NOVELAUDIT: A BENCHMARK FOR OPEN-ENDED NARRATIVE CONSISTENCY AUDITING IN FULL-LENGTH NOVELS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated stories and scripts create a practical need for narrative consistency auditing. We begin with complete human-authored novels, where apparent contradictions may have distant explanations. Existing tasks emphasize supplied claims, predefined error categories, or detected errors, and largely omit explained reversals from required discovery outputs. We introduce NOVELAUDIT, which asks agents to discover explained reversals and unresolved inconsistencies without supplied cases and support them with role-structured quotations. We construct references for 13 principal and 11 extended novels; evaluation separates task completion, case discovery, and explanatory sufficiency. On the principal novels, we compare DeepSeek across six runtimes, with supplementary Qwen and GPT–Codex conditions. Differences in runtime and model–runtime adaptation are clearest in completion rate, while single attempts discover limited, changing case subsets. Complementary discoveries across rounds expand coverage at increasing input cost per new match. Human audits link Explanation-Path F1 to explanatory sufficiency and identify valid predicted cases missed by matching, distinguishing discovery from evidence quality.

1 INTRODUCTION

Advances in long-form generation, planning, and revision support sustained narrative production (Yang et al., 2022; 2023; Bai et al., 2025). Human–AI writing systems support collaborative and professional work on stories and scripts (Lee et al., 2022; Ippolito et al., 2022; Mirowski et al., 2023), while research systems coordinate scenes and audiovisual production for films (Zhao et al., 2025; Wu et al., 2025). Yet generated narratives exhibit plot contradictions and inconsistent character states (Li et al., 2026; Aluru et al., 2026), and writers report difficulties preserving style and voice (Ippolito et al., 2022). These uses and failures motivate tools that locate inconsistencies and determine whether the narrative resolves them. We focus on state consistency; style and overall literary merit remain outside scope.

Models and agents increasingly evaluate generated text, using ratings, debate, or agentic review (Liu et al., 2023; Zheng et al., 2023; Chan et al., 2024; Zhuge et al., 2025). Narrative evaluators assess story quality, preferences, and criticism, including novel-length inputs (Chen et al., 2022; Chhun et al., 2024; Yang & Jin, 2025; Fein et al., 2026). However, weaknesses in evaluation explanations and novel-summary faithfulness judgments raise questions about their evidence-finding ability (Chhun et al., 2024; Kim et al., 2024). *Can current agents discover potentially contradictory state reversals throughout a complete long novel, locate the evidence, and determine whether each reversal is explained?*

Long narratives require integrating dispersed information (Kočiský et al., 2018; Karpinska et al., 2024), while long-context models remain sensitive to evidence position, separation, and absent literal matches (Liu et al., 2024; Tian et al., 2025; Modarressi et al., 2025). Existing tasks detect narrative errors (Ahuja et al., 2025; Li et al., 2026; Aluru et al., 2026), verify supplied novel claims (Karpinska et al., 2024), or extract relation-specific lists (Singhania et al., 2025). They do not jointly require discovering unspecified reversals across complete novels, retaining explained cases, and supplying

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.

separate evidence for opposing states and their resolution. Error-only outputs cannot distinguish an overlooked reversal from one found and adequately explained.

We begin with pure text and established human-authored novels: inspectable works with relatively strong expected consistency (Section 3.3). For example, *Dracula*'s apparent old-to-young reversal is explained by feeding evidence and a rule about restored youth (Figure 1). Agents should discover both the reversal and its explanation. Paired explanation deletions (INJ) simulate missing explanations in weak generated narratives (Section 3.4); generated novels and multimodal works remain future settings (Section 6).

NOVELAUDIT gives agents a complete novel and task definition without supplying entities, state dimensions, cases, or an error taxonomy. Here, "open-ended" refers to discovering candidate cases from the full novel without a predefined list of targets, claims, or text pairs. Agents discover qualifying reversals, label them EXPLAINED or UNRESOLVED, and quote role-specific evidence. Ordinary progression does not qualify. A hidden, finite reference records alternative endpoint anchors and complete explanatory routes, separating discovery, labels, and evidence support.

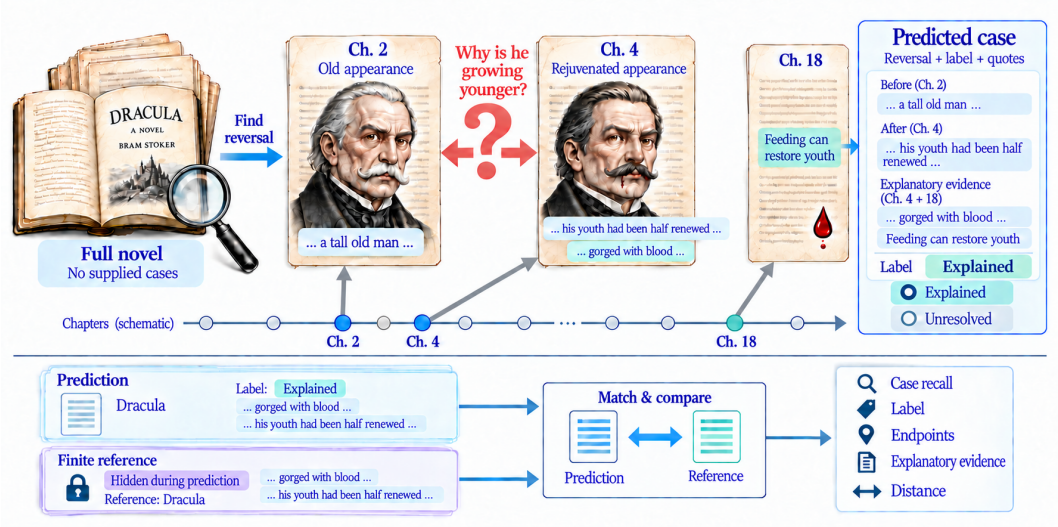


Figure 1: NOVELAUDIT task and evaluation. The Dracula example links opposing appearances to feeding evidence and an explanatory rule. Predictions contain the state reversal, label, and role-specific quotations; a separate reference remains hidden during prediction. Chapter positions are schematic; full passages appear in Appendix A.1.

We evaluate DeepSeek across six native runtimes, with Qwen and GPT–Codex supplementary conditions. Model–runtime adaptation primarily affects completion rate; even reliable configurations have low case recall. Repeated attempts discover complementary subsets, adding cases while missing earlier discoveries, with increasing input cost per new match. Human audits associate Explanation-Path F1 with explanatory sufficiency and identify valid predicted cases missed by matching, requiring case recall and evidence quality to be interpreted together.

Our contributions are:

- **A task for narrative consistency auditing.** Open-ended discovery in complete novels retains explained reversals and unresolved inconsistencies, with role-structured endpoint and explanatory evidence.
- **References for 24 full-length novels.** Candidate pooling and full-text adjudication cover 13 principal and 11 extended novels; performance evaluation uses the principal collection. Stopping depends on time and cost. The fixed reference supports consistent evaluation and lets other researchers add or revise cases in subsequent versions under the same definitions.
- **An evaluation protocol and empirical characterization.** A common output contract separates completion, discovery, and explanatory sufficiency; system comparisons, repeated discovery, and human audits characterize agent behavior and score interpretation.

2 Related Work

Narrative consistency and error discovery. FlawedFictions detects short-story continuity errors without hints and identifies both conflicting sentence sets; ResolvedNegs adds resolving context and treats repaired stories as negatives (Ahuja et al., 2025). ConStory-Bench elicits taxonomy-defined errors with quotations and locations in generated stories (Li et al., 2026). ATLAS checks grounded screenplay world states, excluding explained changes from error judgments (Aluru et al., 2026). In complete novels, explanations may lie far from opposing states. NovelAudit retains explained reversals alongside unresolved inconsistencies, requiring endpoint evidence and, for explained reversals, complete explanatory routes. These requirements distinguish omission, discovery, and sufficient explanation. Where ResolvedNegs adds explanations, our paired edits target their deletion within the same narrative.

Long-form narrative generation. Re3 and DOC use structured planning to improve coherence (Yang et al., 2022; 2023); LongWriter extends generation length, and StoryWriter coordinates writing agents (Bai et al., 2025; Xia et al., 2025). Collaborative writing and script systems connect these capabilities to creative practice (Lee et al., 2022; Ippolito et al., 2022; Mirowski et al., 2023); MovieDreamer and MovieAgent extend generation to films (Zhao et al., 2025; Wu et al., 2025). NovelAudit tests the auditing capability these narratives require, using inspectable human-authored novels and paired explanation deletions that simulate missing explanations in weak generated narratives.

Long-context verification and information discovery. LongBench benchmarks long-context tasks (Bai et al., 2024); NarrativeQA asks narrative questions (Kočiský et al., 2018). Evidence position, separation, and absent literal matches challenge long-context models (Liu et al., 2024; Tian et al., 2025; Modarressi et al., 2025). NoCha verifies supplied novel claims (Karpinska et al., 2024); FABLES evaluates novel-summary faithfulness (Kim et al., 2024). Recall Them All extracts objects for supplied subjects and relations using repeated-search unions (Singhania et al., 2025); Tan et al. (2026) also study evidence unions. In NovelAudit, agents receive very long, sometimes multi-volume novels without supplied subjects or state dimensions. They must choose what to investigate and locate opposing states and explanations across chapters.

Models and agents as narrative evaluators. HANNA, StoryER, prompted LLM evaluation, LongStoryEval, and LitBench study story quality, ratings, criticism, or preferences (Chhun et al., 2022; Chen et al., 2022; Chhun et al., 2024; Yang & Jin, 2025; Fein et al., 2026). General evaluation uses LLM ratings, multi-agent debate, and agentic assessment (Liu et al., 2023; Zheng et al., 2023; Chan et al., 2024; Zhuge et al., 2025). Narrative consistency auditing additionally requires discovering cases and substantiating judgments throughout the source. NovelAudit distinguishes incomplete execution, omitted cases, and insufficient evidence across model–runtime configurations, which an overall story rating does not separately measure.

Evidence organization and evaluation. FIND uses inconsistency lists and lexical matching (Lovering et al., 2026); FEVER allows alternative complete evidence sets (Thorne et al., 2018). Rationale evaluation studies agreement, faithfulness, redundancy, and unannotated support (DeYoung et al., 2020; Carton et al., 2020; Ni et al., 2021); inference judgments vary with interpretation (Pavlick & Kwiatkowski, 2019). In novels, opposing-state evidence and explanatory passages serve different roles, and explanations may require multiple passages. NovelAudit records alternative anchors and complete explanatory routes, separating case identity, label correctness, endpoint support, and explanatory sufficiency, with human review of selected judgments.

3 The NovelAudit Benchmark

3.1 Open-Ended Auditing and Case Qualification

The input is a full-length novel X. Without access to the reference, the agent discovers and returns a set of cases $\widehat{\mathcal{C}}$. It receives no supplied entities, state dimensions, candidate pairs, cases, or fixed inconsistency taxonomy. A case contains a subject, a state dimension, opposing before and after

states, their source evidence, a label, and any explanatory evidence. The public output contract appears in Appendix A.2.

Qualification before adjudication. Review proceeds through five reader-facing tests: (1) the endpoints concern the same subject and state dimension; (2) they express opposing states of the same explicit proposition; (3) elapsed time alone does not make them ordinary, naturally compatible progression; (4) with the explanatory material hidden, a careful reader would reasonably ask why the change occurred and suspect discontinuity; and (5) specific source evidence does or does not fully explain that reversal. Tests 1–4 qualify the candidate; only Test 5 selects EXPLAINED or UNRESOLVED.

Ordinary progression does not qualify, but a reversal need not be logically impossible across time: beliefs and commitments can reverse. Boundary examples appear in Table 40.

What the labels mean. An EXPLAINED reversal has a complete explanatory route supported by the text. An UNRESOLVED inconsistency in the reference lacks such a route after full-text review within the benchmark's adjudicated scope. An agent's predicted UNRESOLVED label may reflect missed explanatory evidence and requires independent evaluation.

Prediction, reference, and scoring. A prediction is the complete set of cases returned by an attempt. The scorer uses a separately constructed, finite reference to evaluate the agent's case discovery, labels, and evidence; case counts do not rate the novel's quality.

3.2 Role-Structured Reference Evidence

Each reference endpoint has a state claim and independently sufficient alternative anchors (Appendix Figure 28). An explanation requires one complete route (OR). A route combines all requirements (AND); equivalent passages provide alternative support for each (OR). Agents supply one minimal sufficient route per predicted explained reversal; partial evidence from different routes is not pooled. This separates endpoint support from explanatory sufficiency (Section 4); quotation selection and full-text review appear in Appendix G.

3.3 Corpus, Discovery, and Reference Freeze

Human-authored novels provide complete, inspectable pure-text narratives without first generating and screening uneven outputs or aligning cross-modal evidence. Established works, including classics, offer relatively strong expected consistency; their explained reversals test discovery of both apparent contradictions and their resolutions. We construct references for 13 principal and 11 extended novels, using the same task and evidence representation. Performance evaluation uses only the principal English/Chinese panel, including texts exceeding a million tokens. References are dominated by explained reversals; corpus and case statistics appear in Appendices C and D.

Discovery, stopping, and source eligibility. We pool anonymized candidates from initial DeepSeek attempts across multiple runtimes and later GPT-5.6 attempts with Codex. We merge equivalent cases and adjudicate labels and evidence through full-text review. Principal discovery stops after repeated low yield across evidence-distance bands or at a round cap, reflecting time and cost constraints. After eligibility and freeze checks, the principal reference includes initial human, GPT-only, and shared GPT/DeepSeek cases, excluding DeepSeek-exclusive cases. Formal evaluation predictions do not enter initial pooling; agents receive no reference material. Other researchers can add or revise cases in subsequent versions under the same requirements; this version remains fixed for reproducible comparisons. Construction and corrections appear in Appendices G and E.7.

3.4 Paired Text Edits Targeting Explanatory Evidence

Each principal novel has one edited copy (INJ), simulating missing explanations in weak generated narratives. Edits target explanatory routes for deletion while seeking to preserve endpoints, case identity, and unrelated cases. Original/INJ references map the same cases for paired comparison; copies do not add independent novels. Residual explanations remain in audited targets (Section 5.5), so INJ is an edited-text diagnostic. Target selection and editing checks appear in Appendix G.5.

4 Evaluation Protocol and Validation

4.1 Case Matching and Component Metrics

The scorer uses quote-grounded lexical matching to check target compatibility and same-role evidence thresholds before global one-to-one assignment. Contiguity-Augmented Anchor Coverage (CAAC) rewards anchor coverage and contiguous matches, penalizing excess quotation length. Matching and case-quality scoring share evidence signals, making case recall sensitive to evidence expression (Appendix H).

For Explanation-Path F1, the scorer matches predicted explanatory passages one-to-one to requirements within a complete route using CAAC weights. For route s with K_s requirements, P predicted passages, and maximum matching weight W_s,

$$F_{\text{route}} = \max_s \frac{2W_s}{P + K_s}. \tag{1}$$

The best recorded route is selected; unmatched passages increase P without increasing W_s. The case-quality score (Distance-Aware Case Evaluation, DACE) is

$$q_j = 0.25M_j + 0.20L_j + 0.25A_j + 0.30E_j, \tag{2}$$

where M indicates matching, L label correctness, and A the geometric mean of before/after CAAC. For correctly labeled explained reversals, E is Explanation-Path F1; for correctly labeled unresolved inconsistencies with empty explanatory evidence, $E = A$; otherwise $E = 0$. Wrong labels set $L = E = 0$; unmatched reference cases have all components zero.

4.2 From Case-Quality Score to Composite

Case recall is the fraction of included cases in the frozen reference that receive a match. Case precision P_{case} is the fraction of predicted cases that receive a match; unmatched predicted cases need not be invalid. Matched-label accuracy uses only matched cases; label recall and further diagnostics are defined in Appendix H. For the included frozen reference set G, $Q_{\text{weighted}} = \sum_{j \in G} w_j q_j / \sum_{j \in G} w_j$. Weights depend on the shortest complete evidence span and band density (Equation 15); unmatched cases retain their denominator weights.

$$\text{Composite} = 100 \frac{2P_{\text{case}}Q_{\text{weighted}}}{P_{\text{case}} + Q_{\text{weighted}}}. \tag{3}$$

Scores are computed per attempt, with N/A for failures; completion, aggregation, and metric availability are reported separately (Section 5; Appendix H).

4.3 Human Review and Its Scope

Independent blinded challenge-pair review assesses correspondence, endpoint support, and explanatory sufficiency; separate probability audits examine matched pairs and unmatched predicted cases. Sampling, exposure, and adjudication are documented in Appendices B, E.11, and E.17. Among qualified, identity-confirmed explained pairs, both reviewers' judgments align with mean Explanation-Path F1 ordered from sufficient through partial to insufficient explanations (Appendix F.3). This supports Explanation-Path F1, not Composite as a whole. Qualification and explanatory interpretation remain sources of disagreement (Appendix F.5).

5 Experiments and Findings

5.1 Experimental Conditions and Statistical Analysis

Models and runtimes. We evaluate DeepSeek V4.1 Flash in six native runtimes: Claude Code (CC), DeepSeek Harness (DSH), Codex Harness (CXH), Qwen Code (QWC), OpenCode (OC), and mini-SWE-agent (MSA), each with three attempts on all 13 principal novels. Native tools, loops, and context management remain part of the configurations. Each attempt receives the read-only novel, canonical task, public output contract, and empty scratch space, without references or prior

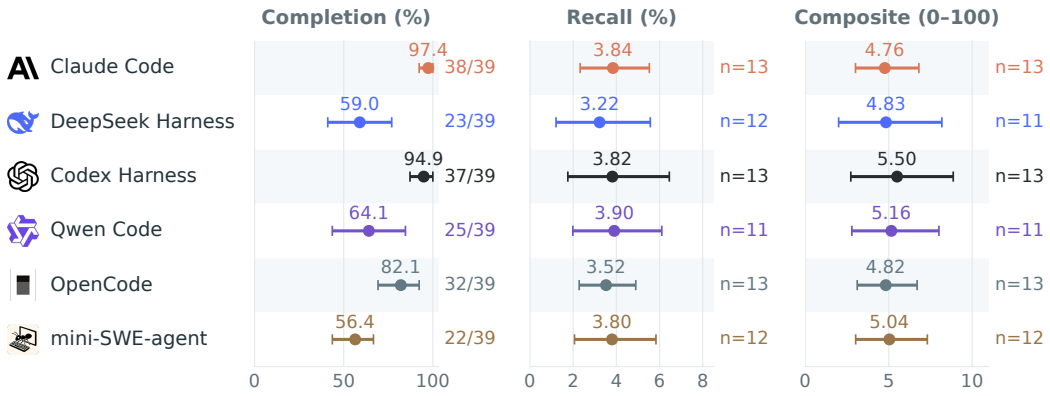


Figure 2: DeepSeek V4.1 Flash across six runtimes: all 13 Original novels, three attempts each (39 per runtime). Fractions show successful/total attempts; n counts novels with scorable outputs for the metric. $n < 13$ means some novels had none across all three attempts. Scores average within, then across novels; bars show 95% book-bootstrap intervals.

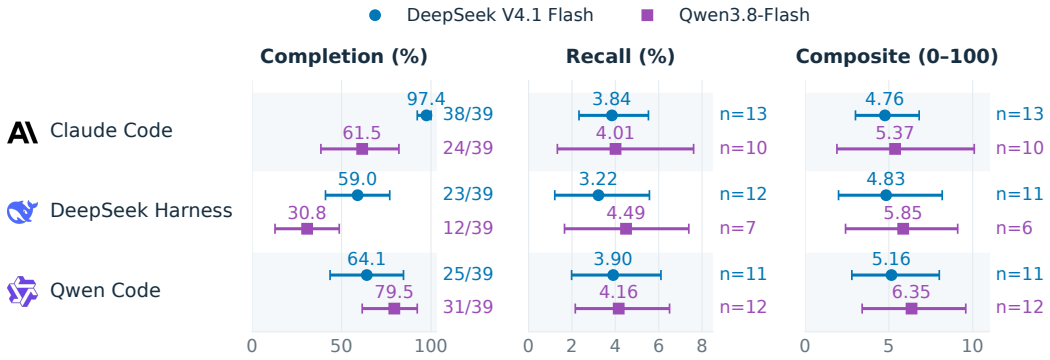


Figure 3: DeepSeek V4.1 Flash and Qwen3.8-Flash across three shared runtimes on Original novels. Aggregation, intervals, and support annotations follow Figure 2.

predictions. Qwen3.8-Flash uses CC, DSH, and QWC under the shared DeepSeek/Qwen context, output, and request allowances; GPT-5.6 with Codex uses native budgets and `xhigh` reasoning. Original/INJ pairs, budget retries, and construction-linked outputs remain separate cohorts. Budgets and service/output processing appear in Appendix I; Table 4 inventories qualified attempts, including failures.

Aggregation and uncertainty. Completion rate uses all qualified attempts. We average defined scores from successful attempts within each novel, then give each contributing novel equal weight. Valid empty predictions retain zero case recall; failures have no evaluation score. We bootstrap novels to compute intervals; paired comparisons use common support. For discovery unions, we deduplicate reference IDs per novel, then average cumulative case recall across novels. Support sensitivity and resource accounting appear in Appendices I and E.14.

5.2 Completion and Model–Runtime Adaptation

Completion rate shows the clearest model–runtime differences (Figures 2 and 3). Under DeepSeek, Claude Code leads, closely followed by Codex Harness; mini-SWE-agent completes least reliably. Successful predictions have similar mean case recall and Composite with substantial variability (Table 5).

Claude Code shows strong cross-model adaptability, while native runtimes favor their own models: DeepSeek outperforms Qwen in completion in DeepSeek Harness, and Qwen outperforms DeepSeek

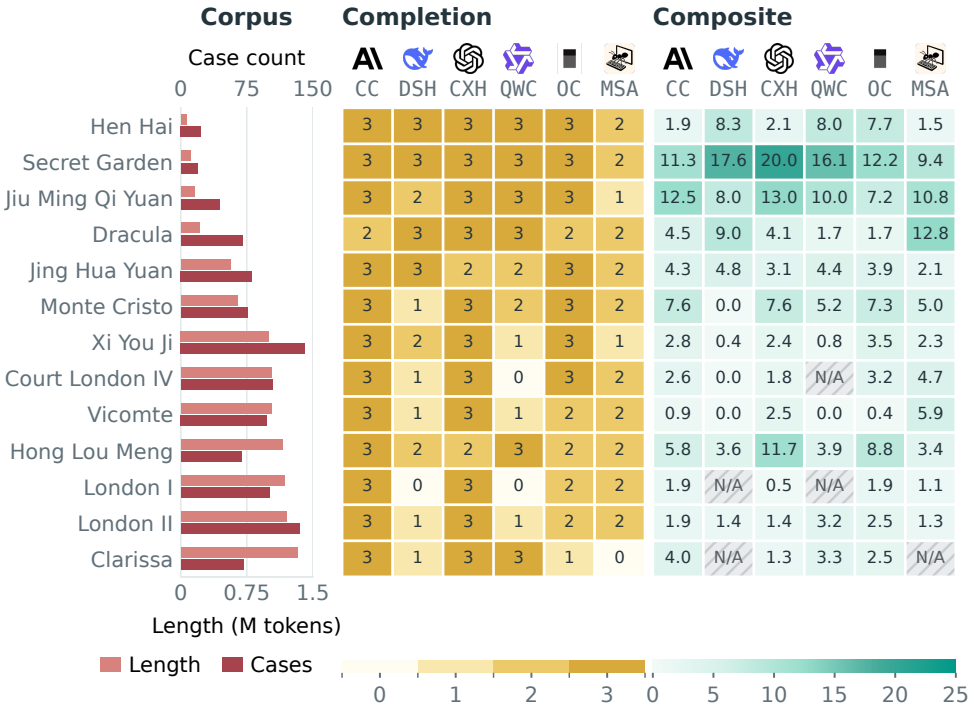


Figure 4: DeepSeek V4.1 Flash on Original novels, ordered by text length. Left: length (bottom axis, million `cl100k_base` tokens) and reference cases (top axis). Center: successful attempts out of three. Right: mean defined Composite across attempts; hatched N/A indicates no defined score.

in Qwen Code. Claude Code and Qwen Code reverse completion order between models, making runtime selection configuration-dependent.

Across the shared runtimes, Qwen has slightly higher mean case recall and Composite than DeepSeek, but a higher completion rate only in Qwen Code. Under DeepSeek, the configuration with the highest completion rate does not lead in Composite. Even reliable configurations have low case recall, so reducing failed attempts alone does not ensure discovery. These findings persist after reference corrections, rematching, and source expansion. Initial-human and shared cases remain more readily matched than GPT-only cases (Appendices E.6 and E.7).

5.3 Novel-Level Variation behind Aggregate Performance

Figure 4 aligns length, reference case count, completion, and Composite. Longer novels generally have more cases but fewer per unit of text. Counts, obtained under the stopping protocol, do not grow proportionally with length: *Xi You Ji* has the most cases, whereas the longest novel, *Clarissa*, has fewer than several shorter works. Length and case count distinguish reading volume from discovery scope.

Completion declines with length differently across runtimes. Claude Code and Codex Harness remain reliable throughout, whereas DeepSeek Harness and OpenCode complete less reliably on longer novels. Qwen Code completes every attempt on *Clarissa* but none on the shorter, case-richer *London I* and *Court London IV*. Mini-SWE-agent fails even on short novels and always on the longest. There is no shared length cutoff across configurations.

Composite is highest on *The Secret Garden* for several runtimes and generally lower on long novels, yet the shortest, *Hen Hai*, is not the easiest. Rankings also vary by novel: mini-SWE-agent leads Composite on *Dracula* and *Vicomte* despite poor overall completion. Aggregate rankings conceal shared difficult novels and runtime-specific differences.

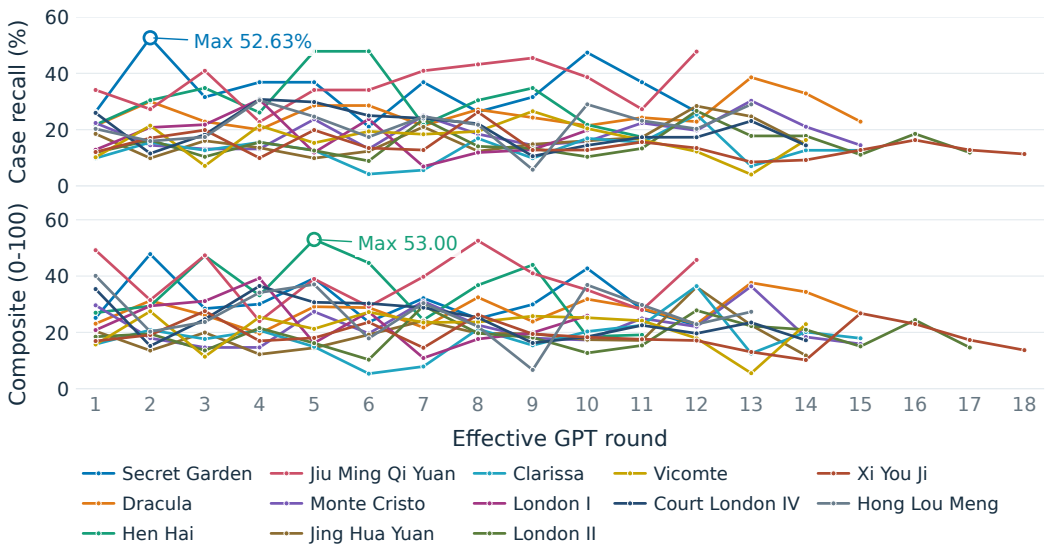


Figure 5: GPT-5.6 with Codex: single-attempt case recall (top) and Composite (bottom) against fixed final references. Colors identify the 13 novels. Each line follows a novel's effective GPT rounds during reference construction and ends at its last available prediction; outlined points mark the global maxima.

Length and case count are broadly negatively associated with completion and, more consistently, Composite. Claude Code and Codex Harness complete all attempts on case-rich *Xi You Ji* and *London II*, yet score poorly; *Clarissa*'s smaller reference does not make its long text easy. Larger references demand broader discovery, while long text remains difficult even with fewer cases. Unmatched references remain in Composite's case-quality denominator, so low case recall directly reduces it.

Qwen–Claude Code completes the same number of attempts in each round, but the successful novels change (Appendix E.3). Matched-label accuracy uses only matched predicted cases and is affected by class imbalance; high accuracy cannot explain away low case recall (Appendices E.5 and F.1).

5.4 Single Attempts Discover Limited, Changing Case Sets

Formal GPT–Codex rounds have similar mean case recall, yet every later round adds cases for every novel and can miss earlier discoveries on unchanged text (Table 7; Appendix E.8). Cumulative case recall grows through complementary subsets: a stable average score does not imply a stable discovery set.

The longer construction-linked trajectories in Figure 5 show case recall and Composite rising and falling, with peaks at different rounds across novels and later predictions often below earlier ones. All rounds are scored against the same final reference constructed under the stopping protocol; fluctuations therefore reflect changes in predictions. These construction rounds remain separate from formal rounds (Appendix E.12). Additional attempts expand coverage while individual predictions remain incomplete, so discovery should track case identities alongside scores.

Across DeepSeek's runtimes, most cases in the discovery union occur under only one runtime. Repeated attempts account for most cumulative case recall gains; mixing runtimes adds less at the same attempt count (Appendix E.13). Merging the three attempts' submitted cases also raises mean Composite for every DeepSeek runtime (Figure 18).

Output volume falls well short of the reference case count even if every predicted case is valid, distinct, and matched; low case recall is not solely a matching artifact. Human audits likewise find more distinct valid cases across attempts than any single prediction contains (Appendix E.17.1). Across the tested configurations, combining complementary discoveries expands coverage and raises mean Composite, while individual attempts return limited, changing subsets.

Later formal GPT rounds require more input tokens per new match; DeepSeek/Qwen combinations show diminishing additions, partly offset by runtime diversity. Failures and successful attempts without new matches both consume resources. High completion and increasing cumulative case recall at low input cost are different objectives (Appendix E.14).

5.5 Evidence Quality and the Limits of Case Matching

Human review confirms case identity, qualification, endpoint support, and explanatory sufficiency for some predicted cases that remain unmatched. Matching gates reject these cases; assignment competition does not explain their rejection. Conversely, identity confirmation is more common than support for every evidence role (Table 41; Appendix E.17.1). Identity and evidence sufficiency are distinct judgments.

Endpoint constraints block incorrect correspondences but reject human-supported explanations with weak endpoint wording alignment. Relaxing matching raises Composite; endpoint constraints lower it (Table 39). Under fixed rules, alternative human-supported passages substantially change some predictions' Composite while leaving others unchanged. Scores thus depend on evidence expression and should be read alongside human judgments (Appendices F.2 and F.7).

Human review endorses sampled reference routes but finds residual explanations in most targeted edited examples. Complete explained-to-unresolved transitions are uncommon; an explained prediction after editing need not be wrong. Original/INJ results therefore describe edited-text behavior with retained explanations, rather than complete removal of explanatory information (Appendix E.8).

6 Discussion and Limitations

From existing novels to generated narratives. Consistency evaluation for generated stories and scripts motivates our pure-text starting point: existing novels and paired edits. Future work can independently adjudicate references for generated novels to enable direct evaluation and compare naturally occurring errors with INJ's targeted explanation deletions. Short dramas and films require evidence across dialogue, images, and sound.

Case discovery and explanatory sufficiency. We extend complete-list discovery (Singhania et al., 2025) to case qualification and explanatory evidence. Role-structured evidence makes case identity, evidence sufficiency, and case matching separately inspectable. Context-dependent interpretation and missed matches call for independent review with explicit context conditions.

Measurement and scope. The 13 evaluated novels were also used to construct the reference and may occur in pretraining; 11 extended novels remain unevaluated. GPT-only discoveries dominate the principal reference. Low case recall persists after source expansion and corrections, without establishing source neutrality or completeness. Scores combine discovery and evidence expression with coupled label/evidence credit and chosen weights. Explanation-Path F1 associations validate neither new matches nor system rankings. Reviewer exposure and adjudication conditions vary; sampling intervals do not account for reviewer error. Residual explanations complicate interpretation of INJ label switches. Runtime/service conditions, output processing, and successful-novel selection affect comparisons; missing usage prevents monetary-cost comparisons. Held-out novels, equal-opportunity construction, and specialized reading or memory baselines remain untested.

7 Conclusion

NovelAudit benchmarks case discovery and explanatory sufficiency across complete novels. Its reference retains explained reversals and unresolved inconsistencies; its protocol separates task completion, case discovery, and evidence quality. Runtime choice and model–runtime adaptation primarily affect completion rate. Across formal DeepSeek experiments and GPT discovery rounds, each attempt finds only a limited, changing subset of a novel's cases. Combining complementary discoveries across rounds expands coverage under the tested configurations, with rising input cost per new match. Human audits link Explanation-Path F1 to explanatory sufficiency and reveal missed matches, incomplete evidence, and residual explanations after editing.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, retrieve and identify related work, support research ideation and execution, and draft sections of the paper. Research assistance included refining the methodology and experimental design, implementing and debugging methods and supporting software, aggregating and analyzing experimental outputs, interpreting results, and preparing scientific figures through analysis and plotting code. AI tools also helped check literature and references, revise the manuscript and LaTeX source, review consistency between claims and experimental evidence, and check presentation and formatting. Automated feedback from ICLR's Paper Assistance Tool (PAT) was also used to identify issues in the draft; selected revisions followed checks against the underlying records.

The reported result figures were rendered programmatically from recorded experimental outputs. An AI image-generation tool was used to create and refine the conceptual task-overview and reference-construction illustrations. Models proposed candidate cases from the human-authored novels during reference construction, as described in the methods. AI tools also helped organize review materials and locate passages in the source texts; human reviewers made the substantive judgments themselves. The authors take responsibility for the research decisions, reference annotations, analyses, code, figures, claims, and final manuscript.

ETHICS STATEMENT

The intended use is evidence-based assistance for readers and authors, not an automatic judgment of literary merit or author ability. A predicted unresolved case is not proof of an authorial error. We do not produce a novel-level consistency score. The collection uses historical literary texts. Redistribution permissions are edition-, translation-, and source-specific; benchmark availability does not imply a blanket license for all underlying texts. Explanation-deleted copies are experimental interventions and must be distinguished from the original works.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the task, reference representation, evaluation protocol, and experimental conditions. The appendices provide the public output schema (Appendix A.2), human-review protocols and their limitations (Appendices B and F), and reference construction, scoring, runtime, and prompt details (Appendices G–J).

REFERENCES

Kabir Ahuja, Melanie Sclar, and Yulia Tsvetkov. Finding flawed fictions: Evaluating complex reasoning in language models via plot hole detection. In *Second Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.11900.

Aayush Aluru, Chloe Ho, Muhammad Hammouri, Kerry Luo, Myra Malik, Ryan Lagasse, Arjun Bahuguna, and Vasu Sharma. From personas to plot: Character-grounded multi-agent story generation for long-form narratives. *arXiv preprint arXiv:2607.00918*, 2026. URL https://arxiv.org/abs/2607.00918.

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

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongWriter: Unleashing 10,000+ word generation from long context LLMs. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kQ5s9Yh0WI.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp.

9294–9307, 2020. doi: 10.18653/v1/2020.emnlp-main.747. URL https://aclanthology.org/2020.emnlp-main.747/.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. StoryER: Automatic story evaluation via ranking, rating and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1739–1753, 2022. doi: 10.18653/v1/2022.emnlp-main.114. URL https://aclanthology.org/2022.emnlp-main.114/.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5794–5836, 2022. URL https://aclanthology.org/2022.coling-1.509/.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142, 2024. doi: 10.1162/tacl_a_00689. URL https://aclanthology.org/2024.tacl-1.62/.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. LitBench: A benchmark and dataset for reliable evaluation of creative writing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7740–7755, 2026. URL https://aclanthology.org/2026.eacl-long.362/.

Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. Creative writing with an AI-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030*, 2022. URL https://arxiv.org/abs/2211.05030.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17048–17085, 2024. doi: 10.18653/v1/2024.emnlp-main.948. URL https://aclanthology.org/2024.emnlp-main.948/.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.01261.

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

Mina Lee, Percy Liang, and Qian Yang. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022. doi: 10.1145/3491102.3502030. URL https://doi.org/10.1145/3491102.3502030.

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 *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8400–8428, 2026. doi: 10.18653/v1/2026.findings-acl.410. URL https://aclanthology.org/2026.findings-acl.410/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Charles Lovering, Seth Ebner, Brandon Smock, Michael Krumdick, Muhammad Saad Rabbani, Ahmed Muhammad, Varshini Reddy, and Chris Tanner. On finding inconsistencies in documents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33523–33564, 2026. doi: 10.18653/v1/2026.findings-acl.1675. URL https://aclanthology.org/2026.findings-acl.1675/.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. doi: 10.1145/3544548.3581225. URL https://doi.org/10.1145/3544548.3581225.

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

Ansong Ni, Matt Gardner, and Pradeep Dasigi. Mitigating false-negative contexts in multi-document question answering with retrieval marginalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6149–6161, 2021. doi: 10.18653/v1/2021.emnlp-main.497. URL https://aclanthology.org/2021.emnlp-main.497/.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019. doi: 10.1162/tacl_a_00293. URL https://aclanthology.org/Q19-1043/.

Sneha Singhania, Simon Razniewski, and Gerhard Weikum. Recall them all: Long list generation from long novels. In *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities*, pp. 133–142, 2025. URL https://aclanthology.org/2025.lm4dh-1.13/.

Nelvin Tan, Yaowen Zhang, James Asikin Cheung, Fusheng Liu, Yu-Ching Shih, and Dong Yang. Improved evidence extraction and metrics for document inconsistency detection with LLMs. *arXiv preprint arXiv:2601.02627*, 2026. URL https://arxiv.org/abs/2601.02627v2.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Runchu Tian, Yanghao Li, Yuepeng Fu, Siyang Deng, Qinyu Luo, Cheng Qian, Shuo Wang, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Huadong Wang, and Xiaojiang Liu. Distance between relevant information pieces causes bias in long-context LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 521–533, 2025. doi: 10.18653/v1/2025.findings-acl.28. URL https://aclanthology.org/2025.findings-acl.28/.

Ellen M. Voorhees, Ian Soboroff, and Jimmy Lin. Can old TREC collections reliably evaluate modern neural retrieval models? *arXiv preprint arXiv:2201.11086*, 2022. URL https://arxiv.org/abs/2201.11086.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent CoT planning. *arXiv preprint arXiv:2503.07314*, 2025. URL https://arxiv.org/abs/2503.07314.

Haotian Xia, Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. StoryWriter: A multi-agent framework for long story generation. *arXiv preprint arXiv:2506.16445*, 2025. URL https://arxiv.org/abs/2506.16445.

Dingyi Yang and Qin Jin. What matters in evaluating book-length stories? a systematic study of long story evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16375–16398, 2025. doi: 10.18653/v1/2025.acl-long.799. URL https://aclanthology.org/2025.acl-long.799/.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4393–4479, 2022. doi: 10.18653/v1/2022.emnlp-main.296. URL https://aclanthology.org/2022.emnlp-main.296/.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3378–3465, 2023. doi: 10.18653/v1/2023.acl-long.190. URL https://aclanthology.org/2023.acl-long.190/.

Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Daniel Campos. On the reliability of test collections for evaluating systems of different types. *arXiv preprint arXiv:2004.13486*, 2020. URL https://arxiv.org/abs/2004.13486.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. MovieDreamer: Hierarchical generation for coherent long visual sequences. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.16655.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-judge: Evaluate agents with agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80569–80611, 2025. URL https://proceedings.mlr.press/v267/zhuge25a.html.

A CASE EXAMPLE AND PUBLIC CONTRACT

In appendix tables, bold marks numerical maxima, including ties, within comparable conditions. Alternative settings of the same quantity are compared within each row; other metrics are compared down columns. Slash-separated counts are compared separately; totals, identifiers, and interval endpoints are not ranked.

A.1 FULL PASSAGES FOR THE INTRODUCTORY CASE

This illustrative *Dracula* reversal has the label EXPLAINED under the reader-facing tests. The excerpts below reproduce the endpoint and explanatory passages from the source text used in the benchmark.

Before, Chapter 2.

> Within, stood a tall old man, clean shaven save for a long white moustache, and clad in black from head to foot, without a single speck of colour about him anywhere.

After, Chapter 4.

> There lay the Count, but looking as if his youth had been half renewed, for the white hair and moustache were changed to dark iron-grey; the cheeks were fuller, and the white skin seemed ruby-red underneath; the mouth was redder than ever, for on the lips were gouts of fresh blood, which trickled from the corners of the mouth and ran over the chin and neck. Even the deep, burning eyes seemed set amongst swollen flesh, for the lids and pouches underneath were bloated. It seemed as if the whole awful creature were simply gorged with blood.

Explanation, Chapter 18 (original spelling and grammar preserved).

> The vampire live on, and cannot die by mere passing of the time; he can flourish when that he can fatten on the blood of the living. Even more, we have seen amongst us that he can even grow younger; that his vital faculties grow strenuous, and seem as though they refresh themselves when his special pabulum is plenty. But he cannot flourish without this diet; he eat not as others.

The subject and state dimension are unchanged; old versus rejuvenated appearance is a genuine reversal, not ordinary aging. Without the explanation, a reader would reasonably ask why it occurred. The after passage observes feeding, and the later explanation establishes its rejuvenating effect. Thus the candidate qualifies but is explained. The example also shows why distance must include the complete explanation rather than only the two endpoint chapters.

A.2 ACTUAL PUBLIC OUTPUT

The minimal state proposition defines the case conceptually. In the public prediction, subject names, endpoint quotations, and the reason express this proposition; it is not a separate field. The required result has a single top-level `cases` array. The following is a schematic shape, not a valid prediction containing real evidence:

```
{
  "cases": [{
    "targets": ["subject name"],
    "before": {"quote": "continuous source passage"},
    "after": {"quote": "continuous source passage"},
    "label": "explained",
    "bridge_evidence": [{"quote": "continuous source passage"}],
    "reason": "whether and why the reversal is explained"
  }]
}
```

The other label is `unresolved`; its `bridge_evidence` field must be an empty array. Targets are natural-language names. These fields are exact: a case contains only `targets`, `label`, `before`,

`after`, `bridge_evidence`, and `reason`. Each endpoint and each explanatory evidence object contains only `quote`. There is no optional public state-dimension field, chapter/source locator, reference case ID, or internal memory record. Extra fields violate the final public output contract. An agent runtime may use source locators internally, but must resolve them into the public quote fields without inventing or repairing semantic content. Usage belongs to `run.json`, not `result.json`; evaluation belongs to `scores.json`. Missing historical usage is unknown, not zero.

The reference is richer than the prediction. It records subject aliases, state claims, alternative anchors with chapter locations, explanatory evidence requirements and alternative explanatory routes, decision, label, and adjudication notes. Only included cases enter the evaluation reference. Excluded working records are retained during discovery but do not contribute to metrics; ambiguous and pending records are also outside the evaluation reference. Pending labels are null, and pending content prevents completion of reference construction.

A.3 Deterministic Output Projection and Cohort Boundaries

Later discovery and the secondary GPT study use a shared deterministic projection from the native final message to the public schema. It extracts a unique JSON object or array, optionally enclosed in a Markdown fence, and permits bounded repairs of trailing commas and evidence objects missing their array wrapper. Recognized aliases map `subjects`/`target` to `targets`, `bridge` to `bridge_evidence`, and `findings` to `cases`. Evidence strings, single objects, and lists are normalized into the required role-specific shape; non-public fields are removed, and label whitespace and case are normalized. A missing reason receives fixed non-scoring text.

For endpoint arrays, the first nonempty quote in original order is retained; `bridge_evidence` arrays retain their valid quotes. Repeated identical target names are deduplicated. A case lacking recoverable targets, either endpoint, or a valid label is dropped with its original index recorded, while other recoverable cases remain. If a nonempty candidate collection yields no recoverable cases, projection fails rather than fabricating a successful empty prediction. The projection invokes no model and does not rewrite quotes, infer labels, merge cases, or remove duplicate cases. The raw final message and projection diagnostics distinguish submitted content from projected output. GPT native discovery additionally constrains its native final output with a JSON schema.

This output condition belongs to later discovery cohorts. The fixed principal Original cohort predates it and retains its original completion outcomes; a later projection cannot replace a failed attempt in that cohort. Structurally recoverable candidate availability and semantic correctness remain separate measurements.

B Challenge-Pair Human Review Protocol

The validation study asks whether independent human judgments agree with automatic case identity and evidence evaluation. It does not revise reference labels, tune thresholds to favor a system, or establish a novel-level score. All pairs were checked against the frozen principal references and official scorer before final reporting. The sample draws on eligible historical external-runtime predictions, supplemented by public predictions collected under prediction isolation where too few near-threshold pairs were available. These validation inputs are separate from the principal runtime comparison.

B.1 Sampling Allocation and Blinding

Table 1 describes a stratified challenge sample of 96 distinct prediction–reference pairs across all 13 principal novels. These pairs involve 81 distinct reference cases, so they are not 96 independent reference cases. The four strata are mutually exclusive. The 32 selected matches also receive evidence-role and explanatory sufficiency review. Sampling within each language covers multiple runtimes, both labels, the available distance bands, and at least three novels where feasible. Selected matches span score margins; near-threshold pairs cover case-score, target, and endpoint guards. Hard negatives distinguish same-subject but different-proposition cases and nearby but different reversals. Historical rejection labels alone do not establish a clean negative control. Shortages are reported rather than filled with duplicates or unrelated easy examples.

Table 1: Achieved stratified challenge sample for blinded identity and evidence review. Every pair received two independent blinded judgments.

Sampling stratum	EN	ZH
Scorer-selected matches, also reviewed for evidence	**18**	**14**
Near-threshold unmatched pairs	**18**	**14**
Same-case controls with different wording	9	7
Hard negatives	9	7
Total	54	42

Reviewers see the language, novel, reference claims and admissible evidence routes, predicted targets and role-specific quotes, neutral source locations, adjacent context, and access to the full novel. They know which side is the reference, preserving its richer evidence structure. They do not see model or runtime identity, attempt information, cost or date, either label, the prediction's reason, reference IDs or decisions, source membership, or the scorer's assignment, scores, margins, and ambiguity flags.

B.2 Judgments, Independence, and Analysis

Case identity is judged as same, different, or ambiguous. Each endpoint and explanatory evidence quote receives a role judgment of correct, partial, wrong, or ambiguous; the combined explanatory evidence receives sufficient, partial, insufficient, or ambiguous. Empty evidence is not automatically sufficient. These judgments concern identity and evidence, not a new adjudication of the hidden predicted label.

Two reviewers independently judged all 96 pairs, yielding 192 judgments. Neither reviewer's answers were shown to the other before submission. Case identity agrees on 90/96 pairs (Cohen's $\kappa = 0.839$): 48/54 in English and 42/42 in Chinese. Agreement by stratum is 32/32 for selected matches, 27/32 for near-threshold unmatched pairs, 15/16 for same-case controls, and 16/16 for hard negatives. All six case-identity disagreements retain both original judgments; the scorer does not adjudicate them.

For the 32 selected pairs, before-role judgments agree on 32/32, after-role judgments on 29/32, and explanatory sufficiency on 24/32. Individual explanatory-role judgments agree on 33/41. These denominators distinguish pair-level and role-level agreement; the case-identity kappa does not measure evidence sufficiency.

A suspected reference error is recorded separately; the validation exercise does not alter the frozen reference or erase unrelated items. Three reference issues raised by one reviewer were not confirmed by the other and did not trigger reference changes. All 96 pairs remain in the case-identity agreement denominator. We report the observed role and sufficiency judgments without selecting a new binary threshold to manufacture agreement. A calibrated population estimate of scorer error requires further sampling. Because sampling deliberately emphasizes challenging pairs, a pooled agreement proportion cannot estimate population prevalence or open-ended prediction precision. The latter requires a separate independent audit of unmatched predictions, without adding accepted discoveries to the same evaluation denominator.

Composite is supplementary in this study and does not support a principal ordering. We report the official parameters without choosing them after seeing a ranking. The common-novel sensitivity analysis concerns support selection and does not establish scorer-parameter robustness.

C Dataset Composition

Principal collection: 13 original novels. The panel comprises *The Secret Garden*; *Dracula*; *Hen Hai*; *Jiu Ming Qi Yuan*; *The Count of Monte Cristo*; *Jing Hua Yuan*; *Clarissa*; *The Mysteries of London*, First and Second Series as separate novels; *The Vicomte de Bragelonne*; *The Mysteries of the Court of London*, Fourth Series; *Xi You Ji*; and the 120-chapter *Hong Lou Meng*. These texts form the Original runtime comparison and the source panel for the 13 paired explanation-deleted copies.

Extended collection: 11 original novels. The collection comprises *Wuthering Heights*; *Middlemarch*; *The History of Sir Charles Grandison*; *Varney the Vampyre*; the 70-chapter *Shui Hu Zhuan*; *Sui Tang Yan Yi*; *The Mysteries of the Court of London*, Second and Third Series as separate novels; *Jiu Wei Gui*; *Jin Ping Mei*; and *Peng Gong An*. The collection uses the same semantic case and evidence definitions, with a flexible, time-bounded discovery depth. The principal saturation-style stopping rule is not imposed on these novels. All 11 references are frozen; their 806 included cases comprise 747 explained and 59 unresolved cases. These texts are outside the reported runtime evaluations.

Counting and comparison units. The construction collection contains 24 originals, of which 13 form the frozen, evaluated principal panel. Edited variants remain paired conditions within the 13 principal novels; they do not increase the independent-novel count. Principal construction, extended construction, Original evaluation, and the intervention study retain distinct denominators. Table 2 reports the extended reference counts; principal statistics appear in Appendix D.

Table 2: Frozen extended references: included cases (N), explained (E), unresolved (U), and excluded records (X). Reference construction is complete for each novel. The 822 stored records comprise 806 included cases and 16 exclusions; runtime performance is not reported for this collection. Court II and III denote the Second and Third Series of *The Mysteries of the Court of London*.

Novel	Language	N	E	U	X
Wuthering Heights	EN	52	52	0	0
Middlemarch	EN	47	47	0	**5**
Sir Charles Grandison	EN	66	63	3	1
Varney the Vampyre	EN	79	69	10	1
Shui Hu Zhuan (70 chapters)	ZH	80	78	2	0
Sui Tang Yan Yi	ZH	**108**	**105**	3	2
Court II	EN	94	91	3	3
Court III	EN	83	82	1	3
Jiu Wei Gui	ZH	36	33	3	0
Jin Ping Mei	ZH	71	58	13	0
Peng Gong An	ZH	90	69	**21**	1
Total	6 EN / 5 ZH	806	747	59	16

D PRINCIPAL DATASET STATISTICS

Table 3 reports the revised principal reference. The 13 texts contain 9.67 million canonical `cl100k_base` tokens, with seven exceeding one million tokens. There are 1,032 included cases: 990 explained and 42 unresolved. Thirteen excluded records and one ambiguous record are outside the scoring denominator. Of the 467 originally edited targets, 466 remain eligible; edited references contain 524 explained and 508 unresolved cases. The edited novel texts are unchanged. Cases sharing a novel are not independent samples. Eleven extended originals broaden construction scope, but do not contribute cases or attempts to these principal statistics.

Table 3: Principal corpus and reference. Tokens are in thousands under a common analysis tokenizer, not provider billing; the total uses unrounded token counts. N counts included cases, U unresolved cases, and T selected intervention targets. Distance counts use minimum complete-span normalized distance: local/medium/long (L/M/H).

Novel	Lang.	Tokens (k)	N	U	T	L/M/H
Secret Garden	EN	107.5	19	0	9	10/2/7
Dracula	EN	213.2	70	5	30	48/16/6
Hen Hai	ZH	63.4	23	1	10	13/3/7
Jiu Ming Qi Yuan	ZH	158.9	44	2	19	33/8/3
Monte Cristo	EN	644.2	76	7	31	44/15/17
Jing Hua Yuan	ZH	562.6	81	4	35	64/5/12
Clarissa	EN	**1323.1**	71	1	34	36/20/15
London I	EN	1180.2	101	1	48	58/23/20
London II	EN	1200.1	135	0	**66**	89/**25**/21
Vicomte	EN	1029.4	98	0	49	64/18/16
Court London IV	EN	1028.1	104	2	50	66/16/22
Xi You Ji	ZH	1002.9	**141**	**15**	55	**104**/14/**23**
Hong Lou Meng	ZH	1154.8	69	4	30	48/14/7
Total	8 EN / 5 ZH	9668.6	1032	42	466	677/179/176

E ADDITIONAL EMPIRICAL RESULTS

Table 4: Formal qualified attempts, including eligible failures. Historical construction outputs are separate; budget retries retain the original attempts.

Condition	Qualified	Successful	Failed
DeepSeek Original	**234**	**177**	57
DeepSeek INJ	**234**	174	**60**
Qwen Original	117	67	50
GPT–Codex Original	40	39	1
GPT–Codex INJ	39	39	0
Main experiment panel	664	496	168
Targeted budget retries	41	23	18
Total	705	519	186

E.1 COVERAGE LIMITS AND COMPLEMENTARY DISCOVERY

The evidence for limited discovery does not rest on low automatic scores alone. We first bound what the submitted case counts could cover under perfect matching, then report gains from repeated formal predictions and complementary cases confirmed by human review. These checks address coverage of the fixed reference; the validity of exact Composite values is examined separately in Appendix F.

Output-count bound. Under the fixed one-to-one protocol, each attempt's recall is bounded by $\min(|\widehat{\mathcal{C}}|, |G|)/|G|$. Applying the original within-novel then between-novel aggregation gives Table 6.

Table 5: DeepSeek Original: 234 qualified attempts (177 successful, 57 failed). Metrics are book-macro percentages except Composite (0–100); recall has 95% book-bootstrap intervals. B counts novels with defined case recall. Precision/Composite use 11 novels/22 attempts for DSH and 13/36 for CXH; other supports match recall. Each runtime has 39 attempts.

Runtime	Success	Completion	B	Recall [95% CI]	Precision	Q	Composite
Claude Code	**38/39**	**97.4**	**13**	3.84 [2.32, 5.53]	**37.8**	2.75	4.76
DeepSeek Harness	23/**39**	59.0	12	3.22 [1.21, 5.57]	21.4	2.32	4.83
Codex Harness	37/**39**	94.9	**13**	3.82 [1.75, 6.45]	32.9	**2.86**	**5.50**
Qwen Code	25/**39**	64.1	11	**3.90** [1.98, 6.10]	35.7	**2.86**	5.16
OpenCode	32/**39**	82.1	**13**	3.52 [2.28, 4.90]	25.2	2.55	4.82
mini-SWE-agent	22/**39**	56.4	12	3.80 [2.06, 5.83]	32.1	2.63	5.04

It quantifies the upper bound on case recall from the existing predicted case counts, without assuming that unmatched outputs are invalid or extrapolating reviewed samples into corrected scores.

Table 6: Reported-case upper bounds under one-to-one matching, on the same successful supports as the main table. Both case recall columns are book-macro percentages.

Runtime	Novels/attempts	Revised recall	Output-count ceiling
CC	**13/38**	3.84	9.49
CXH	**13**/37	3.82	10.85
DSH	12/23	3.22	10.97
MSA	12/22	3.80	11.34
OC	**13**/32	3.52	**14.06**
QWC	11/25	**3.90**	10.32

This bound grants every submitted case validity, a distinct identity, and a match whenever the reference has room for one. It uses no lexical gate or explanation score. Even this favorable assignment leaves single-attempt coverage limited in every DeepSeek runtime. The bound concerns the case units and finite denominator of this benchmark; it is neither an estimate of corrected recall nor validation of every reference case. Completion in Table 4 is also measured without case matching.

Formal repeated discovery. The GPT Original cohort contains 40 qualified attempts: 39 successful predictions across three rounds and 13 novels, plus one recorded failed attempt. Table 7 uses these formal predictions, separately from the construction-linked histories. For each novel, cumulative recall counts distinct matched reference IDs across rounds, then averages equally over novels. Later rounds add cases in every novel while sometimes missing earlier discoveries on unchanged text (Appendix E.8).

Table 7: GPT formal Original discovery across three rounds. Single-attempt and cumulative case recall are book-macro percentages over 13 novels. Composite (0–100) scores each round's predictions, with 95% book-bootstrap intervals. Counts are distinct novel–reference IDs; these union counts describe matched IDs rather than a merged prediction.

Round	Single (%)	New cases	Union cases	Union (%)	Composite [95% CI]
R1	17.06	**146**	146	17.06	18.36 [13.88, 23.30]
R2	**18.37**	98	244	26.97	19.23 [15.52, 23.19]
R3	18.24	61	**305**	**32.79**	**20.74** [15.84, 26.03]

The union remains matcher-dependent: it establishes complementarity among credited cases, not a human-corrected discovery rate. The DeepSeek merged-prediction analysis provides a separate check on the combined submitted evidence under the fixed scorer (Figure 18); it does not replace this union of matched IDs.

Human-confirmed complementarity. The matched- and unmatched-output audits (Appendix E.17.1) provide a separate check on complementary discovery. The 189 final correct pairs and three valid corrected identities join R1's 66 agreed identities into 258 distinct attempt–case groups. After novel–case deduplication and merging DR027/DR058, the accepted sample contributes 111 case groups, the rejected sample 48, and their union 134; 23 occur only on the audited rejected side. These are verified sample sets, not design-weighted estimates of the population's unique cases. Jiu Ming Qi Yuan has 13 pooled verified groups versus at most 12 predicted cases in any single attempt; London II has ten versus at most nine. Thus no single existing prediction can contain either entire verified pool, even granting validity to its unreviewed outputs.

Resource interpretation. Repeated attempts expand coverage in the tested configurations, with increasing input per new formal GPT match (Appendix E.14). The equal-attempt analysis compares repeated and mixed native runtimes; it does not compare repeated attempts with a specialized single-attempt reading, retrieval, or memory strategy at the same budget. The observed gains therefore describe a demonstrated use of additional attempts, without establishing that this use is optimal.

E.2 Novel-Level Support and Failure Modes

Figure 4 shows every principal novel–runtime condition. Successful attempts do not guarantee nonempty predictions or a defined precision denominator. The valid empty DSH output on *Clarissa* remains a zero-recall observation, so DSH recall retains 12 novels/23 attempts. Precision and Composite exclude this attempt and use 11 novels/22 attempts. CXH has one valid empty prediction but other successful attempts on that novel, retaining 13 novels/36 attempts for precision and Composite. Zero recall from a valid output remains an observation; an absent successful output has no conditional evaluation score. This distinction prevents failures from disappearing when comparing completion rates, while avoiding a semantic interpretation of execution errors.

The 57 failures comprise 25 public-output format failures, nine request-budget failures, six missing final results, five runtime out-of-memory failures, five runtime exits without a determined cause, three native action-format failures, two transport failures, one event-stream failure, and one context rejection. Categories follow the first supported causal error, rather than reclassifying unexplained exits from their request counts. Principal attempts retain their collection-time output extraction condition; later projection is not used to retroactively convert these failures into successes.

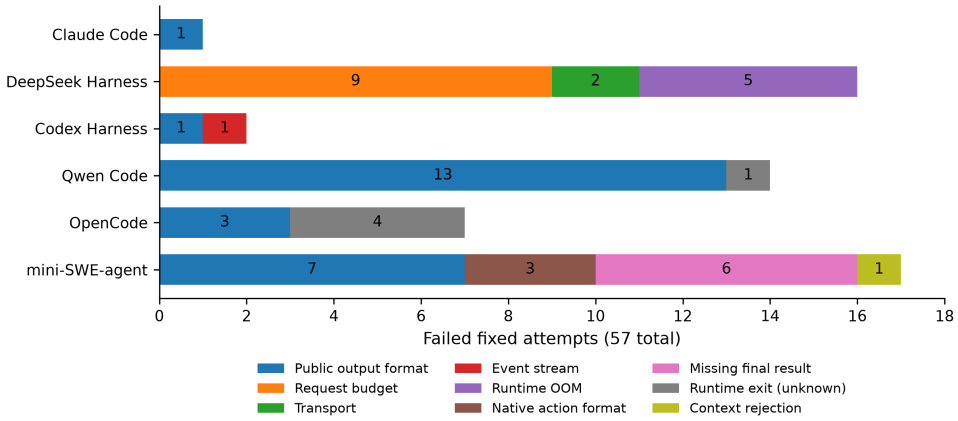


Figure 6: First-cause categories for all 57 principal failures. A runtime can fail to produce a valid prediction without revealing whether its narrative reasoning was correct.

E.3 Three-Round Qwen Runtime Robustness Check

Qwen Original contains 117 qualified attempts: 67 successful and 50 failed, across CC, DSH, and QWC on 13 novels in three rounds. The finalized R1 contributes 18 successes and 21 failures, R2 contributes 26 and 13, and R3 contributes 23 and 16. Ten earlier R1 success cells lacking uniform raw-

output records were replaced unconditionally under the fixed recollection decision; four replacement failures remain failures. Thus old R1 means and the former 47-success two-round aggregate are superseded.

R2 has 51 physical provider-invoking executions, including 12 explicit account-quota interruptions followed by continuations; only its 39 finalized cells enter this matrix. The separate successful London II follow-up does not replace its R1 failure. R3 retains five QWC attempts that produced valid predictions before a later temporary-directory cleanup error. All 67 successful predictions have revised-reference scores; old scores remain archived. Two valid empty R2 predictions, CC–Xi You Ji and DSH–London II, have zero recall but undefined precision and Composite.

Table 8: Final Qwen completion by novel and round. S: valid public prediction; F: qualified failure. Each round has 39 fixed attempts.

	R1			R2			R3		
Novel	CC	DSH	QWC	CC	DSH	QWC	CC	DSH	QWC
The Secret Garden	S	F	S	S	S	F	S	S	S
Dracula	F	F	F	S	F	S	F	F	S
Hen Hai	S	S	S	S	S	S	S	S	S
Jiu Ming Qi Yuan	S	F	S	S	S	S	S	F	S
The Count of Monte Cristo	F	F	F	S	S	S	S	F	S
Jing Hua Yuan	S	F	S	F	F	S	S	F	S
Clarissa	S	F	S	S	F	S	F	F	S
The Mysteries of London I	F	F	F	F	F	F	F	F	F
The Mysteries of London II	F	F	F	F	S	S	F	F	S
The Vicomte de Bragelonne	S	F	F	F	F	S	S	F	S
The Court of London IV	F	F	S	F	S	S	F	S	S
Xi You Ji	S	F	S	S	F	S	S	F	S
Hong Lou Meng	S	S	S	S	S	S	S	F	S

Table 9: Final Qwen round summaries. Completion uses 13 attempts per row. Semantic metrics average defined scores from successful attempts within each novel, then across novels; reduced supports are explicit. Values are percentages except Composite (0–100).

Round	Runtime	Success	Completion	Recall	Precision	Weighted Q	Composite
R1	CC	**8/13**	**61.54**	5.05	18.82	4.31	6.30
R1	DSH	2/**13**	15.38	**8.70**	**33.93**	**5.66**	**9.48**
R1	QWC	**8/13**	**61.54**	4.27	22.61	3.76	6.18
R2	CC	8/**13**	61.54	**4.52**	27.13 ($n = 7$)	**3.52**	**6.92** ($n = 7$)
R2	DSH	7/**13**	53.85	3.39	11.45 ($n = 6$)	2.62	4.18 ($n = 6$)
R2	QWC	**11/13**	**84.62**	3.47	**27.95**	3.42	5.90
R3	CC	8/**13**	61.54	3.43	13.68	2.96	4.51
R3	DSH	3/**13**	23.08	**7.05**	30.63	**5.43**	**9.12**
R3	QWC	**12/13**	**92.31**	4.62	**31.13**	3.88	6.79

CC's 15 failures comprise 13 provider input-content refusals and two final JSON/schema failures. DSH's 27 failures comprise 15 request-budget exhaustions, six ordinary HTTP 429 errors, three HTTP 400 invalid requests, and three attempts without a valid JSON answer. QWC's eight comprise seven input-content refusals and one output-content refusal. These first causes differ from DeepSeek: its DSH failures emphasize request budget (nine) and memory (five), whereas QWC has 13 public-output format failures.

Same-novel status agreement for R1/R2, R2/R3, and R1/R3 is 9/13, 9/13, 11/13 for CC; 8/13, 9/13, 10/13 for DSH; and 8/13, 12/13, 9/13 for QWC. CC has five novels successful in every round and three failing in every round. Two successes and two failures exchange places in each consecutive comparison, despite unchanged 8/13 round completion.

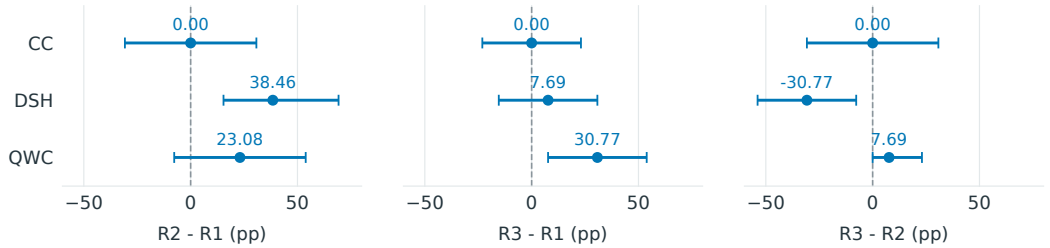


Figure 7: Qwen completion differences across rounds on all 13 novels. Points are means; bars are descriptive 95% book-bootstrap intervals (100,000 resamples).

The corresponding DeepSeek CC/DSH/QWC subset has 117 attempts (86 successful, 31 failed); it is reused from the principal comparison, not added again to the 664-attempt inventory. Across three rounds, DeepSeek completion is CC 38/39, DSH 23/39, QWC 25/39, compared with Qwen's 24/39, 12/39, and 31/39. Composite differences below are Qwen minus DeepSeek, in points on the 0–100 scale, using same-runtime, same-round novels where both models have successful predictions with calculable scores. Brackets give descriptive 95% book-bootstrap intervals. CC: R1 2.42 [-4.36, 10.47] (8 novels), R2 0.31 [-3.71, 4.33] (6 novels), R3 -3.66 [-8.58, 0.17] (8 novels); DSH: R1 0.64 [-0.13, 1.42] (2 novels), R2 -0.33 [-1.69, 0.73] (3 novels), R3 0.83 [-5.73, 7.39] (2 novels); QWC: R1 3.83 [0.04, 7.80] (7 novels), R2 0.72 [-5.30, 8.26] (7 novels), R3 1.24 [-3.55, 6.06] (7 novels). Only Hen Hai succeeds across all three Qwen runtimes and all three rounds. These small conditional supports do not establish a stable semantic ordering. Model alias, provider behavior, collection schedules, and native request trajectories are parts of the observed service configurations; the comparison is not a causal model-weight interaction test.

E.4 Support Sensitivity, Components, and Usage

The common-ten-novel analysis excludes *Clarissa*, *London I*, and *Court London IV*, because at least one runtime lacks a successful output there. In runtime order CC, DSH, CXH, QWC, OC, MSA, common-novel recall is 4.36%, 3.87%, 4.72%, 4.06%, 3.97%, 4.17%. This restriction improves comparability of support, but selects novels on successful completion and is not an unbiased substitute for the full fixed matrix. All 15 pairwise full-support intervals include zero for recall, mean case-quality score, distance-weighted case-quality score, and Composite.

Table 10: Supplementary components, book-macro percentages. Composite is on a 0–100 scale and is not used to declare a winner. Composite uses 11 novels/22 attempts for DSH and 13/36 for CXH; other supports equal recall support.

Runtime	Q_{weighted}	Q_{long}	Strict anchor	Matched label	Composite
Claude Code	2.63	1.77	0.61	96.15	4.76
DeepSeek Harness	2.62	3.53	0.49	**100.00**	4.83
Codex Harness	**3.22**	**3.57**	**1.04**	**100.00**	**5.50**
Qwen Code	2.91	2.72	0.97	**100.00**	5.16
OpenCode	2.74	3.11	0.83	98.08	4.82
mini-SWE-agent	2.78	3.43	0.38	99.40	5.04

Matched-label accuracy has supports of 13/32, 8/17, 13/29, 10/22, 13/28, and 12/18 novels/attempts, respectively. The reference's label imbalance and the small matched subsets limit the meaning of near-perfect conditional label accuracy. Long-distance quality is computed on each novel's long stratum; it does not establish a causal effect of distance, which is confounded with case content and discovery.

Additional recorded resource fields were recovered from terminal usage and call logs, including failed attempts across all six runtimes. Appendix E.14 reports known totals and remaining field-level gaps; missing standardized metadata alone does not imply unavailable underlying usage.

Table 11: Label recall and matched-evidence diagnostics. Values are book-macro percentages; parentheses give effective novels/attempts. E/U recall requires both identity and correct label among all reference cases of that label. Endpoint alignment conditions on matched cases; explanatory evidence recall is soft requirement recall on matched explained cases, not Explanation-Path F1 or label recall over all reference cases.

Runtime	E recall	U recall	Before	After	Evidence
Claude Code	3.75 (13/38)	7.50 (10/29)	35.56 (13/32)	46.11 (13/32)	51.07 (13/32)
DeepSeek Harness	3.21 (12/23)	2.69 (9/18)	36.42 (8/17)	**69.83** (8/17)	38.71 (7/15)
Codex Harness	3.65 (13/37)	5.59 (10/28)	40.06 (13/29)	50.11 (13/29)	55.72 (13/29)
Qwen Code	3.79 (11/25)	6.77 (8/20)	41.14 (10/22)	56.44 (10/22)	47.59 (10/21)
OpenCode	3.18 (13/32)	**14.21** (10/25)	**49.76** (13/28)	55.06 (13/28)	44.94 (13/26)
mini-SWE-agent	**3.80** (12/22)	3.89 (9/16)	26.99 (12/18)	35.70 (12/18)	**57.50** (12/18)

E.5 Label Imbalance and Rare-Class Support

Explained cases constitute 990/1,032 (95.9%) of the full reference, but this is not the appropriate reference accuracy for the selected matched subset. We hold every existing identity match fixed and assign the explained label to all matched cases. For each attempt, the always-explained (AE) accuracy is the fraction of its matched reference cases labeled explained. Actual and AE accuracy use exactly the same nonempty-match attempts, average replicates within novels, and then weight supported novels equally. The gain is computed within each novel before averaging and novel resampling. This is a conditional classification diagnostic, not a rerun or rematched agent baseline.

Table 12: Matched-label accuracy with the revised-reference matches held fixed. Actual and always-explained (AE) are book-macro percentages on identical novels/successful attempts (B/R). The paired gain is Actual minus AE in percentage points; brackets give 95% book-bootstrap intervals. M counts matched case-attempt instances, not distinct cases.

Runtime	B/R	M	Actual	AE	Gain [95% CI]
CC	**13/32**	**86**	**96.15**	92.74	3.42 [-7.91, 13.68]
DSH	8/17	45	**100.00**	77.08	**22.92** [4.17, 47.92]
CXH	**13**/29	69	**100.00**	92.09	7.91 [1.50, 15.60]
QWC	10/22	50	**100.00**	90.17	9.83 [0.00, 25.50]
OC	**13**/28	71	**98.08**	85.96	12.12 [1.26, 23.53]
MSA	12/18	49	**99.40**	98.36	1.04 [0.00, 3.12]

Table 13: Unresolved-label recall: a hit requires a match and a correct label. Macro estimates (percent) and 95% intervals resample supported novels; B/R gives novels/successful attempts with unresolved reference cases. Pooled counts use reference case-attempt opportunities. Union counts distinct cases matched with the correct label at least once over unresolved reference cases in supported novels; the full panel contains 42 unresolved reference cases.

Runtime	B/R	Macro [95% CI]	Pooled hits/opps. (%)	Union
CC	**10/29**	7.50 [1.67, 15.00]	5/**121** (4.13)	5/**42**
DSH	9/18	2.69 [0.00, 5.93]	3/82 (3.66)	3/41
CXH	**10**/28	5.59 [0.67, 13.38]	7/118 (5.93)	5/**42**
QWC	8/20	6.77 [1.56, 13.54]	5/76 (6.58)	4/39
OC	**10**/25	**14.21** [3.60, 26.95]	**11**/114 (9.65)	**7/42**
MSA	9/16	3.89 [0.00, 9.44]	2/65 (3.08)	2/41

All six point estimates exceed AE, but gain intervals are wide and three include zero (CC, QWC, MSA). In particular, DSH's 100% accuracy covers only 45 matched case-attempt instances across 8 novels/17 attempts; it does not show reliable discovery of unresolved cases. Its unresolved label recall is 2.69% [0.00, 5.93], based on 3/82 case-attempt opportunities. Each runtime correctly recovers only 2–7 distinct unresolved cases. These sparse diagnostics do not support a rare-class ranking.

Macro unresolved recall first averages correct unresolved hits divided by unresolved reference within each supported attempt and novel. Pooled recall instead sums these numerators and denominators, counting each successful attempt's reference separately; its observations are correlated and it is not the macro estimand. Union counts deduplicate case IDs within each novel across attempts. Novels with no reference cases labeled unresolved are N/A, and failed attempts have no evaluation score. Runtime-specific union denominators reflect which novels have successful output. The revised full panel has 42 unresolved cases; London II now has none, so its unresolved stratum is N/A rather than a zero observation. The percentile intervals retain the empirical novel support, including zero-hit novels; with so few hits, a zero lower endpoint does not establish absence of capability.

Pooled classification on identical matched supports. The pooled accuracies in Table 14 sum correct labels and matched occurrences before division. Their AE comparator must use that same support and aggregation, rather than the book-macro estimates above. Let m_i be the matched count, c_i the correct-label count, and e_i the count of matched reference cases labeled explained in attempt i. Then pooled actual accuracy is $\sum_i c_i / \sum_i m_i$, AE is $\sum_i e_i / \sum_i m_i$, and gain is their difference.

Table 14: Original pooled label diagnostics on frozen Main matches. Actual and AE use identical matches within each condition. Accuracies are percentages and gains are percentage points; U counts correctly labeled unresolved occurrences and distinct novel–case IDs. Conditions have different matched supports.

Condition	Matches	Actual	AE	Gain	Correct U	Unique U
DS Original	370	99.19	91.08	8.11	**33**	**14**
Qwen Original	126	98.41	84.92	**13.49**	18	11
GPT Original	**470**	**99.57**	**98.51**	1.06	7	5

Repaired labels and matching change these supports. DS has 367 correct labels and 337 matched reference cases labeled explained; Qwen has 124 and 107; GPT has 468 and 463. GPT correctly labels seven unresolved occurrences spanning five cases. An always-explained baseline retains incorrectly labeled explained matches in its numerator. The rare-class change reflects revised labels and supports, not new outputs. Table 14 uses the original 300-request DS/Qwen conditions and GPT C8, with the 500-request retries reported separately.

E.6 Reference Provenance Sensitivity

Main retains cases with initial-human or GPT discovery support; cases found only by DeepSeek are retained in the expanded references. This removes DeepSeek-exclusive contributions from the principal denominator, while leaving shared contributions in it. It does not make Main source-neutral for GPT or equalize historical discovery opportunities. The analysis below tests how the known additions affect the same DeepSeek predictions.

The same 177 successful principal predictions are evaluated against the repaired Main (1,032 cases), Main plus 26 DeepSeek-exclusive cases (1,058), and Full (1,066). Full contains eight later-manual additions: four already in the old supplement and four omissions established by post-evaluation review. The latter four never enter Main or +DS. All 57 failed attempts remain unavailable for semantic scoring. Main partitions into 136 initial-human, 848 GPT-only, and 48 shared cases. Existing predictions and scoring parameters remain unchanged; matching is recomputed against each reference variant.

Figure 8 reports all six runtimes. All 15 paired intervals contain zero for recall, Q, weighted Q, and Composite under every variant. This preserves the low case recall without establishing runtime equivalence or removing unknown contributor bias.

Source-conditioned rediscovery can reflect case salience, unequal historical discovery opportunities, or affinity between a contributor's evidence expression and subsequent outputs. These are alternative hypotheses, since source membership is itself selected by earlier discovery. In retrieval, simulated pooling demonstrates possible bias toward contributing system types (Yilmaz et al., 2020), while reassessment of TREC-8 finds almost unchanged rankings after expanding its judgments (Voorhees et al., 2022). These complementary findings motivate measuring source sensitivity rather than

assuming its direction. Our result establishes stability under the observed additions; equal-opportunity construction and noncontributor evaluation would test the explanations for the source-slice differences.

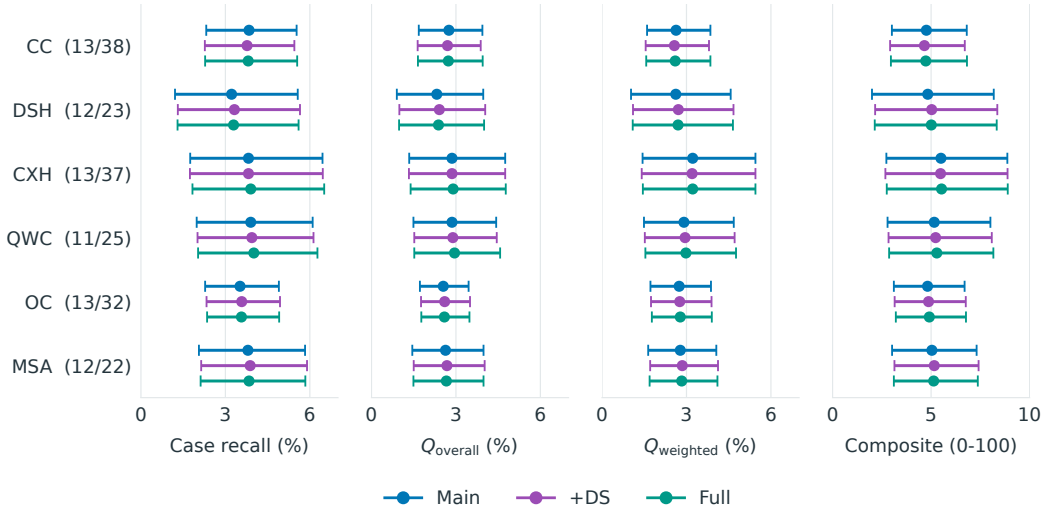


Figure 8: Reference-source variants across six DeepSeek runtimes. Points and bars show book-macro estimates and 95% book-bootstrap intervals. Row labels give novels/successful attempts for recall and quality; Composite uses 11/22 for DSH and 13/36 for CXH. Main, +DS, and Full contain 1,032, 1,058, and 1,066 cases.

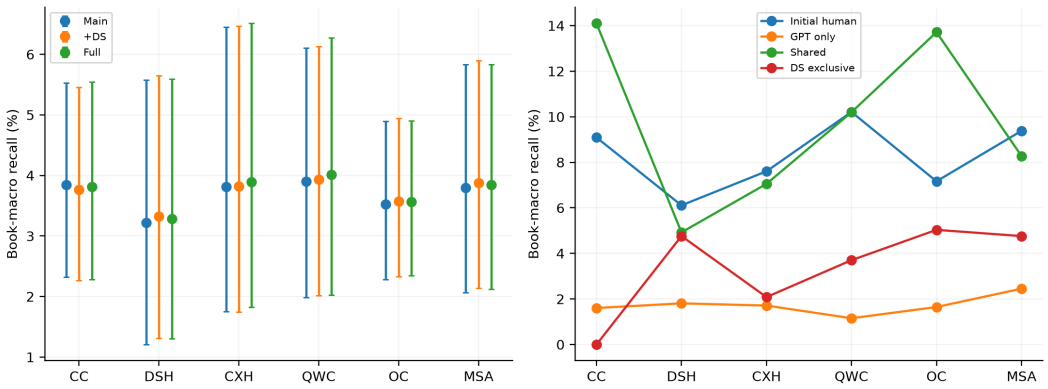


Figure 9: Sensitivity to known reference provenance, with 95% book-bootstrap intervals. Left: Main and expanded denominators with 95% intervals. Right: descriptive source-conditioned partial-reference recall point estimates; support differs by slice and runtime (Table 15). Source slices are descriptive; their differences do not identify model preference or case difficulty.

Table 15: Source-conditioned case recall (percent). Parentheses give effective novel counts, not numbers of independent cases. Initial-human/GPT-only/shared/DeepSeek-exclusive contain 136/848/48/26 cases across 13/13/12/8 novels, respectively.

Runtime	Initial-human	GPT-only	Shared	DS-exclusive
CC	9.09 (13)	1.60 (13)	**14.10** (12)	0.00 (8)
DSH	6.11 (12)	1.81 (12)	4.92 (11)	4.76 (7)
CXH	7.61 (13)	1.71 (13)	7.05 (12)	2.08 (8)
QWC	**10.21** (11)	1.15 (11)	10.20 (10)	3.70 (6)
OC	7.17 (13)	1.65 (13)	13.72 (12)	**5.03** (8)
MSA	9.38 (12)	**2.45** (12)	8.28 (11)	4.76 (7)

All source slices are partial references; their precision and Composite are N/A. The eight later-manual cases include four post-evaluation additions. Relative to revised +DS, Full credits five additional matches across four distinct cases; one is reported twice. This is recovered credit for existing predictions, not new model discovery. Slice supports also differ because not every runtime succeeds on every novel. Source-conditioned case recall varies substantially, associating provenance with matched discoverability without identifying why. Historical opportunities, case content, language, and evidence expression can all contribute. In particular, GPT-only's low recall does not establish that GPT discovered intrinsically harder cases.

These results test known denominator variants, not reference truth or matcher validity. GPT-only still accounts for 82.2% of Main, the two added groups are small, and all variants share the same quotation-based matcher. Thus the analysis cannot remove unknown contributor bias or estimate population false-match and missed-match rates.

London I's ML116, confined to Full, has been reframed as a false claim of continuing disfigurement followed by the unveiling of a healed face. Medical treatment and testing affection explain the concealment and revelation. Current scores use that repaired reference; the earlier audit records retain the description they assessed.

E.7 Reference Repairs and Full Rematching

Human-reviewed label, duplicate, qualification, and endpoint repairs are summarized in Table 16. The revised references contain 1,032 Main cases (990 E/42 U), 1,058 in +DS, and 1,066 in Full. New Full-only entries are distinguished from existing-case repairs below.

Table 16: Reference issues and implemented repairs. The two Dracula IDs are records of one reversal. Main and +DS receive existing-case repairs; ML116 is Full-only.

Record	Issue	Implemented repair
JHY052; London I ML110; London II ML050	Review finds sufficient explanations for three unresolved cases.	Explained routes added in both text conditions.
DR027 and DR058	Duplicate expressions of one information-secrecy reversal.	DR058 merged into DR027.
JMQY040	Provisional option followed by a final choice, without the required reversal.	Ineligible case removed.
HLM053	Original endpoints concern different state dimensions.	Previously excluded case restored with attachment endpoints.
London I ML116	A mistaken scar belief was treated as objective fact.	Full-only endpoint repaired.
XYJ140	Existence of a fan did not establish its expected effect.	Before endpoint specifies the fire-extinguishing property.

Four review-discovered omissions receive new Full-only entries: HH028 (U108), JMQY050 (U174), and London II ML159/ML160 (U098/U100). They are post-evaluation manual additions, not new initial seeds. Main and +DS receive none of them. All four original outputs match their new Full

entries with correct labels; the author's judgments that U174 and U098 still need explanatory or endpoint support remain separate from match credit.

The reference-repair analysis uses 519 successful predictions from 705 qualified attempts, plus 91 construction-linked historical outputs. These are a subset of the completed 636-output rematching, which also included 26 successful context-retention probe outputs outside this paper's scope. Parameters, thresholds, and predictions are unchanged; candidate edges, global one-to-one assignments, evidence alignment, distance weights, and scores were recomputed. The 186 failures in the reported scope retain unavailable evaluation scores. Original scores and human returns remain intact. The expanded historical trajectories in Figure 5 additionally use 89 available predictions scored against the same revised Main reference (Appendix E.12).

The largest Main runtime changes are 0.090 percentage points in recall, 0.052 points in Q, and 0.092 Composite points. All 15 runtime intervals still include zero under each reference variant. Full gains five matches across four distinct cases relative to repaired +DS, while Main changes from 369 to 370 matches and retains 161 distinct DS cases. GPT's formal three-round union changes from 307 to 305 cases (32.79% novel macro), and its historical union from 688 to 686 (69.55%). These are changes in how existing outputs are credited, not newly generated discoveries.

The repaired INJ reference contains 524 explained and 508 unresolved cases, with 466 eligible edited targets. JMQY040's original text edit is retained but no longer scored as a valid target. Surviving-route findings still prevent interpreting every INJ predicted explained reversal as an error. Current paired Composite changes are -0.324 [-1.128, 0.427] for DS and -6.108 [-8.524, -3.408] for GPT.

Audit sample sizes, inclusion probabilities, weights, judgments, and estimates below remain tied to the versions actually reviewed. In particular, the 1,033-case catalog, 803 unmatched predicted cases, and 369 matched pairs are historical sampling populations; they are not silently replaced by the revised scoring denominators.

E.8 Rediscovery and Edited-Text Checks

INJ was designed to remove recorded explanations while retaining the opposing endpoints. Full-text review found surviving explanatory routes in most of the purposively selected targets (Table 18). This prevents treating the construction label switch as a verified semantic change. The paired score tables that follow report behavior under these edited texts and retained construction labels; they do not validate sensitivity to successful explanation removal. The Original completion, output-count, and repeated-discovery results do not depend on this manipulation.

The P5/P6 mechanical counts and audit materials in this subsection retain the original 467-target, pre-repair scoring condition. They document sample selection and the judgments actually returned; they are not the current intervention result table. The three forward Original dropout counts remain unchanged after repair.

Original-to-Original rediscovery (P5). We reuse the 467 frozen edit targets and all three successful GPT rounds. Each comparison conditions on targets matched and labeled explained in its source Original round. Table 17 reports all forward round pairs alongside same-round Original-to-INJ pairs. The forward Original comparisons lose 108/194 matched hits in aggregate, but those 194 observations reuse targets and rounds. All six directed Original comparisons lose 253/426; the reverse comparisons are 50/76, 50/78, and 45/78. These are descriptive rediscovery counts, not independent-case estimates.

Table 17: GPT rediscovery under frozen matching. O denotes Original and I denotes INJ; E/U are destination labels. Each denominator contains source-round explained hits. No edit is made between Original rounds.

Comparison	Source E hits	Destination E	Destination U	Unmatched
$O_1 \to O_2$	59	26	0	33
$O_1 \to O_3$	59	27	0	32
$O_2 \to O_3$	76	**33**	0	43
$O_1 \to I_1$	59	9	3	47
$O_2 \to I_2$	76	7	2	**67**
$O_3 \to I_3$	**78**	15	**4**	59

For the 173 Original hits that become unmatched in INJ (129 distinct targets), we inspect all 3,107 corresponding prediction–target pairs using the unchanged scorer kernels and recorded parameters. No instance has an eligible edge to that target, so none is lost solely to global assignment. For 28 instances, at least one prediction passes the subject and single-endpoint guards but fails the normalized combined threshold; for 145, none passes both guards. INJ target references are unresolved, so eligibility uses $T/3 + 2S/3 > 0.50$, $T \geq 0.50$, and maximum endpoint alignment at least 0.30. These mechanical categories neither establish semantic absence nor estimate a matcher false-negative rate. Subtracting Original and INJ dropout rates does not identify a pure deletion effect: supports, texts, versions, and scheduling differ.

Mechanical localization and author verification (P6). The 101 INJ matched-explained instances cover 85 targets across 13 novels. Only 31 had a paired Original explained hit; the other 70 were unmatched in Original. Their 132 explanatory evidence quotations yield 44 exact full matches, 71 full matches after character normalization, 12 ellipsis-fragment matches, and five unlocated quotations in INJ. Normalization uses NFKC, case folding, and alphanumeric characters; fragment matches do not establish a continuous quotation, order, or semantic sufficiency. First matches of short fragments may be unrelated occurrences, so locations must be checked in context.

We select all 15 targets with at least one instance whose explanatory evidence quotations are not all fully located, plus HH002 and HH005, whose quotations overlap Original explanatory evidence anchors. All related matched-explained rounds are included: 17 targets from nine novels, 23 predicted cases, and 30 quotations. Selection is purposive and outcome-informed; neither it nor author verification estimates the composition of all 85 targets, all 101 predicted cases, or all 467 edited targets. The audit is separate from the 96-pair independent blinded scorer study.

AI tools helped organize the review materials and locate source passages; the author made the initial item-level judgments. Two later reviewers independently returned the 17-target assessments without seeing each other's answers or the old author decisions. Reviewer A had seen Original routes for all 17 targets; B reported prior exposure for two overlapping Reference-audit targets. Final author adjudication resolved the disputed fields and two conflicts with the earlier record. The final values below use those returns and decisions, with raw judgments retained separately. This is not an Original-naive blinded manipulation check (Appendix E.17). For each output, the record distinguishes endpoint qualification, semantic target identity, quotation existence, sufficiency of the model's cited evidence, and other routes found in the full INJ text. Model reason text is not accepted as textual evidence; additional causal passages found during review do not count as evidence the model supplied. Original routes and deletion records are checked alongside the edited text, with changes from preliminary judgments retained.

Table 18: Author-confirmed findings for the purposive P6 sample. S: model evidence sufficient; I: insufficient; N/A: target qualification or prediction identity prevents that judgment. A surviving route can exist even when the model's evidence is insufficient. Counts are sample descriptions, not population estimates.

Target	Model evidence by round	Edited-text target finding
DR022	R3: I	Sufficient route survives
HH002	R1, R3: S	Sufficient route survives
HH005	R1–R3: I	Sufficient route survives
JMQY003	R1, R3: I	Sufficient route survives
JMQY038	R3: S	Sufficient route survives
JMQY040	R1: N/A	Ineligible: proposal followed by decision
JHY032	R3: S	Sufficient route survives
C008	R1: S; R2, R3: I	Sufficient route survives
C013	R2: I	No complete explanatory route found in audit
C069	R2: S	Sufficient route survives
ML040	R3: I	Sufficient route survives
VB067	R1: S	Sufficient route survives
VB078	R1: N/A	Route survives; prediction mismatched
CL017	R1: S	Sufficient route survives
CL021	R2: S	Sufficient route survives
CL031	R2: S	Sufficient route survives
HLM064	R2: S	Sufficient route survives

The 30 quotations comprise 13 continuous quotations and 17 sets of present fragments. Across 23 predicted cases, cited evidence is sufficient in 11, insufficient in ten, and inapplicable in two. Nine of the ten insufficient outputs nevertheless concern targets with sufficient routes elsewhere in INJ. Fifteen targets retain sufficient explanations; C013 has no complete explanatory route found in the audit, and JMQY040 is ineligible. VB078 belongs to the 15 surviving-route targets, but its prediction describes a different reversal. These are the final two-reviewer plus author-adjudicated results, superseding the earlier 10/11/2 output counts and 13/14/3 quotation categories.

What the counterexamples distinguish. In *Hong Lou Meng* HLM064, the deleted narrator explanation is not the only causal account: Wang's accusation that the maid has corrupted Baoyu remains beside the slap and the preceding exchange. The model's quotation omits an intervening action without an ellipsis, yet retains sufficient causal content. Quotation fidelity and explanation sufficiency therefore differ. In *Clarissa* C013, by contrast, cited passages show Lovelace's possible innocence to the reader and an assurance Clarissa previously distrusted, without establishing new information reaching her; no complete explanatory route was found. In *The Vicomte de Bragelonne* VB078, the prediction begins with feigned courtship much earlier than the target's jealous vow to extinguish love: a shared later endpoint does not make them the same reversal. JMQY040 is ineligible because its before is a proposal within a discussion and its after is the resulting decision, not a reversal of an established position.

These findings preserve the frozen records but invalidate an unconditional reading of INJ predicted explained reversals as semantic errors. They concern shared edited texts and thus limit both C7 and C8 intervention claims; the audit does not label individual C7 outputs. The frozen scores are preserved. In a separate bounded exclusion analysis, removing the 15 residual-route targets and JMQY040 leaves 1,353 GPT and 4,564 DeepSeek paired opportunities; complete explained-to-unresolved transitions remain nine and three. The reviewed VB078 prediction–target mismatch is additionally excluded at the pair level in its own sensitivity slice; it does not change either switch count. This exclusion check does not establish successful explanation removal for the remaining targets.

E.9 DeepSeek Original/INJ Diagnostics

The C7 and C8 score tables below use the revised construction references. The edited-text audit in Appendix E.8 determines their interpretation: matching a target's retained unresolved label is agreement with the construction record, not independent confirmation that the edited text lacks an explanation. The intervention matrix contains 234 qualified INJ attempts: 174 successful and 60

failed. Together with the fixed Original inventory of 234 attempts (177 successful, 57 failed), it yields 146 same-runtime, same-novel, same-replicate successful pairs. R1, R2, and R3 contain 51, 47, and 48 such pairs. Semantic metrics condition on these actual pairs; the other 88 cells remain part of completion and are not assigned semantic zeros.

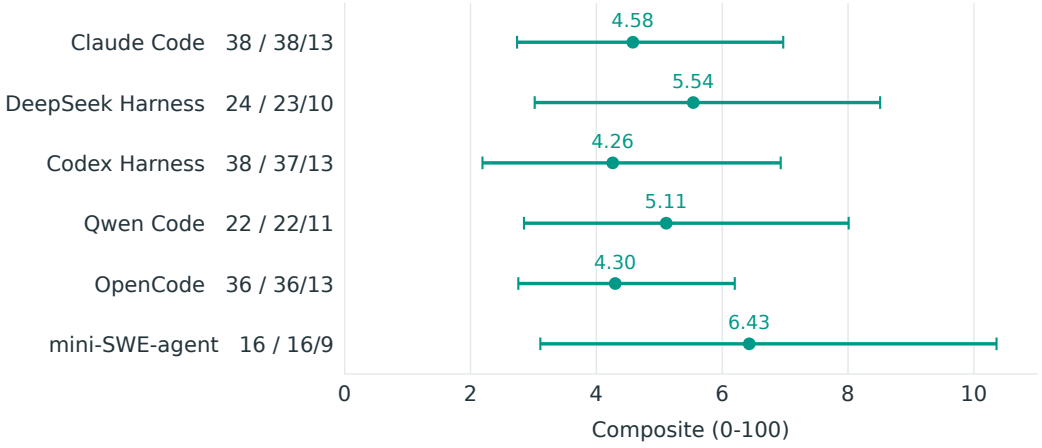


Figure 10: DeepSeek INJ Composite under revised construction labels: book-macro means and 95% book-bootstrap intervals. Row labels give successful attempts / attempts with defined Composite / supported novels; each runtime has 39 qualified attempts. Paired supports are reported separately in Table 19.

Figure 10 reports absolute INJ scores. They must not be subtracted from Original means on different supports to replace the paired differences. One successful output each from DSH and CXH is empty, leaving Composite undefined; failures are also excluded from semantic means, rather than scored zero.

Table 19: DeepSeek C7 paired results under revised references. O/INJ success counts retain 39 attempts per runtime; Pair counts same-replicate successful pairs. Differences have 95% book-bootstrap intervals: recall in percentage points, Composite on 0–100. $E \to U$/O-hit counts switches among Original explained hits; correctness here is relative to construction labels, whose validity is challenged by the audit.

Runtime	O success	INJ success	Pair	Δ Recall	Δ Composite	$E \to U$/O-hit
CC	**38/39**	**38/39**	**37**	-0.42 [-1.79, 0.89]	-0.26 [-2.37, 1.80]	0/**27**
DSH	23/**39**	24/**39**	18	**0.98** [-0.84, 2.65]	-0.28 [-2.99, 1.93]	**1**/12
CXH	37/**39**	**38/39**	36	0.09 [-0.93, 1.49]	-1.13 [-2.61, 0.45]	**1**/21
QWC	25/**39**	22/**39**	17	-0.18 [-1.49, 0.87]	-0.63 [-2.77, 1.44]	**1**/13
OC	32/**39**	36/**39**	30	0.48 [-0.54, 1.50]	-0.80 [-2.07, 0.29]	0/24
MSA	22/**39**	16/**39**	8	0.74 [-1.67, 4.27]	**1.95** [-2.68, 6.54]	0/5

Table 20: DeepSeek C7 target transitions in 146 successful pairs. Entries count 4,727 correlated target-attempt opportunities. Rows are Original predictions and columns are INJ; unmatched is distinct from either label.

Original → INJ	Explained	Unresolved	Unmatched
Explained	18	3	81
Unresolved	0	2	0
Unmatched	**66**	**17**	**4540**

The three full transitions occur once in R1, twice in R2, and not in R3. By runtime, DSH, CXH, and QWC each contribute one; CC, OC, and MSA contribute none. Counts are too small and too correlated to rank runtimes. Across three rounds, the number of distinct INJ targets recovered with the frozen unresolved label at least once is 1, 5, 5, 3, 7, and 1 for CC, DSH, CXH, QWC, OC, and MSA, respectively; the corresponding Original explained unions are 22, 14, 20, 14, 18, and 13. A union measures cumulative label recall, not single-attempt response.

The same pairs contain 5,892 non-target opportunities. Of 207 correctly identified Original controls, 65 remain correct and 142 are unmatched in INJ; none becomes incorrect. Among Original-unmatched controls, 141 become correct and one becomes incorrect. This instability provides a control for rediscovery noise, without isolating its mechanism from lexical matching or open search.

INJ R1 is globally serial. R2 begins in one serial queue, then an authorized stop occurs after eight successes and one qualified failure; the remaining fixed cells continue in six runtime-specific serial queues. R3 uses the same six-queue structure from the start. The manually stopped R2 QWC–Dracula attempt is not a qualified failure; its authorized new Run ID occupies that fixed cell. No other failure is replaced. Concurrency and timing are therefore reported conditions and do not modify attempt or pairing denominators.

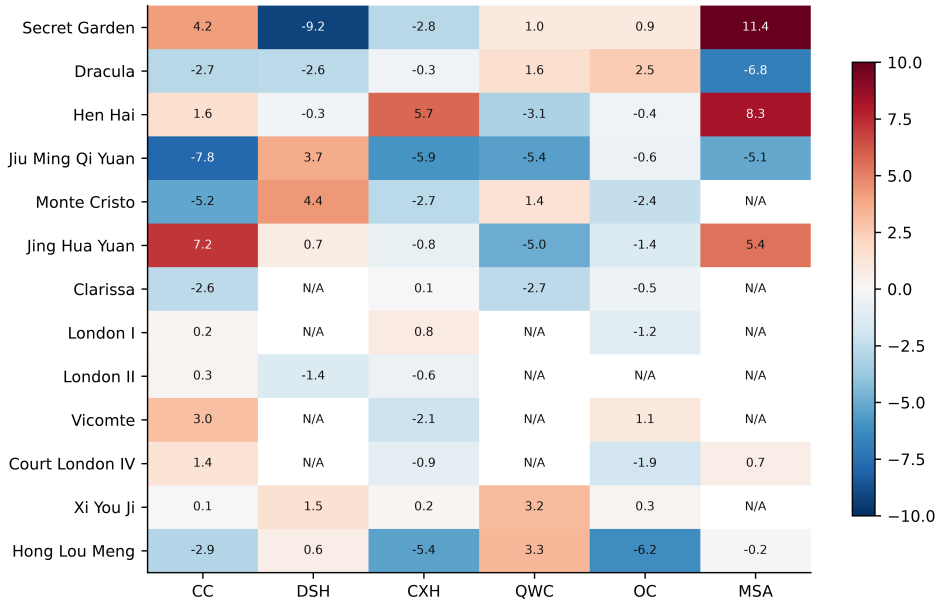


Figure 11: Per-novel mean Composite change (INJ minus Original) among successful same-replicate DeepSeek pairs. N/A denotes no successful pair and remains unavailable rather than zero. The mixture of directions and unequal support prevents a stable semantic ranking.

E.10 Formal GPT Original/INJ Diagnostics

The secondary study comprises 79 qualified attempts: 78 successful scored outputs and one provider-capacity failure in Original R3 for *Xi You Ji*. A separately authorized replacement succeeds, giving 13 Original/INJ pairs in each of three rounds. The failed attempt remains recorded with no evaluation score. Each successful prediction passes the frozen public-output projection and is scored against its condition-specific reference.

Table 21: GPT target transitions under revised case matching. Entries count 1,398 correlated target-attempt opportunities, not independent cases. Rows are Original; columns are INJ. Explained in INJ is not automatically a semantic error.

Original → INJ	Explained	Unresolved	Unmatched
Explained	31	9	172
Unresolved	0	0	1
Unmatched	**69**	**12**	**1104**

Table 22: Six GPT conditions, each with 13 successful predictions. Recall and matched-label accuracy are book-macro percentages; Composite is on a 0–100 scale with 95% book-bootstrap intervals. Target hits are pooled counts among 466 eligible cases with construction labels: explained in Original, unresolved in INJ; the latter labels are challenged for audited targets. The inventory additionally retains one failed Original R3 attempt.

Condition	Predictions	Recall	Matched label	Composite [95% CI]	Targets
Original R1	299	17.06	99.04	18.36 [13.88, 23.30]	59/**466**
Original R2	329	**18.37**	**100.00**	19.23 [15.52, 23.19]	76/**466**
Original R3	**348**	18.24	99.57	**20.74** [15.84, 26.03]	**77**/**466**
INJ R1	222	13.37	**68.16**	**13.62** [8.94, 18.76]	6/**466**
INJ R2	203	13.53	62.73	12.81 [7.95, 18.60]	6/**466**
INJ R3	**256**	**14.30**	60.31	13.58 [9.85, 17.75]	**9**/**466**

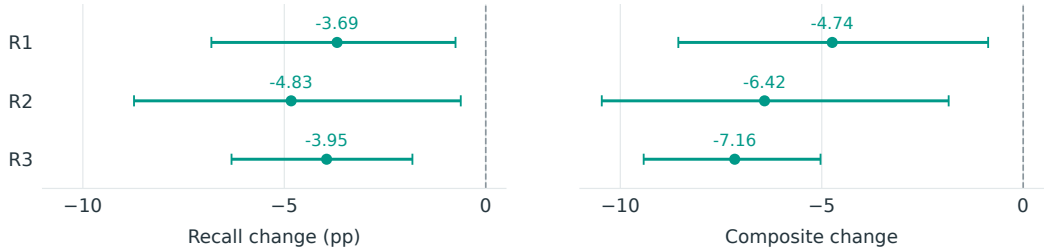


Figure 12: GPT paired INJ minus Original differences. Points are book-macro means; bars are 95% book-bootstrap intervals over 13 paired novels, computed from unrounded scores.

Averaging rounds within each novel gives Original Composite 19.44 [15.43, 23.56] and INJ 13.34 [9.48, 17.67], with a paired difference of -6.11 [-8.52, -3.41]. INJ Composite is lower in 10/13, 11/13, and 13/13 novels in R1–R3. The difference reflects both output changes and condition-specific reference labels and evidence structure. With fixed parameters, reference structure still determines matching eligibility, label/evidence credit, and distance weights. The difference therefore does not isolate behavioral response to explanation removal or any component's causal contribution. Table 22 and Figure 13 report the revised scores.

Of the 212 Original explained target-attempt hits, 9 (4.25%) switch to the revised unresolved label, 31 (14.62%) remain explained, and 172 (81.13%) are unmatched after editing. The nine full transitions are 0.64% of all 1,398 opportunities. INJ has 21 frozen-label unresolved hits (1.50%), 12 of which were unmatched in the corresponding Original round. Its 681 predicted cases contain 60 unresolved labels overall; declaring a case unresolved is not the same as correctly identifying an edited target.

Original rounds recover 142 distinct explained targets at least once; INJ recovers 17 with the construction unresolved label (17/466, 3.65% pooled). Cumulative label recall is not single-attempt performance. Because the rounds revisit the same novels and targets, neither the 39 novel-round pairs nor the 1,398 target-attempt opportunities provide that many independent units for uncertainty estimates.

Codex CLI changes from `0.154.0-alpha.6.2` in R1 to `0.155.0-alpha.2.6` in R2/R3. Original R2 and R3 run simultaneously as two internally serial queues; INJ R2 and R3 run successively with two queues inside each round. Both schedules allow at most two concurrent model processes, but their allocation differs. These native-version and scheduling differences limit interpretation across rounds. The small upward Original point-estimate sequence is not evidence of learning.

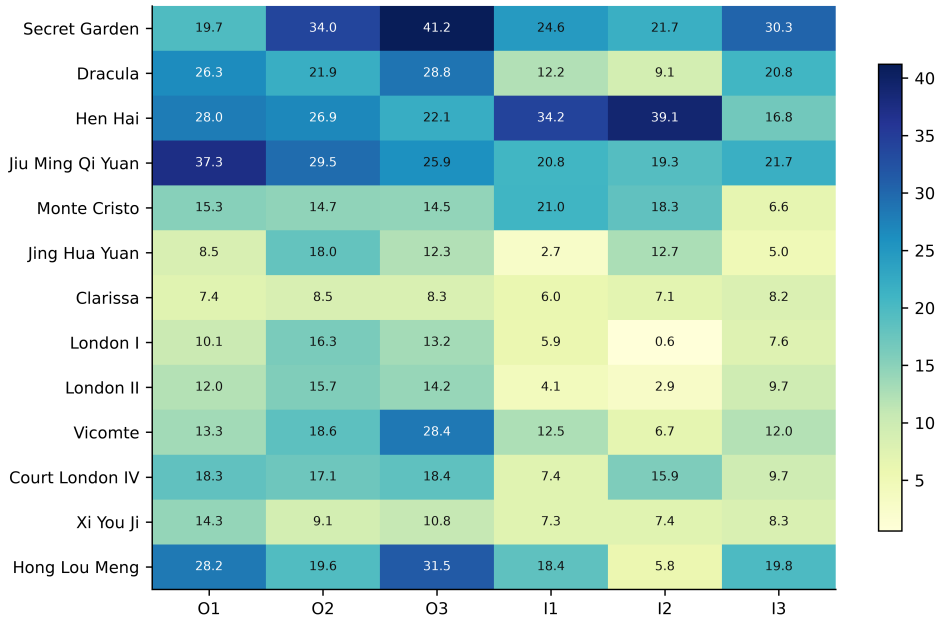


Figure 13: Per-novel Composite across the six completed GPT conditions: O1–O3 and I1–I3 denote Original and INJ rounds, scored against their corresponding revised references. This descriptive heatmap retains within-novel variation and does not compare GPT with the principal DeepSeek runtimes.

The formal cohort still uses a model family that contributed to the reference and construction novels, so input isolation does not establish independent held-out generalization. The construction-linked comparison is reported separately in Appendix E.12.

E.11 Human Judgments by Stratum and Evidence Role

Case-identity agreement is 48/54 for English and 42/42 for Chinese; all six disagreements remain in the 96-pair denominator. The preassigned primary reviewer identifies 15/16 same-case controls as the same and 15/16 hard negatives as different.

Among the 32 selected pairs, the primary reviewer judges all 32 before roles correct, 31 after roles correct and one partial. The 41 explanatory-role judgments comprise 26 correct, ten partial, and five wrong. Explanatory sufficiency judgments are 18 sufficient, eight partial, and six insufficient. Thus high selected-match agreement coexists with weaker explanatory sufficiency judgments. Near-threshold same-case judgments do not authorize silently changing the matching threshold or attributing all unmatched predicted cases to model failure; an independently sampled error analysis is needed for population estimates.

E.12 Historical Discovery Diagnostics

Figure 5 uses all 180 available scored GPT predictions linked to construction records across the 13 principal novels. Each novel's effective GPT rounds are indexed from its first available prediction, excluding DeepSeek rounds; lines end at the last available prediction without imputation. Both metrics use the fixed final Main reference. This extends the previously scored common-round subset

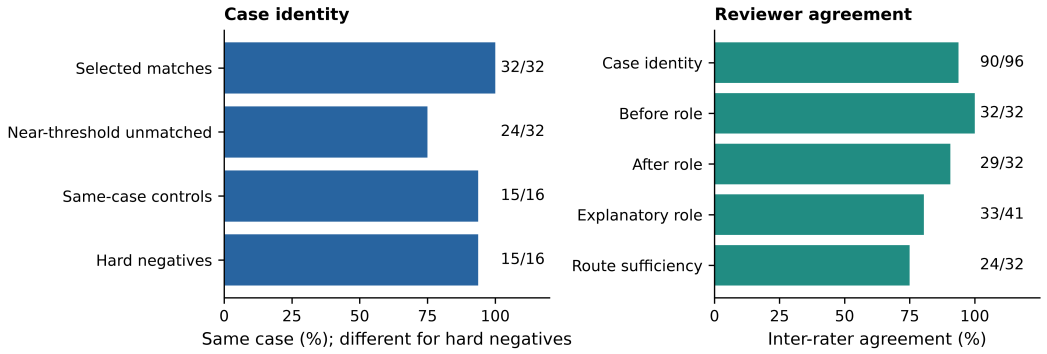


Figure 14: Independent human validation. Left: the preassigned primary reviewer's judgments; near-threshold pairs judged the same case expose missed correspondences. Hard negatives show the different-case proportion. Right: raw agreement between two independent reviewers, with different denominators for case identity, endpoints, individual explanatory roles, and explanatory sufficiency. These are challenge-sample diagnostics.

with 89 predictions; rescoring the original 91 reproduces their existing metrics and matched case sets. Eight other historical directories lack public predictions and are omitted from these score trajectories, so the 180 predictions are not a complete qualified-attempt inventory. The common-round analyses below retain their original 91-prediction scope; no cumulative analysis is added for the expanded trajectories.

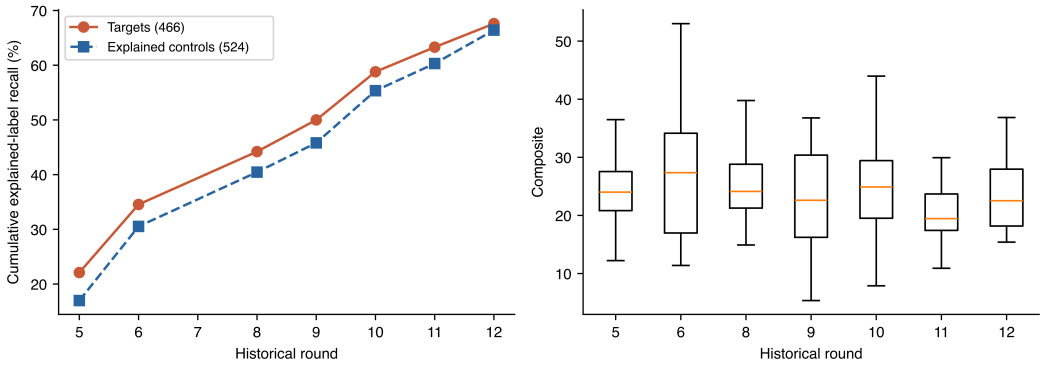


Figure 15: Seven shared historical rounds (91 predictions). Left: cumulative union of cases matched and labeled explained, split into future intervention targets and non-target explained controls. Right: descriptive per-novel Composite distributions; boxes show quartiles and medians, with whiskers to 1.5 interquartile ranges and outliers omitted. These predictions are linked to reference construction and are not a model leaderboard.

The common rounds are R5, R6, R8, R9, R10, R11, and R12, each covering all 13 principal novels. R7 lacks a corresponding output for *The Secret Garden* and is therefore outside this common-round intersection. The shared subset contains 2,102 predicted cases. Under the revised reference, cumulative explained-label recall on intervention targets rises from 103/466 to 315/466; explained controls rise from 89/524 to 348/524. The book-macro Composite is 24.45 [21.27, 27.87], compared with 19.44 [15.43, 23.56] for the formal Original mean; the paired formal-minus-historical difference is -5.01 [-6.86, -3.04]. This contrast is construction-linked, not a controlled change in model ability. Model, budgets, native runtime, and output conditions also differ from the principal DeepSeek comparison. No request, token, or monetary-cost values are imputed for historical predictions lacking standardized usage metadata.

E.13 Discovery Dynamics and Output Scale

For each novel, an attempt's discovery set contains the reference IDs in its saved confirmed matches. New cases are the set difference from earlier rounds; cumulative case recall is the union size divided by that novel's reference size, averaged equally over 13 novels. Thus 305/1,032 is not the book-macro percentage 32.79%. Formal GPT Original uses 40 qualified attempts (39 successful, one failed); the common-round historical subset contains 91 scored outputs over the seven common construction rounds. Their unions are separate.

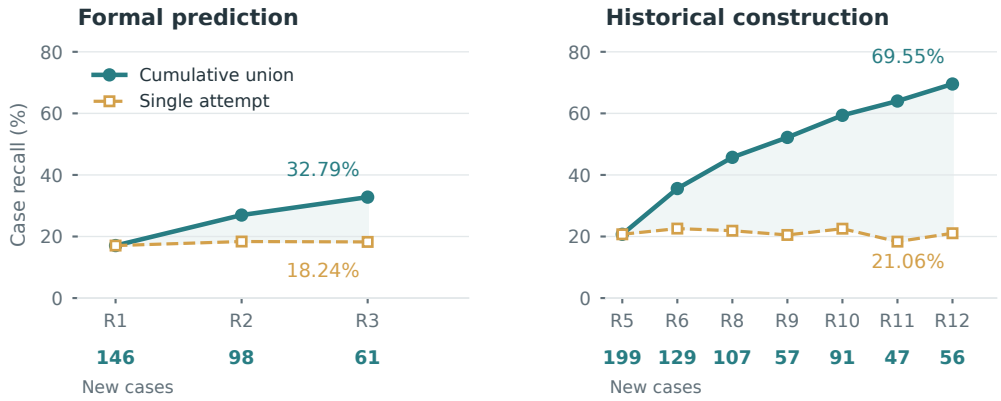


Figure 16: GPT–Codex single-round and cumulative case recall. Each panel averages case recall equally over 13 novels; annotations count new novel–case pairs. Formal and construction-linked historical attempts form separate unions.

Table 23: Historical values underlying Figure 16; formal rounds appear in Table 7. Single-attempt and cumulative case recall are book-macro percentages; new and union counts are distinct novel–reference IDs.

Cohort	Round	Single (%)	New cases	Union cases	Union (%)
Historical	R5	20.72	**199**	199	20.72
Historical	R6	**22.57**	129	328	35.58
Historical	R8	21.88	107	435	45.72
Historical	R9	20.50	57	492	52.20
Historical	R10	22.54	91	583	59.37
Historical	R11	18.34	47	630	64.01
Historical	R12	21.06	56	**686**	**69.55**

The full DeepSeek union contains 161 cases; 871 are not matched by any of the 234 attempts. Counts covered by exactly one through six runtimes are 94, 34, 19, 3, 6, and 5. Runtime unions are CC 68, DSH 35, CXH 53, QWC 39, OC 54, MSA 42. Each runtime contributes 12–26 cases absent from the other five in this cohort.

For a fixed k attempts per novel, we enumerate every choice of k rounds of one runtime, or k distinct runtimes with one round each. Within each strategy and novel, all choices receive equal weight, without selecting the best-performing choice. For DeepSeek, same-runtime choice counts at $k = 1, 2, 3$ are 18/18/6; mixed-runtime counts are 18/135/540. Failed attempts contribute no set of matched cases but retain their attempt slot. The two strategies have identical expected inclusion of each cell, hence equal expected successful-attempt counts and equal known usage at each k. Cumulative case recall differences use paired novel resampling (100,000 draws; seed 20260916).

A success-only common-support check retains novels with at least k successes for every runtime. DeepSeek has 10/5/0 eligible novels at $k = 1/2/3$; Qwen has 5/3/1. Different k values therefore do not form one success-only learning curve. For DeepSeek's five-novel $k = 2$ subset, the difference in cumulative case recall between mixed-runtime and repeated-runtime attempts is -0.206 [-0.648,

Table 24: Equal-attempt exploration. Cumulative case recall is a book-macro percentage; differences are percentage points with descriptive 95% book-bootstrap intervals.

Model	k	Repeated runtime	Mixed runtimes	Difference [95% CI]
DeepSeek	1	**3.056**	**3.056**	0.000 [0.000, 0.000]
DeepSeek	2	5.228	**5.313**	0.084 [-0.046, 0.230]
DeepSeek	3	6.794	**7.135**	**0.341** [0.031, 0.675]
Qwen	1	**2.513**	**2.513**	0.000 [0.000, 0.000]
Qwen	2	4.326	**4.549**	0.223 [0.031, 0.491]
Qwen	3	5.837	**6.230**	**0.393** [0.066, 0.766]

0.237] points, so the all-attempt gain does not imply a universal conditional advantage of switching runtimes.

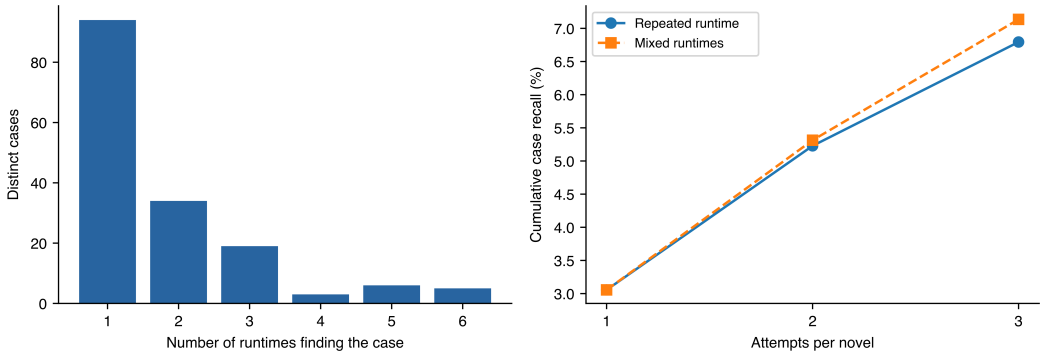


Figure 17: DeepSeek discovery overlap and equal-attempt combinations. Failures occupy fixed cells and contribute no matched cases; common-success supports are separate.

Composite after merging three attempts. For each principal novel and DeepSeek runtime, we concatenate all public cases from successful outputs in R1–R3, in round and original case order, and remove only exactly identical case objects. Targets, labels, endpoint quotations, explanatory evidence, and reasons are all included in the equality test; no evidence is rewritten or selected using the reference. No exact duplicates occur in these outputs, so all 1,172 submitted cases remain. The unchanged language-specific scorer then matches the merged prediction one-to-one against the same frozen Main reference. Differently worded repetitions can compete for the same reference and remain in the precision denominator if unmatched. This is a scored merged prediction, distinct from the union of independently matched reference IDs used for cumulative recall.

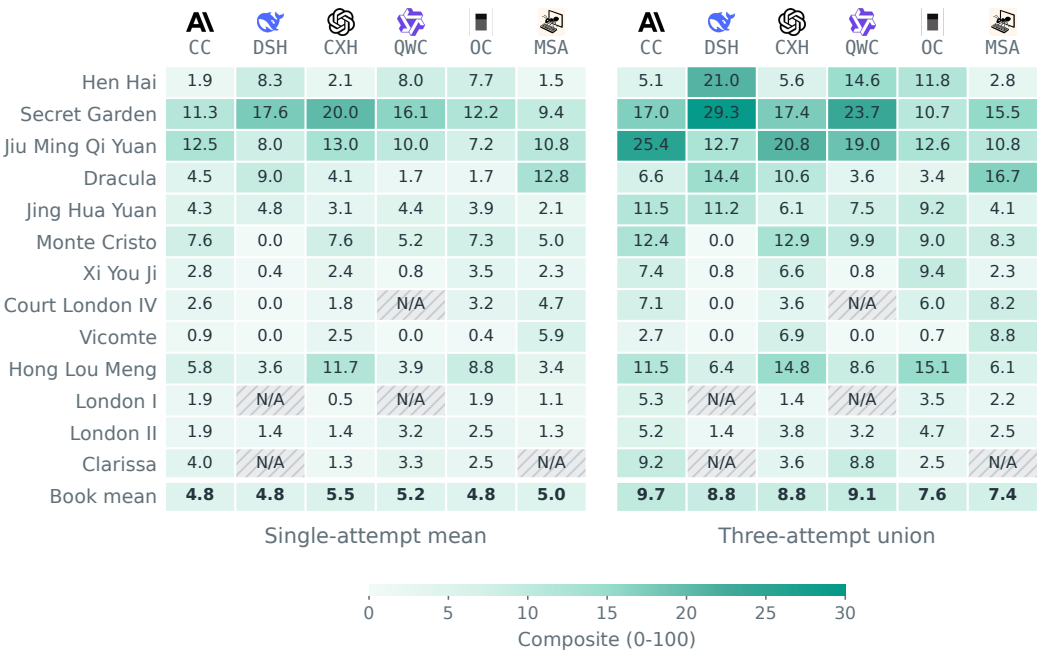


Figure 18: DeepSeek Original Composite: mean defined single-attempt scores (left, reproducing Figure 4) and rescored three-attempt case unions (right). Both panels use the same novel order, runtime order, and color scale. Book means use defined cells; hatched N/A marks unavailable Composite.

The fixed matrix still contains 234 qualified attempts (177 successful, 57 failed). Both panels have 73 defined novel–runtime scores: four cells lack a successful output, and DSH–*Clarissa* has only an empty prediction. All six runtime-level means increase after merging. Union Composite exceeds the single-attempt mean in 61 cells and the best single-attempt score in 44. A single available nonempty output can yield no merging benefit, and repeated expressions can offset some coverage gains through precision. Thus the gain comes from complementary submitted cases under the same scoring rules, rather than averaging scores or selecting the best evidence per reference case.

Output scale. Formal GPT outputs average 25.03 cases (median 22, range 7–53), historical GPT 23.10 (median 23, range 8–62), and DeepSeek 6.62 (median 6, range 0–22); 160/177 successful DeepSeek predictions contain at most ten cases. These are final candidate counts, including unmatched candidates. Correlations use each novel's mean successful output count; DeepSeek pools available successful attempts from all six runtimes within the novel.

Table 25: Output-count scale and robustness. r_N correlates output count with frozen reference count; r_L uses input characters. Full-panel correlations have 13 novels; EN/ZH have eight/five. LOO gives the range after omitting each novel in turn.

Configuration	r_N	r_L	EN r_N	ZH r_N	LOO r_N
GPT formal	**0.850**	**0.606**	**0.922**	0.826	0.808–0.895
GPT historical	0.839	0.512	0.868	**0.850**	0.793–0.873
DeepSeek aggregate	0.202	-0.073	0.153	0.331	-0.038–0.347

DeepSeek runtime-specific r_N values are CC 0.028, DSH 0.313, CXH -0.454, QWC -0.295, OC 0.432, MSA 0.043, on 11–13 supported novels. The weak aggregate correlation does not imply the same association within every runtime. Character length is also language-dependent; the within-language checks reduce that particular mixing concern.

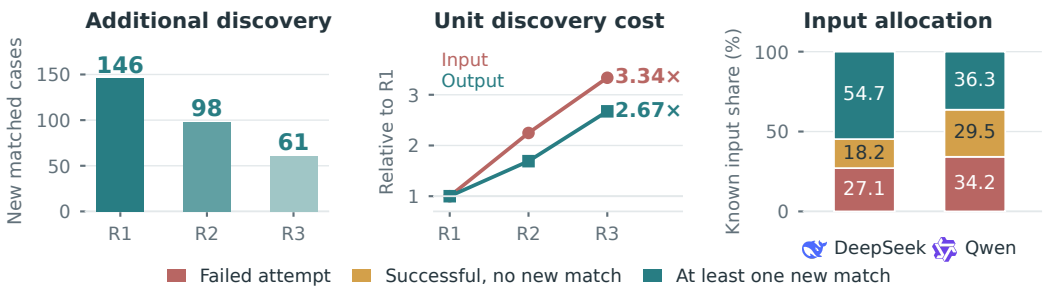


Figure 19: Recorded discovery costs: GPT additions (left), tokens per addition relative to R1 (middle), and DeepSeek/Qwen known input shares (right). Novelty is within runtime across rounds; zero-new usage counts, missing usage is not imputed, and model cohorts are separate.

E.14 Resource Use and the Cost of Additional Discovery

The Original analysis contains 391 qualified attempts: 283 successful and 108 failed. The resource tables additionally include 273 INJ attempts, totaling 664 qualified attempts (496 successful, 168 failed) with distinct condition labels. The 41 targeted retries have separate resource totals in Appendix E.15. Calls count recorded model requests. Input includes cache reads; output includes recorded reasoning and intermediate generation, so it is not prediction length. GPT provides aggregate tokens but no comparable call count. Missing usage is retained field by field; a known call count is not discarded merely because tokens are absent.

Chronological novelty compares each configuration's R1, R2, and R3 within each novel. Every attempt's known usage enters the numerator, including failed and successful zero-new attempts. Ratios divide cohort resource totals by new novel–case counts; zero-new denominators remain undefined. In equal-weight combinations, every selected attempt is also averaged as the final addition. Expected resource use is identical between strategies, not a hard token cap on every individual combination.

Table 26: GPT formal marginal discovery. M and k denote million and thousand tokens. The R3 capacity failure has no known usage; no missing cost is imputed.

Round	New cases	Input M	Output M	Input M/new	Output k/new
R1	**146**	115.354	0.520	0.790	3.56
R2	98	**173.971**	**0.590**	1.775	6.02
R3	61	160.759	0.581	**2.635**	**9.52**

For DeepSeek, symmetric repeated-runtime mean additions are 20.56, 15.61, 12.33. Third-addition known input per new case is 10.52M, versus 9.17M for mixed runtimes.

For Qwen, symmetric repeated-runtime mean additions are 14.00, 10.89, 9.11. Third-addition known input per new case is 15.07M, versus 13.33M for mixed runtimes. Diminishing set-union additions follow symmetric sampling; the measured differences quantify how runtime diversity alleviates overlap. Chronological sequences need not be monotonic (DeepSeek–CC adds 24/19/25 cases; Qwen–QWC adds 14/20/17).

DeepSeek's 57 failed attempts consume 27.10% of known input; its 42 successful zero-new attempts consume another 18.16%. Qwen's corresponding counts are 50 and 22, accounting for 34.18% and 29.48%. Novelty here is additional frozen-reference cumulative case recall within a runtime, not the total value of a prediction.

DeepSeek Original lacks all token usage for three attempts and call counts for two; INJ lacks both for two. Qwen has known tokens for every attempt, but ten QWC attempts contain eleven calls without usage, so the totals can still be incomplete. One GPT Original failure lacks tokens; all 79 GPT attempts lack comparable calls. Cache and reasoning fields are retained where available. Terminal

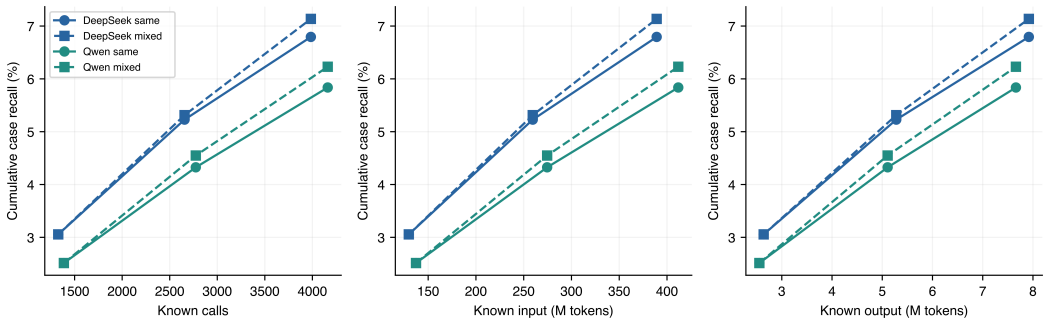


Figure 20: Cumulative case recall against known calls, input, and output. Expected usage is the same for repeated and mixed-runtime strategies at a fixed attempt count.

Table 27: Same-novel, same-round successful resource pairs. Ratios or differences are computed after within-novel averaging, then summarized by their median across novels. Output tokens include the recorded generation process.

A / B	Novels/pairs	Calls ratio	Input ratio	Output ratio	Cases $A - B$	Matches $A - B$
DS DSH / CC	**12**/22	**3.21**	1.50	**10.76**	1.00	-0.50
Qwen DSH / CC	5/9	2.26	1.41	7.86	**4.50**	0.00
Qwen / DS CC	9/**23**	1.28	1.35	1.27	1.00	-1.00
Qwen / DS DSH	4/7	.95	1.06	.86	2.17	**0.75**
Qwen / DS QWC	10/21	1.86	**3.21**	1.06	1.50	**0.75**

Table 28: Original known resource totals. Attempts/successes retain failed attempts. Token totals use millions; input availability gives attempts with known totals.

Condition	Runtime	Attempts/success	Calls	Input M	Output M	Input available
DS Original	CC	**39/38**	2,842	384.738	2.447	38/**39**
DS Original	CXH	**39**/37	4,604	499.307	4.414	**39/39**
DS Original	DSH	**39**/23	**7,239**	**549.594**	**20.353**	38/**39**
DS Original	MSA	**39**/22	2,534	357.826	3.079	38/**39**
DS Original	OC	**39**/32	5,004	439.169	13.723	**39/39**
DS Original	QWC	**39**/25	1,681	104.514	3.491	**39/39**
GPT Original	Codex	40/39	—	450.084	1.692	39/40
Qwen Original	CC	**39**/24	2,762	386.273	2.503	**39/39**
Qwen Original	DSH	**39**/12	**6,803**	**481.553**	**16.804**	**39/39**
Qwen Original	QWC	**39/31**	2,919	367.607	3.675	**39/39**

Table 29: INJ known resource totals. Attempts/successes retain failed attempts. Token totals use millions; input availability gives attempts with known totals.

Condition	Runtime	Attempts/success	Calls	Input M	Output M	Input available
DS INJ	CC	**39/38**	2,757	449.627	2.602	**39/39**
DS INJ	CXH	**39/38**	4,139	432.166	3.990	**39/39**
DS INJ	DSH	**39**/24	**7,668**	**577.387**	**19.599**	**39/39**
DS INJ	MSA	**39**/16	2,488	349.558	2.879	37/**39**
DS INJ	OC	**39**/36	4,778	434.831	13.352	**39/39**
DS INJ	QWC	**39**/22	1,395	78.110	3.438	**39/39**
GPT INJ	Codex	39/39	—	371.940	1.560	39/39

usage does not isolate compaction costs from total input/output. These records support resource-use patterns, not provider-independent monetary prices.

E.15 Targeted 500-Request Retries

The supplementary retry condition targets previously selected budget-limited cells: 14 DS Original attempts (DSH and OC), 12 DS INJ attempts (DSH), and 15 Qwen Original attempts (DSH). There are 41 qualified attempts: 23 successful and 18 failed. These attempts use a 500-normal-request allowance with the existing finalization opportunity; the 300-request matrices, failures, and quality tables remain separate. Of the 23 valid outputs, 19 have matches, three are nonempty with zero matches, and one is empty. All successful predictions have frozen scores; deterministic projection preserves their case counts without repair or dropped cases.

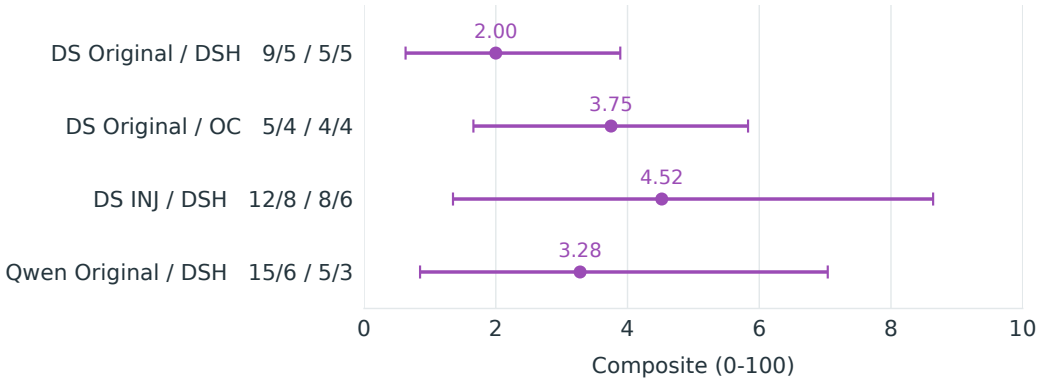


Figure 21: Composite for targeted 500-request retries: book-macro means and 95% book-bootstrap intervals. Row labels give attempts / successes / attempts with defined Composite / supported novels. These selected supports differ from the full 300-request panel.

Figure 21 complements matched and newly discovered case counts with Composite. The empty Qwen prediction has undefined Composite; the three nonempty zero-match outputs receive zero. Recovery and additional discoveries do not by themselves establish higher Composite or a causal budget effect.

Table 30: Targeted retries and marginal discoveries. Qualified attempts include successes and failures. Predicted cases and matches count occurrences across successful predictions, including repeated cases across attempts. New IDs are relative to the same-model, same-text union across all original runtimes; the text conditions are not pooled.

Condition	Attempts	Success	Failed	Predicted cases	Matches	New IDs
DS Original	14	**9**	5	**96**	**20**	**10**
DS INJ	12	8	4	80	18	7
Qwen Original	**15**	6	**9**	68	8	5

The revised matched unions grow from 161 to 171 cases for DS Original, 164 to 171 for DS INJ, and 74 to 79 for Qwen. Within DS Original, DSH adds four cases to the prior six-runtime union and OC adds six; relative to their own earlier outputs they add ten and nine, respectively. Recovery therefore includes both new discoveries and cases already found by another runtime.

Input includes cache reads and output includes reasoning and intermediate generation. DS INJ lacks usage for one failed attempt with a corrupted session log. Four OC metadata records report zero usage, but their existing per-call logs record a total of 58.973M input and 1.961M output; the table uses these logs. The new-case denominator deduplicates all additions in the retry batch. The no-new input share compares each attempt to the fixed prior pool, without allocating overlap among retry

Table 31: Known resource use across all qualified retry attempts, including failures. Token totals and input/new use millions. No-new input includes failures and successes whose matches are all in the fixed prior pool.

Condition	Calls	Input M	Output M	Input/new	No-new input	Available
DS Original	**4,418**	**343.876**	**10.812**	34.388	73.66%	14/14
DS INJ	3,604	313.090	10.456	44.727	48.27%	11/12
Qwen Original	3,215	242.179	8.221	**48.436**	**89.16%**	**15/15**

attempts by order. Earlier 300-request failures separately consumed 338.914M/242.680M/277.853M known input for DS Original/INJ and Qwen; those costs are not included in this table.

Fifteen successful trajectories stop within 300 calls and eight exceed 300, including one 301-call finalization already possible under the original allowance. These retries do not include concurrent 300-request reruns of the same cells, so recovery is not a controlled estimate of the effect of 200 extra calls. Qwen's five new cases all come from one Jing Hua Yuan attempt using 439 calls and 26.264M input. A Clarissa attempt uses 487 calls, 52.010M input, and 2.013M output yet reports no cases. Two DS OC attempts on Vicomte and Hong Lou Meng use 111 calls each and 12.234M combined input, adding six cases. Request count and discovery yield are therefore distinct outcomes.

DS failures comprise four request-limit, three memory, one scratch-permission, and one session-log failure. Qwen failures comprise one timeout, two HTTP 400 errors, two ordinary single-queue HTTP 429 errors, and four missing valid JSON predictions. Four separate Qwen executions were excluded: two initial concurrent HTTP 429 interruptions and two author-paused executions. Three record 279 calls, 17.180M input, and 0.414M output; one pause has no usage. Their costs are disclosed separately from the 15 qualified attempts. The retries add seven frozen explained-to-unresolved INJ target matches, six predicted explained; the one predicted unresolved inconsistency lacks a same-cell successful Original output, adding no observed complete paired transition.

E.16 Separating Failed Attempts, Zero Matches, and Evidence Scores

All 664 qualified attempts are classified into four exhaustive attempt outcomes. The 496 successful predictions contain 4,622 predicted case occurrences and 1,657 matches; occurrences repeat cases across attempts and are not independent novel observations.

Table 32: Attempt outcomes and prediction matching. "Zero" means nonempty output with no match; empty successful predictions are a separate column.

Condition	Attempts	Failed	Empty	Zero	Has match	Matches/ predicted cases
DS Original	**234**	57	**2**	**29**	146	370/1,172
Qwen Original	117	50	**2**	14	51	126/572
GPT Original	40	1	0	0	39	**470**/976
DS INJ	**234**	**60**	**2**	18	**154**	372/**1,221**
GPT INJ	39	0	0	0	39	319/681
Total	664	168	6	61	429	1,657/4,622

Every nonempty zero-match attempt has zero eligible candidate edges. Its case precision and Composite are zero, whereas matched-label accuracy is unavailable. Empty predictions have zero case recall/Q and unavailable case precision/Composite; failed attempts have no evaluation score. These distinctions avoid merging unavailable measurements with valid numerical zeros.

For a successful attempt, let G be its reference count and M its match count. The frozen mean case-quality score has the exact decomposition

$$1 - Q = \frac{G - M}{G} + \frac{.20(M - \sum L)}{G} + \frac{.25(M - \sum A)}{G} + \frac{.30(M - \sum E)}{G}. \tag{4}$$

The sums range over matched cases. Unmatched-reference loss accounts for 98.80%–99.08% of the DeepSeek runtimes' mean total quality shortfall, while direct matched-label loss is at most 0.022

Table 33: Pooled matched-occurrence diagnostics. Label and strict endpoint percentages use all matches; counts of maximum Explanation-Path F1 use correctly labeled explained matches.

Condition	Matches	Correct label	Both endpoints strict	Maximum Explanation-Path F1
DS Original	370	367/370 (99.19%)	79/370 (21.35%)	56/334
Qwen Original	126	124/126 (98.41%)	30/126 (23.81%)	16/106
GPT Original	**470**	**468/470** (99.57%)	**128/470** (27.23%)	**82/461**
DS INJ	372	270/372 (72.58%)	95/372 (25.54%)	16/210
GPT INJ	319	219/319 (68.65%)	85/319 (26.65%)	25/188

percentage points. This is an algebraic description of the score, not a decomposition of internal model failures. Labels gate E, and human judgments distinguish semantic adequacy from lexical alignment.

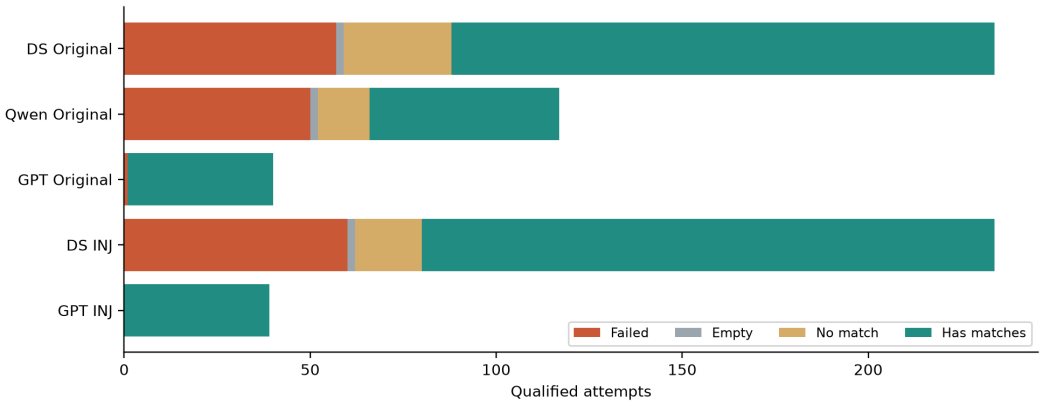


Figure 22: The four attempt outcomes by model and text condition.

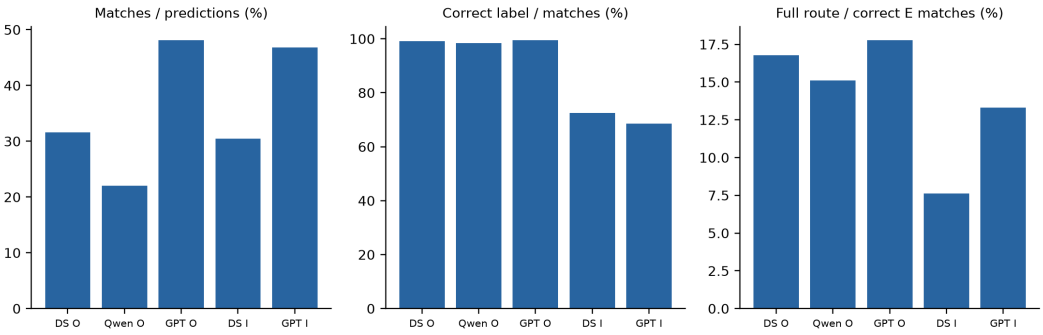


Figure 23: Matching, conditional labels, and evidence alignment use their own denominators; they are not successive independent stages of one funnel.

E.17 Unmatched-Output, Reference, and Follow-up Human Audits

The additional audits complement the original 96-pair blinded challenge study. All numbers in this subsection describe the pre-repair sampling populations and returned judgments. Revised matching does not change their samples or weights; the output-count bound in Appendix E.1 separately uses the current reference. Across these studies, AI tools helped organize review materials and locate source passages; human reviewers made the substantive judgments themselves. Sampling, blinding, prior exposure, and adjudication conditions are reported separately below.

Table 34: Distinct human-review studies. Reported agreement precedes author adjudication; final decisions and original returns are retained separately.

Study	Sampling object	Review process
Scorer challenge	96 pairs / 81 distinct references	Two independent blinded reviewers; 192 pair judgments; no final consensus substituted for agreement.
Unmatched (R1)	216 of 803 unmatched DeepSeek predicted cases	Two reviewers each inspect the prediction, same-novel Main catalog, and source text; 432 returns. Disagreements retained.
Reference (R2-D1)	78 of 1,033 Main cases, stratified by novel and frozen label	Two reviewers, followed by author resolution of disputed fields. Explicit labels and source identities hidden; given routes remain visible.
Edited text (P6)	17 selected targets / 23 predicted cases / 30 quotations	Initial author review, then two later reviews and final author adjudication; Original-route exposure recorded.
Targeted follow-ups (P11/P15)	Ten questions / 18 predicted cases; later eight groups / 15 predicted cases	Informed author review of earlier returns and historical records, with full-text checks; original probability-sample statistics retained.
Accepted matches (P16)	202 of 369 DS Original pairs in 114 strata	Two independent returns with prior exposure recorded; 29 author decisions and 173 unchanged agreed items form the final sample.

Unmatched-output probability sample. The 177 successful DeepSeek Original predictions contain 1,172 predicted cases, including 803 unmatched occurrences. We sample $n_h = \min(3, N_h)$ without replacement from each of 73 nonempty runtime–novel strata, for 216 items. The issued same-novel catalog is Main (1,033 cases), not the 1,063-case Full supplement. Each item receives weight N_h/n_h. Both reviewers supply their own item-level decisions and participation declarations; neither return is silently chosen when the two disagree.

For a fixed binary category, the estimate is $\hat{p} = \sum_h (N_h/N)\bar{y}_h$. The prespecified conservative pointwise 95% Hoeffding half-width is

$$\epsilon = \sqrt{\frac{\log 40}{2} \sum_{h:0<n_h<N_h} \frac{(N_h/N)^2}{n_h}}, \tag{5}$$

with endpoints clipped to $[0, 1]$; census strata contribute no sampling error. For this sample, $N = 803$ and $\epsilon = 10.66$ percentage points. These intervals reflect the sampling design, not reviewer error or simultaneous coverage of all categories.

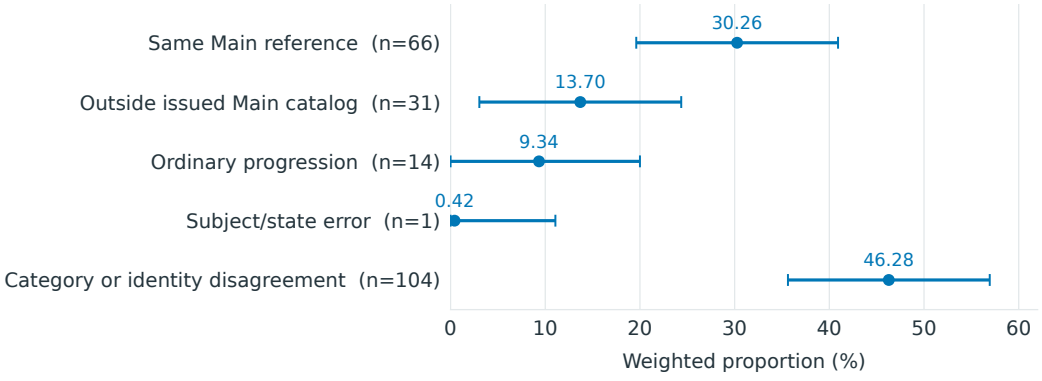


Figure 24: Unmatched-output audit: design-weighted proportions and pointwise 95% bounds over 803 occurrences. Row labels give sample counts; category or identity disagreements remain separate.

Individually, reviewer A assigns 94 items to same-reference (weighted 41.10% [30.44, 51.75]) and B assigns 81 (35.74% [25.08, 46.40]). Category agreement is 113/216 ($\kappa = .3631$), with one of these items having different reference IDs; identity-consistent category agreement is therefore 112/216. Agreement on the multicategory qualification field is 139/216 ($\kappa = .2953$), and explanatory sufficiency agreement is 121/216 ($\kappa = .3398$). Binary qualification endorsement is distinguished in Appendix F.5.

The 66 agreed same-reference occurrences span 61 attempts, 48 distinct novel–reference IDs, and all six runtimes and 13 novels. No identified reference is covered by another output from its attempt. Of these occurrences, 62 have agreed endpoint support, 39 have agreed sufficient explanations, and 37 satisfy qualification, identity, both endpoints, and explanation together. These 37 span 12 novels and all six runtimes. They are sample counts, not an estimated unique-case correction to recall.

How supported expressions lose matching credit. For Athos (U032/VB011), the before quote is a continuous nine-word part of a 90-word reference anchor, and the after quote contains 15 words within an 88-word anchor. Their alignment scores are .275 and .378. The model explains the reversal through Athos's accusation of royal deception; the reference quotes the king's intervention in Raoul's intended marriage. Both reviewers accept the explanation, but route alignment is .061 and combined match score .253. Other cases use different state passages (U037/JMQY004), divide endpoints differently (U061/XYJ025), or name additional participants (U128/JHY009). Among all 66 pairs, combined-threshold, state-guard, and subject-guard failures occur in 61, 15, and four cases, with overlap; all fail before assignment.

Reference probability sample and adjudication. We sample six cases per novel: up to two unresolved, with the remainder explained, using sorted IDs and seed 20260921. The result is 59 explained and 19 unresolved cases; weights preserve the frozen 988/45 class totals. Two reviewers check endpoints, qualification, labels, supplied routes, and duplicates. Both report no prior item exposure for this part. Raw label agreement is 75/78 ($\kappa = .8897$); duplicate agreement is 77/78 ($\kappa = 0$ with one reviewer using only one category). Final author resolution retains all 78 qualifications and endpoints, changes three frozen unresolved labels to explained, and identifies one duplicate/merge issue. The weighted label disagreement is 4.5/1,033 (.44% [0.00, 19.36]); the duplicate-issue estimate is 16.5/1,033 (1.60% [0.00, 20.53]). The same pointwise-bound formula gives 18.93-point half-width for $N = 1,033$. All 65 supplied routes for the 59 frozen explained cases receive sufficient judgments from both reviewers; a route is not an independent case. The three newly explained cases have additional routes found during review and are not counted among those 65. This checks sampled reference entries and supplied routes, with no prior item exposure reported for that study. Reviewers nevertheless saw the proposed cases and routes: the study assesses their support, not independent open-ended reconstruction of the reference. Its agreement cannot substitute for qualification reliability on novel model proposals or establish the neutrality of the case-discovery process.

Edited-text review and final decisions. The P6 reviewers each assess the 17 targets, 23 predicted cases, and 30 quotations. A reports Original-route exposure for all 17 targets; B reports exposure for two, caused by overlap with the Reference sample. Both declare no access to the other's answers or earlier author decisions. Pre-adjudication output-sufficiency agreement is 21/23 ($\kappa = .8296$), and quotation agreement is 29/30 ($\kappa = .9346$). The author resolves disputed fields and conflicts with two earlier decisions, rather than supplying a third independent full-sample judgment. Final target results are 15 residual routes, one route not found, and one ineligible target; output results are 11 sufficient, ten insufficient, and two inapplicable. Full item results and the exclusion analysis appear in Appendix E.8.

Targeted follow-up on identity and qualification. The 31 agreed Main-external outputs form 24 endpoint-based correspondence units; these units are not 24 validated new cases. Historical comparison identifies nine outputs related to excluded cases and three related to an existing identity boundary. Ten first-author questions then cover 18 predicted cases, of which 14 come from the Main-external set. Among those 14, eight are ordinary or otherwise ineligible reversals, three have unresolved identity evidence, one is merged with an existing case, and two involve endpoint repair or reframing. Seventeen Main-external outputs remain unadjudicated. The other four queried outputs concern three identity disagreements and one explanation judgment; the original 104 identity/category disagreements retain 101 unresolved items.

Seven of the 18 queried outputs qualify; three supply sufficient model explanations and four supply only partial explanations. Baoyu's new before passage expresses attachment to Daiyu, aligning with later detachment; the old before expressed a general view of women. Diana's Full supplement instead used Warrington's mistaken belief before unveiling as evidence of an actual scar, although the unveiling establishes that no scar remains. The healing and test of affection must be distinguished from the mistaken belief. These endpoint problems occur in predictions and reference descriptions alike.

For the clear-identity cases, both crying-scene outputs in The Secret Garden give Colin's identity and the instruction that servants must keep him secret. Tang Guichen's return cites the father's letter and pavilion but omits the uncle's promise to return with her; Xiren's two explanatory evidence sentences omit dismissal, arranged marriage, and the reason for abandoning suicide. An audit-record statement that Mary changed her view of the moor after experiencing it directly is corrected against the text: she learned through others' accounts after a limited nighttime view. The original returned document is retained, while the analysis uses the corrected fact.

These results link identity, qualification, expression, and explanation without treating all unmatched output as reference omission. Persistent inferential disagreement has precedents (Pavlick & Kwiatkowski, 2019); our concrete checks determine which problem each selected example exhibits rather than labeling every disagreement irreducible.

Additional targeted omissions. A later priority review resolves 15 predicted cases, after identifying one further item as already present in +DS/Full. Eight of the 15 qualify and seven do not; the eight qualified outputs provide two sufficient, five partial, and one insufficient original explanation. Four constituted independent cases absent from both Main and Full at review time: U108 in Hen Hai (upright conduct versus selling the Xuande censer), U174 in Jiu Ming Qi Yuan (claimed direct observation versus hearing through a window), and U098/U100 in London II (the family's understood marriage history versus a concealed legal marriage; denial versus admission of the same robber's identity). The original explanations for U108 and U100 are sufficient. The other four qualified outputs include three existing identities and one additional expression supporting pooled-discovery counts. These informed, selected judgments do not replace the original R1 probabilities or imply that all remaining disputed outputs are omissions. The four now enter Full only under HH028, JMQY050, and London II ML159/ML160. Their five credited matches are reported in Appendix E.7; the historical R1 probabilities and original evidence judgments remain unchanged.

E.17.1 Probability Audit of Accepted Matches

The population is all 369 accepted prediction–reference occurrences from 177 successful DS Original attempts, not the 1,033 reference cases. The other 803 predicted cases form the separate R1 population. Sampling strata are runtime, novel, and assignment-score band: $Z < 0.20$ versus $Z \geq 0.20$, where $Z = (\widetilde{C} - \tau)/(1 - \tau)$. This partitions 13 novels and 69 runtime–novel combinations into 114 nonempty strata. Allocation starts with $\min(2, N_h)$ per stratum, then adds items to reduce the prespecified conservative error bound until its half-width is below ten points. The final sample contains 202 pairs, including 56 census strata, drawn uniformly without replacement within strata using seed 20260923. Inclusion probabilities are n_h/N_h and weights N_h/n_h. Low/high bands contribute 89/113 sampled pairs from populations of 141/228; no review answer enters allocation.

Review proceeds from output endpoints, to paired reference endpoints, to explanations and other outputs from the same attempt. Runtime, model, automatic score, original IDs, labels, and reference adjudication reasons are masked; the accepted-batch attribute is not hidden. Both reviewers declare independent completion without the other's answers or the internal decoding. Reviewer A had previously reviewed reference/edited-text material and marks prior exposure yes for all 202 items; B participated in R1 and marks it uncertain throughout. Of the sampled items, 158 had been visible as other outputs on R1 pages. Visibility does not prove inspection, but there is no clearly unexposed reviewer subgroup. This is distinct from the original blinded challenge study.

Qualification and endpoint judgments concentrate heavily on positive categories; near-zero endpoint κ coexists with high raw agreement because the few negative judgments occur on different items. Twenty-eight items disagree on at least one primary dimension. They form 24 author question

Table 35: Initial accepted-sample agreement, before author decisions. Weighted percentages use the sampling design; κ describes the unweighted sample.

Dimension	Agreed/sample	Weighted agreement	Cohen κ
Identity	196/**202**	97.15	.7144
Output qualification	196/**202**	97.02	.6133
Reference qualification	**201**/**202**	**99.73**	.0000
Before support	196/**202**	97.70	-.0151
After support	198/**202**	97.56	-.0075
Explanation sufficiency	187/**202**	92.28	**.7782**
Within-attempt duplication	**201**/**202**	99.46	.0000

groups; one supplementary identity item brings author decisions to 29. The other 173 use unchanged agreement on all seven dimensions. Final decisions never replace the initial agreement statistics.

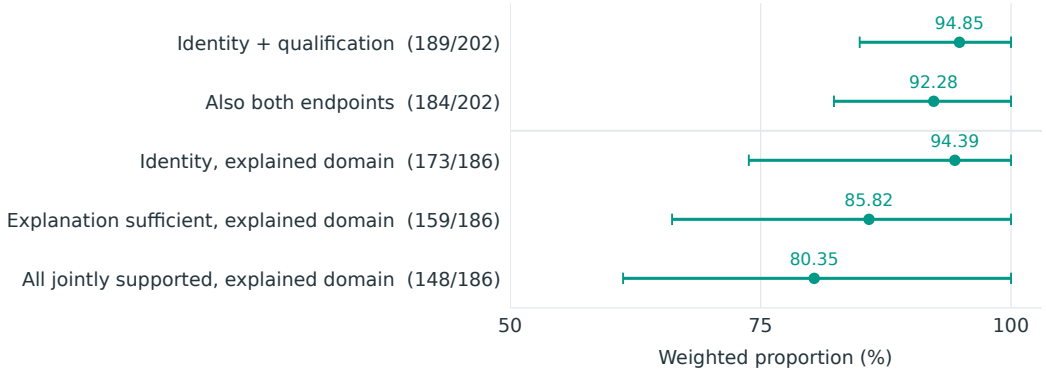


Figure 25: Accepted-match audit: design-weighted proportions and 95% sampling intervals. Row labels give supported/sample counts. The first two rows use all accepted occurrences; the last three use the frozen explained domain. Joint support requires identity, qualification, both endpoints, and sufficient explanatory evidence.

Overall identity has weighted numerator 350 and denominator 369. The 336 frozen explained occurrences contribute 186 sampled pairs; their estimated weighted domain size is 338.5, and complete support has numerator 272. Domain estimates are ratios of weighted totals, not unweighted sample proportions or division by 336. Overall binary estimates use the same stratified Hoeffding form as R1, now with $N = 369$ and half-width 9.96 points. Explained-domain intervals combine separate 97.5% bounds for numerator and denominator using Bonferroni, yielding conservative 95% ratio bounds; census strata contribute no sampling uncertainty. Finite-population corrected standard errors provide supplementary diagnostics. These bounds do not include reviewer error and are not powered to establish six-runtime rankings.

Final judgments of explanatory sufficiency comprise 159 sufficient, 14 partial, six insufficient, and seven inapplicable because output qualification fails. The 16 sampled unresolved pairs are excluded from this explanation-domain calculation. Three valid, fully explained outputs match the wrong reference: M127 should map to G011 rather than G015, M122 to JMQY010 rather than JMQY026, and M047 to XYJ031 rather than XYJ057. They remain errors of the original automatic pairing. Ten other pairs cannot be mapped one-to-one: seven outputs are ineligible, while three overlap only part of the reference endpoint or event group. One of these, M055, validly contrasts a true fan's fire-extinguishing effect with the observed increase in flames; XYJ140's before only states possession of a fan. The case is valid but the reference endpoint is insufficient. M062 and another same-attempt output duplicate the same impersonation case. The combined audit case sets and their comparison with single-attempt output counts are reported in Appendix E.1.

F MEASUREMENT VALIDITY AND TASK BOUNDARIES

This appendix connects the original human judgments to Explanation-Path F1, matching decisions, and task boundaries. These are retrospective analyses of existing outputs and audit materials, not new model attempts or independent human replications. The unmatched and matched-pair audits retain their historical 1,033-case reference; the full alternative-scoring panel uses the repaired 1,032-case Main reference. Their populations and denominators are kept separate. Targeted source-text diagnostics by an AI assistant are identified separately from the original human judgments and author adjudications.

The audits address three different questions: whether a prediction identifies the same qualified case, whether its cited evidence is sufficient, and how matching selection affects attempt-level Composite. Agreement on one question does not validate the others. We present missed matches and aggregate sensitivity first, followed by the conditional explanation-score association and the existing qualification decisions. Sampling and reviewer exposure for each study are summarized in Table 34.

F.1 MISSED MATCHES, FIXED IDENTITIES, AND LABEL SELECTION

Both reviewers identify 66 unmatched predicted cases as qualified discoveries of an existing reference case; 62 also have supported endpoints, 39 sufficient explanations, and 37 both. All 66 fail the original matching gates before global assignment; none loses to another predicted case for the same reference. The failed gates overlap: combined score 61, state-side support 15, and target compatibility four. All 37 fully supported predicted cases agree with their frozen labels. Label error and inadequate explanation therefore do not account for these 37 missed matches. These omissions establish a concrete loss of valid credit. They do not supply a runtime-level correction to recall or Composite: the sample covers unmatched occurrences, judgments can disagree, and corrected matching may reassign other pairs. The output-count ceiling in Appendix E.1 answers the narrower question of whether perfect matching of the existing outputs could yield broad coverage. It does not dismiss the measurement errors documented here.

Table 36: Label agreement with the frozen reference under fixed human identity and qualification judgments. Matched and unmatched samples have different frames, weights, and exposure conditions; their percentages cannot be pooled.

Historical domain	n	Agreeing n	Weighted agreement
Rejected: both reviewers endorse identity and qualification	66	64	97.94%
Accepted: both endorse identity and both objects' qualification	187	185	99.42%
Accepted: adjudication endorses identity and both objects' qualification	**189**	**187**	**99.43%**

Labels were hidden in the identity audit, whose judgments are not independent human label ground truth. The rejected domain includes only three frozen unresolved cases. The original matcher does not read the predicted label directly, but uses its associated evidence structure. For a predicted unresolved inconsistency, the schema requires empty explanatory evidence. Against a reference case labeled explained this gives $C = .2T + .4S$, so the combined gate requires $S > 1 - .5T$ in addition to the other gates. Such predicted cases can still match: M057 and M185 did so despite label disagreement. Among 173 adjudicated qualified explained pairs, 96 (estimated mass 168 of 319.5) need the observed route contribution to pass the combined gate. This is arithmetic on fixed scores, not an intervention or causal effect of labels. High conditional label agreement does not establish independent label-classification ability.

F.2 MATCHING STRUCTURE, SAMPLE SELECTION, AND COMPOSITE

Fixed alternative rules. We compare the default protocol in Appendix H with two separate offline analyses. Relaxed matching halves all four gates: explained normalized score $> .20$, unresolved normalized score $> .25$, target compatibility $T \geq .25$, and $\max(B, H) \geq .15$. It retains the combined-score formula, reference-label normalization, and assignment-weight formula, whose

values change with the thresholds. Lower gates therefore need not preserve the original global assignment.

The endpoint-constrained rule removes explanatory evidence from matching. With $S = \sqrt{BH}$ and $I = (T + 2S)/3$, it requires

$$T \geq .50, \qquad B \geq .30, \qquad H \geq .30, \qquad I > .40, \tag{6}$$

and assigns weight $(I - .40)/.60$ to an eligible edge. Labels, explanatory evidence, reasons, and claims are not identity features. The lexical kernels, tokenizers, global one-to-one assignment algorithm, and post-match quality formula remain unchanged. Reassignment can nevertheless change the selected pair's quality. This jointly changes the endpoint guard, explanation compensation, and label-dependent threshold. It does not isolate the effect of one rule or prove semantic identity through lexical endpoints.

Table 37: Selection of the original or human-supported target under alternative rules, on historical audit subsets. Counts are unweighted and are not pooled across the separate matched-pair and unmatched-case sampling frames.

Subset	n	Default	Relaxed	Endpoint
Unmatched: both endorse identity and qualification	66	0	**49**	8
Unmatched: additionally both support endpoints and explanation	37	0	**25**	5
Accepted: both support identity, qualification, endpoints, and explanation	**146**	**146**	144	88

An AI assistant separately audited 29 targeted identity conflicts against the source texts, reference entries, existing human reviews, and author adjudications. These are diagnostic assistant judgments, not a new independent human annotation round. Under these judgments, the endpoint rule blocks all 24 relaxed targets classified as incorrect, but preserves only two of six default pairs classified as correct. Four relaxed targets remain boundary cases and one is a supported reassignment; the endpoint rule ultimately selects three supported targets and leaves 26 predicted cases unmatched. These materials were used in diagnosis and development, so the check is retrospective, not independent held-out validation. The targeted counts do not estimate the error rate of all relaxed matches; other newly selected pairs are not certified by passing the rule.

Where the original A1 sample goes. Keeping each original pairing and its judgments fixed, the endpoint rule retains 121/202, rejects 80, and reassigns one. Within the adjudicated explanation-validity domain the counts are 108 retained, 65 rejected, and zero reassigned (Table 38). Evaluating the same fixed pair preserves its original before/after scores, raw Explanation-Path F1, label-gated evidence, and case-quality score. An explanation term set to zero inside the matching edge means "unused for identity," not zero explanation quality.

Table 38: Original adjudicated A1 groups under the endpoint rule. Retention is design-weighted; explanation means remain scores of original pairs, including rejected pairs. None is reassigned in this domain.

Sufficiency	n	Retained	Rejected	Retention	Original mean	Retained mean	Rejected mean
Sufficient	**153**	**89**	**64**	60.3%	**.530**	**.457**	**.641**
Partial	14	13	1	94.6%	.214	.216	.174
Insufficient	6	6	0	**100.0**%	.024	.024	—

For sufficient explanations, retention has approximate 95% interval [53.6%, 66.9%] and conservative interval [40.3%, 86.9%]. Conservative intervals for the partial and insufficient groups are both [0%, 100%]. Sufficient-group retention is 61.2% under reviewer A and 61.4% under B. The retained groups preserve the descriptive sufficient > partial > insufficient ordering under all three judgments; this selected subset does not validate new matches. The 153 adjudicated sufficient explanations do not have the same domain as the 146 pairs jointly supported by both reviewers on identity, qualification, endpoints, and explanation (Table 37).

All 65 rejected adjudicated-domain pairs have an endpoint score below .30, including 64 sufficient explanations. Sufficient retained/rejected pairs have weighted geometric endpoint means .680/.224 and mean Explanation-Path F1 values .457/.641. Their mean case-quality scores are .757/.698, with label credit 1 in both groups: stronger endpoint alignment offsets lower Explanation-Path F1 in Equation 2. The original matcher can compensate for weak endpoints through explanation overlap; the endpoint rule instead requires each side to pass. Neither these score patterns nor a failed lexical gate establishes that every rejection stems from a short quote or a reference defect. The sole reassignment, M047 from XYJ057 to XYJ031, agrees with the prior author resolution; it is in A's domain but outside B's and adjudication's. It remains a reassignment, not retention, and old-pair quality is not transferred to the new target.

Full-panel Composite sensitivity. The panel contains 391 eligible attempts: 283 successes and 108 failures (DeepSeek: $234 = 177 + 57$; Qwen: $117 = 67 + 50$; GPT: $40 = 39 + 1$). We rescore the 283 existing successful predictions without model calls or retries. All three rules use the same current 1,032-case Main reference and unchanged post-match scoring formulas. Composite averages calculable scores from successful attempts within each novel, then across available novels; failures and undefined scores are not imputed as zero. DeepSeek Harness has successful predictions on 12 DeepSeek and seven Qwen novels, but Composite is calculable on only 11 and six, respectively.

Table 39: Book-macro Composite (0–100) under three matching rules. "Novels" counts calculable support. Different novel sets and execution conditions preclude treating these rows as a fair cross-model or cross-runtime ranking.

Model	Runtime	Novels	Default	Relaxed	Endpoint
DeepSeek	Claude Code	**13**	4.7633	**9.1616**	2.8939
DeepSeek	Codex Harness	**13**	5.5009	**10.4553**	4.0590
DeepSeek	DeepSeek Harness	11	4.8346	**9.2054**	3.4336
DeepSeek	mini-swe-agent	12	5.0373	**10.5427**	2.6203
DeepSeek	OpenCode	**13**	4.8239	**9.4553**	3.6500
DeepSeek	Qwen Code	11	5.1555	**9.2214**	4.0129
Qwen	Claude Code	10	5.3704	**10.4267**	3.0610
Qwen	DeepSeek Harness	6	5.8540	**13.9435**	2.9672
Qwen	Qwen Code	12	6.3521	**11.9050**	4.6646
GPT	Codex native	**13**	19.4434	**29.4690**	12.2646

Matched outputs change from 370 to 765 to 255 for DeepSeek, 126 to 287 to 83 for Qwen, and 470 to 777 to 306 for GPT (default, relaxed, endpoint). Relaxed matching increases every reported Composite mean; the endpoint rule decreases every mean. These directions do not establish greater validity: the audits show both erroneous credit and rejection of supported pairs. Unchanged post-match scoring does not imply unchanged system-level measurement because matching selects the population receiving credit. The default results are therefore retained, with alternative rules reported as structural and sensitivity diagnostics rather than human-corrected scores or a replacement leaderboard. Retaining the prespecified default fixes the reported protocol; it does not establish that its thresholds are semantically optimal. No rule is selected because it produces higher scores. The human subsets expose false credit and missed credit, but do not provide an independent calibration of whole-system Composite or its ordering across runtimes. The completion and output-count findings therefore have a different evidential basis from exact score differences.

F.3 Conditional Association with Explanation-Path F1

Population and domain. All 202 sampled matched pairs were linked to their original predictions, scores, issued review materials, two reviewer returns, and existing author adjudications. The frame contains 369 matched pairs in 114 runtime–novel–matching-score strata (Appendix E.17.1). We retain inverse-inclusion weights $w_i = N_h/n_h$. The explanation-validity domain requires a frozen explained label, human-confirmed case identity with the original reference, and qualification of both prediction and reference. It does not additionally condition on both endpoints being judged supported. The domain is defined separately for reviewer A, reviewer B, and adjudication; their judgments are

not pooled. Frozen unresolved cases are excluded, so their permitted empty explanatory evidence sets are not treated as explanation failures.

The three domains contain 175, 173, and 173 sampled pairs, with estimated population masses 320.0, 319.5, and 319.5. The full sample contains 186 frozen explained pairs, with estimated mass 338.5 rather than the known frame count of 336: labels were not sampling strata, and domain denominators are not recalibrated to that count. Reviewer A reported prior exposure to all 202 items; reviewer B marked item-level exposure uncertain. Of the sampled predicted cases, 158 had previously been displayed as other predicted cases from the same attempt in the unmatched audit. This establishes availability, not actual reading. Author adjudication is not a third independent review.

Estimation and uncertainty. For a domain and sufficiency group indicated by d_{ig} and an original score $z_i \in [0, 1]$, we estimate

$$\widehat{D}_g = \sum_i w_i d_{ig}, \qquad \widehat{T}_g = \sum_i w_i d_{ig} z_i, \qquad \widehat{\mu}_g = \widehat{T}_g / \widehat{D}_g. \tag{7}$$

The main score is raw Explanation-Path F1 (Equation 1), before the label gate used in the case-quality score. Linearized 95% intervals use $e_i = d_{ig}(z_i - \widehat{\mu}_g)/\widehat{D}_g$ and $\widehat{V} = \sum_h N_h^2(1 - n_h/N_h)s_{e,h}^2/n_h$, with estimate $\pm 1.96\sqrt{\widehat{V}}$ clipped to the parameter range. Out-of-domain samples remain in their original strata with zero influence; census strata have no sampling variance. Contrasts use differences of influence values, retaining covariance under the shared design.

We also report conservative finite-sample intervals. For stratified sampling without replacement, Hoeffding bounds jointly bound the numerator and denominator totals with radius

$$a_K = \sqrt{\frac{\log(2K/.05)}{2} \sum_{h \in \mathcal{H}_*} \frac{N_h^2}{n_h}}, \tag{8}$$

where $\mathcal{H}_*$ contains noncensus strata whose known labels permit domain membership. A mean uses $K = 2$ bounded totals and a contrast $K = 4$. Ratio bounds combine $(\widehat{T}_g - a_K)/(\widehat{D}_g + a_K)$ and $(\widehat{T}_g + a_K)/(\widehat{D}_g - a_K)$, clipped to $[0, 1]$; a nonpositive denominator lower bound gives upper bound 1. Contrasts combine the joint ratio bounds without assuming independent groups. Coverage is pointwise, not simultaneous over all reported quantities. These intervals cover sampling uncertainty in the fixed historical frame, not human error or generalization to new novels. A sparse group with zero estimated variance is not assigned a zero-width approximate interval.

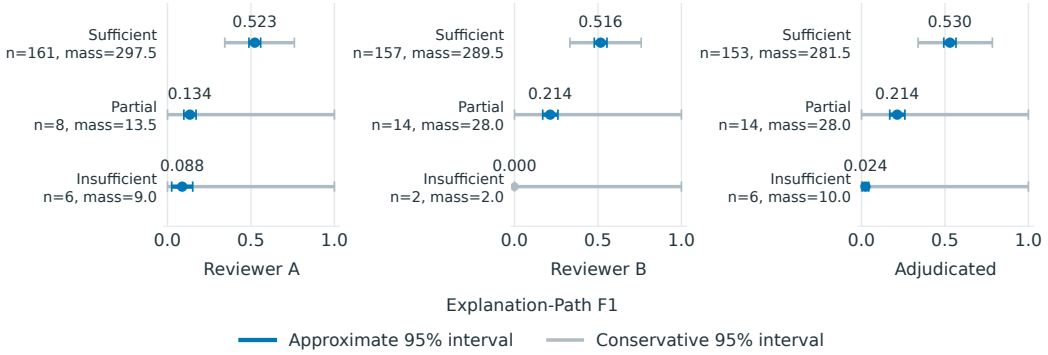


Figure 26: Explanation-Path F1 by human sufficiency judgment: design-weighted means with approximate and conservative 95% intervals. Row labels give sample counts and estimated population masses. The approximate interval is omitted for reviewer B's sparse insufficient group.

The sufficient-minus-noncomplete contrast is .408 for A, .316 for B, and .366 for adjudication. Their approximate intervals are [.363, .452], [.264, .368], and [.315, .416]; conservative intervals are [-.669, .781], [-.680, .781], and [-.673, .806]. All conservative contrast intervals include zero. The

adjudicated sufficient-minus-partial contrast is .316 (approximate interval [.261, .370]; conservative interval [-.673, .806]), so the descriptive ordering does not depend only on the two empty explanatory evidence sets in the insufficient group. Sufficient explanations nevertheless span $[0, 1]$, with weighted median .500 and interquartile range [.313, .703] (Figure 27). The group means support a conditional association, not a semantic probability or an individual decision threshold.

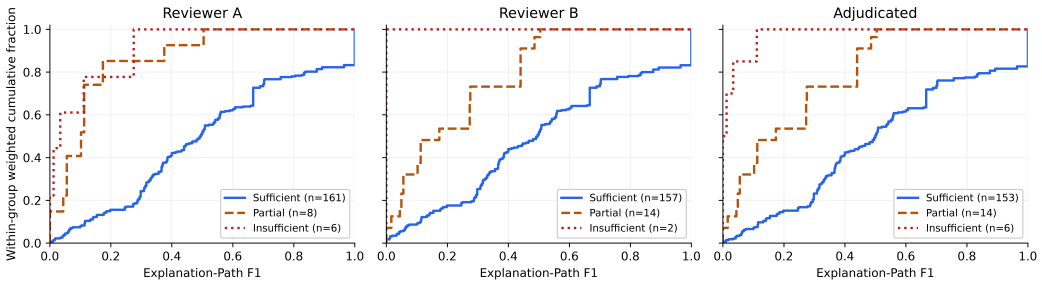


Figure 27: Original design-weighted empirical score distributions within each reviewer's explanation-validity domain. Curves describe scores conditional on existing sufficiency judgments; they are not accuracy curves or simultaneous confidence bands.

Individual failures and case-quality scores. M037/JHY044 is sufficient under both reviewers and adjudication but has raw Explanation-Path F1 of .000. Its explanatory evidence does not align with the annotated route, whose requirements span more than the submitted evidence. This identifies an expression/scope discrepancy without establishing which semantic sufficiency standard is correct. M134/HLM032 is partial under all three judgments but scores .504: overlapping evidence proposes an invitation without fully explaining the later acceptance. It is a partial, not an insufficient, explanation; the highest adjudicated insufficient score is .111. These selected examples are not population error estimates. Adjudicated sufficient/partial/insufficient mean case-quality scores are .734/.694/.612. Because q also credits identity, label, and endpoints, it is neither explanatory sufficiency nor attempt-level Composite.

F.4 Previously Adjudicated Boundary Examples

Table 40 distinguishes reversal qualification from submitted explanatory sufficiency under the existing definition. It summarizes prior adjudications rather than changing the eligibility standard. Under the five tests in Section 3.1, the first question is whether the two endpoints oppose the same proposition about the same subject and state dimension. A proposal followed by a decision, or evaluations of different actions, need not meet this test. For a qualifying reversal, the existence of an explanation determines its label; it does not remove the case from discovery. Finally, the sufficiency of the model's submitted evidence is judged separately from explanations found elsewhere in the novel. Evidence located during review is not credited as originally submitted model evidence.

For U148, the conditional first state and later response occur in Chapters 13 and 25. The original reviewers both endorsed qualification before author rejection. For U100, the brief after quote relies on the adjacent question to identify the person and robbery. Describing this context does not make it a longer quotation originally supplied by the model. Its omission refers to the historical Main audit catalog, not the later Full inventory. The denial is near the start of the novel; the admission and explanation occur in the later testimony.

For U003, Chapters 19 and 120 establish the endpoints. The omitted explanation includes Baoyu's departure to religious life, Lady Wang's decision to send Xiren home, her brother and sister-in-law's marriage arrangements, and concern about harm to Lady Wang and her brother that leads her to abandon suicide. Incomplete submitted evidence does not make the reference unresolved. U104 contrasts Chapters 1 and 9: the same subject does not make etiquette and addiction the same state dimension.

A supplementary exclusion, U199 in *Jiu Ming Qi Yuan* (Chapters 5–6), makes the same distinction. Guixing praises Tianlai for obeying his father, then becomes hostile after a fabricated insult to his

Table 40: Four task boundaries illustrated by existing author adjudications. These selected examples do not estimate error prevalence.

Type / item	Proposition and endpoints	Decision and decisive boundary
Ordinary change; U148, *The Secret Garden*	Colin sometimes dislikes his mother's smiling portrait when ill and miserable; later he is no longer angry at her smile.	Excluded. The first endpoint is conditional, and the change is compatible with changed circumstances (tests 3–4 fail). This does not exclude all emotional or belief changes.
Qualified, sufficient; U100, *The Mysteries of London II*	Georgiana denies under oath that Rainford committed the Hounslow robbery, then acknowledges that he did.	The submitted admission of perjury to save him sufficiently explains the conflicting testimony. Deeper pressure motivating the perjury is not another necessary role here.
Qualified, partial; U003, *Hong Lou Meng*	Xiren insists on remaining in the Jia household until death rather than being redeemed; later she leaves for marriage.	HLM014 is retained. The submitted explanatory evidence omits necessary departure/marriage arrangements and reasons for abandoning suicide; a sufficient explanation exists in the novel.
Different state dimensions; U104, *Hen Hai*	Chen Bohe displays proper childhood etiquette; later he describes his opium addiction as incurable.	Excluded. Etiquette and dependence concern different propositions (tests 1–2 fail). Explaining his life trajectory does not make these endpoints opposite.

own father. These evaluations concern different actions and can coexist. Explaining the anger does not establish opposing states of one proposition. The author excludes the case despite initial shared qualification endorsement. These examples explain existing decisions; they do not replace the original agreement statistics or establish reference completeness.

Table 41: Semantic checks separate identity, explanation, and textual alignment. Descriptions summarize audited source passages; full records and sample definitions appear in Appendices E.17 and E.8.

Example	What the check establishes
Athos (VB011)	Shorter state quotations and a different royal betrayal passage support the same reversal; both reviewers accept the evidence, but the predicted case remains unmatched.
Xiren (HLM014)	The case identity is clear, but two cited sentences omit the causal steps connecting dismissal, marriage arrangements, and abandoning suicide.
King's abdication (JHY032)	Deleting the explicit demand to yield the throne leaves military defeat, the threatened city, and forced retreat as an explanatory route.
Baoyu / Diana	An alternative passage about Baoyu's earlier attachment aligns the endpoints on the same state dimension; Diana's earlier reference endpoint treated a mistaken scar belief as fact; both references have been repaired.

F.5 Disagreement Structure

We compare the original fields of all 216 unmatched-output audits, including their 145 disagreement records. Table 42 partitions the sample; its unweighted counts are not estimates of unique cases or of the full frame of 803 predicted cases. The original 104 category/identity disagreements comprise 103 category differences and one within-category reference-ID difference (U058). Qualification-field differences total $77 = 47 + 30$, overlapping that partition; only the 47 represent different endorsement decisions. Collapsing the original qualification field to endorsement versus nonendorsement gives agreement on $216 - 47 = 169$ items (78.2%). This binary agreement proportion describes a different partition from the multicategory $\kappa = .2953$; neither is a post-clarification reliability estimate.

Explanation judgments differ on 95 items: 61 within the original 104 and 34 within the additional 41. Of those 41, 38 have shared qualification endorsement, including 32 with different explanation

Table 42: Mutually exclusive groups of original unmatched-output judgments. The first three rows sum to 104; adding the fourth gives 145.

Group	n	Interpretation
One reviewer endorses qualification	47	A alone: 16; B alone: 31. Four nonendorsements are boundary cases, not unequivocal rejections.
Neither endorses it; reasons differ	30	Ordinary progression, state-dimension error, insufficient evidence, and definition boundary are distinct reasons.
Both endorse it; category or correspondence differs	27	Existing/outside-Main: 19; existing/split–merge: four; outside-Main/split–merge: three; different IDs within one category: one.
Category and required identity agree; endpoints or explanation differ	41	23 existing cases, 15 outside-Main candidates, two ordinary changes, and one state-dimension error.
All compared judgment fields agree	**71**	Includes category, required identity, qualification, endpoints, explanation, and duplication; not necessarily written reasons or semantic correctness.

judgments. Explanatory sufficiency thus remains a distinct difficulty even when qualification, category, and required identity agree.

Table 43: Original agreement statistics on 216 unmatched predicted cases. Dimensions overlap; κ is unweighted and is not revised after author adjudication.

Dimension	Different	Same	Cohen's κ
Qualification category	77	139	.2953
Output category	**103**	113	.3631
Before-endpoint support	24	192	.0310
After-endpoint support	16	200	.2113
Submitted explanatory sufficiency	95	121	.3398
Within-attempt duplication	1	**215**	**.6646**

High direct endpoint agreement coexists with concentrated marginals: A/B mark before support in 204/202 items and after support in 202/209. Chance correction can therefore produce low κ, without removing observed disagreement. The reviewers' category counts also differ (Table 44). Neither uses reference-issue or cannot-determine. Separate qualification endorsement totals are 140 and 155; sufficient/partial/insufficient/not-applicable explanation counts are 113/16/10/77 and 98/62/12/44. These marginals do not establish that one reviewer is uniformly stricter or more accurate.

Table 44: Original unmatched-output category usage. Each column totals 216.

Category	Reviewer A	Reviewer B
Same existing reference	**94**	**81**
Candidate outside supplied Main catalog	39	70
Split/merge	7	4
Ordinary progression	21	55
Subject/state-dimension error	33	6
Insufficient endpoint evidence	6	0
Definition boundary	16	0

Existing adjudications distinguish the issues. U058 has shared existing-case classification but different reference IDs; the author resolution identifies JHY037. U003 has agreed identity HLM014 but sufficient versus partial explanation judgments; adjudication retains identity and judges the submitted explanation partial. Both reviewers endorsed U148 and U199, later excluded by author adjudication. Agreement is not correctness. These resolutions used prior judgments and do not constitute fresh independent reviews or demonstrate improved agreement after clarification.

F.6 Evidence Expression, Context, and Diagnostic Mechanisms

Sufficiency, minimality, and recorded alternatives. The canonical prompt requires a faithful clause or passage that still supports its evidence role (Appendix J); it does not reward shortening a quote after support is lost. Necessary speaker attribution, a question needed to interpret an answer, or a temporal qualifier may therefore be retained. Reference construction likewise seeks independently sufficient endpoint anchors and complete explanatory routes, collecting confirmed alternatives when found. Prediction and reference have different reporting duties: a predicted explained reversal supplies one minimal sufficient route, whereas the reference can record several. A sufficient collected route need not exhaust possible explanations, as the residual-explanation audit demonstrates. Neither construction nor the historical audits establish that every anchor is minimally sufficient or that all sufficient alternatives have been collected.

CAAC uses recorded anchor length in its coverage denominator and permits partial alignment (Appendix H). Its excess-length factor is one up to three times that anchor length. These choices combine coverage of annotated evidence with resistance to quote stuffing. They do not equate reference length with semantic necessity: lower coverage can reflect omitted necessary evidence, dispensable context, or a different sufficient expression. Conversely, shared words need not identify the same reversal. These alternatives require source-based diagnosis.

What the historical review conditions establish. The challenge materials supplied source locations and adjacent context, with access to the novels. Subsequent unmatched-output and matched-pair instructions required full-text checks and distinguished qualification, identity, endpoints, and submitted explanatory sufficiency. In the matched-pair review, finding an explanation in the novel did not make it evidence submitted by the model; a different sufficient submitted route could nevertheless be recognized. The records establish available information and intended credit rules, not identical actual context use by both reviewers. They did not separately test minimality by removing portions of each quotation. Interpreting a submitted passage through necessary context and supplying absent state or causal evidence are distinct, but their boundary can require judgment. The observed disagreements do not identify context use as their cause, and no new independent review after clarification is reported.

Selected source-text diagnostics and attribution. An AI assistant examined 11 existing DeepSeek predicted cases from five novels: U013, U021, U032, U090, U128, M065, M073, and M177; same-target comparisons U078 and M172; and cross-case control U005. Selection targeted existing score and gate patterns, not a probability sample. The diagnostic consulted the full texts, reference entries, prior reviewer reasons, and author decisions, retaining the original human fields. It is neither a fresh human annotation round nor a held-out validation. The fixed-pair scores use the historical 1,033-case audit reference, separate from the current 1,032-case Main panel. Of 42 submitted quotations, 41 locate as continuous source passages; M065 contains the exception described below.

Table 45: Selected mechanisms under existing human correspondence judgments. Interpretations are assistant source-text diagnostics, not new independent human labels or population frequency estimates. Token counts use the respective scorer tokenizer; the best endpoint comparisons incur no excess-length penalty.

Items / reference	Recorded score pattern	Diagnostic distinction
U032, U078 / VB011	Before .275 / .367; 9 / 14 tokens against a 90-token anchor.	The longer excerpt adds Athos's name, Comte de la Fere. Length and explicit identity change together; necessary royal-relationship context remains.
U021, M172 / HH002	Before .007 / .400.	Retrospective self-reproach and earlier direct narration describe prior avoidance. The former is a different source passage, not merely a shorter excerpt of the anchor.
M073 / JHY058	After score 0; prediction 25 versus reference 22 Chinese bigram tokens.	Bian Bi speaks in person; the reference instead quotes his introduction in the same scene. The longer alternative has no shared bigrams and fails the endpoint gate.
U128 / JHY009	Before/after .910/.938; subject .154.	Both aliases already occur in the reference. Combining them in one predicted name and listing two additional participants lowers normalized subject compatibility.
U005 / VB020; relaxed VB084	Combined score .2717 for VB020, .2754 for VB084.	Leaving royal service and returning differs from promising to remain and later resigning. Shared participants do not establish the same direction or episode.

Other items prevent a uniform reference-omission explanation. U090 cites an examination outcome, with nearby context identifying the successful candidate. M177 instead quotes a pre-examination intention: this can support a changed commitment without itself establishing the completed actions named by the endpoint. For U013 and M065, transfer to prison and later continued imprisonment raise temporal-scope questions relative to the first-entry anchor. M065's explanatory evidence also omits the intervening speech tag *murmured he*, making it non-contiguous despite Explanation-Path F1 of .964. These observations qualify the interpretation of prior human support; they do not overwrite the original judgments or automatically certify new reference anchors.

U078 illustrates a different information problem. Its short declaration of love for La Vallière occurs in both Raoul's speech to Athos and Louis's later admission. One review's search note points to the former occurrence, while the other review's reason uses the latter. Both nevertheless endorse sufficiency. This documents source-location ambiguity, not a measured disagreement caused by context. The companion quotation and source scene remain relevant to interpreting the predicted case.

Credit carried by an inspected pair. Under the historical reference, M073's assigned case-quality score falls from .650 to zero under the endpoint rule; its prediction Composite changes from 4.2803 to .9694. Cancelling only this default pair, while fixing all other assignments, yields the same 3.3109-point drop. For M065, cancelling its inspected pair lowers Composite by 1.2077, whereas the Composite of the complete prediction under the endpoint-constrained rule drops by 6.7182. The latter change therefore cannot be attributed entirely to that item. For local cancellation, matched count changes from m to $m-1$, predicted case count n is fixed, and distance-weighted case-quality score changes by $-w_i q_i / \sum_j w_j$; Composite is recomputed from the resulting precision and quality. This is credit accounting without reassignment, not a corrected evaluation. Effects are non-additive, and the quotation and proposition issues above preclude treating all removed credit as independently verified model capability.

F.7 Fixed-Rule Endpoint Substitutions and Composite

Design and semantic support. We construct counterfactual versions of four saved DeepSeek predictions concerning two references: U032/U078 for Athos's reversal toward Louis XIV (VB011), and U021/M172 for Dihua's transition from avoidance to nursing Bohe in *Hen Hai* (HH002). The original unmatched-output or matched-pair reviewers both supported each item's identity, qualification, and

endpoints. Within each same-case pair, we exchange only before quotes, only after quotes, or both. Targets, labels, reasons, explanatory evidence, other predicted cases, reference evidence, and scoring parameters remain fixed. All eight original endpoint quotations locate uniquely and contiguously in the novels. We rerun candidate construction, global assignment, and aggregation for each complete prediction under the three fixed rules: 12 baseline replays and 36 counterfactual evaluations, not new model attempts. Every other predicted case retains its assignment.

Prior reviewers assessed the original endpoints in narrative context; they did not independently reassess the counterfactual predictions. Semantic support is conditional on those existing endpoint judgments. These two cases are development examples, not held-out semantic validation. The two novels' references are unchanged from the historical audit version. Runtime-level changes use current Main results, replacing only the specified attempt and preserving all other attempts and calculable novel supports.

Assignment and post-match quality. Under the default rule, replacing U021's retrospective before passage with M172's direct narration raises before alignment from .0071 to .4000 and restores HH002. Composite rises from 6.1051 to 13.6668. The reciprocal substitution removes M172's match and lowers Composite from 10.4707 to 3.1365. Replacing only M172's after preserves the match but raises alignment from .5208 to .9375, case-quality score from .6171 to .6561, and Composite to 10.8914. These distinguish a discontinuity at the matching gate from a post-match quality change.

In contrast, all Athos substitutions remain unmatched to VB011 under the default rule, leaving each prediction's Composite unchanged: subject compatibility (.5) and low Explanation-Path F1 remain fixed. Under the endpoint rule, adding the name-bearing before prefix restores U032, while removing it loses U078. The relaxed rule retains the same targets in these selected predictions but changes their evidence quality; this does not negate the cross-case errors diagnosed elsewhere.

Table 46: Complete prediction Composite (0–100) under fixed scoring rules and endpoint substitutions. All other fields and predicted cases remain fixed. These are selected counterfactual predictions, not new model attempts or corrected official scores.

Prediction	Rule	Original	Before only	After only	Both
U032	Default	**0.0000**	**0.0000**	**0.0000**	**0.0000**
U032	Relaxed	6.7266	6.7678	6.7440	**6.7880**
U032	Endpoint	0.0000	1.9948	0.0000	**2.0143**
U078	Default	**5.3079**	**5.3079**	**5.3079**	**5.3079**
U078	Relaxed	**10.3133**	10.2674	10.2923	10.2492
U078	Endpoint	**7.4240**	5.3079	7.4034	5.3079
U021	Default	6.1051	**13.6668**	6.1051	13.2719
U021	Relaxed	17.6382	**18.9737**	17.5851	18.5843
U021	Endpoint	6.1051	**13.6668**	6.1051	13.2719
M172	Default	10.4707	3.1365	**10.8914**	3.1365
M172	Relaxed	16.0607	14.9857	**16.4799**	15.0428
M172	Endpoint	10.4707	3.1365	**10.8914**	3.1365

Propagation to runtime-level Composite. Table 47 changes one attempt at a time in the current Main panel. Prediction-level changes are averaged within the affected novel, then across calculable novels. The two Hen Hai predictions each belong to a three-attempt novel cell and an 11-novel runtime support, so their prediction changes are divided by 3×11. The Athos DeepSeek Harness cell has only one calculable attempt; its change is not also divided by three. The resulting means were checked by complete reaggregation. The scenarios are separate, not jointly applied corrections or estimated population effects.

Table 47: Effect of one endpoint substitution on book-macro Composite of a DeepSeek runtime. All other saved attempts and novel supports are held fixed. Qwen Code is a runtime here; all four underlying predictions use DeepSeek. Differences use unrounded values.

Prediction	Field	Rule	Runtime	Original	Changed	Δ
U021	Before	Default	Qwen Code	5.1555	**5.3847**	**+0.2291**
M172	Before	Default	DeepSeek H.	**4.8346**	4.6123	-0.2222
M172	After	Default	DeepSeek H.	4.8346	**4.8473**	+0.0128
U032	Before	Endpoint	DeepSeek H.	3.4336	**3.6150**	+0.1813
U078	Before	Endpoint	mini-SWE	**2.6203**	2.5321	-0.0882

What this comparison establishes. Previously supported expression choices can affect both assignment and post-match quality under fixed rules, reaching prediction and runtime Composite. The comparisons do not isolate word count: the Athos before prefix adds a name, while the Hen Hai before passages contrast retrospective admission with direct narration. The unchanged Athos default results also preclude a general claim that longer quotations restore credit. This diagnostic does not establish minimality, improved independent reviewer agreement, error prevalence, cross-model bias, or a reliable alternative scorer. Context-sensitive equivalence and sufficiency remain concrete targets for independent measurement validation; the existing outputs, human returns, and official scores are unchanged.

G Reference Construction Details

G.1 Role-Structured References and Full-Text Review

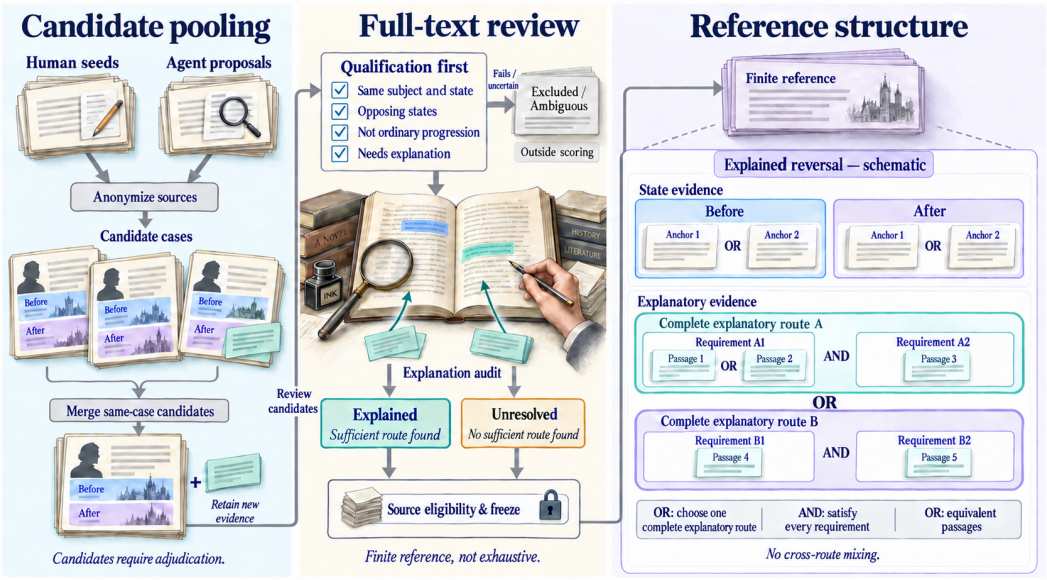


Figure 28: Reference construction and evidence organization. Candidate pooling precedes full-text adjudication and freeze. Alternative anchors and complete explanatory routes are OR choices; requirements within a route combine by AND. The expanded case is schematic.

The reference separates case identity from its admissible textual realizations. Before and after each contain a state claim and one or more independently sufficient source anchors. Explanatory evidence contains alternative complete explanatory routes. Within a set, all requirements must be satisfied; within a requirement, alternative anchors are interchangeable. The structure is therefore OR over routes, AND over requirements, and OR over equivalent passages. Scoring cannot pool incomplete portions of different recorded routes for maximum Explanation-Path F1. A different combination may be retained as an additional route only if source-based review establishes that it independently explains the reversal; absence from the current reference alone does not establish insufficiency.

Quotations must be concise while sufficient for their evidence role. Necessary context identifying the subject, speaker, referent, or temporal scope may be retained; brevity does not excuse missing state or causal evidence. Minimal sufficiency depends on the role and the context needed to interpret the passage. Explanatory evidence cannot manufacture opposing endpoints, and an action alone does not establish changed private belief. Qualification, correspondence to the same reference case, and sufficiency of submitted evidence are separate judgments (Table 40).

The review protocol requires active search across the full novel, not just verification of a model's citations. Search covers names and aliases, both directions of the state claim, relevant objects, causal events, recollections, letters, testimony, and later reinterpretations. Additional passages are retained only when they independently support the same endpoint or evidence role; a new route must fully explain the reversal. Grounding checks establish that a passage is in the source, not that its interpretation is correct. Two proposed requirements should be merged if one natural, minimal sufficient passage fulfills both, so that annotation granularity does not create an artificial penalty for concise predictions.

Only `include` records enter the evaluation reference. `exclude`, `ambiguous`, and `pending` do not; a pending record has no assigned label. Model proposals remain pending until adjudicated and cannot enter the evaluation reference simply by entering the working reference. The independent human scorer study checks prediction–reference correspondence; it does not establish exhaustive reference coverage or substitute for the construction adjudication.

G.2 Multi-Source Discovery and Candidate Anonymization

A discovery attempt receives the novel and common task, but no reference or prior case list. Proposals are anonymized and grouped into semantic case equivalence classes before adjudication. The workflow distinguishes new case equivalence classes, supplements to existing cases, duplicates, and ineligible proposals. A predicted case in an existing equivalence class is not automatically discarded: it may add a valid endpoint, an alternative complete explanatory route, or a boundary correction requiring owner review. Supplements do not count as new case equivalence classes.

Construction and evaluation have different objectives. Construction seeks valid coverage, heterogeneity, and useful marginal discovery; it need not equalize historical attempt counts. Discovery comprises an initial multi-runtime phase and a subsequent GPT-only phase. Initial DeepSeek rounds use Claude Code, Codex Harness, and DeepSeek Harness; subsequent rounds each use one GPT-5.6 plus Codex attempt under prediction isolation. Historical membership, failures, and adjudications remain attached to their respective construction cohorts. GPT discovery outputs do not enter the principal runtime comparison. Evaluated DeepSeek outputs were excluded from initial pooling and source attribution; later human-reviewed repairs to existing case equivalence classes are disclosed separately (Appendix E.7). Output extraction conditions are reported by cohort (Appendix A.3).

G.3 Stopping, Coverage, and Freeze

For the principal panel, stopping is evaluated separately for each novel and each local, medium, and long evidence-distance band, using the shortest complete span defined in Section 4. Within a discovery round, complementary outputs are merged into one case equivalence class representative and pure same-round duplicates are folded into that representative. Counting then occurs *before* filtering against existing case equivalence classes. For round r and band k, $N_{r,k}$ counts all normalized, classifiable representatives: new case equivalence classes, existing-case duplicates, supplements, cases requiring owner review, and classifiable non-cases. Structurally unclassifiable rejections are reported separately. Only adjudicated included *new* case equivalence classes in the working candidate set contribute to $I_{r,k}$, before final principal-source eligibility screening. Supplements can improve reference evidence, but do not contribute to this numerator. The number of newly added reference records is a separate bookkeeping total, not the stopping denominator.

A band's successful discovery opportunity has low yield when

$$N_{r,k} = 0 \quad \text{or} \quad \left(N_{r,k} \geq 5 \text{ and } \frac{I_{r,k}}{N_{r,k}} < 0.20 \right) \quad \text{or} \quad (0 < N_{r,k} < 5 \text{ and } I_{r,k} = 0)\,. \tag{9}$$

Exactly 20% is not low yield. Only a complete, readable execution of the round's fixed discovery opportunities updates the streak: low yield increments it, and non-low yield resets it. A failed round leaves the previous streak unchanged. An individual included case does not reset a streak when its round still meets the low-yield condition. Pending decisions or an unverified discovery status prevent finalization. A principal-panel novel stops early when all three bands have two consecutive successful low-yield opportunities, or at the fixed maximum of 20 discovery rounds. Historical rounds without verifiable opportunity evidence remain historical records and do not retroactively advance this streak.

An absent medium or long band additionally requires prediction-isolated directed review; a lack of model proposals alone is insufficient. Stopping is an empirical statement about the contributors, configuration, and review protocol, not exhaustive enumeration of an open case space. The round cap is an administrative stopping basis, not evidence of saturation; London II and Xi You Ji reached this cap. Later valid discoveries within the authorized construction are not discarded to enforce a case quota. Final freeze additionally requires completed coverage review, adjudication, source eligibility, source grounding, and contract checks. The resulting reference is a finite audited evaluation set.

Source screening is separate from these historical yield counts. The 26 known DeepSeek-exclusive cases were all retained as included in the pre-freeze working candidate sets and in the inclusion records of discovery rounds 1–4. Final screening removed them from the principal scoring denominator without subtracting them from those historical records. Those early rounds also do not retroactively advance the later verified low-yield streak. The records therefore do not support attributing premature stopping to this final exclusion; neither the stopping rule nor this check establishes exhaustive coverage.

New independent cases found in evaluated predictions do not enter Main or +DS; they enter the explicitly post-evaluation Full supplement only. Existing-case repairs based on human review are separately disclosed, with full rematching of the unchanged predictions (Appendix E.7). The principal Original panel permits reference-blind predictions to be collected before the initial reference freeze, provided the agent-visible conditions are fixed and the predictions are isolated from pooling, source attribution, and reference revision until that freeze (Section 5). This exception does not apply to injected texts or future independently constructed held-out evaluations.

G.4 Principal and Extended Collections

The reported benchmark evaluates 13 principal novels; 11 additional originals belong to the extended construction collection (Appendix C). Both use the same five case tests, label definitions, full-text adjudication, and role-structured evidence representation. English sources include Project Gutenberg editions and curated historical-text transcriptions; source and edition differences remain limitations. The revised principal panel contains 1,032 included reference cases and supports the fixed runtime comparison and paired intervention study. The frozen extended collection broadens textual coverage, with discovery depth bounded by available construction time rather than requiring the principal panel's two-consecutive-low-yield stopping criterion.

A time-bounded end to extended discovery does not establish saturation. All 11 extended references have completed coverage review and contain 806 included cases: 747 explained and 59 unresolved. Sixteen excluded records do not enter this count (Table 2). These finite references use the same adjudication and source-grounding requirements; their freeze does not establish exhaustive coverage. Principal statistics report language, length, label balance, intervention targets, and evidence-distance distributions. Known-provenance sensitivity is reported in Appendix E.6. Extended runtime evaluation remains future work, separate from the reported principal performance denominators.

G.5 Explanation-Targeted Editing and Its Validity Limits

The study pairs each of the 13 principal novels with an edited copy. Edits were designed to remove sufficient explanations for 467 originally selected cases while preserving opposing endpoints. Removing ineligible JMQY040 leaves 466 valid scoring targets; its historical edit remains in the unchanged text. The intended question was whether an agent changes its judgment after explanation removal while continuing to identify and ground the same reversal. The post-output audit in Appendix E.8 establishes that this intended manipulation was not achieved for at least 15 audited targets.

Target selection seeks a feasible near-balanced label distribution within each novel, rather than a random sample of errors. The number of selected explained cases depends on the original label counts and the availability of edits that preserve endpoints and unrelated cases. Post-edit labels are construction judgments, not independently established truth.

The intended intervention changes explanation availability, not subject identity, the minimal state proposition, or the temporal incompatibility of the endpoints. A qualifying edited case must retain its original case boundary and lack a surviving complete explanatory route, including an alternative route elsewhere in the novel. Merely deleting an explanatory evidence field from the reference would not create such an intervention. Full-text checking during construction sought to distinguish successful removal from residual routes and incidental qualification changes; it did not establish their absence.

Original and edited texts have separate references and reference-blind agent outputs. Current paired scores use repaired existing-case references, including the removal of JMQY040; they remain diagnostics rather than verified explanation-removal effects. The original mappings, scores, audit materials, and fixed-assignment exclusion analysis are retained as historical records. Surviving explanations and a reviewed prediction identity error limit interpretation; the audit does not justify relabeling every edited case or treating every predicted explained reversal as wrong. Unaffected reference cases provide a collateral-stability check during construction; the current prediction analysis does not establish population-level absence of collateral effects. Edited copies are derived conditions of the same 13 novels, not 13 additional independent novels. Their results remain separate from natural-text auditing and do not estimate the prevalence of naturally occurring errors.

H Detailed Scoring Protocol

The scorer first aligns cases, then evaluates the selected pairs. This is *quote-grounded lexical matching*, not an oracle of semantic identity. Prediction labels and reasons are not direct matching features; reference structure and schema-constrained evidence availability determine which signals can contribute.

H.1 Tokenization

EN applies Unicode NFKC and casefold, then extracts contiguous Unicode letter/digit sequences; punctuation and underscores are not tokens. ZH uses overlapping adjacent Han-character bigrams, retaining an isolated Han character as one unit. Punctuation, spaces, underscores, and Markdown marks do not break Han runs, whereas non-Han letters and digits do; non-Han word-like sequences retain the EN treatment. ZH does not perform simplified/traditional conversion, synonym expansion, or semantic normalization. The formulas and thresholds are otherwise shared. These rules specify the implementation, not demonstrated cross-language measurement equivalence.

H.2 CAAC and Explanation-Path F1

For a prediction passage p and reference anchor r, let L and N be their token counts, m their longest common subsequence length, and ℓ their longest contiguous match. Contiguity-Augmented Anchor Coverage is

$$\mathrm{CAAC}(p, r) = \left(0.75\frac{m}{N} + u(\ell, N)\right) \min\left(1, \frac{3N}{L}\right). \tag{10}$$

The contiguity bonus is 0.25 for a full contiguous reference match or $\ell \geq 11$, 0.20 for $8 \leq \ell < 11$, 0.15 for $5 \leq \ell < 8$, and zero otherwise; full coverage takes precedence for short anchors. Empty token sequences score zero. EN uses normalized word-like tokens; ZH uses a separate tokenizer (Section H.1). The excess-length factor is one when $L \leq 3N$; partial coverage can receive credit. This permits necessary adjacent context while suppressing quote stuffing, consistent with concise sufficient evidence. However, N is the recorded anchor length, not an estimate of semantically necessary length. A sufficient shorter passage or a different sufficient expression can receive limited overlap credit. CAAC does not recognize arbitrary paraphrases or prove entailment; adding context within the length allowance does not itself establish semantic support.

For an explained reference route s, let K_s be its number of requirements and P the number of predicted explanatory evidence quotes. Each quote–requirement edge uses the maximum CAAC over that requirement's alternative anchors. Let W_s be the maximum total one-to-one edge weight: each quote serves at most one requirement, and each requirement receives at most one quote. Precision is W_s/P when defined, recall is W_s/K_s, and Explanation-Path F1 follows Equation 1 (internal compatibility name `Setwise CAAC-EF1`). An empty prediction for a nonempty route scores zero. Select the route with highest F1, then highest aligned weight, then fixed route order. Missing requirements and wrong-role alignment limit W_s; unmatched submitted quotes increase P without increasing W_s. Each recorded set is scored separately, never pooling fragments across sets for maximum Explanation-Path F1. Whether an unrecorded combination forms a sufficient explanation requires source-based semantic assessment.

H.3 Case Matching

Target alignment. For predicted target t and reference surface s, form the harmonic mean of $\mathrm{CAAC}(t, s)$ and $\mathrm{CAAC}(s, t)$, or zero when their sum is zero. Take the best surface within each reference alias group. A maximum-weight one-to-one alignment of predicted targets to reference groups gives weight U^T; then $T = U^T/|\widehat{\mathcal{T}}|$, where $|\widehat{\mathcal{T}}|$ is the number of predicted target entries. This penalizes unrelated or duplicated predicted targets without requiring every auxiliary reference participant to be named.

For predicted case i and reference case j, T_{ij} is target compatibility, using one-to-one alignment of predicted names with reference alias groups. Let B_{ij} and H_{ij} be the best same-role before and after CAAC scores and $S_{ij} = \sqrt{B_{ij}H_{ij}}$. For reference cases labeled explained, R_{ij} is Equation 1; for

reference cases labeled unresolved it is zero. Define

$$C_{ij} = 0.20T_{ij} + 0.40S_{ij} + 0.40R_{ij}, \tag{11}$$

$$C_j^{\max} = 0.60 + 0.40\mathbf{1}[y_j = \text{explained}], \tag{12}$$

$$\widetilde{C}_{ij} = C_{ij}/C_j^{\max}. \tag{13}$$

The structural normalization avoids requiring explanatory evidence from reference cases labeled unresolved. An eligible edge must satisfy

$$\widetilde{C}_{ij} > \tau_{y_j}, \qquad T_{ij} \geq 0.50, \qquad \max(B_{ij}, H_{ij}) \geq 0.30, \tag{14}$$

where $\tau_{\text{explained}} = 0.40$ and $\tau_{\text{unresolved}} = 0.50$. Global one-to-one assignment maximizes the sum of $Z_{ij} = (\widetilde{C}_{ij} - \tau_{y_j})/(1 - \tau_{y_j})$ over eligible edges. Equal objectives are resolved by match count, then fixed index order. Ambiguous components remain recorded; unselected duplicates remain unmatched.

Coupling and measurement tradeoffs. The public schema requires empty explanatory evidence for a predicted unresolved inconsistency. Against a reference case labeled explained, that predicted case therefore lacks the explanatory evidence contribution available to a predicted explained reversal, even though its label is not a direct edge feature. The endpoint guard requires at least one endpoint to reach 0.30, not both: with enough other evidence, a case can match despite a weak second endpoint. Geometric endpoint quality and strict anchor diagnostics assess this remaining deficit. Case matching is therefore neither pure semantic discovery nor a certificate of full grounding.

Target compatibility divides by the number of predicted targets (Section H.3); additional names without reference counterparts can lower it. A concise but sufficient quote can have low coverage of a longer reference anchor, while excessive context receives the length penalty. Chinese normalization does not equate arbitrary simplified/traditional variants. Conversely, overlapping entities and passages need not imply the same state proposition, which is not an explicit matching feature. These are mechanisms, not measured error frequencies. The 96-pair challenge study does not estimate their population rates, and an unmatched-output audit alone cannot estimate false matches among selected matches. Thresholds are fixed defaults, not established cross-domain optima.

H.4 Case-Quality Score and Distance (DACE)

For each reference case j, Equation 2 combines match M_j, correct label L_j, endpoint quality A_j, and explanatory evidence E_j. An unmatched reference case has all four terms zero. For a matched pair, $M_j = 1$, $A_j = \sqrt{B_{ij}H_{ij}}$, and L_j checks label agreement on that selected match. A wrong label gives $L_j = E_j = 0$. With a correct explained label, $E_j = F_{\text{route}}$. With a correct unresolved label, $E_j = A_j$ if the predicted explanatory evidence is empty, and zero otherwise. Empty explanatory evidence therefore does not earn full evidence credit unless both endpoints attain full lexical alignment credit. The component weights and the rule $E = A$ allocate credit; the components do not measure independent abilities.

Soft anchors are complemented by strict diagnostics. Each side independently needs some reference anchor with CAAC ≥ 0.70 and contiguous match $\ell \geq \min(5, N)$. The search is existential across that side's anchors, not restricted to whichever anchor maximizes CAAC.

Complete-span distance. For an explained case, choose one before anchor, one after anchor, one complete evidence set, and one anchor for every requirement in that set. The span is the maximum chapter index minus the minimum among those chosen passages. Minimize over all such complete choices. For unresolved cases, minimize over the two endpoint choices alone. It is invalid to pick a convenient subset of explanatory evidence requirements or to select distance based on the evaluated prediction. Derived distance takes precedence over a stale cached distance. If chapter information is invalid, distance-dependent metrics become unavailable locally; matching, labels, anchors, evidence, and mean case-quality score remain computable where defined.

Normalized evidence span and distance weighting (BNDW) uses the shortest *complete* reference evidence span d_j, in chapter-index differences: both endpoints plus one sufficient route for explained cases, and both endpoints for unresolved cases. For a novel with n_b chapters,

$$\rho_j = \frac{\max(0, d_j - 1)}{\max(1, n_b - 2)}, \qquad w_j = \frac{1 + \sqrt{\rho_j}}{\sqrt{n_{b,k(j)}}}. \tag{15}$$

The bands are local ($\rho \leq 0.05$), medium ($0.05 < \rho \leq 0.20$), and long ($\rho > 0.20$); $n_{b,k}$ counts reference cases in that band. Density damping reduces the dominance of a populous band without discarding legitimate local cases. Both $Q_{\text{overall}} = |G|^{-1} \sum_j q_j$ and $Q_{\text{weighted}} = \sum_j w_j q_j / \sum_j w_j$ are retained. Novel length and evidence distance are distinct variables, and chapter normalization does not make search difficulty identical across novels.

H.5 Reference Construction Status and Metric Availability

With a construction-complete, frozen reference and nonempty predictions, P_{case} is matched predicted cases divided by all predicted cases: *precision against the frozen reference*. It is not independently adjudicated real-world precision. Composite is defined in Equation 3. Both inputs are proportions in $[0, 1]$, before conversion to displayed percentages. Composite is computed per attempt, then defined scores from successful attempts are averaged within each novel and available novels are weighted equally. Applying its formula to macro-averaged inputs would give a different quantity. Failed attempts have N/A evaluation metrics; completion is reported separately. When both terms are defined and zero, Composite is zero. With a partial reference, case precision and Composite are N/A; case recall, label, and evidence metrics remain available when their own conditions hold. Missing distance affects distance-dependent metrics only. Local ambiguity does not erase otherwise valid scores.

Label recall over all reference cases and label-stratified case-quality scores accompany accuracy conditional on a match. A missing label or distance stratum is N/A, not zero. Empty predictions yield zero case recall and mean case-quality score when the reference is nonempty, but undefined case precision and Composite. With no reference cases, case recall and mean case-quality score are N/A; for a construction-complete, frozen reference and nonempty predictions, case precision is zero. Per-novel values and effective denominators must be shown; the reporting design below does not change these mathematical conditions.

H.6 Challenge-Pair Human Review

Agreement with source text and deterministic assignment do not establish agreement with human judgments. The blinded study therefore examines selected matches, near-threshold unmatched pairs, independently confirmed same-case controls with different wording, and hard negatives in both English and Chinese. Selected matches also receive endpoint-role, explanatory-role, and explanatory sufficiency judgments. System identity, labels, reasons, and scorer decisions are hidden from reviewers; source context and full-novel search remain available. Appendix B specifies the 96-pair challenge sample and independent double-review procedure. This scorer study is separate from construction adjudication and does not estimate open-ended prediction precision or update the frozen evaluation denominator.

I Runtime Conditions and Statistical Estimands

The evaluation examines case discovery, label decisions, complete evidence, reliable task completion, and recorded resource usage under six native agent runtime conditions. Original-text auditing measures performance on natural narratives. Paired edited-text attempts were intended to test responses to explanation removal; residual explanations make them diagnostics rather than validated interventions (Appendix E.8).

I.1 Comparison Units and Common Conditions

The model is fixed to DeepSeek V4.1 Flash, using the canonical API name `deepseek-flash`. Six runtimes—Claude Code, DeepSeek Harness, Codex Harness, Qwen Code, OpenCode, and mini-SWE-agent—form six model–runtime configurations. Qwen Code is a runtime name and uses the same DeepSeek model. The principal Original comparison assigns three fixed attempts to each runtime–novel condition: $6 \times 13 \times 3 = 234$ attempts in three replicate blocks of 78. The attempt allocation is fixed independently of output success. GPT-5.6 plus Codex under prediction isolation supplies separate construction diagnostics and a secondary paired intervention study; neither enters

Table 48: Six principal runtime conditions. Version identifiers specify native builds; local protocol mappings remain part of the evaluated condition.

Runtime	Native method retained	Mechanical mapping
Claude Code `2.1.227`	Official agent loop, default tools, and context management	Anthropic-compatible provider, streaming, request accounting, and file isolation
DeepSeek Harness `47f94385`	Headless native-tools loop, session history, and termination	Provider configuration, request accounting, shell isolation, and same-session finalization
Codex Harness `cbe85e11`	Native `codex exec`, tools, sandbox, and compaction	Responses–Chat bridge, request/usage accounting, and finalization mapping
Qwen Code `0.23.0`	Bare headless loop and native tool environment	Isolated configuration, OpenAI-compatible proxy, stream parsing, and request accounting
OpenCode `1.18.29`	Native build agent, tools, and session loop	OpenAI-compatible proxy, event parsing, request accounting, and file isolation
mini-SWE-agent `2.4.6`	Default templates, query–execute–observe loop, Bash action, and submission	LiteLLM provider mapping, Docker environment, and submitted-result extraction

this principal comparison. The completed three-round Qwen-model comparison is supplementary; neither it nor older V4-model attempts enters the principal fixed-model ranking.

Qwen uses the public alias `qwen3.8-flash`, thinking enabled, 1M context, 32K output, and 300 requests. Its provider exposes no dated model snapshot, so backend revision identity across collection dates is unknown. GPT uses `gpt-5.6-sol` with `xhigh` reasoning and native service budgets. Its recorded Codex CLI versions are `0.154.0-alpha.6.2` in R1 and `0.155.0-alpha.2.6` in R2/R3; these budgets are not equated to the DS/Qwen allowance. The native versions in Table 48 describe the configured builds; uniform historical per-attempt version evidence is unavailable for every runtime and failed attempt.

Every principal agent receives the same complete read-only novel, canonical task prompt, and public output contract, with an initially empty writable scratch space. References, prior predictions, case lists, scorer output, repository history, user history, and user skills, rules, or external MCP servers are unavailable. Model-invoked shell commands have no network access; provider connectivity serves model inference only. The common configuration enables thinking, with a 1,000,000-token context limit, 32,000 output tokens per provider request, and 300 normal provider requests. Request accounting includes retries, compaction, and subagent calls where supported. A 300-second request or stream-idle timeout applies; there is no whole-novel wall-clock deadline.

Budget disclosure states the common allowance without prescribing a reading strategy. After normal requests are exhausted, at most one extra, tool-free finalization request can produce the public answer from the existing context. An explicit provider `max_tokens` or `length` termination permits at most one same-attempt, tool-free recovery, without rereading or exploration. It consumes the normal allowance if available, or the same single extra finalization opportunity at exhaustion. Repeated truncation or failed recovery does not permit another repair loop. All calls and their usage are counted.

I.2 Native Methods and Output Conditions

Table 48 identifies the native runtime objects and necessary protocol mappings. Native system prompts, tools, context management, and agent loops remain method differences. Adapters implement shared conditions without choosing cases, labels, or evidence, adding semantic instructions, or changing case correctness. A valid low-scoring or empty output is an observation, not a reason for extra semantic assistance.

The Original cohort retains the output extraction and validation condition under which its attempts were collected. A later shared deterministic output projection is used in DeepSeek INJ, the final Qwen matrix, and the secondary GPT study (Appendix A.3); it is not retroactively applied to recover failed principal attempts. Although projection makes no semantic judgments, its treatment of malformed outputs can affect task completion and candidate availability. Comparisons therefore distinguish output handling conditions as well as native versions.

The Original DeepSeek cohort lacks separately saved raw final messages for all 234 attempts, so a whole-cohort retrospective common-projection check is unavailable. All 174 successful INJ outputs have projection records; one records deletion of a nonpublic before-evidence field. GPT's 78 successful raw final JSON outputs equal the public results, with one missing projection log. The finalized Qwen R1 replaced ten cells unconditionally to obtain uniform raw-output records; all 67 successes across three rounds have final messages and projection records. Representation-level repairs include `bridge_evidence` field mapping, removal of nonpublic fields, and fixed non-scored reason text. They retain all candidate cases. These observed differences are disclosed as part of the collection conditions, not estimated as a causal projection effect.

This fixed-model comparison estimates differences between complete runtime conditions. It cannot identify a model main effect, a model–runtime interaction, cross-model ranking stability, or the causal contribution of an individual tool or memory mechanism. Protocol bridges remain part of each evaluated runtime condition.

I.3 Prediction Isolation and Fixed Cohorts

Principal Original predictions can be collected before the initial reference freeze because the agent-visible inputs, runtime conditions, budgets, and success criteria are independent of that hidden reference. During initial construction, operational checks are limited to exit state, public prediction availability, and schema validity; predictions are not used for semantic review, pooling, source attribution, or reference revision. Subsequent human review corrected existing case equivalence classes, with full rematching disclosed in Appendix E.7. Newly identified independent cases from evaluated outputs enter Full only, not Main or +DS.

Each attempt uses a separate agent session; any resulting prediction belongs to that attempt. The principal matrix uses one concurrent model attempt, and failures retain their original identity without replacement to fill a success quota. Collection dates and order expose possible service-time effects. Reused predictions are eligible only when their actual model, novel, prompt, runtime, tools, budgets, output contract, and isolation are substantively equivalent to the comparison condition. Eligibility is determined independently of success or favorable scores. Missing configuration evidence remains a disclosed limitation. All eligible outputs are evaluated against the same revised principal reference and unchanged scorer parameters; intermediate partial-reference scores are outside the formal comparison.

I.4 Source Eligibility

The principal reference admits initial human case equivalence classes, GPT-only discoveries, and case equivalence classes independently discovered by both GPT and DeepSeek. DeepSeek-only case equivalence classes are excluded from the principal denominator and retained in the complete adjudicated collection for sensitivity analysis. Source eligibility is determined by an audit of case equivalence classes, not the first submitter. Missing independent-discovery evidence cannot establish eligibility, and later human additions are not retrospectively called initial seeds. Each reference slice is identical across the six runtime conditions. Construction yields remain distinct from balanced evaluation and cannot estimate a model effect.

We rescore the same 177 successful predictions against 1,032-case Main, a 1,058-case variant restoring 26 DeepSeek-exclusive cases, and 1,066-case Full. Full adds eight later-manual cases, four identified after evaluation; those four do not enter Main or +DS. Existing-case corrections and full rematching are disclosed in Appendix E.7. Completion and model outputs are unchanged. The initial-human, GPT-only, shared, DeepSeek-exclusive, and later-manual slices are partial references: case precision and Composite are N/A for these slices. Metrics average successful replicates within a runtime–novel cell and then novels equally; 100,000 book-bootstrap resamples provide intervals. Appendix E.6 reports each slice's support and the resulting descriptive comparisons.

I.5 Completion and Scores on Successful Attempts

Completion rate and evaluation scores form separate primary result panels. Let S_{hb} be the successful public-output attempts for runtime h on novel b, out of the three fixed attempts. For metric q, let

$S^q_{hb} \subseteq S_{hb}$ contain attempts on which that metric is defined. The novel-level summaries are

$$C_{hb} = \frac{|S_{hb}|}{3}, \qquad \overline{q}_{hb} = \frac{1}{|S^q_{hb}|} \sum_{r \in S^q_{hb}} q_{hbr}. \tag{16}$$

The second quantity is N/A when S^q_{hb} is empty. A failed attempt has no evaluation score; it is not converted into an empty-case prediction or a zero-accuracy observation. An empty prediction is valid and follows the metric's ordinary denominator rules. Qualified attempt totals, successful and failed attempts, and metric-specific availability are reported together.

We macro-average within-novel summaries over their eligible novels. Paired comparisons use common novels with defined scores for both runtimes and disclose their effective support; a sensitivity analysis restricts all six runtimes to a shared set of novels. Novel bootstrap intervals retain all three attempts within each resampled novel. Thus novels, rather than individual attempts or detected cases, are the sampling units. Missing evaluation scores do not erase other computable metrics.

Evaluation metrics include case recall, case precision against the frozen reference, mean and distance-weighted case-quality scores, label recall, long-distance case-quality scores, strict anchors, and Explanation-Path F1. Available requests and token counts accompany these results; unavailable token, latency, and monetary cost measurements are not inferred. Composite is shown with its components and prespecified sensitivity checks if used for ranking (Appendix B). The challenge sample examines selected unmatched pairs for lexical matching failures; the additional probability samples describe unmatched-output composition and Reference validity, retaining uncertainty and disagreement (Appendix E.17). Completion failures distinguish the first supported causal error, such as provider, context, memory, transport, format, or budget failure; an operational failure alone does not establish a reasoning deficit.

I.6 Paired Explanation-Deletion Interventions

The DeepSeek C7 study pairs the existing Original fixed cells with three new INJ attempts for every runtime–novel condition. Completion rate uses all 234 attempts per side. Score differences require the same runtime, novel, and replicate to have successful predictions on both sides; the 146 observed pair differences are first averaged within novel and then bootstrapped by novel. R1 is globally serial. R2 begins serial and continues in six runtime-specific queues after the authorized stop; R3 uses six queues. These differences are execution conditions, not exclusion criteria or isolated causal factors.

The secondary GPT study has three paired rounds per novel. Analyses match Original and INJ at the novel and predefined-target levels; an unmatched edited case is lost matched discovery, not a correct unresolved judgment. Cross-round comparisons first average each condition's three values within a novel and then bootstrap paired novel differences. The aggregate transition counts and cumulative label recall are descriptive, not independent-case inference. One provider-capacity failure and its separately authorized successful replacement are both disclosed; this secondary exception does not change the principal DeepSeek no-replacement protocol. All principal and intervention comparisons use the 13-novel panel; the extended collection contributes no attempts.

J Canonical Task Prompt

The following is the common task prompt supplied with each complete novel. Native runtime system instructions, tool definitions, source mounting, and budget disclosure remain part of the runtime conditions described above. The task prompt provides neither reference cases nor a case-count quota. Instructions to read the complete text specify the task; they do not establish that a successful run actually inspected every passage. Wording and example code fences are preserved from the supplied prompt; visual line wrapping below is for typesetting only.

```
You are a Judge Agent auditing one complete unfamiliar novel for
long-horizon internal consistency. Read the complete novel supplied by
the Harness and return auditable cases grounded in that novel alone. Do
not use any human reference, gold answer, scorer output, or prior
prediction.
```

```
A case is not every state change. It is a potentially contradictory state
reversal: the before and after evidence concern the same subject and the
same state axis, express substantially opposing states around the same
proposition, and would make a careful reader reasonably suspect
inconsistency if the intervening explanation were hidden.

Candidate qualification comes before bridge search. Apply these tests in
order:

1. Do before and after concern the same subject and the same state axis,
defined at the narrowest shared issue supported by the endpoint text and
context?
2. Do they express a meaningful reversal around that shared issue? The
two endpoints may use different forms, such as evaluation, belief, advice,
intention, commitment, or enacted practical stance, only when the text
supports reading them as opposing positions on the same narrow
proposition. Do not infer that an inner belief changed merely from
outward action. Exclude different semantic roles such as plan versus
execution, expectation versus fulfilment, authorization versus
performance, or learning after reading.
3. Are the two endpoints naturally compatible merely because time passes?
Ordinary movement, plan fulfilment, expected arrival, and recovery after
treatment are normal plot development and must be excluded.
4. If the intervening evidence were hidden, would a careful reader
clearly ask why the state suddenly changed and reasonably suspect
narrative discontinuity without an explanation?
5. Is there concrete textual bridge evidence that explains the whole
reversal, rather than merely repeating an endpoint or providing nearby
background?

Use a two-stage bridge-isolation decision. In the qualification stage,
bracket any existing or newly discovered bridge evidence. Use only the
before and after passages plus the endpoint context necessary to identify
the subject, state, references, and endpoint time, and complete tests 1
through 4. Do not import a rule or causal condition from bridge evidence
back into the endpoints and then declare the endpoints naturally
compatible. Only after the candidate passes qualification may the
labeling stage use bridge evidence.

If tests 1 or 2 fail, do not emit a case. Test 4 may help resolve a
borderline alignment under test 2, but reader surprise cannot rescue
endpoints that lack a shared subject, narrow state axis, or opposing
positions. If the change is naturally compatible under test 3 and would
not require a non-trivial explanation under test 4, do not emit a case.
Tests 1 through 4 determine case qualification. Only after a candidate
qualifies under tests 1 through 4 may test 5 bridge evidence determine
its label.

"Naturally compatible merely because time passes" means that no
additional story rule, special ability, hidden event, or causal mechanism
is needed: a reader who knows only that time passed would not ask for an
explanation. If compatibility requires a story-specific rule or
non-trivial condition, such as a supernatural ability, a form restriction,
a concealed identity, deception, amnesia, or outside intervention, that
rule or condition is bridge evidence and must not be used to exclude the
case. For example, a plan followed by its execution, learning after
reading, or recovery after ordinary treatment is normal development. By
contrast, when an ability is available at one endpoint but unavailable at
the other and only a textual story-specific constraint reconciles them,
retain the reversal as a candidate; the constraint may make the case
`explained`.
```

```
Do not interpret these tests as requiring a formal logical contradiction
or identical grammatical forms at both endpoints. People, objects,
relationships, locations, beliefs, knowledge claims, advice, intentions,
promises, positions, identities, abilities, persistent physical states,
and story rules can all define a state axis. A private evaluation
followed by contrary advocacy or enforcement may qualify as a reversal in
practical stance when both endpoints clearly address the same narrow
issue and the hidden transition calls for a non-trivial explanation.
Describe only the supported practical reversal; do not claim that the
private evaluation itself changed unless the text says so. The fact that
some states can change over time does not by itself explain a significant
reversal.

Use exactly one label, but only for a reversal that has already passed
the candidate-qualification tests:

- `explained`: the potentially contradictory reversal qualifies as a case,
and concrete textual bridge evidence sufficiently explains the whole
reversal.
- `unresolved`: the potentially contradictory reversal qualifies as a
case, but the complete novel provides no sufficient bridge evidence to
explain it.

Bridge evidence determines only the label. It must never turn an ordinary,
naturally compatible plot development into an `explained` case.

Quotes must faithfully preserve the cited passage's words and meaning.
Shorter faithful clauses are allowed. Do not invent evidence. A quote may
be repeated in another evidence role when the same passage genuinely
serves both roles. Do not produce a novel-level score and do not infer
that the whole novel is consistent merely because no unresolved case was
found.

Use tool calls, private reasoning, and the writable `scratch/` directory
as needed while reading. None of that working material is part of the
answer. Once the complete novel has been read, consolidate once and
immediately return the final JSON. The sole visible final output must be
one JSON object whose only top-level key is `cases`. Do not output target
lists, observations, candidates, ledgers, summaries, progress reports,
analysis, plans, explanations of your process, or Markdown before or
after that object. Do not restart the analysis or repeat candidate
brainstorming. Keep each quote to the shortest faithful clause or passage
that still supports its evidence role, keep each reason concise, and
ensure the complete JSON fits comfortably within the available output
budget. These brevity requirements do not impose a case-count limit:
include every supportable case you identified rather than omitting cases
arbitrarily.

Return only one JSON object with exactly this public structure and no
Markdown code fence:

```json
{
  "cases": [
    {
      "targets": ["canonical subject name", "another subject when
      needed"],
      "before": {"quote": "faithful earlier quote"},
      "after": {"quote": "faithful later quote"},
      "label": "explained",
      "bridge_evidence": [
        {"quote": "faithful quote explaining the transition"}
      ],
      "reason": "concise evidence-grounded judgment"
    }
```
```

```
  ]
}
```

For `explained`, `bridge_evidence` must be non-empty. For `unresolved`, it must be empty. For an explained case, output one minimal sufficient bridge path. Use multiple `bridge_evidence` items only when they are jointly necessary to establish that path; do not enumerate redundant quotations or multiple alternative paths. `targets` contains natural-language names, not IDs assigned by a Harness. Return an empty `cases` array when no supportable case is found.