Can AI writing be salvaged? Mitigating Idiosyncrasies and Improving Human-AI Alignment in the Writing Process through Edits

Tuhin Chakrabarty
tchakrabarty@salesforce.com
Salesforce AI Research
USA

Philippe Laban
Salesforce AI Research
USA

Chien-Sheng Wu
Salesforce AI Research
USA

Abstract

LLM-based applications are helping people write, and LLM-generated text is making its way into social media, journalism, and our classrooms. However, the differences between LLM-generated and human-written text remain unclear. To explore this, we hired professional writers to edit paragraphs in several creative domains. We first found these writers agree on undesirable idiosyncrasies in LLM-generated text, formalizing it into a seven-category taxonomy (e.g. clichés, unnecessary exposition). Second, we curated the LAMP corpus: 1,057 LLM-generated paragraphs edited by professional writers according to our taxonomy. Analysis of LAMP reveals that none of the LLMs used in our study (GPT4o, Claude-3.5-Sonnet, Llama-3.1-70b) outperform each other in terms of writing quality, revealing common limitations across model families. Third, building on existing work in automatic editing we evaluated methods to improve LLM-generated text. A large-scale preference annotation confirms that although experts largely prefer text edited by other experts, automatic editing methods show promise in improving alignment between LLM-generated and human-written text.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Natural language generation.*

Keywords

Human-AI collaboration, Large Language Models, Design Methods, Text Editing, Natural Language Generation, Evaluation, Writing Assistance, Generative AI, Homogenization, Alignment, Behavioral Science

ACM Reference Format:

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025. Can AI writing be salvaged? Mitigating Idiosyncrasies and Improving Human-AI Alignment in the Writing Process through Edits. In . ACM, New York, NY, USA, 33 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Artificial Intelligence (AI) has the potential to revolutionize how we write, communicate, and express ideas [68]. Recent studies have demonstrated the potential of large language models (LLMs) in assisting with various writing tasks, including argumentative [24, 65], scientific [43], and creative writing [16, 51, 77, 78, 117]. Aligning LLMs with human preferences [86] has enabled their transformation into user-friendly tools for non-technical users, such as Google's WorkSpace Labs, Grammarly, and Sudowrite. However, to truly benefit society, AI writing assistants must enhance human creativity and expression rather than homogenize content or diminish linguistic diversity [39, 52]

While LLM-based writing assistants have the potential to improve writing quality and increase author productivity, they also introduce an algorithmic monoculture [58]. Padmakumar and He [87] and Anderson et al. [3] discuss how writing with LLMs unintentionally reduces content diversity, leading to homogenization. This homogenization occurs not just at the semantic level but also at syntactic (structural), lexical [59], and stylistic levels [103]. For instance, prior work from Chakrabarty et al. [16], Ippolito et al. [51] has shown how LLM-generated text is often hackneyed and rife with clichés while failing to demonstrate rhetorical complexity and often revealing the subtext —a phenomenon known as "telling instead of showing" [15]. Additionally, LLM-generated texts are typically full of redundant exposition, overwrought metaphors, and florid descriptions due to verbosity bias during preference labeling [101].

Current AI-assisted writing tools are powered by pre-trained language models that are refined through reinforcement learning from human feedback (RLHF) [121]. RLHF transforms human preferences into training data to guide language models toward desired outcomes. The most common type of feedback used with RLHF is binary preferences between pairs of examples sampled from one or more language models [19] (See Fig 1 Traditional Alignment). However, this approach has a drawback. The paired outputs may differ in numerous ways and could be equally flawed in containing idiosyncrasies. Asking an annotator to choose between two undesirable outputs does not improve alignment[1]. [14, 44]. We argue that alignment training needs to be aware of how desirable any individual response is, regardless of its preference relationship. Editing undesirable portions of a response can be seen as an effective mechanism for enhancing alignment (See Fig 1 Alignment via Edits). An LLM-generated response that has been edited typically contains

[1] In the current design of RLHF annotators are not allowed to not pick either

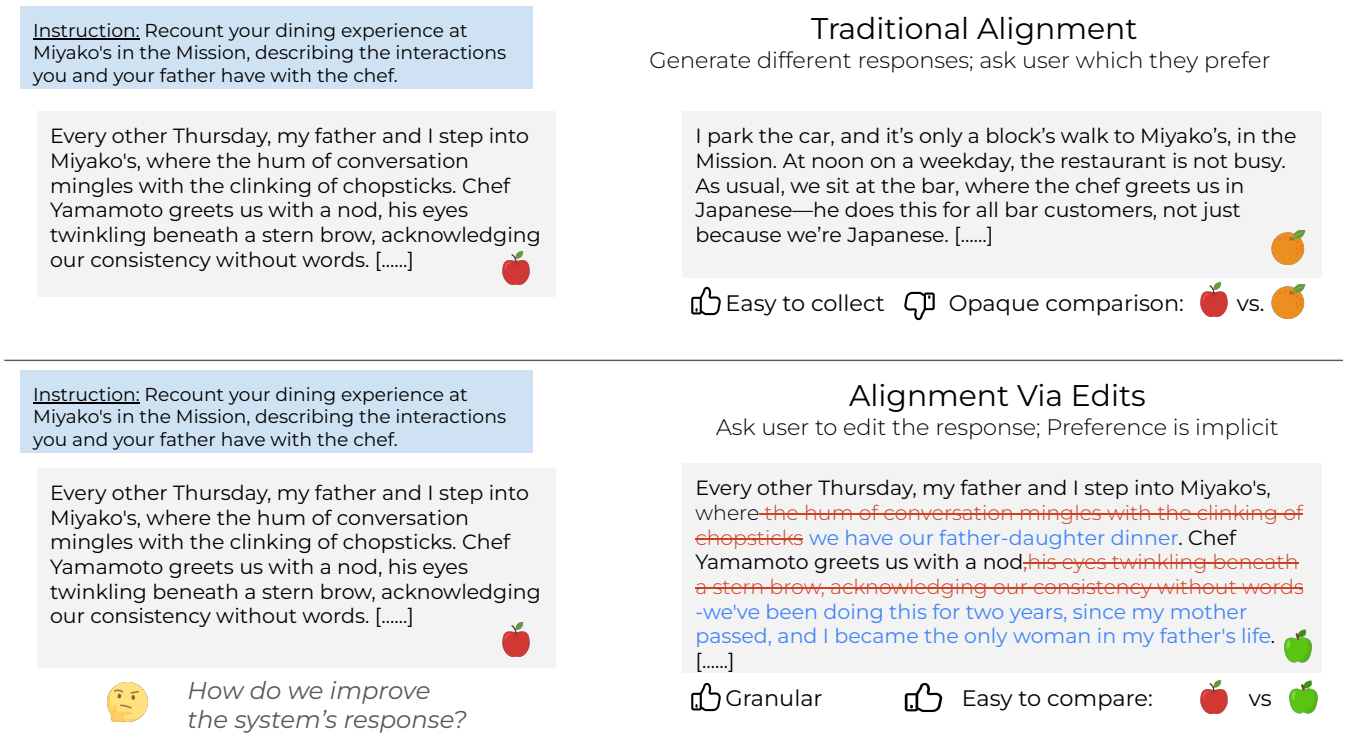


Figure 1: To align models to human preferences, human annotators are typically shown two responses and asked to choose the one they prefer. (i) The top portion of the Figure shows Traditional Alignment: it is often hard to compare two responses that differ widely. (ii) The bottom portion of the Figure shows Alignment via Edits where the original response is edited, allowing for a more granular comparison, with the edited version of the text naturally preferred over the original response.

fewer undesirable traits and can be paired with the original LLM-generated response for preference ranking (*edited* > *original*). The challenge, however, lies in consistently identifying and implementing edits that enhance quality while aligning with human preferences for effective writing. Prompting techniques that encourage the model to self-edit [72] have shown promise; however, they do not work for long-form or paragraph-level writing [88]. The primary reason for this is that LLMs do not inherently know what aspects of the writing need improvement, how extensively they should edit, or how to make changes that align with any given writer's expertise level and style.

To address these problems, we first create a comprehensive taxonomy of edit categories based on expert writing practices. We then recruit 18 writers to edit LLM-generated text using categories from our taxonomy. We define edits as changes that alter, replace, or refine specific phrases, clauses, or sentences within a larger text. We restrict our focus to generating text in literary fiction and creative non-fiction, as these genres challenge LLMs with their creativity, emotional nuance, and sophisticated language use. We focus on paragraph-level edits, as they balance granularity and scope, reducing costs and annotator fatigue. Paragraphs capture style and context better than sentences, enabling more cohesive improvements. Given LLMs' limitations in long-term discourse coherence, paragraph-level enhancements facilitate human-AI collaboration. This approach allows humans to guide the overall structure and flow, while AI handles the lower-level details [24]. Finally, while expert writer's edits are valuable for identifying idiosyncrasies, this approach may not scale for large text volumes. To address this, we design few-shot prompts that use writer's edits to identify problematic spans in LLM-generated paragraphs and suggest improvements. This method aims to enhance overall paragraph quality at scale. Our work adds to the existing rich literature [40, 53, 70, 116] on using completion edits to improve alignment in different tasks. To summarize our contributions:

- We propose a comprehensive edit taxonomy informed by expert writing practices that serve as a useful framework to identify and mitigate the idiosyncrasies in LLM-generated text.
- We release the LAMP (Language model Authored, Manually Polished) corpus containing 1057 <instructions, response> pairs grounded in real-world writing tasks such as Literary Fiction or Creative Non-Fiction. These responses originally generated by LLMs are further edited by 18 professional writers using the above-mentioned taxonomy, resulting in 8035 fine-grained edits (Section 4).
- We present a thorough analysis of the process of editing LLM-generated text, offering insights into how expert writers edit them, to what extent the edits differ in quantity, how the

distribution of edit categories varies across text generated by different model families, and whether LLM generated text contain any specific stylistic idiosyncrasies (Sections 4 and 5.3).
- Building on prior work Hayes et al. [47], Scardamalia [102] we conduct an empirical investigation that tests if LLMs can automatically detect and rewrite their own idiosyncrasies. Our statistically significant results show an encouraging preference trend $Writer - edited > LLM - edited > LLM - generated$ suggesting that edits improve human-AI alignment in the writing process.
- Finally, we discuss how LLM edits can both mimic and differ from edits provided by professional writers, and what future LLM-based writing support tools can do to improve the co-writing experience.

Our code, data, and experimental setup is available at [2].

2 RELATED WORK

2.1 Text Editing in HCI

Text editing is the process of modifying written content using specialized software. HCI research on text editing aims to improve digital writing tools' efficiency and usability. Word processors have long allowed flexible editing functions [92]. Systems like Soylent [7], MicroWriter [110], and WearWrite [80] developed interfaces for crowd-based editing, focusing on task breakdown, cost management, and minimizing delays. Robertson and Black [99] proposes a goal-fate analysis model for text editing behavior, supported by data showing distinct plan units in editing tasks, with the potential for intention-based user assistance. Tyler et al. [113] investigated text editing skill progression and effective training methods. Rosson [100] examined the impact of experience on editing behavior, questioning if users naturally develop optimal strategies or plateau, noting that experienced users tend to develop more efficient editing heuristics than novices. Reza et al. [98] present ABScribe, a novel interface that streamlines the process of generating and comparing writing variations using Large Language Models, addressing challenges in existing text-editing workflows and improving writer's efficiency and satisfaction. Zhou and Sterman [119] suggest that imperfect AI text suggestions can promote deeper engagement in rewriting, potentially preserving the writer's authenticity and creative ownership. Park and Lee [91] found that providing rationales for edits in collaborative writing was generally preferred by participants, despite no significant differences in survey results. This led to design recommendations for effective collaboration. Dang et al. [21] propose a text editor with continuously updated paragraph-wise summaries as margin annotations to help users plan, structure, and reflect on their writing process.Laban et al. [61] introduce InkSync, an LLM-based editing interface suggesting executable document edits. It uses a three-stage approach (Warn, Verify, Audit) to reduce factual errors and enhance editing accuracy, efficiency, and user experience compared to standard chat interfaces. In contrast to existing research, we focus on edits as a method to improve human-AI alignment in writing assistance. Our work characterizes the undesirable aspects of AI writing informed by expert consensus and designs an approach to mitigate these through text editing.

2.2 Text Editing in NLP

NLP research has explored various text editing tasks [18, 62, 93]. Adding to it the advent of Large Language Models has enabled AI-assisted writing tools [12, 50]. Faltings et al. [34] release the WikiDocEdits dataset and propose an interactive text generation setting in which a user interacts with the system by issuing commands to edit existing text. Raheja et al. [95] proposed an instruction-based editing system using fine-tuned language models. Shu et al. [104] developed strategies for cross-sentence rewriting and introduced the OpenRewriteEval benchmark. Reid and Neubig [97] modeled multi-step editing processes to better mimic human content creation and improve performance on various tasks. Kim et al. [55] presented a system that iteratively improves fluency, clarity, coherence, and style by detecting editable spans and their corresponding edit intents, then instructing a revision model to refine the text. Yang et al. [115] developed a taxonomy and classifier for Wikipedia edit intentions. Following them Du et al. [29] created a multi-domain corpus of revised text with annotated edit intentions. Unlike existing work, we create a resource for text editing that caters to challenging writing tasks (literary fiction and creative nonfiction). Our data consists of 8035 fine-grained edits that are annotated by creative writing experts and we further show how recent advances in few-shot learning can help models improve their own writing by learning from edits provided by the writers.

2.3 Issues in AI Writing

Prior work has highlighted several issues in AI-generated text. Chakrabarty et al. [15, 16], Ippolito et al. [51], Marco et al. [74], Mirowski et al. [78] show how LLM-generated text is often hackneyed and rife with clichés, lacks nuance, subtext, and rhetorical complexity. Recent work from Mirowski et al. [77] shows LLMs fail to act as good creativity support tools for comedy writing and mostly resort to producing bland and biased comedy tropes. They further highlight how existing moderation strategies used in safety filtering and instruction-tuned LLMs reinforce hegemonic viewpoints by erasing minority groups and their perspectives in writing. In summarizing short stories Subbiah et al. [108] demonstrate how LLMs struggle with specificity and interpretation of difficult subtext. In a similar vein, Tian et al. [112] shed light on how LLM-generated stories are homogeneously positive and lack tension. Compared to existing work we create a fine-grained taxonomy highlighting the issues in AI writing and further create a large-scale corpus to fuel research in this direction. We also develop automated methods to identify and mitigate issues in AI writing at scale.

2.4 Human AI alignment in Writing

Lee et al. [64] highlight how AI tools have transformed writing processes, establishing new criteria for future AI writing assistants. In a similar vein Li et al. [68] reveal that while users benefit from AI assistance in productivity and confidence, potential drawbacks include decreased accountability and diversity in writing. LLMs used in writing assistance can significantly influence human-authored content. Hohenstein and Jung [49] found LLM-generated

[2]https://github.com/salesforce/creativity_eval/tree/main/Writing_Alignment

text suggestions can affect a human writer's emotional tone. Arnold et al. [6] showed predictive text encourages predictable writing. Anderson et al. [4] and Laban et al. [61] found LLMs like ChatGPT helped users generate more detailed ideas, but outputs were less semantically distinct across users [87], and participants felt less responsible for their produced ideas. Recent work from Pan et al. [88] demonstrates language models can enhance outputs via feedback. However, methods like *Iterative Self-refinement scenarios*, using another language model as an evaluator, may result in reward hacking, where the model exploits the evaluator's flaws. For alignment training, it's crucial to consider the absolute desirability of each potential response, not just how responses compare to one another in terms of preference. Towards this in our work, we create pairs consisting of an initial LLM-generated response and its refined counterparts that by nature are more contrastive (or closely comparable). Our results show that such a pairing results in improved alignment and agreement during preference ranking.

3 DESIGN CONSIDERATIONS TO IMPROVE AI WRITING

"The secret to good writing is good editing.It's what separates hastily written, randomly punctuated, incoherent rants from learned polemics and op-eds, and cringe-worthy fan fiction from a critically acclaimed novel " [46]. In this section, we outline the design principles and desiderata that guided our approach to improving AI writing through textual edits.

Design Principle 1: Develop a comprehensive edit taxonomy grounded in expert writing practices. This principle emphasizes creating a comprehensive taxonomy of edit categories [33] rooted in an expert editor and writer's practices. Prior work has shown that experts and novices define revising in very different ways with experts attending more systematically to different aspects of the text than novices [9, 100, 106, 113]. By developing such a taxonomy, we aim to provide an approach to analyzing and enhancing LLM-generated text. It also allows for a more granular understanding of the specific areas where AI writing may fall short and enables targeted improvements. Sommers [106] found that "experienced writers have a second objective; a concern for their readership". Grounding the taxonomy in expert writing practices ensures that the edits align with the standards of high-quality writing and are acceptable to its readers. Finally, this principle also acknowledges the complexity of the editing process, recognizing that different categories of edits may be required at various levels of the text, from sentence-level corrections to broader structural changes [47, 106, 115].

Design Principle 2: Leverage edits to balance both meaning preservation and substantive semantic changes. Preserving the core meaning and intent of the original text is crucial to maintaining coherence and faithfulness to the initial ideas. On the other hand, introducing substantive semantic changes is often required to adhere to the quality and characteristics of good writing. Prior work on edit taxonomies focuses on low-level syntactic operations [33] or semantic edits [22, 62, 115] tailored to specific websites like Wikipedia. LLM-generated text often benefits from syntactic edits. These edits (primarily meaning preserving) enhance readability by diversifying sentence structures, expanding vocabulary choices, and minimizing repetitive phrasing. Consequently, semantic edits (both meaning preserving and changing) in AI writing are important for enhancing specificity or reducing unnecessary flourishes and clichés that can otherwise obscure meaning. Our methodology aims to navigate the tension between maintaining original meaning and introducing necessary improvements to mitigate AI-specific writing quirks.

Design Principle 3: Utilize edits as a mechanism for enhancing human-AI alignment in writing. Current AI writing systems are developed using pre-trained language models (LMs) refined through human interaction, employing supervised learning and reinforcement learning (RL) techniques. Reinforcement learning from human feedback (RLHF) [121] is a key approach, transforming human input into training data to guide LMs toward desired outcomes. The most common type of feedback used with RLHF is binary preferences between pairs of examples sampled from one or more Language Models [19]. However, a learned preference ordering can fail to converge to the true one when the desirability of examples depends on noise [42]. Following recent work in preference learning [25, 30], we evaluate edits as a mechanism for enhancing alignment. An LLM-generated response that has been edited typically contains fewer undesirable traits and can be paired with the original LLM-generated response for preference ranking ($edited > original$).While contemporaneous works [25, 30] have conducted preliminary efforts to incorporate edits for improved preference data collection, we evaluate this approach in the context of creative writing.

4 LARGE SCALE DATA COLLECTION PROCESS

We aim to create a valid test-bed to evaluate the quality of LLM-generated text on realistic writing tasks that require creative skill. We follow a three-step approach illustrated in Figure 2: (1) First we select original paragraphs of human-written text from trusted publication venues, (2) Second we reverse-engineer each of these paragraphs into a writing instruction. Because each instruction originates from an existing human-written paragraph within a piece of creative writing, this simulates real-world writing situations. (3) Third, we prompt several LLMs to generate responses to each of the writing instructions. In the following subsections, we first detail each of these steps, and then describe the formative study we conducted to develop a taxonomy of idiosyncrasies in LLM-generated text.

4.1 Collecting Instruction and Response Pairs

To curate source material, we select five well-regarded [3] publication venues, listed in Table 1, that publish pieces in different domains ranging from fiction to food writing and internet advice. For each venue, we manually extract between 100-700 pieces of writing and isolate individual paragraphs. We then manually review these paragraphs, ensuring they are long enough and can stand alone as coherent pieces of writing without requiring additional context. In total, we selected approximately 1200 paragraphs following this procedure. The Literary Fiction genre has a larger representation (80%) in our selection, while the creative non-fiction genres have a smaller representation.

[3] https://www.pewresearch.org/politics/2012/09/27/section-4-demographics-and-political-views-of-news-audiences/

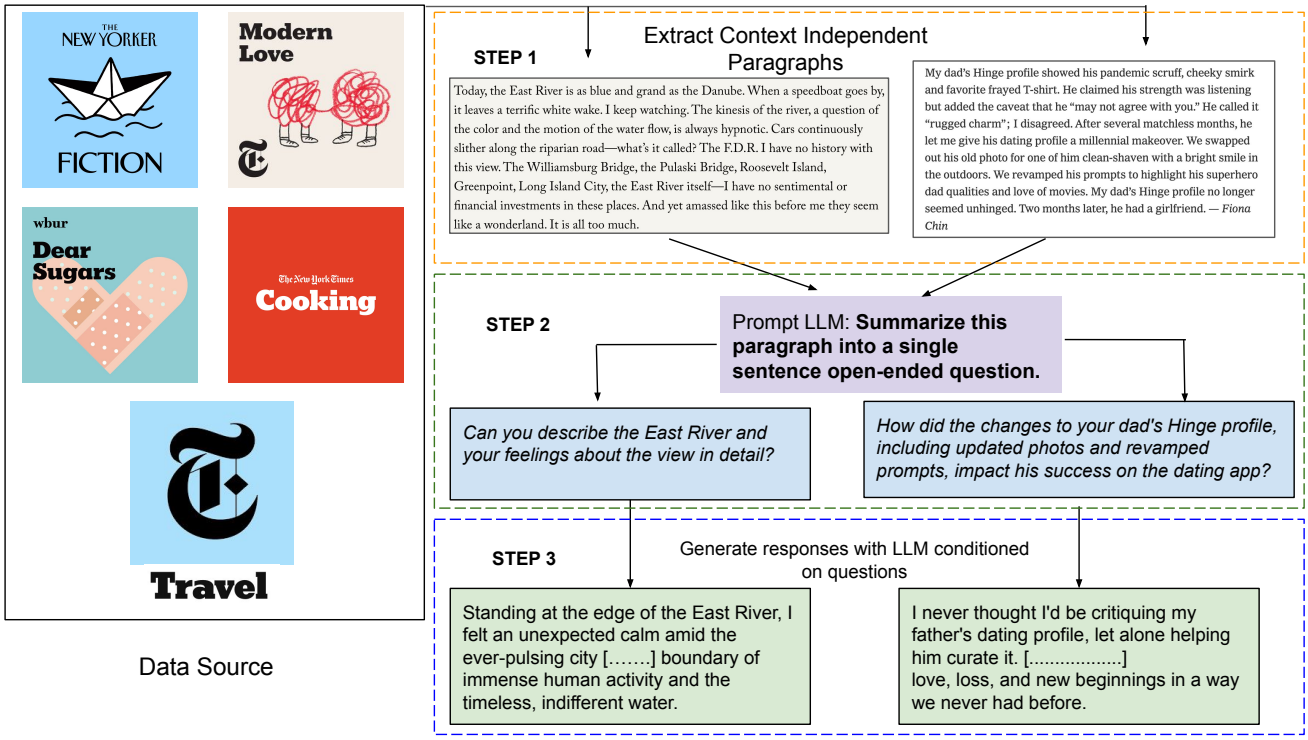


Figure 2: The pipeline for data creation. Step 1) Extracting context-independent paragraphs from our respective sources Step 2) Using an LLM to automatically generate instructions for corresponding human-written text Step 3) Use the generated instructions grounded in real-world writing to elicit responses from LLMs to create <instructions,response> pairs

Next, we follow Li et al. [67]'s approach of *Instruction Backtranslation* to automatically generate instructions corresponding to each of the selected paragraphs. Specifically, we prompt an LLM (i.e., GPT4o) to summarize each paragraph into an open-ended question. Questions obtained through back-translation (see examples in Table 1) can be interpreted as realistic writing instructions. We manually verified the generated instructions, filtering out questions that were ill-formed or overly specific, yielding a total of 1,057 writing instructions.

Finally, we use the generated instructions to elicit responses from three state-of-the-art LLMs: OpenAI's GPT-4o [83], Anthropic's Claude-3.5-Sonnet [5], and Meta's Llama 3.1-70b [30]. Each LLM is used to generate responses to one-third of our instruction data. We ensure that each LLM responds to instruction across all domains in equal proportion. To generate high-quality responses, we provide each LLM with the writing instruction, as well as the genre and source, and instruct it to adhere to the style of the venue. The prompt further specifies: *"Try your best to be original, avoiding clichés or overused tropes. Do not use ornamental language and focus on nuance, simplicity, and subtext"* (See Prompt in Appendix A Table 14). Through this process, we obtain 1,057 writing <instructions, response> pairs, with responses averaging 205 words. This collection of instructions and LLM-generated responses serves as the foundation for the three studies we conducted: the formative study, the full-scale editing annotation, and the preference annotation.

4.2 Formative study: formulating the taxonomy for fine-grained edits

Our formative study observed writers with copy-editing experience as they edited LLM-generated text in the creative writing domain. We aimed to identify common edit categories. The study consisted of three phases. First, participants were individually briefed via video conference on the study's objectives. Next, they accessed a web application (Figure 10) to view <instruction, response> pairs from our dataset (Section 5). For each sample, participants highlighted problematic response spans, suggested rewrites, and tagged each span with a free-form category to characterize the issue. We recruited eight participants for the formative study, with each completing annotations for 25 samples.

For participant selection in our formative study, we limited involvement to individuals with established expertise in creative writing. Participants were required to have completed a Master of Fine Arts (MFA) in Creative Writing and were recruited through mailing lists from MFA writing programs in the United States. This aligns with prior work from Chakrabarty et al. [15], using the Consensual Assessment Technique [2], which emphasizes the importance of recruiting domain experts. During the initial video call, we confirmed participants' familiarity with copy-editing and informed

Domain	Source & Genre	Example Seed Paragraph with Generated Instruction	#N
Fiction	The NewYorker (Literary Fiction)	The sunset is a red-gold rumpus on the western sky. It has rained. The crow tosses itself from branch to branch, pole to pole, glistening on its pace, and she follows. They are soon far from where they began, streets unfamiliar to her, an older part of town,[..... ....]. A man reading. Old Christmas tree in a corner. It feels secret. The sky is clearing overhead. She feels secret, too. She feels tremendous. Instruction : **Can you describe a vivid scene at sunset that transitions into nighttime, incorporating elements of nature, urban surroundings, and personal observations?**	815
Creative Non-Fiction	NYTimes (Travel Writing)	Prague, the Czech capital, is finding a new balance between preserving its past and embracing the future, improving many of its important historic sites while making striking additions to its skyline. [.....] Stop by for a coffee, hit up one of the many great new bakeries or visit a charismatic old beer hall as you explore a city that is clearly entering its prime. Instruction : **How is Prague balancing its historical preservation with modern development while enhancing local amenities and vibrant neighborhoods outside Old Town?**	110
	NYTimes (Food Writing)	The origins of the fruit sandwich are believed to go back to Japan's luxury fruit stores & the fruit parlors attached to them. This version comes from Yudai Kanayama, a native of Hokkaido who runs the restaurants the Izakaya NYC and Dr Clark in New York. [.....] .The sandwich looks like dessert but isn't, or not exactly; it makes for a lovely little meal that feels slightly illicit, as if for a moment there are no rules Instruction : **How did Yudai Kanayama reinvent the traditional Japanese fruit sandwich to create a unique culinary experience?**	83
	NYTimes (Personal Essay)	My dad's Hinge profile showed his pandemic scruff, cheeky smirk and favorite frayed T-shirt. He claimed his strength was listening [.....] We revamped his prompts to highlight his superhero dad qualities and love of movies. My dad's Hinge profile no longer seemed unhinged. Two months later, he had a girlfriend. Instruction : **How did the changes to your dad's Hinge profile, including updated photos and revamped prompts, impact his success on the dating app?**	19
	Dear Sugar (Internet Advice)	What is a prestigious college? What did attending such a school allow you to believe about yourself? What assumptions do you have about the colleges that you would not describe as prestigious? What sorts of people go to prestigious colleges and not [.....] I believe our early experiences and beliefs about our place in the world inform who we think we are and what we deserve and by what means it should be given to us. Instruction : **How do your beliefs and assumptions about educational privilege and the type of schools people attend shape your current view of yourself and others?**	30

Table 1: Venues of source material used to extract real-world examples of creative writing, along with example seed paragraphs and the generated instructions and the number of samples per genre.

them they would edit LLM-generated texts. We recruited participants through UserInterviews [4], a professional freelancing platform, paying $75 USD for study completion. Editing a response took 4-6 minutes, with all participants finishing within two hours. Table 2 shows diverse demographics and professional backgrounds of the recruited creative writers.[5] In total, the eight participants edited 200 paragraphs, annotating roughly 1,600 edits attributed to 50 distinct initial categories. We used this data as the foundation for our next analysis, which aimed to develop a taxonomy for categorizing edits.

[4]https://www.userinterviews.com

[5]The research was conducted at an institution without a formal IRB approval process. However, an Ethical Practices team reviewed the work and study protocols. No personally identifiable information (PII) was collected or shared during data collection, and participants were offered compensation regardless of study completion.

ID	Profession	Gender	Age	Educational Background
W1	Writer & Editor at Magazine	Male	28	MFA in Fiction
W2	Writer & Fellow at Fine Arts Work Center	Male	29	MFA in Fiction
W3	MFA Fiction Student	Male	31	MFA in Fiction
W4	Writer	Female	30	MFA in Non-Fiction
W5	Writer	Female	35	MFA in Fiction
W6	MFA Poetry Student	Female	27	MFA in Poetry
W7	Writer & Journalist	Female	35	MFA in Non-Fiction
W8	MFA Fiction Student	Male	26	MFA in Fiction

Table 2: Pilot study: background of participants.

4.3 From initial to final categorization of edits

We observed significant semantic overlap among the 50 initial categories used by participants, suggesting potential consolidation into a unified taxonomy. For instance, "Show don't tell" (W4) corresponded to "Unnecessary because implied" (W6). Using a general inductive approach for qualitative data analysis [111], we synthesized these 50 initial categories into a comprehensive, fine-grained taxonomy of edits. First, two authors independently bucketed these categories into initial low-level groups. Through iterative discussions, these groups were refined to reduce overlap and establish shared groupings. The refined low-level groups were then aggregated into high-level categories. Each high-level category was assigned a name reflecting its generalized representation.

The aggregation process yielded 7 distinct edit categories, presented in Table 3 along with contributing participant IDs. Final categories were retained only if derived from initial categories identified by at least four participants, ensuring majority representation in editing feedback [6]. It's worth noting that not every LLM-generated response exhibits all these idiosyncrasies. The formative study's objective was not to establish the relative prevalence of each category. Instead, this taxonomy serves as a useful framework when considering the categories of edits to apply to LLM-generated content. The categorization provides a structured approach to refining such text.

4.4 Final Taxonomy for Fine-Grained Edits

Here we describe our final taxonomy for fine-grained edits. Table 5 shows examples of edits in each of these categories defined below

4.4.1 ***Cliché.*** Clichés in writing are pejoratively characterized as phrases, ideas, or sentences overused to the point of losing their original impact or meaning. They often use vivid analogies or exaggerations from everyday life to describe abstract concepts. While occasionally effective when used sparingly, the frequent use of clichés in writing is generally viewed as a sign of inexperience or lack of originality [38]. Replacing clichés with fresh, original language improves the writing and engages readers more effectively.

4.4.2 ***Unnecessary/Redundant Exposition.*** Unnecessary or redundant exposition refers to the inclusion of excessive, repetitive, or implied information in writing. This common pitfall often involves restating the obvious or providing details that add little value. In a conversation with W2 they said *"I'm adding a category of edit called "fluff" - this is a common term in the writing world to refer to unnecessary filler"*. Effective writing embraces the principle of "show, don't tell," allowing readers to infer meaning from context rather than relying on explicit explanations [11, 17, 56, 82]. Impactful writing, often allows the core message to shine through without being obscured by unnecessary verbiage.

4.4.3 ***Purple Prose.*** In literary criticism, purple prose refers to excessively elaborate writing that disrupts the narrative flow by attracting undue attention to its flamboyant style [1]. This can detract from the text's overall appreciation. Such writing is often difficult to read, using sprawling sentences, abstract words, and excessive adjectives, adverbs, and metaphors to convey little information. Careful editing can trim purple prose by replacing ornate language with more direct expressions, resulting in clearer writing that preserves narrative flow and the author's voice.

4.4.4 ***Poor Sentence Structure.*** Poor sentence structure reduces the clarity and readability of writing [8, 57, 76]. A lack of proper transitions can make the text feel disjointed and hard to follow. Editing for clarity [23] often reveals that it's better to split a convoluted thought into two sentences, rather than forcing it into one [63]. Run-on sentences, characterized by multiple independent clauses improperly connected, are also frequent problems in AI writing [15]. These, very long and complex sentences can overwhelm the reader, making the core message difficult to grasp. Edits that reduce these problems lead to more coherent and fluent text.

4.4.5 ***Lack of Specificity and Detail.*** Lack of specificity and details in writing often stems from a writer's tendency to rely on broad generalizations [71]. This overly general approach fails to engage readers, leaving them unable to visualize scenes or connect with any given writing on a deeper level. Good writing often focuses on adding vivid details that create a clear mental image [26, 37, 60], contextualizing information to give it relevance [94], and deepening the internality of characters or subjects [20, 36]. Additionally, developing a unique voice through carefully chosen words and phrases can inject personality into the writing, making it more engaging and distinctive [48, 81]. Edits belonging to this category typically make the text longer as writers add more details to make the text engaging.

4.4.6 ***Awkward Word Choice and Phrasing.*** Awkward phrasing can significantly reduce writing quality, often confusing or disengaging readers. This issue typically involves misused or disproportionate use of certain words [59], unclear pronoun references, or an overuse of passive voice. In an email, W1 pointed out "*Another little observation to share: a very common phrasing in these excerpts*

[6] Only 5% of edit categories were not included in the 7 categories as they did not have enough coverage

Final Category	Initial Categories	Participants
Cliche	*Cliched image from old westerns, Cliche, Hackneyed*	W1,W2,W3,W4, W5,W6,W7,W8
Unnecessary/Redundant Exposition	*Repetition of what has already been stated, Unnecessary, Show don't tell, Repetition, Cut Unnecessary, Unnecessary because implied, Over exposition, Fluff, Slim down, Trying to cut things down, Concision*	W1,W2,W3,W4, W5,W6,W7,W8
Purple Prose	*Too wordy, Purple Prose, Ornamental, Very Verbose, Clunky Unnecessarily wordy, Simplify, Overwrought, Mixed metaphor*	W1,W3, W4 W5,W7,W8
Poor Sentence Structure	*Structure, Transition, Editing for clarity, Better to split up into two sentences, Run-on sentence, Very Long and Complex Sentence*	W1, W5, W7,W8
Lack of Specificity and Detail	*Lacks specificity, Overly General, More details to help move the reader, Added details, Creating a scene, Contextualizing information, Deepening internality, Needs to be more specific, Adding Voice*	W1,W2,W3,W4 W5,W6,W8
Awkward Word Choice and Phrasing	*Word Choice, Pronoun Clarity, Passive, Awkward Word Choice, Wrong choice of word, Rewording, Rephrasing, Weird Phrasing, Inelegant*	W1,W2,W3,W4 W5,W6,W7,W8
Tense Inconsistency	*Fragment sounds weird-is it past or present?, Wrong Tense, Inconsistent Tense*	W1, W5, W7,W8

Table 3: Our final taxonomy for fine-grained edits to mitigate idiosyncrasies in AI writing

is 'seem to _(verb)_'. This is not technically wrong it's just inelegant, something many writing teachers have told me to avoid. Unless there is some specific uncertainty or doubt about the verb action, it's always preferred to just use the verb without 'seem' (ex. from the current excerpt I have up: 'amplified' is better than 'seemed to amplify')". Editing plays a crucial role in refining these elements. Through careful revision, writers can identify and replace imprecise or ill-fitting words with more appropriate alternatives, ensuring each term accurately conveys the intended meaning.

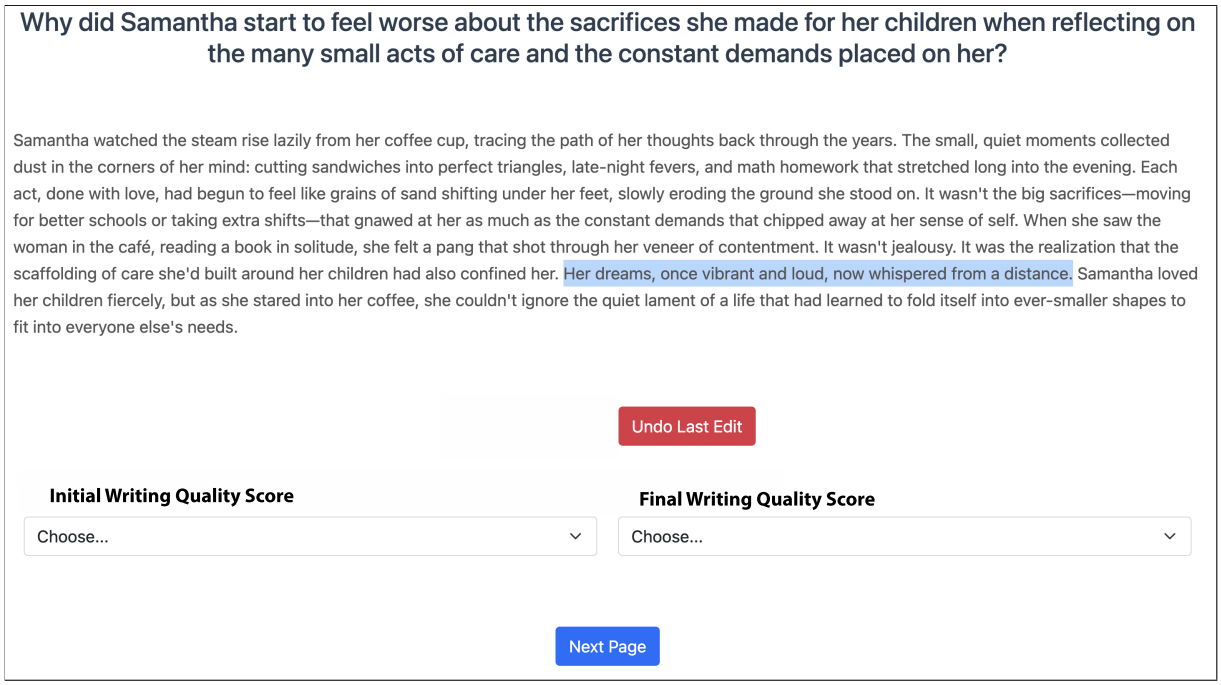
4.4.7 ***Tense Inconsistency.*** Tense inconsistency is a prevalent issue in writing. It occurs when a writer inadvertently shifts between past, present, and future tenses often even within the same paragraph or sentence. This grammatical misstep can make the timeline of events unclear and detract from the overall coherence of the text. Careful editing plays a crucial role in addressing this issue. By paying close attention to verb forms and temporal indicators, editors can improve writing that deals with tense inconsistency [79].

4.5 Collecting Edits on LLM Generated responses

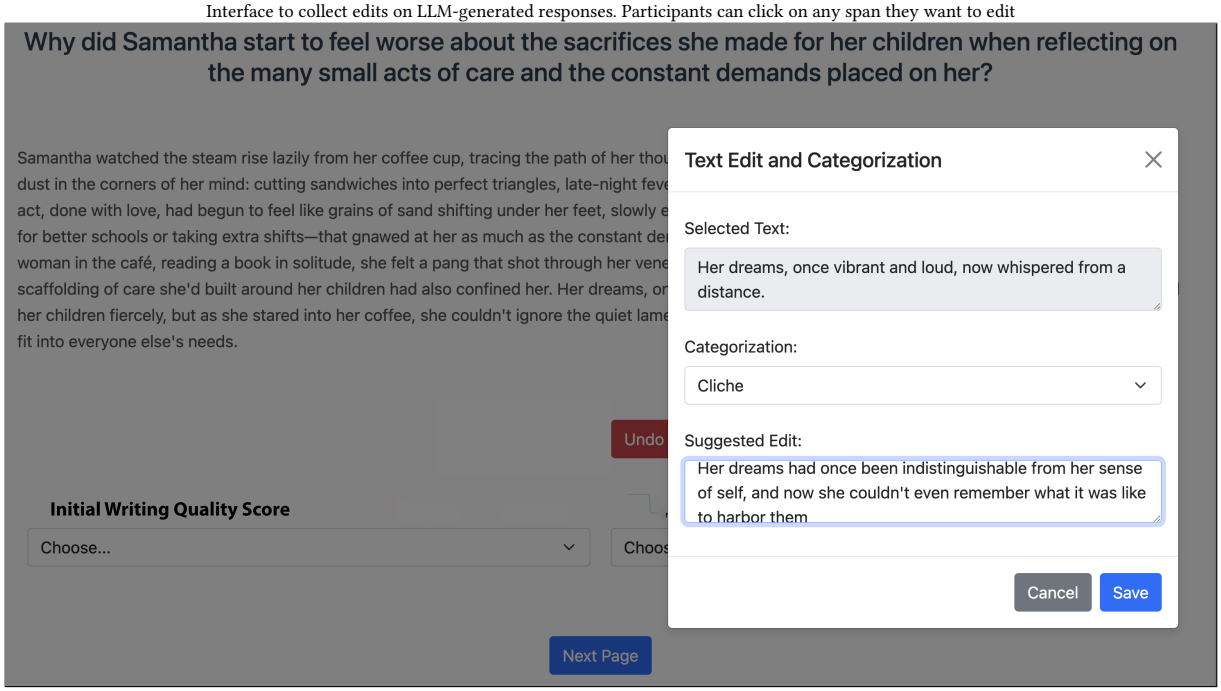
With the finalized taxonomy of edits, we next conducted a larger-scale annotation study. The purpose of this study was to collect edits from writers on LLM-generated responses, categorizing them according to the established taxonomy. We validated the taxonomy's comprehensiveness by consulting writers who were not part of the formative study but participated in the editing task. These writers had the option to select an "Other" category and provide its name if an edit didn't fall into any of the seven established categories. The writers rarely chose the *Other* category, doing so in only 10 out of 8,035 cases. These rare exceptions fell into categories such as "Repetitive Sentence Structure," "Confusing, Unclear or Incomplete Action/Meaning," and "Mixed Metaphors". We also separately asked the writers if the categories encompassed all traits they encountered while editing these paragraphs and whether they would suggest any additions. Through email exchanges, all writers confirmed the taxonomy's comprehensiveness.

This task followed a similar format to the formative study where participants were provided access to an editing interface (Figures 3) populated with instructions and LLM-generated responses. In this interface, participants could select any span of text in the response and suggest a rewrite. Unlike the formative study, participants had to choose from the seven predefined categories in our taxonomy for each edit, rather than entering free-text categories. Participants received training about the taxonomy via email before beginning annotation. The training incorporated example edits for each category, akin to those in Table 5. Participants had no set limit on edits per response but were urged to improve the text as they saw fit. The interface logged all edits chronologically and offered an undo feature, enabling us to track the entire editing process, not just the final product.

After completing their edits, participants assigned two scores to the sample: an *Initial Writing Quality Score* (IWQS) for the original response quality, and a *Final Writing Quality Score* (FWQS) for the post-edit quality. Both used a 1-10 scale, with 1 being the lowest and 10 being the highest quality. The scores were incorporated to add a quantitative dimension to the qualitative process of editing. Additionally, the self-reported writing quality scoring system serves as a signal for writers to recognize their own improvements, set



Interface to collect edits on LLM-generated responses. Participants can click on any span they want to edit



Pop-up edit window to input edit and label the category of edit

Figure 3: Interface to collect edits from writers on LLM-generated text

personal goals, and develop intrinsic motivation for enhancing their work.

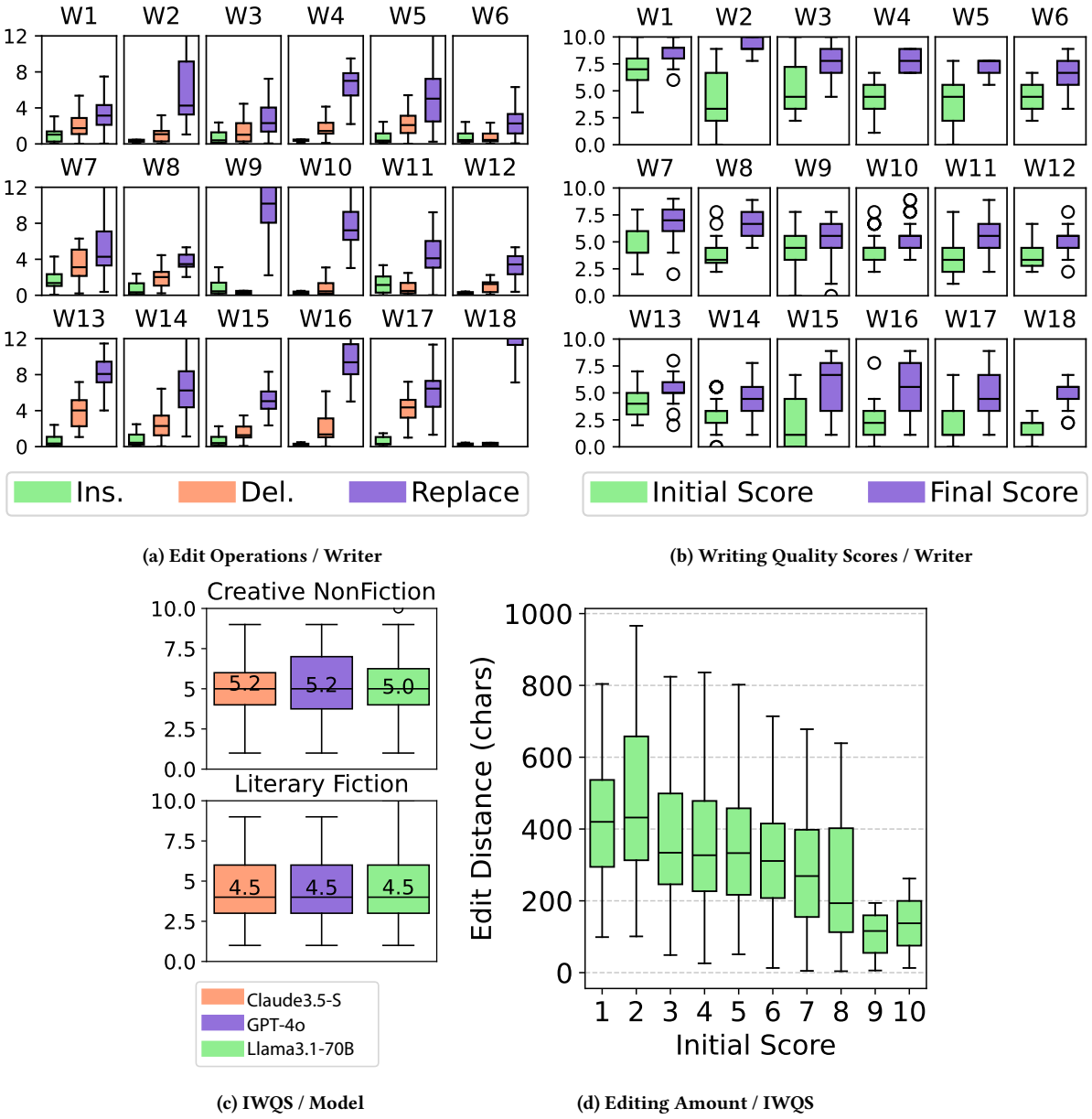
Editing is a personal, time-consuming task, with edit quality dependent on participants having sufficient time to carefully read and consider improvements. To ensure quality, we maintained communication with all recruited participants. Participants completed the task in batches of 25, which typically took 3 hours, and were compensated $100 USD for each batch. We recruited 18 writers with formal creative writing backgrounds from MFA mailing lists for

ID	Profession	Gender	Age	Educational Background	Responses_Edited
W1	Writer & Editor at Magazine	Male	28	MFA in Fiction	123
W2	Writer & Fellow at Fine Arts Work Center	Male	29	MFA in Fiction	109
W3	Writer & Teacher	Male	31	MFA in Fiction	119
W4	MFA Fiction Student & Translator	Male	30	MFA in Fiction	25
W5	Writer	Female	35	MFA in Poetry	77
W6	MFA Poetry Student	Female	27	MFA in Poetry	23
W7	Writer & Journalist & MFA Fiction Student	Female	35	MFA in Fiction	71
W8	MFA Fiction Student	Male	26	MFA in Fiction	23
W9	Writer & Editor	Male	30	MFA in Fiction	25
W10	Writer & Creative Writing Instructor	Female	28	MFA in Fiction	25
W11	Writer	Female	27	MFA in Poetry	25
W12	Writer & High School Teacher	Male	33	MFA in Fiction	24
W13	Writer & Editor	Female	29	MFA in Non-Fiction	25
W14	MFA Fiction Student	Male	26	MFA in Fiction	24
W15	Poet	Female	28	MFA in Poetry	125
W16	Writer & Director	Male	31	MFA in Literary Arts	122
W17	Writer	Non-Binary	28	MFA in Poetry	25
W18	Screenwriter & MFA Literary Arts Student	Female	27	MFA in Literary Arts	67

Table 4: Background of participants who provide span level edits on LLM generated responses

Categorization	Paragraph with rewrites
Cliche	As Sarah stepped off the bus, the scent of pine and damp earth enveloped her. [.....] In the kitchen, she found herself reaching for the cabinet where her mother always kept the coffee, only to stop short. ~~The realization that she was alone here, truly alone, settled over her like a heavy blanket.~~ **This time, though, she was alone. Her mother would never come back.**She sank into a chair at the old oak table[....]
Unnecessary / Redundant Exposition	As Mingus and Dylan stepped out of the car, [...] The Brooklyn-Queens Expressway ~~loomed above, a concrete behemoth that cast long shadows over the desolate landscape.~~ **cast a long shadow.**[...] For a moment, he stood there, lost in thought, as the city seemed to hold its breath around him.
Lack of Specificity and Detail	~~Dr. Arthur Steiger's fall from grace began with a series of whispered concerns among his colleagues at Cormac General Hospital.~~ **Pain was Dr. Arthur Steiger's forte. Not inflicting it, that is, but resolving it. Whenever a patient had problem, whether a tear in a tendon, a sprain, a knock, a headache, a broken bone – it was Dr. Steiger that knew what to do.**The small-town pain specialist had always been known for his compassionate approach, but as opioid addiction rates climbed in the community [....]
Poor Sentence Structure	~~As the night wore on, Z.'s laughter grew louder, his words slurring together like a sloppy melody. N. and I exchanged a knowing glance, our concern simmering beneath the surface.~~**Z. was drinking more and more as the night went on. He laughed more loudly. His words started to slur, blurring one into the next.I looked at N., who knew what I was thinking. We were going to have to take care of him.** At first, it was just a slight stumble, a misstep that could be brushed off as a joke. But as the hours passed,[...]
Purple Prose	My mother cried not just because twenty grand vanished into the ether[.....].All of it vanished~~, cycling back through her mind, not as numbers but memories of scraped knees she bandaged alone and birthdays where her absence was felt more acutely than her presence.The sobs emerged from this deep well of unspoken expectations, leaving behind a residue of weary resilience and a few hopeful echoes yet unwilling to completely extinguish.~~ **She cried. She cried deep from this well of scraped knees she bandaged alone and birthdays she missed to work. She cried for unfairness. She cried without relief.**
Awkward Word Choice and Phrasing	I remember the city as a place of perpetual twilight, where the sky ~~seemed to hover~~ **hovered** between dawn and dusk [....] glass towers, and the ~~sound~~ **music** of sirens [.....] bodega on the opposite side still ~~sold~~ **reeked of** warm beer and stale cigarettes. The people were a blur of faces, each with their own story of [....]
Tense Inconsistency	As the sun dipped below the horizon, Elliot found himself engulfed by the growing darkness on Route 7. The first snowflakes ~~began to drift~~ **drifted** down from the heavens,[...]

Table 5: Example of Edit types from our data



(a) Edit Operations / Writer

(b) Writing Quality Scores / Writer

(c) IWQS / Model

(d) Editing Amount / IWQS

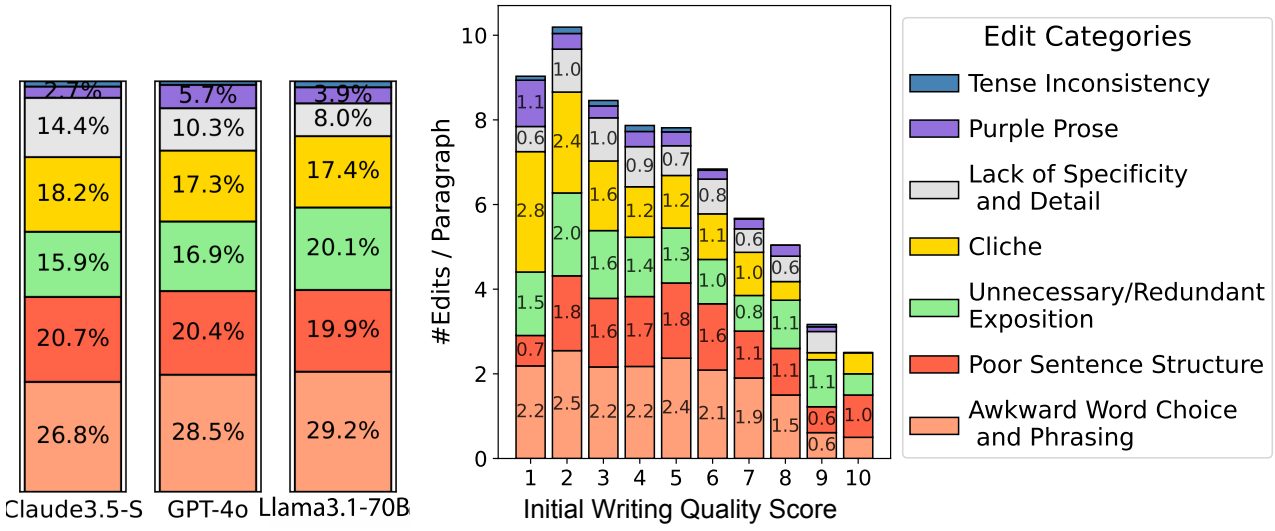
Figure 4: Analysis of 1,057 paragraphs edited by 18 Writer participants, analyzing: (a) the edit operations they perform (insertions, deletions, etc.), (b) the writing quality scores they assign, (c) comparing writing quality scores across LLMs, (d) the relationship between IWQS and editing amount.

our study, including 3 participants from the formative study (Table 4). Over 2.5 months, these writers edited LLM-generated responses based on their availability. Due to staggered start times, the number of edited samples varied among participants (see details in Table 4). In total, each of the 1,057 <instruction, response> pairs we had prepared was edited by at least one participant, and 50 responses were edited by three participants, allowing us to study similarities and differences that occur when multiple writers edit the same response. The next section details the analysis we performed on the 8,000+ collected edits.

5 The LAMP Corpus

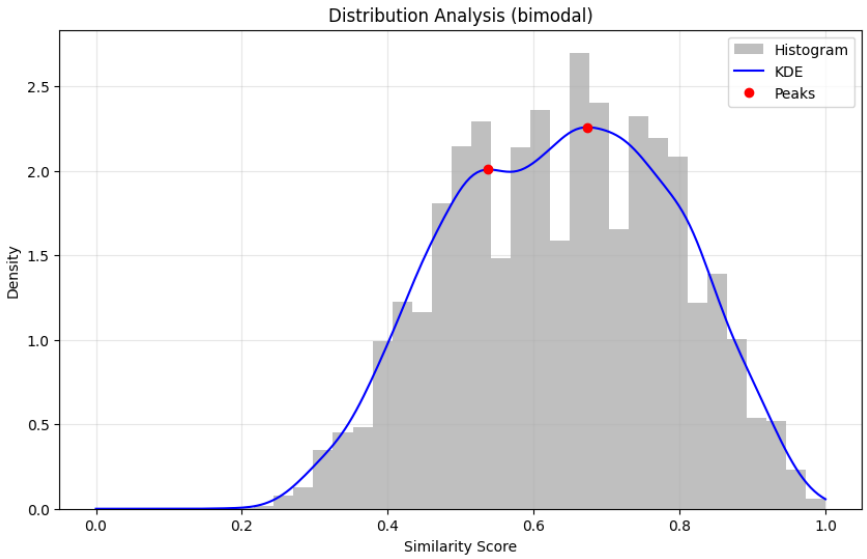
5.1 Overall Statistics

We created the LAMP Corpus by collaborating with 18 writers who edited 1,057 LLM-generated paragraphs, gathering about 8 edits per paragraph, totaling 8,035 fine-grained edits. The data includes paragraphs from Claude3.5 Sonnet (368), GPT4o (393), and Llama3.1-70B (296). Figures 4-5 present analyses of the LAMP Corpus, offering insights into how professional writers edit LLM-generated text and revealing a surprising lack of difference in writing quality across different model families [120].



(a) Edit Categories / Model

(b) Edit Categories / Score



(c) Distribution of semantic similarity scores between original and edited span

Figure 5: (a) the categories of edits they implement, and (b) the relationship between writing quality scores and error categories. (c) Distribution of semantic similarity scores for the edits in the dataset

Meaning Preserving	During the quarantine, the days stretched like endless corridors, each more indistinguishable from the last	0.72
	The days blurred into themselves during the quarantine, and I couldn't tell one from the other	
	She glanced down the hallway, suddenly aware of how quiet it was for a Tuesday evening	0.79
	It was eerily quiet for a Tuesday evening	
Meaning Changing	Sophia took her smoking breaks in the back garden, a ritual she kept as precise as the time on the old clock in her kitchen	0.52
	One of the great comforts of old age was the ability to stop caring what other people thought	
	a brief reprieve from the unsaid words that floated between them	0.48
	time to think of what to say. She hadn't told them her boyfriend was black	

Table 6: Examples of Meaning Preserving vs Meaning Changing Edits. Each example is a pair of original and edited span. Last Column shows the semantic similarity (BERT scores)

	W3	W12	W16
Original	,the numbers glaring back at me like an unsolvable riddle	glaring back at me like an unsolvable riddle	,the numbers glaring back at me like an unsolvable riddle
Category	Cliche	Cliche	Unnecessary/Redundant Exposition
Edited	. The numbers stared back	barreling over one another as they raced to some unseemly height.	-
Original	and an unsettling sense of mystery that gnawed at me more than the inexplicable weight itself	Her words felt like a placeholder for an answer neither of us had yet. I walked out with a slip for blood tests and an unsettling sense of mystery that gnawed at me more than the inexplicable weight itself.	-
Category	Purple Prose	Cliche	-
Edited	-	But when I saw her turn to go, whispering in the halls with a colleague, I knew there was still something she had yet to tell me.	-

Table 7: Original spans selected by 3 writers from the same paragraph. '-' denotes the span was deleted while editing

We analyze the editing process by examining *edit operations*: insertion, deletion, or replacement. An edit is an insertion if it deletes no characters or adds 40+ characters net. Conversely, it's a deletion if it adds no characters or removes 40+ characters net. All other edits are tagged as replacements. We choose a threshold of 40 characters (roughly 10 words), to avoid labeling edits with minor length changes ("a" to "the") as insertions or deletions. Figure 4a shows edit operations by participant for each paragraph. Replacements are most frequent (74%), followed by deletions (18%) and insertions (8%). Editing styles vary: some participants primarily use replacements (W2, W9, W10, W16, W18), while others employ deletions more often (W1, W5, W7, W8, W13, W17). Insertions are uncommon across all participants. To quantify meaning-preserving vs. meaning-changing edits, we calculate semantic similarity between original and edited text using BERT score[118]. Using a threshold of 0.6 [7], we classify edits with similarity > 0.6 as meaning-preserving. Of 6468 non-deletion edits, 70% are meaning-preserving, with the rest meaning-changing. This finding supports our Design Principle 2. Figure 5c shows the distribution of the semantic similarity scores for the edits in LAMP. [8]. While the distribution is bimodal (with peaks around 0.6 and 0.75), these modes don't align well with the 0.6 threshold. This suggests that meaning preservation versus meaning change exists more on a continuum rather than as two clearly distinct categories.

The annotation interface allowed participants to provide Initial and Final Writing Quality Scores (IWQS and FWQS) for each paragraph, ranging from 1 to 10. Figure 4b shows the distribution of these scores for each participant, revealing significant variability (e.g., W1's median IWQS is 7, W18's is 2). Calibration of writing quality scores is a known challenge, and we follow prior work in normalizing the scores into *z-Scores* by subtracting the mean, dividing by the standard deviation of the scores for each participant [45, 73], and re-scaling them to the 1 to 10 range. Subsequent analyses use these normalized scores.

[7]This threshold was decided by manually analyzing 100 edits

[8]However we note that with a threshold of 0.6, this appears to be quite a fuzzy/arbitrary distinction rather than a clear separation

We compute an *edit distance* between the original LLM-generated text and the final edited text, by calculating a character-level Levenshtein distance [66] between the two strings of texts. The edit distance measures the "amount of editing work" performed by a writer. Figure 4d shows a negative correlation between edit distance and IWQS (Pearson's $r = -0.31$), indicating that higher perceived text quality (high IWQS) requires less editing, while lower IWQS necessitates more editing.

Figure 4c shows the average IWQS for each LLM on creative non-fiction and fiction writing tasks. Writers were unaware of which model generated each text, and tasks were shuffled to avoid bias. This analysis estimates the writing quality of the three models in both domains. Comparing model scores, we find no significant difference in writing quality across the three models. GPT-4o and Claude 3.5-Sonnet perform slightly better on creative non-fiction instructions (average 5.2) compared to Llama3.1-70B (5.0), though the difference is not statistically significant. All models show a slight decrease in performance for fictional instructions, with an average IWQS of 4.5. This suggests fiction writing may be more challenging for LLMs than creative non-fiction. These findings differ from task-oriented benchmarks that reveal performance gaps between models in areas like factual or logical reasoning. **Our results indicate that, when it comes to creative writing, writers perceived no significant qualitative differences among the texts generated by large language models (LLMs) such as GPT-4, Claude 3.5 Sonnet, and Llama 3.1 70B.**

Figure 5a displays edit categories applied by writers to texts from three LLMs. The distribution is similar across models, with the most common categories being *Awkward Word Choice and Phrasing* (28%), *Poor Sentence Structure* (20%), *Unnecessary/Redundant Exposition* (18%), and *Clichés* (17%). Minor differences include GPT-4o using more purple prose and Llama3.1-70B generating more unnecessary exposition. Overall, **LLMs across the three model families exhibit similar idiosyncrasies that are edited out in similar proportions by professional writers.** Figure 5b illustrates the relationship between edit categories and IWQS. Higher IWQS scores correspond to fewer total edits, with texts rated 2 averaging

10.2 edits and those rated 10 receiving 2.4 edits, confirming that higher-quality texts need less editing. This trend however varies across edit categories: "Unnecessary/Redundant Exposition" and "Lack of Specificity and Detail" remain relatively constant, while the number of "Awkward Word Choice and Phrasing" and "Cliché" edits decrease as IWQS increases, suggesting a stronger correlation with perceived writing quality.

5.2 *Writers differ greatly in the amount of editing they do*: But to what extent?

The writer's approaches to editing vary based on personal or organizational philosophy. Some prioritize preserving the original voice and make minimal changes to preserve authenticity [109]. Others may take a more interventionist stance, heavily revising to align with their vision or house style. Additionally, some writers might make fewer but more impactful changes, while some might make numerous small revisions. To quantify this, we asked 3 writers (W3, W12, and W16) to edit a subset of the same 50 paragraphs from the LAMP Corpus. As expected, these three writers differed in the amount of editing they did. W3 did 9.4 edits on average while W12 and W16 did 6.0 and 6.3 edits on average. On average the span level precision (see Section 6.1 for more detail on the metric) between the 3 writers was 0.57 suggesting a moderately significant agreement.

Table 7 shows how sometimes writers select the exact same problematic span but assign different categories. A span that both W3 and W16 selected "*, the numbers glaring back at me like an unsolvable riddle*" (Table 7 Row 1) was categorized differently (Table 7 Row 2). Yet both categorizations can be correct interpretations. When one relies on overused phrases or clichés, they often state the obvious or provide information that readers can easily infer implicitly. This results in redundant or superfluous exposition that doesn't add value to the narrative. Other times writers may select the same category but with only partial overlap on the selected span (Table 7 Row 1 W3 vs W12). Looking at (Table 7 Row 4 and 5; W3 vs W12) there is a partial overlap in the selected span "*and an unsettling sense of mystery that gnawed at me more than the inexplicable weight itself*". However, the selected categories are Purple Prose and Cliché respectively. Here again, it should be noted that Purple Prose is a style of writing that can be original or cliché, depending on its usage, context, and frequency. Not all elaborate writing is overused, but when certain ornate phrases or styles become too common, they can cross the line into cliché territory. W16 however did not edit this span.

We also highlight that diversity in edits among writers such as selecting different spans or rewriting it in an individualistic style is a positive aspect that prevents homogenization while still improving LLM-generated text as shown by our results in Section 6.2.

5.3 Are there any specific stylistic idiosyncrasies in LLM generated responses?

Recent work from Shaib et al. [103] uses syntactic patterns with Part-of-speech [9] as abstract representations of texts, that can capture more subtle repetitions than mere text memorization. They find that language models tend to use repetitive syntactic templates more often than humans and these patterns can help evaluate style memorization in language models. Following their experiments we consider Part-of-speech templates of length $n \epsilon \{5, 6, 7, 8\}$ in LLM-generated responses as well as the original seed human-written paragraphs (Table 1). We looked at the 50 most common templates in LLM-generated responses and found that 15 templates do not occur as frequently in original human-written seed paragraphs. Table 8 shows representative sequences corresponding to particular syntactic patterns present in higher proportion in LLM-generated responses. These sequences constitute categories of **Clichés, Unnecessary/Redundant Exposition or Poor Sentence Structure** and are often heavily edited by writers in our study.

To better understand idiosyncrasies, we examined awkward words/phrases occurring disproportionately in LLM-generated responses. For instance, Figure 6 shows how a word like *unspoken* occurs in about 15% of LLM-generated responses. Similarly phrases such as *weight of*, *sense of*, *mix of* occur very rarely or not at all in original seed paragraphs (Table 1) while they occur frequently in LLM-generated responses. We also found peculiar and uncommon phrases generated by LLMs across several responses such as *air was thick*, *hung in the air*, *eyes darting*, *a sense of unease (grew/growing/settles) in the pit of (her/my) stomach*. The most surprising finding is that all 3 LLMs generate these idiosyncratic words/phrases **suggesting possible overlap/mixture in instruction tuning data across model families or one model trained on synthetic data generated from another model** [120].

6 AUTOMATIC DETECTION AND REWRITING OF LLM IDIOSYNCRASIES

While several automated editing approaches exist for improving LLM outputs at scale, we evaluate their effectiveness specifically for addressing idiosyncrasies in creative writing. Building on Hayes et al. [47] and Scardamalia [102], we develop techniques to separate detection and rewriting tasks, evaluating them using LAMP Corpus annotations. Given automated evaluation limitations for text editing[27], we conduct a large-scale preference annotation study with LAMP Corpus writers, comparing human and LLM-produced edits. To accommodate methods that require training samples, we split our data: 146 of 1057 LAMP Corpus paragraphs for training, the rest for testing.

6.1 Automatic detection of problematic spans in LLM-generated text

We formulate the problem of detecting problematic spans in LLM-generated text as a *multi-span categorical extraction* problem. In other words, given a paragraph of LLM-generated text, the method must output a list of non-overlapping spans present in the original text, and assign a category to each extracted span (from the list of categories of the LAMP Corpus).

To evaluate various methods, we use the span-level precision metric, a common metric used in NLP tasks requiring comparison of extracted spans [96]. Span-level precision measures the degree of overlap between predicted spans and reference or ground truth spans (in our case spans collected as a part of the editing process from writers). The overlap is measured at the character level, such

[9] https://www.sketchengine.eu/blog/pos-tags/

Syntactic Pattern	% of Times Edited	Representative Sequence
DT NN IN NN CC	54%	a mix of pride and, a mix of fear and, a sense of protection and, a sense of wonder and, a means of connection and, a pang of nostalgia and, a pang of disappointment, but, a flicker of hope or, a blend of relaxation and, a blend of curiosity and, the power of storytelling, the power of empathy, and, a web of belonging and [....]
NN IN NN CC NN	35	scene of chaos and destruction, mix of desperation and resolve, mix of relief and gratitude, torn between curiosity and caution, breakfast of bread and jam, sense of calm and normalcy, perception of loyalty and identity, glimmer of fear and vulnerability, meaning of protection and care, story of struggle and resilience, blend of fear and hope, [....]
DT JJ NN IN PRP$	40	a constant reminder of his, the mundane routine of our, the intricate tapestry of its, the subtle shift in their, the potential weight of its, a quiet sigh as her, a small acknowledgment of their, the upcoming chapter of her, a silent battle between his, a complex blend of their, the subtle shift in her, the unspoken plea in her [...]
DT NN IN JJ NN	27	the fabric of daily life, a moment of genuine connection, a life of absolute relaxation, the face of inevitable loss, the weight of past grievances, a state of constant unease, a residue of weary resilience, a sea of unspoken expectations, a mask of controlled concern, the weight of unresolved history, a foundation of silent understanding, [....]
IN DT NN IN NN CC	45	with a mix of wariness and, by the hum of traffic and, in a flurry of pursuit and, into a world of precision and, in a gesture of comfort and, in a storm of pain and, for the sake of stability and, in the rhythm of routine or, in the magic of family and, like the depth of understanding and, with a sense of nuance and [...]

Table 8: Idiosyncratic Sequences following certain syntactic patterns in LLM generated responses that are edited by writers. These syntactic patterns do not occur in the human written seed paragraphs

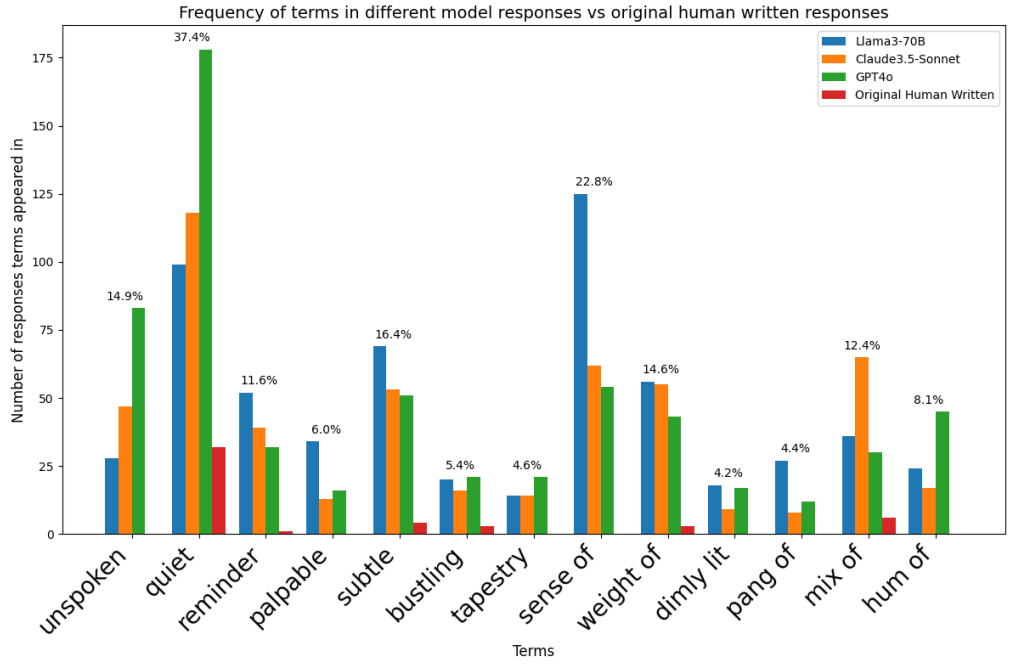


Figure 6: Distribution of peculiar and odd words and phrases occurring in LLM-generated text vs. human-written text in the LAMP Corpus.

that spans that partially overlap will get precision scores that reflect the amount of overlap between the two spans. High span-level precision indicates that the model is precise in identifying the correct boundaries of relevant text spans without over-predicting. We implement two precision metric variants: General and Categorical Precision. General Precision credits span selection regardless of category assignment, while Categorical Precision requires correct category assignment. We use a precision-based metric (like BLEU

[90]) rather than recall-based (like ROUGE [69]) as LLM-based methods tend to over-generate spans, which recall doesn't penalize. Our focus is measuring the overlap between generated and ground truth spans. Appendix A.2 provides a simplified example where General and Categorical Precision are computed, illustrating the suitability of these metrics for this evaluation setting.

	General			Categorical		
Expert Agreement	0.57			0.23		
Detector LLM	n=2	n=5	n=25	n=2	n=5	n=25
Claude3.5-Sonnet	0.43	**0.46**	0.44	0.20	0.21	0.20
Llama3.1-70b	0.42	0.45	0.38	0.16	0.21	0.14
GPT-4o	0.44	**0.46**	**0.46**	0.17	0.18	0.17

Table 9: Results of automated methods for detecting problematic spans in LLM-generated text, compared to agreement levels between three experts. Results report Precision scores for various LLMs used for detection, using a 2,5,25 examples in the few-shot prompt instruction, reporting both a General Precision and the stricter Categorical Precision.

We implement few-shot LLM-based methods [10] that have demonstrated competitive performance on tasks across several disciplines, often using fewer than 100 examples [10]. Our experiment varies the number of few-shot examples (2, 5, and 25) with the 2-shot prompt in Appendix A.1 and tests Llama3.1-70B, GPT-4o, and Claude3.5-Sonnet. As part of the collection of the LAMP Corpus, 50 paragraphs were edited independently, and we computed General and Categorical Precision on this set to estimate expert-expert agreement. Table 9 summarizes results. The best General Precision (0.46) is achieved by Claude-3.5 Sonnet and GPT-4o with a 5-shot prompt, below the expert agreement level (0.57). LLM-based methods can identify problematic spans with significant expert overlap, but improvement is possible. Performance improves from 2-shot to 5-shot prompts but plateaus thereafter (Claude-3.5 Sonnet and GPT-4o achieve similar or better performance with 5-shot vs. 25-shot prompts).

Categorical Precision is consistently lower than General Precision for both LLM-based methods and writer-writer comparison, suggesting that even when problematic spans are commonly identified, category agreement may differ. Section 5.2 explores this finding in detail, with examples of spans that could be reasonably assigned multiple categories. Table 10 illustrates the contrast between automatic editing and writer-selected edits by showing a paragraph with problematic spans and categories identified by a writer versus an LLM.

6.2 Automatic rewriting of problematic spans in any LLM-generated text

To propose improvements for detected problematic spans, we use few-shot prompting [10] with LLMs as well. We design prompts for each of the seven edit categories, incorporating examples of rewrites from writers. Each prompt includes a category definition, 25 examples from the LAMP Corpus with original paragraphs, a single problematic span for the category, the expert-proposed rewrite, and finally the input paragraph with the target span to be rewritten. The prompts for each category are listed in Appendix A.3.

The detection and rewriting methods can form a two-step pipeline for editing paragraphs. Detection identifies problematic spans and assigns categories while rewriting uses category-specific prompts to revise each detected span. A final step replaces all problematic spans in the original paragraph with their rewrites. Unlike the detection task, we do not evaluate the rewriting stage in isolation. Instead, we judge the complete pipeline that edits an entire paragraph (by detecting and rewriting multiple spans) through manual evaluation with 12 writers that annotated the LAMP Corpus. We describe this manual experiment next.

6.3 Evaluating Automatic Editing of LLM-generated Text

To evaluate editing quality, we design an evaluation task where participants read three variants of a paragraph and rank them in terms of overall preference: (1) an unedited **LLM-generated** paragraph from the LAMP Corpus, (2) the **Writer-edited** version from the LAMP Corpus, and (3) an **LLM-edited** version using our pipeline to detect and rewrite problematic spans. 6 out of 18 of our experts were unavailable during the preference annotations. Some of them were busy with their full-time jobs while others wanted to take a break after the edit task. Given the concern for proper turn around we re-hired 12 of the 18 experts who had participated in creating the LAMP Corpus for this evaluation.

We split the LLM-edited variant into two further sub-conditions:

- **Writer Detected and LLM Rewritten**: In this condition, the pipeline skips automatic detection of problematic spans, relying only on reference spans selected by the writer during manual editing. It runs solely the rewriting stage, simulating an oracle setting where problematic spans are manually provided. This condition is coded as **LLM-edited-Oracle**
- **LLM Detected and LLM Rewritten**: In this condition, the two-step pipeline is entirely automatic, with the automatically detected spans being provided to the automatic rewriting module. This condition fully automates editing of the paragraph and is coded as **LLM-edited-full**.

Table 11 shows examples of LLM-edited paragraphs under both sub-conditions. Although the examples here showcase 'Replacement' and 'Deletion' edits, analysis on a larger corpus of 200 LLM-edited paragraphs confirms that the automatic editing pipeline mirrors the edit type distribution of expert editors (shown in Figure 4a): roughly 65% 'replace' edits, 25% 'delete' edits, and 10% 'insert' edits. This surface-level analysis confirms the automatic pipeline's ability to mimic the editing process of experts measured in terms of relative proportions of character insertion and deletion.

In our pilot evaluation, we initially included all four conditions for annotation. However, ranking four paragraphs proved challenging for participants, especially when distinguishing between their second and third preferences. Based on this feedback, we redesigned the task to have participants judge only three conditions in each annotation. We always included the **LLM-generated** and **Writer-edited** paragraphs and alternated between including **LLM-edited-oracle** and **LLM-edited-full** paragraphs. While traditionally pairwise rankings tend to be easier for humans to annotate,

[10] Model fine-tuning experiments on LAMP Corpus are left for future work

Writer	Jackson leaned back in his office chair [, **staring out**$_{\text{Poor Sentence Structure}}$] through the expansive glass windows of the high-rise building. Below him, the city churned with [**relentless**$_{\text{Awkward Word Choice and Phrasing}}$] energy, [**a tapestry of modernity threaded with the hum of traffic,**$_{\text{Cliché}}$] the glow of digital billboards, [**and the unceasing flow of pedestrians. His desk, a**$_{\text{Poor Sentence Structure}}$] chaotic mixture of reports and coffee cups [, **contrasted sharply with the sleek orderliness of the cityscape**$_{\text{Poor Sentence Structure}}$] [**Despite the apparent advancement that framed his daily life, Jackson felt an inexplicable disconnection. His**$_{\text{Unnecessary/Redundant Exposition}}$] mind often wandered to an untouched box of old photographs at home, depicting simpler times – wooden houses, dirt paths, faces etched with stories of a bygone era. Those images spoke to him more than the polished chrome and glass now surrounding him.[**He couldn't help but yearn for the intangible warmth of community gatherings under open skies, unhurried conversations, and the tangible reality of a slower pace. Here, in the heart of progress, he was a mere observer, detached from the legacy beneath the steel and concrete, his own inclinations buried beneath the layers of modernity.**$_{\text{Cliché}}$]
LLM	Jackson leaned back in his office chair, staring out through the expansive glass windows of the high-rise building. Below him, the city [**churned with relentless energy**$_{\text{Cliche}}$].The city was [**a tapestry of modernity threaded with the hum of traffic, the glow of digital billboards, and the unceasing flow of pedestrians**$_{\text{Purple Prose}}$]. His desk, a [**chaotic mixture of reports and coffee cups**$_{\text{Cliche}}$] contrasted sharply with **the sleek orderliness.** [**of the cityscape**$_{\text{Cliche}}$]. Despite the apparent advancement that framed his daily life, Jackson felt **an inexplicable** [**disconnection**$_{\text{Cliche}}$] His mind often wandered to an untouched box of old photographs at home, depicting simpler times – [**wooden houses, dirt paths, faces etched with stories of a bygone era.**$_{\text{Cliche}}$] He couldn't help but yearn for [**the intangible warmth of community gatherings under open skies, unhurried conversations, and the tangible reality of a slower pace**$_{\text{Purple Prose)}}$] Here, in the [**heart of progress**$_{\text{Cliche}}$], he was a mere observer, [**detached from the legacy beneath the steel and concrete, his own inclinations buried beneath the layers of modernity.**$_{\text{Purple Prose)}}$]

Table 10: Example of problematic spans identified by a writer and by a LLM.

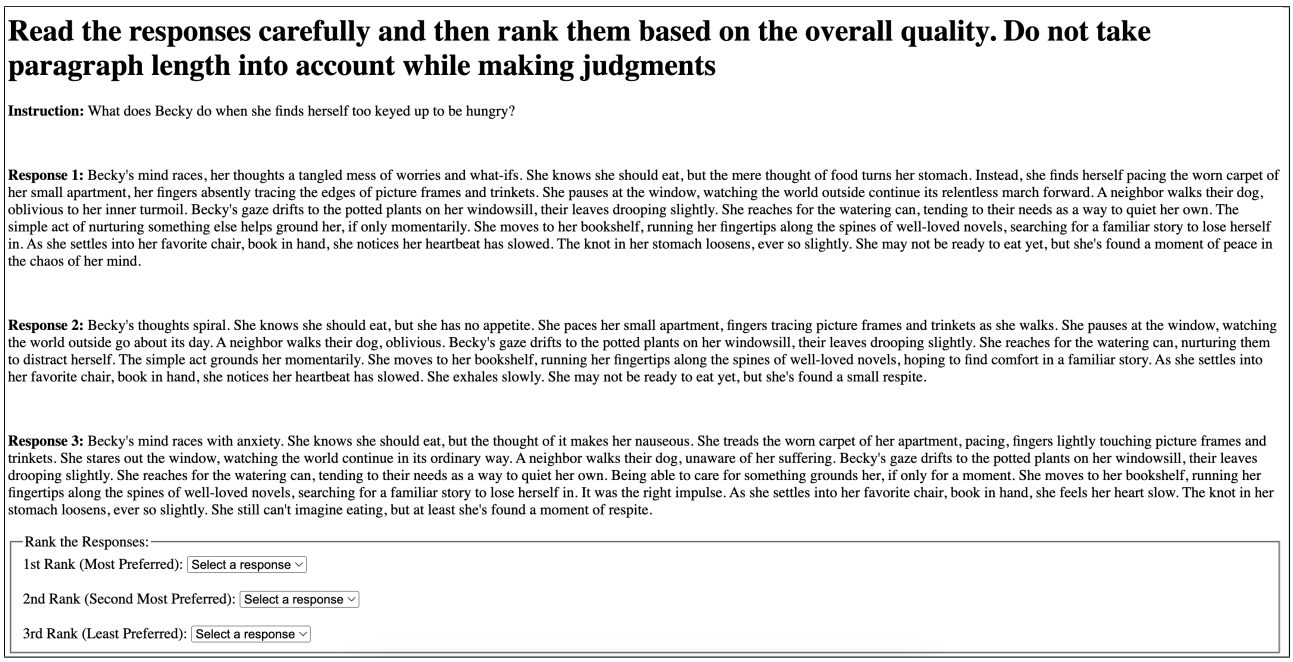


Figure 7: Interface used by participants to read through variants of a paragraph (one LLM-generated, one manually edited by an expert, one edited by an LLM-based system), and rank them in terms of preference.

three-way ranking allowed direct comparison between the key conditions we wanted to evaluate (LLM-generated v.s. Writer-edited v.s. LLM-edited), making the relative preferences clearer in a single

Writer Detected and LLM Rewritten	Dust settled on my window sill, ~~a quiet testament to the daily upheaval outside~~ **a reminder of the ongoing transformation outside**. Yellow machines trundled along the narrow street, ~~their relentless growl a constant soundtrack~~ **growling incessantly**. I watched the workers in their neon vests, ~~like bright insects against the concrete hive~~ **bright and busy against the concrete backdrop**, orchestrating a ballet of drills and hammers. Every so often, a stone ~~house crumbled~~ **mansion fell** to give way to sleek modernism, steel skeletons reaching shyly towards a haze-obscured sky. ~~The trees lining the avenue, long sentries guarding our history, stood fewer each day~~ **The trees lining the avenue dwindled each day**. Neighbors ~~nodded at one another with tight smiles~~ **exchanged tight smiles,** masking the void ~~of vanished gardens and familiar creeks~~ **left by vanished gardens**. ~~Change buzzed in the air, metallic and cold, seeping into conversations and dreams.~~ **Change felt cold and metallic, entering conversations and dreams**. I ~~found myself wondering if we were~~ **wondered if we were** building towards ~~something or away from it~~**something or abandoning it.**
LLM Detected and LLM Rewritten	~~Becky's mind races, her thoughts a tangled mess of worries and what-ifs.~~ **Becky's thoughts spiral**. She knows she should eat, but ~~the mere thought of food turns her stomach.~~**she has no appetite**. Instead, she ~~finds herself pacing the worn carpet of her small apartment, her fingers absently tracing the edges of picture frames and trinkets.~~ **paces her small apartment, fingers tracing picture frames and trinkets as she walks.** She pauses at the window, watching the world outside ~~continue its relentless march forward~~ **go about its day.** A neighbor walks their dog, ~~oblivious to her inner turmoil~~ **oblivious**. Becky's gaze drifts to the potted plants on her windowsill, their leaves drooping slightly. She reaches for the watering can, ~~tending to their needs as a way to quiet her own~~**nurturing them to distract herself.**~~The simple act of nurturing something else helps ground her, if only momentarily.~~ **The simple act grounds her momentarily.** She moves to her bookshelf, running her fingertips along the spines of well-loved novels, ~~searching for a familiar story to lose herself in~~ **hoping to find comfort in a familiar story.** As she settles into her favorite chair, book in hand, she notices her heartbeat has slowed. ~~The knot in her stomach loosens, ever so slightly.~~ **She exhales slowly.** She may not be ready to eat yet, but ~~she's found a moment of peace in the chaos of her mind.~~ **she's found a small respite.**

Table 11: Table showing LLM-edited paragraphs on both Oracle Writer predicted spans as well as LLM predicted spans

LLM-generated	Writer-edited	LLM-edited-full
2.55	1.47	1.99

LLM-generated	Writer-edited	LLM-edited-oracle
2.47	1.53	1.99

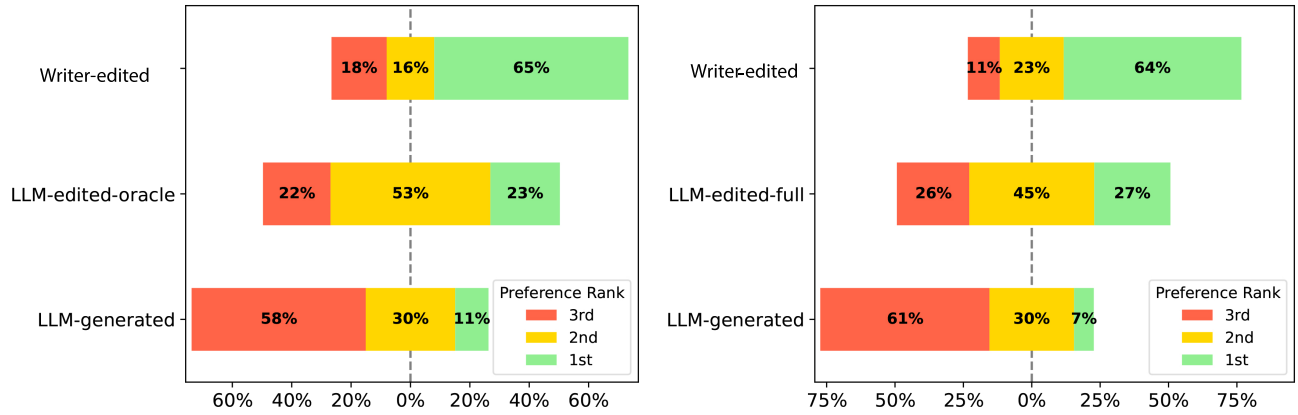
Table 12: Average Ranking across 600 preference judgments. LLM-edited > LLM-generated (p-value: 1.3e-11 for Writer Predicted spans; 2.8e-13 for LLM Predicted spans) and Writer-edited > LLM-generated (p-value: 1.1e-26 for Writer Predicted spans; 1.17e-31 for LLM Predicted spans) using Wilcoxon signed-rank test

evaluation rather than having to combine multiple pairwise comparisons. Additionally, we want to avoid any familiarity effects or biases that could arise from evaluators reading the same text multiple times, which would likely occur when collecting overlapping pairwise comparisons.

We note that to obtain automatic edits of a paragraph, we used the same LLM that had originally been used to generate the paragraph.[11] While not optimal, as a single LLM might offer slightly better detection and rewriting capabilities, this approach allows us to simplify the experiment conceptually and also test our hypothesis if edits lead to overall better alignment without relying on a single model family. We assess if using an LLM in a multi-stage pipeline (drafting, problem detection, rewriting) can enhance overall writing quality. Future work could potentially optimize this editing pipeline further, possibly yielding better results for LLM-edited conditions.

To ensure fairness, paragraph variants are displayed in a shuffled order and anonymized, and participants were not informed about the difference between the paragraphs (i.e. whether they are edited). For the curious reader, Figure 7 provides the interface used for the annotation task, including three variants of a paragraph. To conduct our experiments, we selected a total of 200 paragraph triplets (100 including an **LLM-edited-oracle** paragraph, and 100 including an **LLM-edited-full** paragraph) selecting samples from the LAMP Corpus's test set. Preference judgments were collected in batches of 25-35 paragraph triplets, with participants paid $35/hour. To account for potential subjectivity and calculate agreement and reliability, three experts judged each triplet, totaling 600 annotated

[11] In other words, we used GPT-4o in the two-step pipeline to generate edits to paragraphs that were originally generated by GPT-4o.



(a) Samples involving oracle detection (N=300)

(b) Samples involving automated detection (N=300)

Figure 8: Distribution of rankings for each variant in the preference annotation study. Annotators read three variants of a paragraph (Writer-edited, LLM-generated, and either LLM-edited-oracle or LLM-edited-full) and ranked them by preference (1st, 2nd, 3rd). The distribution indicates how often each variant was ranked as best (1st), second best (2nd), or worst (3rd).

preference rankings. To ensure the validity of the results, **no participant reviewed paragraphs they had seen or edited in past tasks, and only judged paragraphs edited by other experts**.

To analyze the reliability of the results we calculate inter-annotator agreement using Kendall's W (also known as Kendall's coefficient of concordance) [35] which ranges from 0 (no agreement) to 1 (complete agreement) to evaluate agreement amongst participants. Our annotation achieves an overall agreement of **0.505**, suggesting a moderate level of agreement across all participants. This moderate agreement underscores the subjective nature of judging writing quality while suggesting that certain differences are distinctive enough to be consistently preferred by multiple participants.

Table 12 and Figure 8 summarize the preference evaluation results, showing average ranks across 600 annotations. Overall, the **Writer-edited condition is most preferred, a sign that expert-edited text is unrivaled in terms of writing quality**, being marked as the most-preferred paragraph variant 65% of the time and achieving an average rank of 1.5. Next, the LLM-edited variants come in second, with an average rank of 1.99 for both the LLM-edited-oracle and LLM-edited-full conditions. Surprisingly, the condition that leveraged the oracle span from writers ranks almost identical to the condition with automatically detected spans. This provides evidence that detection of problematic spans is not the bottleneck in improving writing quality, and instead the **rewriting module (which is common to both conditions) is what dictates the overall performance of an automated text-editing pipeline**. Finally, the original **LLM-generated** paragraphs achieve the worst ranking performance, being least preferred 60% of the time, and achieving an average rank of 2.51.

In summary, our experiment validates the potential benefit of automatic editing to improve writing quality: although automatic editing does match the quality of edits provided by professional writers, LLM-edited text is significantly preferred to LLM-generated text by expert writers (Design Principle 3). In other words, this experiment shows that **LLMs can improve the quality of their writing in a fully automatic way, by first generating a draft, selecting problematic spans, and then rewriting such spans**. These results align with previous findings [32, 40, 89, 114] showing that iterative refinement and editing can improve LLM outputs.

6.4 Qualitative Insights into LLM edits

Given how LLM-edited paragraphs are often preferred over default LLM-generated ones and sometimes even Writer edited paragraphs, we inspect if there are specific categories of edits where LLMs perform better, more so if that caters to a certain genre. Table 13 shows examples of original spans edited by LLMs that vary across genre and category. In general, there isn't a domain or category where LLM edits are better than the rest except for `Purple Prose`. To edit `Purple Prose` models mostly need to learn how to simplify the text (i.e., generate a paraphrase in simple language). LLMs are effective at Sentence Simplification [28] so this doesn't come across as a surprising finding. Categories like `Poor Sentence Structure` and `Unnecessary Exposition` deliver mixed results. Sometimes they are effective. For instance Row 7 in Table 13 shows how the model improves the structure by breaking the run-on very long sentence. However, for the other example in Row 6, the human edit is *Then he hands his card, always leaving a 22% tip* while the model edit is a mere paraphrase. Similarly, under the `Unnecessary Exposition` category, the LLM is very good at editing sentences with the following structure [Main clause] [comma] [exposition] where they simply remove the exposition (as learned from Writer edits). However, the bigger issue in this category is revealing the subtext. When a given input span doesn't conform to the structure above or consists of multiple sentences (Row 9, Table 13) LLMs don't align with human edits. For `Cliché`, models mostly rewrite it in simpler language typically with fewer words which while being effective isn't what humans do. This also shows the fundamental overlap between `Cliché` and `Purple Prose` as shown in Section 5.2. Finally, models are mostly ineffective at addressing `Lack of Specificity and Detail` (more details in section 7.2). These findings demonstrate the challenge of editing text across any domains

Genre	Category	Original Span	LLM edited span
Literary Fiction	Lack of Specificity and Detail	seems to embody the city's frustrations and disappointments	seems to embody the city's frustrations and disappointments, his eyes a deep well of anger and desperation, his face a topographic map of the city's corruption and neglect.
Food Writing		Gone are the individual ramekins, replaced by a single, generous vessel that invites sharing.	Gone are the individual ramekins, replaced by a deep-dish pie plate that invites communal indulgence. The golden crust cradles a sea of silky custard, its surface a glassy expanse of caramelized sugar.
Literary Fiction	Purple Prose	confusion clouded his understanding like the haze hanging over the park	he remained confused
Food Writing		It's the democratization of ingredients that sets the bánh mì apart	It's the variety of ingredients that sets the bánh mì apart
Internet Advice		But pain has a way of cracking us open, making space for new possibilities we couldn't have imagined before.	But pain can open us up to new possibilities.
Travel Writing	Poor Sentence Structure	He tips generously but modestly, as if acknowledging a pact of respect	He tips generously yet with restraint, honoring a mutual respect
Literary Fiction		The gulf between her vision for her future and her parents' hopes wasn't something that could be bridged with words alone, and Bella realized that sometimes growth requires difficult choices.	Bella realized that her vision for the future and her parents' hopes were irreconcilable. Growth, she understood, often demands difficult choices.
Internet Advice	Unnecessary Exposition	, a silent reminder of everything it once meant	-
Personal Essay		Now, when life deals its inevitable blows, I think of those stones. I remember that sometimes it's the cracks that let the light in, and that even the most jagged edges can be smoothed by time and patience.	Now, when life deals its inevitable blows, I think of those stones.
Travel Writing	Cliché	As the weekend comes to a close, reflect on the kaleidoscope of experiences you've had, and how each neighborhood has contributed to the vibrant tapestry that is Los Angeles.	As the weekend comes to a close, you'll have a new appreciation for the city's diversity.
Food Writing		a dish that feels both familiar and unexpected, much like discovering a hidden alley in a well-known city. It's	a familiar dish with a surprising twist. It's
Literary Fiction		, watching his chest rise and fall with labored breaths	-
		I noticed a vase of wilting flowers on the nightstand, forgotten in the weight of more pressing concerns	The flowers on the nightstand had wilted

Table 13: LLM edited spans from different categories across different sub-genres of writing

that require deeper emotional resonance or cultural commentary and the importance of rethinking the design of alignment for tasks with subjective or fuzzy rewards.

7 DISCUSSION

7.1 How is editing human writing different from LLM-generated text?

Editing human writing and LLM-generated text presents distinct challenges and requires different approaches. Human writing often contains nuanced expressions, personal style, and contextual references that reflect the author's unique voice and experiences. Editors must preserve these elements while refining clarity, structure, and coherence. In contrast, LLM-generated text may lack consistent tone and exhibit repetitive patterns (Section 5). We asked writers to explain the differences in editing LLM-generated text compared to Human-written text. Several writers mentioned that LLM writing often required more extensive editing, mainly to remove unusual and sometimes nonsensical metaphors, inappropriate use of complex vocabulary that doesn't fit the context, and improving an

overall tone that comes across as impersonal and mechanical. In an exchange, W3 noted "*I edit a lot of prose for my magazine but one thing that stuck with me as I was editing these paragraphs are the massive amount of cliché, histrionic descriptions, and direct exposition of intended meanings rather than effective representation. Indeed very strange*". They observed that the types of edits needed for LLM-generated text were often similar, but the sheer volume of necessary changes was higher than that of human writing. LLM-generated content's repetitive nature paradoxically made the human editors feel robotic while trying to improve it. By the edits that catered to deep emotional resonance or cultural commentary,

7.2 How well can LLMs mimic edits from writers?

Our preference ranking results in Section 6.3 indicate that automatically edited paragraphs frequently rank second and sometimes first. This raises questions about LLMs' ability to analyze textual patterns and generate content closely resembling a given writer's edit. For the span "Janet lay in bed each night, her mind a whirlpool of restless thoughts", both LLM and writer identified it as cliché. The LLM edited it to "Janet lay in bed each night, unable to sleep" while the writer changed it to "Each night, Janet lay prone in her bed and unable to sleep". LLMs can also split run-on sentences and improve poor structure. For "Sarah froze, realizing it was her high school sweetheart, Alex, whom she hadn't seen in over a decade", the LLM edited it to "Sarah froze. It was Alex, her high school sweetheart. She hadn't seen him in over a decade", similar to the writer's edit. However, LLMs sometimes replace clichés with other clichés or fail to remove unnecessary exposition. The most challenging edit category is Lack of Specificity and Detail where LLMs often fail to add engaging details. For "Her irritation slowly morphed into a strange, disconnected calm", the writer added "After all, the noise just meant that she wasn't the only one awake at this hour." The model's edit was less effective: "Her irritation slowly morphed into a strange, disconnected calm. The repetitive thump-thump-thump became almost hypnotic, lulling her into a trance-like state". Much evocative detail in human writing comes from lived sensory experiences. LLMs on the other hand lack the grounded understanding that helps humans select vivid, emotionally resonant details. Additionally, when we combine autoregressive objectives with post-hoc adjustment through RLHF (typically designed to prevent toxic/harmful text generations) models often default to common, generic descriptions rather than specific ones since those are "safer" predictions. However, we note one potential limitation in our experiments is our reliance on few-shot instructions, requiring the model to learn rewriting from only a few examples. Training on the entire LAMP Corpus or more data might improve edit quality.

7.3 What recommendations can we provide for future LLM-based writing support tools that aspire to improve the co-writing experience?

Eminent author Curtis Sittenfeld calls LLM writing the literary equivalent of fat-free cookies [12]. LLMs are proficient at producing sentences that are grammatically correct and devoid of spelling errors. Beyond that, LLMs require extensive learning to effectively assist humans in improving their writing. In his essay *Politics and English Language* [85], George Orwell said "Never use a long word when a short one will do." LLM writing transgresses this simple rule by overusing lofty words. Clichés are bound to slip into even the best human writing, but when it comes to LLMs it simply cannot write without them. We believe this is partially a drawback of the technology behind LLMs. When an LLM calculates the probability of one word following another, clichés become very likely, because they've appeared so many times before. This explains why every other generated response is rife with clichés despite our prompt explicitly asking LLMs to avoid clichés and overused tropes (Table 14). LLMs need to learn how to identify and write without clichés such that it is engaging to every single reader. Overwriting is a bigger problem than underwriting. The rule for most writers is, "If in doubt, cut it." [46] The Pulitzer Prize-winning writer John McPhee has called the process "writing by omission." [75]. To become a better writer LLMs need to learn how to avoid unnecessary exposition. Last but not least, structure is what good writing hangs on [46]. Long, run-on sentences are hard to read, and LLMs need to know when and how to split effectively to better manage flow and clarity.

7.4 What are the potential long-term effects on language evolution and writing styles as LLM becomes more prevalent and how can aligned editing tools help?

The increasing prevalence of large language models (LLMs) could significantly impact language evolution and writing styles over time. There's potential for more homogenized writing as people rely on LLM-generated content, possibly leading to a reduction in linguistic diversity and individual voice [105]. However, well-designed editing tools aligned with expert writing practices could help counteract these effects. Such tools could encourage more nuanced and sophisticated language use, preserve stylistic diversity, and promote critical thinking about word choice and sentence structure. By highlighting elements of expert writing, these tools could elevate overall writing quality while still allowing for personal expression, potentially steering language evolution towards greater clarity, precision, and effectiveness in communication [68].

7.5 Why are there no significant differences in perceived writing quality or types of edits needed across texts generated by different large language models?

The lack of significant quality differences between LLM writing is somewhat unexpected and warrants deeper examination. Several factors however may explain this surprising finding. Books play a crucial role in the training of generative AI systems. Their long, thematically consistent paragraphs provide information about how to generate coherent and fluent text. All three models (GPT-4, Claude 3.5, Llama 3.1) are pre-trained on Books3corpus [54] which constitute a bigger portion of Pile [41] pre-training data. Additionally, all LLMs rely on Scale AI for collecting preference data and there is

[12] https://www.nytimes.com/2024/08/20/opinion/beach-read-ai.html

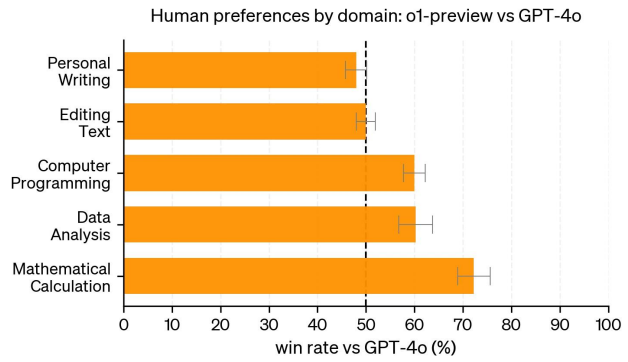


Figure 9: Preference Evaluation Results between GPT4o and GPT4-o1

very likely a significant overlap in workers who annotate the preferences as recently discussed [107]. Creative writing may also present an inherent "quality ceiling" for current LLM architectures, where all models encounter similar limitations in maintaining a consistent narrative voice and avoiding clichés—a pattern also observed by [112] in their analysis of narrative generation. Our syntactic pattern analysis (Section 5.3) reveals remarkably similar templates and phrasings across models, further supporting the hypothesis of shared training foundations. While our methodology focused on concrete, actionable edit categories, we acknowledge that more subtle stylistic differences or subjective "vibes" may exist between models not captured by our current taxonomy [31]. The consistency in writing quality across models ultimately points to broader questions about the current limitations of LLMs in creative writing tasks, rather than differences between specific implementations. Recent preference evaluation results from OpenAI's GPT4-o1 technical report [84] (See Figure 9) corroborate our findings where there are no significant differences in GPT4o vs GPT4o1 for personal writing or editing text.

8 Limitations

While our study provides valuable insights into improving LLM-generated text through expert editing, there are several limitations to consider. Our study was conducted with 18 MFA-trained creative writers. While this ensured a high level of expertise, it may limit the generalizability of our findings. Future research could expand the participant pool to encompass diverse cultural backgrounds and writing traditions. The editing data primarily comes from literary fiction and creative non-fiction, making the identified idiosyncrasies and editing strategies potentially less applicable to other genres like technical writing, journalism, or scientific writing. Future work can expand on the line of work by including a broader range of writing styles and purposes. The selected LLMs (GPT-4, Claude 3.5 Sonnet, and Llama 3.1) are among the most advanced models, but they might not fully represent the entire spectrum of AI writing abilities. It should be noted that the evaluation of writing quality is inherently subjective, even with multiple annotators and inter-annotator agreement calculations. Experts may disagree on what constitutes an improvement, potentially influencing our results and their interpretation.Our automated methods for detecting and rewriting problematic spans relied on few-shot learning with a limited number of examples. While this approach showed promise, it may not fully capture the complexity and nuance of expert editing and training a model on the entire LAMP Corpus or additional data is required.

It should be noted that while paragraph-level editing provides a balance between granularity and context, it may miss broader structural or thematic issues that become apparent only when considering longer pieces of writing. Last but not least, our study relied on professional writers editing AI-generated text for monetary compensation, which may have influenced the quality and nature of the edits. Editing one's own work typically involves more personal investment than editing text for pay, potentially leading to less motivation for substantial improvements [13]. Additionally, the repetitive nature of editing multiple AI-generated paragraphs could lead to fatigue, especially if the content is perceived as uninteresting or lacking in creativity. This fatigue could result in less thorough or thoughtful edits as the task progresses. Finally, another potential limitation of LLM-based editing is the risk of hallucinated and factually inconsistent information being introduced during the editing process. We did not study hallucinations as part of our work, due to the chosen domain focus of fictional writing which has less stringent factuality requirements, yet prior work has documented that LLM-based text editing can introduce factual errors [61].

9 Conclusion

In this work, we present a comprehensive approach to mitigating idiosyncrasies and improving human-AI alignment in the writing process through expert editing. We i) develop a taxonomy of edit categories grounded in established writing practices, ii) create the LAMP corpus containing over 8,000 fine-grained edits by professional writers on LLM-generated text, and iii) design methods for automatic detection and rewriting of problematic spans. Our analysis reveals several key findings. Professional writers identify consistent categories of edits needed to improve AI writing. Surprisingly, there are no significant differences in perceived writing quality or types of edits needed across texts generated by different large language models (GPT-4, Claude 3.5, Llama 3.1). Automated methods using few-shot prompting can detect and rewrite problematic spans in LLM-generated text, though far from matching human expert performance. Finally, in terms of preference evaluations, writers consistently rank text edited by other writers highest, followed by LLM-edited text, with unedited LLM-generated text ranking lowest. As AI text generation becomes more prevalent, developing robust editing and alignment techniques will be crucial to ensure AI systems produce high-quality writing that meets human standards and enhances creativity and linguistic diversity.

References

[1] 2018. Purple prose. https://en.wikipedia.org/wiki/Purple_prose
[2] Teresa M Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology* 43, 5 (1982), 997.

[3] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C&C '24)*. Association for Computing Machinery, New York, NY, USA, 413–425. https://doi.org/10.1145/3635636.3656204

[4] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*. 413–425.

[5] Anthropic. 2024. Claude-3.5-Sonnet. (2024). https://www.anthropic.com/news/claude-3-5-sonnet

[6] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 128–138. https://doi.org/10.1145/3377325.3377523

[7] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (jul 2015), 85–94. https://doi.org/10.1145/2791285

[8] John R Bormuth. 1966. Readability: A new approach. *Reading research quarterly* (1966), 79–132.

[9] Robert J Bracewell et al. 1978. The development of audience awareness in writing. (1978).

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[11] Janet Burroway, Elizabeth Stuckey-French, and Ned Stuckey-French. 2019. *Writing fiction: A guide to narrative craft*. University of Chicago Press.

[12] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[13] Kris Byron and Shalini Khazanchi. 2012. Rewards and creative performance: a meta-analytic test of theoretically derived hypotheses. *Psychological bulletin* 138, 4 (2012), 809.

[14] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).

[15] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or Artifice? Large Language Models and the False Promise of Creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 30, 34 pages. https://doi.org/10.1145/3613904.3642731

[16] Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity Support in the Age of Large Language Models: An Empirical Study Involving Professional Writers. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C &C '24)*. Association for Computing Machinery, New York, NY, USA, 132–155. https://doi.org/10.1145/3635636.3656201

[17] Anton Chekhov. 1984. *The selected letters of Anton Chekhov*. Macmillan.

[18] Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making Sentences Stand-Alone. *Transactions of the Association for Computational Linguistics* 9 (2021), 447–461. https://doi.org/10.1162/tacl_a_00377

[19] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[20] Gregory Currie. 1990. *The nature of fiction*. Cambridge University Press.

[21] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 98, 13 pages. https://doi.org/10.1145/3526113.3545672

[22] Johannes Daxenberger and Iryna Gurevych. 2013. Automatically Classifying Edit Categories in Wikipedia Revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (Eds.). Association for Computational Linguistics, Seattle, Washington, USA, 578–589. https://aclanthology.org/D13-1055

[23] Marshall L Dermer, Shannon L Lopez, and Paul A Messling. 2009. Fluency training a writing skill: Editing for concision. *The Psychological Record* 59 (2009), 3–20.

[24] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1044, 18 pages. https://doi.org/10.1145/3613904.3642134

[25] Karel D'Oosterlinck, Winnie Xu, Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri. 2024. Anchored Preference Optimization and Contrastive Revisions: Addressing Underspecification in Alignment. *arXiv preprint arXiv:2408.06266* (2024).

[26] Mark Doty. 2014. *The art of description: World into word*. Graywolf Press.

[27] Yao Dou, Philippe Laban, Claire Gardent, and Wei Xu. 2023. Automatic and Human-AI Interactive Text Generation. *arXiv preprint arXiv:2310.03878* (2023).

[28] Yao Dou, Philippe Laban, Claire Gardent, and Wei Xu. 2024. Automatic and Human-AI Interactive Text Generation (with a focus on Text Simplification and Revision). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, Luis Chiruzzo, Hung-yi Lee, and Leonardo Ribeiro (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3–4. https://doi.org/10.18653/v1/2024.acl-tutorials.2

[29] Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding Iterative Revision from Human-Written Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3573–3590. https://doi.org/10.18653/v1/2022.acl-long.250

[30] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[31] Lisa Dunlap, Krishna Mandal, Trevor Darrell, Jacob Steinhardt, and Joseph E Gonzalez. 2024. VibeCheck: Discover and Quantify Qualitative Differences in Large Language Models. *arXiv preprint arXiv:2410.12851* (2024).

[32] Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, and Fabio Petroni. 2022. Editeval: An instruction-based benchmark for text improvements. *arXiv preprint arXiv:2209.13331* (2022).

[33] Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College Composition & Communication* 32, 4 (1981), 400–414.

[34] Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text Editing by Command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 5259–5274. https://doi.org/10.18653/v1/2021.naacl-main.414

[35] Andy P Field. 2005. Kendall's coefficient of concordance. *Encyclopedia of statistics in behavioral science* 2 (2005), 1010–11.

[36] David Fishelov. 1990. Types of character, characteristics of types. *Style* (1990), 422–439.

[37] Kristie S Fleckenstein. 1991. Inner Sight: Imagery and Emotion in Writing Engagement. *Teaching English in the Two-Year College* 18, 3 (1991), 210–16.

[38] Nigel Fountain. 2012. *Clichés: Avoid them like the plague*. Michael O'Mara Books.

[39] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. 2024. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244* (2024).

[40] Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning llm agents by learning latent preference from user edits. *arXiv preprint arXiv:2404.15269* (2024).

[41] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).

[42] Yang Gao, Dana Alon, and Donald Metzler. 2024. Impact of preference noise on the alignment performance of generative language models. *arXiv preprint arXiv:2404.09824* (2024).

[43] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1002–1019.

[44] Sayan Ghosh, Tejas Srinivasan, and Swabha Swayamdipta. 2024. Compare without Despair: Reliable Preference Evaluation with Generation Separability. *arXiv preprint arXiv:2407.01878* (2024).

[45] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. 33–41.

[46] Harry Guinness. 2024. How to Edit Your Own Writing. (2024). https://www.nytimes.com/2020/04/07/smarter-living/how-to-edit-your-own-writing.html

[47] John R Hayes, Linda Flower, Karen A Schriver, James Stratman, Linda Carey, et al. 1987. Cognitive processes in revision. *Advances in applied psycholinguistics* 2 (1987), 176–240.

[48] Laraine Herring. 2007. *Writing Begins with the Breath: Embodying Authentic Voice*. Shambhala Publications.

[49] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190.

[50] Daphne Ippolito, Liam Dugan, Emily Reif, Ann Yuan, Andy Coenen, and Chris Callison-Burch. 2022. The Case for a Single Model that can Both Generate Continuations and Fill-in-the-Blank. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2421–2432. https://doi.org/10.18653/v1/2022.findings-naacl.185

[51] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers. *arXiv preprint arXiv:2211.05030* (2022).

[52] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3544548.3581196

[53] 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*. 9420–9435.

[54] Makena Kelly. 2023. *Books3 is the Internet's Most Controversial AI Dataset*. https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/ Accessed: 2024-11-18.

[55] Zae Myung Kim, Wanyu Du, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Improving Iterative Text Revision by Learning Where to Edit from Other Revision Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9986–9999. https://doi.org/10.18653/v1/2022.emnlp-main.678

[56] Stephen King. 2000. *On writing: A memoir of the craft*. Simon and Schuster.

[57] George R Klare. 1974. Assessing readability. *Reading research quarterly* (1974), 62–102.

[58] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118.

[59] Dmitry Kobak, Rita González Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into ChatGPT usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016* (2024).

[60] Anežka Kuzmičová. 2014. Literary narrative and mental imagery: A view from embodied cognition. *Style* 48, 3 (2014), 275–293.

[61] Philippe Laban, Jesse Vig, Marti A Hearst, Caiming Xiong, and Chien-Sheng Wu. 2023. Beyond the chat: Executable and verifiable text-editing with llms. *arXiv preprint arXiv:2309.15337* (2023).

[62] Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A Dataset for Document-Level Simplification of Wikipedia Pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10674–10695. https://doi.org/10.18653/v1/2023.acl-long.596

[63] Janet Lane and Ellen Lange. 1999. *Writing clearly: An editing guide*. Heinle & Heinle Boston, MA.

[64] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–35.

[65] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[66] Vladimir I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet physics. Doklady* 10 (1966), 707–710.

[67] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259* (2023).

[68] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1048, 25 pages. https://doi.org/10.1145/3613904.3642625

[69] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[70] Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony Liu, and Soroush Vosoughi. 2022. Second thoughts are best: Learning to re-align with human values from text edits. *Advances in Neural Information Processing Systems* 35 (2022), 181–196.

[71] Susan Peck MacDonald. 1986. Specificity in context: some difficulties for the inexperienced writer. *College Composition & Communication* 37, 2 (1986), 178–194.

[72] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36 (2024).

[73] Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2022. LENS: A learnable evaluation metric for text simplification. *arXiv preprint arXiv:2212.09739* (2022).

[74] Guillermo Marco, Julio Gonzalo, Ramón del Castillo, and María Teresa Mateo Girona. 2024. Pron vs Prompt: Can Large Language Models already Challenge a World-Class Fiction Author at Creative Text Writing? *arXiv preprint arXiv:2407.01119* (2024).

[75] John McPhee. 2015. Omission. *The New Yorker* (14 September 2015). https://www.newyorker.com/magazine/2015/09/14/omission Accessed: [Insert access date here].

[76] Bonnie JF Meyer. 2003. Text coherence and readability. *Topics in language Disorders* 23, 3 (2003), 204–224.

[77] Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A Robot Walks into a Bar: Can Language Models Serve as Creativity SupportTools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1622–1636.

[78] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 355, 34 pages. https://doi.org/10.1145/3544548.3581225

[79] Marc Moens. 1987. Tense, aspect and temporal reference. (1987).

[80] Michael Nebeling, Alexandra To, Anhong Guo, Adrian A. de Freitas, Jaime Teevan, Steven P. Dow, and Jeffrey P. Bigham. 2016. WearWrite: Crowd-Assisted Writing from Smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3834–3846. https://doi.org/10.1145/2858036.2858169

[81] Nancy Nelson and Montserrat Castelló. 2012. Academic writing and authorial voice. In *University writing: Selves and texts in academic societies*. Brill, 33–51.

[82] William Noble. 2013. *Show Don't Tell: A Writer's Guide*. Vol. 6. The Write Thought, Inc.

[83] OpenAI. 2024. GPT-4o. (2024). https://openai.com/index/hello-gpt-4o/

[84] OpenAI. 2024. *Learning to Reason with LLMs*. Technical Report. OpenAI. https://openai.com/index/learning-to-reason-with-llms/

[85] George Orwell. 2013. *Politics and the English language*. Penguin UK.

[86] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[87] Vishakh Padmakumar and He He. 2023. Does Writing with Language Models Reduce Content Diversity? *arXiv preprint arXiv:2309.05196* (2023).

[88] Jane Pan, He He, Samuel R Bowman, and Shi Feng. 2024. Spontaneous Reward Hacking in Iterative Self-Refinement. *arXiv preprint arXiv:2407.04549* (2024).

[89] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics* 12 (2024), 484–506.

[90] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[91] So Yeon Park and Sang Won Lee. 2023. Why "why"? The Importance of Communicating Rationales for Edits in Collaborative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 616, 25 pages. https://doi.org/10.1145/3544548.3581345

[92] Roy D Pea and D Midian Kurland. 1987. Chapter 7: Cognitive Technologies for Writing. *Review of research in education* 14, 1 (1987), 277–326.

[93] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 34. 480–489.

[94] Luqman M Rababah. 2022. Contextualization to enhance students' writing ability. *Theory and Practice in Language Studies* 12, 11 (2022), 2316–2321.

[95] Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text Editing by Task-Specific Instruction Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5274–5291. https://doi.org/10.18653/v1/2023.findings-emnlp.350

[96] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[97] Machel Reid and Graham Neubig. 2022. Learning to Model Editing Processes. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3822–3832. https://doi.org/10.18653/v1/2022.findings-emnlp.280

[98] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. ABScribe: Rapid Exploration & Organization of Multiple Writing Variations in Human-AI Co-Writing Tasks using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1042, 18 pages. https://doi.org/10.1145/3613904.3641899

[99] Scott P. Robertson and John B. Black. 1983. Planning units in text editing behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '83)*. Association for Computing Machinery, New York, NY, USA, 217–221. https://doi.org/10.1145/800045.801614

[100] Mary Beth Rosson. 1983. Patterns of experience in text editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '83)*. Association for Computing Machinery, New York, NY, USA, 171–175. https://doi.org/10.1145/800045.801604

[101] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076* (2023).

[102] Marlene Scardamalia. 1983. The development of evaluative, diagnostic, and remedial capabilities in children's composing. *The psychology of written language: A developmental approach* (1983), 67–95.

[103] Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and Measurement of Syntactic Templates in Generated Text. *arXiv preprint arXiv:2407.00211* (2024).

[104] Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: An instruction-tuned large language model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18970–18980.

[105] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631, 8022 (2024), 755–759.

[106] Nancy Sommers. 1980. Revision strategies of student writers and experienced adult writers. *College Composition & Communication* 31, 4 (1980), 378–388.

[107] Wall Street Journal Staff. 2024. Alexandr Wang: Scale AI. *The Wall Street Journal* (2024). https://www.wsj.com/tech/ai/alexandr-wang-scale-ai-d7c6efd7 Accessed: 2024-11-18.

[108] Melanie Subbiah, Sean Zhang, Lydia B Chilton, and Kathleen McKeown. 2024. Reading Subtext: Evaluating Large Language Models on Short Story Summarization with Writers. *arXiv preprint arXiv:2403.01061* (2024).

[109] G Thomas Tanselle. 1976. The editorial problem of final authorial intention. *Studies in Bibliography* 29 (1976), 167–211.

[110] Jaime Teevan, Shamsi T. Iqbal, and Curtis von Veh. 2016. Supporting Collaborative Writing with Microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2657–2668. https://doi.org/10.1145/2858036.2858108

[111] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[112] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are Large Language Models Capable of Generating Human-Level Narratives? *arXiv preprint arXiv:2407.13248* (2024).

[113] Sherman W. Tyler, Steven Roth, and Timothy Post. 1982. The acquisition of text editing skills. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems* (Gaithersburg, Maryland, USA) *(CHI '82)*. Association for Computing Machinery, New York, NY, USA, 324–325. https://doi.org/10.1145/800049.801803

[114] Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, and Greg Durrett. 2024. Learning to Refine with Fine-Grained Natural Language Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 12281–12308.

[115] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying Semantic Edit Intentions from Revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2000–2010. https://doi.org/10.18653/v1/D17-1213

[116] Zonghai Yao, Benjamin Schloss, and Sai Selvaraj. 2023. Improving Summarization with Human Edits. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2604–2620.

[117] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 841–852.

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

[119] David Zhou and Sarah Sterman. 2024. Ai.llude: Investigating Rewriting AI-Generated Text to Support Creative Expression. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C & C '24)*. Association for Computing Machinery, New York, NY, USA, 241–254. https://doi.org/10.1145/3635636.3656187

[120] Yilun Zhou, Caiming Xiong, Silvio Savarese, and Chien-Sheng Wu. 2024. Shared Imagination: LLMs Hallucinate Alike. *arXiv preprint arXiv:2407.16604* (2024).

[121] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

A Appendix

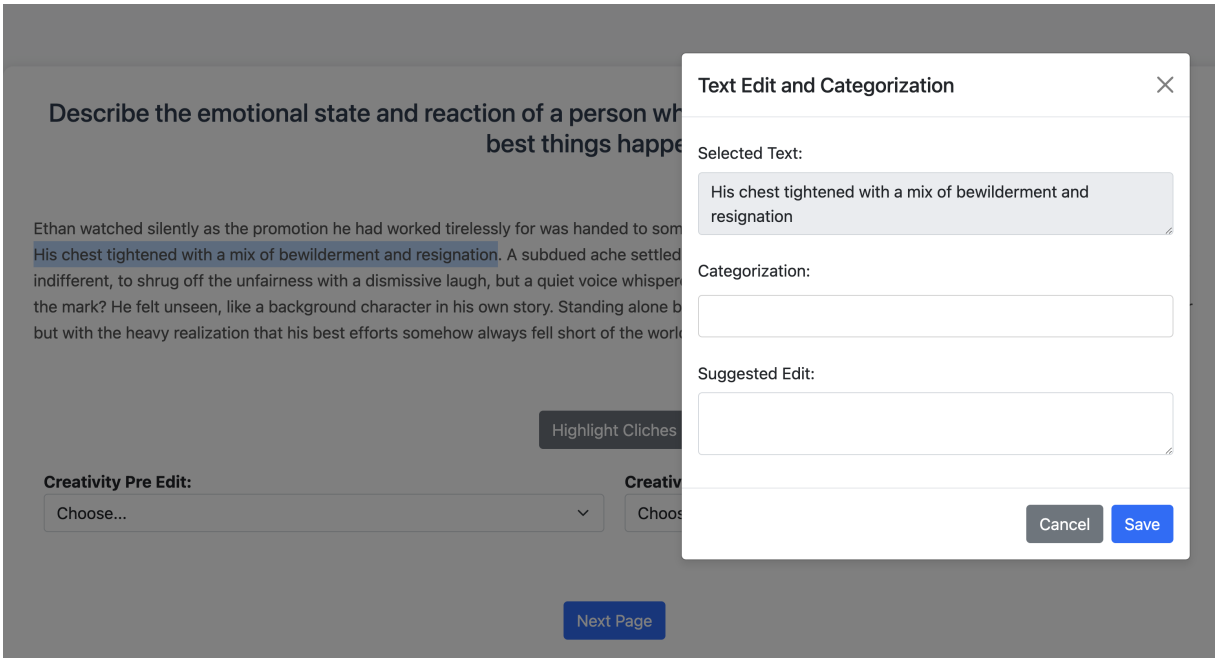


Figure 10: Interface for formative study to collect fine-grained labels

Instruction Prompt	Summarize this paragraph into a single sentence open-ended question.\n **{{paragraph}}**
	Summarize this paragraph into a single sentence open-ended instruction.\n **{{paragraph}}**
Response Prompt	Imagine you are a fiction writer for the NewYorker. Now write a paragraph (10-15 sentence) as a response to the following question. Try your best to be original, avoiding clichés or overused tropes. Do not use ornamental language and focus on nuance, simplicity, and subtext. Start directly with your response. \n **{{instruction}}**
	Imagine you are a writer for the New York Times Modern Love section. Now write a paragraph (10-15 sentence) as a response to the following question. Try your best to be original, avoiding clichés or overused tropes. Do not use ornamental language and focus on nuance, simplicity, and subtext.Start directly with your response \n **{{instruction}}**
	Imagine you are a writer for the New York Times Cooking section. Now write a paragraph (10-15 sentence) as a response to the following question. Try your best to be original, avoiding clichés or overused tropes. Do not use ornamental language and focus on nuance, simplicity, and subtext.Start directly with your response \n **{{instruction}}**
	Imagine you are a writer for the New York Times Travel section. Now write a paragraph (10-15 sentence) as a response to the following question. Try your best to be original, avoiding clichés or overused tropes. Do not use ornamental language and focus on nuance, simplicity, and subtext.Start directly with your response \n **{{instruction}}**
	Imagine you are a beloved female Internet advice columnist whose trademark is deeply felt and frank responses grounded in your own personal experience. Now write a paragraph (10-15 sentence) as a response to the following question.Try your best to be original, avoiding clichés or overused tropes. Do not use ornamental language and focus on nuance, simplicity, and subtext.Start directly with your response \n **{{instruction}}**

Table 14: Prompts for generating instructions and responses

A.1 Idiosyncracy Span Detection Prompt

Idiosyncracy Span Detection Prompt

You are given a paragraph of writing, and your goal is to provide feedback by selecting spans of text in the writing that could be improved and assign each problematic span to an error category. Below, we list the 7 error categories that you can choose from.
You are also provided 2 examples of paragraphs that were annotated by professional writers, which you can use to better understand the task and the error categories.
Error Categories:
- "Awkward Word Choice and Phrasing": Suggestions for better word choices or more precise phrasing to enhance clarity and readability.
- "Cliche": The use of hackneyed phrases or overly common imagery that lacks originality or depth.
- "Poor Sentence Structure": Feedback on the construction of sentences, recommending changes for better flow, clarity, or impact.
- "Unnecessary/Redundant Exposition": Redundant or non-essential parts of the text that could be removed/rephrased for conciseness.
- "Lack of Specificity and Detail": Need for more concrete details or specific information to enrich the text and make it more engaging.
- "Purple Prose": Identifying parts of the text that are seen as unnecessary ornamental and overly verbose.
- "Tense Consistency": Comments pointing out inconsistencies in verb tense that need to be addressed for uniformity.

Example 1: Input Text

Output:
Example Output in JSON format.

Example 2:
(Similar to example 1)
Rules:
- Number of Spans – You can provide feedback on multiple spans, and multiple spans can have the same category.
- Span must be verbatim – The span you select must be verbatim from the paragraph, otherwise, the feedback will not be provided to the user.
- No Overlap – Spans should not overlap, and one span should not include the other.
- Single Category – Each span should have exactly one category from the categories listed above.

Paragraph:

PARAGRAPH

A.2 Precision Metrics Explanation and Example

We illustrate the General and Categorical Precision on a simple example and justify the choice of the metric.

Imagine we have the following sentence, that has been annotated by a human annotator: On this dark and stormy night, her heart skipped a beat as she was afraid of what was to come.

> ANNOTATION = On this **dark and stormy night**[CLICHÉ], her heart skipped a beat **as she was afraid of what was to come.**[UNNECESSARY EXPOSITION]

Now let's imagine that System 1 and System 2 have produced the following predictions:

> SYSTEM 1 = On this **dark and stormy night, her heart skipped a beat as she was afraid of what was to come.**[CLICHÉ]
>
> SYSTEM 2 = On this dark and **stormy night**[CLICHÉ], her heart skipped a beat **as she was afraid of what was to come.**[CLICHÉ]

We extract the annotated spans:

- Span 1: characters [9,30]; category: CLICHÉ
- Span 2: characters [57, 94]; category: UNNECESSARY EXPOSITION

System 1 produced a single span:

- Span 1: characters [9, 94]; category: CLICHÉ

System 2 produced two spans:

- Span 1: characters [19,30]; category: CLICHÉ
- Span 2: characters [57, 94]; category: CLICHÉ

We can first compute General Precision, which disregards the category of the spans. It is the overlap between predicted spans and annotated spans, divided by the total amount of predicted characters:

- General Precision (System 1) = ((30-9) + (94-57)) / (94-9) = 0.68
- General Precision (System 2) = ((30-19) + (94-57)) / ((30-19) + (94-57)) = 1.0

System 2 achieves a higher precision, as all the spans it predicted were included in the manual annotation. On the other hand, System 1 predicted a larger span that included the annotated span, but also additional characters, causing a lower precision score.

When consider Categorical Precision, overlap is only considered as valid if the overlapping spans coincide in category. The scores would be:

- Categorical Precision (System 1) = (1*(30-9) + 0*(94-57)) / (94-9) = 0.25
- Categorical Precision (System 2) = (1*(30-19) + 0*(94-57)) / ((30-19) + (94-57)) = 0.23

System 1 achieved higher categorical precision by fully overlapping with the annotated CLICHÉ span, while System 2 only partially overlapped. Both systems incorrectly categorized the second span, resulting in lower precision scores. Precision scores can be inflated by reducing predictions, but our LLMs weren't instructed to optimize for precision. In fact, they tend to select more spans than human annotators, leading to high recall but potentially lower precision. We focus on precision to penalize systems that produce too many or overly large spans.

A.3 Rewriting Prompts

A cliché is a saying, idea, or element of an artistic work that has become overused to the point of losing its original meaning or effect, even to the point of being weird, irritating, or bland

You will be given example of 25 paragraphs with spans that count as Cliche and suggested edits that either **REWRITES THE CLICHE or SIMPLY REMOVES IT**.

Your task will then be to suggest edits (either spans or empty string) that gets rid of the cliche while making the resulting paragraph coherent, given a new paragraph and highlighted span of Cliche from it. Do not simply paraphrase or use fancy ornamental language; Try to keep each sentence short. Look at the examples carefully

**IT IS VERY IMPORTANT TO MAKE SURE THAT YOUR EDITED TEXT ONCE ADDED TO THE PARAGRAPH READS COHERENTLY AND GRAMMATICALLY CORRECT.

For instance if you replace text within tags with a longer span; please make sure the following text after the edit, is its continuation. A simple way to ensure this is to ensure that the edited span has the same casing and punctuation at the beginning and end as that of the original span.

PLEASE FOLLOW THE OUTPUT SCHEMA AS THE EXAMPLES BELOW AND DO NOT RETURN ANYTHING OTHER THAN THE EDITED SPAN WITHIN QUOTES

Example 1

Paragraph: Matthews had lived in the Valley all his life, and its rhythms and secrets were etched into his being like the lines on a well-worn map. He knew [...]
Original Span: "like the lines on a well-worn map"
Edited Span: "like creases in an old pocket map"

.
.
.

Example 18

Paragraph: Husna sat at the ancient wooden [....] The room was a bubble of quiet concentration, the only sounds the clacking of the typewriter, the rustling of paper, and the occasional whistle of the teakettle in the adjoining kitchen.
Original Span: "The room was a bubble of quiet concentration, the only sounds the clacking of the typewriter, the rustling of paper, and the occasional whistle of the teakettle in the adjoining kitchen."
Edited Span: "The room was quiet. The outside world did not exist. At times, Husna tapped her foot. Shah Sahib coughed and she would stop. The typewriter never did."
.
.

Example 25

Paragraph: Last night, I dreamt of an [....] She didn't speak, but her eyes communicated a haunting mix of sadness and knowing, as if she heldthe weight of forgotten secrets. I felt a [...]
Original Span: "communicated a haunting mix of sadness and knowing, as if she held"
Edited Span: "conveyed"

Table 15: Prompt to rewrite Cliche

Poor sentence structure refers to writing that is difficult to understand or lacks clarity due to issues with how sentences are constructed. It encompasses issues like run-on sentences, fragments, misplaced or dangling modifiers, lack of variety, overuse of passive voice, improper parallelism, and unclear pronoun references, all of which impede clear communication and reader comprehension

You will be given examples of 25 paragraphs with text within tags that shows poor sentence structure and suggested edits that either **REWRITES WITH IMPROVED SENTENCE STRUCTURE**.

Your task will then be to suggest edits that rewrite the text within the span tags with better sentence structure while making the resulting paragraph coherent, given a new paragraph and highlighted span of poor sentence structure from it. Do not use fancy ornamental language; Look at the examples carefully and do not output anything after closing quotes.

**IT IS VERY IMPORTANT TO MAKE SURE THAT YOUR EDITED TEXT ONCE ADDED TO THE PARAGRAPH READS COHERENTLY AND GRAMMATICALLY CORRECT. For instance, if you replace text within tags with a longer span; please make sure the following text after the edit, is its continuation.

PLEASE FOLLOW THE OUTPUT SCHEMA AS THE EXAMPLES BELOW AND DO NOT RETURN ANYTHING OTHER THAN THE EDITED SPAN WITHIN QUOTES
.
.
.
Example 4

Paragraph: As the night wore on, Z.'s laughter grew louder, his words slurring together like a sloppy melody. N. and I exchanged a knowing glance, our concern simmering beneath the surface.At first, it was just a slight stumble, a misstep that could be brushed off as a joke. [.....]

Original Span: "As the night wore on, Z.'s laughter grew louder, his words slurring together like a sloppy melody. N. and I exchanged a knowing glance, our concern simmering beneath the surface."
Edited Span: "Z. was drinking more and more as the night went on. He laughed more loudly. His words started to slur, blurring one into the next. I looked at N., who knew what I was thinking. We were going to have to take care of him.".
.
.

Example 13

Paragraph: As I step into the quiet, garden-facing room on the second floor, I'm struck by the sense of stillness that pervades the space. The occupants, an elderly couple, sit motionless in their armchairs, their [....]
Original Span: "As I step into the quiet, garden-facing room on the second floor, I'm struck by the sense of stillness that pervades the space"
Edited Span: "A sense of stillness pervades the garden-facing room on the second floor" .
.
.

Example 25

Paragraph: Chef Amelia raced [.....] She plastered on a polite smile, determined not to let her personal history interfere with her professional duties.As Daniel approached, plate in hand, Amelia steeled herself [.....]
Original Span: "She plastered on a polite smile, determined not to let her personal history interfere with her professional duties."
Edited Span: "She shot a dutiful smile for anyone who was looking. This was an important night, and she wasn't going to let the past get in the way of a job well done."

Table 16: Prompt to rewrite Poor Sentence Structure

Unnecessary or redundant exposition in writing refers to providing excessive explanatory information that doesn't contribute meaningfully to the story, characters, or overall narrative.

You will be given example of 25 paragraphs with text within tags that count as unnecessary/redundant exposition and suggested edits that either **REWRITES IT IN FEWER WORDS or SIMPLY REMOVES IT**.

Your task will then be to suggest edits that rewrites the text within the span tags correcting the unnecessary /redundant exposition while making the resulting paragraph coherent, given a new paragraph and highlighted text within of unnecessary/redundant exposition. Do not simply paraphrase or use fancy ornamental language or repeat the same thing in the edited span; Look at the examples carefully.

**IT IS VERY IMPORTANT TO MAKE SURE THAT YOUR EDITED TEXT ONCE ADDED TO THE PARAGRAPH READS COHERENTLY AND GRAMMATICALLY CORRECT.

For instance if you replace text within tags with a shorter span; please make sure the following text after the edit, is its continuation. Simple way to ensure this is to make sure that the edited span has the same casing and/or punctuation at the beginning and end as that of the original span.

PLEASE FOLLOW THE OUTPUT SCHEMA AS THE EXAMPLES BELOW AND DO NOT RETURN ANYTHING OTHER THAN THE EDITED SPAN WITHIN QUOTES

.
,
Example 2
Paragraph: In spring, when the first buds unfurled [...] embrace of varenyky dinners provided comfort against the chill , each bite narrating a history of resilience and hope. It was through [...]
Original Span: ", each bite narrating a history of resilience and hope"
Edited Span: ""
.
.
.

Example 18

Paragraph: As Oghi watched his mother-in-law, Mrs. Kim, he felt a subtle sense of unease settle in the pit of his stomach.It wasn't just the uncharacteristic behavior itself - [...]
Original Span: "As Oghi watched his mother-in-law, Mrs. Kim, he felt a subtle sense of unease settle in the pit of his stomach."
Edited Span: "Oghi watched his mother-in-law Mrs. Kim with heightening unease."

.
.

Example 23

Paragraph: The small room [....] They teased and corrected each other's recollections , creating a tapestry of resilience and camaraderie.It wasn't all smooth-sharp words resurfaced around old wound, [....]
Original Span: ", creating a tapestry of resilience and camaraderie"
Edited Span: ""

.
.

Table 17: Prompt to rewrite Unnecessary or redundant exposition

Lack of Specificity and Detail in writing refers to the absence of concrete and specific information, which can make the text feel vague and unengaging. The need for more concrete details or specific information is crucial to enrich the text and make it more engaging. Specificity helps to create vivid imagery, provides clarity, and connects with the reader on a deeper level. doesn't contribute meaningfully to the story, characters, or overall narrative.

You will be given example of 25 paragraphs with text within tags that lacks specificity and detail and suggested edits that either **REWRITES WITH SPECIFICITY AND DETAIL**.

Your task will then be to suggest edits that rewrites the text within the span tags with specificity and detail that is engaging while making the resulting paragraph coherent, given a new paragraph and highlighted span of lack of specificity and detail from it. Do not simply paraphrase or use fancy ornamental language; Look at the examples carefully and do not output anything after closing quotes.

**IT IS VERY IMPORTANT TO MAKE SURE THAT YOUR EDITED TEXT ONCE ADDED TO THE PARAGRAPH READS COHERENTLY AND GRAMMATICALLY CORRECT. For instance if you replace text within tags with a longer span; please make sure the following text after the edit, is its continuation. Simple way to ensure this is to make sure that the edited span has the same casing and punctuation at the beginning and end as that of the original span.

PLEASE FOLLOW THE OUTPUT SCHEMA AS THE EXAMPLES BELOW AND DO NOT RETURN ANYTHING OTHER THAN THE EDITED SPAN WITHIN QUOTES

Example 1
Paragraph: Sarah Mitchum's marriage appeared outwardly conventional, but subtle tensions simmered beneath the surface. She and [.....] leaving Sarah feeling increasingly isolated within her own marriage.
Original Span: "within her own marriage."
Edited Span: ".Their marriage had run its course. There was no coming back."
.
.
.
Example 15

Paragraph: Dr. Arthur Steiger's fall from grace began with a series of whispered concerns among his colleagues at Cormac General Hospital.The small-town pain specialist had always been known [....]
Original Span: "Dr. Arthur Steiger's fall from grace began with a series of whispered concerns among his colleagues at Cormac General Hospital."
Edited Span: "Pain was Dr. Arthur Steiger's forte. Not inflicting it, that is, but resolving it. Whenever a patient had problem, whether a tear in atendon, a sprain, a knock, a headache, a broken bone– it was Dr. Steiger that knew what to do."

.
.
Example 21

Paragraph: Mila sat on her porch a week after the storm had hit, sipping lukewarm tea. [....] Each night it grew louder, shifting from a whisper to a groan, but she had dismissed it, too tired from long days at work. [.....]
Original Span: "it grew louder, shifting from a whisper to a groan, but she had dismissed it, too tired from long days at work"
Edited Span: "lying like blanched spinach in her IKEA bed, trying not to think about another day of writing emails with someone else's signature on them and pretending not to care what John Blanchett, CEO of Executive Industries thought of her blouse–in other words,another day as John's executive assistant–"

.
.

Table 18: Prompt to rewrite Lack of Specificity and Detail

In literary criticism, purple prose is overly ornate prose text that may disrupt a narrative flow by drawing undesirable attention to its own extravagant style of writing, thereby diminishing the appreciation of the prose overall. Purple prose is characterized by the excessive use of adjectives, adverbs, and metaphors.

You will be given example of 25 paragraphs with text within tags that has purple prose in it and suggested edits that either **REWRITES THEM WITH SIMPLER WORDS OR REMOVES IT**.

Your task will then be to suggest edits that rewrites the text within the span tags altering the purple prose while making the resulting paragraph coherent, given a new paragraph and highlighted span of purple prose from it. Do not simply paraphrase or use fancy ornamental language; Look at the examples carefully and do not output anything after closing quotes.

**IT IS VERY IMPORTANT TO MAKE SURE THAT YOUR EDITED TEXT ONCE ADDED TO THE PARAGRAPH READS COHERENTLY AND GRAMMATICALLY CORRECT. For instance if you replace text within tags with a longer span; please make sure the following text after the edit, is its continuation. Simple way to ensure this is to make sure that the edited span has the same casing and punctuation at the beginning and end as that of the original span.

PLEASE FOLLOW THE OUTPUT SCHEMA AS THE EXAMPLES BELOW AND DO NOT RETURN ANYTHING OTHER THAN THE EDITED SPAN WITHIN QUOTES

.
.
Example 2

Paragraph: Fruto never intended to stir anything beyond the melting pot of their weekly card game.But when the chatter turned to the dry monotony of their jobs, Fruto found himself blurting out, [....]
Original Span: "Fruto never intended to stir anything beyond the melting pot of their weekly card game."
Edited Span: "Fruto hadn't meant to disrupt the routine of their weekly card game."
.
.
.

Example 16

Paragraph: My mother cried, [....] All of it vanished, cycling back through her mind, not as numbers but memories of scraped knees she bandaged alone and birthdays where her absence was felt more acutely than her presence. The sobs emerged from this deep well of unspoken expectations, leaving behind a residue of weary resilience and a few hopeful echoes yet unwilling to completely extinguish.
Original Span: ", cycling back through her mind, not as numbers but memories of scraped knees she bandaged alone and birthdays where her absence was felt more acutely than her presence. The sobs emerged from this deep well of unspoken expectations, leaving behind a residue of weary resilience and a few hopeful echoes yet unwilling to completely extinguish."
Edited Span: "She cried. She cried deep from this well of scraped knees she bandaged alone and birthdays she missed to work. She cried for unfairness. She cried without relief." .
.

Example 24

Paragraph: As they navigated their final year of high school, Maya and Jake found themselves at a crossroads, their educational paths diverging like tributaries of a river.[....]
Original Span: "As they navigated their final year of high school, Maya and Jake found themselves at a crossroads, their educational paths diverging like tributaries of a river."
Edited Span: "The final year of high school was pulling Maya and Jake in different directions."
.
.

Table 19: Prompt to rewrite Purple Prose