MindCopilot: Towards Formalizing and Evaluating Granular Human-LLM Co-Writing

Youqing Fang[1,2*], **Yinhao Tang**[1,2*], **Yanan Sun**[2†], **Jiangning Liu**[2], **Ziyi Wang**[2], **Xun Zhao**[2], **Bin Liu**[1], **Weiming Zhang**[1], **Kuikun Liu**[2], **Wenwei Zhang**[2] and **Kai Chen**[2†]
[1]University of Science and Technology of China
[2]Shanghai AI Laboratory
{fangyq, tangyinhao}@mail.ustc.edu.cn, {sunyanan, chenkai}@pjlab.org.cn

Abstract

Recent writing assistants are increasingly shifting from passive, prompt-driven interaction to proactive, suggestion-based completion, which integrates localized continuations into the writing flow and reduces coordination burden. However, existing evaluations simply focus on output quality, failing to capture how users accept, edit, or repair suggestions in real-time interaction, and thus obscuring the true usability of proactive co-writing systems. To address this gap, we adopt a sequential, behavior-centered view of interactive writing and formalize co-writing as a Human-in-the-Loop Markov Decision Process, modeling writing as an interaction shaped by user acceptance and editing decisions. Based on this formulation, we introduce the Co-Writing Fidelity Suite, an interaction-aware metric suite that captures both user–assistant alignment and cognitive editing effort, including Hierarchical Acceptance Rate and Knowledge-aware Editing Distance. We conduct a large-scale simulation study across 16 writing domains, using 1,688 controlled continuation queries sampled from different writing stages. Our analysis reveals systematic effects of interaction structure on acceptance behavior and editing cost. A follow-up user study with 30 participants confirms that these behavioral patterns align with real user experience. Together, our findings demonstrate that interaction-aware evaluation provides insights beyond output-only metrics and informs the design of more effective proactive writing assistants. Resources are available at https://github.com/fangyouqing/MindCopilot.

1 Introduction

Driven by rapid advances in Large Language Models (LLMs) [Liu *et al.*, 2025; Comanici *et al.*, 2025; Singh *et al.*, 2025], digital writing workflows have been reshaped, enabling users to draft, revise, and expand text with unprecedented ease [Anthropic, 2025; OpenAI, 2025]. In most existing systems, writing assistance is realized through *prompt-driven interaction*, where the model remains largely *passive* until explicitly invoked [Bai *et al.*, 2022; Pasch and Ha, 2025]. Authors issue discrete requests, receive complete responses, and manually integrate them into an evolving document. While effective for content generation, this paradigm places the burden of initiative and coordination squarely on the user, offering limited support during the continuous flow of writing [Reza *et al.*, 2025; Mysore *et al.*, 2025; Dhillon *et al.*, 2024; Li *et al.*, 2026].

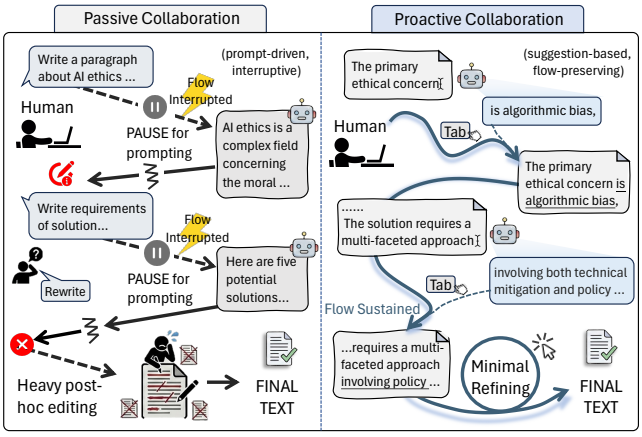

Figure 1: Two paradigms of LLM-assisted writing. Prompt-driven interaction interrupts writing flow and shifts coordination burden to the user (left), whereas suggestion-based interaction integrates proactive, localized completions into the writing process, sustaining flow and reducing post-hoc editing (right).

Recent writing interfaces increasingly adopt a *suggestion-based interactive completion* paradigm [Lee *et al.*, 2022; Lee *et al.*, 2022; Laban *et al.*, 2024], in which the assistant acts more *proactively* within the writing process. Rather than waiting for prompts, the system surfaces localized continuations directly in the text, and authors iteratively accept, revise, or reject them. This mode of collaboration preserves human agency while reframing writing as a recurrent decision process—one centered on *suggestion review, selective acceptance, and post-hoc editing*. Despite its growing practical relevance, how users behave in such proactive co-writing settings remains poorly understood.

† Corresponding author.

This gap is mirrored in prevailing evaluation practices for writing assistants. Widely used paradigms—such as LLM-as-a-Judge or Arena-style comparisons—assess generated text largely in isolation [Chung *et al.*, 2025; Li *et al.*, 2025; Liao *et al.*, 2025; Ying *et al.*, 2025; Fein *et al.*, 2025], abstracting away the human actions required to integrate suggestions into an ongoing document. As a result, generation quality is often conflated with usability: metrics capture how good a continuation looks, but not whether it aligns with author intent or meaningfully reduces editing effort. Such evaluations fall short for interactive writing, where acceptance behavior and editing cost are as central as surface-level text quality.

To address this limitation, we formalize interactive co-writing as a *Human-in-the-Loop Markov Decision Process (HitL-MDP)*. In this formulation, writing is modeled as a sequential decision problem: the assistant generates context-aware suggestions, the user responds through accept/reject/edit actions, and these actions drive state transitions that shape future assistance. This abstraction provides a principled framework for analyzing how different interaction structures influence user behavior over time, while admitting a operationalization in which acceptance decisions can be instantiated as a learnable component for downstream optimization.

Building on this framework, we introduce the **Co-Writing Fidelity Suite**, a set of interaction-aware metrics designed to evaluate human–AI collaboration along two complementary dimensions: alignment and interaction friction. First, we propose **Hierarchical Acceptance Rate (HAR)**, which treats acceptance not as a binary outcome but as a structured, multi-stage decision, reflecting the cascading filters users implicitly apply during suggestion review. Second, we introduce **Knowledge-aware Editing Distance (KED)**, which quantifies post-acceptance editing effort by weighting edits according to their cognitive cost, recognizing that correcting knowledge-dense content or factual errors demands substantially more effort than surface-level revisions.

Using these metrics, we conduct a large-scale simulation study across multiple interaction paradigm. We curate a dataset of 60 human-authored articles spanning 16 writing domains and construct 1,688 controlled continuation queries sampled from different writing stages. By varying interaction structures while holding content constant, we analyze how factors such as interaction history affect acceptance behavior and editing effort. Guided by insights from these simulations, we further design an adaptive interaction policy and implement a real-world writing interface. A subsequent user study with 30 participants shows that behavioral patterns observed in simulation closely align with subjective user experience.

Our contributions are threefold:

- We cast interactive co-writing as a Human-in-the-Loop sequential decision process, providing a principled abstraction for modeling user acceptance, rejection, and editing behaviors over time.
- We introduce the Co-Writing Fidelity Suite, a set of interaction-aware metrics—including Hierarchical Acceptance Rate and Cognitive-Weighted Editing Distance—that quantify user–assistant alignment and the cognitive effort required to integrate suggestions.
- We conduct a large-scale empirical study combining controlled simulation and real-user experiments, demonstrating that interaction-aware evaluation based on the Co-Writing Fidelity Suite reveals behavioral patterns invisible to output-only metrics.

2 Related Work

2.1 Human-AI Co-Writing: From Tool to Collaborative Agency

The co-writing paradigm is evolving from passive, auxiliary local completion [Lee *et al.*, 2022; Coenen *et al.*, 2021] to mixed-initiative collaboration that frames AI as a "co-operative partner" preserving user ownership [Ashkinaze *et al.*, 2025; Zhang *et al.*, 2025; He *et al.*, 2025; Yang *et al.*, 2022a]. However, evaluation methodologies lag behind. Traditional metrics fail to capture dynamic user experience [Shen and Wu, 2023], while recent attempts at quantifying effort[Devatine and Abraham, 2024] or keystroke intent [Le *et al.*, 2025] suffer from coarse granularity and rigid definitions. There remains a notable absence of fine-grained, entity-aware metrics to measure cognitive friction in complex interactions.

2.2 Long-form Writing: Control and Alignment

With improved context capabilities, generation tasks now encompass document-level long-form content. While works employ outline-driven, recursive strategies to ensure long-range coherence[Mirowski *et al.*, 2023; Yang *et al.*, 2022b; Yang *et al.*, 2023; Bai *et al.*,]—evaluated via coarse-grained dynamic metrics [Du *et al.*, 2025; Wu *et al.*, ; Que *et al.*, 2024]—they predominantly rely on "black-box" monolithic generation. This approach leads to "loss of control" and "cognitive misalignment", forcing users into extensive post-generation revisions by ignoring the need for progressive confirmation. The absence of structured interaction protocols prevents the establishment of cognitive consensus during intermediate stages; as a result, even high-quality outputs often diverge from user intent, significantly increasing the burden of post-editing.

2.3 Creative Evaluation: Quality vs. Intent

While generative capabilities continue to expand, defining effective creative collaboration in open-domain scenarios remains elusive. Lacking a single gold standard, current approaches depend heavily on preference learning and "Arena modes". [Fein *et al.*, 2025; Fang *et al.*, 2025; Chung *et al.*, 2025]. Yet, these metrics—rooted in "win rates" or "external judging"—reveal a fundamental disconnect: quality does not equal satisfaction. In co-creation, user acceptance is driven less by writing quality than by alignment with the user's specific creative conception. Existing Arena benchmarks fail to capture this alignment with the user's Original Intent. Although AI offers diverse perspectives, it often misses the user's core message, necessitating a paradigm shift from binary "good/bad" evaluations to the study of "acceptance and editing behavior patterns".

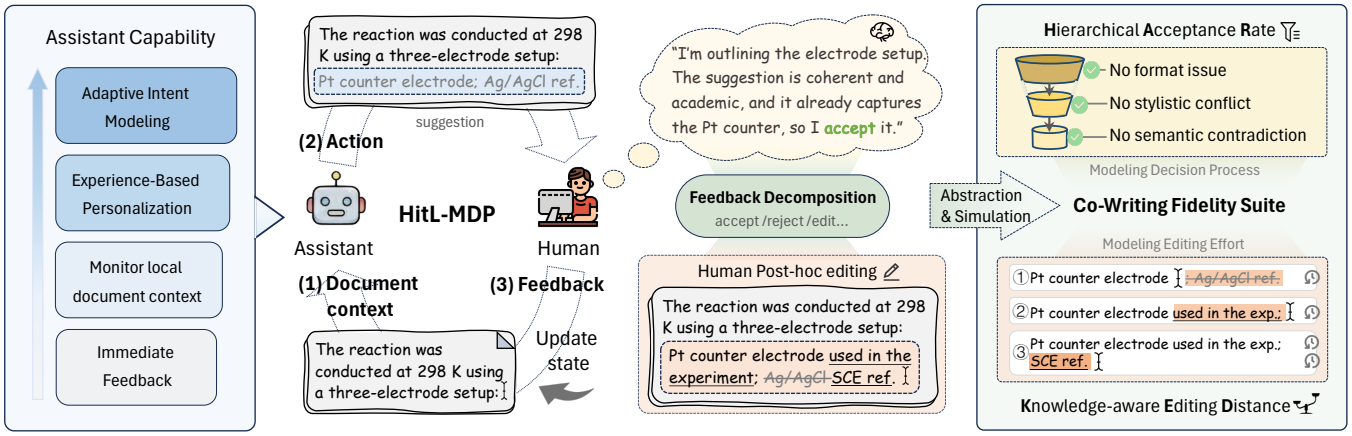

Figure 2: **Co-writing as a Human-in-the-Loop decision process.** The center illustrates co-writing as a HitL-MDP (Sec.3), where the assistant proposes context-aware suggestions and the human responds actions that update the document state. Human feedback is decomposed into acceptance followed by post-hoc editing. The right panel shows the Co-Writing Fidelity Suite (Sec.4): HAR models structured acceptance decisions, while KED models editing effort, with darker regions indicating higher cognitive cost for knowledge-dense edits.

3 Co-Writing Formulation

3.1 Proactive Collaboration

We frame interactive auto-completion as a mixed-initiative co-writing setting, where a human author and a writing assistant collaboratively compose a document via fine-grained, suggestion-based interactions. In contrast to prompt-driven generation, the assistant proactively monitors the evolving document and offers localized textual continuations during writing, without requiring explicit user prompts.

Formally, given a document state s_t, the assistant samples a suggestion

$$a_t \sim \pi_\theta(a \mid s_t), \tag{1}$$

where $a_t \in \mathcal{A}$ denotes a bounded continuation, designed to remain local in scope, concise in length, and aligned with the surrounding writing style. Crucially, these continuations are non-binding. The assistant surfaces candidate text, while all document updates are determined by the human author.

3.2 Human-in-the-Loop MDP

Upon receiving a suggestion, the human author provides feedback that governs how the suggestion is treated. Such feedback span a range of document-editing actions, including direct adoption, user-mediated modification, explicit dismissal, and other interactions that alter the proposal's realization. Collectively, these forms of feedback are drawn from a structured yet open-ended user response space $\mathcal{U}$, and determine whether and how the suggested content contributes to the subsequent document state.

This asymmetric interaction induces a recurrent decision-making loop. At step t, the assistant observes the current document state $s_t \in \mathcal{S}$ and generates a suggestion $a_t \in \mathcal{A}$. The human author then responds with a form of feedback $u_t \in \mathcal{U}$, which specifies how the suggestion is handled. The document is updated accordingly, yielding a new state s_{t+1}, upon which the assistant conditions to produce the next suggestion. Over time, document evolution emerges from the repeated interplay between assistant suggestions and human feedback.

We model this iterative process as a Human-in-the-Loop Markov Decision Process (HitL-MDP), which makes explicit the division of roles between proposing and deciding. Formally, the process is defined as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{U}, \mathcal{T} \rangle$. At each step, given the current state s_t, the assistant samples a suggestion according to $a_t \sim \pi_\theta(\cdot \mid s_t)$, after which the human selects a form of feedback u_t. State transitions are governed by the transition function

$$s_{t+1} = \mathcal{T}(s_t, a_t, u_t), \tag{2}$$

reflecting the fact that all document updates are enacted through human actions.

This formulation highlights a key structural property of interactive co-writing: while the assistant influences the process by proposing candidate continuations, control over state transitions resides entirely with the human author. As a result, learning and evaluation in this setting must be grounded in observable human responses to system proposals.

4 Co-Writing Fidelity Suite

Modeling mixed-initiative co-writing as a HitL-MDP makes user acceptance and editing actions explicit. Yet these behaviors are costly, noisy, and ill-suited for large-scale or comparative evaluation across interaction paradigm. To address this limitation, we adopt a *human simulator* perspective, in which user feedback is abstracted into behavioral proxies grounded in observed co-writing patterns. Under this view, a useful suggestion is not defined by fluency alone, but by whether it (i) satisfies the user's implicit acceptance criteria and (ii) minimizes the cognitive effort required for integration.

Motivated by this perspective, we introduce the **Co-Writing Fidelity Suite**, a set of interaction-aware metrics that approximate human acceptance behavior and post-acceptance editing cost within the HitL-MDP framework. Specifically, we propose **Hierarchical Acceptance Rate (HAR)** to measure intent-aligned acceptance, and **Knowledge-aware Editing Distance (KED)** to quantify the cognitive effort incurred after acceptance.

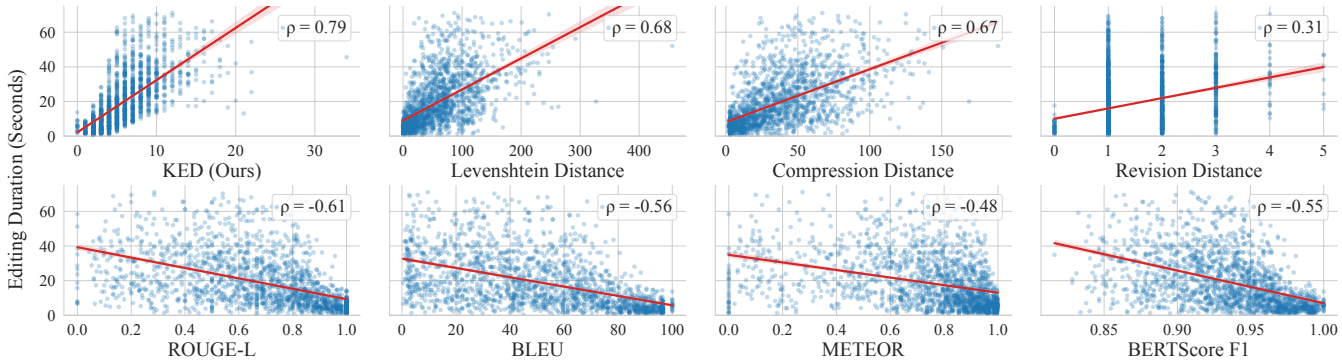

Figure 3: Editing duration versus editing distance and similarity metrics. Each point represents a human post-editing instance; red lines show linear regression trends. KED aligns most strongly with observed editing time, whereas surface-level distances and similarity metrics correlate weakly with human editing effort.

Metrics	DeepSeek-V3		GPT-5.1		Gemini-2.5-Pro	
	Align	κ	Align	κ	Align	κ
Baselines (Single-Dimension & Holistic)						
$\text{LLM}_{\text{Logic}}$	51.6	0.072	53.5	0.032	56.6	0.020
$\text{LLM}_{\text{Style}}$	67.3	0.106	51.4	0.102	57.7	0.018
$\text{LLM}_{\text{Semantic}}$	69.0	0.158	74.6	**0.184**	65.7	0.180
$\text{LLM}_{\text{Holistic}}$	48.8	0.014	78.3	0.063	64.1	0.196
Hierarchical Ablation (Ours)						
HAR_{L1}	59.9	0.126	76.3	0.045	67.0	**0.269**
$\text{HAR}_{\text{L1+L2}}$	60.3	0.104	77.0	0.070	79.8	0.243
HAR_{Full}	**69.3**	**0.166**	**81.1**	0.154	**82.2**	0.254

Table 1: Alignment rate (%) and inter-annotator agreement (Cohen's κ). Best results are **bolded**; second-best are underlined.

Metric	Pearson	Spearman
KED (ours)	**0.686**	**0.786**
Levenshtein Distance	0.587	0.684
Compression Distance	0.582	0.670
Revision Distance	0.305	0.306
ROUGE-L	-0.498	-0.612
BLEU	-0.499	-0.565
METEOR	-0.356	-0.482
BERTScore	-0.469	-0.546

Table 2: Correlation between automatic evaluation metrics and human post-editing effort, measured by editing duration. Pearson (r) and Spearman (ρ) correlations are reported.

4.1 Hierarchical Acceptance Rate

Motivation and Definition. In proactive co-writing, acceptance is a structured decision rather than a binary judgment. Prior work in cognitive psychology and behavioral decision theory suggests that humans evaluate complex options through sequential filtering processes, progressively eliminating candidates that fail increasingly stringent criteria rather than making a single holistic judgment [Lee *et al.*, 2024; Simon, 1955; Tversky, 1972; Payne *et al.*, 1993]. Users thus implicitly apply a sequence of filters—ranging from basic correctness to higher-level semantic alignment—before integrating a suggestion.

We model this process with *Hierarchical Decision Criteria*. Given an ordered set of criterias $\{\phi_1, \phi_2, \ldots, \phi_K\}$, a suggestion is accepted only if it passes all criterias, where failure at any stage terminates the process. Early-stage criteria capture immediate usability (e.g., grammaticality, continuation correctness), while later stages assess stylistic and semantic alignment, including required entity coverage. When applied across a collection of model-generated suggestions, this hierarchical decision process induces a well-defined acceptance probability, which is formalized as *Hierarchical Acceptance Rate*. This conservative design ensures stable and interpretable acceptance judgments across interaction rounds.

Alignment with Human Evaluation. We evaluate HAR on 1.2K expert-annotated instances. Each instance contains a writing context, a model-generated continuation, and a gold reference continuation; annotators determine whether the suggestion should be accepted using a checklist-style protocol that mirrors realistic co-writing decisions.

We compare HAR against two representative LLM-as-a-Judge alternatives: *single-dimension judges* that decide acceptance based only on logic, style, or semantic correctness, and a *holistic judge* that outputs one-shot accept/reject without explicitly modeling intermediate criteria. We report (i) *alignment rate* with expert decisions and (ii) *inter-annotator agreement* measured by Cohen's κ.

Table 1 shows that HAR yields the most reliable human-aligned acceptance estimates across backbone models. In terms of alignment, the full hierarchical design (HAR_{Full}) achieves the best performance for all three backbones (e.g., 69.3/81.1/82.2 for DeepSeek-V3/GPT-5.1/Gemini-2.5-Pro), while single-dimension judges vary substantially across model families, indicating limited robustness when acceptance is reduced to a single criterion.

Agreement results further support the benefit of explicit criteria: HAR_{Full} achieves the highest κ on DeepSeek-V3 (0.166) and remains competitive on GPT-5.1 (0.154), and on Gemini-2.5-Pro the hierarchical variants markedly improve agreement (up to 0.269), suggesting that structuring acceptance into ordered criteria reduces ambiguity compared to

Level	Interaction Paradigm	Interaction Scope	Key Capability
L0	User-Initiated Collaboration	Reacting to explicitly stated user intent	Demand Responsiveness
L1	Stateless Proactive Collaboration	Proactively predicting user intent from local context	Immediate Feedback
L2	Stateful Proactive Collaboration	Proactively predicting user intent from interaction history	History-Aware Personalization
L3	Adaptive Proactive Collaboration	Adaptively predicting user intent via strategies	Adaptive Intent Modeling

Table 3: Interaction paradigms organized by increasing levels of initiative, contextual awareness, and adaptivity. The progression highlights the transition from reactive to proactive intent inference in human–AI co-writing.

one-shot holistic decisions.

Overall, these results are consistent with the view that acceptance in interactive co-writing is better characterized as a *hierarchical decision process* rather than a monolithic judgment: making intermediate criteria explicit not only improves alignment with expert choices, but also produces more consistent acceptance decisions.

4.2 Knowledge-aware Editing Distance

Motivation and Definition. Acceptance alone does not imply low interaction cost. In realistic co-writing, users accept a suggestion yet invest substantial effort in post-editing to reconcile it with their underlying intent. Such effort is frequently driven by correcting semantic mismatches, factual inaccuracies, or discourse-level inconsistencies, which are poorly captured by surface-level edit-distance or similarity metrics.

We introduce *Knowledge-aware Editing Distance* (KED) to quantify this post-acceptance editing cost. KED measures the distance between a model-generated suggestion and the final user-edited text, while weighting edit operations according to their cognitive demands: semantic rewrites and factual corrections incur higher costs than local or surface-level edits. As a result, KED is designed to reflect not just textual change, but the mental effort required to repair a suggestion into an acceptable continuation.

Alignment with Human Evaluation. We evaluate KED using writing logs from the CoAuthor dataset [Lee *et al.*, 2022], which records human revisions to AI-generated suggestions together with timestamps. Following prior work[Devatine and Abraham, 2024], we use *editing duration* as a behavioral proxy for cognitive effort, as longer interaction time is consistently associated with increased mental workload. We compare KED against a diverse set of baselines, including character-level distances (Levenshtein Distance), structure-aware distances (Compression Distance[Devatine and Abraham, 2024], Revision Distance[Ma *et al.*, 2024]), and similarity-based metrics (ROUGE-L, BLEU, METEOR, and BERTScore). For each metric, we compute Pearson and Spearman correlations with observed editing duration.

Figure 3 illustrates the relationship between editing duration and metrics. KED exhibits the strongest monotonic and linear correlations with editing time, indicating that it more faithfully reflects the cognitive cost experienced by users during post-editing. In contrast, conventional edit distances show weaker correlations, as uniform edit costs overemphasize superficial changes. Similarity-based metrics display moderate to weak negative correlations, confirming that high textual similarity does not necessarily imply low editing effort.

Model	Level	Overall		Scientific		Creative	
		HAR(%) ↑	KED ↓	HAR(%) ↑	KED ↓	HAR(%) ↑	KED ↓
DeepSeek-V3.2	L1	6.99	10.18	8.18	11.70	5.35	7.11
	L2	10.37	9.97	12.88	11.39	6.90	6.31
Gemini-2.5-Flash	L1	13.39	9.68	14.83	11.11	11.41	7.20
	L2	16.05	10.24	20.04	11.79	10.56	6.35
Gemini-2.5-Pro	L1	14.17	**9.17**	15.15	11.00	12.82	6.20
	L2	17.89	9.39	19.12	**10.81**	**16.20**	7.03
GPT-5.1	L1	17.59	11.28	21.47	12.05	12.25	9.46
	L2	**19.61**	10.50	**23.62**	12.14	14.08	**6.06**

Table 4: Results of LLMs under Level-1 and Level-2 interaction paradigm across scientific and creative domain.

Table 2 quantifies these trends. KED achieves the highest Pearson and Spearman coefficients among all evaluated metrics, demonstrating both sensitivity to absolute editing effort and consistency in ranking instances by difficulty. All reported correlations are statistically significant.

Taken together, our findings indicate that model suggestions vary in quality even under universal acceptance. Explicitly modeling the cognitive burden of different edit types is essential for capturing the friction users experience after accepting an assistant-generated suggestion.

5 Co-Writing Experiment

5.1 Interaction Paradigms

We organize co-writing systems into four interaction paradigms, as summarized in Table 3. These paradigms differ in how initiative is distributed between the user and the assistant, what contextual signals are available for intent inference, and whether the system adapts its intervention behavior over time. Together, they characterize a structured progression from passive assistance to proactive and adaptive collaboration, providing a unified lens for analyzing user acceptance and editing behavior.

L0: User-Initiated Collaboration. In this paradigm, the assistant responds only to explicitly stated user intent and provides assistance on demand. All initiative remains with the user, and the system does not anticipate or infer intent. L0 emphasizes demand responsiveness and serves as a reference condition, particularly in user-facing evaluations.

L1: Stateless Proactive Collaboration. L1 introduces system-initiated assistance by proactively predicting immediate user intent from local document context. Suggestions are generated independently at predefined points, without access to prior interaction outcomes. This paradigm provides

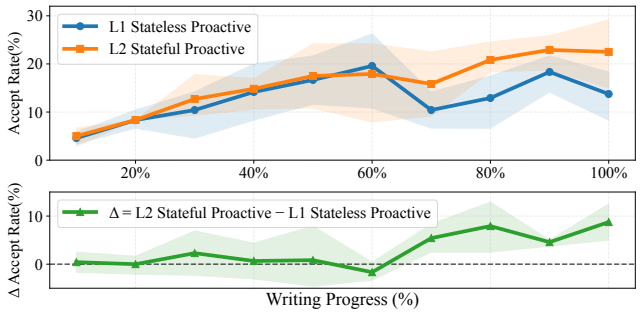

Figure 4: Acceptance rate as a function of writing progress for L1 and L2 proactive interaction. L2 exhibits increasing advantages in later stages, with Δ curves and confidence intervals shown below.

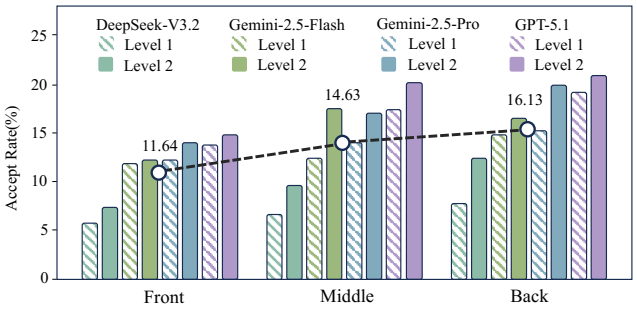

Figure 5: Paragraph-level acceptance rates increase from early to late paragraphs across models and interaction settings.

immediate feedback while isolating the effect of proactive intervention.

L2: Stateful Proactive Collaboration. L2 extends L1 by conditioning intent prediction on accumulated interaction history, including previous human feedback. This enables experience-based personalization, allowing the assistant to progressively align with the user's evolving intent. Comparing L2 with L1 isolates the behavioral impact of interaction memory.

L3: Adaptive Proactive Collaboration. L3 further introduces adaptive intent modeling, where the assistant dynamically determines when and how to provide suggestions. Intervention itself becomes a strategy-level decision, making interaction events endogenous to the system policy. We therefore treat L3 as an analysis-driven paradigm, informed by insights from L1 and L2, and reserve its evaluation for downstream system design and user studies.

In the following sections, we focus quantitative analysis on L1 and L2, where interaction events are well-defined and directly comparable, and use the resulting insights to inform adaptive collaboration in L3.

5.2 Quantitative Evaluation

Setup. We evaluate multiple language models under different interaction paradigms using a controlled dataset of 60 human-authored articles across 16 domains. Articles are segmented into 1,688 continuation queries sampled from different writing stages, with fixed ground-truth continuations across settings to isolate the effect of interaction structure.

Results and Analysis. As shown in Table 5, GPT-5.1 achieves the highest HAR overall, while Gemini-2.5-Pro attains the lowest KED, reflecting complementary strengths in intent prediction and post-editing efficiency. Moving from L1 to L2 further yields systematic gains in acceptance and reduced editing effort, underscoring the value of stateful intent modeling.

- **Intent prediction accuracy increases as writing progresses.** As shown in Figure4, both stateless (L1) and stateful (L2) proactive settings exhibit an upward trend in acceptance rate from early to later writing stages, indicating that richer local context facilitates more accurate intent inference. Notably, the performance gap between L2 and L1 widens in the later stages of writing, where accumulated interaction history becomes available. The consistently positive Δ curve, together with its confidence intervals, suggests that conditioning on prior user decisions enables more effective anticipation of user intent beyond what can be inferred from local context alone. These results highlight the growing value of interaction memory as co-writing proceeds and user intent becomes increasingly shaped by earlier choices.
- **Paragraph-level acceptance increases monotonically with writing progress.** As shown in Figure 5, acceptance rates computed from paragraph-level continuations consistently rise from the front to the middle and back sections of an article across all evaluated models and interaction settings. This monotonic trend holds for both stateless and stateful proactive paradigms, as well as across different model families, indicating that the effect is not model-specific. The increasing acceptance toward later paragraphs suggests that author intent becomes progressively more explicit and constrained at the paragraph level, making intent inference more reliable as the document structure and thematic direction solidify.
- **Acceptance rates are consistently lower in Creative than Scientific domains.** This pattern is consistently observed across models and interaction settings (Table4) and is characteristic of the higher semantic openness of creative writing.

Based on the above analyses, L3 Collaboration should treat proactivity as a cost-sensitive policy conditioned on interaction history, writing stage, and domain characteristics. Such adaptivity aligns system initiative with the evolving predictability of user intent while limiting unnecessary editing overhead.

5.3 User Study

To validate whether the behavioral differences observed in offline evaluation translate into real user experience, we conduct a controlled user study with 30 participants recruited from universities and technology companies, using an online interactive writing system that instantiates all four interaction paradigms with a unified LLM backend (GPT-5.1). User experience is assessed using Likert-scale measures of perceived workload, usefulness, customization, and adapt-

Level	Window Switches ↓(count)	Editing Time ↓ (min)	Acceptance Rate ↑ (%)
L0	15.76	46.42	-
L1	8.46	32.85	18.55
L2	8.92	**31.07**	21.62
L3	**8.15**	31.28	**27.87**

Table 5: User study results based on interaction logs, reporting mean window switch counts, editing time, and acceptance rate averaged across all users under different interaction levels.

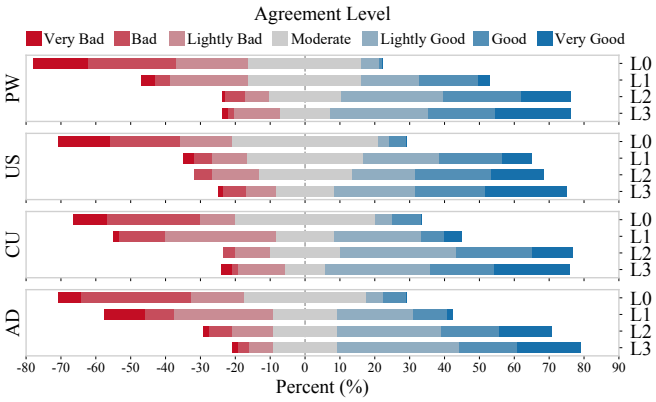

Figure 6: User questionnaire results across four interaction systems for perceived workload(PW), usefulness(US), customization(CU), and adaptability(AD). Blue indicates more favorable ratings, while red indicates less favorable ratings.

ability, following established human–AI interaction guidelines [Amershi *et al.*, 2019; Google, 2021; Apple, 2023; Pasch and Ha, 2025], and supplemented with qualitative feedback.

As shown in Figure 6 and Table 5, both subjective ratings and interaction logs exhibit consistent trends. Proactive paradigms (L1–L3) are preferred over user-initiated interaction (L0), with stateful and adaptive settings (L2, L3) achieving the lowest perceived workload, highest usefulness, and strongest adaptability. Interaction logs further corroborate these perceptions: proactive collaboration substantially reduces window switching and task completion time, while acceptance rates increase from stateless to stateful and adaptive settings. Together, these results confirm that the gains identified in quantitative analysis reflect genuine improvements in user experience rather than artifacts of offline metrics.

5.4 Learning with Interaction-Aware Rewards

Finally, we conduct reinforcement learning experiments to examine the practical learnability of our interaction-aware metric under the proposed formalization. We conduct reinforcement learning experiments on Qwen3-4B [Yang *et al.*, 2025] using HAR as the sole reward for continuation prediction, considering two training variants. In *stateless RL*, the model is optimized using the original query context, without access to prior interaction history. In *stateful RL*, interaction history is incorporated into the context during training. As shown in Table 6, both RL variants substantially improve HAR and reduce KED compared to the base model. Moreover, stateful RL consistently outperforms stateless RL,

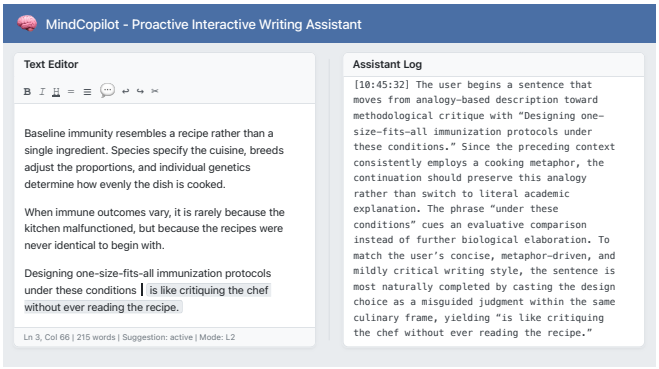

Figure 7: A stateful co-writing interface in which the assistant leverages interaction history to infer the user's preference for analogy-driven critique and predicts a style-consistent continuation.

Model	L1 Collaboration		L2 Collaboration	
	HAR(%) ↑	KED ↓	HAR(%) ↑	KED ↓
Qwen3-4B	1.48	14.79	0.95	15.66
+ Stateless RL	8.30	12.64	7.47	12.38
+ Stateful RL	**8.86**	**12.56**	**9.73**	**12.28**

Table 6: Performance of Qwen3-4B models after Stateless RL and Stateful RL. L1 and L2 Collaboration are defined in Table 3.

particularly under L2 collaboration, indicating that modeling interaction history further enhances learning. These results confirm that HAR can be effectively used as a reward for training, providing empirical support that our Human-in-the-Loop MDP formulation admits practical optimization and yields measurable gains in interactive writing performance.

6 Discussions and Conclusion

This work advances the study of human–AI co-writing by shifting the focus from isolated output quality to interaction-driven user behavior. By formalizing interactive completion as a Human-in-the-Loop Markov Decision Process, we provide a framework for analyzing how acceptance and editing actions reveal alignment and friction during writing.

The proposed Co-Writing Fidelity Suite evaluates collaboration through behavioral signals. *Hierarchical Acceptance Rate* captures the structured nature of human acceptance decisions, while *Knowledge-aware Editing Distance* quantifies post-acceptance cognitive effort. Across expert annotation, large-scale simulation, and a controlled user study, these metrics demonstrate stronger alignment with human judgment and user experience than output-only measures.

Using these metrics as analytical tools, we identify consistent behavioral patterns across interaction settings and writing stages, and derive interaction strategies that better match users' evolving intent. Importantly, systems informed by these metric-driven analyses yield higher perceived usefulness and lower cognitive burden in real writing tasks, supporting the external validity of the proposed evaluation approach.

Acknowledgements

We thank the anonymous reviewers and area chair for their helpful comments. This project is supported by the Shanghai Artificial Intelligence Laboratory.

Contribution Statement

Youqing Fang* and Yinhao Tang* contributed equally.

References

[Amershi *et al.*, 2019] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[Anthropic, 2025] Anthropic. Economic index – us usage, 2025.

[Apple, 2023] Apple. Human Interface Guidelines: Machine Learning. https://developer.apple.com/design/human-interface-guidelines/machine-learning, 2023. Accessed: 2026-01-13.

[Ashkinaze *et al.*, 2025] Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. How ai ideas affect the creativity, diversity, and evolution of human ideas: evidence from a large, dynamic experiment. In *Proceedings of the ACM collective intelligence conference*, pages 198–213, 2025.

[Bai *et al.*,] 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, 2024. *URL https://arxiv.org/abs/2408.07055*.

[Bai *et al.*, 2022] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[Chung *et al.*, 2025] John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yi Wang, Yuqian Sun, Tiffany Wang, Shm Garanganao Almeda, Brett A Halperin, Yuwen Lu, and Max Kreminski. Literarytaste: A preference dataset for creative writing personalization. *arXiv preprint arXiv:2511.09310*, 2025.

[Coenen *et al.*, 2021] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. Wordcraft: a human-ai collaborative editor for story writing. *arXiv preprint arXiv:2107.07430*, 2021.

[Comanici *et al.*, 2025] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[Devatine and Abraham, 2024] Nicolas Devatine and Louis Abraham. Assessing human editing effort on llm-generated texts via compression-based edit distance. *arXiv preprint arXiv:2412.17321*, 2024.

[Dhillon *et al.*, 2024] Paramveer S Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. Shaping human-ai collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.

[Du *et al.*, 2025] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

[Fang *et al.*, 2025] Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, et al. Creation-mmbench: Assessing context-aware creative intelligence in mllm. *arXiv preprint arXiv:2503.14478*, 2025.

[Fein *et al.*, 2025] Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing. *arXiv preprint arXiv:2507.00769*, 2025.

[Google, 2021] Google. People + AI Design Guidebook. https://pair.withgoogle.com/guidebook/patterns, 2021. Accessed: 2026-01-13.

[He *et al.*, 2025] Jessica He, Stephanie Houde, and Justin D Weisz. Which contributions deserve credit? perceptions of attribution in human-ai co-creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2025.

[Laban *et al.*, 2024] Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. Beyond the chat: Executable and verifiable text-editing with llms. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23, 2024.

[Le *et al.*, 2025] Khanh Chi Le, Linghe Wang, Minhwa Lee, Ross Volkov, Luan Tuyen Chau, and Dongyeop Kang. Scholawrite: A dataset of end-to-end scholarly writing process. *arXiv preprint arXiv:2502.02904*, 2025.

[Lee *et al.*, 2022] 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*, pages 1–19, 2022.

[Lee *et al.*, 2024] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–35, 2024.

[Li *et al.*, 2025] Yunwen Li, Shuangshuang Ying, Xingwei Qu, Xin Li, Sheng Jin, Minghao Liu, Zhoufutu Wen,

Tianyu Zheng, Xeron Du, Qiguang Chen, et al. Coig-writer: A high-quality dataset for chinese creative writing with thought processes. *arXiv preprint arXiv:2510.14763*, 2025.

[Li *et al.*, 2026] Pengcheng Li, Jie Zhang, Tianwei Zhang, Han Qiu, Weiming Zhang, Nenghai Yu, Wenbo Zhou, et al. State-dependent safety failures in multi-turn language model interaction. *arXiv preprint arXiv:2603.15684*, 2026.

[Liao *et al.*, 2025] Jianxing Liao, Tian Zhang, Xiao Feng, Yusong Zhang, Rui Yang, Haorui Wang, Bosi Wen, Ziying Wang, and Runzhi Shi. Rlmr: Reinforcement learning with mixed rewards for creative writing. *arXiv preprint arXiv:2508.18642*, 2025.

[Liu *et al.*, 2025] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[Ma *et al.*, 2024] Yongqiang Ma, Lizhi Qing, Jiawei Liu, Yangyang Kang, Yue Zhang, Wei Lu, Xiaozhong Liu, and Qikai Cheng. From model-centered to human-centered: Revision distance as a metric for text evaluation in llms-based applications. *arXiv preprint arXiv:2404.07108*, 2024.

[Mirowski *et al.*, 2023] 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*, pages 1–34, 2023.

[Mysore *et al.*, 2025] Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. Prototypical human-ai collaboration behaviors from llm-assisted writing in the wild. *arXiv preprint arXiv:2505.16023*, 2025.

[OpenAI, 2025] OpenAI. How people are using chatgpt, 2025.

[Pasch and Ha, 2025] Stefan Pasch and Sun-Young Ha. Human–ai interaction and user satisfaction: Empirical evidence from online reviews of ai products. *International Journal of Human–Computer Interaction*, pages 1–22, 2025.

[Payne *et al.*, 1993] John W Payne, James R Bettman, and Eric J Johnson. *The adaptive decision maker*. Cambridge university press, 1993.

[Que *et al.*, 2024] Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, et al. Hellobench: Evaluating long text generation capabilities of large language models. *arXiv preprint arXiv:2409.16191*, 2024.

[Reza *et al.*, 2025] Mohi Reza, Jeb Thomas-Mitchell, Peter Dushniku, Nathan Laundry, Joseph Jay Williams, and Anastasia Kuzminykh. Co-writing with ai, on human terms: Aligning research with user demands across the writing process. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–37, 2025.

[Shen and Wu, 2023] Hua Shen and Tongshuang Wu. Parachute: Evaluating interactive human-lm co-writing systems. *arXiv preprint arXiv:2303.06333*, 2023.

[Simon, 1955] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118, 1955.

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

[Tversky, 1972] Amos Tversky. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281, 1972.

[Wu *et al.*,] Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, et al. Writingbench: A comprehensive benchmark for generative writing, march 2025. *arXiv preprint arXiv:2503.05244*.

[Yang *et al.*, 2022a] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray Lc. Ai as an active writer: Interaction strategies with generated text in human-ai collaborative fiction writing. In *Joint proceedings of the ACM IUI workshops*, volume 10, pages 1–11. CEUR-WS Team, 2022.

[Yang *et al.*, 2022b] Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*, 2022.

[Yang *et al.*, 2023] 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)*, pages 3378–3465, 2023.

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

[Ying *et al.*, 2025] Shuangshuang Ying, Yunwen Li, Xingwei Qu, Xin Li, Sheng Jin, Minghao Liu, Zhoufutu Wen, Xeron Du, Tianyu Zheng, Yichi Zhang, et al. Beyond correctness: Evaluating subjective writing preferences across cultures. *arXiv preprint arXiv:2510.14616*, 2025.

[Zhang *et al.*, 2025] Shuning Zhang, Hui Wang, and Xin Yi. Exploring collaboration patterns and strategies in human-ai co-creation through the lens of agency: A scoping review of the top-tier hci literature. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–43, 2025.