SCHOLAWRITE: A Dataset of End-to-End Scholarly Writing Process

Anonymous ACL submission

Abstract

Writing is a cognitively demanding activity that requires constant decision-making, heavy reliance on working memory, and frequent shifts between tasks of different goals. To build writing assistants that truly align with writers' cognition, it is necessary to capture and analyze the complete thought process behind how writers transform ideas into final texts. We present `ScholaWrite`, the first dataset of *end-to-end scholarly writing*, tracing the multi-month journey from initial drafts to final manuscripts. The dataset traces nearly 62K LaTeX-based edits from five computer science preprints over four months and is enriched with fine-grained annotations of cognitive writing intentions. We demonstrate the value of ScholaWrite through three complementary contributions: (1) analysis of real-world writing behavior reveals that scholarly writing is highly non-linear and multi-intentional, blending rapid drafting bursts with cognitively sustained writing sessions; (2) evaluations of current large language models show that they struggle to provide meaningful support throughout the human writing process; and (3) models finetuned on SCHOLAWRITE demonstrate improved alignment with human writing workflows. SCHOLAWRITE underscores the value of capturing scientists' cognitive writing process and provides actionable insights and resources for the development of future writing assistants.

1 Introduction

Scientific writing is one of the most cognitively demanding forms of human communication. Researchers must articulate novel ideas with precision and structure, iteratively refining arguments, evidence, and phrasing over time (Jourdan et al., 2023; Kallestinova, 2011; Bourekkache, 2022). Unlike linear text generation, scholarly writing is a complex cognitive process, writers continually plan, draft, and revise across multiple stages and intentions (see Figure 1). Yet, despite the growing integration of large language models (LLMs) into research workflows, our understanding of how human scientists actually write remains limited. Without such understanding, it is difficult to build writing assistants that genuinely align with the cognitive processes of their users.

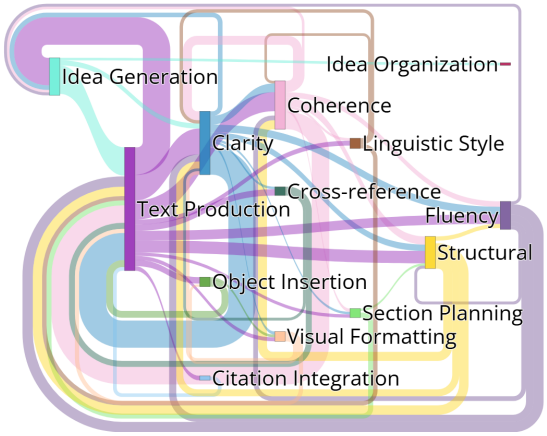


Figure 1: An example scholarly writing process with annotated writing intents in SCHOLAWRITE: it is iterative, non-linear, and switches frequently between multiple activities, tools, and intents over a long range of time.

Recent efforts in writing assistance have leveraged LLMs for tasks such as revision or feedback generation (Du et al., 2022b; Liang et al., 2024). However, these models largely operate autoregressively, producing text from left to right, while human writing unfolds through recursive, non-linear cycles of ideation, organization, and refinement (Flower and Hayes, 1981). This fundamental cognitive gap raises a key question: how can we empirically capture and model the end-to-end scholarly writing process so that future LLMs can support, rather than substitute, human cognition?

Existing work studies writing through revision logs or keystroke analyses (Leijten and Van Waes, 2013; Koo et al., 2023), but it primarily targets short tasks or version level comparisons of completed drafts (Du et al., 2022b; Jiang et al., 2022; D'Arcy et al., 2024), reflecting a broader lack of

datasets that capture scholarly writing as a longitudinal and intention driven process and thereby limiting fine grained analysis of writing behavior, evaluation of writing assistants, and training of models aligned with real writers' workflows.

To bridge this gap, we present SCHOLAWRITE, the first dataset capturing the end-to-end scholarly writing process, from initial drafts to final manuscripts, through in-situ keystroke logging and cognitive annotation. Rather than treating the dataset as an endpoint, we demonstrate its value through three complementary uses:

(1) Analysis of scholarly writing behavior. Our analyses reveal that scholarly writing is highly *nonlinear* and *multi-intentional*: more than half of all writing sessions involve three or more intertwined intentions. Writers dedicate most time to producing and revising text for clarity and coherence, while cognitively heavy tasks, such as idea organization or visual structuring, occur in fewer but longer, more focused sessions.

(2) Benchmarking cognitively aligned writing assistance. Evaluations of current LLMs show that they can imitate surface-level edits but *fail to predict next intentions* or *sustain cognitively complex revisions*, highlighting gaps between human writing processes and the current capabilities of LLMs in providing meaningful assistance.

(3) Training LLMs for cognitive support. LLMs finetuned on SCHOLAWRITE exhibit consistent gains in next-intention prediction and *intention-conditioned output alignment*. When deployed in iterative writing, these models adopt more human-like patterns – frequently alternating between Planning, Implementation, and Revision across stages – despite persistent gaps in surface-level text quality.

2 Related Work

Cognitive Theory of Writing Process Research on writing has shifted from analyzing final texts to examining cognitive processes across writing phases (Diederich, 1974; Krapels, 1990; MacArthur and Graham, 2016). Building on this process-oriented approach, Flower and Hayes (1981) outline three core sub-processes: *planning*, *translating*, and *reviewing*. These dynamic, nonlinear stages inform our work, where we extend this model into a finer-grained taxonomy of cognitive writing patterns in scholarly communication. Koo et al. (2023) proposed a taxonomy of scholarly writing based on keystroke data from short, 30-minute research-plan sessions. Expanding on this, we collect larger-scale, longitudinal keystroke data spanning months and culminating in published research. Expert-reviewed and grounded in prior theory, our taxonomy captures end-to-end cognitive trajectories of scholarly writing.

Keystroke Loggers for Scholarly Writing Keystroke logging tools (e.g., Inputlog) enable unobtrusive observation of digital writing (Chan, 2017; Johansson et al., 2010; Leijten and Van Waes, 2013; Lindgren and Sullivan, 2019). Yet, most operate in closed ecosystems like MS Word and are ill-suited for LaTeX-based academic writing or extended sessions. To address this, we built systems that securely record real-time keystrokes from Overleaf over months while preserving privacy. This workflow supports span-level annotation of writing intentions, providing a natural setting to study long-term cognitive writing activities across document sections.

Writing Datasets Publicly available datasets from previous work tend to track linguistic style changes or grammatical edits during revision (Du et al., 2022b; Jiang et al., 2022; Ito et al., 2019; Mita et al., 2022), while others capture edits based on feedback and peer review (D'Arcy et al., 2024; Jourdan et al., 2024; Kuznetsov et al., 2022) or citation generation (Kobayashi et al., 2022; Narimatsu et al., 2021). Also, most focus on specific sections of papers (e.g., abstracts, introductions) (Du et al., 2022b; Mita et al., 2022). In contrast, our dataset covers all cognitive phases of writing over an extended period and all sections. Furthermore, those existing corpora often compare multiple versions of final manuscripts from preprint databases (Jiang et al., 2022; Du et al., 2022b; D'Arcy et al., 2024; Jourdan et al., 2024), missing how those manuscripts evolved (Jourdan et al., 2023). To address this, we build a keystroke-based corpus that captures real-time progression of publications.

3 SCHOLAWRITE Dataset

SCHOLAWRITE represents the first large-scale publicly available dataset capturing the cognitive dynamics of end-to-end scholarly writing. It links every keystroke to the author's underlying writing intention, allowing direct observation of how ideas evolve into finalized text. Each annotated entry follows a standardized schema connecting text changes, contextual metadata, and intention

labels, as shown below, illustrating a Fluency edit, in which the writer corrects "expct" to "aspect."


```
{ "Project": 1,
   "timestamp": 1702958491535,
   "author": "author1",
   "before text": "One important expct of
       studying a LLMs is ..",
   "after text": "One important aspect of
       studying LLMs is ..",
   "label": "fluency" }
```


3.1 Data Collection & Annotation Process

Keystroke Collection in Overleaf To capture scholarly writing as it naturally unfolds, we developed a Chrome extension that records the end-to-end writing process on Overleaf, from individual keystrokes to their cognitive interpretation, without interrupting authors' workflow. Please refer to Appendix A.2 for detailed descriptions.

Participant Recruitment We recruited 10 graduate students in computer science at an R1 university in the United States, each actively writing manuscripts for peer-review AI and NLP conferences using Overleaf. The collection period spanned four months (November 2023 - February 2024). Our study has been approved by the IRB institution of the authors. Please refer to Appendix A.1 for more detailed descriptions.

Annotation Interface To decode the nuanced intentions behind each writing action, we annotate each keystroke collected from the Chrome extension, using a specialized interface. This interactive annotation interface that visualizes writing activity over time and across authors and files. Please see Appendix B for further annotation details

3.2 Intention Taxonomy

Building on cognitive theories of writing (Flower and Hayes, 1981) and recent empirical studies on scholarly revision (Du et al., 2022b; Koo et al., 2023), we developed a taxonomy of 15 distinct writing intentions grouped into three overarching categories (Table 1[1]). Following Pustejovsky et al. (2017), two annotators conducted iterative open coding on initial subsets of keystrokes to inductively identify recurring cognitive patterns. Each span of edits was treated as a *meaningful writing unit*, such as composing a sentence or improving phrasing, and assigned an appropriate intention label. Disagreements were resolved through multiple rounds of discussion with a cognitive linguist, who refined label definitions and boundaries to ensure coherence.

[1]Refer to Appendix Figure 7 for the full table with examples

1st	Intention	Definition
PLANNING	Idea Generation	Formulate and record initial thoughts and concepts.
	Idea Organization	Select the most useful materials and visually demarcate generated ideas.
	Section Planning	Initially create sections and sub-level structures.
IMPLEMENTATION	Text Production	Translate ideas into full language, either from the writer's own reasoning or borrowed sentences from external sources.
	Object Insertion	Insert visual representations of arguments (e.g., figures, tables, equations, lists).
	Citation Integration	Incorporate bibliographic references and link them using citation commands.
	Cross-reference	Link sections, figures, tables, or other elements via referencing commands.
	Macro Insertion	Incorporate predefined commands or packages into a LaTeX document.
REVISION	Fluency	Fix grammatical or syntactic errors in text or LaTeX commands.
	Coherence	Improve logical connections across sentences, paragraphs, or document elements.
	Clarity	Improve semantic relationships to make text more straightforward and concise.
	Structural	Improve information flow by modifying the placement of text or objects.
	Linguistic Style	Modify wording or style to better reflect the writer's stylistic preferences.
	Scientific Accuracy	Update or correct scientific evidence (e.g., numbers, equations).
	Visual Formatting	Modify stylistic formatting of text, objects, and citations.

Table 1: The developed taxonomy of scholarly writing processes in SCHOLAWRITE.

To validate the taxonomy, we applied two principles from taxonomy design (Nickerson et al., 2013; Kundisch et al., 2021): mutual exclusivity (each edit corresponds to one primary intention) and collective comprehensiveness (taxonomy covers all observed behaviors). After iterative refinement, we achieved a weighted F1 inter-annotator agreement of 0.71 on a 1K-keystroke subset, confirming strong reliability. See details in Appendix B.

Post-processing After annotation, we post-processed the data to improve usability and ensure privacy. All personal identifiers were removed, and keystrokes labeled as non-informative artifacts were filtered out (See details in Appendix A.3).

Project	1	2	3	4	5
# Authors	1 (3)	1 (4)	1 (3)	1 (4)	9 (18)
# Keystrokes	14,217	5,059	6,641	8,348	27,239
# Words added	17,387	23,835	7,779	12,448	57,511
# Words deleted	11,739	15,158	2,308	7,621	25,853

Table 2: Statistics of writing actions per Overleaf project in SCHOLAWRITE. "# Authors" indicates the number of participants who contributed data, with parentheses showing the total number of actual authors.

3.3 Overview

The final SCHOLAWRITE dataset contains 61,504 keystroke-based writing actions collected across five Overleaf projects, each of which culminated in an arXiv preprint authored by graduate researchers. Every action is annotated with a cognitive writing intention drawn from the 15-category taxonomy. Table 2 summarizes key statistics across projects. In total, the dataset captures over 62K text edits, encompassing more than 118K words added or deleted across multi-month writing trajectories. Each project contains thousands of temporally ordered keystrokes, providing a detailed chronicle of the writing process at the sentence, paragraph, and document levels.

3.4 Analytical Units: Capturing Writing Flow

Although SCHOLAWRITE records writing activity at millisecond, analysis of keystroke writing benefits from organizing this continuous stream into cognitively coherent units. We define two complementary units: *intention sessions* and *sessions*.

Intention session. An intention session is a continuous period during which the writer focuses on a *single cognitive goal*, such as producing text, before switching to another intention. These sessions typically last from a few seconds to several minutes, capturing a concentrated burst of focused cognitive activity.

Session. A session *aggregates all writing actions within a continuous working period*, regardless of how many intentions are involved. It represents the natural rhythm of writing, how authors weave together planning, drafting, and revising while working on a particular section or idea.

These two units capture both micro-level cognitive focus and macro-level writing flow, forming the backbone of our analyses in §4.

4 Analysis of Scholarly Writing Behavior

SCHOLAWRITE offers an unprecedented window into the dynamic, cognitive process of scholarly writing. Our analysis pursues two complementary goals: (1) to uncover how scholarly writing intentions evolve, interact, and shift across time, and (2) to inform the design of cognitively aligned writing assistants capable of supporting real-world research workflows. We focus on three research questions:

- (§4.1) What writing intentions dominate the scholarly writing process?
- (§4.2) How are different intentions intertwined and sequenced in practice?
- (§4.3) How do these patterns change across phases of manuscript development?

4.1 Which Writing Intentions Dominate?

We examine which writing intentions account for the largest share of time and cognitive effort during scholarly writing.

Time Distribution across Intentions Overall, writing activity is dominated by Text Production and Revision, while Planning activities plays a comparatively smaller role (See Appendix Table 10 for the average share of time spent per intention). These results suggest that scholars devote most of their time to articulating and refining text rather than explicit planning. The considerable effort spent on revision and formatting underscores that much of scholarly writing involves clarifying and visually shaping ideas, an aspect often overlooked in computational models of writing.

Session Duration and Cognitive Effort We analyze the duration of each *intention session*. Intention sessions associated with planning and structural work – such as Idea Organization, Object Insertion, and Visual Formatting – tend to last longer and exhibit greater variability, reflecting the sustained cognitive effort required for reorganizing content or managing visual elements. (See Appendix Table 11 for intention session level time statistics).

In contrast, Text Production, while accounting for the largest share of overall writing time, typically unfolds in short, frequent bursts. Together, these patterns reveal two complementary cognitive modes in scholarly writing: (1) rapid, iterative drafting episodes, and (2) longer, focused periods of planning or visual composition that demand sustained attention.

4.2 How Are Writing Intentions Intertwined?

Scholarly writing rarely unfolds linearly. Instead, it involves recursive alternations between complementary intentions, writers generate ideas, structure them, produce text, and immediately refine it. As seen in Figure 1 (and other projects in Appendix Figure 12), the intertwined processes of writing reveals dense bidirectional flows across intentions.

Transition Patterns The transition probability matrix (Appendix Figure 13) shows that Planning and Implementation are tightly coupled. Planning activities, such as *Idea Generation*, are frequently followed by *Text Production*. Within Implementation, transitions predominantly return to Text Production, whereas Revision intentions exhibit strong recursive loops – particularly among *Clarity*, *Fluency*, and *Coherence* – reflecting iterative refinement of language and argumentation. Overall, the results indicate that planning, implementation, and revision are not distinct stages but interdependent cycles continually revisited throughout the writing process.

Multitasking and Alternation Patterns Writing sessions are highly multitasking in nature. As shown in Figure 2, 57% of all sessions involve three or more distinct intentions, reflecting frequent switching between cognitive modes within a single sitting. This dynamic interplay reveals how writers manage multiple goals in parallel.

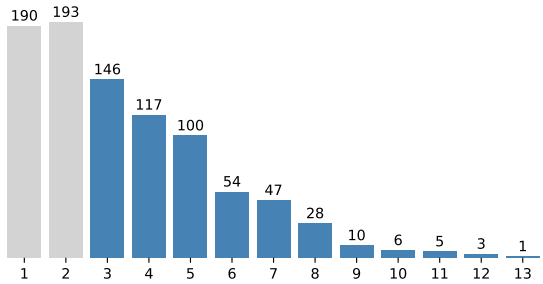


Figure 2: Distribution of the number of intentions per writing session. The x-axis shows the number of intentions within a session, and the y-axis shows the number of sessions exhibiting that count.

4.3 Are writing patterns stable throughout the process, or do they change across phases?

Scholarly writing evolves as authors move from idea formation to refinement. To examine how writing behavior changes across the writing process, we compare the *early phase* (first third of project duration) with the *late phase* (final third).

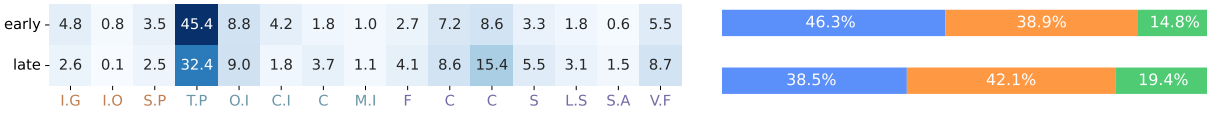
Intention	Early	Late	Trend
Idea Gen.	1.6 (23)	**2.5 (11)**	longer, fewer
Idea Org.	**2.6 (2)**	—	early-only
Sect. Plan.	1.6 (13)	**1.0 (17)**	more, shorter
Text Prod.	2.1 (221)	2.0 (227)	stable
Obj. Insert.	2.4 (28)	**3.4 (42)**	up both
Cite Int.	1.4 (16)	1.8 (6)	fewer later
Cross-ref.	0.7 (8)	**1.2 (18)**	more later
Clarity	1.5 (43)	**1.3 (99)**	more, quicker
Structural	2.4 (8)	**1.4 (29)**	more later
Ling. Style	3.5 (3)	**1.7 (12)**	concise edits
Vis. Form.	3.6 (16)	**1.7 (43)**	quick fixes

Table 3: Early-late phase shift in writing intentions (mean duration [min]; counts in parentheses). Late writing involves more frequent, shorter revision and formatting sessions.

Phase-dependent intention patterns Writing intentions exhibit clear stage-dependent shifts (Figure 3a). Planning intentions are more frequent early, reflecting front-loaded structuring. Within Implementation, *Citation Integration* is higher early, while *Cross-referencing* increases late as authors add visuals and refine internal links. Revision grows sharply in the late phase, consistent with end-stage polishing for coherence and presentation.

Table 3 summarizes session-level differences. In Planning, *Idea Generation* and *Idea Organization* sessions decrease in count but lengthen, suggesting deeper cognitive engagement later. *Section Planning* rises as writers reorganize material for submission. For Revision session counts increase substantially, though durations shorten, indicating rapid, localized edits characteristic of final revisions.

How intention interactions change over time? Figure 14 in Appendix compares transition probabilities across phases. Early writing features strong flows into Implementation, whereas late writing shifts toward Revision, marking the transition from building content to refining it. Some intention paths also evolve – for example, transitions from Macro Insertion lead to Idea Generation early but to Section Planning late, as layout decisions become more prominent. Writers also juggle more intentions later in the process as shown in Figure 3b, reflecting increased multitasking and integration of planning, drafting, and revising.



(a) Share of time by intention across early and late phases.

(b) Distribution of number of intentions per session.

Figure 3: Dynamics across early and late phases of writing. (a) The share of time devoted to each intention shifts from planning to revision as writing progresses. (b) Later sessions involve more overlapping intentions (blue for 1-2, orange for 3-5, and green for >5 intentions), reflecting higher cognitive integration.

5 Benchmarking and Improving AI Writing Assistants

Scientific writing is a cognitively demanding, non-linear, and phase-dependent process in which writers continuously shift between intentions over time. Using SCHOLAWRITE, we therefore evaluate and improve cognitively aligned writing assistance along three complementary dimensions: a model's ability to predict writers' next intentions (§5.1), to align its outputs with human writing actions at different temporal granularity (§5.2), and to test how intention prediction and output alignment interact during iterative writing (§5.3) (Figure 4).

5.1 Can models anticipate what writers do next?

Because writing involves frequent task-switching, an effective assistant should anticipate the writer's next move. Using SCHOLAWRITE, we evaluate whether state-of-the-art language models can infer writers' upcoming intentions from ongoing writing context, and we further examine the extent to which this capability improves through finetuning.

Setup & Metrics This task takes the "before-text" from a keystroke pair and context prompt, and predicts the writing intention to apply for the subsequent actions.

We evaluate GPT-5 (OpenAI, 2025) and Qwen2.5-14B-Instruct (Yang et al., 2024) in a zero-shot setting for next intention prediction.

To assess the impact of finetuning on SCHOLAWRITE, we use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and Llama-3.1-8B-Instruct (Dubey et al., 2024) as trainable baselines, fine-tuning each on the SCHOLAWRITE training split. See Appendix G.2 for finetuning details.

To evaluate model performance, we used a weighted F-1 score[2].

[2]Weighted F-1 was chosen to address skewed label distribution.

Findings *Finding 1: Current state-of-the-art language models struggle to accurately anticipate writers' intentions during the writing process.* When evaluated on next-intention prediction, state-of-the-art models such as GPT-5 and Qwen-2.5-14B-Instruct still perform poorly on this task, achieving weighted F1 scores below 0.5 (Table 4).

Finding 2: Finetuning on SCHOLAWRITE *substantially improves models' ability to predict writers' intentions.* Regardless of the intricate nature of the task itself[3], all models finetuned on SCHOLAWRITE show an improved performance compared to their baselines (Table 4). BERT and RoBERTa achieved the most improvement, while LLama-8B-Instruct showed a modest improvement after fine-tuning. Those results demonstrate the effectiveness of our SCHOLAWRITE dataset to align language models with writers' intentions.

5.2 Can models align their output with human writing actions?

To assist writers effectively, a model must not only infer their intentions but also generate text aligned with those goals. We evaluate output alignment at two temporal granularities: at the *session level*, which assesses whether model-generated text is consistent with the overall writing actions observed in a human session, and at the *keystroke level*, which examines alignment at the level of fine-grained, moment-to-moment writing actions.

Setup & Metrics For each session, models receive the "before-text" and either the sequence of intentions observed in that session (session-level) or just a single intention (keystroke-level), then generate the corresponding "after-text." Output alignment is evaluated using two complementary metrics: semantic similarity, measured by BERTScore-F1 (Zhang et al., 2019), and operational similarity, measured by Levenshtein distance.

[3]Each model predicts the next intention using only the "before" text, while human annotators consider multiple "before and after" edits. Moreover, the chosen next intention is not necessarily the only correct one.

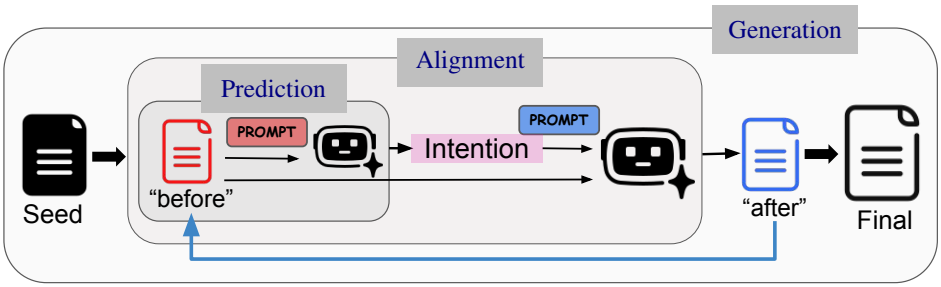


Figure 4: Overview of our AI writing assistant evaluation setup: intention prediction (§5.1), output alignment (§5.2), and iterative writing (§5.3).

	BERT	RoBERTa	Llama-8B	GPT-5	Qwen-14B
Base	0.04	0.02	0.12	0.41	0.08
+ SW	**0.64**	**0.64**	**0.13**	–	–

Table 4: Weighted F1 scores for next-intention prediction. "+SW" = fine-tuned on SCHOLAWRITE.

	BERTScore-F1	Levenshtein
Llama-8B-Zero	0.91	0.64
Llama-8B-SW	**0.99**	**0.98**

Table 5: Keystroke-level output alignment results

	Automatic Evaluation			Human Evaluation		
	Lexical Diversity	Topic Consistency	Intention Coverage	Accuracy	Alignment / Fluency / Coherence	Relevance
Llama-8B-Zero	0.20	0.66	6.75	**38.4**	**3**	**3**
Llama-8B-SW	**0.43**	**0.67**	**9.75**	16.2	0	1.75

Table 6: Automatic and human evaluation results for iterative writing across four seed documents. Accuracy reports the average number of generated keystrokes with correctly inferred intentions. Other metrics report the number of human evaluators who agreed on the model's performance.

5.2.1 Coarse-grained alignment (session-level)

Session-level alignment is necessary to test whether models can align with human writing goals at the level of overall writing outcomes.

We report main results using GPT-5 as the representative model and evaluate alignment with human writing using BERTScore-F1. Results for other model (Qwen2.5-14B-Instruct) and metric (Levenshtein distance) appear in the Appendix F.3 and show consistent patterns.

Finding *Model-human alignment degrades as writing sessions become longer or cognitively more complex* (Appendix Figure 15). As shown in Appendix Figure 16, GPT-5 (and Qwen) perform best on Revision intentions (such as *Clarity*, *Fluency*) but struggle with Planning and Implementation, which demand deeper reasoning and structural organization. Alignment further drops when multiple intentions co-occur within a session, indicating models' limited ability to integrate overlapping cognitive goals. Overall, current LLMs excel at surface refinement but remain weak at sustaining the intertwined reasoning processes that drive human scholarly writing.

5.2.2 Fine-grained alignment (keystroke-level)

While session-level alignment evaluates overall writing outcomes, keystroke-level alignment allows us to study how models can be fine-tuned to align with the fine-grained writing actions that give rise to those outcomes.

We fine-tune Llama-3.1-8B-Instruct (Llama-8B-SW) and compare it against the vanilla Llama-3.1-8B-Instruct (Llama-8B-Zero). All models use the same train-test split (80%, 20%) across experiments. Training details are provided in Appendix G.

Finding *Models finetuned on* SCHOLAWRITE *generate outputs that are more accurately aligned with human writing actions and better positioned within the writing workflow.* Table 5 shows that (Llama-8B-SW) achieves higher BERTScore-F1 and Levenshtein ratio than the base model, indicating improved semantic alignment and closer correspondence to human editing actions. This suggests that SCHOLAWRITE fine-tuning helps models generate outputs that better match both the content and the placement of human revisions.

5.3 Does alignment persist across iterative writing?

Iterative writing requires jointly predicting writers' intentions and generating text that is aligned with those intentions across multiple steps of revision. We use this setting to test whether our SCHOLAWRITE-fine-tuned model can sustain intention-aware, aligned generation over extended writing trajectories.

Setup & Metrics During iterative self-writing, models process LaTeX-formatted seed document (as "before-text") with a context prompt to predict the next writing intention, then revises the text ("after-text") accordingly given prompt. The revised document then serves as the new seed for the next iteration. This process repeats until a set iteration limit (e.g., 100) is reached.

We compare the same finetuned Llama-8B-SW model (from 5.2.2) to Llama-8B-Zero. Also, seed documents were derived from LaTeX-formatted abstracts of four award-winning NLP papers on diverse topics (Zeng et al., 2024; Lu et al., 2024; Du et al., 2022a; Etxaniz et al., 2024), as shown in Appendix Listings 1-4.

We evaluated *lexical diversity* (unique-to-total token ratio), *topic consistency* (cosine similarity between seed and final output), and *intention coverage* (unique writing intentions used over 100 iterations). For **human evaluation**, three native English speakers[4] with LaTeX expertise assessed outputs from Llama-8B-Zero and Llama-8B-SW on *accuracy* (alignment with predicted intention), *alignment* (similarity to human writing), *fluency* (grammatical correctness), *coherence* (logical structure), and *relevance* (connection to the seed document) (Refer to Appendix G.7).

Finding *Finetuning improves alignment with human writing behavior in iterative settings, even though surface-level text quality remains limited.* Llama-8B-SW shows consistent gains on automatic measures (Table 6). At the process level, it also displays writing activity patterns that more closely resemble human behavior, frequently alternating between implementation and revision and engaging all three high-level writing processes over time (Appendix Figure 23). In contrast, Llama-8B-Zero tend to remain within a single dominant stage throughout iterative writing.

Human evaluation reveals that, despite these process-level improvements, Llama-8B-SW still lags in surface-level fluency, logical coherence, and consistent realization of predicted intentions (Appendix Figure 22). Nevertheless, it produces more relevant content in some settings, aligning with observed gains in topic consistency.

Together, these results suggest that fine-tuning on SCHOLAWRITE enables models to better emulate the process of human scholarly writing, even as challenges in text quality remain.

6 Conclusion and Future Work

This work introduces SCHOLAWRITE, the first dataset to capture the *end-to-end cognitive process* of scholarly writing through unobtrusive, in-situ logging of real writing activity paired with a cognitively grounded annotation taxonomy. By focusing on how text is produced, rather than solely on the final output, SCHOLAWRITE provides a foundation for studying, evaluating, and training writing assistants grounded in real human writing workflows.

Using SCHOLAWRITE, we demonstrate that scholarly writing is fundamentally non-linear and multi-intentional, and phase-depedent. We further show that, despite strong text generation capabilities, current large language models remain misaligned with human writing workflows, struggling to support writing at the process level. Finally, we find that models trained on SCHOLAWRITE exhibit improved alignment with human writing behavior, suggesting that exposure to process-level supervision enables models to better adapt to evolving writing goals.

Together, SCHOLAWRITE contributes: (1) a **fine-grained lens** on the cognitive dynamics of scholarly writing; (2) a **benchmark** for evaluating LLMs' alignment with human writing intentions; and (3) a **training resource** for developing models that adapt to evolving, multi-intent writing behavior.

Future work will extend beyond keystroke logs to include pre-writing ideation, collaboration, and multimodal composition (e.g., figure and data integration). By tracing the full scholarly writing lifecycle, we aim to deepen our understanding of human cognitive patterns and to design AI writing assistants that not only complete text, but truly *complement* how humans think and write.

[4]The IAA scores are 0.84 (SW) and 0.76 (Zero) for accuracy, all 100% for alignment, fluency, coherence, and 49.8% (SW) and 100% (Zero) for relevance.

Limitations and Ethical Considerations

We acknowledge several limitations in our study. First, the SCHOLAWRITE dataset is **currently limited to the computer science domain**, as LaTeX is predominantly used in computer science journals and conferences. This domain-specific focus may restrict the dataset's generalizability to other scientific disciplines. Future work could address this limitation by collecting keystroke data from a broader range of fields with diverse writing conventions and tools, such as the humanities or biological sciences. For example, students in humanities usually write book-length papers and integrate more sources, so it could affect cognitive complexities.

Second, our dataset includes **contributions from only 10 participants, resulting in five final preprints on arXiv**. This small-to-medium sample size is partly due to privacy concerns, as the dataset captures raw keystrokes that transparently reflect real-time human reasoning. To mitigate these concerns, we removed all personally identifiable information (PII) during post-processing and obtained full IRB approval for the study's procedures. However, the highly transparent nature of keystroke data may still have discouraged broader participation. Future studies could explore more robust data collection protocols, such as advanced anonymization or de-identification techniques, to better address privacy concerns and enable larger-scale participation. We also call for community-wise collaboration and participation for our next version of our dataset, SCHOLAWRITE 2.0 and encourage researchers to contact authors for future participation.

Furthermore, **all participants were early-career researchers** (e.g., PhD students) at an R1 university in the United States. Expanding the dataset to include senior researchers, such as post-doctoral fellows and professors, could offer valuable insights into how writing strategies and revision behaviors evolve with research experience and expertise. Despite these limitations, our study captured an end-to-end writing process for 10 unique authors, resulting in a diverse range of writing styles and revision patterns. The dataset contains approximately 62,000 keystrokes, offering fine-grained insights into the human writing process, including detailed editing and drafting actions over time. While the number of articles is limited, the granularity and volume of the data provide a rich resource for understanding writing behaviors. Prior research has shown that detailed keystroke logs, even from small datasets, can effectively model writing processes (Leijten and Van Waes, 2013; Guo et al., 2018; Vandermeulen et al., 2023). Unlike studies focused on final outputs, our dataset enables a process-oriented analysis, emphasizing the cognitive and behavioral patterns underlying scholarly writing.

Third, **collaborative writing is underrepresented** in our dataset, as only one Overleaf project involved multiple authors. This limits our ability to analyze co-authorship dynamics and collaborative writing practices, which are common in scientific writing. Future work should prioritize collecting multi-author projects to better capture these dynamics. Additionally, the dataset is **exclusive to English-language writing**, which restricts its applicability to multilingual or non-English writing contexts. Expanding to multilingual settings could reveal unique cognitive and linguistic insights into writing across languages.

Finally, the human evaluation process in Section G.7 was determined as exempt from IRB review by the authors' primary institution, while the data collection using our Chrome extension program was fully approved by the IRB at our institution. Importantly, no LLMs were used during any stage of the study, except for grammatical error correction in this manuscript.

References

Samir Bourekkache. 2022. English for specific purposes: writing scientific research papers. case study: Phd students in the computer science department.

Sathena Chan. 2017. Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia*, 7:1–27.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2024. ARIES: A corpus of scientific paper edits made in response to peer reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6985–7001, Bangkok, Thailand. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Paul B Diederich. 1974. Measuring growth in english.

Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022a. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 96–108, Dublin, Ireland. Association for Computational Linguistics.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022b. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Linda Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387.

Hongwen Guo, Paul D Deane, Peter W van Rijn, Mo Zhang, and Randy E Bennett. 2018. Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, 55(2):194–216.

Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. Diamonds in the rough: Generating fluent sentences from early-stage drafts for academic writing assistance. *arXiv preprint arXiv:1910.09180*.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the human revision process in scientific writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9420–9435, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Roger Johansson, Åsa Wengelin, Victoria Johansson, and Kenneth Holmqvist. 2010. Looking at the keyboard or the monitor: relationship with text production processes. *Reading and writing*, 23:835–851.

Léane Jourdan, Florian Boudin, Richard Dufour, and Nicolas Hernandez. 2023. Text revision in scientific writing assistance: An overview. *arXiv preprint arXiv:2303.16726*.

Léane Jourdan, Florian Boudin, Nicolas Hernandez, and Richard Dufour. 2024. Casimir: A corpus of scientific articles enhanced with multiple author-integrated revisions. *arXiv preprint arXiv:2403.00241*.

Elena D Kallestinova. 2011. How to write your first research paper. *The Yale journal of biology and medicine*, 84(3):181.

Keita Kobayashi, Kohei Koyama, Hiromi Narimatsu, and Yasuhiro Minami. 2022. Dataset construction for scientific-document writing support by extracting related work section and citations from PDF papers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5673–5682, Marseille, France. European Language Resources Association.

Ryan Koo, Anna Martin, Linghe Wang, and Dongyeop Kang. 2023. Decoding the end-to-end writing trajectory in scholarly manuscripts. *arXiv preprint arXiv:2304.00121*.

Alexandra Rowe Krapels. 1990. Second language writing: An overview of second language writing process research.

Dennis Kundisch, Jan Muntermann, Anna Maria Oberländer, Daniel Rau, Maximilian Röglinger, Thorsten Schoormann, and Daniel Szopinski. 2021. An update for taxonomy designers: methodological guidance from information systems research. *Business & Information Systems Engineering*, pages 1–19.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, and 1 others. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Eva Lindgren and Kirk Sullivan. 2019. *Observing writing: Insights from keystroke logging and handwriting*, volume 38. Brill.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Liang Lu, Peirong Xie, and David Mortensen. 2024. Semisupervised neural proto-language reconstruction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14715–14759, Bangkok, Thailand. Association for Computational Linguistics.

Charles A MacArthur and Steve Graham. 2016. Writing research from a cognitive perspective.

Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Towards automated document revision: Grammatical error correction, fluency edits, and beyond. *arXiv preprint arXiv:2205.11484*.

Hiromi Narimatsu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minami, and Hirotoshi Taira. 2021. Task definition and integration for scientific-document writing support. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 18–26.

Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359.

OpenAI. 2025. Gpt-5 is here. Accessed: 2025-10-3.

James Pustejovsky, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. *Handbook of Linguistic Annotation*, pages 21–72.

Nina Vandermeulen, Elke Van Steendam, Sven De Maeyer, and Gert Rijlaarsdam. 2023. Writing process feedback based on keystroke logging and comparison with exemplars: Effects on the quality and process of synthesis texts. *Written Communication*, 40(1):90–144.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A More About Data Collection Process

A.1 Participant Recruitment & Demographics

We recruited ten graduate students in the computer science department who actively prepared their manuscripts in Overleaf, an online LaTeX editor, and who aimed to submit their manuscripts to peer-reviewed conferences. We held a consent procedure with each participant through a 30-minute virtual meeting remotely. After the consent process, we installed on their computers a Chrome extension program that we designed and implemented only for this study and asked for the ID number of only the Overleaf projects that the participants agreed to share as their manuscripts. We provided all participants with a $100 Amazon gift card per project, which could be divided among the authors involved in the project. Our data collection process is approved by the IRB of the authors' primary institution.

All ten participants of our study are graduate students who currently study computer science domain at an accredited university in the United States. Out of the ten, two of them identified themselves as native English speakers, and the remaining participants identified themselves as proficient in English in terms of writing. Also, two of the ten participants attained a Master of Science degree in Computer Science with several publication experiences, and the remaining eight of them are currently PhD students with extensive research experiences.

A.2 Technical Details of System Implementations

We built a custom Chrome extension (Appendix Figure 6) that unobtrusively logs real-time keystroke trajectories from Overleaf. The extension operates in the background once participants consent and authenticate via unique credentials. Every time a key-up event occurs, the system captures the visible text within the user's Overleaf editor, compares it with the previous version using Google's diff_match_patch algorithm, and stores the resulting text differences alongside metadata such as timestamp, file name, and author ID. To protect privacy, only project IDs pre-approved through participant consent were collected; any non-consented Overleaf projects were automatically filtered out. All data were stored on a secure institutional server and anonymized before analysis.

When a key-up event fires in a browser, the extension collects the writer's viewable texts in the code editor panel[5]. When each of these actions[6] occurs, the extension uses '`diff_match_patch`' package[7] to generate an array of differences between two subsequent texts (i.e., Figure 5). Then, the extension will send the array along with metadata (e.g., time stamp, author ID, etc.) to the backend server.

For any Overleaf project that consists of multiple LaTeX files, we also collected all keystrokes from subfiles associated with the main LaTeX file. Our comprehensive data collection process captures the end-to-end writing processes of the participating authors across all parts of Overleaf projects. This approach ensures that our dataset reflects the full scope of scholarly writing including edits made in auxiliary files such as files of each section, appendix, bibliography, etc.

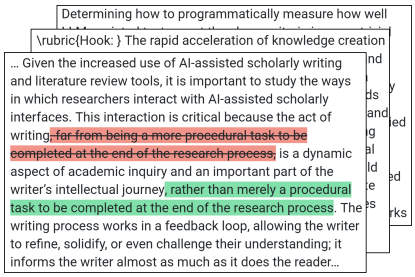


Figure 5: The array of differences between two subsequent texts, generated by `diff_match_patch`

We explain the technical implementation details of the two systems for the data collection process. For the Chrome extension, we implemented a backend application using Flask and Python and stored all keystroke data in the MongoDB database.

For the annotation interface, we used HTML/CSS and JavaScript for the client side and Flask for the backend. All data for the annotation interface was retrieved from the MongoDB database used in the Chrome extension system.

Privacy Concerns To prevent any issue of private data collection, we designed the backend of our Chrome extension to fetch only the IDs of the Overleaf projects that participants consented to share during the recruitment process and filter out participants' keystroke data from any unauthorized

[5]To prevent privacy concerns, the extension filters out keystroke data from any unauthorized Overleaf projects. Please see Appendix A.2 for more details.

[6]Example actions are (1) inserting a space/newline; (2) copy/paste; (3) undo/redo; (4) switching files and (5) scrolling a page.

[7]https://github.com/google/diff-match-patch

projects. We used Google Sheet API to retrieve ID information that we collected during the recruitment process.

A.3 Data Post-Processing

For use during the annotation phase, each keystroke entry from the raw collection includes the following fields: (1) a valid file name; (2) a valid writing action that triggered keystroke logging (e.g., copy, paste, typing, etc.); (3) a valid array of differences to enable visualization of writing trajectories; and (4) the line numbers in the Overleaf editor. Data entries annotated with a valid intention label (i.e., labels except 'artifact') and having a difference array length of fewer than or equal to 300 are then used for model training.

Regarding the additional postprocessing for public use, we took the following steps to post-process our data with the annotations to promote the usability of our dataset and prevent any privacy issues. For the annotation data, we only include the keystroke changes, anonymized project ID, timeframe information, and anonymized author's name (e.g., 0, 1, 2, etc.) from the metadata. Then, we extract the before and after texts from the differences array. We also include the annotated intention label for each entry.

Then, we analyzed any 'artifact' generated due to natural keyboard/mouse activities or user switching files, and we discarded them as they are not informative to any writing intention in our taxonomy. Lastly, to prevent any privacy issues we removed keystrokes containing any private author information such as names, affiliations, contact information, and any personally identifiable information (PID) from the collected keystrokes. Instead, we replaced those with an arbitrary command (e.g., '`\anonymous`').

B More About Writing Intention Annotations & Taxonomy

B.1 Annotator Recruitment

Due to privacy concerns, we did not hire external freelancers with expertise, rather the two corresponding authors of this paper annotated the data are graduate students who possess extensive scholarly writing experiences in natural language processing and data annotation skills. The raw keystroke data collected by our Chrome extension could potentially contain personally identifiable information, such as specific content edits or metadata that could reveal the identity of the authors. To ensure the confidentiality and ethical handling of sensitive information, we restricted access to the data to the authors only. This annotation process was also authorized by the IRB of the authors' institution. Please note that the final dataset which will be released publicly is ensured not to contain any PII information through several sophisticated post-processing steps.

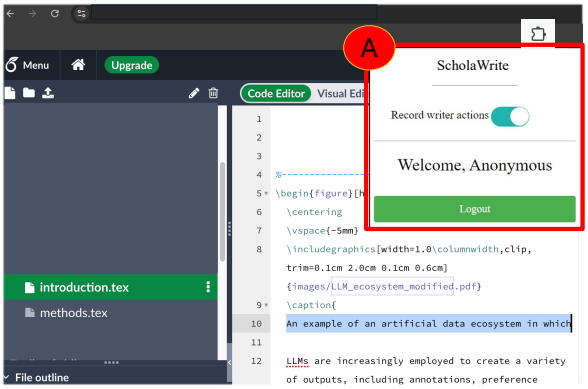


Figure 6: The Chrome extension interface (A) on the Overleaf project, where it collects real-time keystrokes in the Overleaf editor (highlighted).

B.2 Annotation Interface

The annotation process transforms raw writing traces into interpretable cognitive data, forming the foundation of the SCHOLAWRITE taxonomy and subsequent analyses. Each keystroke record contains metadata, such as file name, type of action, text differences between two states, and line number in the Overleaf editor, which allows precise reconstruction of how writers iteratively develop and refine their manuscripts. This process is crucial for developing a detailed taxonomy of the end-to-end scholarly writing process, which serves as the basis for annotating the collected keystrokes. To perform annotations, we processed those raw keystroke entries by file name, type of writing actions, an array of differences between two subsequent texts, and line number in an Overleaf editor (Figure 7). Annotators can navigate a timeline of keystrokes, examine spans of related edits (e.g., drafting a sentence or clarifying an argument), and assign one or more intention labels from a predefined taxonomy via a dropdown menu.

B.3 Detailed Annotation Process

The two authors (or annotators) collaborated with a cognitive linguist to develop a codebook and review the results of the annotations. Also, the

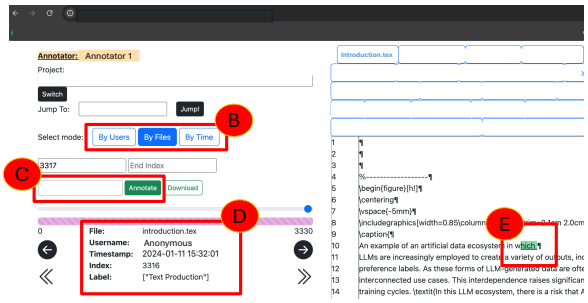


Figure 7: Annotation interface. During the annotation stage, annotators can click a viewing mode of the collected keystroke data (B). By right-clicking to navigate the timeline of keystroke trace in the interactive panel on the right side (E), annotators can choose an intention label under the drop-down menu (C). They can also view the meta-information of each annotated keystroke (D).

two annotators conducted an iterative open coding approach to identify several unique writing intentions from keystrokes and developed a codebook of intention labels ("ground-truth labels") within each high-level process (Planning, Implementation, and Revision) based on the findings from Flower and Hayes (1981); Koo et al. (2023). Using this codebook, those annotators re-labeled each span of keystrokes with the corresponding label during the annotation process.

The annotators were fully informed about all the labels and had complete access to them when annotating each data point. The annotation process for all the labels is the same: First, they view through multiple consecutive data points and identify which high-level label occurs (e.g., Planning, Implementation, or Revision). Once annotators have identified the current high-level label, attempting to identify where it ends. Then, they decide on the low-level label within the high-level label (e.g., idea generation or organization under the Planning stage, etc.). Finally, they identify the interval for low-level labels and annotate data points in the interval with the identified low-level label. If a keystroke does not deliver any insight, then label it as an 'artifact.'

We calculated **inter-annotator agreement** using the weighted F1 score in a multi-label, multi-class setting, which is suitable for our complex annotation schema involving multiple labels per instance. The weighted F1 score achieved was 0.71, indicating a high level of agreement between the annotators.

C More About the Taxonomy of Scholarly Writing Process

During the **planning** stage, the writers engage in a process of generating and organizing raw ideas, arguments, or content structures that were not introduced in the previous trajectory. Based on the plan, the writers **implement** their plan by drafting full sentences and paragraphs and structuring the contents tangibly. At the same time, the writers enter the **revision** stage by improving the quality of their implemented sentences and LaTeX objects in terms of linguistic styles, format, or information accuracy. Particularly, spans of keystrokes whose intentions involved any changes but did not change the meaning of original texts are classified as **Revision**. For those edits that show changes in the meaning, we considered them as **Implementation**. Furthermore, if an author repeatedly adds, removes, and revises text back and forth until a sentence is completed, we consider this process as part of **text production**. Any subsequent changes made to the sentence after it is finished are considered **revision**. Table 7 presents the comprehensive, complete definitions of each intention of end-to-end scholarly writing process, identified from SCHOLAWRITE dataset.

D SCHOLAWRITE dataset statistics

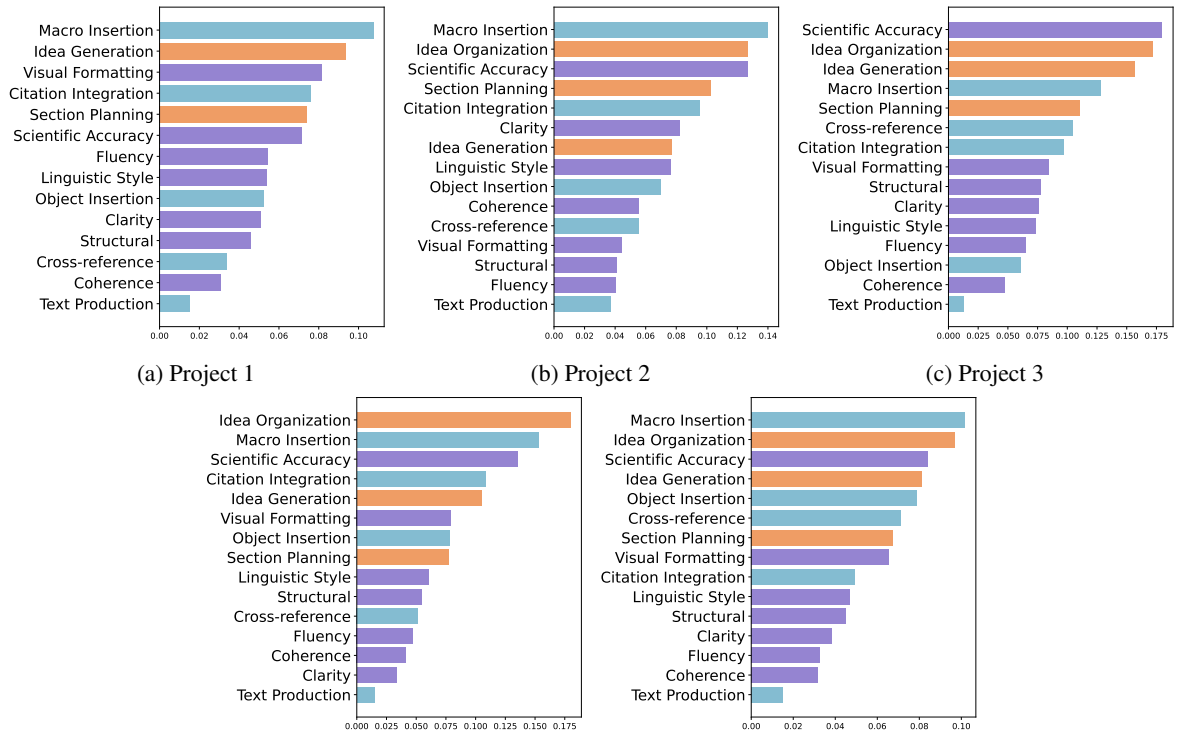
Table 9 shows the distribution of intention labels per Overleaf project from SCHOLAWRITE.

Figures 8 to 11 show several characteristics of human writing process, analyzed from SCHOLAWRITE DATASET: (1) Figure 8 - the average Wasserstein distance between each intention distribution and uniform distribution; (2) Figure 9 - distribution of labels over time; (3) Figure 10 for high-level intention distribution over time; and (4) Figure 11 for intention-wise writing activity distribution over time.

E More about Analyses

Figure 14 shows transition probability between pair of intentions in early stage and later stage.

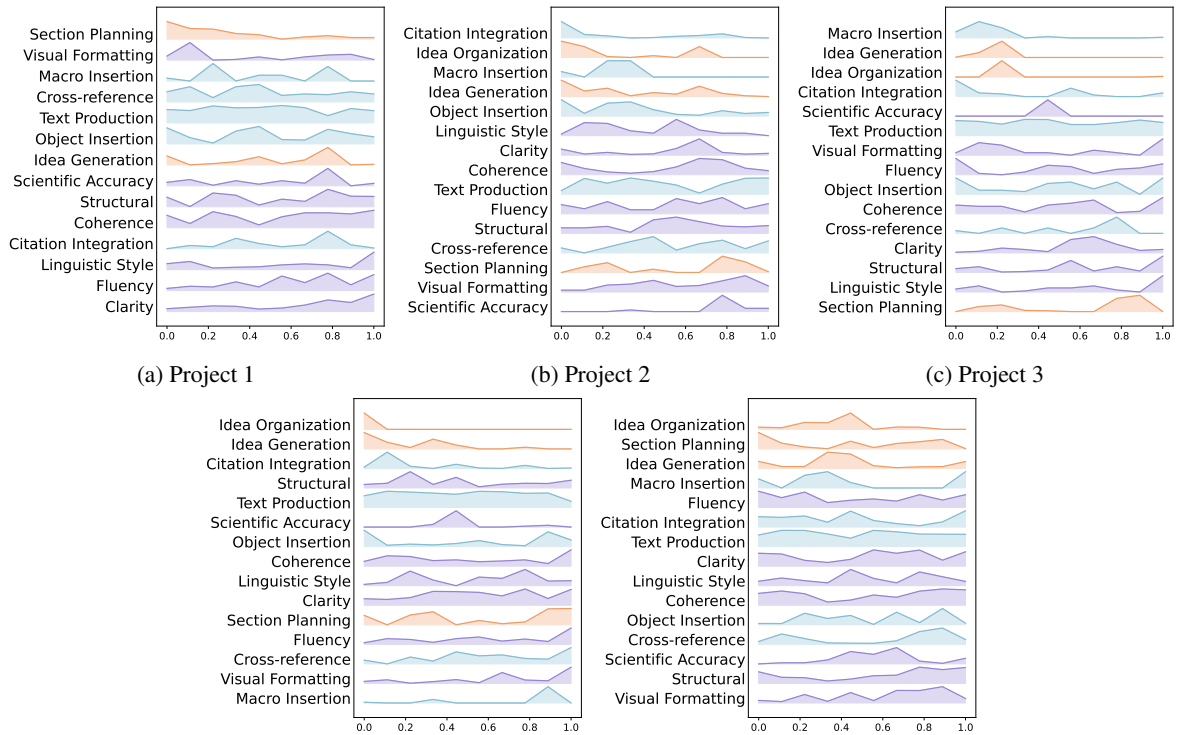
Table 12 shows top 6-gram writing intention sequences by session coverage (%)



(a) Project 1 (b) Project 2 (c) Project 3

(d) Project 4 (e) Project 5

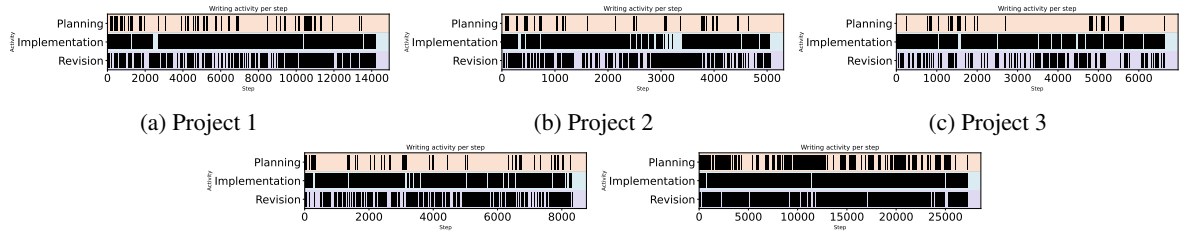
Figure 8: Wasserstein distance to uniform distribution for each distribution of writing intentions. Orange, Blue, and Purple represent Planning, Implementation, and Revision writing actions, respectively.



(a) Project 1 (b) Project 2 (c) Project 3

(d) Project 4 (e) Project 5

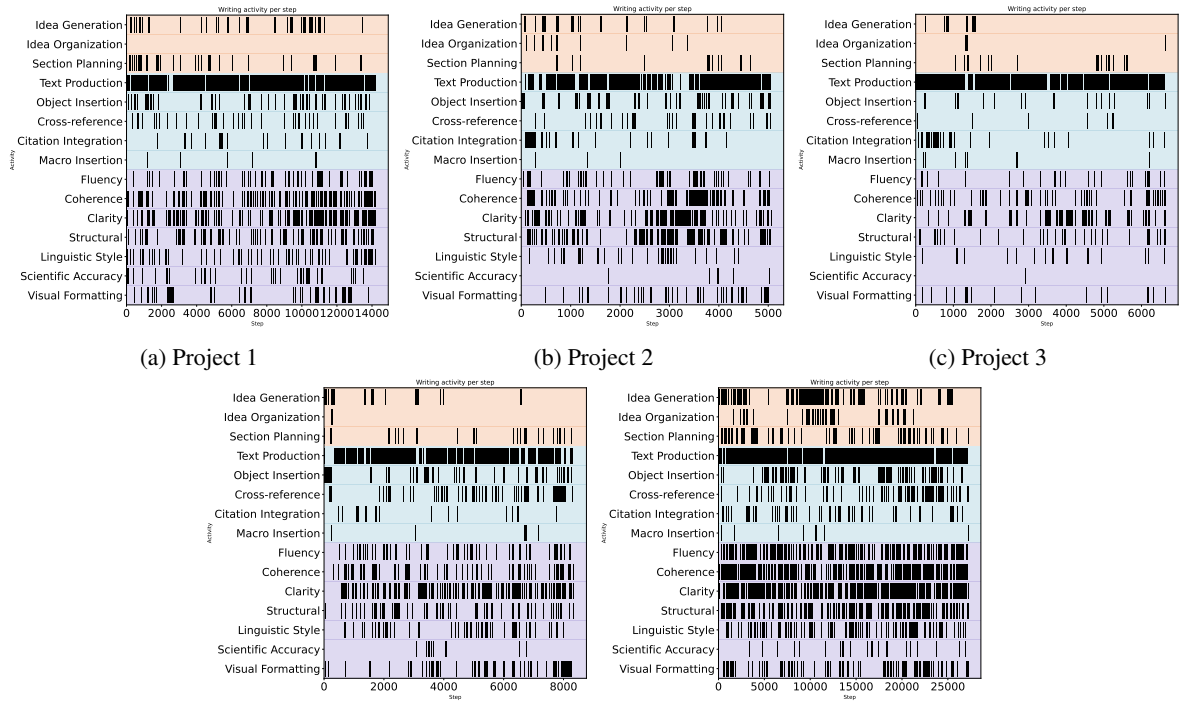
Figure 9: Distribution of labels over time across projects. Orange, Blue and Purple represent Planning, Implementation, and Revision writing actions respectively. The writing actions are sorted in ascending order, top to bottom, according to their distribution mean.



(a) Project 1 (b) Project 2 (c) Project 3

(d) Project 4 (e) Project 5

Figure 10: Distribution of high-level intention activities over time. Orange, Blue and Purple represent Planning, Implementation, and Revision writing actions respectively.



(a) Project 1 (b) Project 2 (c) Project 3

(d) Project 4 (e) Project 5

Figure 11: Distribution of Per-intention writing activities over time. Orange, Blue and Purple represent Planning, Implementation, and Revision writing actions respectively.

1st	Intention	Definition	An example action	Prop.
PLANNING	Idea Generation	Formulate and record initial thoughts and concepts.	writing down keywords of a paragraph beforehand (e.g., "*..%[Comment out] main point: artifacts lack in human subjectivity..*")	7.0%
	Idea Organization	Select the most useful materials and demarcate those generated ideas in a visually formatted way.	Linking the generated ideas into a logical sequence and spacing out between ideas (e.g., "..% (1) need diff. stress testing...%%[Spacing] (2) exp. setup? ")	0.5%
	Section Planning	Initially create sections and sub-level structures.	Putting section-related LaTeX commands (e.g., `\section`, `\paragraph`)	2.2%
IMPLEMENTATION	Text Production	Translate their ideas into full languages, either from the writers' language or borrowed sentences from an external source.	Generating subsequent sentences with the author's own idea (e.g., "... GPT-4 (OpenAI, 2023) explains the data ... Our approach is built on top of GPT-4...")	57.4%
	Object Insertion	Insert visual claims of their arguments (e.g., figures, tables, equations, footnotes, lists)	e.g., `\begin{figure}[h] \centering \includegraphics{figure_A.pdf} \end{figure}`	4.6%
	Citation Integration	Incorporate bibliographic references into a document and systematically link these references using citation commands.	Inserting a new BibTeX object in the bibliography file and adding the object name to an existing `\cite{}` on the Related Work section	1.7%
	Cross-reference	Link sections, figures, tables, or other elements within a document via referencing commands.	Putting a command `\label{figure-1}` to a figure and referencing it by calling `\ref{figure-1}`	1.1%
	Macro Insertion	Incorporate predefined commands or packages into a LaTeX document to alter its formatting.	Putting a `\usepackage{minted}` for formatting a LLM prompt	0.2%
REVISION	Fluency	Fix grammatical or syntactic errors in the text or LaTeX commands.	"We desig~~ining~~ned several experiment setups for the LLM evaluations as described.."	1.4%
	Coherence	Logically link (1) multiple sentences within the same paragraph; (2) any two subsequent paragraphs; or (3) objects to be consistent.	"Each comment was annotated by three different annotators ~~, which~~ and we achieved high inter-annotator agreement."	3.3%
	Clarity	Improve the semantic relationships between texts to be more straightforward and concise.	"..relevant studies have examined ~~one of the several textual styles~~ one aspect of texts, the formality, ..."	11.5%
	Structural	Improve the flow of information by modifying the location of texts and objects.	"..human alignment to compare the alignment between lexicon-based preferences and humans' original preferences. First, we calculated the score from each human participant ~~to compare the alignment between..~~"	3.7%
	Linguistic Style	Modify texts with the writer's writing preferences regarding styles and word choices, etc.	"We ~~believe~~ posit that ..."	1.6%
	Scientific Accuracy	Update or correct scientific evidence (e.g., numbers, equations) for more accurate claims.	"..Pearson's r correlation (~~0.78~~0.68; $p < 0.01$)"	0.7%
	Visual Formatting	Modify the stylistic formatting of texts, objects, and citations	`\cite` → `\citet`, `\textbf` → `\textsc`, etc.	3.2%

Table 7: The developed taxonomy of scholarly writing process in SCHOLAWRITE.

F More about Experiments

F.1 Predicting Next Writing Intention

F.1.1 Training environments

BERT & RoBERTa We fine-tuned BERT and RoBERTa with the following hyperparameter setups: (1) a learning rate of $2e^{-5}$; (2) training batch size per device of 8; (3) evaluation batch size per device of 8; (4) the number of training epochs of 10; and (5) a weight decay of 0.01. For each model, it took approximately 3.5 hours on one NVIDIA RTX A6000.

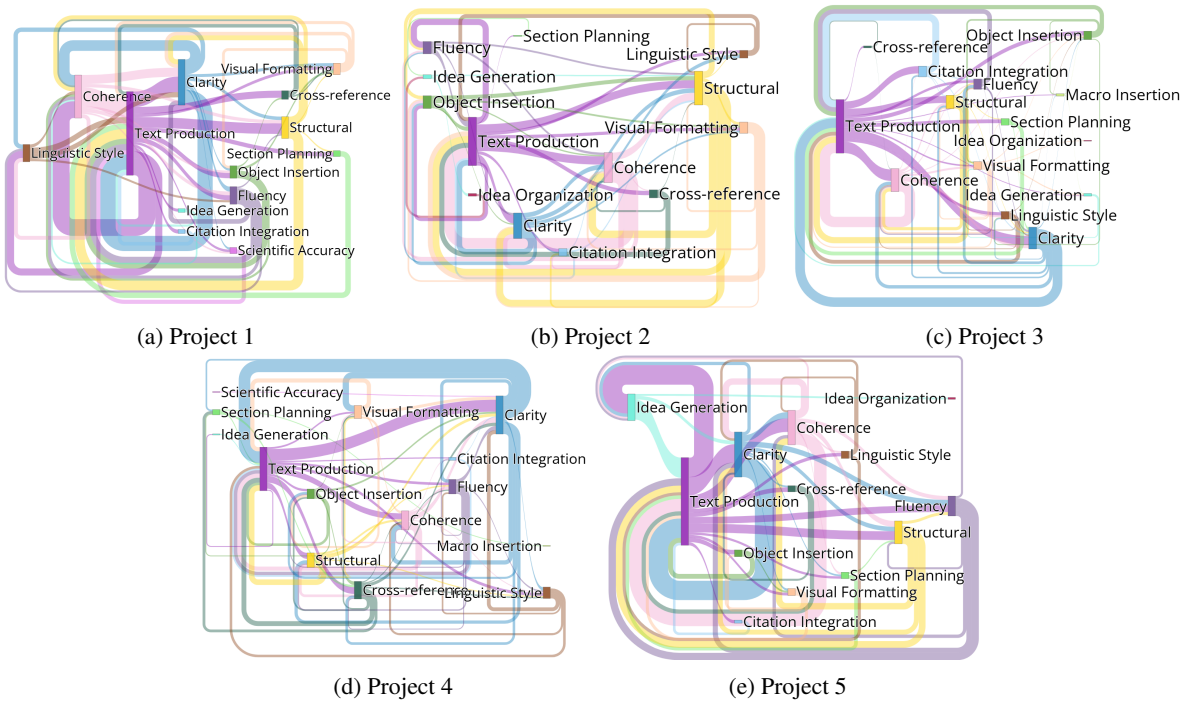
Llama For all experiments, we used baseline models of 4-bit quantized Llama-8B-Instruct[8], using unsloth library[9].

For **the intention prediction task** (Sec. 5.1), here are the hyperparameter setups for the Llama models: (1) only one epoch of training; (2) a weight decay of 0.01; (3) warm-up steps of 5; (4) learning rate of $2e^{-4}$; and (5) AdamW 8-bit optimizer. Due to computational constraints, we were able to run only one epoch for fine-tuning Llama models on our SCHOLAWRITE dataset. For Llama-8B, it took approximately 8 hours on one RTX A5000.

GPT-4o We used the `GPT-4o-2024-08-06` version.

[8] https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit

[9] https://github.com/unslothai/unsloth



(a) Project 1 (b) Project 2 (c) Project 3

(d) Project 4 (e) Project 5

Figure 12: Sankey diagrams representing the intention flow of each project. Figure (a) to (d) generated from all intentions. Figure (e) generated from first 10K intentions due to computational constraint.

Label		Subsequent label	Probability
Idea Generation	→	Text Production	0.52
Idea Organization	→	Idea Generation	0.34
Section Planning	→	Text Production	0.33
Text Production	→	Clarity	0.20
Object Insertion	→	Text Production	0.32
Citation Integration	→	Text Production	0.37
Cross-reference	→	Text Production	0.36
Macro Insertion	→	Idea Generation	0.29
Fluency	→	Text Production	0.30
Coherence	→	Text Production	0.34
Clarity	→	Text Production	0.35
Structural	→	Text Production	0.27
Linguistic Style	→	Text Production	0.29
Scientific Accuracy	→	Text Production	0.34
Visual Formatting	→	Text Production	0.25

Table 8: Probability of inter-connections between writing intentions in SCHOLAWRITE. For example, in 34% of instances where an author engaged in "Idea Organization," the subsequent intention was "Idea Generation."

	1	2	3	4	5
Idea Generation	515	130	116	309	3255
Idea Organization	0	45	25	9	231
Section Planning	182	57	111	201	773
Text Production	9267	2438	5109	4478	14031
Object Insertion	583	383	62	486	1300
Cross-reference	141	112	13	292	458
Citation Integration	75	151	69	127	245
Macro Insertion	16	7	51	29	33
Linguistic Style	233	75	42	201	411
Coherence	422	242	126	193	1021
Clarity	1249	645	721	1180	3301
Scientific Accuracy	307	15	2	24	95
Structural	359	506	105	257	1042
Fluency	116	90	46	135	476
Visual Formatting	752	163	43	427	567

Table 9: Distribution of intention labels annotated across all five Overleaf projects.

F.1.2 Details About Finetuning Process

The fine-tuning prompt included all possible labels with definitions, task instructions, the "before-text" chunk, and the corresponding human-annotated intention label, asking the model to predict the intention label based on the "before-text". Differences in prompts were limited to only task instructions (see §G.3 for prompt details).

To achieve optimal performance while minimizing memory usage, we employed QLoRA (Dettmers et al., 2024) to fine-tune all linear modules of a 4-bit quantized Llama. During fine-tuning, we utilized the `'train_on_response_only'` function provided by the `unsloth` library, which masks the task instructions, intention label definitions, and "before" text with -100s. This ensures the model is

Intention	% Time	Intention	% Time
Idea Generation	4.8	Fluency	3.3
Idea Organization	0.6	Coherence	**6.9**
Section Planning	2.6	Clarity	**13.4**
Text Production	**40.3**	Structural	4.0
Object Insertion	**8.4**	Linguistic Style	2.6
Citation Integration	2.4	Scientific Accuracy	1.2
Cross-reference	2.4	Visual Formatting	**6.1**
Macro Insertion	0.8		

Table 10: Average percentage of total writing time spent. Implementation intentions dominate overall effort, while Revision tasks show sustained cognitive load.

Intention	Mean	SD	Max	Count
Idea Generation	1.87	1.94	11.40	76
Idea Organization	**3.71**	4.58	13.79	7
Section Planning	1.37	1.57	9.36	38
Text Production	2.20	2.53	**28.14**	**724**
Object Insertion	**2.88**	3.05	21.61	108
Citation Integration	1.70	1.53	7.53	34
Cross-reference	1.08	0.77	3.98	38
Macro Insertion	1.43	1.05	3.71	7
Fluency	0.66	0.18	1.06	10
Coherence	1.46	1.71	9.62	55
Clarity	1.64	1.61	9.53	243
Structural	1.50	1.43	8.42	53
Linguistic Style	2.05	2.55	10.81	24
Scientific Accuracy	1.97	1.60	6.44	20
Visual Formatting	**2.22**	2.66	14.25	87

Table 11: Session-level time statistics in minutes. Longer sessions for Idea Organization, Object Insertion, and Visual Formatting reflect deeper cognitive engagement.

trained exclusively on the response portion of the fine-tuning prompt (i.e., the predicted intention label), without being influenced by the instructional components of the input. The model was fine-tuned for one epoch with a batch size of 2, 4 gradient accumulation steps, and the AdamW 8-bit optimizer.

F.2 Results

Table 13 reports the per-class F1 scores for the fine-tuned BERT model.

F.2.1 Prompt Templates

Prediction Prompt for Llama-8B-Zero models

"Here are all the possible writing intention labels:

- Idea Generation: Formulate and record initial thoughts and concepts.
- Idea Organization: Select the most useful materials and demarcate those generated ideas in a visually formatted way.

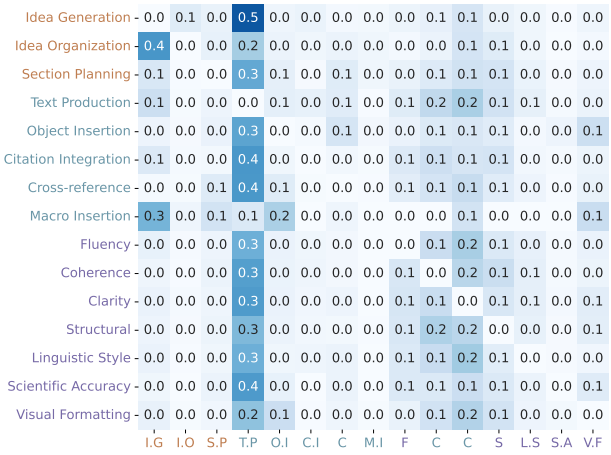
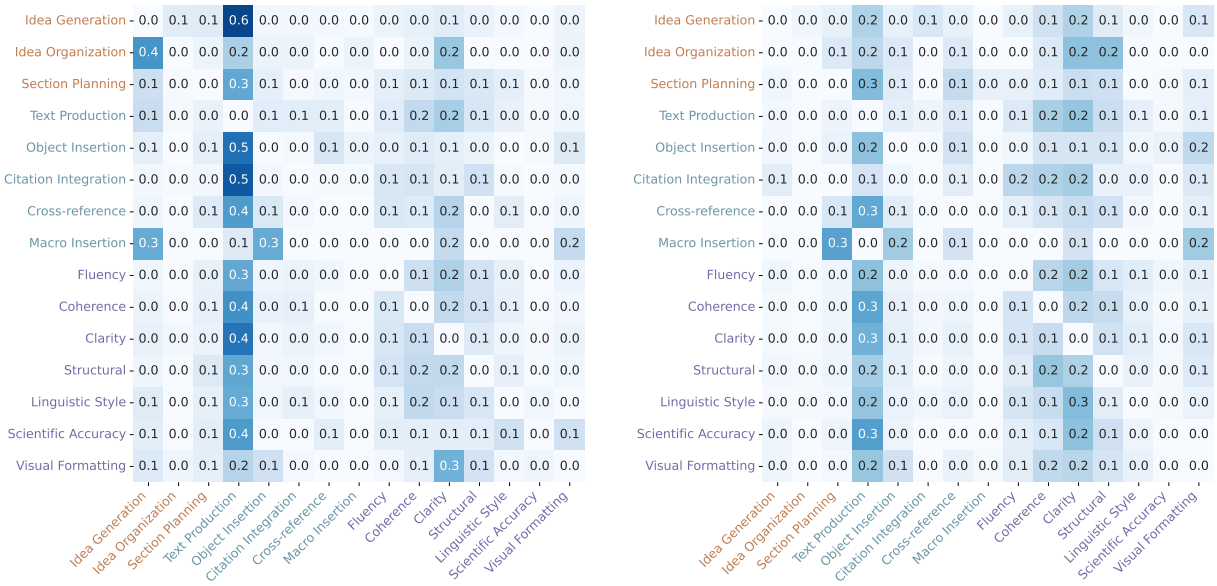


Figure 13: Transition probability matrix between writing intentions. Each cell shows the likelihood that a session with the current intention (y-axis) is followed by a session with the next intention (x-axis).

- Section Planning: Initially create sections and sub-level structures.
- Text Production: Translate their ideas into full languages, either from the writers' language or borrowed sentences from an external source.
- Object Insertion: Insert visual claims of their arguments (e.g., figures, tables, equations, footnotes, itemized lists, etc.).
- Cross-reference: Link different sections, figures, tables, or other elements within a document through referencing commands.
- Citation Integration: Incorporate bibliographic references into a document and systematically link these references using citation commands.
- Macro Insertion: Incorporate predefined commands or packages into a LaTeX document to alter its formatting.
- Fluency: Fix grammatical or syntactic errors in the text or LaTeX commands.
- Coherence: Logically link (1) any of the two or multiple sentences within the same paragraph; (2) any two subsequent paragraphs; or (3) objects to be consistent as a whole.
- Structural: Improve the flow of information by modifying the location of texts and objects.



(a) Early stage

(b) Late stage

Figure 14: Transition probabilities across time. (a) Early sessions vs. (b) Late sessions.

6-gram Sequence	**Percentage**
Text Production → Clarity → Text Production → Clarity → Text Production → Clarity	3.22
Clarity → Text Production → Clarity → Text Production → Clarity → Text Production	3.22
Text Production → Object Insertion → Text Production → Object Insertion → Text Production → Object Insertion	2.34
Object Insertion → Text Production → Object Insertion → Text Production → Object Insertion → Text Production	2.34
Idea Generation → Idea Organization → Idea Generation → Idea Organization → Idea Generation → Idea Organization	2.05

Table 12: Top 6-gram writing intention sequences by session coverage (%)

Writing Intention	**F1-score**
Idea Generation	0.72
Idea Organization	0.47
Section Planning	0.83
Text Production	0.63
Object Insertion	0.79
Citation Integration	0.52
Cross-reference	0.10
Macro Insertion	0.67
Fluency	0.13
Coherence	0.32
Clarity	0.59
Structural	0.39
Linguistic Style	0.27
Visual Formatting	0.44
Accuracy	0.56
Macro Avg	0.49

Table 13: Per-class F1 scores for BERT-finetuned.

- Clarity: Improve the semantic relationships between texts to be more straightforward and concise.

- Linguistic Style: Modify texts with the writer's writing preferences regarding styles and word choices, etc.

- Scientific Accuracy: Update or correct scientific evidence (e.g., numbers, equations) for more accurate claims.

- Visual Formatting: Modify the stylistic formatting of texts, objects, and citations.

Identify the most likely next writing intention of a graduate researcher when editing the following LaTex paper draft. Your output should only be a label from the list above.

Input: `before_text`

Output: "

Prediction Prompt for Llama-8B-SW models

"Here are all the possible writing intention labels:

- Idea Generation: Formulate and record initial thoughts and concepts.

- Idea Organization: Select the most useful materials and demarcate those generated ideas in a visually formatted way.

- Section Planning: Initially create sections and sub-level structures.

- Text Production: Translate their ideas into full languages, either from the writers' language or borrowed sentences from an external source.
- Object Insertion: Insert visual claims of their arguments (e.g., figures, tables, equations, footnotes, itemized lists, etc.).
- Cross-reference: Link different sections, figures, tables, or other elements within a document through referencing commands.
- Citation Integration: Incorporate bibliographic references into a document and systematically link these references using citation commands.
- Macro Insertion: Incorporate predefined commands or packages into a LaTeX document to alter its formatting.
- Fluency: Fix grammatical or syntactic errors in the text or LaTeX commands.
- Coherence: Logically link (1) any of the two or multiple sentences within the same paragraph; (2) any two subsequent paragraphs; or (3) objects to be consistent as a whole.
- Structural: Improve the flow of information by modifying the location of texts and objects.
- Clarity: Improve the semantic relationships between texts to be more straightforward and concise.
- Linguistic Style: Modify texts with the writer's writing preferences regarding styles and word choices, etc.
- Scientific Accuracy: Update or correct scientific evidence (e.g., numbers, equations) for more accurate claims.
- Visual Formatting: Modify the stylistic formatting of texts, objects, and citations.

Identify the most likely next writing intention of a graduate researcher when writing the following LaTex paper draft. Your output should only be a label from the list above.

Input: `before_text`
Output: "

Classification Prompt for GPT-4o "You are a classifier that identify the most likely next writing intention. You will be given a list of all possible writing intention labels with definitions, and an in-progress LaTex paper draft written by a graduate student. Please strictly follow user's instruction to identify the most likely next writing intention.

Here are the verbalizers of all the possible writing intention labels:

- Idea Generation: Formulate and record initial thoughts and concepts.
- Idea Organization: Select the most useful materials and demarcate those generated ideas in a visually formatted way.
- Section Planning: Initially create sections and sub-level structures.
- Text Production: Translate their ideas into full languages, either from the writers' language or borrowed sentences from an external source.
- Object Insertion: Insert visual claims of their arguments (e.g., figures, tables, equations, footnotes, itemized lists, etc.).
- Cross-reference: Link different sections, figures, tables, or other elements within a document through referencing commands.
- Citation Integration: Incorporate bibliographic references into a document and systematically link these references using citation commands.
- Macro Insertion: Incorporate predefined commands or packages into a LaTeX document to alter its formatting.
- Fluency: Fix grammatical or syntactic errors in the text or LaTeX commands.
- Coherence: Logically link (1) any of the two or multiple sentences within the same paragraph; (2) any two subsequent paragraphs; or (3) objects to be consistent as a whole.
- Structural: Improve the flow of information by modifying the location of texts and objects.
- Clarity: Improve the semantic relationships between texts to be more straightforward and concise.

- Linguistic Style: Modify texts with the writer's writing preferences regarding styles and word choices, etc.

- Scientific Accuracy: Update or correct scientific evidence (e.g., numbers, equations) for more accurate claims.

- Visual Formatting: Modify the stylistic formatting of texts, objects, and citations.

Identify the most likely next writing intention of a graduate researcher when editing the following LaTex paper draft. Your output should only be a label from the list above.

Input: `before_text`
Output: "

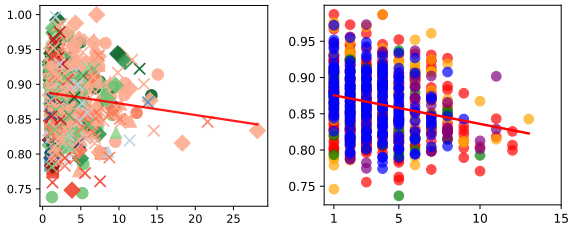
F.2.2 Discussion about Next Intention Prediction Results

According to Table 4, none of them is reaching 0.7 in F1. This is likely due to the intricate nature of the task. The model is asked to predict the next intention by only giving the before text. In the annotation task, the annotator labels the data by looking through multiple consecutive before and after text pairs to determine the current intention rather than looking at the current text to predict the next intention. Another reason is that the next intention chosen by the author does not necessarily mean that it is the only correct intention. We also noticed that BERT and RoBERTa perform much better than the vanilla Llama-8b-Instruct. This is likely due to Llama 8b being under-trained as they have a larger size of parameters, while we fine-tuned it on our data for only one epoch.

F.3 Output Alignment across Writing Sessions

F.3.1 Prompt Templates

Single-intention Session Output Generation
You are an expert writing assistant.
Your role is to improve user-provided text according to specific criteria.
The input will be LaTeX text.
Always output only the improved LaTeX text, with no explanations or notes.
Improve the following LaTeX text according to this criteria:
- {intention}: {intention definition}
Text to improve:
{before_text}



(a) Duration vs. alignment (b) Intents vs. alignment

Figure 15: **Model alignment patterns.** (a; minutes-vs-alignment) Longer writing sessions show lower alignment, indicating higher cognitive complexity. (b; #-intents-vs-alignment) Alignment decreases as more intentions intertwine within a session.

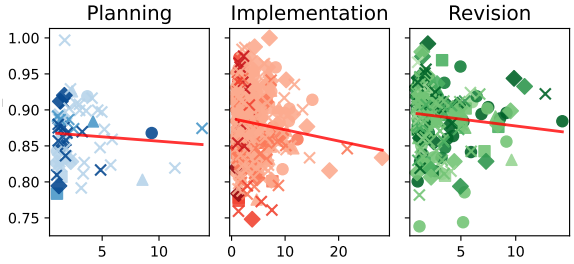


Figure 16: Writing duration (minutes) vs. alignment across different intention categories

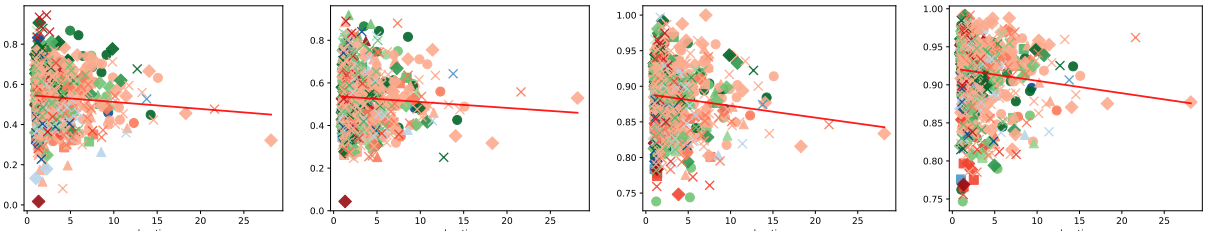
Multi-intention Session Output Generation
You are an expert writing assistant.
Your role is to improve user-provided text according to specific criteria.
The input will be LaTeX text.
Always output only the improved LaTeX text, with no explanations or notes.
Improve the following LaTeX text according to these criteria:
{list of intentions and their definitions} Text to improve:
{before_text}

F.3.2 Results

Figure 17 shows the relationship between a single-intention session duration and alignment score (either Levenshtein ratio or BERTScore-F1) for two models, GPT-5 and Qwen-2.5-14B.

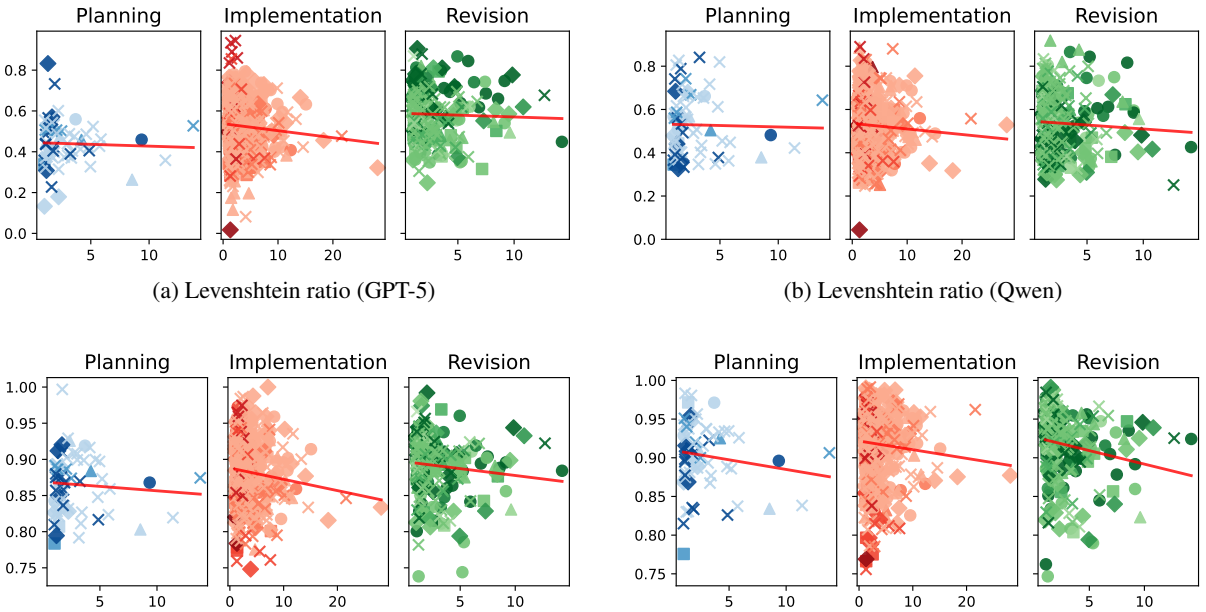
Figure 18 shows the relationship between a single-intention session duration and alignment score (either Levenshtein ratio or BERTScore-F1) for two models, GPT-5 and Qwen-2.5-14B across different intention categories.

Figure 19 shows the relationship between number of writing intentions in a multi-intention session and alignment score (either Levenshtein ratio or BERTScore-F1) for two models, GPT-5 and Qwen-2.5-14B.



(a) Levenshtein ratio (GPT-5) (b) Levenshtein ratio (Qwen) (c) BERTScore-F1 (GPT-5) (d) BERTScore-F1 (Qwen)

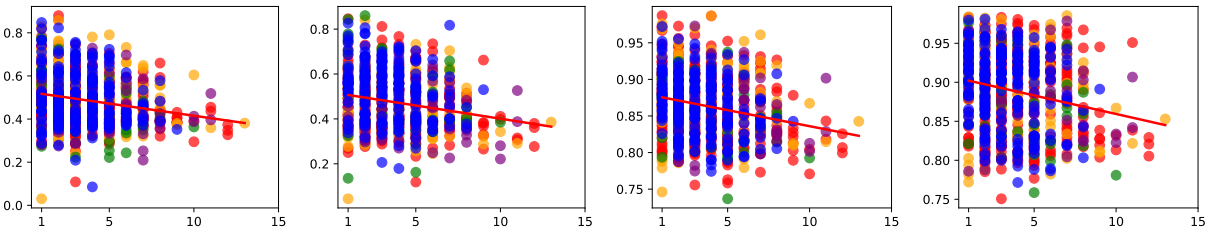
Figure 17: Writing duration (minutes) vs. model alignment scores (Levenshtein ratio and BERTScore-F1)



(a) Levenshtein ratio (GPT-5) (b) Levenshtein ratio (Qwen)

(c) BERTScore-F1 (GPT-5) (d) BERTScore-F1 (Qwen)

Figure 18: Writing duration (minutes) vs. model alignment scores (Levenshtein ratio and BERTScore-F1) across different intention categories



(a) Levenshtein ratio (GPT-5) (b) Levenshtein ratio (Qwen) (c) BERTScore-F1 (GPT-5) (d) BERTScore-F1 (Qwen)

Figure 19: Number of writing intentions vs. alignment scores (Levenshtein ratio and BERTScore-F1) for GPT-5 and Qwen-2.5-14B

G More about SCHOLAWRITE Evaluation

We envision SCHOLAWRITE as a valuable resource for training language models and improving future writing assistants for scholarly writing. To evaluate its usability, we conducted experiments training LLMs to mimic the complex, non-linear writing processes of human scholars.

Iteratively generating scholarly writing actions from scratch (called Iterative Self-Writing), mirroring the human writing process: This task focuses on how well the model trained on our dataset can replicate the actual iterative writing and thinking process of scholars, and whether the generated text achieves higher quality compared to LLM-prompted writing.

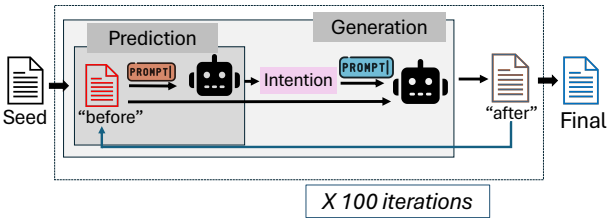


Figure 20: The overview of next writing intention prediction task (Prediction box) and iterative self-writing task setup (the whole pipeline).

Setups During iterative self-writing (Figure 20), a fine-tuned model processes LaTeX-formatted seed document (as "before-text") with a context prompt to predict the next intention, then revises the text ("after-text") accordingly given prompt. The revised document then serves as the new seed for the next iteration. This process repeats until a set iteration limit (e.g., 100) is reached. All models use the same train (80%)-test (20%) split across experiments. Figure 24 for training details.

We fine-tune Llama3.1-8B-Instruct (Llama-8B-SW) and compare it to vanilla Llama-8B-Instruct (Llama-8B-Zero) and GPT-4o. Also, seed documents were derived from LaTeX-formatted abstracts of four award-winning NLP papers on diverse topics (Zeng et al., 2024; Lu et al., 2024; Du et al., 2022a; Etxaniz et al., 2024), as shown in Appendix Listings 1-4.

Metrics We evaluated *lexical diversity* (unique-to-total token ratio), *topic consistency* (cosine similarity between seed and final output), and *intention coverage* (unique writing intentions used over 100 iterations). For **human evaluation**, three native English speakers with LaTeX expertise assessed outputs from Llama-8B-Zero and Llama-8B-SW on *accuracy* (alignment with predicted intention), *alignment* (similarity to human writing), *fluency*(grammatical correctness), *coherence*(logical structure), and *relevance* (connection to the seed document) - Refer to Appendix G.7. Accuracy was judged per iteration, while other metrics used pairwise comparisons. Inter-annotator agreement (IAA)[10] was measured using Krippendorff's alpha for accuracy and percentage agreement for others.

Results Figure 21 shows that Llama-8B-SW consistently produced the most lexically diverse words, generated the most semantically aligned topics (Seeds 1 & 2), and covered the most writing intentions (except Seed 3). These results underscore the value of SCHOLAWRITE in improving scholarly writing quality generated by language models.

However, our human evaluation (Figure 22) revealed that Llama-8B-SW generated less human-like writing, in terms of fluency and logical claims. It also struggled with generating texts aligned with the predicted intentions. See Appendix Tables 15 to 18 for more details. Despite the weaknesses, Llama-8B-SW still produced more relevant content (Seed 2), which aligns with topic consistency trends in Figure 21, highlighting the usefulness of SCHOLAWRITE dataset in certain contexts.

Moreover, Llama-8B-SW exhibited the most human-like writing activity patterns over time (Figure 23), which frequently switches between implementation and revision and covers all three high-level processes. Llama-8B-Zero and GPT-4o tend to remain in a single high-level stage throughout all 100 iterations of self-writing (see Appendix 26 and 27 for details). Compared to Appendix Figure 11, which depicts frequent transitions across all three stages in an early draft (e.g., the first 100 steps), Llama-8B-SW most closely replicates human writing behaviors in iterative writing tasks. These findings reinforce the potential of SCHOLAWRITE in helping LLMs emulate human scholarly writing processes.

G.1 Training environments

BERT & RoBERTa We fine-tuned BERT and RoBERTa with the following hyperparameter setups: (1) a learning rate of $2e^{-5}$; (2) training batch size per device of 8; (3) evaluation batch size per device of 8; (4) the number of training epochs of 10; and (5) a weight decay of 0.01. For each model, it took approximately 3.5 hours on one NVIDIA RTX A6000.

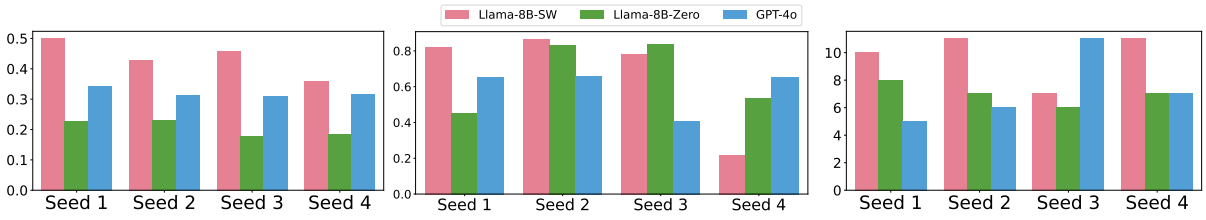
Llama For all experiments, we used baseline models of 4-bit quantized Llama-8B-Instruct[11], using unsloth library[12].

For **the 'after' text generation subtask** from the iterative self-writing experiment, we used the following hyperparameter setups for the fine-tuned Llama-8B-Instruct: (1) only one epoch of training; (2) a weight decay of 0.01; (3) warm-up steps of 10; (4) learning rate of $3e^{-4}$; and (5) AdamW 8-bit optimizer. Due to computational constraints, we were

[10]The IAA scores are 0.84 (SW) and 0.76 (Zero) for accuracy, all 100% for alignment, fluency, coherence, and 49.8% (SW) and 100% (Zero) for relevance.

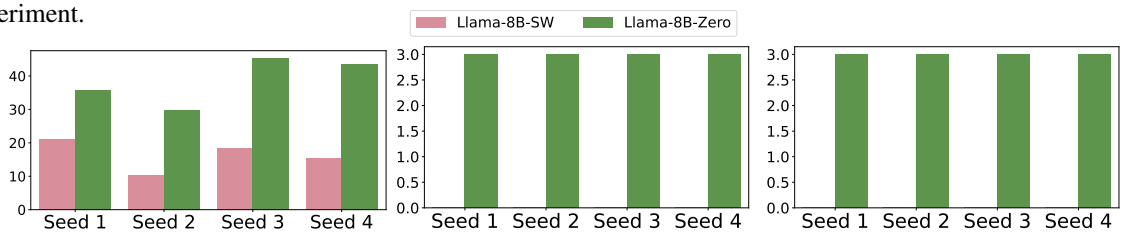
[11]https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit

[12]https://github.com/unslothai/unsloth

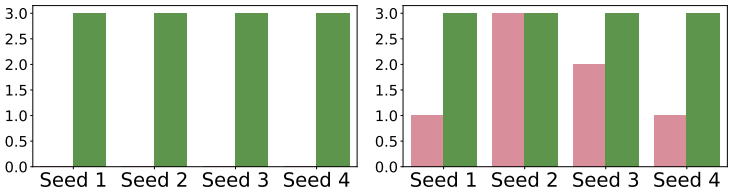


(a) Lexical Diversity (b) Topic Consistency (c) Intention Coverage

Figure 21: Metric scores of the final writing output of models after 100 iterations of the iterative self-writing experiment.



(a) Accuracy (b) Coherence (c) Fluency



(d) Alignment (e) Relevance

Figure 22: Human evaluation results of iterative writing outputs of models

Iter.	Llama-8B-SW	Llama-8B-Zero	GPT-4o
25	[..*Editing Abstract*..] but rather should be used to improve the flow of information to avoid information overload~~,~~. (**Text Production**)	[..*Editing Table*] Acceptance rate (%) & 75 & 8~~7.5~~8.2 // Revision time (minutes) & 45 & 2~~9~~8.5 Human effort (minutes) & 60 & 4~~3~~2 ... (**Scientific Accuracy**)	[..*Same as the 24th iteration*] The ~~efficiency~~ ...consider the structural flowchart in Figure \ref{fig:system-architecture}, which outlines... [*Inserting the figure*] \label{fig:system-architecture} (**Object Insertion**)
51	[..*Editing Abstract*..] but rather should be used to improve the flow of information, offering ~~teach~~previously trained to a load more related information over the-load. (**Clarity**)	\section{Impact of the Proposed System} The proposed system, $\mathcal{R}3$, has the potential to impact the writing process in several ways.... \section{Future Research Direction}... (**Structural**)	\bibitem{jones2020one_shot} Jones, L., \& Green, D. (2020). \bibitem{brown2021collaboration} Brown, E., \& Davis, M. (2021). \bibitem{garcia2021revision_metrics} Garcia, I., \& Lopez, R. (2021). (**Object Insertion**)
100	[..*Same as the 99th iteration*] \end{document} (**Macro Insertion**)	[..*Same as the 99th*] \usepackage[margin=1in]{geometry} % Customizes page margins \usepackage{hyperref} % Enables hyperlinks (**Fluency**)	[..*Same as the 93th iteration*] \bibliography{references} \bibliographystyle{plain} ~~% References~~\begin{thebibliography}{} (**Cross-reference**)

Table 14: Example model outputs at different iterations from the seed 2

able to run only one epoch for fine-tuning Llama models on our SCHOLAWRITE dataset. Also, it took approximately 12 hours on one NVIDIA L40s.

GPT-4o We used the `GPT-4o-2024-08-06` version. The iterative writing with 100 iterations took approximately 1 hour on each seed.

G.2 Details About Finetuning Process

The structure of the "after" text differs slightly to help the model learn where and what edits to make. We used the `diff_match_patch` library to generate a word-level difference array between the "before" and "after" texts. Special tokens (`<same>`, `</same>`, `<del>`, `</del>`, `<add>`,

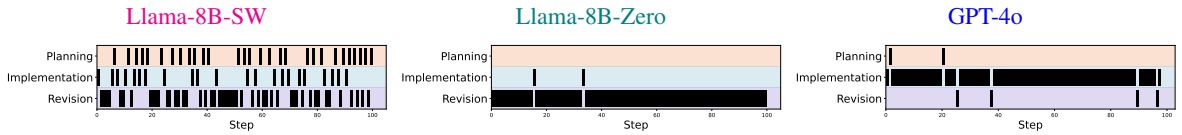
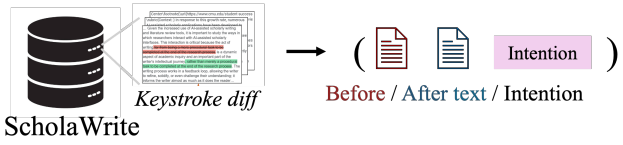
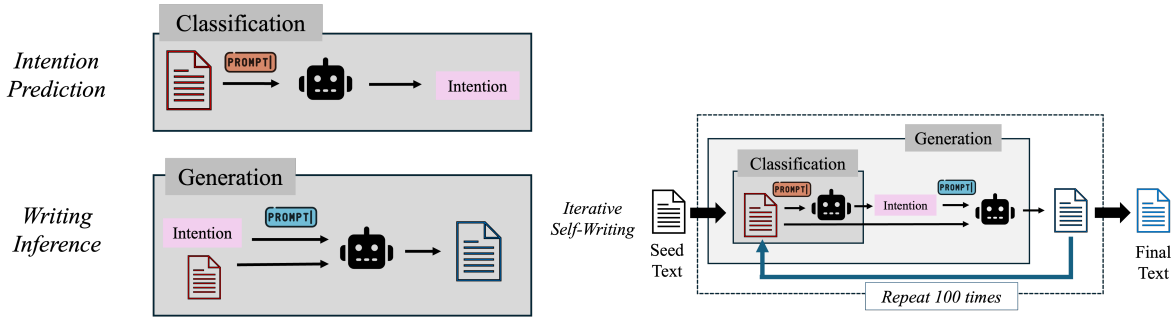


Figure 23: Distribution of high-level writing activities on seed 3 over time by models.



(a) SCHOLAWRITE Dataset Preparation



(b) Finetuning Setup

(c) Iterative Self-writing

Figure 24: Experiment Setup for the iterative self-writing

</add>) were added to the tokenizer, and the difference array was converted into text wrapped with these tokens. For example, given a "before" text of "Bad dog" and an "after" text of "Good dog", the difference array would be `[(-1, 'Bad'), (1, 'Good '), (0, 'dog')]`. This is converted into: `<del>Bad </del><add>Good </add><same>dog</same>`. This transformation was applied only to the "after" text, while the "before" text remained as plain LaTeX text.

For finetuning, we randomly split the SCHOLAWRITE dataset into training (80%) and testing (20%) sets. From each intention label in the test set, we sample up to 300 keystroke entries due to budget constraints.

For intention prediction, we fine-tune Llama3.1-8B-Instruct on SCHOLAWRITE training set (LLAMA-8B-SW-PRED") and compare it to baseline models (Llama-8B-Instruct and GPT-4o) from Section 5.1. For "after-text" generation, we fine-tune another Llama3.1-8B-Instruct model ("LLAMA-8B-SW-GEN") using the same dataset, with Llama3.1-8B-Instruct and GPT-4o as baselines. The fine-tuning prompt includes task instructions, a verbalizer from human-annotated labels, and "before-text." While prompts were standardized, task instructions varied by model.

The fine-tuning prompt included task instructions, a verbalizer derived from human-annotated labels, and the "before" text. For fine-tuning, we used QLoRA (Dettmers et al., 2024) to optimize all linear modules of a 4-bit quantized model while maintaining a small memory footprint. Additionally, the `embed_tokens` and `lm_head` modules were set as trainable and saved in the final checkpoint. To focus training on the response portion (the "after" text), we used the `train_on_response_only` function, masking the task instructions, verbalizer, and "before" text with `-100s`. This ensures the model learns to generate the "after" text without being influenced by instructional input. The model was trained for one epoch with a batch size of 1, 4 gradient accumulation steps, and the AdamW 8-bit optimizer.

During iterative writing, we performed 100 iterations, treating the model's output under one intention as a single iteration. If the intention predicted by the classification model (fine-tuned Llama3.1-8B-Instruct, as described in Sec 5.1) matched the current predicted intention, the model was prompted to edit the text again. In this case, the newly generated output was not treated as final output in the iteration, and the iteration did not proceed. We moved to the next iteration only when the intention prediction model generated a different intention label than the previous one.

For model setups, we created three pairs of models for each prediction and generation subtasks as

follows:

- LLama-8B-SW: LLama-8B-Instruct fine-tuned on SCHOLAWRITE dataset (Prediction) & LLama-8B-Instruct fine-tuned on SCHOLAWRITE dataset (Generation), independently
- LLama-8B-Instruct: Vanilla LLama-8B-Instruct (Prediction) & Vanilla LLama-8B-Instruct (Generation), independently
- GPT-4o: GPT-4o inference (Prediction) & GPT-4o inference (Generation)

Also, due to budget constraints, models had different revision strategies. LLAMA-8B-SW-* and LLAMA-8B-INSTRUCT continued revision until the next predicted intention changed. GPT-4o, however, moved to the next iteration regardless. We refer to the fine-tuned Llama-8B model as Llama-8B-ScholaWrite (or Llama-8B-SW) and the vanilla model as Llama-8B-Zero.

G.3 Prompt Templates

G.3.1 Generation Prompt for Llama-8B-Zero models

"You are a computer science researcher with extensive experience of scholarly writing. Here, you are writing a research paper in natural language processing using LaTeX languages.

You currently want to "`put the verbalizer of the predicted intention label`" (`e.g., "initially create sections and sub-level structures" if the predicted label was section planning`).

Below is the paper you have written so far. Please strictly follow the writing intention given above and insert, delete, or revise at appropriate places in the paper given below.

Your writing should relate to the paper given below. Do not generate text other than paper content. Do not describe the changes you are making or your reasoning.

Input: `before_text` "

G.3.2 Generation Prompt for Llama-8B-Zero models

"You are a computer science researcher with extensive experience in scholarly writing. Here, you are writing a research paper in natural language processing using LaTeX.

You currently want to "`put the verbalizer of the predicted intention label`" (`e.g., "initially create sections and sub-level structures" if the predicted label was section planning`).

Below is the paper you have written so far. Given the paper information below and the corresponding scholarly writing intention, please revise or add to the text to fulfill this writing intention.

You may insert, delete, or revise text at appropriate places in the given paper.

Please provide a complete output. Do not generate text that is nonsensical or unrelated to the given paper information.

Input: `before_text` "

G.3.3 Generation Prompt for GPT-4o

" You are a computer science researcher with extensive experience of scholarly writing. Here, you are writing a research paper in natural language processing using LaTeX languages.

Your writing intention is to "`put the verbalizer of the predicted intention label`" (`e.g., "initially create sections and sub-level structures" if the predicted label was section planning`).

Below is the paper you have written so far. Please strictly follow the writing intention given above and insert, delete, or revise at appropriate places in the paper given below.

Your writing should relate to the paper given below. Do not generate text other than paper content. Do not describe the changes you are making or your reasoning. Do not include sidenotes. Your output should only be the paper draft in latex, without the "'latex delimiters.

Input: `before_text`
Output: "

G.4 Seed Documents for Iterative Self-Writing

We present the four seed documents that we used for the iterative self-writing experiments, as shown in Listings 1 to 4.

G.5 Definition of Quantitative Metrics for Iterative Self-Writing

- *Lexical diversity*: Assess the unique tokens model generated in the final iteration of writing, measured by the number of unique tokens divided by the total tokens generated.
- *Topic consistency*: Cosine similarity between the seed document and output from the final iteration of writing.

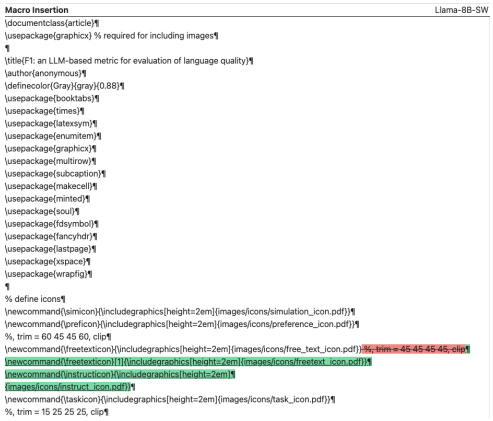
```
\begin{document}
\maketitle

\title{How Johnny Can Persuade LLMs to Jailbreak
     Them: Rethinking Persuasion to Challenge AI
     Safety by Humanizing LLMs}
\author{}
\date{}

\begin{abstract}
Most traditional AI safety research has
     approached AI models as machines and
     centered on algorithm-focused attacks
     developed by security experts. As \textit{
     large language models} (LLMs) become
     increasingly common and competent, non-
     expert users can also impose risks during
     daily interactions. This paper introduces a
     new perspective on jailbreaking LLMs as
     human-like communicators to explore this
     overlooked intersection between everyday
     language interaction and AI safety.
     Specifically, we study how to persuade LLMs
     to jailbreak them. First, we propose a
     persuasion taxonomy derived from decades of
     social science research. Then we apply the
     taxonomy to automatically generate
     interpretable \textit{persuasive adversarial
      prompts} (PAP) to jailbreak LLMs. Results
     show that persuasion significantly increases
      the jailbreak performance across all risk
     categories: PAP consistently achieves an
     attack success rate of over $92\%$ on Llama
     2-7b Chat, GPT-3.5, and GPT-4 in $10$ trials
     , surpassing recent algorithm-focused
     attacks. On the defense side, we explore
     various mechanisms against PAP, find a
     significant gap in existing defenses, and
     advocate for more fundamental mitigation for
      highly interactive LLMs.
\end{abstract}

\end{document}
```

Listing 1: An example seed document (Zeng et al., 2024) as shown in LaTeX codes to begin iterative self-writing.

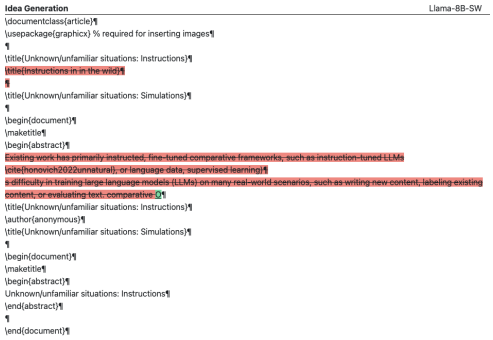
- *Intention coverage*: Assess the diversity of the model's writing intention, measured by the number of unique labels predicted through the entire 100 iterations divided by all 15 intended labels available in our taxonomy.

G.6 Iterative Training Sample Outputs

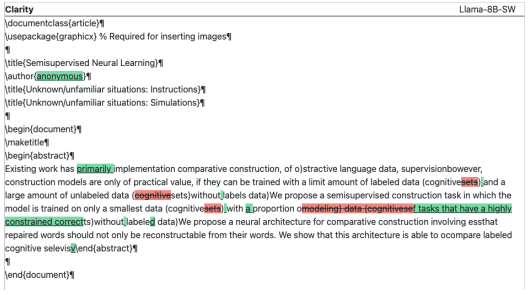
Figure 25 presents the sample outputs from the fine-tuned Llama-8B model during the iterative self-writing experiment. The model successfully was able to add several LaTeX commands to put some custom icon images ("Macro Insertion"). Also, it successfully revised several words and phrases in a paragraph for better clarity ("Clarity"). However,



(a) Llama writing inference for Macro Insertion activity. The model successfully added several LaTeX commands for custom actions.



(b) Llama writing inference for Idea Generation. The model failed to provide generated ideas and instead deleted abstract.



(c) Llama writing inference for Clarity. The model successfully revised words and phrases for clearer delivery.

Figure 25: Sample outputs from Llama-8B-SW during the self-writing experiment.

```
\begin{document}
\maketitle

\title{Read, Revise, Repeat: A System
    Demonstration for Human-in-the-loop
    Iterative Text Revision}
\author{}
\date{}

\begin{abstract}
Revision is an essential part of the human
    writing process. It tends to be strategic,
    adaptive, and, more importantly, \textit{
    iterative} in nature. Despite the success of
     large language models on text revision
    tasks, they are limited to non-iterative,
    one-shot revisions. Examining and evaluating
     the capability of large language models for
     making continuous revisions and
    collaborating with human writers is a
    critical step towards building effective
    writing assistants. In this work, we present
     a human-in-the-loop iterative text revision
     system, $\mathcal{R}$ead, $\mathcal{R}
    $evise, $\mathcal{R}$epeat (\textsc{$\
    mathcal{R}3$}), which aims at achieving high
     quality text revisions with minimal human
    efforts by reading model-generated revisions
     and user feedbacks, revising documents, and
     repeating human-machine interactions. In \
    method, a text revision model provides text
    editing suggestions for human writers, who
    can accept or reject the suggested edits.
    The accepted edits are then incorporated
    into the model for the next iteration of
    document revision. Writers can therefore
    revise documents iteratively by interacting
    with the system and simply accepting/
    rejecting its suggested edits until the text
     revision model stops making further
    revisions or reaches a predefined maximum
    number of revisions. Empirical experiments
    show that \method can generate revisions
    with comparable acceptance rate to human
    writers at early revision depths, and the
    human-machine interaction can get higher
    quality revisions with fewer iterations and
    edits.
\end{abstract}
\end{document}
```

Listing 2: An example seed document (Du et al., 2022a) as shown in LaTeX codes to begin iterative self-writing.

it struggled with understanding the definition of "Idea Generation," and the model just deleted all paragraphs instead.

Table 14 shows sample model outputs at different iterations (25th, 51st, and 100th), from the seed document (Du et al., 2022a) as shown in Listing 2.

```
\begin{document}
\maketitle

\title{Semisupervised Neural Proto-Language
    Reconstruction}
\author{}
\date{}

\begin{abstract}
Existing work implementing comparative
    reconstruction of ancestral languages (proto
    -languages) has usually required full
    supervision. However, historical
    reconstruction models are only of practical
    value if they can be trained with a limited
    amount of labeled data. We propose a
    semisupervised historical reconstruction
    task in which the model is trained on only a
     small amount of labeled data (cognate sets
    with proto-forms) and a large amount of
    unlabeled data (cognate sets without proto-
    forms). We propose a neural architecture for
     comparative reconstruction (DPD-
    BiReconstructor) incorporating an essential
    insight from linguists' comparative method:
    that reconstructed words should not only be
    reconstructable from their daughter words,
    but also deterministically transformable
    back into their daughter words. We show that
     this architecture is able to leverage
    unlabeled cognate sets to outperform strong
    semisupervised baselines on this novel task.
\end{abstract}

\end{document}
```

Listing 3: An example seed document (Lu et al., 2024) as shown in LaTeX codes to begin iterative self-writing.

G.7 Human Evaluation of Iterative Self-Writing Experiment

G.7.1 Study Procedures

Human evaluation was conducted on the outputs of two models: the LLama-8B-Instruct[13](Llama-8B-Zero) and its finetuned counterpart (Llama-8B-SW). Each evaluation session lasted approximately two hours and was conducted via a Zoom call. Participants were three graduate students of an R1 University in the United States, with extensive experience in Overleaf-based writing, and they were compensated with a US 40 dollar gift card for their effort.

Before the evaluation began, the author shared their screen to present the following information:

- A brief explanation of the research and an overview of the task (e.g., evaluating outputs from two models) using Google Slides.

[13]'Unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit'

```
\begin{document}
\maketitle

\title{Latxa: An Open Language Model and
    Evaluation Suite for Basque}
\author{}
\date{}

\begin{abstract}
We introduce Latxa, a family of large language
    models for Basque ranging from 7 to 70
    billion parameters.
Latxa is based on Llama 2, which we continue
    pretraining on a new Basque corpus
    comprising 4.3M documents and 4.2B tokens.
    Addressing the scarcity of high-quality
    benchmarks for Basque, we further introduce
    4 multiple choice evaluation datasets:
    EusProficiency, comprising 5,169 questions
    from official language proficiency exams;
    EusReading, comprising 352 reading
    comprehension questions; EusTrivia,
    comprising 1,715 trivia questions from 5
    knowledge areas; and EusExams, comprising
    16,774 questions from public examinations.
    In our extensive evaluation, Latxa
    outperforms all previous open models we
    compare to by a large margin. In addition,
    it is competitive with GPT-4 Turbo in
    language proficiency and understanding,
    despite lagging behind in reading
    comprehension and knowledge-intensive tasks.
     Both the Latxa family of models, as well as
     our new pretraining corpora and evaluation
    datasets, are publicly available under open
    licenses. Our suite enables reproducible
    research on methods to build LLMs for low-
    resource languages.
\end{abstract}

\end{document}
```

Listing 4: An example seed document (Etxaniz et al., 2024) as shown in LaTeX codes to begin iterative self-writing.

- An explanation of all intention labels and their corresponding definitions in our taxonomy.
- A walkthrough of the evaluation web application, including (1) information displayed in the user interface (UI); (2) the locations on the UI where the participants should focus, such as the intention label; and (3) how to move between different seed documents.
- A tutorial on how to complete the evaluation of the five metrics (accuracy, alignment, fluency, coherence, relevance) using the provided Google Sheet.

During the evaluation, participants were required to share their screen while using the evaluation web application in their browser. The author remained muted but monitored the participants' shared screens and the Google Sheet to ensure the process was proceeding smoothly. The author only interacted with participants to address questions, resolve technical issues, or clarify instructions. No unsolicited interaction was allowed.

The model identities were hidden from the participants. Instead, the models were labeled as "Model 1" and "Model 2." Specifically, "Model 1" corresponded to the Llama-8B-SW, and "Model 2" corresponded to the Llama-8B-Zero. The evaluation web app displayed two text boxes side by side, with an intention label predicted by the respective model shown in the top-left corner of each text box. Please refer to Figure 28 as a screenshot of the user interface that we developed for the human evaluation process.

After completing the evaluation, the authors thanked the participants for their time and effort, marking the conclusion of the Zoom call.

G.7.2 Definitions of Evaluation Metrics

Here are the full description of our human evaluation metrics for the iterative self-writing experiment:

- *Accuracy*: Out of 100, how many is the number of generated outputs that align with the provided intention?
- *Alignment*: Which model's whole writing process throughout the entire 100 iterations looks more human-like behaviors?
- *Fluency*: Which model's final writing sounds more grammatically correct?
- *Coherence*: Which model's final writing sounds more logical?
- *Relevance*: Does the final writing from each model contain related contents to the original seed document?

G.7.3 Results & Discussion

Tables 15 to 18 present the human evaluation of results for each seed document. We observe that the fine-tuned Llama-8B-SW did not outperform the baseline vanilla counterpart (Llama-8B-Zero) across all metrics for all four seed settings.

This discrepancy may be attributed to the way the prompt used for training isolates individual writing actions from the continuous, interconnected process typical of human writing. A single writing

action involves referencing the text before making an edit, the text after the edit, and the intention behind the edit. In human writing, these actions are cognitively and logically linked as part of a cohesive sequence. However, our model struggles to capture these connections due to the prompt structure, potentially causing it to become stuck in a local minimum.

Metrics	Model	Evaluator 1	Evaluator 2	Evaluator 3
Accuracy	SW	43	3	17
	Zero	47	22	38
Alignment	SW			
	Zero	X	X	X
Fluency	SW			
	Zero	X	X	X
Coherence	SW			
	Zero	X	X	X
Relevance	SW	Yes	No	No
	Zero	Yes	Yes	Yes

Table 15: Human evaluation results for the seed document 1.

Metrics	Model	Evaluator 1	Evaluator 2	Evaluator 3
Accuracy	SW	26	0	5
	Zero	48	12	29
Alignment	SW			
	Zero	X	X	X
Fluency	SW			
	Zero	X	X	X
Coherence	SW			
	Zero	X	X	X
Relevance	SW	Yes	Yes	Yes
	Zero	Yes	Yes	Yes

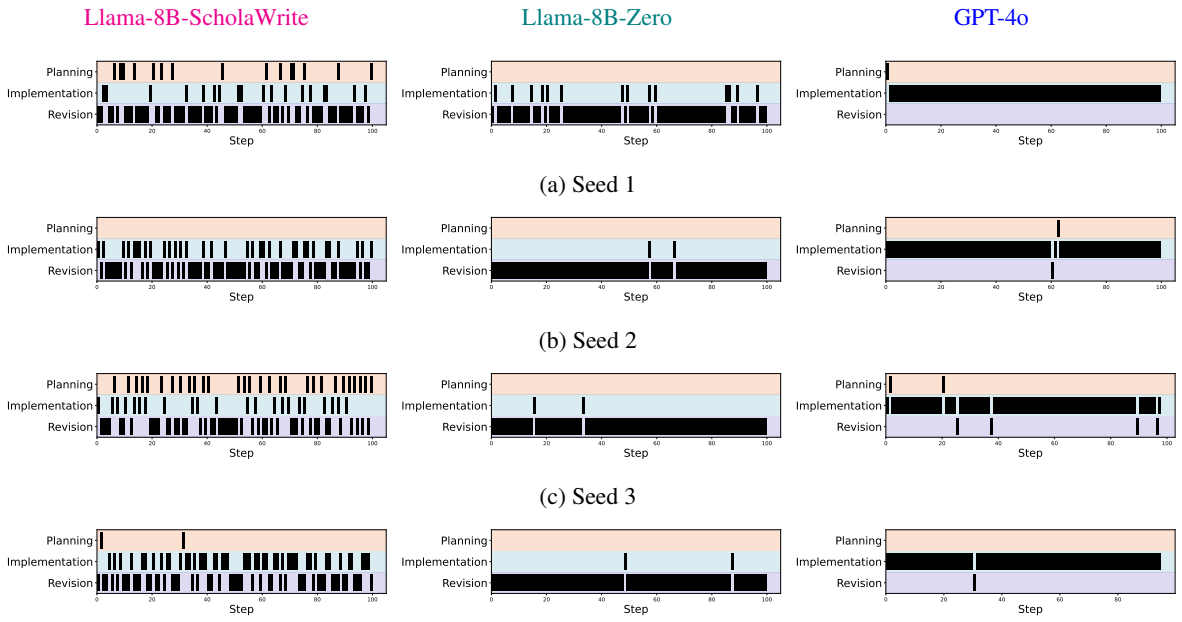
Table 16: Human evaluation results for the seed document 2.

Metrics	Model	Evaluator 1	Evaluator 2	Evaluator 3
Accuracy	SW	52	0	3
	Zero	70	23	43
Alignment	SW			
	Zero	X	X	X
Fluency	SW			
	Zero	X	X	X
Coherence	SW			
	Zero	X	X	X
Relevance	SW	Yes	Yes	No
	Zero	Yes	Yes	Yes

Table 17: Human evaluation results for the seed document 3.

Metrics	Model	Evaluator 1	Evaluator 2	Evaluator 3
Accuracy	SW	37	3	6
	Zero	60	22	48
Alignment	SW			
	Zero	X	X	X
Fluency	SW			
	Zero	X	X	X
Coherence	SW			
	Zero	X	X	X
Relevance	SW	Yes	No	No
	Zero	Yes	Yes	Yes

Table 18: Human evaluation results for the seed document 4.



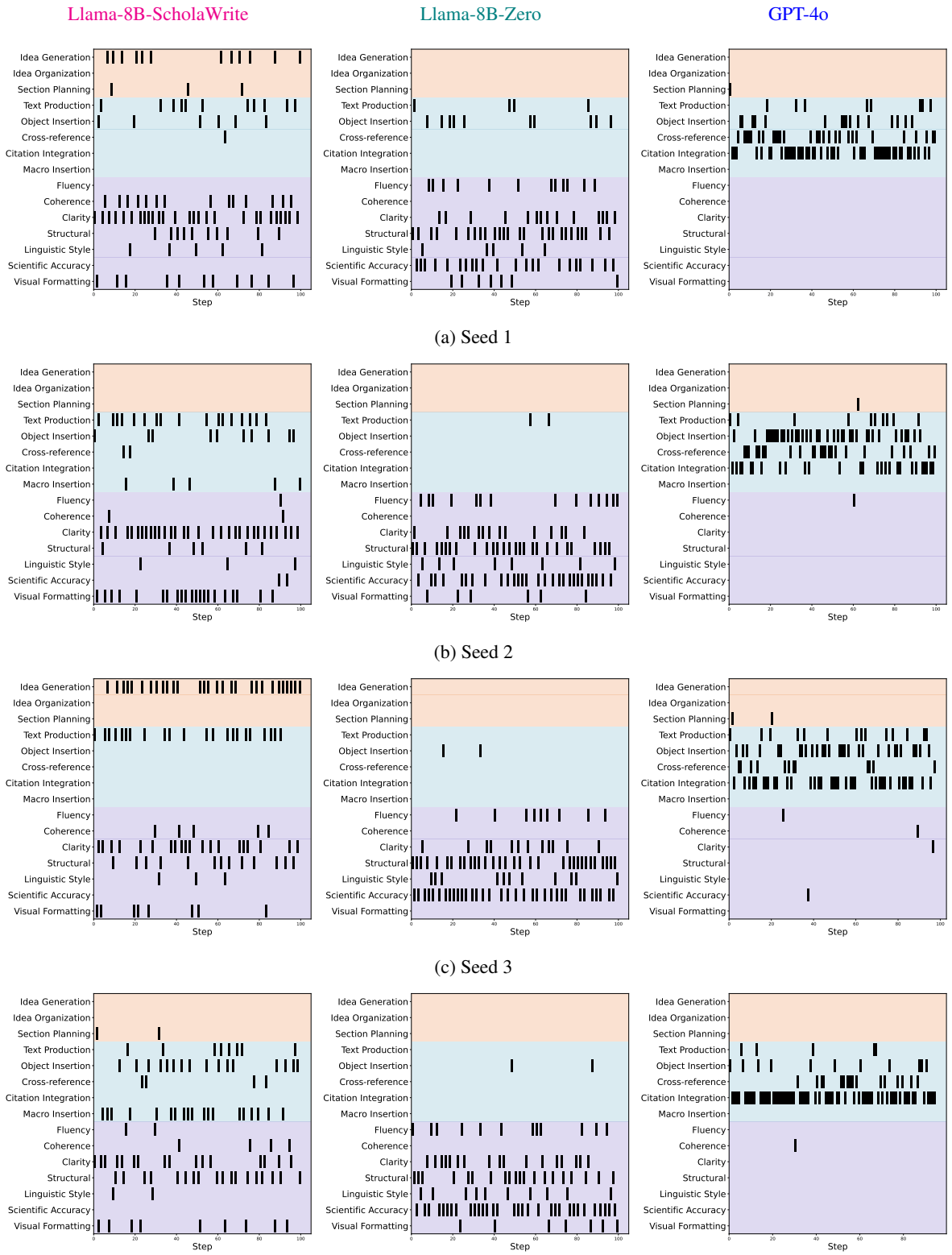
(a) Seed 1

(b) Seed 2

(c) Seed 3

(d) Seed 4

Figure 26: Distribution of high-level writing activities over time by models - Llama-8B-ScholaWrite (left); Llama-8B-Zero (middle); GPT-4o (right). Orange, Blue, and Purple represent Planning, Implementation, and Revision writing actions respectively.



(a) Seed 1

(b) Seed 2

(c) Seed 3

(d) Seed 4

Figure 27: Distribution of Per-intention writing activities over time by models - Llama-8B-ScholaWrite (left); Llama-8B-Zero (middle); GPT-4o (right). Orange, Blue, and Purple represent Planning, Implementation, and Revision writing actions respectively. We observe different writing patterns by model during the entire 100 iterations.

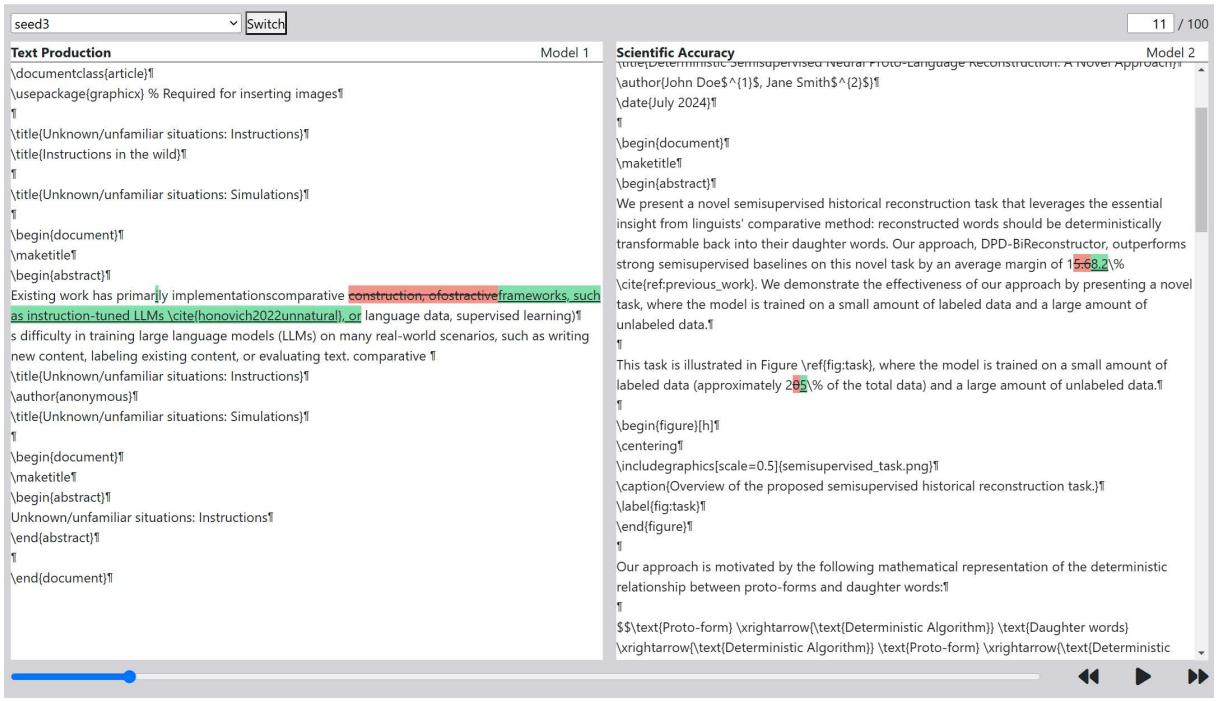


Figure 28: The user interface for the human evaluation process.