Process-Oriented Evaluation of AI-Assisted Scientific Writing

Patrick Queiroz Da Silva[♣] **Sanchaita Hazra**[♡] **Doeun Lee**[♣]
Sachin Kumar[♣] **Bodhisattwa Prasad Majumder**[♢]
[♣]The Ohio State University, Columbus, OH
[♡]University of Utah, Salt Lake City, UT
[♢]Allen Institute for AI, Seattle, WA

Project page: github.com/skai-research/process-revision-sci-writing
Emails: dasilva.30@osu.edu, sanchaita.hazra@utah.edu

Abstract

Bad writing hinders the publication of science. The role of artificial intelligence (AI) in generating and editing scientific texts remains unsettled. Abstracts serve as the critical gateway to scientific manuscripts, often shaping readers' interest. We inspect how individuals revise AI-generated abstracts compared to human-authored abstracts when incentivized to communicate scientific content. Using 869 keystroke-level edit logs with 240k total edits, we construct behavioral labels and measure linguistic properties of edit bursts to investigate the edit trajectories. AI abstracts exhibit higher sentence-level agency, whereas human-authored abstracts outperform in global coherence, even with edits. Experts engage in stigmatic behavior, switching their strategy from predominantly restructuring to substitution when AI source is disclosed. Language Models (LMs) improve edit outcomes through a mix of local and global features, but still actively struggle with global coherence. Both humans and LMs often target the weakest sections of abstracts, but fail to improve stronger areas. Our large-scale process-oriented evaluation highlights the perks and pitfalls of both human and LM editing processes as machine-generated texts emerge in scientific communication.

1 Introduction

Scientific texts are increasingly being shaped by Language Models (LMs). By helping reduce time and (often) effort required for drafting and revision, LMs are being used by researchers to accelerate key writing tasks (Liang et al., 2025b; Kusumegi et al., 2025). AI-assisted scientific writing is often perceived as more readable, which can boost writers' confidence and lead to fewer edits (Markowitz, 2024; McMinn et al., 2025; Hazra et al., 2026). To boot, studies find that AI-generated scientific abstracts appear credible and, with minor editing, are often accepted by reviewers (Gao et al., 2023; Holland et al., 2024; Hazra et al., 2026). At the same time, other work shows AI-generated scientific texts to lack in detail and insight (Tang et al., 2023; Hadan et al., 2024; Peters & Chin-Yee, 2025).

These diverging observations point to a key missing piece: how authors revise AI-generated scientific text. Revision and refinement lie at the heart of the writing process, offering a rich window into how authors strategize, evaluate, and refine text (Flower & Hayes, 1981; Du et al., 2022). Recent efforts in revision research in scientific writing have largely focused on human revision corpora (Jiang et al., 2022a), comment-to-edit alignment (D'Arcy et al., 2024), or paragraph-level automated revision (Jourdan et al., 2025).

Research focusing on AI aid in scientific writing has, so far, focused on final judgments of quality (Hazra et al., 2026; Tang et al., 2023; Hadan et al., 2024; Peters & Chin-Yee, 2025) and detectability (Sarzaeim et al., 2023; Ladha et al., 2023; Flitcroft et al., 2024). Character-level edits help to recover revision events, annotate writing processes, predict text quality, and

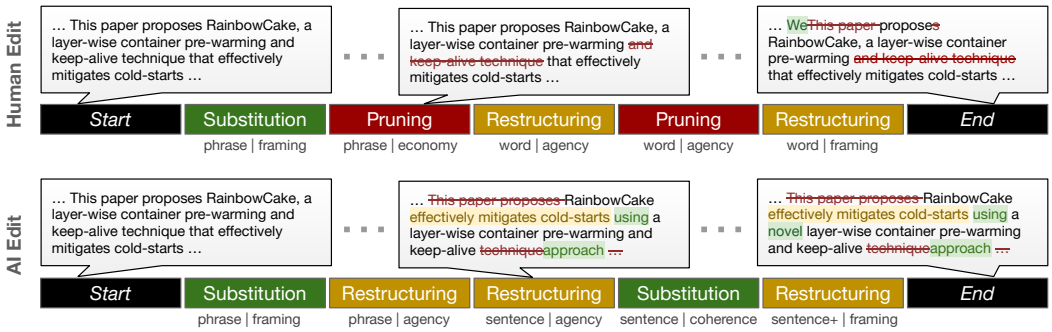


Figure 1: **Ebb and flow of scientific writing.** Human editors capitalize most improvements through substitution and pruning, but eventually fail to alter distributional differences between AI- and human-authored abstracts. LMs offer consistent improvements in sentence-level quality features, but enforcing human behavioral patterns degrades quality. We highlight edit behaviors in the sequence of colored boxes (more on subsection 2.2). Text below colored boxes indicates the scope of the change, followed by a '|', followed by the intent of the linguistic change (more on subsection 2.3).

study how writers allocate effort across revision—critical information unavailable in final outputs (Conijn et al., 2020; Velentzas et al., 2024; Tian et al., 2025; Miletić et al., 2022). Models like IteraTeR (Du et al., 2022), arXivEdits (Jiang et al., 2022b), ParaRev (Jourdan et al., 2025), among others, focus on revisions at sentence-, word-, and paragraph-levels to establish the crucial role of revisions in scientific writing with AI; however, two lingering questions remain.

First, the existing models evaluate either final text quality or local revision behavior, without explicitly relating the two. Second, the revisions are often represented as isolated edit operations or aligned draft pairs. This obscures how draft-level properties drive subsequent edits and how sequences of edits alter a document over time. Instead of evaluating only final outputs or cataloging edits in isolation, ***we analyze how properties of the current draft predict subsequent revision behavior and how those revisions alter the evolving text by jointly modeling the text state and the revision state.***

We leverage large-scale experiment data from Hazra et al. (2026) where participants are segregated into two pools: authors and reviewers. Authors revise an abstract, and reviewers assess if the edited version faithfully reflects the content of the original draft. Importantly, the data corpus captures timestamped keystroke-level edits performed while editing the abstracts. For this study, we infer edit trajectories from these keystroke-level logs by transforming noisy timestamped keyboard and cursor operations into a behaviorally interpretable taxonomy of bursts. Using these metrics, we study fine-grained edit trajectories and analyze distributional differences in revision behavior between human and AI-authored texts. We summarize our contributions as follows:

- We motivate the joint modeling of text and revision state for scientific writing to formulate human editing strategies and evaluate their efficacy on editing both human-authored and AI-generated abstracts.
- We observe that human editors uphold local agency ($p < .01$) and structure ($p < .01$) in AI abstracts, but global coherence falls behind ($p < .01$) in human abstracts. Multiple threads of evidence suggest human editors do not close the distributional gap between AI and human-authored abstracts, hinting at the risk of macro-level behavior homogenization.
- We demonstrate the mixed capabilities of LM editors. They excel locally ($p < .001$), but struggle at the discourse level ($p > .05$); improve the weakest abstracts to average ($p < .001$), but degrade high-quality ones ($p < .05$), indicating LM editors are situationally useful in improving the quality of scientific texts.

2 Data and methods

Keystroke-level edits unearth the sequential process of editing. We act on keystroke logs containing timestamped keyboard and cursor operations performed while editing a piece of scientific text (abstracts) at the character level. Editing behavior emerges from the interplay between a starting text state and a set of revisions that modify it. Our methodological approach involves producing behaviorally meaningful sequences, followed by quantifying the linguistic properties of the texts in sequence.

2.1 Data

We borrow the data from Hazra et al. (2026), who use abstract composition as an element of scientific writing. Hazra et al. (2026) select 45 published papers from notable conference venues in Computer Science. For each paper, they produce 45 respective AI-generated abstracts, controlling for the content and quality of the information.

Group	Variable	Summary
Activity	Entry pause (s)	34.00 (308.23)
	Burst duration (s)	16.00 (108.52)
	Edits per burst	58.25 (129.23)
Navigation	No movement	95.0 (13.0)
	Backward	1.4 (6.9)
	Forward	3.6 (11.0)
Action	Insert	43.9 (41.6)
	Delete	32.9 (38.5)
	Substitute	23.2 (31.8)
Scope	Word	9.2 (28.7)
	Phrase	61.5 (48.0)
	Sentence+	29.3 (45.0)

Table 1: Edit burst descriptive statistics for the revision data from Hazra et al. (2026). Entry pause and burst duration are reported as median (SD). All remaining variables are reported as mean (SD).

The core experiment involves a 2 × 2 between-subject design with the abstract provenance in one arm and its disclosure in the other. Participants with relevant experience are split into author (domain experts) and reviewer (domain experts) pools. Authors are randomly assigned to one of the four treatments. Authors are incentivized to revise the provided abstracts to ensure their edited abstracts are accepted by at least two independent reviewers. Reviewers compare the edited abstracts against the original abstract to provide accept/reject judgments[1]. Edits are captured and timestamped at the keystroke level.

The data thus comprises 45 pairs of abstracts from the selected seed papers: 45 original human abstracts and 45 of their AI-generated counterparts. Each author is randomly assigned to edit three abstracts, generating 891 keystroke-level edit logs. We dropped 22 editing sessions that were deemed AI-assisted. Finally, we were left with 869 keystroke-level edit logs corresponding to 236,033 total edits. The data also contained additional anonymized demographic information about the participants, including education. The descriptive statistics of this data are reported in Table 1.

2.2 Revision process

Analyzing edit trajectories at the character level is noisy and difficult to interpret. Instead, we convert these singular events into behaviorally separable ***bursts***. Common practice segments bursts using pause thresholds (Leijten & Van Waes, 2013). In our pipeline, we focus on creating bursts that represent contiguous local revisions. Thus, we allow both strong relocation and pauses greater than two seconds to segment the bursts. We take additional measures to detect and merge edge cases, e.g., pausing within a word. In total, we create 3,941 bursts after segmentation.

Burst dimensions. We aggregate events within bursts into four dimensions, largely inspired by standard keystroke logging metrics (Baaijen et al., 2012).

[1]For more details about data collection and experiment design, refer to Section 3 in Hazra et al. (2026).

Family	Behavior label	Definition
Expansion	Phrasal expanding	Phrase scope, insert-only
Expansion	Sentence+ expanding	Sentence+ scope, insert-only
Expansion	Sentence+ elaborate	Sentence+ scope, insert with minor delete/substitute
Pruning	Backward word prune	Single-word delete with backward navigate
Pruning	Backward span prune	Phrase/sentence scope delete with backward navigate
Pruning	Forward word prune	Single-word delete with forward navigate
Pruning	Forward span prune	Phrase or sentence scope delete with forward navigate
Pruning	Phrasal prune	Phrase scope, deletion-only
Pruning	Phrasal condense	Phrase scope delete/substitute
Substitution	Forward word replace	Single-word replace with forward navigate
Substitution	Phrasal replace	Phrase scope, balanced insert/delete
Substitution	Phrasal rewrite	Phrase scope, mixed insert/delete/substitute
Substitution	Phrasal substitute	Phrase scope, substitute-only
Substitution	Phrasal insert+substitute	Phrase scope substitute with added insert
Restructuring	Sentence+ restructure	Sentence+ scope, mixed insert/delete/substitute
Restructuring	Cross-scope restructure	Mixed-scope, mixed insert/delete/substitute

Table 2: Edit-burst behavior labels. Sentence+ denotes sentence-level or larger editing spans.

- ***Pauses*** delineate text production vs physical inactivity. We compute two metrics related to time: (1) the average time in seconds prior to starting a burst, and (2) the average pause time within a burst. Both pause metrics are log-transformed due to extreme positive skew. We additionally track (3) the number of edits made within a burst.
- ***Navigation*** tracks relative cursor position as a proxy for attention and grounds the movement of edits. We measure the share of (1) movement among forward, (2) backward, and (3) no movement, quantified as a proportion of the entire text.
- ***Actions*** describe the edit operations: (1) insert, (2) delete, and (3) substitute. We calculate the share among insert, delete, and substitute edits.
- ***Scope*** elaborates the breadth of text necessary for understanding the context of an edit. We categorize scope into (1) word-level, (2) phrase-level, and (3) sentence+ using the pre- and post-burst text difference.

Defining burst behaviors. To better interpret the twelve continuous features per burst (three from each of the four burst dimensions above), we cluster the variables with a Gaussian mixture model, choosing up to twenty clusters. The final cluster count is determined by cluster fitness with the Bayesian Information Criterion (BIC) and additional stability analysis. Perfectly collinear features (e.g. action share) are transformed via an isometric log-ratio prior to clustering.

In total, the algorithm chose sixteen clusters. We manually label each according to the centroid averages on the twelve features, naming according to dominant action mix, dominant scope, and salient navigation direction; we focus on observable editing behavior rather than assuming latent states. The edit burst behaviors, described in Table 2, fall into four general families with frequencies in parentheses: expansion (28%), pruning (23%), substitution (30%), and restructuring (19%). **Expansion**, **pruning**, and **substitution** relate to edits which primarily add, remove, or replace material, respectively. **Restructuring** reworks content through multiple edit actions at a broader scope. The cluster fit metrics and feature means per cluster are available in Tables 4 to 6 in Appendix D.

2.3 Linguistic text properties

Editing inherently perturbs the linguistic properties of a text. The readability of a scientific text improves when contextual material appears in the topic position and new or salient information is placed in the stress position, where readers expect it (Halliday, 1967; Gopen & Swan, 1990). Clarity, likewise, is enhanced when grammatical subjects mention concrete "characters" and verbs express actions directly, rather than obscuring them through nom-

Δ Dimension	β_{source}	$\beta_{\text{pre\_edit}}$	ΔR^2
Agency	−0.05	−0.34**	0.10
Economy	−0.02	−0.37**	0.12
Structure	0.02	−0.39**	0.14
Coherence	−0.13**	−0.39**	0.08
Framing	0.00	−0.23**	0.03

Table 3: Effect sizes of text source and pre-edit linguistic properties on the edit outcome at the burst level, and the difference in variance explained between the pre-edit and source. $*p < .05$, $**p < .01$.

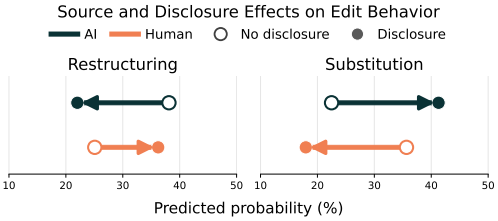


Figure 2: Edit behavior changes in those with masters+ education moderated by abstract source (human vs AI) and disclosure (no disclosure vs disclosure)

inalizations (Swales, 2014; Williams & Bizup, 2014). We thus need to create an outcome that captures the effect of the behavioral states established in subsection 2.2. We group 25 computational linguistic metrics, derived from Hazra et al. (2026), into five theory-driven dimensions to improve the broader interpretability of results. We focus on descriptive dimensions calculable from the text. Please see Table 7 in Appendix D for a complete description of all 25 variables.

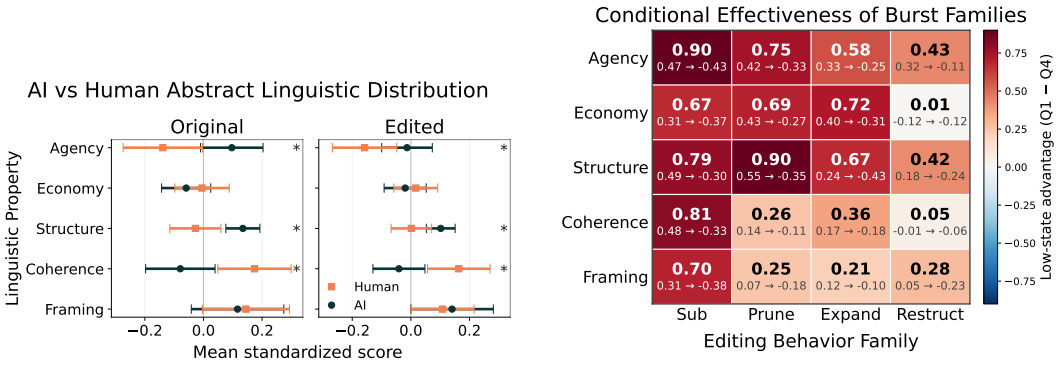
All dimensions are coded such that higher scores tend to indicate better writing. Each individual metric is standardized in reference to the initial sample collected in (Hazra et al., 2026) for comparability. Thus, the median value for each metric is zero, and a score of one represents a single standard deviation above the average. We take a mean of the individual metrics to compute the final dimension score. The linguistic text properties are defined as follows:

- **Agency** clarifies the role between the actors (subjects) and their actions (verbs) (Hyland, 2002; 2005; Gopen & Swan, 1990; Halliday & Matthiessen, 2014). Clear writing makes it easy to identify who is doing what.
- **Economy** quantifies how efficiently a text conveys complex information without unnecessary compression (Halliday, 2004; Biber & Gray, 2013). It represents the tradeoff between information-rich writing and reading difficulty. Hallmarks of poor economy include nominalizations, dense noun clusters, and filler words.
- **Structure** describes how the arrangement of words and phrases leads to the primary grammatical relationships (Gopen & Swan, 1990; Gibson, 1998; Williams & Bizup, 2014). Strong writing prioritizes both succinct links between its core components and balanced composition.
- **Coherence** measures how well a text introduces new information and connects threads (Graesser et al., 2004; Halliday & Hasan, 2014). Best practice suggests pushing new information later within a sentence, and keeping topics similar across nearby sentences.
- **Framing** tracks the placement of context and claims (Swales, 1990; Hoey, 2000). Balanced positioning allows the reader to understand a point and its significance best.

3 Behavioral and linguistic analyses on edit logs

3.1 Pre-edit linguistic features best explain edit trajectories.

We fit separate linear models predicting the standardized change in each of the linguistic properties from abstract provenance (AI vs. Human) and the corresponding standardized pre-edit score. We additionally controlled for the burst position within the session, and the source paper of the abstract. In all five metrics, pre-edit linguistic properties substantially outweigh provenance in predictive power. Lower initial scores were associated with larger improvements, evidenced by the negative pre-edit metric effect sizes ($\beta_{\text{pre\_edit}}$) ranging from -0.23 to -0.39 (all $p < .01$) shown in Table 3. In contrast, provenance explained relatively little variance; only coherence showed a small significant effect ($\beta_{\text{source}} = -0.13$, $p < .01$).



(a) (b)

Figure 3: (a) Linguistic properties of original and edited AI-and-human authored abstracts. (b) Editing behavior efficacy in low and high scoring regions of the abstracts. For the top value, higher means more beneficial in weaker states. The bottom values are the mean Q1 and Q4 respectively.

Even so, the abstract source explained only a small portion of unique variance compared to pre-edit metrics ($R^2_{\text{source}} = 0.004$, $R^2_{\text{pre\_edit}} = 0.085$)[2].

3.2 Education moderates behavior related to AI disclosure.

Multinomial regression accounting for education level reveals several significant differences in editing behavior. Those with bachelors or lower use the four burst categories with relative similarity for AI and human authored abstracts. However those with masters or higher education (37% of the data) vary behaviorally depending on both source and disclosure. Figure 2 shows that for AI abstracts, disclosure is associated with severe drops in restructuring ($\Delta = -16.1\%$, $p < .001$) in favor of substitution ($\Delta = +18.8\%$, $p < .001$). Disclosing human source had the opposite effect: restructuring ($\Delta = +11.1\%$, $p < .05$) and substitution ($\Delta = -17.7\%$, $p < .01$). This reflects the common stigma against AI text. Editors are only willing to make restructuring changes to AI text without disclosure, and much rather perform substitutions when they know the text is AI. The full statistics for all education levels and contrasts are available in Table 9.

3.3 Human revision does not improve final abstract quality.

Human and AI abstract linguistic distributions remain divergent. Hazra et al. (2026) reports significant differences in number of edits across source and disclosure, which suggests certain linguistic improvements may be sensitive to the treatment condition. Because revision is a consequence of the initial properties of the abstract, we first conduct a paired t-test between human-AI abstract counterparts, visualized in Figure 3a. AI abstracts were significantly stronger than human abstracts in sentence-level agency ($d_z = 0.45, p = .008$) and structure ($d_z = 0.54, p = .004$), but underperformed in global coherence ($d_z = -0.47, p = .007$). The only two significant changes post-editing occurred in AI, where Agency decreased ($d_z = -0.60, p = .001$) and Economy increased ($d_z = 0.46, p = .009$). However, the original distributional differences between human and AI authored abstracts remain, despite an average of 300 individual edit actions, and 6 bursts per abstract editing session. Please refer to Tables 11 to 13 in Appendix E for the full statistics.

Pre-edit linguistic features moderate behavior efficacy. We further investigate the result from subsection 3.1. Pre-edit regions with lower scores show the strongest improvement. We calculate the mean improvement in the highest (Q4) and lowest (Q1) quartiles, attributed to

[2]Additional details on the partial R^2 are shown in Table 8 in Appendix E.

their respective burst behavior family. Figure 3b shows that burst behaviors vary in efficacy, moderated by the linguistic property and its starting state. Substitution most strongly and broadly repairs weak states, suggesting that replacing text in place is the strongest method to improve low-quality starting states. Pruning is selectively strong for agency, economy, and structure. Intuitively, deleting text should help de-clutter and improve the structure of bloated text. Adding new text is similarly helpful for sentence-level metrics, but not as strong as pruning; likely because expansion without pruning means the original text dilutes the performance improvements. Restructuring shows a relatively weak effect in repairing low-quality regions of the abstract. Finally, all of the Q4 behavior efficacies are negative, suggesting that humans tend to hinder the highest quality abstracts. The full metrics and confidence intervals are reported in Table 10 in Appendix E.

Implications: Human editors capitalize on the lower quality regions of abstracts. They tend to make the most improvements through substitution and pruning, rather than heavier restructuring. While those with a masters' degree or above are willing to restructure AI text, disclosure is associated with a reversed effect, inducing significantly more substitution. Ultimately, human editors fail to alter the distributional differences between AI- and human-authored abstracts.

4 Language models as editing assistants

Poor human editing outcomes seen in subsection 3.3 suggest ample opportunity for the growing adoption of LMs as assistants (Liang et al., 2025a). We test two common applications of AI for scientific abstracts to identify whether LMs can enhance or compensate for deficiencies in human editing. First, revising an already written abstract (zero-shot editing), and second, revising a specific component of the abstract (scoped assistant). We experiment with the latest versions of GPT (5.4) and Claude (Opus 4.6), reflecting upon the systems people are most likely to use for similar tasks in the current time.

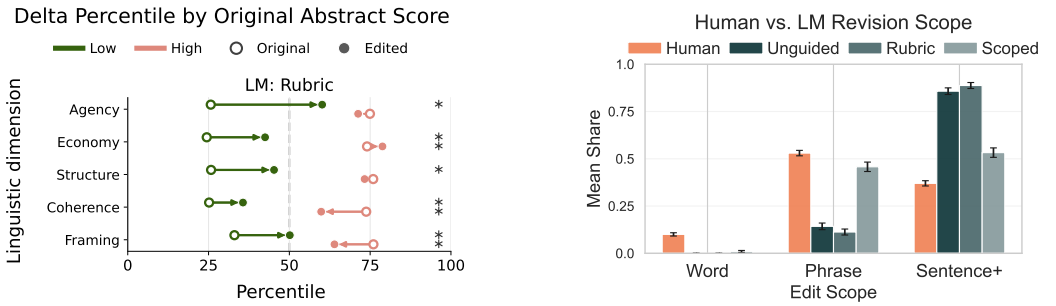
4.1 Use case 1: LM as a zero-shot editor

We prompt the LM to edit the 45 abstract pairs to improve their quality using two separate task descriptions.

- **Unguided** provides general instructions about scientific abstracts (the same instructions seen by participants in Hazra et al. (2026)). **Unguided** represents how a user might interact with an LM naturally.
- **Rubric** provides a rubric capturing all linguistic properties described in subsection 2.3 (see Table 7 in Appendix D). **Rubric** grounds the LM in metrics that make an abstract stronger. Detailed prompts can be found in subsection F.1

LMs improve locally but struggle globally. As established in subsection 3.3, humans do not improve the linguistic dimensions significantly. In contrast, **Unguided** shows significant improvements in economy, further improved by LMs with **Rubrics** ($d_z = 0.53, p < .001$). **Rubrics** additionally improve agency ($d_z = 0.50, p < .001$) and structure ($d_z = 0.37, p < .001$) during editing. While framing is significant in **Unguided** generation, the differences remain underpowered when a consistent **Rubric** is provided. Congruent with the results from subsection 3.3, LMs as editors largely improve on the same distributions that were originally strong in AI-authored abstracts. Overall, the results imply: while LMs robustly improve sentence-level metrics, they struggle at the discourse level. Complete statistical test results are available in Tables 14 and 15 in Appendix F.

LM editors make bad abstracts average, and good abstracts worse. Figure 4a delineates linguistic improvement of the bottom and top-scoring abstracts. For this analysis, abstracts are split by median based on each linguistic dimension. The lowest-scoring abstracts consistently receive the strongest improvements, lifting them from an average percentile of 26.8% to 46.7%. All individual metric gains tested significant and are available in Table 16 in Appendix F. On the other hand, the highest-scoring abstracts often had no significant gain. Rather, coherence dropped modestly (13.9%, $p < .001$). Economy showed signs of



(a) (b)

Figure 4: (a) LMs fail to improve low-scoring abstracts to above average. Besides economy, high-scoring abstracts suffer or stay the same. (b) LMs consistently performed larger-scoped edits than humans. Prompting with the desired scope improves alignment, but still produces significantly larger edits.

improvement (4.8%, $p < .05$). We find LM editors may enhance the lowest-scoring abstracts up to the mean; they often struggle to maintain consistency and improve the best abstracts, often actively hindering them.

4.2 Use case 2: LM as a scoped assistant

Humans edit in targeted, interpretable bursts. Aligned assistants should edit in a way that is congruent with human intent. We measure LM editing alignment in this case by asking the LMs to improve the abstract scoped to the same chunk of text a human edited. Each human editing burst could correspond to a single phrase, sentence, or the entire abstract. We sample 200 such bursts from the original 90 abstracts using inverse distribution weighting to ensure a diverse set of editing actions, scope, and outcomes. We utilize the same **Unguided** and **Rubric** experiment descriptions as in the zero-shot setup. In addition, we add another description, **Scoped**, that mandates explicitly that the LM must adhere to the exact editing scope of the human (word, phrase, sentence+).

LMs betray editing scope. Figure 4b shows that even when conditioned on the same scope, models often ignore the scope and make wider-reaching changes across sentences. This misalignment is difficult for a human to verify visually, especially with increased text lengths, and can compound over an editing session. **Scoped** results in better alignment with human editing intentions, but still results in significantly more sentence+ revisions. All statistically significant differences are reported in Table 18 in Appendix F.

LMs with scoped edits are weaker at improving the worst abstracts. **Unguided** and **Rubric** LM editors at the burst level follow the same trends as in the zero-shot setup. We point readers to Table 19 in Appendix F for additional details. More interestingly, **Scoped** editors underperform compared to **Rubric** in the lowest scoring abstracts (average $\Delta = -4.6\%$, all individual tests $p < .05$, besides coherence $p = .351$). Contrastingly, for high-scoring abstracts, their scores are not significantly different. This suggests that LMs do not edit their strongest when forced to align with human scope. Further post-training efforts should focus on aligning LMs to faithfully follow human editing intentions. All statistics are reported in Table 17 in Appendix F.

Implications: LMs offer consistent improvements on sentence-level features, but only by bringing the lowest-scoring abstracts to average. Language models also consistently make larger-scoped edits than humans. Forcing LMs to edit closer to human scopes harms performance on the lowest-scoring abstracts.

5 Discussion

Stigma and AI writing. Nature quotes "Tools such as ChatGPT threaten transparent science" (Nature, 2023). Is this stance generalizable? We find that education is a key marker of stigma in responses to AI-assisted scientific writing among individuals with a master's degree. The usefulness of AI for scientific writing clearly hinges on the willingness to adopt and meaningfully engage with it. While we see efficiency gains from using AI assistance in revising bad-to-moderate scientific texts, this resistance may be concentrated and percolated among those with stronger disciplinary writing conventions.

Homogenizing behavior. LMs tend to reduce self-correction and increase irrational decision-making (Cheng et al., 2026; Sharma et al., 2024; Bhattacharya et al., 2026). Our results imply parallels. Human editors focused on the most visible weaknesses in a text, yet these interventions rarely closed the broader quality gap between AI- and human-generated abstracts. This points to an alarming situation in which homogenized AI-generated text may prevail even after human editing. Cognitively, repeated exposure to apparently fluent but suboptimally structured AI-generated scientific texts may explain the tendency towards superficial edits (Abdulhai et al., 2026). LM editors are in no position to rescue; in fact, they can often make the situation worse by degrading the quality of high-quality abstracts. Our findings mirror findings from (Noy & Zhang, 2023; Doshi & Hauser, 2024; Agarwal et al., 2025b) where larger gains with LM assistance are observed with weaker writers or weaker abstracts, but to raise the quality to a stylistically homogeneous baseline.

6 Additional Related work

AI writing and edit-based evaluation. The spread of AI writing tools makes it essential to review how humans interact with, assess, and edit AI-generated texts. AI-generated creative texts exhibit 3x to 10x more stylistic weaknesses than human-authored texts (Chakrabarty et al., 2024). Editing human writing and AI-generated text also presents challenges, as Chakrabarty et al. (2025) observes that although edits for AI-generated texts are broadly similar, they were needed much more extensively compared to human-authored texts. IteraTeR (Du et al., 2022), arXivEdits (Jiang et al., 2022a), ParaRev (Jourdan et al., 2025) contribute to scientific revision being an iterative and structured process, with authors making systematic edits for style, precision, simplification, grammar, and formatting across full-paper versions rather than merely altering final text superficially. ScholaWrite (Le et al., 2025) extends this line of research by providing keystroke-level traces of end-to-end scholarly writing, enabling the study of revision at a finer temporal resolution.

AI for scientific writing. LMs are increasingly shaping scientific writing, from assistant plugins like Writefull[3] being integrated into the editing platforms of Overleaf to tools that support drafting, revision, and even full paper generation (Ito et al., 2019; Bezerra et al., 2021; Wang et al., 2019; Kanna, 2024; Liebling et al., 2025; Kousha & Thelwall, 2025; Lu et al., 2024). At the same time, anecdotal cases of seemingly AI-generated papers and peer reviews, along with broader critiques of generative AI in academia, have raised concerns about authenticity and quality (Cybernews, 2023; DeGeurin, 2024; Oransky & Marcus, 2024; Liang et al., 2024). Corpus-level research further suggests that LLM-generated content is becoming widespread in scientific writing and may be contributing to more standardized or declining prose quality across disciplines (Matsui, 2024; Lemire, 2024; Geng & Trotta, 2024; Kobak et al., 2025). More recently, Hazra et al. (2026) foregrounds the need to understand how humans interact with AI-generated texts, how much agency they retain, how edited AI-generated texts are perceived by an external evaluator, and how these systems reshape scientific communication norms (Reza et al., 2025).

[3] https://www.writefull.com/

7 Conclusion

LMs are reshaping writing and editing, but little is known about how authors revise AI-generated scientific text. We push the community towards a process revision mindset, using bursts in edit trajectories to reveal patterns in editing strategy. We highlight heterogeneous editing patterns by human editors at different scopes (word to sentence), but it is often ineffective to change an AI-generated text extensively, suggesting broader implications of behavior homogenization. While LMs tend to improve sentence-level features during editing, the benefits remain concentrated on the lowest-scoring abstracts, telling a cautionary tale of LM assistance in scientific writing.

References

Marwa Abdulhai, Isadora White, Yanming Wan, Ibrahim Qureshi, Joel Leibo, Max Kleiman-Weiner, and Natasha Jaques. How llms distort our written language. *arXiv preprint arXiv:2603.18161*, 2026.

Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, and Peter Clark. Autodiscovery: Open-ended scientific discovery via bayesian surprise. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=kJqTkj2HhF.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pp. 1–21, 2025b.

Veerle M. Baaijen, David Galbraith, and Kees de Glopper. Keystroke analysis: Reflections on procedures and measures. *Written Communication*, 29(3):246–277, 2012. doi: 10.1177/0741088312451108.

João Ribeiro Bezerra, Luís Fabrício Wanderley Góes, and Wladmir Cardoso Brandão. Newriter: A text editor for boosting scientific paper writing. In *International Conference on Enterprise Information Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235259213.

Haimanti Bhattacharya, Subhasish Dugar, Sanchaita Hazra, and Bodhisattwa Majumder. The good, the bad, and the ugly: The role of ai quality disclosure in deception detection. *Journal of Behavioral and Experimental Economics*, pp. 102555, 2026.

Douglas Biber and Bethany Gray. *Nominalizing the verb phrase in academic science writing*, pp. 99–132. Studies in English Language. Cambridge University Press, 2013.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–33, 2025.

Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic ai decreases prosocial intentions and promotes dependence. *Science*, 391(6792): eaec8352, 2026. doi: 10.1126/science.aec8352. URL https://www.science.org/doi/abs/10.1126/science.aec8352.

Rianne Conijn, Emily Dux Speltz, Menno Van Zaanen, Luuk Van Waes, and Evgeny Chukharev-Hudilainen. A process-oriented dataset of revisions during writing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 363–368, 2020.

Cybernews. Academic cheating surges with ai: Google search exposes researchers using chatgpt. *Cybernews*, 2023. URL https://cybernews.com/news/academic-cheating-chatgpt-openai/.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. ARIES: A corpus of scientific paper edits made in response to peer reviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6985–7001, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.377. URL https://aclanthology.org/2024.acl-long.377/.

Mack DeGeurin. How ai-generated text is flooding scientific journals. *Popular Science*, 2024. URL https://www.popsci.com/technology/ai-generated-text-scientific-journals/.

Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.

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

Madelyn A Flitcroft, Salma A Sheriff, Nathan Wolfrath, Ragasnehith Maddula, Laura McConnell, Yun Xing, Krista L Haines, Sandra L Wong, and Anai N Kothari. Performance of artificial intelligence content detectors using human and artificial intelligence-generated scientific writing. *Annals of Surgical Oncology*, 31(10):6387–6393, 2024.

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

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *NPJ digital medicine*, 6(1): 75, 2023.

Mingmeng Geng and Roberto Trotta. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*, 2024.

Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1): 1–76, 1998.

George D Gopen and Judith A Swan. The science of scientific writing. *American scientist*, 78 (6):550–558, 1990.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. Coh-metrix: analysis of text on cohesion and language. *Behav. Res. Methods Instrum. Comput.*, 36(2): 193–202, May 2004.

Hilda Hadan, Derrick M Wang, Reza Hadi Mogavi, Joseph Tu, Leah Zhang-Kennedy, and Lennart E Nacke. The great AI witch hunt: Reviewers' perception and (mis)conception of generative AI in research writing. *Computers in Human Behavior: Artificial Humans*, 2(2): 100095, August 2024.

M A K Halliday and Ruqaiya Hasan. *Cohesion in English*. Routledge, January 2014.

M. A. K. Halliday and Christian M. I. M. Matthiessen. *Halliday's Introduction to Functional Grammar*. Routledge, Abingdon, Oxon, 4th edition, 2014. ISBN 9781444146608.

Mak Halliday. *Writing Science: Literacy And Discursive Power*. Critical Perspectives on Literacy and Education. Routledge, January 2004.

Michael AK Halliday. Notes on transitivity and theme in english part i. *Journal of linguistics*, 3(1):37–81, 1967.

Sanchaita Hazra, Doeun Lee, Bodhisattwa Prasad Majumder, and Sachin Kumar. Accepted with minor revisions: Value of ai-assisted scientific writing. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, IUI '26, pp. 1–22. ACM, March 2026. doi: 10.1145/3742413.3789140. URL http://dx.doi.org/10.1145/3742413.3789140.

Michael Hoey. *Textual interaction*. Routledge, London, England, September 2000.

Andrew M. Holland, William R. Lorenz, James C. Cavanagh, et al. Comparison of medical research abstracts written by surgical trainees and senior surgeons or generated by large language models. *JAMA Network Open*, 7(8):e2425373, 08 2024. doi: 10.1001/jamanetworkopen.2024.25373.

Ken Hyland. Authority and invisibility. *J. Pragmat.*, 34(8):1091–1112, August 2002.

Ken Hyland. *Metadiscourse: Exploring Interaction in Writing*. 2005.

Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. Diamonds in the rough: Generating fluent sentences from early-stage drafts for academic writing assistance. In *International Conference on Natural Language Generation*, 2019. URL https://api.semanticscholar.org/CorpusID:204800442.

Chao Jiang, Wei Xu, and Samuel Stevens. arXivEdits: Understanding the human revision process in scientific writing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9420–9435, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.641. URL https://aclanthology.org/2022.emnlp-main.641/.

Chao Jiang, Wei Xu, and Samuel Stevens. arxivedits: Understanding the human revision process in scientific writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9420–9435, 2022b.

Léane Jourdan, Florian Boudin, Richard Dufour, Nicolas Hernandez, and Akiko Aizawa. ParaRev : Building a dataset for scientific paragraph revision annotated with revision instruction. In Michael Zock, Kentaro Inui, and Zheng Yuan (eds.), *Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025)*, pp. 35–44, Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics. URL https://aclanthology.org/2025.wraicogs-1.4/.

P Kanna. How much research is being written by large language models, 2024.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025.

Kayvan Kousha and Mike A Thelwall. How much are llms changing the language of academic papers after chatgpt? a multi-database and full text analysis. 2025. URL https://api.semanticscholar.org/CorpusID:281252711.

Keigo Kusumegi, Xinyu Yang, Paul Ginsparg, Mathijs de Vaan, Toby Stuart, and Yian Yin. Scientific production in the era of large language models. *Science*, 390(6779):1240–1243, 2025.

N Ladha, K Yadav, and P Rathore. Ai-generated content detectors: Boon or bane for scientific writing. *Indian Journal of Science and Technology*, 16(39):3435–3439, 2023.

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

Mariëlle Leijten and Luuk Van Waes. Keystroke logging in writing research. *Writ. Commun.*, 30(3):358–392, July 2013.

Daniel Lemire. Will ai flood us with irrelevant papers? *Communications of the ACM*, 67(9): 9–9, 2024.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024.

Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. The widespread adoption of large language model-assisted writing across society. *Patterns*, 6(12), 2025a.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, pp. 1–11, 2025b.

Daniel J. Liebling, Malcolm Kane, Madeleine Grunde-Mclaughlin, Ian J. Lang, Subhashini Venugopalan, and Michael P. Brenner. Towards ai-assisted academic writing. *ArXiv*, abs/2503.13771, 2025. URL https://api.semanticscholar.org/CorpusID:277104179.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

David M Markowitz. From complexity to clarity: How ai enhances perceptions of scientists and the public's understanding of science. *PNAS nexus*, 3(9):pgae387, 2024.

Kentaro Matsui. Delving into pubmed records: Some terms in medical writing have drastically changed after the arrival of chatgpt. *medRxiv*, pp. 2024–05, 2024.

David McMinn, Tom Grant, Laura DeFord-Watts, Veronica Porkess, Margarita Lens, Christopher Rapier, Wilson Q Joe, Timothy A Becker, and Walter Bender. Using artificial intelligence to expedite and enhance plain language summary abstract writing of scientific content. *JAMIA open*, 8(2):ooaf023, 2025.

Aleksandra Miletić, Christophe Benzitoun, Georgeta Cislaru, and Santiago Herrera-Yanez. Pro-text: An annotated corpus of keystroke logs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1732–1739, 2022.

Nature. Tools such as chatgpt threaten transparent science; here are our ground rules for their use. *Nature*, 613(7945):612, 2023. doi: 10.1038/d41586-023-00191-1. URL https://www.nature.com/articles/d41586-023-00191-1.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Ivan Oransky and Adam Marcus. Papers and peer reviews with evidence of chatgpt writing. *Retraction Watch*, 2024. URL https://retractionwatch.com/papers-and-peer-reviews-with-evidence-of-chatgpt-writing/.

Uwe Peters and Benjamin Chin-Yee. Generalization bias in large language model summarization of scientific research. *R. Soc. Open Sci.*, 12(4):241776, April 2025.

Mohi Reza, Jeb Thomas-Mitchell, Peter Dushniku, Nathan Laundry, Joseph Jay Williams, and Anastasia Kuzminykh. Co-writing with ai, on human terms: Aligning research with user demands across the writing process. *arXiv preprint arXiv:2504.12488*, 2025.

Paria Sarzaeim, Arya Doshi, and Qusay Mahmoud. A framework for detecting ai-generated text in research publications. In *Proceedings of the International Conference on Advanced Technologies*, volume 11, pp. 121–127, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

John Swales. *Cambridge applied linguistics: Genre analysis: English in academic and research settings*. Cambridge University Press, Cambridge, England, November 1990.

John M Swales. Genre analysis: English in academic and research settings. cambridge: Cambridge university press, selected 45–47, 52–60. In *The Discourse Studies Reader: Main currents in theory and analysis*, pp. 306–316. John Benjamins Publishing Company, 2014.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158, 2023.

Yu Tian, Scott Crossley, and Luuk Van Waes. The klicke corpus: Keystroke logging in compositions for knowledge evaluation. *Journal of Writing Research*, 17(1):23–60, 2025.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, et al. Logging keystrokes in writing by english learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10725–10746, 2024.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. PaperRobot: Incremental draft generation of scientific ideas. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1980–1991, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1191. URL https://aclanthology.org/P19-1191/.

Joseph M. Williams and Joseph Bizup. *Style: Lessons in Clarity and Grace*. Pearson, Boston, 11th edition, 2014.

A Limitations

Limited sample. The abstracts inherited from (Hazra et al., 2026) are limited to the domain of computer science. Other domains, such as math, physics, or the humanities, may invite differing editing behaviors due to differing underlying text features. Additionally, abstracts are only a small portion of all scientific writing. Understanding how humans revise longer-form scientific writing remains an interesting avenue for future work. However, our analytical framework generalizes to other domains and forms of scientific writing.

Expert editors. We see more interesting results with the editors who are best aligned with the task. Our current sample hosts a rather small pool of experts, and an in-depth exploration with an expert and senior author can be insightful.

B Reproducibility statement

We plan to release the full analysis code, data, and preprocessing scripts upon publication.

C Language model usage disclosure

We leveraged AutoDiscovery (Agarwal et al., 2025a), an LM-based data-driven discovery tool, to perform an initial exploration of our data. While we did not use any specific result from the output, it generally informed our analytic directions. We use LMs (e.g., Grammarly) to check grammatical and structural mistakes.

D Additional details on methods

Additional results relating to Tables 4 to 7.

E Additional details on results

Additional results relating to Tables 8 to 13.

F Additional details on Language Models as editing assistants

Additional results relating to Tables 14 to 19.

Cluster	Log pause	Entry pause	Edits	No Nav	Back	Fwd	Insert	Delete	Sub	Word	Phrase	Sent+
Phrasal expanding	5.918	120.189	45.489	0.986	0.003	0.011	1.000	0.000	0.000	0.000	1.000	0.000
Phrasal substitute	6.084	104.342	16.399	0.982	0.005	0.013	0.000	0.000	1.000	0.000	1.000	0.000
Phrasal prune	6.340	99.230	10.572	0.953	0.014	0.033	0.000	1.000	0.000	0.000	1.000	0.000
Phrasal replace	6.362	121.257	21.733	0.972	0.013	0.015	0.507	0.493	0.000	0.000	1.000	0.000
Sentence+ restructure	6.281	109.023	106.813	0.958	0.015	0.027	0.274	0.439	0.287	0.000	0.000	1.000
Forward word replace	3.784	165.563	4.901	0.884	0.000	0.116	0.370	0.230	0.401	1.000	0.000	0.000
Sentence+ expanding	5.978	142.310	203.030	0.992	0.000	0.008	1.000	0.000	0.000	0.000	0.000	1.000
Forward word prune	0.000	104.514	1.000	0.845	0.000	0.155	0.000	1.000	0.000	1.000	0.000	0.000
Cross-scope restructure	7.180	65.715	40.402	0.869	0.057	0.074	0.379	0.397	0.224	0.129	0.381	0.490
Backward span prune	0.000	185.090	1.000	0.762	0.238	0.000	0.282	0.576	0.142	0.000	0.475	0.525
Phrasal rewrite	5.972	117.693	53.936	0.992	0.002	0.006	0.319	0.331	0.351	0.000	1.000	0.000
Sentence+ elaborate	6.419	120.703	125.610	0.949	0.025	0.026	0.632	0.177	0.191	0.000	0.197	0.803
Forward span prune	0.000	195.389	1.000	0.600	0.000	0.400	0.209	0.663	0.128	0.000	0.452	0.548
Phrasal insert+substitute	6.141	74.764	38.512	0.984	0.006	0.010	0.498	0.000	0.502	0.000	1.000	0.000
Backward word prune	0.963	78.386	1.353	0.768	0.232	0.000	0.132	0.794	0.074	1.000	0.000	0.000
Phrasal condense	6.082	140.812	22.000	0.984	0.005	0.011	0.000	0.497	0.503	0.000	1.000	0.000

Table 4: Cluster means across pause/activity, navigation, action, and scope features.

k	Covariance	BIC	Mean max posterior	Min cluster n	Smallest cluster share	Converged
2	full	22779.49	1.0000	1572	0.3989	True
3	full	8908.56	0.9991	620	0.1573	True
4	full	-2338.91	0.9998	275	0.0698	True
5	full	-8593.44	0.9983	275	0.0698	True
6	full	-9932.09	0.9955	317	0.0804	True
7	full	-18591.95	0.9987	119	0.0302	True
8	full	-20675.32	0.9981	144	0.0365	True
9	full	-23794.34	0.9980	94	0.0239	True
10	full	-27518.06	0.9982	94	0.0239	True
11	full	-30685.60	0.9991	35	0.0089	True
12	full	-31232.85	0.9947	73	0.0185	True
13	full	-34489.82	0.9985	75	0.0190	True
14	full	-36391.13	0.9984	35	0.0089	True
15	full	-35873.42	0.9952	15	0.0038	True
16	**full**	**-43035.56**	**0.9995**	**40**	**0.0101**	**True**
17	full	-40712.56	0.9974	14	0.0036	True
18	full	-40137.37	0.9960	13	0.0033	True
19	full	-42742.12	0.9884	11	0.0028	True

Table 5: Gaussian mixture model selection summary. Cluster size k=16 chosen.

Statistic	ARI vs. reference	Mean max posterior	Min cluster n
count	25.0000	25.0000	25.0000
mean	0.8720	0.9960	19.4800
std	0.0580	0.0030	8.5010
min	0.7450	0.9890	8.0000
255075max	0.9450	1.0000	37.0000

Table 6: Gaussian mixture model stability summary across repetitions with chosen cluster size 16.

Dimension	Metric	Meaning
Agency	lexical-verb ratio	Proportion of verbs that carry semantic weight to all finite verbs.
Agency	passive-voice share	The number of sentences that use passive constructions.
Agency	character-as-subject	How often grammatical subject of a sentence refers to a real actor or "character."
Agency	agentless-passive	passive clauses omitting a by-phrase agent.
Agency	metadiscourse we-density	explicit use of first-person pronouns (*I*/*We*) in sentences that introduce claims.
Economy	nominalization density	Measures how often verbs and adjectives are turned into abstract nouns.
Economy	abstract-congestion	Detects "congested" sentences where an abstract-noun subject is accompanied by additional nominalizations.
Economy	compound-noun rate	Captures the number of long noun strings (three or more consecutive nouns).
Economy	prep. phrase density	Counts *of*/*in*/*for* relations per 100 words.
Economy	filler-word density	Counts filler words like *very*, *actually*, *basically*.
Structure	subject-onset distance	Measures how many tokens occur before the main grammatical subject.
Structure	initial-momentum	How soon the main verb appears relative to the sentence start.
Structure	subject-verb gap	The number of words between a subject and its verb.
Structure	verb-object gap	Tokens separate a verb from its direct object.
Structure	post-verb density	Count of subordinate clauses or heavy additions that follow the main verb.
Coherence	cohesion-chain strength	How well consecutive sentences connect via repeated or echoed content words.
Coherence	old-to-new ordering	The relative position of previously unseen words in each sentence.
Coherence	topic-continuity	Whether successive sentences share the same grammatical subject.
Coherence	stress-position rate	How often the stress position is occupied by a content word.
Framing	pivot-explicitness	Presence of adversative markers (*but*, *however*, *yet*) indicating a shift.
Framing	condition-and-cost	Whether problem statements articulate both the condition and the consequence.
Framing	claim placement	Position where the main contribution or result (e.g., "we show") appears.

Table 7: Text property dimensions and metric definitions. Agency, Economy, and Structure can be interpreted at a sentence level, whereas Coherence and Framing are interpretable only across the broader text. Metrics definitions are derived from (Hazra et al., 2026). Return to subsection 2.3.

Δ Dimension	β_{source}	p_{source}	R^2_{source}	$\beta_{\text{pre\_edit}}$	$p_{\text{pre\_edit}}$	$R^2_{\text{pre\_edit}}$	R^2_{full}
Agency	-0.0463	0.1493	0.0006	-0.3362	0.0000	0.0997	0.1170
Economy	-0.0156	0.6211	0.0001	-0.3660	0.0000	0.1154	0.1275
Structure	0.0210	0.5109	0.0001	-0.3912	0.0000	0.1395	0.1530
Coherence	-0.1268	0.0001	0.0040	-0.3902	0.0000	0.0849	0.0959
Framing	-0.0002	0.9944	0.0000	-0.2286	0.0000	0.0310	0.0422

Table 8: Full results from Table 3, including partial R^2, the R^2 from the entire model, and the p-values.

Education	Contrast	Behavior	Con. left	Con. right	Difference	p_{FDR}
College	AI-withInfo - AI-noInfo	Expansion	0.202	0.193	+0.009	0.9540
College	AI-withInfo - AI-noInfo	Pruning	0.161	0.244	-0.083	0.0329
College	AI-withInfo - AI-noInfo	Restructuring	0.305	0.291	+0.014	0.9540
College	AI-withInfo - AI-noInfo	Substitution	0.332	0.272	+0.060	0.4709
College	H-withInfo - H-noInfo	Expansion	0.251	0.254	-0.003	0.9748
College	H-withInfo - H-noInfo	Pruning	0.170	0.198	-0.027	0.7171
College	H-withInfo - H-noInfo	Restructuring	0.306	0.295	+0.010	0.9540
College	H-withInfo - H-noInfo	Substitution	0.273	0.253	+0.020	0.9540
College	AI-noInfo - H-noInfo	Expansion	0.193	0.254	-0.061	0.3757
College	AI-noInfo - H-noInfo	Pruning	0.244	0.198	+0.046	0.4709
College	AI-noInfo - H-noInfo	Restructuring	0.291	0.295	-0.004	0.9748
College	AI-noInfo - H-noInfo	Substitution	0.272	0.253	+0.019	0.9540
College	AI-withInfo - H-withInfo	Expansion	0.202	0.251	-0.049	0.4709
College	AI-withInfo - H-withInfo	Pruning	0.161	0.170	-0.009	0.9540
College	AI-withInfo - H-withInfo	Restructuring	0.305	0.306	-0.001	0.9816
College	AI-withInfo - H-withInfo	Substitution	0.332	0.273	+0.059	0.4709
Masters+	AI-withInfo - AI-noInfo	Expansion	0.190	0.248	-0.059	0.1169
Masters+	AI-withInfo - AI-noInfo	Pruning	0.177	0.146	+0.032	0.2928
Masters+	AI-withInfo - AI-noInfo	Restructuring	0.220	0.381	-0.161	0.0004
Masters+	AI-withInfo - AI-noInfo	Substitution	0.413	0.225	+0.188	0.0005
Masters+	H-withInfo - H-noInfo	Expansion	0.269	0.175	+0.094	0.0359
Masters+	H-withInfo - H-noInfo	Pruning	0.189	0.218	-0.029	0.4170
Masters+	H-withInfo - H-noInfo	Restructuring	0.362	0.251	+0.111	0.0377
Masters+	H-withInfo - H-noInfo	Substitution	0.180	0.356	-0.177	0.0034
Masters+	AI-noInfo - H-noInfo	Expansion	0.248	0.175	+0.073	0.0724
Masters+	AI-noInfo - H-noInfo	Pruning	0.146	0.218	-0.072	0.0335
Masters+	AI-noInfo - H-noInfo	Restructuring	0.381	0.251	+0.130	0.0115
Masters+	AI-noInfo - H-noInfo	Substitution	0.225	0.356	-0.132	0.0197
Masters+	AI-withInfo - H-withInfo	Expansion	0.190	0.269	-0.080	0.0514
Masters+	AI-withInfo - H-withInfo	Pruning	0.177	0.189	-0.011	0.7159
Masters+	AI-withInfo - H-withInfo	Restructuring	0.220	0.362	-0.142	0.0038
Masters+	AI-withInfo - H-withInfo	Substitution	0.413	0.180	+0.233	0.0001

Table 9: Disclosure effects on burst behavior probabilities. The full results corresponding to Figure 2. ″H″ corresponds to human.

Dimension	Family	Q1	Q4	Q1-Q4	Q1 share	Q1 n
Agency	Substitution	0.47 [0.36, 0.57]	-0.43 [-0.53, -0.31]	0.90 [0.74, 1.03]	44.2%	405
Agency	Pruning	0.42 [0.25, 0.58]	-0.33 [-0.46, -0.20]	0.75 [0.54, 0.95]	25.0%	229
Agency	Expansion	0.33 [0.15, 0.50]	-0.25 [-0.38, -0.13]	0.58 [0.36, 0.78]	26.7%	245
Agency	Restructuring	0.32 [0.00, 0.60]	-0.11 [-0.35, 0.11]	0.43 [0.03, 0.85]	4.1%	38
Economy	Pruning	0.43 [0.24, 0.60]	-0.27 [-0.39, -0.17]	0.69 [0.48, 0.91]	25.0%	229
Economy	Expansion	0.40 [0.19, 0.62]	-0.31 [-0.43, -0.22]	0.72 [0.49, 0.97]	25.5%	234
Economy	Substitution	0.31 [0.21, 0.41]	-0.37 [-0.47, -0.28]	0.67 [0.53, 0.82]	43.5%	399
Economy	Restructuring	-0.12 [-0.37, 0.09]	-0.12 [-0.23, -0.04]	0.01 [-0.28, 0.26]	6.0%	55
Structure	Pruning	0.55 [0.37, 0.70]	-0.35 [-0.47, -0.26]	0.90 [0.71, 1.08]	28.4%	260
Structure	Substitution	0.49 [0.36, 0.62]	-0.30 [-0.36, -0.24]	0.79 [0.66, 0.93]	41.5%	381
Structure	Expansion	0.24 [0.09, 0.40]	-0.43 [-0.55, -0.30]	0.67 [0.48, 0.85]	25.0%	229
Structure	Restructuring	0.18 [-0.38, 0.64]	-0.24 [-0.45, -0.06]	0.42 [-0.17, 0.91]	5.1%	47
Coherence	Substitution	0.48 [0.34, 0.65]	-0.33 [-0.45, -0.21]	0.81 [0.62, 1.02]	42.0%	385
Coherence	Expansion	0.17 [0.04, 0.29]	-0.18 [-0.30, -0.07]	0.36 [0.18, 0.51]	28.4%	260
Coherence	Pruning	0.14 [0.06, 0.27]	-0.11 [-0.22, -0.03]	0.26 [0.14, 0.43]	25.5%	234
Coherence	Restructuring	-0.01 [-0.27, 0.16]	-0.06 [-0.24, 0.07]	0.05 [-0.24, 0.28]	4.1%	38
Framing	Substitution	0.31 [0.21, 0.44]	-0.38 [-0.57, -0.23]	0.70 [0.53, 0.92]	43.3%	400
Framing	Expansion	0.12 [0.04, 0.21]	-0.10 [-0.23, 0.03]	0.21 [0.06, 0.36]	25.9%	239
Framing	Pruning	0.07 [0.03, 0.13]	-0.18 [-0.35, -0.05]	0.25 [0.12, 0.43]	25.8%	238
Framing	Restructuring	0.05 [0.01, 0.18]	-0.23 [-0.63, 0.07]	0.28 [-0.03, 0.72]	5.0%	46

Table 10: Behavior efficacy at the first and last quartile. Mean and 95% confidence interval are reported for each quartile.

Dimension	AI $M(SD)$	Human $M(SD)$	d_z	$t(44)$	p
Agency	0.10 (0.36)	-0.14 (0.46)	0.45	2.99	.008*
Economy	-0.06 (0.29)	-0.00 (0.32)	-0.20	-1.32	.242
Structure	0.13 (0.20)	-0.03 (0.30)	0.54	3.62	.004*
Coherence	-0.08 (0.41)	0.17 (0.43)	-0.47	-3.15	.007*
Framing	0.12 (0.54)	0.15 (0.51)	-0.04	-0.26	.796

Table 11: Human vs AI initial linguistic properties.

Dimension	AI $M(SD)$	Human $M(SD)$	d_z	$t(44)$	p
Agency	0.01 (0.31)	-0.19 (0.38)	0.46	3.07	.018*
Economy	-0.02 (0.25)	0.02 (0.27)	-0.17	-1.17	.309
Structure	0.11 (0.18)	0.00 (0.24)	0.41	2.76	.019*
Coherence	-0.05 (0.31)	0.13 (0.42)	-0.39	-2.63	.019*
Framing	0.13 (0.50)	0.11 (0.39)	0.03	0.18	.855

Table 12: Human vs AI post-edit linguistic properties.

Dimension	Source	Pre M (SD)	Post M (SD)	Mean Change	d_z	$t(44)$	p_{FDR}
Agency	AI	0.10 (0.36)	-0.01 (0.30)	-0.11	-0.60	-4.04	.001*
Agency	Human	-0.14 (0.46)	-0.16 (0.38)	-0.02	-0.10	-0.70	.610
Economy	AI	-0.06 (0.29)	-0.02 (0.25)	0.04	0.46	3.07	.009*
Economy	Human	-0.00 (0.32)	0.02 (0.26)	0.02	0.24	1.61	.303
Structure	AI	0.13 (0.20)	0.10 (0.17)	-0.03	-0.32	-2.16	.060
Structure	Human	-0.03 (0.30)	0.00 (0.24)	0.03	0.24	1.58	.303
Coherence	AI	-0.08 (0.41)	-0.04 (0.30)	0.04	0.22	1.49	.178
Coherence	Human	0.17 (0.43)	0.16 (0.37)	-0.01	-0.06	-0.40	.695
Framing	AI	0.12 (0.54)	0.14 (0.48)	0.02	0.14	0.95	.346
Framing	Human	0.15 (0.51)	0.11 (0.37)	-0.04	-0.16	-1.09	.470

Table 13: Original and edited abstract differences according to a paired t-test.

Dimension	Editor	Mean Diff	Cohen's d_z	p_{FDR}
Agency	Human	-0.001	-0.01	0.918
Agency	LM: Unguided	0.056	0.15	0.137
Agency	LM: Rubric	0.215	0.50	$< .001$
Economy	Human	0.005	0.05	0.320
Economy	LM: Unguided	0.053	0.31	$< .001$
Economy	LM: Rubric	0.111	0.53	$< .001$
Structure	Human	-0.003	-0.03	0.633
Structure	LM: Unguided	0.017	0.08	0.428
Structure	LM: Rubric	0.086	0.37	$< .001$
Coherence	Human	0.018	0.06	0.223
Coherence	LM: Unguided	-0.048	-0.13	0.167
Coherence	LM: Rubric	0.002	0.00	0.952
Framing	Human	0.006	0.03	0.618
Framing	LM: Unguided	0.127	0.22	0.016
Framing	LM: Rubric	0.104	0.17	0.081

Table 14: LM as a zero-shot editor paired t-test.

Dimension	Editor 1	Editor 2	Mean Diff	β	p
Agency	LM: Rubric	Human	0.216	0.836	$< .001$
Economy	LM: Rubric	Human	0.107	0.816	$< .001$
Structure	LM: Rubric	Human	0.094	0.618	$< .001$
Coherence	LM: Rubric	Human	-0.027	-0.088	.811
Framing	LM: Rubric	Human	0.099	0.286	.021
Agency	LM: Rubric	LM: Unguided	0.158	0.392	$< .001$
Economy	LM: Rubric	LM: Unguided	0.055	0.276	.007
Structure	LM: Rubric	LM: Unguided	0.069	0.308	.003
Coherence	LM: Rubric	LM: Unguided	0.046	0.122	.154
Framing	LM: Rubric	LM: Unguided	-0.018	-0.031	.686

Table 15: LM as a zero-shot editor comparative results from a Welch's t-test.

Dimension	Editor	Percentile	Pre mean	Post mean	Δ mean	t	p_{FDR}
Agency	LM: Rubric	Below median	25.69	60.15	34.46	11.48	$< .001$
Agency	LM: Rubric	Above median	74.94	71.28	-3.65	-1.65	0.114
Economy	LM: Rubric	Below median	24.44	42.49	18.05	7.51	$< .001$
Economy	LM: Rubric	Above median	74.04	78.81	4.77	2.11	0.048
Structure	LM: Rubric	Below median	25.79	45.26	19.48	6.53	$< .001$
Structure	LM: Rubric	Above median	75.94	73.29	-2.66	-1.03	0.307
Coherence	LM: Rubric	Below median	25.18	35.65	10.47	3.82	$< .001$
Coherence	LM: Rubric	Above median	73.78	59.85	-13.92	-5.28	$< .001$
Framing	LM: Rubric	Below median	33.01	50.14	17.13	6.70	$< .001$
Framing	LM: Rubric	Above median	76.02	63.96	-12.06	-4.20	$< .001$

Table 16: Paired t-test statistics for pre and post edits, segmented into high and low percentiles. Corresponds to Figure 4a.

Dimension	Initial group	Rubric Δ	Scoped Δ	Δ improvement	95% CI	p_{FDR}
Agency	Low initial	12.89	6.34	-6.55	[-9.95, -3.14]	0.002
Agency	High initial	2.39	2.00	-0.39	[-2.44, 1.66]	0.788
Economy	Low initial	9.36	4.28	-5.08	[-8.05, -2.10]	0.004
Economy	High initial	2.81	2.87	0.06	[-2.38, 2.50]	0.962
Structure	Low initial	11.40	5.67	-5.73	[-9.36, -2.09]	0.007
Structure	High initial	-2.55	-3.46	-0.92	[-3.89, 2.06]	0.682
Coherence	Low initial	5.33	3.50	-1.83	[-4.70, 1.04]	0.351
Coherence	High initial	-5.30	-2.79	2.51	[-0.60, 5.62]	0.228
Framing	Low initial	6.68	2.86	-3.82	[-6.57, -1.08]	0.016
Framing	High initial	-2.49	-1.61	0.88	[-1.62, 3.37]	0.682

Table 17: Direct comparison of percentile improvement under rubric-guided versus scoped prompting. Negative values in Δ improvement indicate smaller gains under scoped prompting.

Experiment	Scope	Human	LM	ΔLM–Human	d_z	p_{FDR}
Unguided	Word	0.10	0.00	-0.10	-0.33	$< .001$
Rubric	Word	0.10	0.00	-0.10	-0.33	$< .001$
Scoped	Word	0.10	0.01	-0.09	-0.28	$< .001$
Unguided	Phrase	0.53	0.14	-0.39	-0.68	$< .001$
Rubric	Phrase	0.53	0.11	-0.42	-0.78	$< .001$
Scoped	Phrase	0.53	0.46	-0.07	-0.14	$< .001$
Unguided	Sentence+	0.37	0.86	0.49	0.93	$< .001$
Rubric	Sentence+	0.37	0.89	0.52	1.03	$< .001$
Scoped	Sentence+	0.37	0.53	0.16	0.39	$< .001$

Table 18: Human and LM burst scope distributions. Positive values indicate the LM uses the corresponding scope more often than humans.

Editor	Dimension	Mean change	d_z	p_{FDR}
Human	Agency	0.01	0.02	0.866
Human	Economy	0.06	0.11	0.207
Human	Structure	0.00	0.01	0.930
Human	Coherence	0.01	0.02	0.866
Human	Framing	-0.01	-0.05	0.722
Unguided	Agency	0.01	0.07	0.247
Unguided	Economy	0.02	0.13	0.020
Unguided	Structure	0.03	0.15	0.007
Unguided	Coherence	-0.01	-0.03	0.772
Unguided	Framing	0.04	0.19	$< .001$
Rubric	Agency	0.12	0.48	$< .001$
Rubric	Economy	0.06	0.38	$< .001$
Rubric	Structure	0.05	0.27	$< .001$
Rubric	Coherence	0.00	0.01	0.866
Rubric	Framing	0.05	0.14	0.0109

Table 19: Paired t-tests of burst-level linguistic scores for human and LM editors.

F.1 Prompts

Unguided
Goal
Your goal is to improve the quality of the scientific abstract.
What is an abstract?
An abstract is a short summary of completed research. It is intended to describe the work without going into great detail. Abstracts should be self-contained and concise, explaining your work as briefly and clearly as possible. An abstract should be able to stand independently from the research paper and still tell the reader something significant. The most important function of an abstract is to help a reader decide if he or she is interested in reading your entire publication.
An effective abstract will contain several key features:
Motivation or problem statement: Why is the research/argument important? What practical, scientific, theoretical or artistic gap is the project filling?

Methods/procedure/approach: What did the researcher actually do to get your results? (e.g. analyzed 3 novels, completed a series of 5 oil paintings, interviewed 17 students)

Results/findings/product: After completing the above procedure, what did the researcher learn/invent/create?

Conclusion/implications: What are the larger implications of the findings, especially for the problem/gap identified previously? Why is this research valuable?

Keep the abstract short: A general rule of abstract length is 150-200 words.

Do not add any new information: If something doesn't appear in the input, then don't put it in the abstract. An abstract is supposed to convey scientific findings, so they have to be precise and factual. Please don't embellish any results or findings.

Original Abstract Below is the original abstract before editing.

{abstract}

Output Format
Please think step by step to complete your goal, then provide the abstract in full. Your response should be in the following JSON format:

```
{
  "reasoning": "the reasoning here",
  "edited_abstract": "the full edited abstract here"
}
```

Rubric
Goal
Your goal is to improve the quality of the scientific abstract.
What is an abstract?
An abstract is a short summary of completed research. It is intended to describe the work without going into great detail. Abstracts should be self-contained and concise, explaining your work as briefly and clearly as possible. An abstract should be able to stand independently from the research paper and still tell the reader something significant. The most important function of an abstract is to help a reader decide if he or she is interested in reading your entire publication.
An effective abstract will contain several key features:
Motivation or problem statement: Why is the research/argument important? What practical, scientific, theoretical or artistic gap is the project filling?
Methods/procedure/approach: What did the researcher actually do to get your results? (e.g., analyzed 3 novels, completed a series of 5 oil paintings, interviewed 17 students)
Results/findings/product: After completing the above procedure, what did the researcher learn/invent/create?
Conclusion/implications: What are the larger implications of the findings, especially for the problem/gap identified previously? Why is this research valuable?
Keep the abstract short: A general rule of abstract length is 150–200 words.

Do not add any new information: If something doesn't appear in the input, then don't put it in the abstract. An abstract is supposed to convey scientific findings, so they have to be precise and factual. Please don't embellish any results or findings.

Rubric for Writing
Goal
Your goal is to improve the quality of the scientific abstract across five dimensions: Agency, Economy, Structure, Coherence, and Framing.
Metric Rubric
Each dimension is scored on a standardized scale relative to a reference distribution of typical pre-edit scientific writing:

- 0 = typical
- positive = better than typical
- negative = worse than typical

Agency
Definition: clarifies the role between the actors (subjects) and their actions (verbs). Clear writing makes it easy to identify who is doing what.

- lexical-verb ratio: proportion of semantically meaningful verbs relative to all finite verbs.
- passive-voice share: the number of sentences that use passive constructions.
- character-as-subject: how often the grammatical subject refers to a real actor.
- agentless-passive: how often passive clauses omit a by-phrase agent.
- metadiscourse we-density: how often writers use first-person pronouns to introduce claims.

Economy
Definition: quantifies how efficiently a text conveys complex information without unnecessary compression.

- nominalization density: Measures how often verbs and adjectives are turned into abstract nouns.
- abstract-congestion: Detects "congested" sentences where an abstract-noun subject is accompanied by additional nominalizations.
- compound-noun rate: Captures the number of long noun strings (three or more consecutive nouns).
- prep. phrase density: Counts of in for relations per 100 words.
- filler-word density: Counts empty hedges and intensifiers like very, actually, basically.

Structure
Definition: describes how the arrangement of words and phrases leads to the primary grammatical relationships.

- subject-onset distance: Measures how many tokens occur before the main grammatical subject.
- initial-momentum: Metric gauges how soon the main verb appears relative to the sentence start.
- subject-verb gap: Measures the number of words between a subject and its verb.
- verb-object gap: Captures how many tokens separate a verb from its direct object.
- post-verb density: Counts subordinate clauses or heavy additions that follow the main verb.

Coherence
Definition: measures how well a text introduces new information and connects threads.

- cohesion-chain strength: Measures how well consecutive sentences connect via repeated or echoed content words.
- old-to-new ordering: Calculates the relative position of previously unseen words in each sentence.
- topic-continuity: Checks whether successive sentences share the same grammatical subject.

- stress-position rate: Measures how often the stress position is occupied by a content-bearing word.

Framing
Definition: tracks the placement of context and claims.

- pivot-explicitness: Checks for clear adversative markers (but, however, yet) signaling a shift.
- condition-and-cost: Examines whether problem statements articulate both the condition and the consequence.
- claim placement: Captures where the main contribution or result (e.g., "we show") appears.

Original Abstract

Output Format
Please think step by step to complete your goal, then provide the abstract in full. Your response should be in the following JSON format:

```
  {
 "reasoning": "<the reasoning here>",
 "edited_abstract": "<the full edited abstract here>"
}
```