.28	13.47	10.36	8.29	18.65	17.62	13.99	24.87	**30.57**
	64–255	19.85	19.85	11.45	17.56	14.50	12.98	16.03	16.79	23.66	**27.48**
	256+	–	–	–	–	–	–	–	–	–	–
MAE↓	0–15	25.94	27.73	31.17	31.24	37.77	31.80	**25.08**	29.18	27.28	26.81
	16–63	43.94	36.71	39.56	43.40	46.15	30.95	30.91	34.31	**30.47**	30.71
	64–255	34.86	36.52	45.34	34.52	36.05	43.23	34.02	**28.28**	39.55	34.55
	256+	–	–	–	–	–	–	–	–	–	–

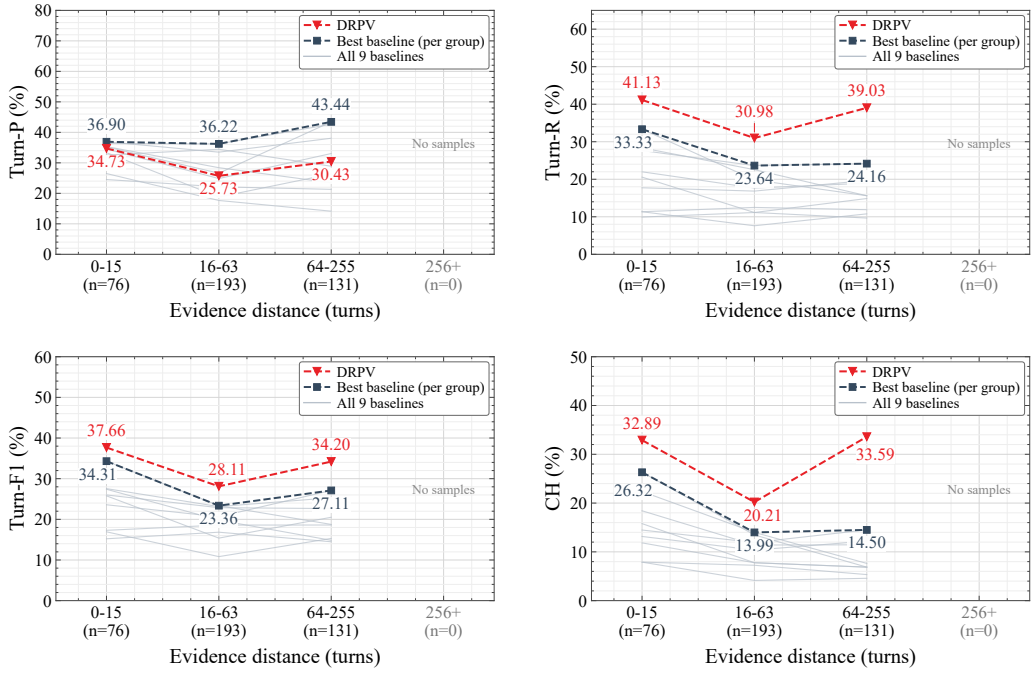

Figure 8: Performance across evidence-distance groups within the 64K–128K source-length bucket.

Table 12: Supplementary metrics by evidence distance within the 128K–256K source-length bucket.

Metric	Distance	Direct	Sliding	NCP	ArcANE	NFS	WSE-S	WSE-V	Flawed	ConStory	DRPV
Flag Rate	0–15	50.00	50.00	31.58	47.37	52.63	92.11	71.05	31.58	78.95	92.11
	16–63	64.17	60.00	32.50	47.50	52.50	80.00	65.00	51.67	86.67	93.33
	64–255	61.06	58.85	31.86	44.69	60.62	84.07	65.93	42.48	85.40	95.13
	256+	87.50	68.75	50.00	56.25	68.75	87.50	62.50	50.00	100.00	100.00
Hit↑	0–15	7.89	5.26	5.26	5.26	13.16	18.42	21.05	5.26	23.68	**34.21**
	16–63	10.83	17.50	8.33	9.17	15.83	25.00	16.67	12.50	29.17	**45.83**
	64–255	14.60	15.04	7.96	9.73	14.60	23.45	22.57	8.85	28.32	**43.81**
	256+	25.00	12.50	12.50	18.75	6.25	12.50	0.00	18.75	37.50	**43.75**
ES↑	0–15	**5.26**	2.63	0.00	0.00	**5.26**	0.00	**5.26**	0.00	0.00	**5.26**
	16–63	6.67	6.67	2.50	5.00	2.50	2.50	3.33	4.17	0.83	**14.17**
	64–255	6.19	7.08	2.65	2.65	3.10	1.77	1.33	5.75	2.65	**9.73**
	256+	**6.25**	0.00	**6.25**	**6.25**	0.00	0.00	0.00	**6.25**	**6.25**	**6.25**
FE↑	0–15	7.89	2.63	2.63	5.26	7.89	7.89	13.16	2.63	5.26	**21.05**
	16–63	8.33	13.33	6.67	5.83	10.83	9.17	8.33	9.17	11.67	**24.17**
	64–255	12.39	12.83	5.31	7.08	8.85	8.41	10.62	8.41	10.62	**22.12**
	256+	12.50	12.50	12.50	12.50	6.25	6.25	0.00	12.50	12.50	**18.75**
FE±1 ↑	0–15	10.53	5.26	5.26	5.26	7.89	7.89	13.16	2.63	7.89	**23.68**
	16–63	8.33	14.17	6.67	5.83	12.50	9.17	9.17	10.00	12.50	**24.17**
	64–255	12.39	13.72	6.19	7.08	8.85	8.41	10.62	8.85	10.62	**23.45**
	256+	12.50	12.50	12.50	12.50	6.25	6.25	0.00	12.50	12.50	**18.75**
FE±3 ↑	0–15	13.16	5.26	5.26	5.26	7.89	13.16	13.16	2.63	7.89	**26.32**
	16–63	8.33	14.17	6.67	6.67	13.33	9.17	9.17	10.00	13.33	**25.83**
	64–255	13.27	15.04	6.64	9.29	10.62	9.29	11.06	10.18	11.06	**24.78**
	256+	12.50	12.50	12.50	12.50	6.25	6.25	0.00	12.50	12.50	**18.75**
MAE↓	0–15	74.21	63.53	68.75	81.28	53.55	72.89	54.89	94.25	57.57	**48.06**
	16–63	85.19	64.74	88.56	81.63	**51.90**	73.31	75.09	74.08	71.78	60.21
	64–255	78.24	63.68	80.94	83.51	62.94	82.86	74.93	73.80	69.23	**61.37**
	256+	142.29	131.73	**110.88**	121.44	142.73	146.64	193.20	124.25	127.69	131.69

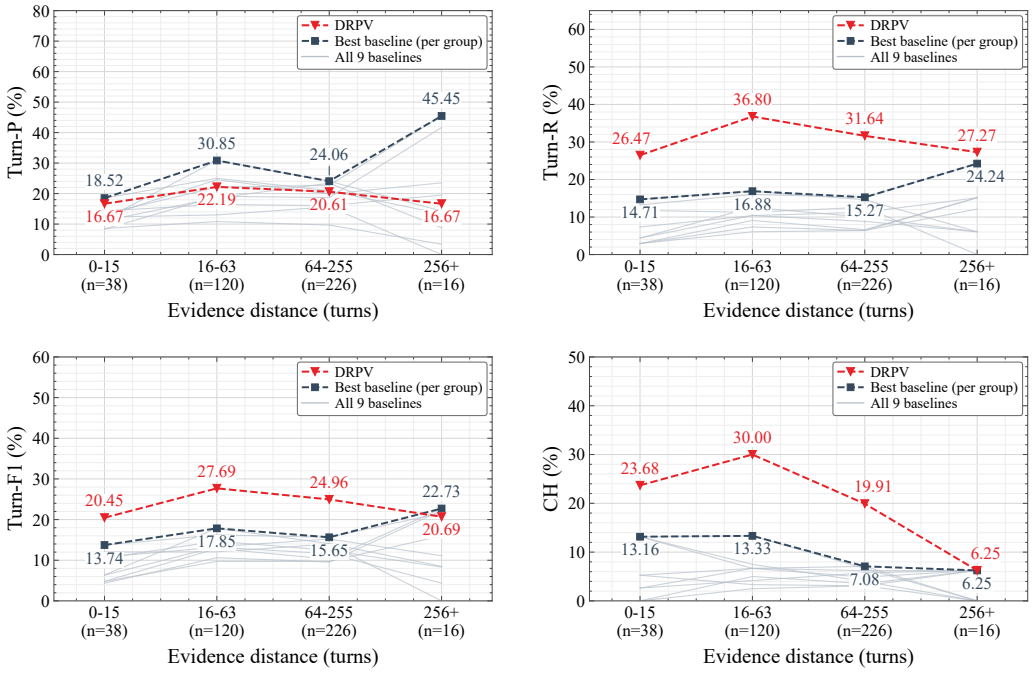

Figure 9: Performance across evidence-distance groups within 128K–256K source-length bucket.

Table 13: Supplementary metrics by evidence distance within the 256K–512K source-length bucket.

Metric	Distance	Direct	Sliding	NCP	ArcANE	NFS	WSE-S	WSE-V	Flawed	ConStory	DRPV
Flag Rate	0–15	36.36	48.48	21.21	27.27	48.48	90.91	57.58	21.21	69.70	100.00
	16–63	42.86	57.14	20.78	40.26	50.65	89.61	59.74	18.18	77.92	98.70
	64–255	43.88	61.60	22.78	30.80	51.48	89.45	65.82	24.47	69.62	98.73
	256+	43.40	62.26	32.08	18.87	37.74	86.79	60.38	22.64	64.15	100.00
Hit↑	0–15	9.09	6.06	0.00	0.00	12.12	12.12	9.09	0.00	21.21	**33.33**
	16–63	5.19	6.49	1.30	3.90	10.39	22.08	12.99	2.60	9.09	**40.26**
	64–255	4.22	8.44	2.95	3.38	8.86	19.83	10.55	2.95	10.55	**29.96**
	256+	5.66	5.66	3.77	0.00	5.66	1.89	3.77	3.77	7.55	**18.87**
ES↑	0–15	3.03	0.00	0.00	0.00	3.03	0.00	3.03	0.00	0.00	**15.15**
	16–63	1.30	2.60	1.30	0.00	5.19	0.00	0.00	2.60	0.00	**9.09**
	64–255	0.84	2.11	1.27	0.42	1.69	0.84	1.27	0.84	0.42	**5.06**
	256+	1.89	0.00	0.00	0.00	1.89	0.00	0.00	**3.77**	**3.77**	0.00
FE↑	0–15	9.09	0.00	0.00	0.00	12.12	0.00	6.06	0.00	12.12	**21.21**
	16–63	1.30	5.19	1.30	0.00	7.79	2.60	5.19	2.60	0.00	**11.69**
	64–255	2.53	5.06	2.11	2.53	6.75	4.64	4.64	2.11	4.22	**8.86**
	256+	**3.77**	1.89	0.00	0.00	**3.77**	1.89	1.89	**3.77**	**3.77**	1.89
FE±1 ↑	0–15	9.09	0.00	3.03	0.00	12.12	3.03	6.06	0.00	12.12	**21.21**
	16–63	3.90	5.19	1.30	0.00	7.79	2.60	5.19	2.60	2.60	**11.69**
	64–255	2.95	5.06	2.53	2.53	7.17	5.91	5.06	2.11	4.22	**8.86**
	256+	**5.66**	1.89	1.89	0.00	3.77	1.89	1.89	3.77	3.77	1.89
FE±3 ↑	0–15	9.09	3.03	3.03	3.03	12.12	3.03	6.06	0.00	12.12	**21.21**
	16–63	5.19	6.49	1.30	1.30	7.79	2.60	5.19	2.60	5.19	**11.69**
	64–255	3.38	5.49	2.53	2.95	7.59	6.75	5.06	2.11	4.64	**9.28**
	256+	**5.66**	3.77	1.89	0.00	3.77	1.89	1.89	3.77	3.77	1.89
MAE↓	0–15	146.50	131.75	210.86	**102.44**	110.12	151.43	181.42	159.86	144.22	109.27
	16–63	138.88	151.89	132.25	142.35	133.10	140.93	153.76	121.79	146.65	**112.33**
	64–255	199.30	143.82	203.54	128.75	**114.10**	133.83	138.70	174.86	151.58	117.37
	256+	213.57	171.00	167.53	225.60	**163.60**	205.83	191.78	200.17	191.12	180.57

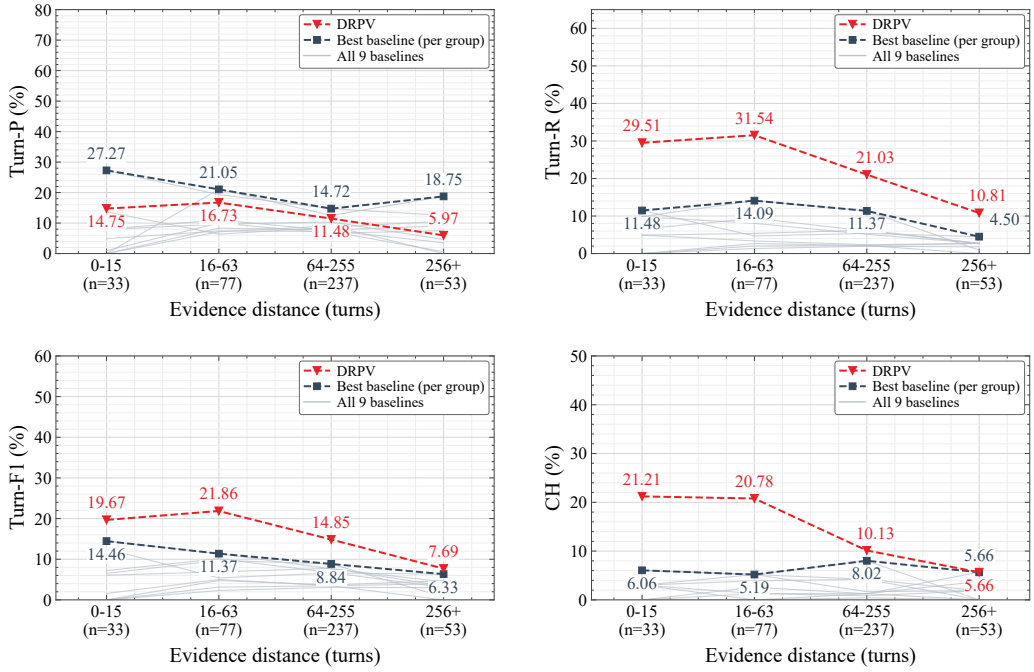

Figure 10: Performance across evidence-distance groups within 256K–512K source-length bucket.

Table 14: Supplementary metrics by evidence distance within the 512K–1M source-length bucket.

Metric	Distance	Direct	Sliding	NCP	ArcANE	NFS	WSE-S	WSE-V	Flawed	ConStory	DRPV
Flag Rate	0–15	34.29	82.86	14.29	25.71	42.86	97.14	65.71	14.29	40.00	100.00
	16–63	24.72	88.76	5.62	30.34	39.33	94.38	68.54	8.99	39.33	96.63
	64–255	29.41	86.63	10.70	24.60	40.64	94.12	72.73	6.42	39.57	96.26
	256+	22.47	82.02	11.24	24.72	28.09	88.76	65.17	12.36	40.45	95.51
Hit↑	0–15	2.86	14.29	2.86	0.00	8.57	17.14	8.57	0.00	5.71	**31.43**
	16–63	0.00	5.62	0.00	1.12	6.74	7.87	4.49	0.00	3.37	**19.10**
	64–255	1.60	8.56	1.07	1.60	4.28	16.04	10.70	1.07	4.81	**18.72**
	256+	3.37	8.99	1.12	2.25	0.00	13.48	10.11	0.00	4.49	**24.72**
ES↑	0–15	0.00	2.86	2.86	0.00	2.86	0.00	0.00	0.00	0.00	**8.57**
	16–63	0.00	0.00	0.00	0.00	2.25	0.00	0.00	0.00	0.00	**3.37**
	64–255	0.53	0.00	1.07	0.00	1.07	0.00	0.00	0.00	2.14	**3.74**
	256+	1.12	0.00	0.00	1.12	0.00	0.00	0.00	0.00	0.00	**2.25**
FE↑	0–15	2.86	**8.57**	2.86	0.00	5.71	0.00	2.86	0.00	2.86	**8.57**
	16–63	0.00	3.37	0.00	0.00	4.49	1.12	0.00	0.00	1.12	**6.74**
	64–255	1.07	2.67	1.07	1.07	2.14	2.14	2.67	0.53	3.74	**9.09**
	256+	3.37	4.49	0.00	1.12	0.00	2.25	4.49	0.00	3.37	**12.36**
FE±1 ↑	0–15	2.86	**11.43**	2.86	0.00	5.71	0.00	2.86	0.00	2.86	8.57
	16–63	0.00	4.49	0.00	0.00	5.62	1.12	0.00	0.00	1.12	**7.87**
	64–255	1.07	3.74	1.07	1.07	2.14	2.14	2.67	0.53	3.74	**9.09**
	256+	3.37	4.49	0.00	1.12	0.00	2.25	4.49	0.00	3.37	**12.36**
FE±3 ↑	0–15	2.86	**14.29**	2.86	0.00	5.71	0.00	2.86	0.00	2.86	8.57
	16–63	0.00	5.62	0.00	0.00	5.62	2.25	1.12	0.00	1.12	**8.99**
	64–255	1.07	5.35	1.07	1.07	2.14	2.14	2.67	0.53	4.28	**9.09**
	256+	3.37	4.49	0.00	1.12	0.00	2.25	4.49	0.00	3.37	**12.36**
MAE↓	0–15	258.42	**185.55**	433.20	304.22	259.33	233.82	223.30	414.40	276.14	186.37
	16–63	520.64	**159.09**	555.80	259.74	244.40	204.42	211.02	336.88	333.00	184.87
	64–255	455.84	193.48	339.15	360.35	261.53	202.95	196.63	258.50	273.85	**159.19**
	256+	363.05	190.37	429.80	292.55	213.00	251.75	244.22	447.91	290.75	**161.01**

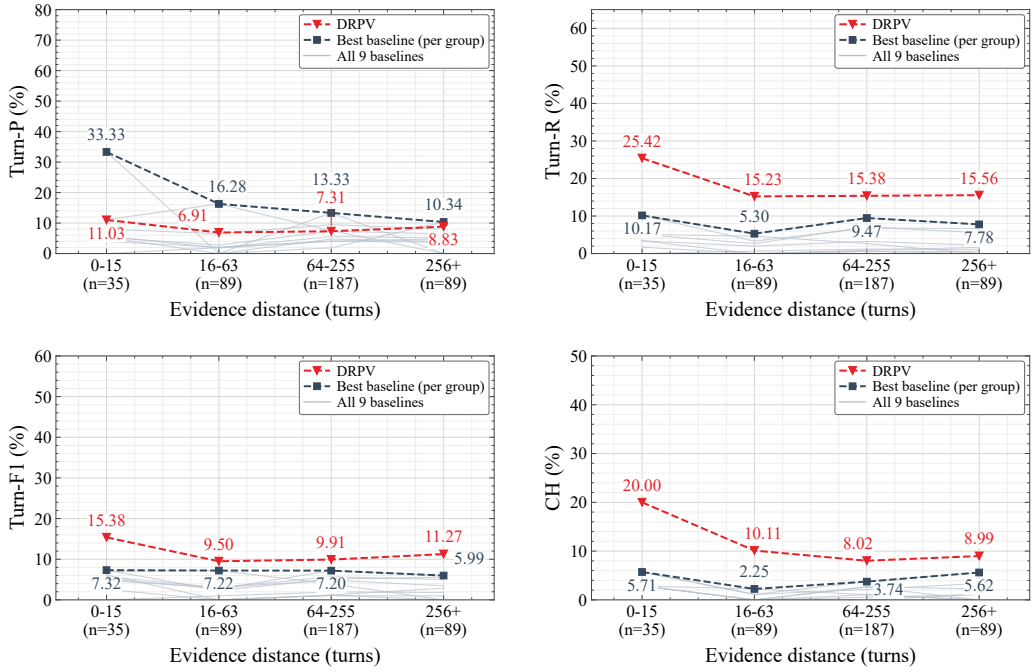

Figure 11: Performance across evidence-distance groups within 512K–1M source-length bucket.

G Detailed Results by Source Length

Tables 15–19 extend the main-text source-length comparison to all 11 evaluation measures defined in Appendix D, with 400 narratives in each of five source-length buckets. We examine precision–recall trends, the gap between detection and error-set recovery, and first-error localization. These comparisons describe associations with source length; differences in narrative composition and evidence distance prevent isolating its independent effect.

G.1 Turn-Level Localization Across Source Lengths

Precision and Turn-F1 decrease across successive source-length buckets for every method, while recall generally follows the same pattern. For DRPV, precision falls from 33.53% below 64K to 7.88% at 512K–1M, and recall declines from 37.50% to 16.21%. Its lower Turn-F1 therefore reflects both a smaller proportion of predictions matching the annotations and reduced recovery of annotated error turns. DRPV retains the highest Turn-F1 in every bucket. In the longest bucket, its recall is nearly twice that of WSE-S (8.24%), the strongest baseline on this metric, although its precision remains below that of NFS (8.56%) and ConStory (8.44%). Its stronger recall thus coexists with limited prediction precision. One possible explanation is that longer histories make it harder to recover relevant evidence and distinguish genuine contradictions from changes supported by intervening events. Missing evidence may reduce recall, while overlooking such changes may produce additional non-gold predictions.

G.2 Narrative Flagging and Error-Set Recovery

Method grouping. We group the evaluated implementations by how they organize conflict search. *Global auditing methods* include Direct, Flawed, and ConStory, which scan full narratives through a bounded number of judging passes, and ArcANE and NFS, which audit selected narrative-wide evidence. *Turn-/block-wise auditing methods* include NCP, WSE-S, and WSE-V, which label turns in batches; Sliding, which combines window-level predictions; and DRPV, which discovers candidates across blocks before selection and verification. This grouping concerns the organization of conflict search rather than the context length of individual requests: WSE receives narrative prefixes, and DRPV verifies selected candidates against complete prefixes.

Flagging versus localization. Flag Rate declines overall with source length for all global auditing methods, whereas the turn-/block-wise methods show rising or non-monotonic trends. These patterns describe the frequency of nonempty predictions rather than localization accuracy. From the shortest to the longest bucket, WSE-S's Flag Rate rises from 69.75% to 93.25%, while its Hit falls from 32.75% to 13.75%. In the longest bucket, DRPV has a Flag Rate of 96.50% but a Hit of only 21.25%. Thus, 75.25% of all narratives in that bucket receive nonempty predictions without any annotated error turn being identified. This gap concerns recovery of the injected errors and is not a narrative-level false-positive rate.

Error-set recovery. Recovering the complete error set remains difficult at greater source lengths. DRPV's CH falls from 27.50% in the shortest bucket to 9.75% in the longest, while ES decreases from 15.75% to 3.75%. The 6.00-percentage-point gap between CH and ES in the longest bucket represents narratives with complete gold coverage accompanied by non-gold predictions. DRPV achieves the highest CH across all five source-length buckets and the highest ES in the four buckets at or above 64K tokens. Below 64K, Direct achieves the highest ES at 18.50%, while ConStory achieves the highest Hit. The best ES across methods decreases from 18.50% in the shortest bucket to 3.75% in the longest. Despite DRPV's relative advantage, exact recovery remains uncommon in the longest narratives. Non-gold predictions refer to predicted turns outside the human-verified error-turn sets.

G.3 First-Error Localization Across Source Lengths

First-error accuracy generally declines with source length. DRPV achieves the highest FE in every bucket, but its score falls from 28.75% below 64K to 9.25% at 512K–1M. In the longest bucket, allowing deviations of one and three turns raises accuracy to 9.50% and 9.75%, respectively. The

small gains from these tolerance windows indicate that predictions within three turns of the annotated first error account for only a small fraction of exact-match failures. Most failures involve either no predicted location or an earliest prediction more than three turns away. Because first-error localization uses the earliest predicted turn, an additional non-gold prediction before the annotated first error can invalidate FE even when the true error turn is also recovered. Conditional MAE increases across successive buckets for every method. For DRPV, it rises from 19.93 to 167.78 turns; the latter is the lowest MAE among the evaluated methods in the longest bucket. However, raw MAE also reflects the larger range of deviations in longer histories and is computed over each method's own subset of nonempty predictions. It should therefore be interpreted alongside FE and Flag Rate, rather than as an isolated measure of degradation with source length.

Table 15: Localization results and Flag Rate for the < 64K source-token bucket ($N = 400$).

	Prediction behavior	Turn-level			Sample-level			First-error			
Method	Flag Rate	Turn-P↑	Turn-R↑	Turn-F1↑	Hit↑	CH↑	ES↑	FE↑	FE±1 ↑	FE±3 ↑	MAE↓
Direct	77.75	42.58	28.99	34.49	33.25	19.75	**18.50**	25.50	27.25	30.50	21.12
Sliding	76.00	42.60	25.27	31.72	29.50	17.50	17.50	23.50	24.00	27.25	21.75
NCP	52.00	39.41	21.28	27.63	24.50	14.00	13.25	18.50	19.25	21.50	20.48
ArcANE	63.00	40.93	21.01	27.77	24.50	16.50	16.25	20.50	22.25	24.75	20.14
NFS	59.25	35.22	14.89	20.93	21.25	7.75	7.00	13.75	16.00	18.25	22.65
WSE-S	69.75	27.70	24.20	25.83	32.75	13.75	5.25	13.75	15.75	19.25	20.33
WSE-V	58.50	32.92	21.01	25.65	28.25	13.50	5.50	16.50	18.25	20.25	**18.48**
Flawed	50.25	**45.51**	18.22	26.02	20.50	13.75	13.50	16.25	17.25	19.00	20.12
ConStory	92.75	25.74	32.18	28.61	**49.50**	18.50	8.00	24.75	27.25	31.00	19.42
DRPV	87.75	33.53	**37.50**	**35.40**	44.00	**27.50**	15.75	**28.75**	**30.00**	**33.25**	19.93

Table 16: Localization results and Flag Rate for the 64K–128K source-token bucket ($N = 400$).

	Prediction behavior	Turn-level			Sample-level			First-error			
Method	Flag Rate	Turn-P↑	Turn-R↑	Turn-F1↑	Hit↑	CH↑	ES↑	FE↑	FE±1 ↑	FE±3 ↑	MAE↓
Direct	63.50	32.86	17.99	23.26	22.00	13.25	12.75	17.25	17.75	18.75	37.68
Sliding	71.25	35.68	18.89	24.71	24.00	12.75	12.25	19.50	21.25	23.50	34.89
NCP	41.25	**38.21**	12.08	18.36	15.25	8.25	7.75	11.50	12.25	13.00	39.79
ArcANE	54.50	29.73	14.14	19.16	18.00	9.00	8.75	13.50	14.50	15.25	37.94
NFS	55.50	23.78	9.38	13.46	15.25	5.00	5.00	10.00	10.50	12.00	41.27
WSE-S	83.75	17.97	21.21	19.46	30.00	13.50	2.50	12.75	14.25	17.00	35.10
WSE-V	71.00	28.47	20.82	24.05	29.25	13.75	4.50	15.25	16.50	19.50	**30.64**
Flawed	45.50	32.14	10.41	15.73	14.00	6.75	6.25	11.00	13.25	15.00	31.20
ConStory	89.50	22.29	24.55	23.36	42.00	11.50	5.75	22.25	24.00	25.00	32.81
DRPV	91.25	29.01	**35.60**	**31.97**	**45.25**	**27.00**	**13.75**	**26.75**	**28.25**	**31.75**	31.19

Table 17: Localization results and Flag Rate for the 128K–256K source-token bucket ($N = 400$).

	Prediction behavior	Turn-level			Sample-level			First-error			
Method	Flag Rate	Turn-P↑	Turn-R↑	Turn-F1↑	Hit↑	CH↑	ES↑	FE↑	FE±1 ↑	FE±3 ↑	MAE↓
Direct	62.00	19.26	10.59	13.66	13.25	6.25	6.25	10.75	11.00	11.75	83.71
Sliding	58.75	**24.61**	10.08	14.30	14.75	6.25	6.25	12.00	13.00	13.75	67.18
NCP	32.75	20.76	6.25	9.61	8.00	2.75	2.50	5.75	6.50	6.75	83.92
ArcANE	46.25	18.93	6.76	9.96	9.50	3.75	3.25	6.75	6.75	8.25	84.56
NFS	57.75	21.58	9.06	12.76	14.50	3.25	3.00	9.25	9.75	11.00	62.92
WSE-S	83.75	9.60	14.92	11.68	23.00	9.00	1.75	8.50	8.50	9.50	81.75
WSE-V	66.00	13.83	11.61	12.62	19.75	5.75	2.25	9.75	10.00	10.25	77.41
Flawed	44.50	22.92	7.40	11.19	10.00	5.00	4.75	8.25	8.75	9.50	77.54
ConStory	85.75	16.21	15.69	15.94	28.50	6.25	2.00	10.50	11.00	11.50	71.71
DRPV	94.50	20.58	**32.53**	**25.21**	**43.50**	**22.75**	**10.50**	**22.50**	**23.50**	**25.00**	**62.77**

Table 18: Localization results and Flag Rate for the 256K–512K source-token bucket ($N = 400$).

	Prediction behavior	Turn-level			Sample-level			First-error			
Method	Flag Rate	Turn-P↑	Turn-R↑	Turn-F1↑	Hit↑	CH↑	ES↑	FE↑	FE±1 ↑	FE±3 ↑	MAE↓
Direct	43.00	3.85	2.80	3.24	5.00	1.25	1.25	3.00	4.00	4.50	185.93
Sliding	59.75	9.23	5.46	6.86	7.50	3.50	1.75	4.25	4.25	5.25	148.25
NCP	23.50	7.26	1.65	2.69	2.50	1.00	1.00	1.50	2.25	2.25	185.44
ArcANE	30.75	7.18	1.65	2.69	2.75	0.25	0.25	1.50	1.50	2.25	138.13
NFS	49.25	**16.61**	5.72	8.51	9.00	2.50	2.50	7.00	7.25	7.50	**122.56**
WSE-S	89.25	6.21	10.29	7.74	17.25	6.25	0.50	3.50	4.50	5.00	145.96
WSE-V	63.25	7.75	6.23	6.91	10.00	4.25	1.00	4.50	4.75	4.75	151.36
Flawed	22.75	13.60	2.16	3.73	2.75	1.75	1.50	2.25	2.25	2.25	168.88
ConStory	70.50	8.84	5.59	6.85	10.75	1.75	0.75	4.00	4.50	5.25	154.70
DRPV	99.00	12.00	**22.24**	**15.59**	**30.75**	**12.50**	**6.00**	**9.50**	**9.50**	**9.75**	124.19

Table 19: Localization results and Flag Rate for the 512K–1M source-token bucket ($N = 400$).

	Prediction behavior	Turn-level			Sample-level			First-error			
Method	Flag Rate	Turn-P↑	Turn-R↑	Turn-F1↑	Hit↑	CH↑	ES↑	FE↑	FE±1 ↑	FE±3 ↑	MAE↓
Direct	27.25	2.94	0.96	1.45	1.75	0.50	0.50	1.50	1.50	1.50	430.16
Sliding	85.75	3.85	6.32	4.79	8.50	3.75	0.25	3.75	4.75	6.00	184.22
NCP	10.00	6.25	0.69	1.24	1.00	0.75	0.75	0.75	0.75	0.75	400.65
ArcANE	26.00	3.89	0.96	1.54	1.50	0.25	0.25	0.75	0.75	0.75	315.03
NFS	37.75	**8.56**	2.61	4.00	4.25	1.25	1.25	2.50	2.75	2.75	249.30
WSE-S	93.25	3.53	8.24	4.95	13.75	2.50	0.00	1.75	1.75	2.00	216.43
WSE-V	69.50	6.00	5.63	5.81	9.00	2.00	0.00	2.50	2.50	2.75	211.92
Flawed	9.00	4.44	0.27	0.52	0.50	0.00	0.00	0.25	0.25	0.25	355.44
ConStory	39.75	8.44	2.75	4.15	4.50	1.50	1.00	3.00	3.00	3.25	290.90
DRPV	96.50	7.88	**16.21**	**10.61**	**21.25**	**9.75**	**3.75**	**9.25**	**9.50**	**9.75**	**167.78**

H SENSITIVITY TO THE VERIFICATION CAP

Coverage-precision trade-off. Figure 5 reports the primary localization metrics across verification caps $K = 1, \ldots, 10$. Across the tested caps, Turn-R increases from 11.91% at $K = 1$ to 31.86% at $K = 10$, while Turn-P decreases from 26.98% to 14.10%. Turn-F1 peaks at 23.08% with $K = 3$. At the default $K = 5$, CH increases from 16.05% to 19.90%, while Turn-F1 decreases to 22.55%. The default therefore provides greater complete gold coverage, with a Turn-F1 0.53 percentage points below the highest value observed in the sweep.

Flagging and error-set recovery. Figure 12 examines prediction frequency alongside partial and exact recovery of the annotated error set. Flag Rate rises from 84.50% at $K = 1$ to 93.80% at $K = 5$, remaining between 94.00% and 94.30% for larger caps. Hit increases from 22.80% at $K = 1$ to 36.95% at $K = 5$ and 40.20% at $K = 10$. In contrast, ES reaches its highest observed value of 9.95% at K = 5 and falls to 4.75% at K = 10. At K = 5, CH is 19.90%, and ES is 9.95%, meaning that 9.95% of all narratives have complete gold coverage accompanied by additional non-gold predictions. Recovering all annotated errors therefore does not necessarily produce an exact prediction set. Flag Rate describes output frequency and is interpreted separately from localization quality.

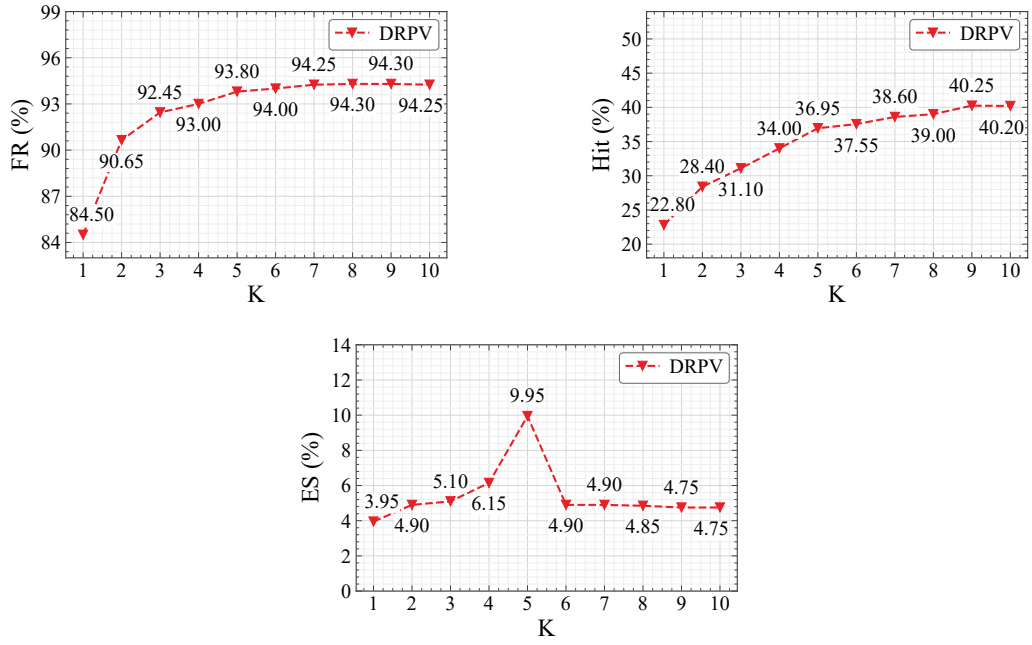

Figure 12: Flag Rate, Any-turn Hit, and Exact Set across verification caps $K = 1, \ldots, 10$. Values are percentages. Flag Rate measures the frequency of valid, nonempty predictions and has no preferred direction; Hit and ES measure recovery of at least one annotated error turn and exact recovery of the annotated error set, respectively.

First-error localization. Figure 13 reports FE, FE±1, FE±3, and conditional MAE. The three accuracy measures attain their highest observed values at $K = 2$: 19.45%, 20.20%, and 22.15%, respectively. FE remains between 19.35% and 19.45% for K = 2–5, indicating relatively little variation over this range despite increasing complete gold coverage. MAE decreases overall from 125.67 turns at $K = 1$ to 80.59 turns at $K = 10$. However, it is computed only over valid, nonempty predictions, and the evaluated subset varies across settings. Figure 12 reports the proportion of narratives included through Flag Rate. Lower mean absolute deviation does not necessarily imply more exact or near-exact matches, even on a fixed evaluation subset. Here, the subset of valid, nonempty predictions also varies across caps. The MAE decrease should therefore be interpreted jointly with first-error accuracy and Flag Rate, rather than as evidence of uniformly improved first-error localization.

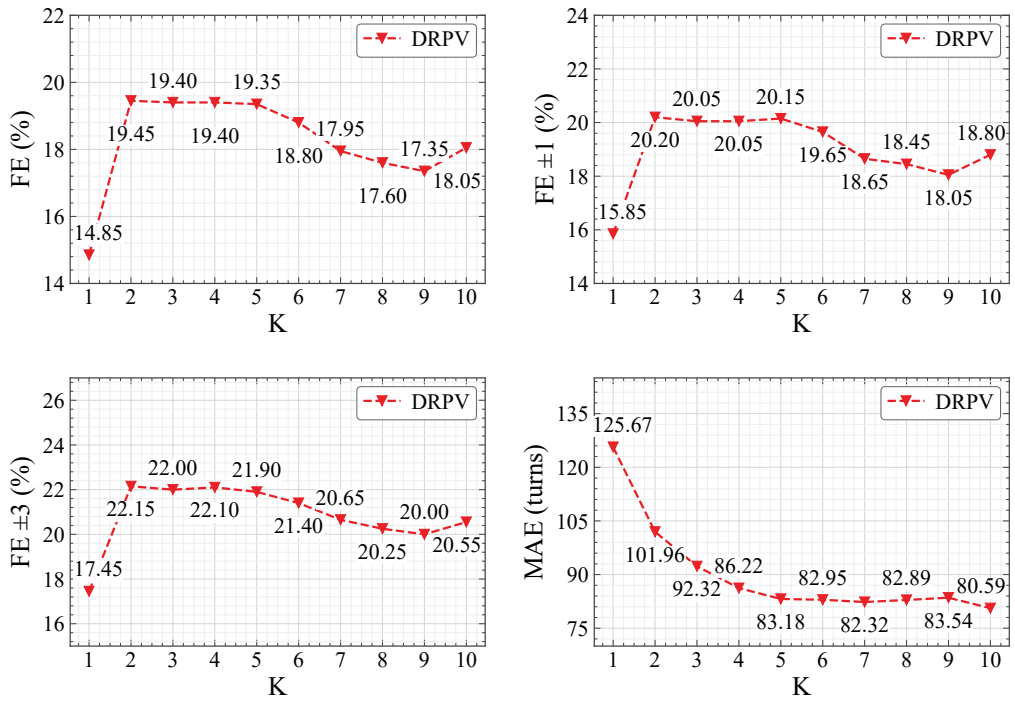

Figure 13: First-error localization across verification caps. FE, FE±1, and FE±3 are percentages evaluated over all narratives. MAE is measured in turns over valid, nonempty predictions only; the Flag Rate is reported in Figure 12. The MAE subset can differ across verification caps.

Verification Workload. Figure 14 reports the average number of logical verifications per completed narrative, including cached verifications and excluding API retries. Moving from K = 3 to K = 5 increases average logical verifications from 2.61 to 3.85, approximately 48%, alongside gains of 3.85 percentage points in CH and 4.85 points in ES, with a 0.53-point reduction in Turn-F1. Increasing K from 5 to 10 raises the average number of logical verifications per completed narrative from 3.85 to 6.00, approximately 56%, while increasing CH by 2.10 percentage points and reducing Turn-F1 and ES by 3.00 and 5.20 points, respectively. These results quantify the additional verification workload associated with higher coverage at larger caps. Logical verification counts describe the verification stage and do not measure end-to-end token usage, latency, or monetary cost.

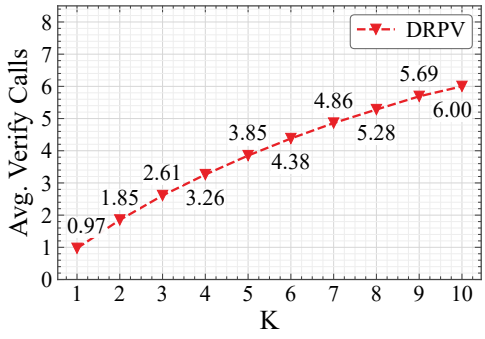

Figure 14: Average logical verifications per successfully completed narrative across verification caps. Counts include cached verifications and exclude API retries.

Table 20: Reported error-turn counts and derived false-alarm summaries. Each entry shows reported error turns/flagged histories (false-alarm rate). Overall counts sum the five source-length buckets.

Method	< 64K ($n = 10$)	64K–128K ($n = 10$)	128K–256K ($n = 10$)	256K–512K ($n = 10$)	512K–1M ($n = 10$)	Overall ($n = 50$)
Direct	12 / 7 (70%)	8 / 5 (50%)	13 / 6 (60%)	22 / 3 (30%)	3 / 2 (20%)	58 / 23 (46%)
Sliding	8 / 6 (60%)	8 / 6 (60%)	7 / 5 (50%)	11 / 5 (50%)	24 / 7 (70%)	58 / 29 (58%)
NCP	7 / 4 (40%)	10 / 3 (30%)	5 / 3 (30%)	4 / 2 (20%)	7 / 2 (20%)	33 / 14 (28%)
ArcANE	7 / 5 (50%)	7 / 4 (40%)	5 / 3 (30%)	5 / 3 (30%)	3 / 2 (20%)	27 / 17 (34%)
NFS	7 / 5 (50%)	6 / 4 (40%)	8 / 4 (40%)	6 / 4 (40%)	7 / 3 (30%)	34 / 20 (40%)
WSE-S	15 / 7 (70%)	14 / 6 (60%)	30 / 8 (80%)	34 / 9 (90%)	36 / 9 (90%)	129 / 39 (78%)
WSE-V	10 / 5 (50%)	9 / 5 (50%)	15 / 6 (60%)	15 / 6 (60%)	16 / 6 (60%)	65 / 28 (56%)
Flawed	8 / 5 (50%)	6 / 4 (40%)	5 / 3 (30%)	2 / 1 (10%)	1 / 1 (10%)	22 / 14 (28%)
ConStory	18 / 8 (80%)	18 / 7 (70%)	14 / 7 (70%)	9 / 5 (50%)	6 / 4 (40%)	65 / 31 (62%)
DRPV	**16 / 6 (60%)**	**22 / 7 (70%)**	**17 / 5 (50%)**	**11 / 4 (40%)**	**9 / 5 (50%)**	**75 / 27 (54%)**

I False Alarms on Reviewed Clean Histories

Evaluation setup. We evaluate DRPV and nine baselines on 50 reviewed pre-injection histories, with 10 histories per source-length bucket. Each method evaluates each history once using the same configuration as in the main localization experiment. This evaluation complements localization on error-injected variants by examining whether evaluators flag contradictions in histories labeled as clean.

Human verification. We manually review all model-reported contradictions against the relevant historical evidence and narrative context. Review follows the same semantic-contradiction criterion used in benchmark construction: conflicting statements must concern established narrative facts under compatible temporal conditions, without a narrative-supported explanation reconciling them. No reported contradiction is confirmed as valid under this criterion. The resulting alarms are therefore treated as false positives relative to the reviewed clean labels.

False-alarm counts and rates. For each method, we count the histories receiving a valid, nonempty predicted error-turn set. A history contributes at most one false alarm, regardless of the number of turns reported. The false-alarm rate is this count divided by the number of evaluated histories: 10 within each source-length bucket and 50 overall. Execution failures remain in the denominators and do not establish correct clean classification. Table 20 reports counts and rates together.

Interpretation. These results complement the localization measures reported on error-injected variants. A lower false-alarm rate indicates less frequent flagging of reviewed clean histories, but does not by itself establish stronger contradiction detection, since a conservative evaluator may also miss injected errors. We therefore interpret false-alarm behavior jointly with localization performance. Comparisons across source-length buckets are descriptive: each bucket contains only 10 histories, and differences in narrative composition prevent isolating a causal effect of length.

J Qualitative Case Studies

We present two illustrative cases of DRPV's error-turn localization with $K = 5$: exact recovery of repeated contradictions and complete error coverage accompanied by false positives. Chinese excerpts are paired with English translations and retain the original turn IDs. Ellipses indicate omitted text, and prediction sets are shown in full.

Case A: Recovering repeated contradictions. *128K–256K; 329 turns.*

Turn	Chinese original	English translation
4	十分钟后，苏世墨趴在地上，林天还是背着手看着他道："还来么？已经趴了三回了。"	Ten minutes later, Su Shimo lay on the ground. Lin Tian, still watching with his hands behind his back, asked, "Again? You've already gone down three times."
260	他想去打听那位与自己从未切磋过的枪皇门人林天的消息。	He wanted news of Lin Tian, the Spear Emperor's disciple, with whom he had never sparred.
261	自己虽然从未和他切磋过，也能从传闻中看得出来。	Although he had never sparred with him, he could tell from what he had heard.

Gold: $\{260, 261\}$; **ConStory:** $\{96\}$; **DRPV:** $\{260, 261\}$.

The historical encounter contradicts both later denials of prior sparring. The logs show that retrieval supplied turn 4 for both targets, and verification accepted both candidates. DRPV therefore recovers the two annotated error turns exactly, whereas ConStory misses both. This case illustrates successful evidence linking and localization of repeated expressions of the same error.

Case B: Complete coverage with extra predictions. *64K–128K; 199 turns.*

Turn	Chinese original	English translation
23	…刘景舟首次以夜游神形象示人。	…Liu Jingzhou appeared as the Night Patrol Deity for the first time.
136	刘景舟微微挥手，化身日游神…	Liu Jingzhou waved slightly and transformed into the Day Patrol Deity…
6	就这九重境界，也才是仙人九境，要想恢复神灵手段，得到第十境的登楼境界…	These nine realms are only the nine immortal realms; regaining divine powers requires reaching the tenth realm, Tower Ascension…
62	炼气士十二重境界，每三境就是一道坎儿。	Qi cultivators have twelve realms, with a threshold every three realms.

Gold: $\{127, 136\}$; **ConStory:** $\{84, 156\}$; **DRPV:** $\{62, 127, 136, 154\}$.

DRPV identifies both injected substitutions of "Day Patrol Deity" for "Night Patrol Deity," but also reports turns 62 and 154. At turn 62, the verifier incorrectly treats "nine immortal realms" as an exhaustive total and accepts a conflict with "twelve realms." Yet turn 6 explicitly mentions a tenth realm, so the cited statements do not establish that conflict. Relative to the annotations, DRPV obtains $P = 50\%$, $R = 100\%$, and $F_1 = 66.67\%$, with $\mathrm{CH} = 1$ and $\mathrm{ES} = 0$ for this case. Complete coverage can thus coexist with extra predictions; access to the complete prefix does not guarantee correct interpretation of a claim's scope.