How LLMs Distort Transform Our Language

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) are increasingly used as writing assistants for tasks such as revising text, generating suggestions, and improving clarity. However, it remains unclear whether these systems preserve users' writing style, tone, or even intended meaning when editing text. In this work, we examine how LLMs alter the semantic and stylistic properties of human writing. Using a dataset of human-written essays and their revisions from before the release of LLMs, we study how asking an LLM to revise the essay based on the human-written feedback induces large changes in the resulting content and meaning. We then conduct a randomized controlled user study to understand how humans actually interact with LLMs when using them for writing. Our study of 50 users reveals that those who use an LLM to assist them showed similar alterations to their writing, and reported that the resulting essay was significantly less creative and not in their voice. Finally, we study how LLM use is already affecting our institutions, such as scientific peer review, altering the criteria for publication and assigning scores. These findings highlight a misalignment between the perceived benefit of AI use and an insidious and consistent underlying semantic change, motivating future work on how widespread AI writing will affect our cultural and scientific institutions.

1 Introduction

Over 1 billion people use Large Language Models (LLMs) weekly, with much traffic devoted to getting help with writing (Chatterji et al., 2025), including revising emails and drafting workplace documents in professional settings (Liang et al., 2024; Sanz-Tejeda et al., 2026) and generating ideas for essays and creative writing (Doshi & Hauser, 2024). LLM-generated text has also rapidly permeated formal institutions, with recent reports highlighting the use of LLMs in drafting UK parliamentary speeches (James, 2025) and 25% of peer reviews likely generated by LLMs at a major academic conference (Emi, 2025). This increasing use of LLMs for writing raises questions about whether these systems are able to preserve a human's writing preferences, tone, stylistic voice, and even the content of what they write about. Prior work suggests that LLMs tend towards homogenization, converging on a narrow set of linguistic patterns across diverse prompts and contexts (Jiang et al., 2025b). When such models are used as writing assistants, this tendency may be amplified through repeated exposure to model-generated suggestions, potentially reducing text diversity and encouraging convergence toward a writing style (Hutchinson et al., 2025).

In this paper, we investigate the extent to which the use of LLMs for writing alters or distorts human writing, in both style and meaning. To understand the use of LLMs by humans for writing, we conduct a human user study with 50 participants, where 30 participants complete an essay-writing task with access to an LLM and the remaining 20 participants complete the same task without access to the LLM. We survey participant attitudes and preferences towards LLMs and writing before and after the study, and analyze stylistic differences among essays written solely by humans, essays written with LLM assistance, and essays generated entirely by an LLM. Our results show that participants who used LLMs report their essays to be significantly less creative and less reflective of their own voice, suggesting the inability of LLMs to meet human writing preferences.

Motivated by findings that roughly two-thirds of LLM-assisted writing involves editing, critiquing, or translating existing text rather than generating text from scratch (Chatterji et al., 2025), we additionally perform a larger-scale quantitative analysis of how LLMs edit human-written essays, compared to how humans edit their own essays. Specifically, we leverage a publicly available dataset of 86 human-written essays, expert feedback, and the resulting human-revised drafts (Kashefi et al.,

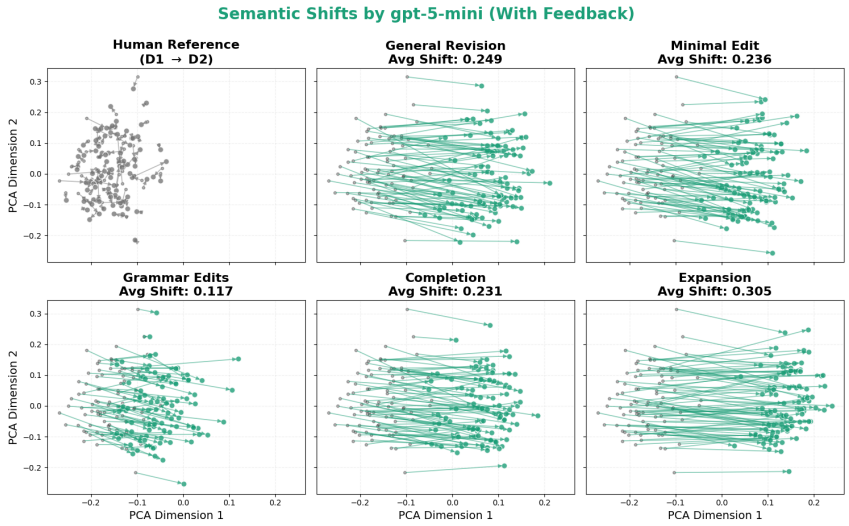


Figure 1: Semantic shifts induced by human and LLM revisions for the Arg Rewrite v2 dataset. Each point pair represents an essay before (D1) and after (D2) revision, embedded using `gemini-004` sentence embeddings and projected into two dimensions via PCA, a common approach for analyzing semantic differences (Dhillon et al., 2015). The left panel shows human revisions, while the remaining panels show revisions produced by prompting the LLM with different edit instructions with access to expert feedback, including prompting the LLM to make minimal edits (top right). Arrows indicate the direction and magnitude of semantic change. Human revisions exhibit smaller semantic shifts, whereas LLM revisions produce larger shifts that are strongly aligned in a common direction, indicating a homogenizing, meaning-altering distortion.

2022), comparing human edits to edits produced by three commonly used LLMs by humans. We measure differences between human-edited and LLM-edited essays along dimensions of semantics, lexical, part-of-speech distributions, emotional tone, and stylistic features. Our results show that using LLMs for editing leads to a dramatic shift from initial drafts and the homogenization of the resulting text. As shown in Figure 1, when humans revise their own writing, they perform edits of much smaller magnitude that take the writing in different directions. In contrast, when LLMs are prompted to edit the essays, they uniformly shift human writing in the same direction, resulting in essays that are not only less diverse but also occupy a region of embedding space where no previous human-written essay exists. This pattern is consistent with prior findings that LLMs encourage convergence towards a shared semantic style rather than preserving individual writing preferences (Jiang et al., 2025a). As our results will show, we find that LLMs increase the usage of both argumentative and emotional language, shifting the meaning to argue more positively for technology.

Finally, we analyze peer reviews from the International Conference on Learning Representations (ICLR) from 2022 to present (González-Márquez & Kobak, 2024), comparing those written by humans with those generated or heavily edited by LLMs. We summarize strengths and weaknesses in LLM-generated versus human-written reviews, and find that using LLMs for writing academic peer reviews does not just reduce the diversity of responses and change the resulting average scores. Rather, LLMs have begun to change what criteria we focus on in evaluating peer-reviewed scientific research in the wild. Thus, LLM use is already shifting our scientific and cultural institutions in insidious ways that are as of yet poorly understood.

In summary, the contributions of our paper are to study how LLMs use shifts in human writing, both in degree and in meaning, by leveraging for the first time:

- A randomized controlled trial studying the output of and preferences of humans as they use an LLM writing assistant under naturalistic conditions
- A comparison of human-edited essays with expert feedback, pre-dating LLMs, to LLMs making revisions on the same essays;
- Data from in-the-wild LLM use permeating scientific peer review in the International Conference for Learning Representations (ICLR) 2026, a significant machine learning venue.

Our findings all point to the same conclusion: if left unchecked, LLMs may inadvertently push people towards a vastly different, homogenized style of writing that does not reflect our preferences, and even subtly steer our writing and decision-making toward different conclusions than those originally

intended. We hope this work underscores the need for further study into these phenomena, as well as the development of writing tools that support clarity and correctness without altering meaning or reducing creative expression of the human voice.

2 Related Work

LLM Homogenization. Large language models have been adopted for use at a dizzying pace, with one of the most dominant usage categories being writing, editing, and generating ideas (Chatterji et al., 2025). However, co-writing with an LLM leads to writing that tends to converge stylistically across creative tasks, decision making, and open-ended text generation (Zhou & Fiedler, 2025; Wang et al., 2025; Agarwal et al., 2025; Jiang et al., 2025a; Xu et al., 2025) even across model families. Recent work provides evidence that this loss of diversity may arise from feedback loops between LLM models and training with paradigms such as RLHF (Murthy et al., 2025). Algorithmic systems can create cultural "lock-in," where the outputs of generative models reinforce their own stylistic priors over time (Hutchinson et al., 2025). These works indicate that homogenization is not a surface-level phenomenon, but a structural property of current LLMs. Our work furthers these claims by analyzing a dataset of writing from before the widespread adoption of LLMs (Chen et al., 2022). We not only show that LLM-prompted edits from a range of different model families consistently shift essays toward a narrower region of the distribution, in contrast to the more varied distribution of human edits, but that LLMs induce semantic, lexical, and emotional differences in writing (Mohammad & Turney, 2013; Boyd et al., 2022).

The Effect of LLMs on Humans. Recent work has begun to examine how sustained interaction with LLMs affects human cognition, preferences, and decision-making. Although LLMs have been shown to improve productivity by assisting in generating ideas, structuring content, summarizing literature, and reducing drafting time (Noy & Zhang, 2023), using LLMs for writing has been shown to reduce creativity in humans (ScienceDaily, 2024; Wang & Fan, 2025; Doshi & Hauser, 2023; Meincke et al., 2025; Anderson et al., 2024), with the resulting writing showing less diversity. Additionally, frequent reliance on AI assistance may carry cognitive costs, with humans who delegate writing to LLMs accumulating "cognitive debt," and producing fluent but less conceptually diverse work (Kim et al., 2025). In everyday communication tools, LLMs increasingly shape how people write as much as what they decide. Empirical studies show that AI-mediated communication alters emotional tone, partner perception, and decision making (Hohenstein et al., 2021; Sabour et al., 2025), diminishes authorship and agency in professional correspondence (Wenker, 2023), and, in cross-cultural contexts, homogenizes prose toward Western stylistic norms, reducing cultural nuance (Agarwal et al., 2025). These findings suggest that the large-scale deployment of AI writing assistants may gradually stabilize and narrow the range of communicative styles that people use. Our work is the first to perform randomized controlled trials to study how humans actually interact with LLMs, to show that not only do LLMs homogenize our writing, but have the ability to influence the views and judgments that humans express when writing.

The Effect of AI on Institutions. As LLM usage becomes embedded in formal institutions, concerns have emerged on its impact on collective judgment and evaluation. Reports show that academic peer reviews and scientific writing contain content heavily generated and/or edited by LLMs (Emi, 2025; Liang et al., 2025). While prior work has quantified whether LLMs are used when writing, our work is the first to examine how LLM-assisted writing in academic contexts changes conclusions reached by humans. Our findings suggest that the use of LLMs in institutional contexts causes not only the homogenization of language but also the decisions that shape scientific outcomes.

3 Experiment Methodology

In this section, we describe the methodology used to quantify semantic, lexical, grammatical, and affective differences between human-written text and LLM-written or LLM-edited text, focusing on three datasets: (1) 50 argumentative essays written with and without the assistance of LLMs by human participants; (2) 86 argumentative essays written by humans and edited independently by humans and by LLMs prompted as editors; and (3) peer reviews from an academic conference. We describe the three settings below.

3.1 DATASETS.

Human Study. We conducted a randomized controlled trial (RCT) to investigate how humans use LLMs when writing an argumentative essay on the question: "Does money lead to happiness?", an open-ended topic that we felt most participants could relate to. We recruited 50 participants from Prolific, a research platform through which participants can voluntarily participate in research surveys and receive compensation. We randomly assigned participants to one of two conditions: (1) a baseline ($n = 20$), where participants were not allowed to use LLMs, or (2) an AI-assisted condition ($n = 30$), where participants were allowed to use an LLM in whatever way they chose while writing. For the AI-assisted group, we recorded the full interaction history with the LLM as well as the final draft written with LLM assistance. To evaluate the impact of LLM intervention, each participant completed a pre-study and post-study questionnaire to measure shifts in attitudes toward LLMs, their own creative agency, and the alignment of model outputs with their preferences when writing. These self-reported ratings allow us to clearly measure if human preferences are truly being met through collaboration with LLMs, and how the human's opinion on the use of LLMs changes as a result of the interaction. We provide the pre-study and post-study questions in Section A.2, and details such as compensation, recruitment process, and duration of the study in Section A.1. All procedures were approved by our Institutional Review Board (IRB).

ArgRewrite-v2. For our analysis of how LLMs vs humans edit essays, we use ArgRewrite-v2 (Chen et al., 2022), a dataset of argumentative writing revisions collected from 86 university students. All participants were tasked to develop an initial argumentative essay draft as the first draft (D1) for or against self-driving cars that could serve as an op-ed piece in a local newspaper. Each draft was then provided with feedback from a human expert to improve the initial draft, which included both coarse-grained (surface vs content) and fine-grained (e.g., claim, evidence, reasoning, word usage, precision, etc.) feedback. In response, participants revised the initial draft based on the expert-provided feedback to form the second draft (D2). Importantly, we use this dataset for our analysis as it was released in 2022, and hence the writing pre-dates the widespread adoption of LLM-based writing assistants like ChatGPT (OpenAI, 2026).

Peer Reviews. For analysis on how in-the-wild LLM use may shift the claims made by those who use them, we analyze peer reviews from the International Conference on Learning Representations (González-Márquez & Kobak, 2024), written by humans with those generated or heavily edited by LLMs in 2026. From 75,000 reviews, reports have found that over 20% of these reviews were LLM-generated, with an additional 39% of papers that used LLMs to edit or generate parts of the text (Emi, 2025). We chose to use this dataset because of (1) the incentive of the reviewers to create high-quality reviews as their professional reputation is on the line, and (2) the high confidence and low false positive rate of the Pangram tool for use as an LLM Detector (Emi & Spero, 2024). For a select number of topic categories, we compare the LLM-edited texts to the fully human-edited texts to conclude the effects that LLM editing has not only on writing style, but on consequential factors such as sentiment of the review (positive vs. negative) and the assigned score, showing that LLM-generated text is already having a meaningful impact on real-world institutions.

3.2 METRICS

PCA of Embedding Representations. To capture changes in semantic meaning between human-edited and LLM-edited drafts, we project each draft into a high-dimensional vector space using an encoder transformer model to generate an embedding. To identify the primary axes of semantic variation across the dataset, we apply Principal Component Analysis (PCA) to the embedding representation. Distances in our PCA plot reflect similarity in meaning or semantic content between essays (Mikolov et al., 2013; Reimers & Gurevych, 2019). In our analysis, we analyze both the magnitude and direction of each of the 86 different essays. Large magnitude edits, all pointing in a similar direction, indicate that the LLM is not merely correcting grammar, but is actively steering diverse human perspectives in a homogenizing way, toward a very different conceptual mode.

Lexical Distributions. To complement our semantic analysis, we also want to quantify the lexical difference between human-edited and LLM-edited draft to understand which words appear, how often they appear, and whether some words are swapped for others. While embeddings capture high-level conceptual change, they may mask changes to specific word choices that make up an individual's unique writing style. Thus, we measure the divergence between the unigram distri-

Changing the Human User's Conclusion	**Human Draft** The transition phase of slowly adding self-driving cars would cause lingering tension. At this time, since the perks do not outweigh the persistent cons, perhaps America is not ready for self-driving cars.	**Grammer Edit | CLAUDE-HAIKU** The transition phase of slowly adding self-driving cars would cause tension and last **until production costs are reduced.** At this time, since the drawbacks do not outweigh the benefits, America is not ready for self-driving cars **until we can implement them efficiently to truly reap their advantages.**
Removal of Human User Voice	**Human Draft** In today's age, **we** are always looking for the 'next thing' in terms of technological advancement. Google is one company that **continuously strives to answer that call.** Recently they have been doing so by researching the technology for and creating self-driving cars. While this may sound like some **futuristic fabrication from Back to the Future, the technology is actually here and being tested even as this piece is being written.**	**Expansion Edit | gemini-2.5-flash** In today's age, **the relentless pursuit** of technological advancement **drives innovation across industries.** Google, for instance, has been at the **forefront** of developing self-driving car technology, **transforming** what once seemed like futuristic fiction into a **tangible reality.** While these autonomous vehicles are currently undergoing extensive testing, **their widespread integration into society presents a complex dual potential.**
Undoing Human-User's Intention for Paragraph	**Human Draft** The topic of self-driving cars is highly debated especially in light of recent event with the **death of an Arizona woman at the hands** of a self-driving Uber. Those for self-driving cars argue that they reduce the rate of accidents as a result of distracted drivers, improve traffic flow, are cost saving in terms of time and monetary value from insurance, and could allow greater access to transportation for the disabled and elderly. **Those against self-driving cars argue the system does not eliminate accidents, that there is a learning curve with the technology, that it would take away profit from other sectors of the economy, and that there are instances in which the technology is not equipped to make a decision. While both sides make valid points, the benefits of adopting a system of self-driving cars rises to the forefront.**	**General Edit | gpt-5-mini** The debate over self-driving cars intensified after the **fatal Arizona collision involving an autonomous Uber, but a careful assessment shows that the benefits of adopting self-driving technology outweigh its risks.** Autonomous vehicles promise to reduce crash rates caused by human error, improve traffic flow and travel efficiency, lower overall societal costs from accidents, and expand transportation access for elderly and disabled populations. **Although the technology is imperfect and will require rigorous development and regulation, its potential to save lives and increase mobility makes continued investment and deployment the responsible course.**

Figure 2: Example of substantial argumentative rewriting by an LLM. Red text highlights segments removed or substantially reframed by the model, while green text highlights segments added to the human draft. The figure illustrates how LLM edits can alter conclusions (top left), remove human colloquialisms and examples (bottom left), and remove content in favor of a biased claim.

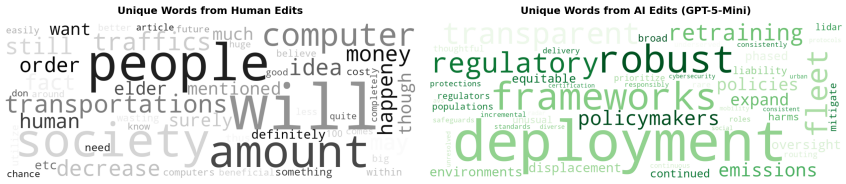


Figure 3: Unique words in human-edited texts (left) versus AI-edited texts using `gpt-5-mini` (right). Word size reflects relative frequency, highlighting stylistic and thematic differences between edits.

butions of the original human drafts and the revisions. We quantify these lexical shifts using the Jensen-Shannon Divergence (JSD) (Menéndez et al., 1997). We represent each draft as a discrete probability distribution over the global vocabulary, where the probability of a token is proportional to its frequency in the text. This approach allows us to quantify lexical change independently of sentence structure or syntax.

For probability distributions P of the initial draft and Q of the revised draft, the JSD is defined as:

$$\mathrm{JSD}(P \,\|\, Q) = \frac{1}{2} D_{\mathrm{KL}}(P \,\|\, M) + \frac{1}{2} D_{\mathrm{KL}}(Q \,\|\, M) \tag{1}$$

where $M = \frac{1}{2}(P + Q)$. By calculating JSD over word counts, we can empirically determine the extent to which LLMs alter the lexical composition of human writing compared to LLM-edited writing. A higher JSD indicates a more pronounced departure from the human's vocabulary.

Measuring change in emotional distribution. To quantify shifts in the affective quality of revisions, we utilize the NRC (National Research Council - Canada) Emotion Lexicon (Mohammad & Turney, 2013). This resource is a list of English words and their associations with eight basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, as well as two sentiments (positive and negative). By computing the density of these emotional markers in both human and AI-revised essays, we can detect whether LLMs systematically change the emotional tone of argumentative writing or introduce specific affective biases, such as an increase in positive sentiment, that were absent in the original human drafts.

LIWC: We also employ the Linguistic Inquiry and Word Count (LIWC) tool (Tausczik & Pennebaker, 2010; Boyd et al., 2022) to measure the psychological, emotional, and cognitive characteristics of the language in human and AI-written text based on word usage. LIWC categorizes words into over 90 semantically and grammatically defined dimensions, including summary variables (e.g., analytical thinking, clout, and authenticity), grammatical categories (e.g., pronouns, prepositions), & psychological processes (e.g., cognitive mechanisms, social processes). LIWC allows us to move

beyond simple surface-level changes and quantify how text edited or written by LLMs alters the meaning of an essay across different categories. For instance, we use the Analytical Thinking and Authenticity metrics to determine if LLM revisions shift a human-users's natural voice toward the formal, detached, and highly structured style commonly associated with generative models.

Qualitative Analyses Using LLM-as-a-Judge We use `gpt-4o` as an LLM-as-a-Judge, a common technique to automate qualitative analyses to determine qualitative attributes for essays and ICLR reviews written by humans and AIs, respectively. For the human study, we use our LLM-as-a-Judge to determine (1) the extent to which the essay agreed or disagreed with the question "Does money lead to happiness?" and (2) the different claims that the essays made to support their claims. For the ICLR review analysis, we use LLM-as-a-Judge to first extract categories of strengths and weaknesses (such as 'novelty') and then label each ICLR review with these selected strengths and weaknesses. Prompts for the LLM-as-a-Judge can be found in Section E.

4 RESULTS

We observe from our human user results that those who used an LLM found that their essay was not in their voice and less creative ($p < 0.01$). To understand why, we perform a controlled experiment where we prompt LLMs to perform edits on essays from a pre-LLM era, revealing dramatic distortions in style, content, and meaning, even when the LLM is prompted only to give grammar edits. We find the distortion of LLMs goes beyond controlled experiments in essay writing, affecting the decisions and rationales of reviewers at top conferences such as ICLR 2026.

(1) Humans Report That Essays using LLMs Do Not Reflect Their Own Voice Figure 4 shows that human users who relied on an LLM to help write their essay reported feeling that their essay was significantly less creative and less in their own voice (p-score¡0.005). Human users also did not report feeling that it was easier to organize their writing, that they struggled less than the human users who did not use an LLM, or feeling less satisfied with the outcome of the essay. This illustrates a fundamental drawback with using preference scores, because although the satisfaction reported by human users did not decrease, reflective of standard preference scores, the participants noticed a significant decrease in creativity and personal voice.

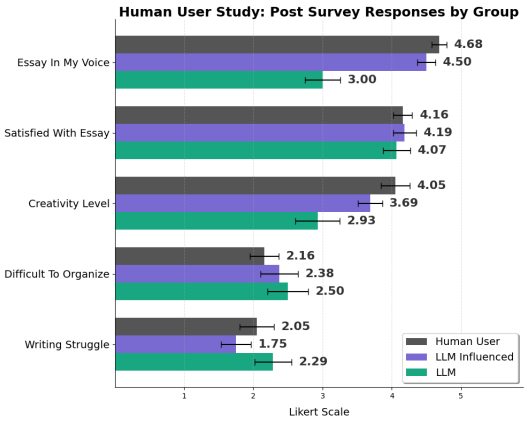


Figure 4: Post-study survey responses reveal that human users consider the responses that they constructed with an LLM are significantly less creative and less in their voice ($p < 0.01$). Despite this, users who used an LLM did not report significantly less satisfaction or struggle when writing their essay.

(2) LLMs Homogenize Writing by Shifting Essays in a Common Semantic Direction. From our human user-study, we found that 16 / 50 users who were allowed to interact with an LLMs chose to instead use the LLM as an information-seeking or writing advice tool. These participants reported writing the essay almost entirely on their own, with an LLM-generated percentage less than 40% from an existing AI-detection software (Emi & Spero, 2024), and 14 of them reporting less than 20% of the essay as written by an LLM. We refer to these participants as *LLM-influenced*, as they had access to LLMs even if they were not included in their final work. We distinguish those participants from those who used the LLM to generate the complete essay, denoted as the *LLM* condition.

To understand how the use of LLMs affects the semantic meaning expressed in human writing, we analyzed semantic shifts using sentence embeddings projected into two dimensions via T-SNE between human writing, LLM influenced writing and LLM writing.

Figure 5 shows the semantic distribution of essays across these three conditions. We find that essays primarily written by humans (in black) and those that are *LLM-influenced* (in purple) are widely spread out throughout the embedding space, occupying a broad region that reflects the diversity of individual perspectives, writing styles, and argumentation. On

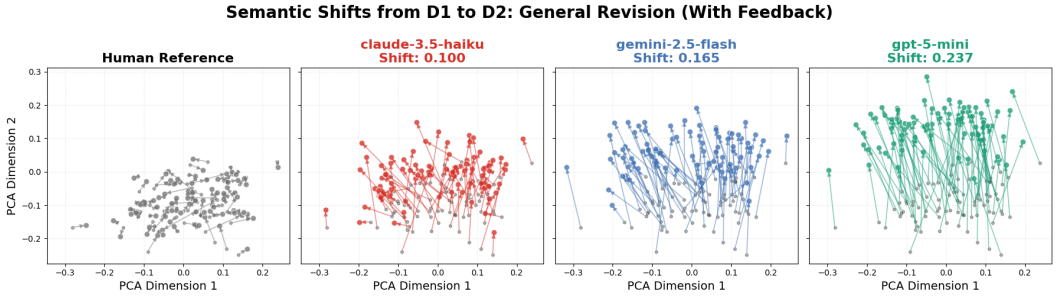


Figure 6: Semantic shifts induced by human and LLM revisions for the Arg Rewrite v2 dataset. Each point pair represents an essay before (D1) and after (D2) revision, embedded using `gemini-004` sentence embeddings and projected into two dimensions via PCA, a common approach for analyzing semantic differences (Dhillon et al., 2015). The left panel shows human revisions, while the remaining panels show revisions produced by different LLMs without access to expert feedback. Arrows indicate the direction and magnitude of semantic change. Human revisions exhibit smaller, more varied semantic shifts, whereas LLM revisions produce larger shifts that are strongly aligned in a common direction, indicating a homogenization effect in semantic space.

the other hand, essays written by LLMs (in green) form a tight cluster in the bottom right quadrant of the space, a space that is not occupied by any of the human-written essays. This clustering shows that LLMs produce text that is semantically different from that produced by humans. We find that *LLM-influenced* is in between human-users and LLMs, suggesting that even minimal exposure to LLM outputs can shift human writing toward LLM semantic patterns.

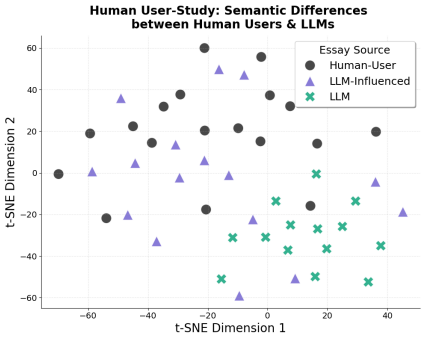


Figure 5: t-SNE of semantic embeddings. LLM essays form a distinct cluster, while LLM-influenced essays lie between human & LLM regions.

To further test whether there is a significant difference between the semantic meaning of human-written text versus LLM-written text, we performed a similar analyses in a controlled setting. Specifically, we prompted three LLMs (`gpt-5-mini`, `gemini-2.5-flash`, `claude-haiku`) to edit original human drafted essays in the ArgRewrite-v2 dataset. Our human user study shows that human users employ LLMs for generating ideas and arguments, expanding their own ideas from existing human-written text, writing a first paragraph and asking the LLM to finish the rest, asking the LLM to review the human-written essay, and asking the LLM to write the entire essay. Hence, we prompted LLMs to perform five revision types: (1) *general* version to comprehensively improve the essay (2) *minimal edits* that make only necessary corrections (3) *grammar* revisions that fix surface errors without changing content (4) *completion* revisions to finish incomplete essays and (5) *expansion* to elaborate on existing ideas. For each revision type, we also varied whether the LLM had access to expert human feedback on the original human draft to maintain a fair comparison with the edited drafts from human users in the ArgRewrite-v2 dataset. However, as human users generally do not have access to expert feedback when improving essays, we also ask the LLM to edit without the expert feedback.

Figure 6 shows the semantic shifts from the initial human draft (D1) to revised drafts (D2) by the human editor for the ArgRewrite-v2 datasets and three LLM editors for the *general* revision case, and Figure 1 shows the embedding distribution across different types of revisions from `gpt-5-mini`. We find that humans make small, multidirectional semantic shifts, with arrows pointing in diverse directions, and the revised essays remaining dispersed throughout the embedding space. In contrast, all three LLMs produce semantic shifts consistently pointed in the same direction, with the magnitude of semantic change largest for `gpt-5-mini` and smallest for `claude-haiku`. Across LLMs, we also find the largest semantic change is between the initial human draft and the final LLM draft when LLMs are prompted to *complete* and *expand* the essay. However, we find concerning levels of shift when LLMs are tasked to perform *minimal* edits to the essay or only edit the essay for *grammar*. Further results in Section C.2, Section C.3, and Section C.4.

In the next sections, we further analyze how these differences between human-written and LLM-written texts actually manifest.

(3) LLMs Make Substantially Larger Lexical Changes Than Humans. We also examined how LLMs alter the specific word choices that make up an individual's writing style, by measuring the lexical divergence between the initial human draft and the revised human and LLM drafts with JSD for the ArgRewrite-v2 dataset. Figure 7 shows the distribution of JSD values for general revisions with expert feedback, with JSD plots for other types of edits found in Section C.5. The human baseline exhibits a tight distribution centered around 0.2-0.3 JSD, indicating that humans make modest, targeted word substitutions while preserving most of their original vocabulary. In contrast, all three LLMs shift the distribution towards the right and produce wider spreads, with *gpt-5-mini* showing the most significant change in divergence of nearly triple the human baseline, with many essays reaching divergences above 0.7. These lexical shifts demonstrate that LLMs replace a much larger fraction of the original writing than humans do when revising their own work. This substitution of words contributes to the loss of individual voice and style, as the unique lexical fingerprint of each writer is overwritten by the given models preferred vocabulary. Further results with different LLM-editing conditions and LLMs found in Section C.5.

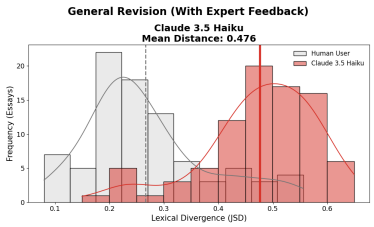


Figure 7: Lexical divergence distribution for *general* revision of `claude-haiku` with expert feedback. Gray bars show the human baseline, while colored distribution show LLM revisions systematically shift rightward. Higher JSD indicates more extensive vocabulary replacement, with LLMs substituting substantially more words than humans.

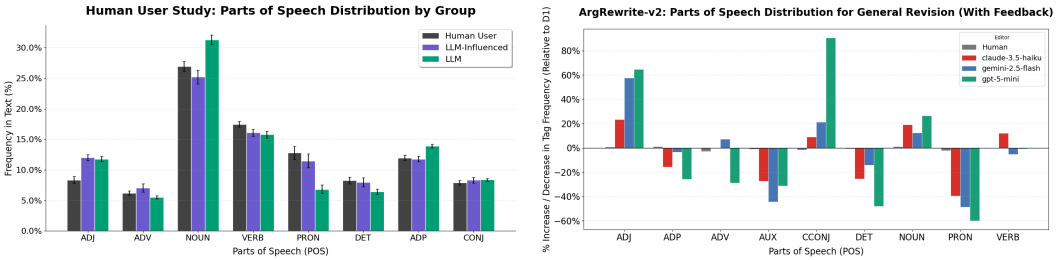


Figure 8: Parts of Speech Distribution. Left: Essays written completely by humans versus written with assistance of LLMs. Right: Essays edited by humans versus edited by 3 different LLMs prompted to make *general* edits. LLMs use more nouns and adjectives when writing/editing, and reduce the use of pronouns.

(4) LLMs Systematically Restructure Grammar Toward a Noun-Heavy, Formal Style. To understand how LLMs change the syntactic structure of writing, we analyzed part-of-speech (POS) distributions in essays written by human users versus essays written by LLMs. Figure 8 (left) shows the relative change in POS tag frequencies for our human-user study. We find there to be 50% decrease in pronouns from human-user essays and essays written with LLMs, signifying a removal of first-person, experience-based argumentation toward impersonal language. We also find that LLM-written essays show a higher percentage of nouns (31% frequency in text) than human-written essays (27% frequency in text), with adjectives showing a similar trend (12% for LLM-written essays, 8% for human-written essays). Figure 8 shows a similar analysis for the ArgRewrite-v2 dataset. We find that humans make minimal grammatical changes typically under 5% for any POS category. In contrast, LLMs systematically restructure sentences toward a more formal style and primarily using nouns in writing. We also find that all three models dramatically increase the use of adjectives compared to the human draft ($57 - 90\%$ increase) and coordinating conjunctions ($13 - 90\%$ increase), while substantially reducing pronouns ($40 - 60\%$ decrease) and determiners ($25 - 50\%$ decrease). This confirms our qualitative findings that LLM edits move writing away from narrative, first-person toward impersonal, academic writing. We find these shifts to be more pronounced when the LLM is prompted to complete or expand the essay. We find `gpt-5-mini` shows the most extreme restructuring, with an 88% increase in coordinating conjunctions and 27% increase in nouns, alongside a 61% decrease in pronouns. This pattern aligns with prior observations that LLMs favor complex, formal constructions over the more direct, personal style typical of human writing. For minimal edits, LLMs still increase use of adjectives by 40-87% and reduce pronouns by 31-56%. Further results with different LLM-editing conditions found in Section C.6.

(5) The use of LLMs for writing increases emotional language. We examined whether LLM edits change the distribution of emotions present in text using the NRC Lexicon (Mohammad &

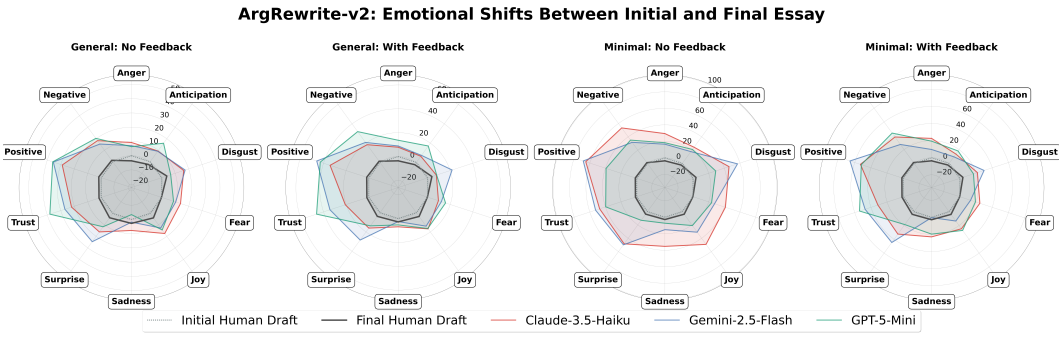


Figure 9: Emotional shifts for general revisions with and without expert feedback. Human baseline (gray) shows minimal affective changes. LLMs dramatically increase positive sentiment (37-54%) and trust-related language (17-53%), while also simultaneously increasing negative sentiment (24-38%). This indicates dramatic increases in the use of emotional language, regardless of the way the LLM is prompted.

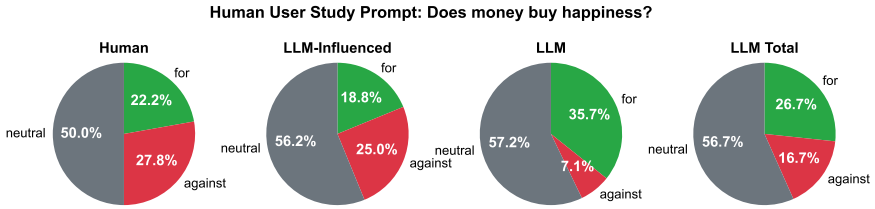


Figure 10: We display the results of using LLM-as-a-Judge to categorize the human study essays based on stance. LLM Total refers to a combination of both the LLM-Influenced and LLM categories. Humans who had the option to interact with the LLM (LLM total) during the RCT are 40% less likely to argue against the idea that money leads to happiness, while humans who leaned heavily on the LLM to write most of the essay are 68% less likely to argue that money leads to happiness.

Turney, 2013) on the ArgRewrite-v2 dataset. Figure 9 shows the change in emotional word density for *general* and *minimal* revisions with and without expert feedback. We find human edits make minimal affective changes (black), with adjustments typically under 5% for any emotion category compared to the initial human draft. On the other hand, LLMs increase the amount of emotional language used in general, with LLMs expressing slightly words relating to positivity and trust labels in the essay. This pattern suggests that LLMs systematically reframe arguments in more positive, optimistic terms, even when the original human text may have been critical, skeptical, or balanced. For essays about self-driving cars (the ArgRewrite-v2 topic), this could mean downplaying concerns about safety, job displacement, or ethical issues in favor of enthusiasm about technological progress, as shown in Figure 2, where the LLM removes mention of the drawbacks of self-driving cars (row 3). Interestingly, we also find that without expert human feedback guiding the LLM edits, the emotional shifts across all categories relative to the initial human draft is a lot more drastic, reinforcing the observation that LLMs increase emotional language in the essay. We specifically find edits by `claude-haiku` to be less pronounced with expert feedback. Further results with different LLM-editing conditions and LLMs found in Section C.7.

(6) LLMs Distort Human Decisions and Rationales. In our human user study, we find that LLMs significantly distort both the conclusions that humans come to and the arguments they use to arrive at their conclusions. In Figure 10, we find that humans who interacted with an LLM to write the essay are 28% more likely to argue that money does lead to happiness, and humans who used an LLM to write most of the essay are 68% less likely to argue against money leading to happiness. This finding illustrates that humans' intent in decision-making is strongly influenced by interacting with an LLM while writing an essay, and the effects are more pronounced the more that humans use LLMs to help them write. In Figure 44, we find that certain types of arguments are more typical of LLMs than humans. Humans are more likely to use arguments related to personal experience, while LLM-written essays are more likely to use statistics and logic. LLM-influenced essays also cite expert opinions, something that human-written essays rarely do. We find these results corroborated in Figure 2 (row 2), where the first-person voice is removed by the LLM-edited draft.

In our qualitative analysis of LLM-edited essays on ArgRewrite-v2, we find similar trends. As shown in the examples in Figure 2, when prompted to edit an essay for grammar, LLMs inadvertently make changes to the claims in the essay. This pattern indicates that LLMs are not simply correcting

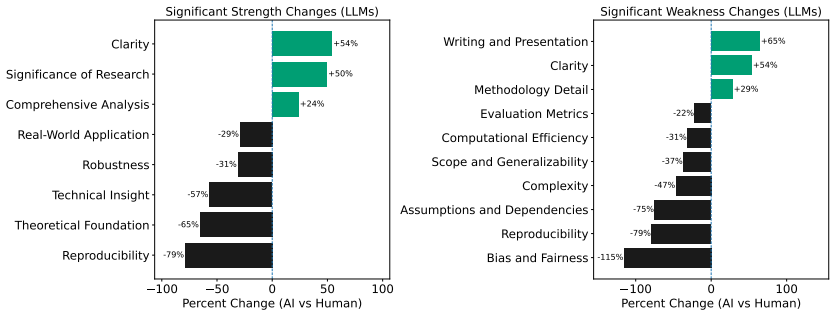


Figure 11: We use LLM-as-a-Judge on 2.2k reviews from ICLR 2026, selected because they were all written on the subject of LLMs, where 1.1k reviews are entirely written by humans, and 1.1k are entirely written by LLMs. We plot the relative frequency of certain strengths and weaknesses occurring in LLM reviews instead of human reviews. Humans are over 50% more likely to comment on clarity and significance of research for both strengths and weaknesses, while LLMs are 79% more likely to comment on reproducibility.

errors or improving clarity, but are fundamentally reorienting the content of diverse human essays toward a shared semantic mode, which, in the case of ArgRewrite, are essays that are in support of self-driving cars, with very similar use of language. The uniformity of these shifts across different LLMs also suggests a convergence toward LLM-preferred linguistic patterns (Jiang et al., 2025a) that may not reflect the original intent or voice of human writers, as the final drafts produced by humans are semantically very different than the LLM-edited drafts.

(7) LLMs Distort Decisions affecting Scientific Institutions When LLMs are employed in the scientific review process, the scores, decisions, and arguments made shift dramatically. In our ICLR review analysis, we observe that LLMs assign scores 10% higher than humans (4.43 for LLM reviews and 4.13 for human reviews). In fig. 11, strengths and weaknesses are dramatically different between LLM-written reviews in comparison to human-written reviews. [1] We use LLM-as-a-Judge to label 2.2k reviews from ICLR 2026 on 1.1k papers on the topic of LLMs, each with one review written entirely by a human and a review written entirely by an LLM. We observe that humans are over 50% more likely to comment on the clarity and significance of research, while LLMs are 79% more likely to comment on reproducibility for both strengths and weaknesses.

5 Discussion

In this paper, we show (1) LLM-edited texts change semantics, lexical distribution, and emotions more significantly than human-edited texts; (2) users with LLM assistants as co-authors find the resulting essay is significantly less creative and not in their voice; and (3) LLMs used to edit conference reviews are affecting the outcomes and decisions at the institution level. Our results show that text generated by LLMs induces significant changes in human writing, in a way that causes many human users to feel that they are losing a sense of creativity and voice in their writing. With LLM-generated text already affecting real-world institutions like scientific peer review, if 1 billion people are currently using LLMs, including politicians in their speeches in parliament, how will this affect our cultural institutions? How will that affect the ways we organize, cooperate, develop scientific research and technology? Will a loss of diversity in thought and culture make us more vulnerable in a way that a loss of ecological diversity makes species more vulnerable to disaster and extinction? We hypothesize that rather than maximizing positive human feedback from a chat interface (similar to the one in our pilot study) as is common in RLHF Ouyang et al. (2022), designers of algorithms need to introduce metrics to minimize alteration magnitude, while maintaining diversity to avoid homogenization effects. Within preference optimization, there is a large body of work centered around optimizing for more diverse preferences Jang et al. (2023); Poddar et al. (2024). However, these methods still require asking for feedback from diverse respondents.

[1] The LLM review versus human review is from designations from (Emi & Spero, 2024), a classifier with a low false positive rate. It is possible for this tool to have errors and to introduce correlations, but the tool focuses largely on stylistic differences to make judgments rather than semantic ones.

Ethics Statement

This work studies how using LLMs for writing and editing of text impacts the content of human writing across various dimensions including semantics, grammar, emotional distributions, and the claims being made. We conducted a human-user study to understand human-user preferences towards writing, and made sure to have proper IRB approval and protocols to make sure for no personally identifiable information to be collected from human-users. We recruited participants via Prolific, asked for consent prior to the study, and compensated participants for their time. Data for ArgRewrite-v2 and academic peer reviews were publicly available on the internet and did not contain any personal identifiable information.

Finally, our contributions provide evidence for the need to design LLMs that preserve human agency when writing. We hope this works encourages research to develop safeguards and improve tools for human writing that preserves the human voice.

Reproducibility Statement

In order to ensure reproducibility, we provide details of our human-user study in Section A.1, containing the instructions provided to participants who were tasked to write and without an LLM, as well as the pre-study and post-study questions provided. Additionally, we provide further user study results in Section B, including quanitative results on the semantic distribution, grammar, and emotional distribution of essays. We also provide the prompts used for our analysis of the ArgRewrite-v2 dataset in Section C.1, as well as full results on the semantic (Section C.3, Section C.2, and Section C.4), lexical (Section C.5), emotional (Section C.7), and parts of speech (Section C.6) analyses, and for robustness, perform quantitative experiments across models and various prompts for editing grounded in our human-user study. Lastly, we provide analysis on several other ICLR categories, and provide the prompts used for the analysis in Section C.8.

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–21. ACM, April 2025. doi: 10.1145/3706598.3713564. URL http://dx.doi.org/10.1145/3706598.3713564.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Creativity and Cognition*, CC '24, pp. 413–425. ACM, June 2024. doi: 10.1145/3635636.3656204. URL http://dx.doi.org/10.1145/3635636.3656204.

Ryan L. Boyd, Aneeta Ashokkumar, Sarah Seraj, and James W. Pennebaker. The development and psychometric properties of liwc-22. Technical report, University of Texas at Austin, Austin, TX, 2022. URL https://www.liwc.app.

Aaron Chatterji, Erik Brynjolfsson, Anton Korinek, Kristina McElheran, Robert Seamans, and Nikolas Zolas. How are people using generative ai? Technical Report 34255, National Bureau of Economic Research, 2025. URL https://www.nber.org/papers/w34255.

Li Chen, Fan Zhang, and Diane Litman. Argrewrite 2.0: A corpus and model for argumentative text revision. *arXiv preprint arXiv:2206.01677*, 2022.

Paramveer S Dhillon, Dean P Foster, and Lyle H Ungar. Eigenwords: spectral word embeddings. *J. Mach. Learn. Res.*, 16:3035–3078, 2015.

Anil R Doshi and Oliver P Hauser. Generative artificial intelligence enhances creativity but reduces the diversity of novel content. *arXiv preprint arXiv:2312.00506*, 2023.

Anil R. Doshi and Oliver P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290. URL https://www.science.org/doi/abs/10.1126/sciadv.adn5290.

Bradley Emi. Pangram predicts 21% of iclr reviews are ai-generated. Pangram Labs Blog, Nov 2025. URL https://www.pangram.com/blog/pangram-predicts-21-of-iclr-reviews-are-ai-generated. Accessed: 2026-01-12.

Bradley Emi and Max Spero. Technical report on the pangram ai-generated text classifier, 2024. URL https://arxiv.org/abs/2402.14873.

Rita González-Márquez and Dmitry Kobak. Learning representations of learning representations. In *Data-centric Machine Learning Research (DMLR) workshop at ICLR 2024*, 2024.

Jess Hohenstein, Dominic DiFranzo, Rene F. Kizilcec, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeff Hancock, and Malte F. Jung. Artificial intelligence in communication impacts language and social relationships. *arXiv*, 2021. URL https://arxiv.org/abs/2102.05756.

Ben Hutchinson, Suchi Saria, and Oren Etzioni. The lock-in hypothesis: Stagnation by algorithm. *arXiv preprint arXiv:2506.06166*, 2025.

Rhiannon James. Labour mps accused of using AI to write parliamentary speeches. The Independent, Sep 2025. URL https://www.the-independent.com/news/uk/home-news/labour-mp-ai-chat-gpt-speeches-tugendhat-b2823339.html. Accessed: 2026-01-12.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *Neural Information Processing Systems (NeurIPS)*, 2025a.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025b. URL https://arxiv.org/abs/2510.22954.

Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. Argrewrite v.2: an annotated argumentative revisions corpus. *Language Resources and Evaluation*, 56(3):881–915, January 2022. ISSN 1574-0218. doi: 10.1007/s10579-021-09567-z. URL http://dx.doi.org/10.1007/s10579-021-09567-z.

Soojin Kim, Chen Li, and David Alvarez-Melis. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing tasks. *arXiv preprint arXiv:2506.08872*, 2025.

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

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, 9:2599 – 2609, 2025. URL https://api.semanticscholar.org/CorpusID:280523748.

Lennart Meincke, Gideon Nave, and Christian Terwiesch. Chatgpt decreases idea diversity in brainstorming. *Nature human behaviour*, pp. 1–3, 2025.

María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del Carmen Pardo. The jensen-shannon divergence. *Journal of The Franklin Institute-engineering and Applied Mathematics*, 334:307–318, 1997. URL https://api.semanticscholar.org/CorpusID:120842983.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11241–11258. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.561. URL http://dx.doi.org/10.18653/v1/2025.naacl-long.561.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586.

OpenAI. Chatgpt. https://chatgpt.com/, 2026. Large language model (GPT-5.2), accessed January 14, 2026.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning, 2024. URL https://arxiv.org/abs/2408.10075.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Sahand Sabour, June M. Liu, Siyang Liu, Chris Z. Yao, Shiyao Cui, Xuanming Zhang, Wen Zhang, Yaru Cao, Advait Bhat, Jian Guan, Wei Wu, Rada Mihalcea, Hongning Wang, Tim Althoff, Tatia M. C. Lee, and Minlie Huang. Human decision-making is susceptible to ai-driven manipulation, 2025. URL https://arxiv.org/abs/2502.07663.

Aránzazu Sanz-Tejeda, Juana Celia Domínguez-Oller, Josep María Baldaquí-Escandell, Raquel Gómez-Díaz, and Araceli García-Rodríguez. The impact of generative ai on academic reading and writing: A synthesis of recent evidence (2023–2025). *Frontiers in Education*, 10, 2026. doi: 10.3389/feduc.2025.1711718. URL https://www.frontiersin.org/articles/10.3389/feduc.2025.1711718.

ScienceDaily. Ai found to boost individual creativity – at the expense of less varied content, 2024. URL https://www.sciencedaily.com/releases/2024/07/240712222127.htm. accessed: 2026-01-13.

Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1): 24–54, 2010. doi: 10.1177/0261927X09351676. URL https://doi.org/10.1177/0261927X09351676.

Jin Wang and Wenxiang Fan. The effect of chatgpt on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12, 05 2025. doi: 10.1057/s41599-025-04787-y.

Weichen Wang, Sihan Ma, Shiyi Yang, Joon Sung Lee, Paras Jain, Ashton Anderson, et al. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 11(32):eadn5290, 2025.

Kilian Wenker. Who wrote this? how smart replies impact language and agency in the workplace. *Telematics and Informatics Reports*, 10:100062, 2023. doi: 10.1016/j.teler.2023.100062. URL https://doi.org/10.1016/j.teler.2023.100062.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35): e2504966122, 2025. doi: 10.1073/pnas.2504966122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2504966122.

Ruiqi Zhou and Klaus Fiedler. The basic b*** effect: The use of llm-based agents reduces the distinctiveness and diversity of people's choices. *arXiv preprint arXiv:2509.02910*, 2025.

A User Study Details

A.1 User Study Recruitment Process, Compensation and Duration

To recruit our participants, we use Prolific, a research platform through which participants can voluntarily participate in research surveys and receive compensation. We give our participants 8 US dollars to participate in the study, with an estimated time of 35 minutes and a maximum time limit of 1 hour. To participate in the study, we require participants to be native English speakers and reside in the US, Canada, Australia, Ireland, the UK, or New Zealand to control for native language as a factor when writing essays with an AI assistant.

A.2 User Study Instructions.

We give the users the following instructions before we have them do the pre-study questions, the writing with an assistant, and the post-study questions:

A.3 LLM Assisted Instructions

Instructions (LLM-Assisted Condition)

You will be writing an essay, and you may use the LLM chat to assist you in the writing process. Use no other LLM than the one provided in this interface, and you may take as much time as you need.
First, answer some pre-study questions about your attitudes toward AI and writing before you begin the essay. You may not use other sources, such as the internet, to inform your essay.
After the study, you will be asked post-study questions about your experience.
The purpose of this study is to understand how people use LLMs for writing in their normal workflow. If you do not usually use LLMs, think of the AI tool as a writing partner: someone to bounce ideas off, ask questions, and get feedback from as you go.

A.4 Without LLM Instructions

Instructions (No-LLM Condition)

First, answer pre-study questions about your writing habits and experiences.
Then write your essay in the text box provided. You are not allowed to use an LLM, AI assistant, or the internet while writing this essay.
Return here once the essay is complete to answer a few questions.
Responses will be reviewed for indications of AI-generated content. Participants found to have used such tools will not receive compensation.

A.5 Pre-Study Questions

For LLM Users

Attitudes Toward AI-Assisted Writing

- I believe AI tools can improve my writing quality. (+)
- I do not expect AI systems to understand my writing style. (-)
- I trust AI systems to provide accurate information. (+)
- I do not believe using AI for writing is acceptable in academic contexts. (-)

Appropriate Uses of AI for Essay Writing

For essay writing, I think AI tools should be used:

- Not at all
- Check grammar, spelling, or clarity
- To offer suggestions on how to improve my writing
- Help brainstorm or outline ideas
- To rewrite my essay from scratch
- To write the entire essay
- Other (please specify): _____

FOR BOTH

Writing Confidence and Habits

- I often struggle with structuring my ideas clearly. (-)
- I feel confident in my ability to write and edit essays on my own. (+)
- I find essay writing time-consuming. (-)
- I usually find essay writing enjoyable. (+)

Frequency of LLM Use

I use LLMs to help me write:

- Daily
- Weekly
- Monthly
- Checked it out a few times
- Never

Purposes for Using LLMs

What do you use LLMs for?

- I don't use LLMs
- General conversation
- Search queries / seeking knowledge
- Learning or understanding new concepts
- Advice
- Writing or editing text
- Work or productivity tasks
- Other (please specify): _____

OPEN-ENDED

Prior Experience With LLMs

Please describe the last time you used a large language model (LLM) such as ChatGPT, Claude, or Gemini. What did you use it for, and in what context (e.g., work, study, personal use)? How helpful was the experience, and why?

A.6 POST-STUDY

FOR BOTH

- I was satisfied with the essay. (+)
- I felt the essay was written in my voice. (+)
- I found it difficult to organize my thoughts while writing. (-)
- Writing this essay was a struggle for me. (-)

How creative do you feel you were in writing the essay?

- Very creative
- Somewhat creative
- Neither creative nor uncreative
- Somewhat uncreative
- Not at all creative

OPEN-ENDED

Please describe your experience writing this essay.
Comment on: How well does the essay reflect your own views and writing style? How much effort did you put into writing it? Did you learn anything during the process?

FOR LLM USERS

Estimated LLM Contribution

What percentage of the document would you say was LLM-generated?

- 0%
- 20%
- 40%
- 60%
- 80%
- 100%

Perceptions of LLM Assistance

The following statements were rated on a Likert scale **?** from *strongly disagree* to *strongly agree*.

- The LLM helped me generate ideas more effectively. (+)
- The model's feedback improved the quality of my essay. (+)
- The LLM's suggestions were irrelevant to my goals. (-)
- I learned something new about writing from using the LLM. (+)
- I felt I had less control of the essay writing process when working with the LLM. (-)

- The model took too much initiative in generating content. (-)
- I felt that the LLM and I were collaborating as partners. (+)
- The model's behavior matched my preferred level of assistance. (+)
- I did not trust the LLM's writing suggestions. (-)
- I would not use this LLM again for a similar writing task. (-)
- Using the LLM made me question what counts as original writing. (-)
- I would disclose AI assistance if submitting this essay academically. (+)

B User Study Results

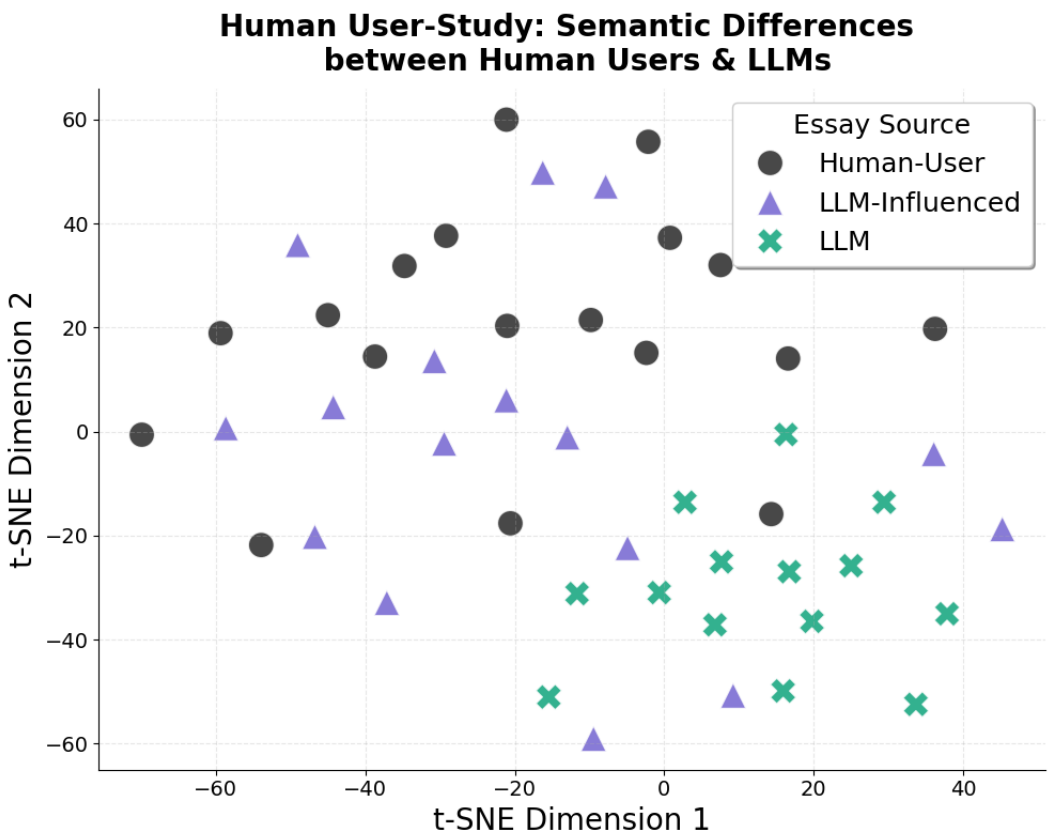
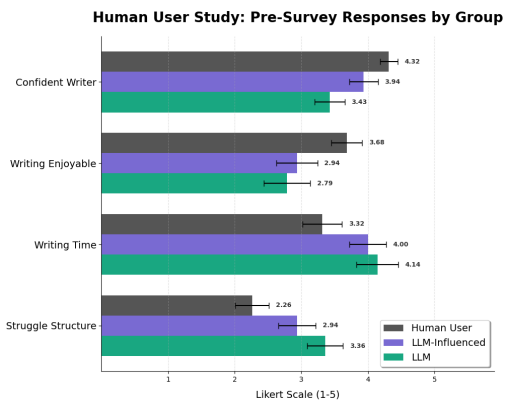
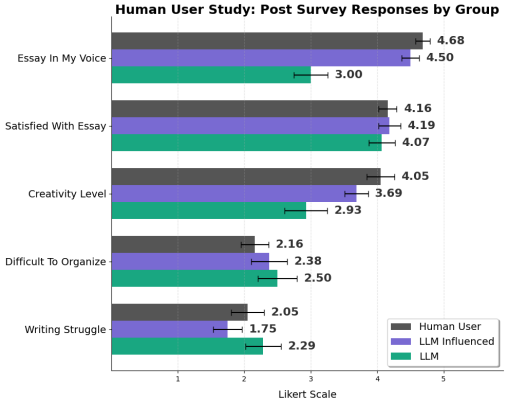


Figure 12: t-SNE visualization of semantic embeddings. AI essays form a distinct cluster, while AI-influenced essays lie between human and AI regions.

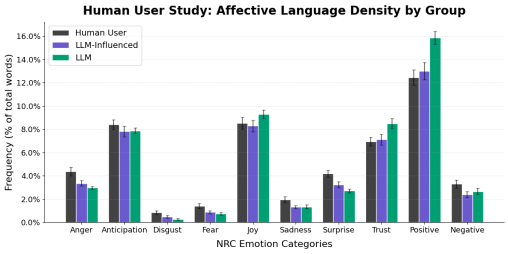


(a) Pre-study survey responses assessing confidence, enjoyment, time, and structural difficulty.

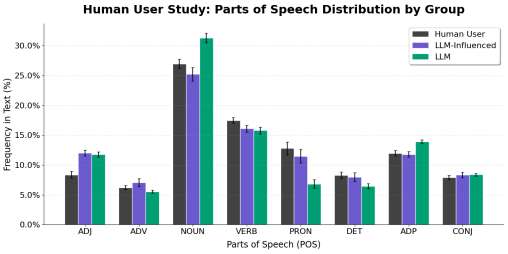


(b) Post-study survey responses evaluating perceived voice, creativity, satisfaction, and writing difficulty.

Figure 13: Pre- and post-study survey results by writing condition. Likert scale responses reveal shifts in confidence, creativity, and perceived writing difficulty after task completion.

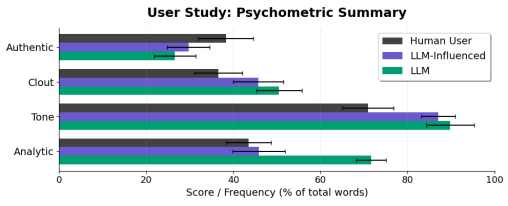


(a) Affective language density using NRC emotion categories.

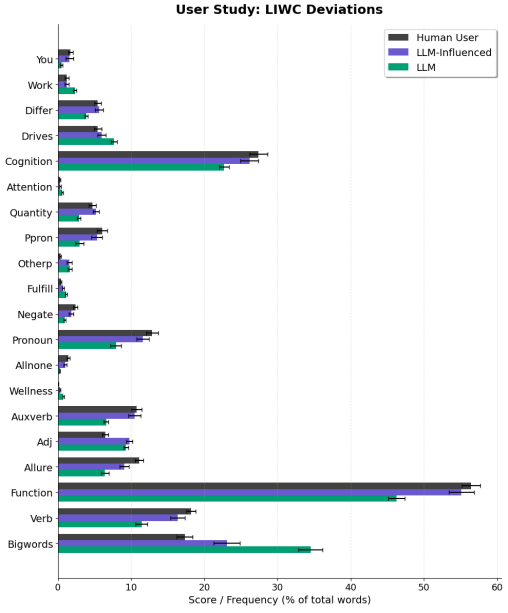


(b) Parts-of-speech distribution by writing condition.

Figure 14: Emotional and syntactic characteristics of essays across writing conditions.



(a) Psychometric dimensions including authenticity, clout, tone, and analytic reasoning.



(b) LIWC linguistic category frequencies across essay groups.

Figure 15: Core linguistic differences between human, AI-influenced, and AI-generated essays.

C ARGREWRITE-V2 ANALYSIS

C.1 GENERATING LLM DRAFTS

LLM Draft Generation Prompts

The following prompts were used to generate LLM drafts:

1. **Expert**: You are given an essay draft. Revise the essay.
2. **Minimal**: You are given an essay draft. Rewrite this essay and keep to a similar word count.
3. **Grammar**: You are given an essay draft. Revise this essay for grammar.
4. **Completion**: You are given an essay draft. The text is only the first paragraph. Finish the rest of the essay.
5. **Expansion**: You are given an essay draft. Expand on the ideas in the following draft with more detail and depth.

C.2 SEMANTIC SHIFTS WITH GEMINI-004 EMBEDDING

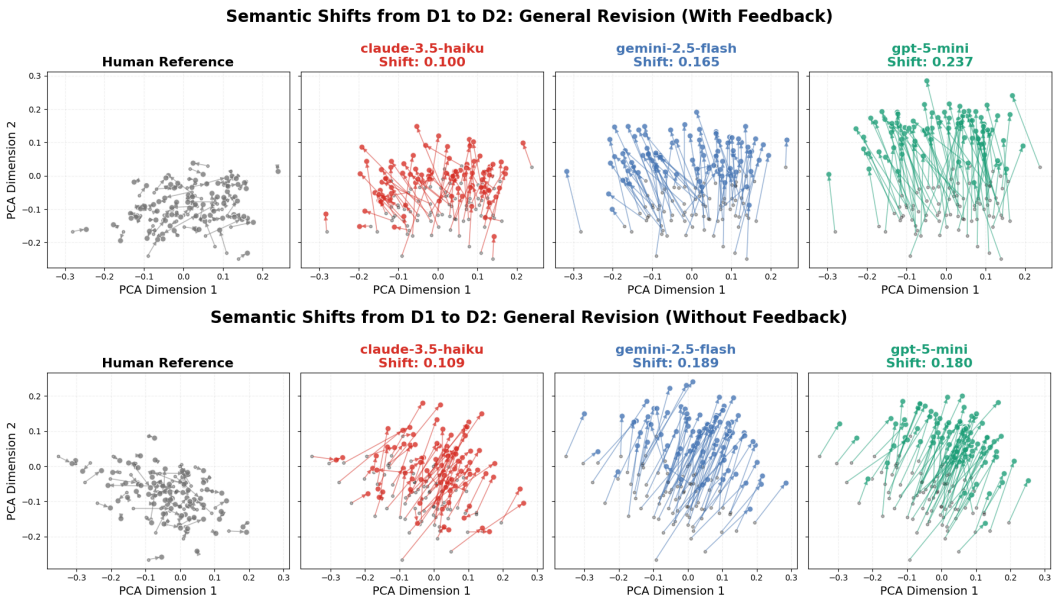


Figure 16: Semantic shifts from D1 to D2 for **General revisions** with **Gemini-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

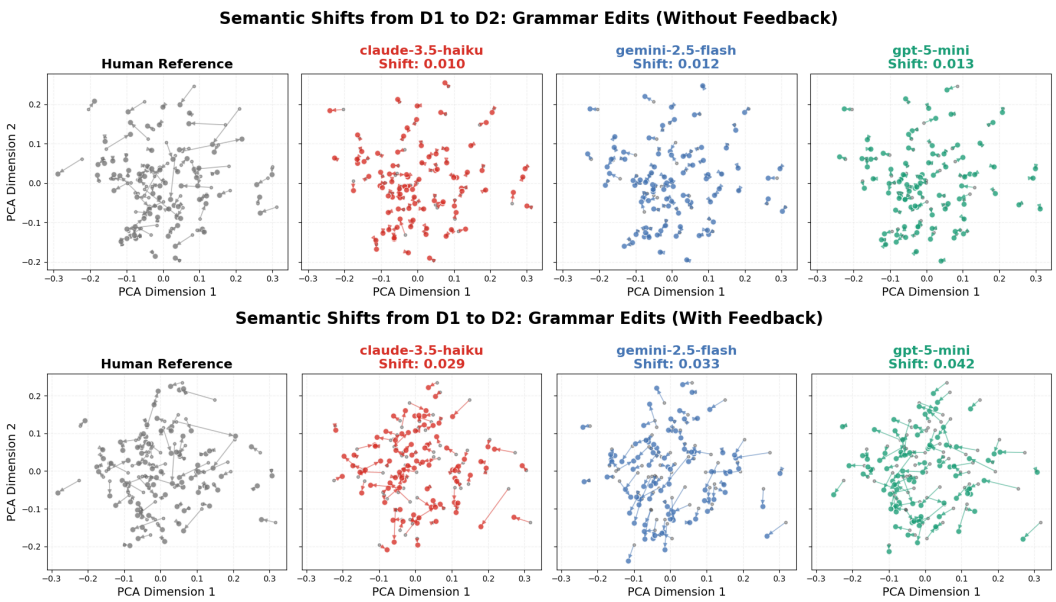


Figure 17: Semantic shifts from D1 to D2 for **Grammar edits** with with **Gemini-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

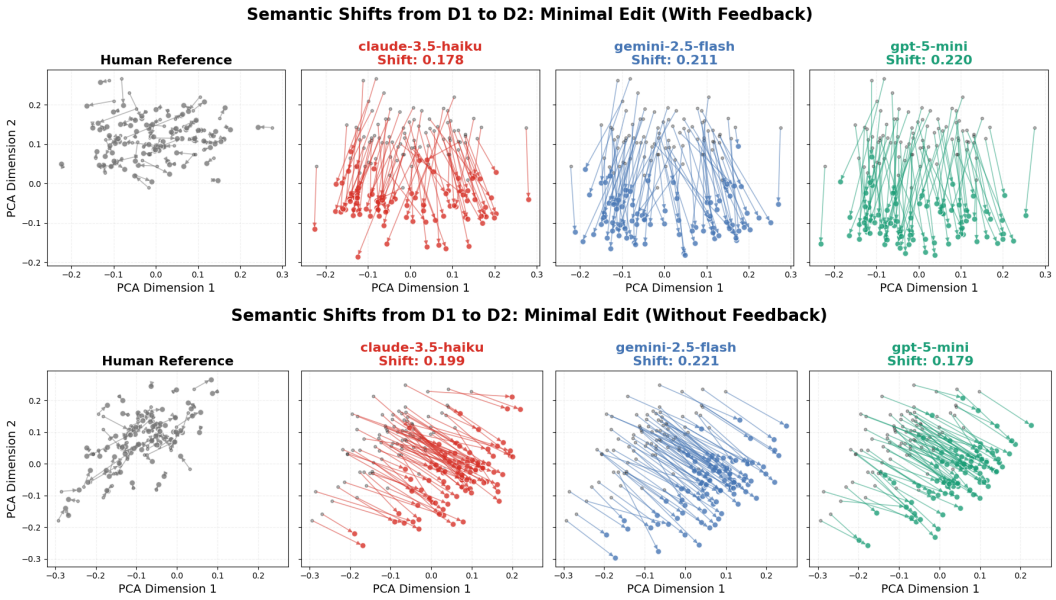


Figure 18: Semantic shifts from D1 to D2 for**Miminal revisions** with **Gemini-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

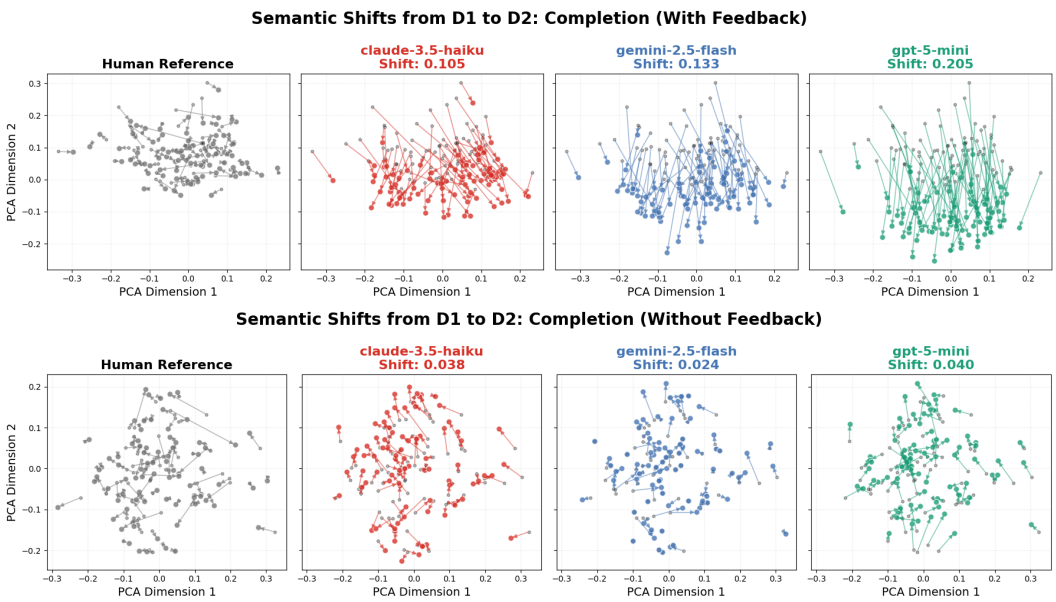


Figure 19: Semantic shifts from D1 to D2 for **Completion revisions** with **Gemini-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

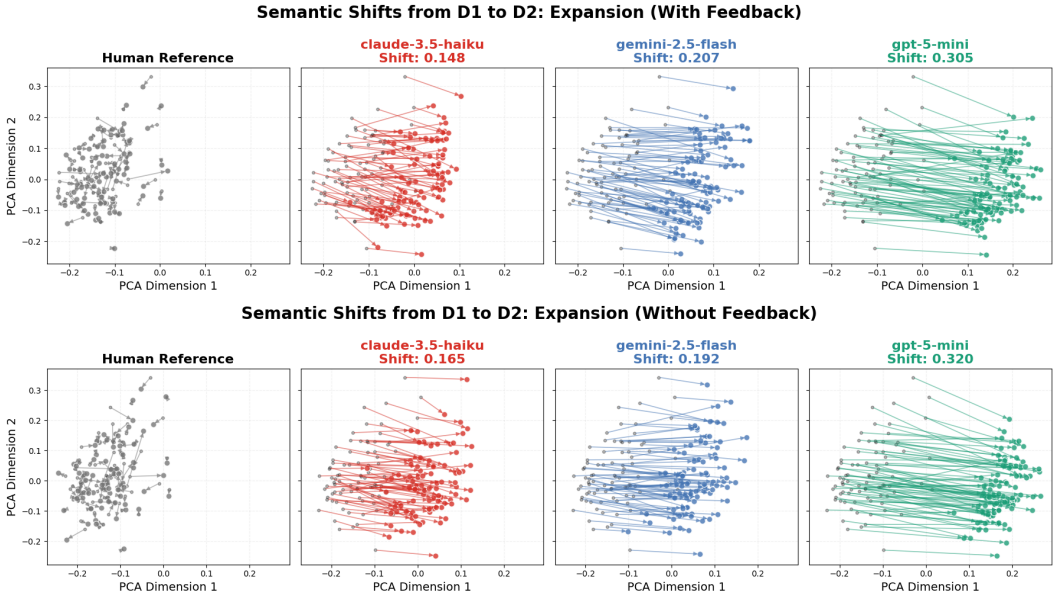


Figure 20: Semantic shifts from D1 to D2 for **Expansion revisions** with **Gemini-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

C.3 SEMANTIC SHIFTS WITH MINILM-L6-V2 EMBEDDING

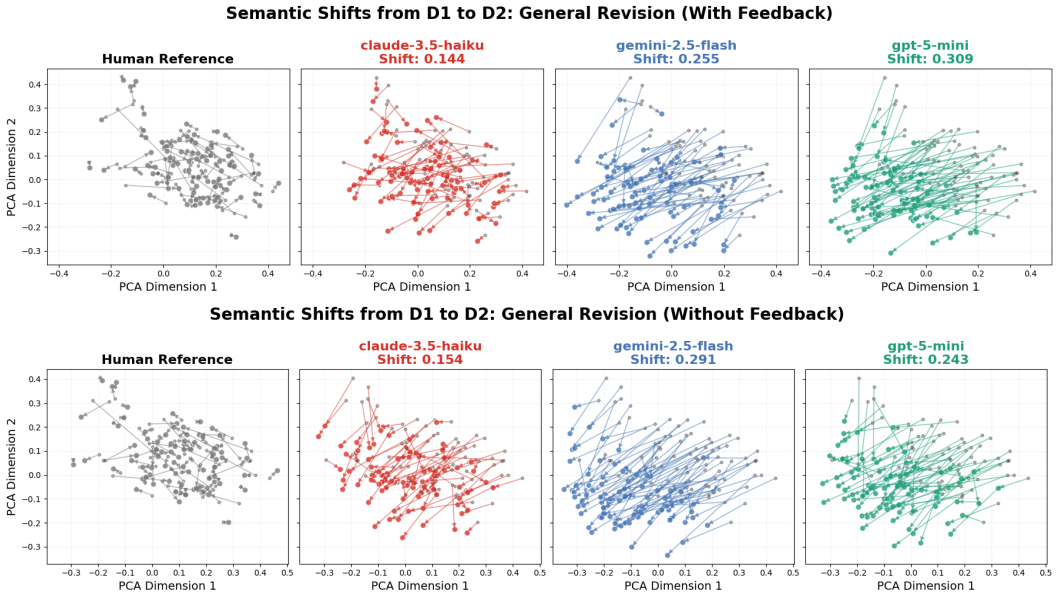


Figure 21: Semantic shifts from D1 to D2 for **General revisions** with **MiniLM-L6-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

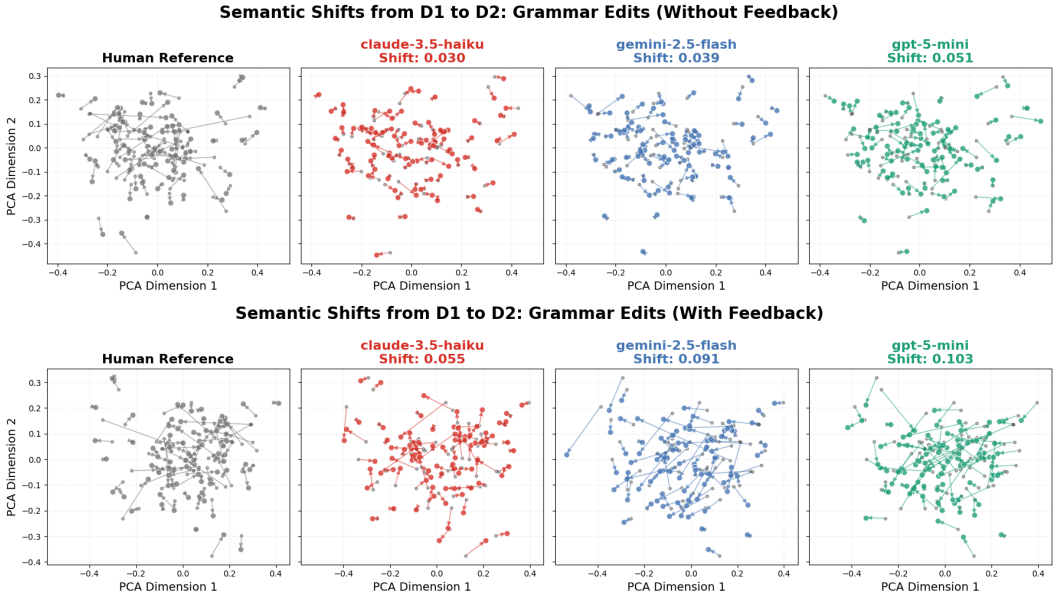


Figure 22: Semantic shifts from D1 to D2 for **Grammar edits** with **MiniLM-L6-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

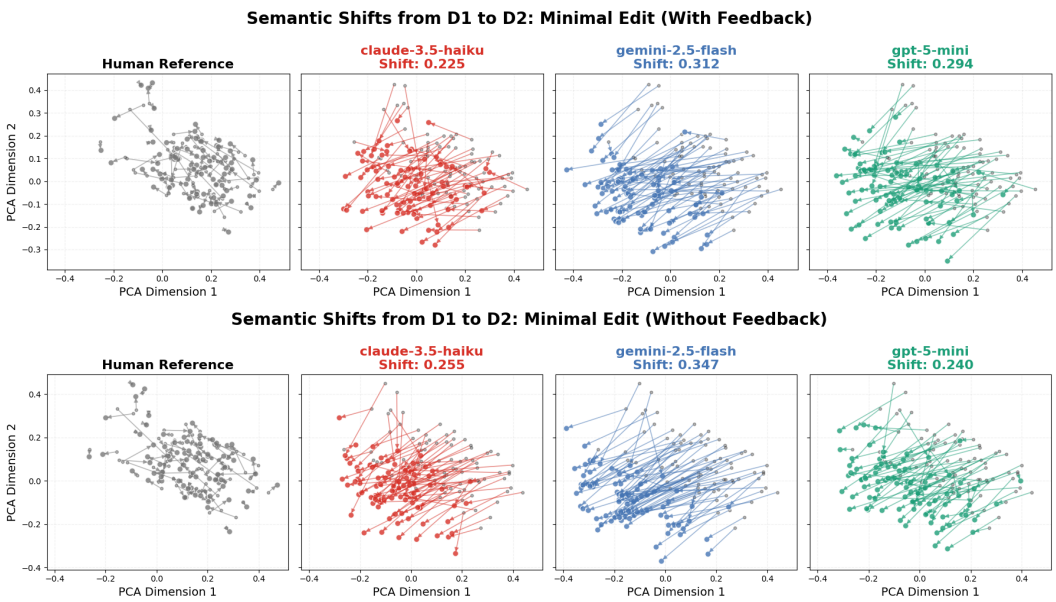


Figure 23: Semantic shifts from D1 to D2 for**Miminal revisions** with **MiniLM-L6-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

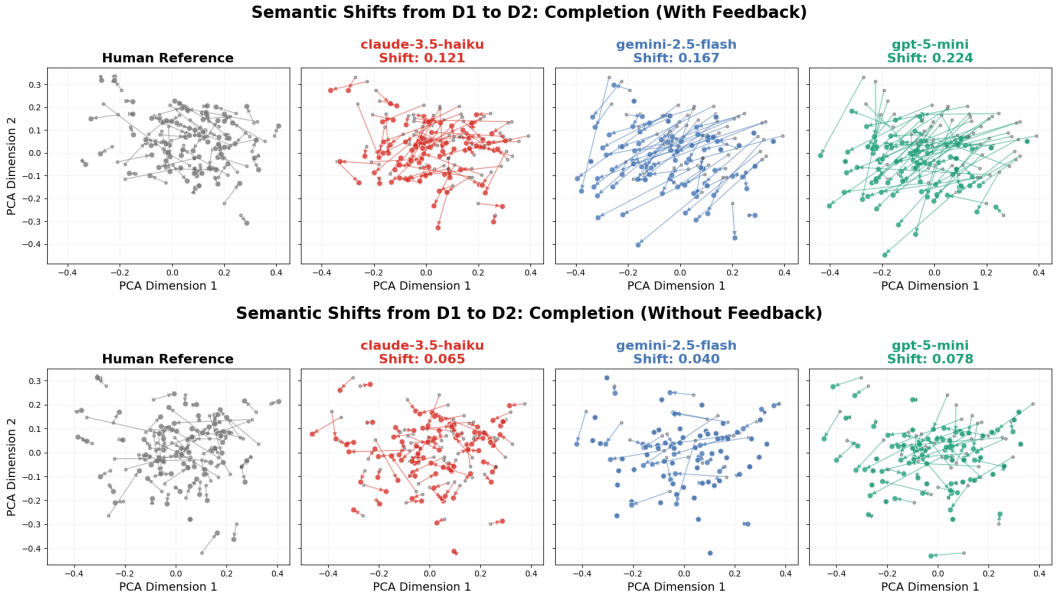


Figure 24: Semantic shifts from D1 to D2 for **Completion revisions** with **MiniLM-L6-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

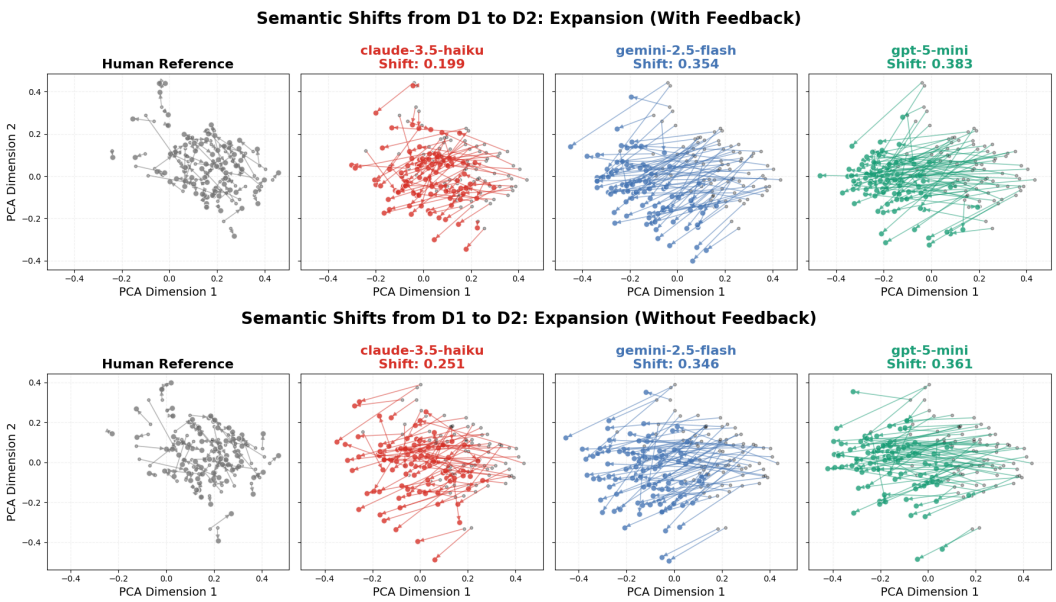


Figure 25: Semantic shifts from D1 to D2 for **Expansion revisions** with **MiniLM-L6-004 Embedding**. Top: with expert feedback. Bottom: without expert feedback.

C.4 Semantic Shifts Across Settings / Model

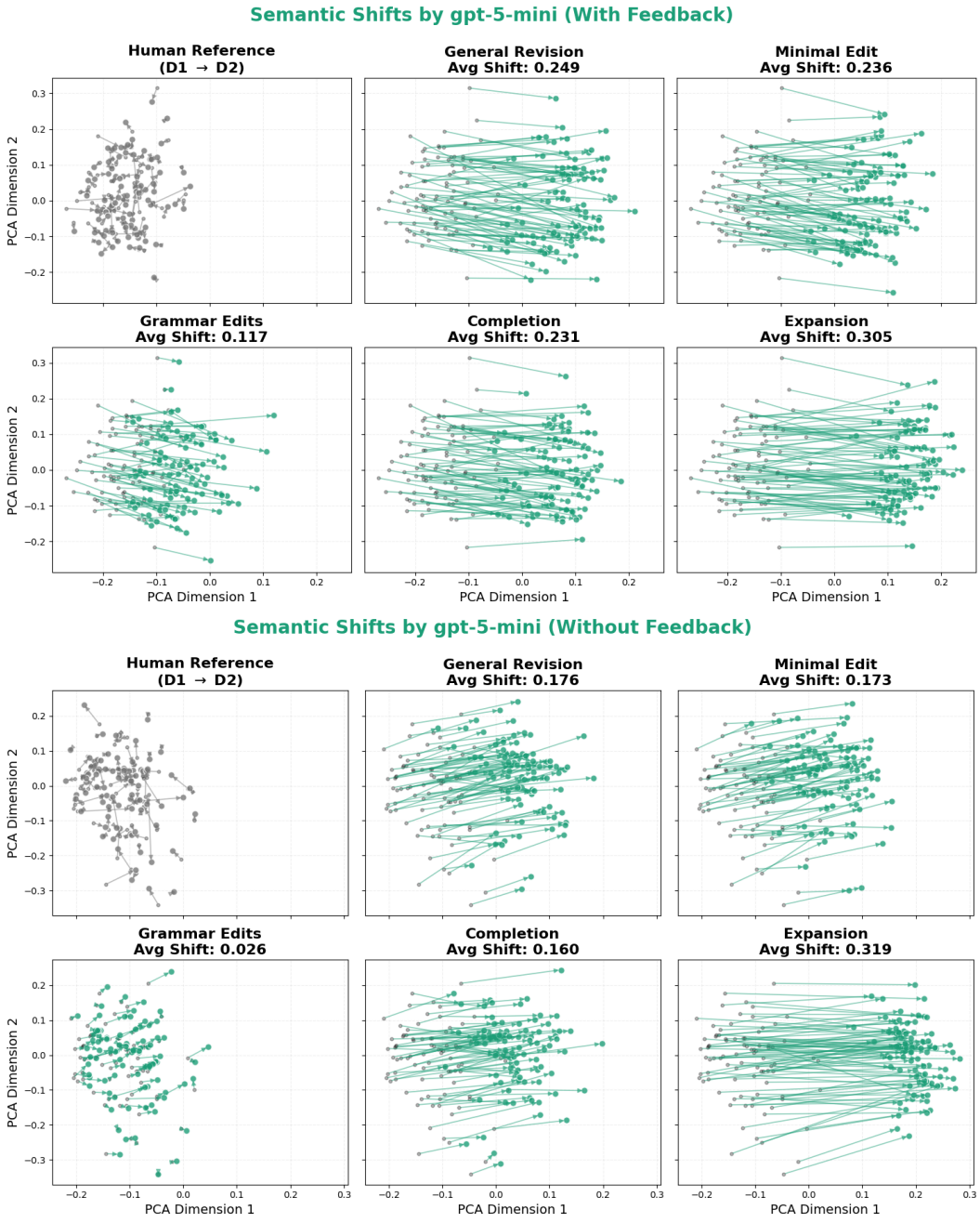


Figure 26: Semantic shifts from D1 to D2 produced by **GPT-5-mini**. Top: revisions with expert feedback. Bottom: revisions without feedback.

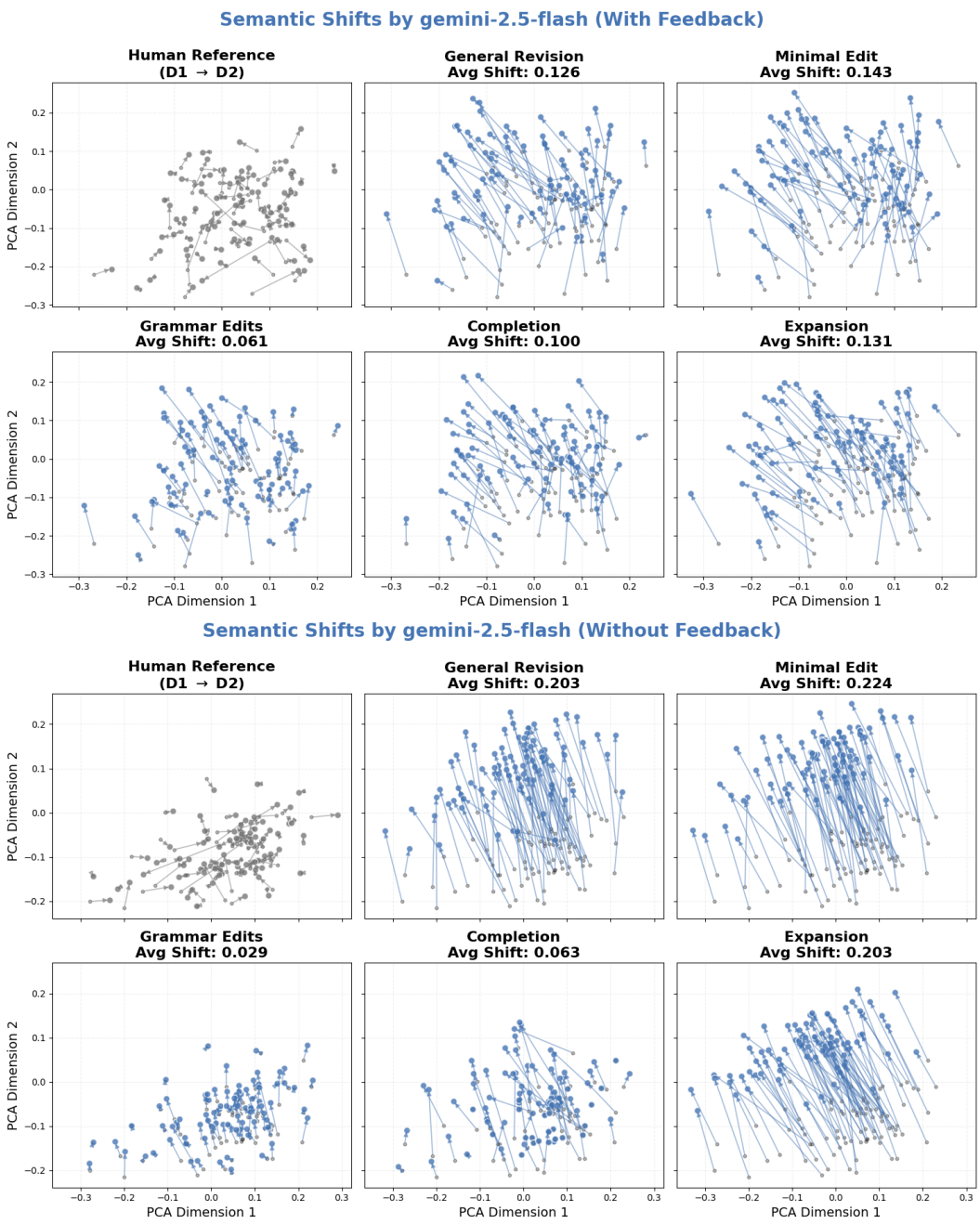


Figure 27: Semantic shifts from D1 to D2 produced by **Gemini-2.5-Flash**. Top: revisions with expert feedback. Bottom: revisions without feedback.

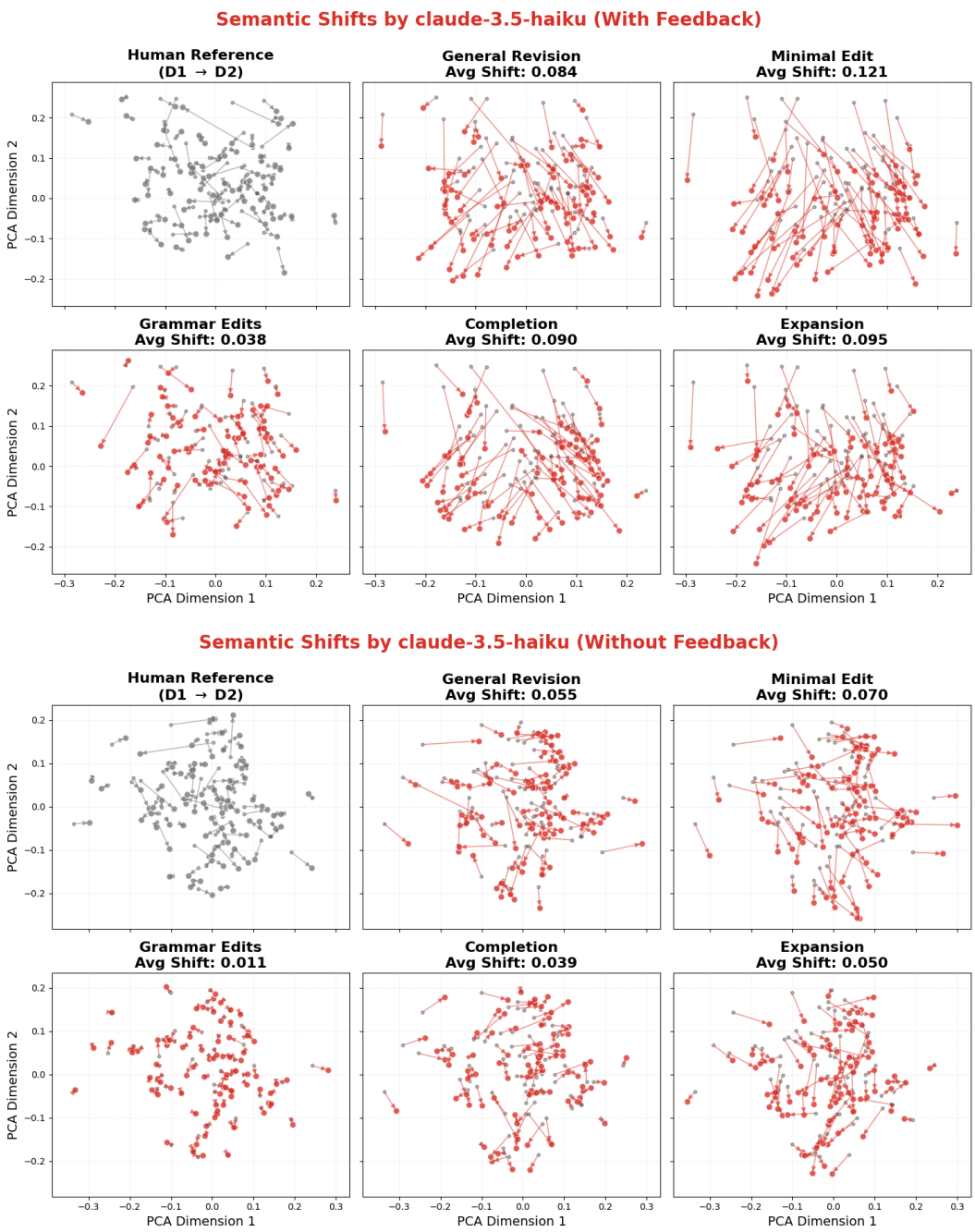


Figure 28: Semantic shifts from D1 to D2 produced by **Claude-3.5-Haiku**. Top: revisions with expert feedback. Bottom: revisions without feedback.

C.5 JENSEN-SHANNON DIVERGENCE

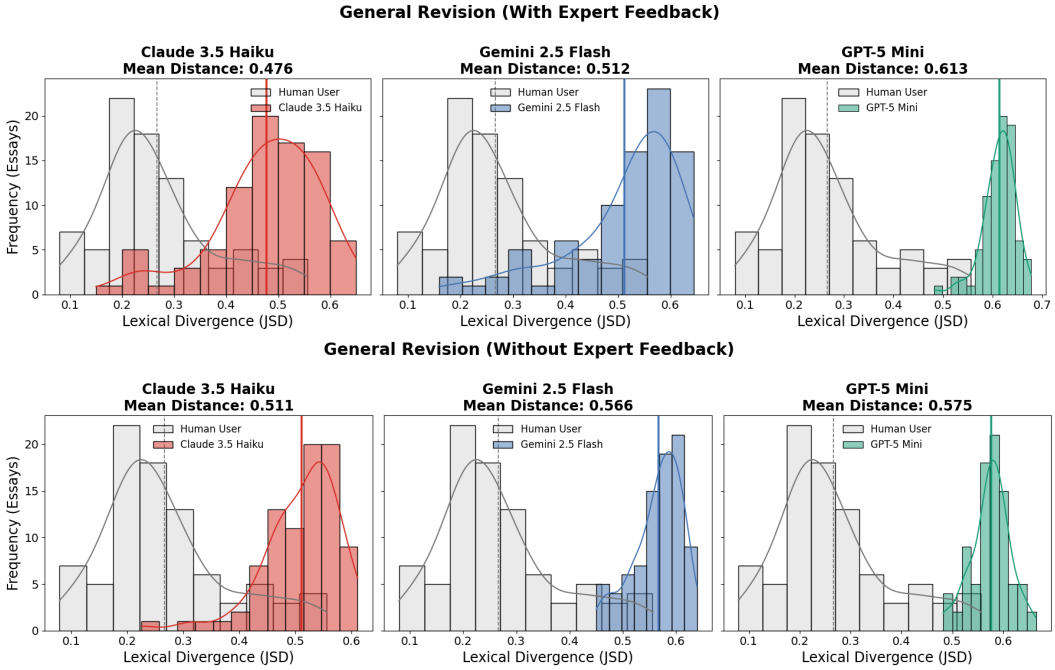


Figure 29: Lexical divergence (JSD) distributions for **General revisions**. Top: with expert feedback. Bottom: without expert feedback.

Figure 30: Lexical divergence (JSD) distributions for **Grammar revisions**. Top: with expert feedback. Bottom: without expert feedback.

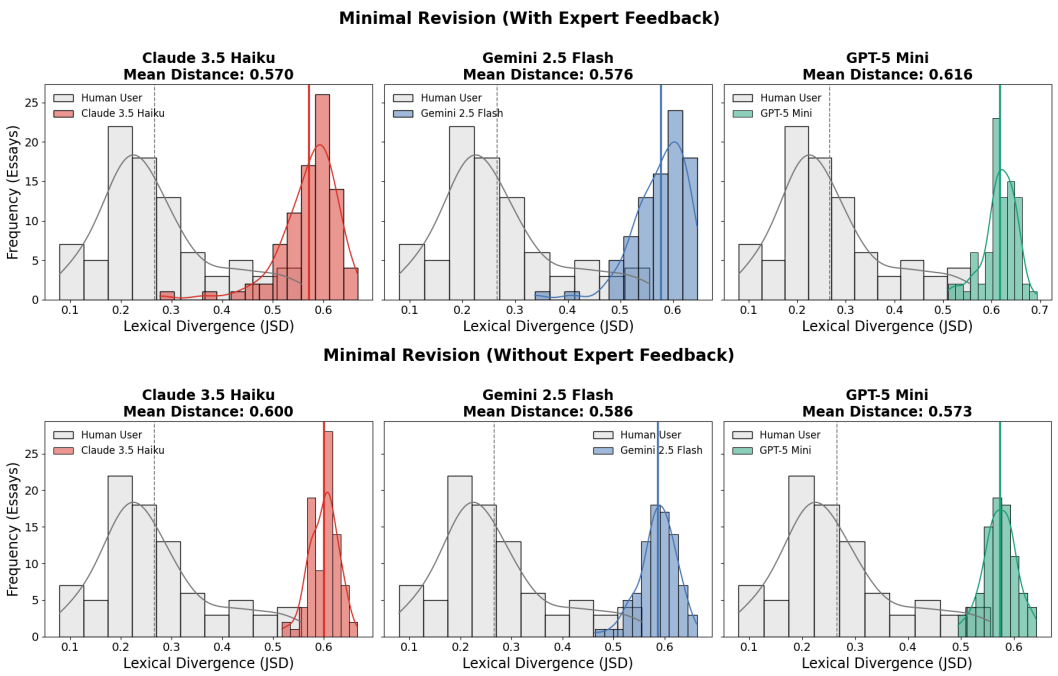


Figure 31: Lexical divergence (JSD) distributions for **Minimal revisions**. Top: with expert feedback. Bottom: without expert feedback.

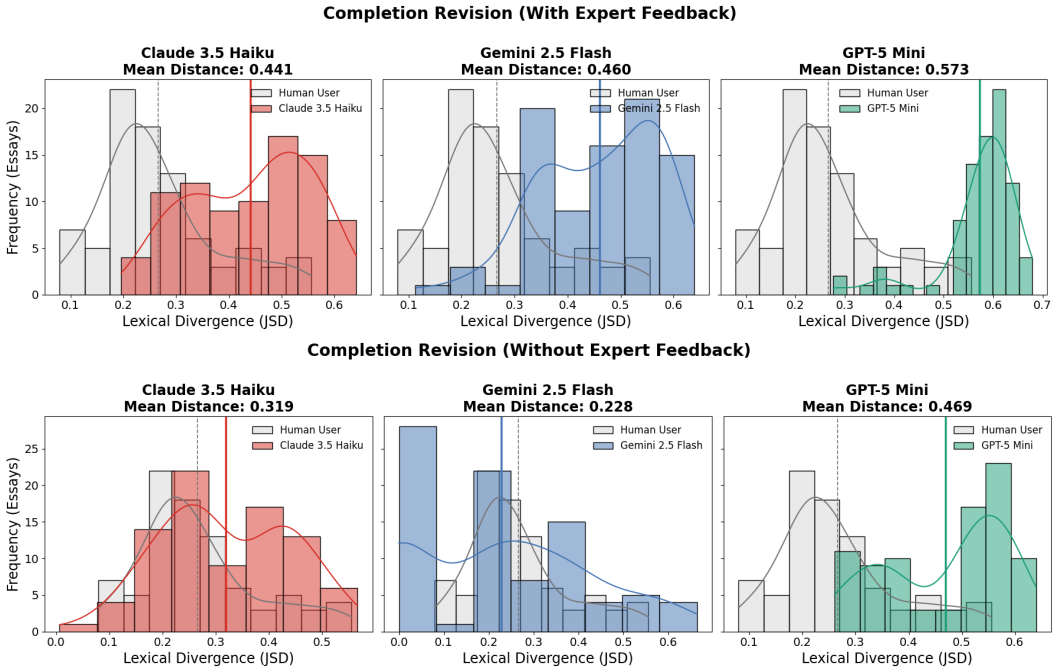


Figure 32: Lexical divergence (JSD) distributions for **Completion revisions**. Top: with expert feedback. Bottom: without expert feedback.

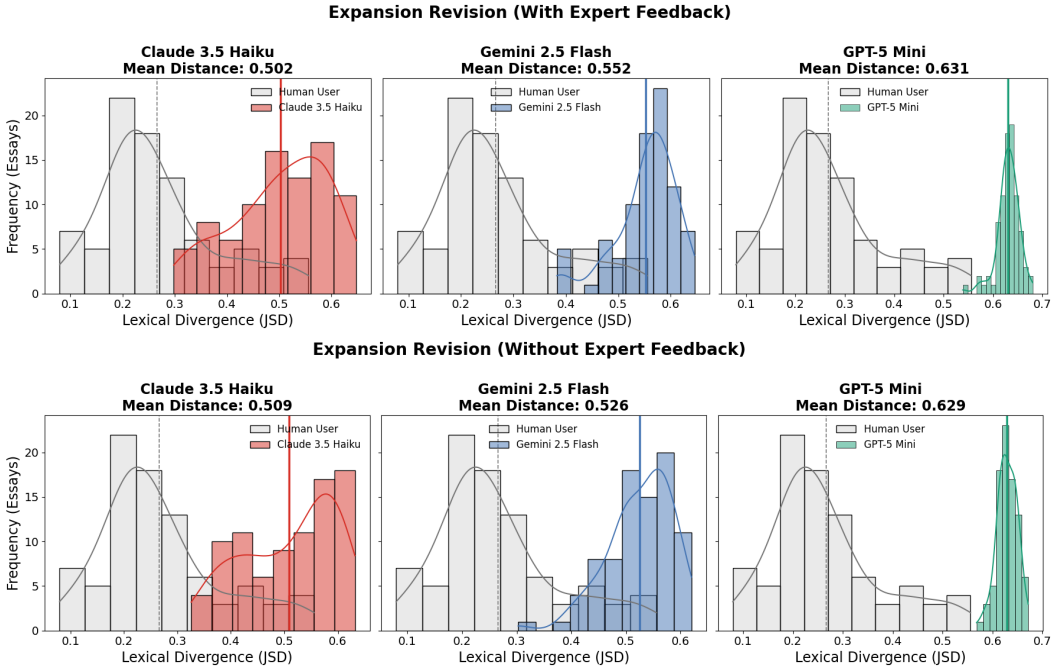


Figure 33: Lexical divergence (JSD) distributions for **Expansion revisions**. Top: with expert feedback. Bottom: without expert feedback.

C.6 POS Distribution

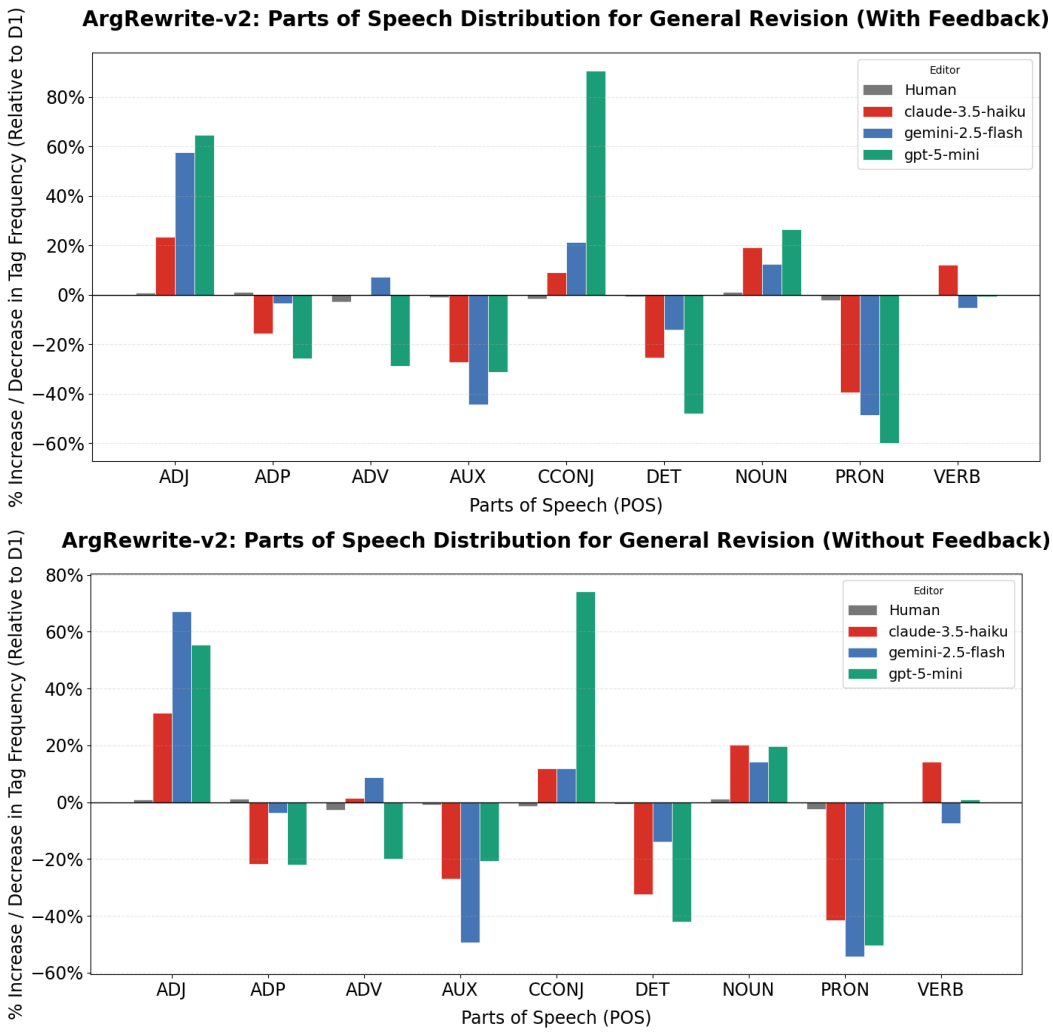


Figure 34: Grammatical shifts (POS) for **General revisions**. Top: with expert feedback. Bottom: without feedback.

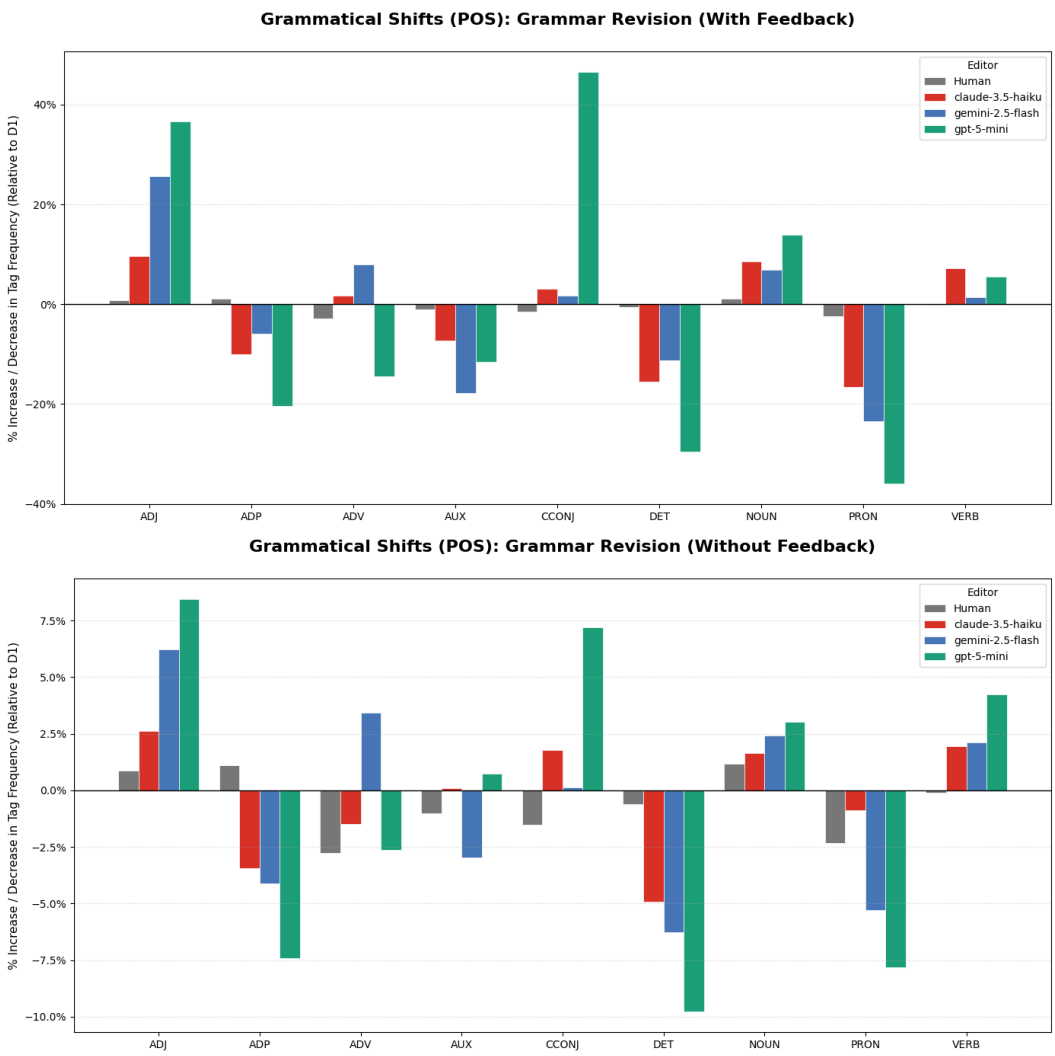


Figure 35: Grammatical shifts (POS) for **Grammar revisions**. Top: with expert feedback. Bottom: without feedback.

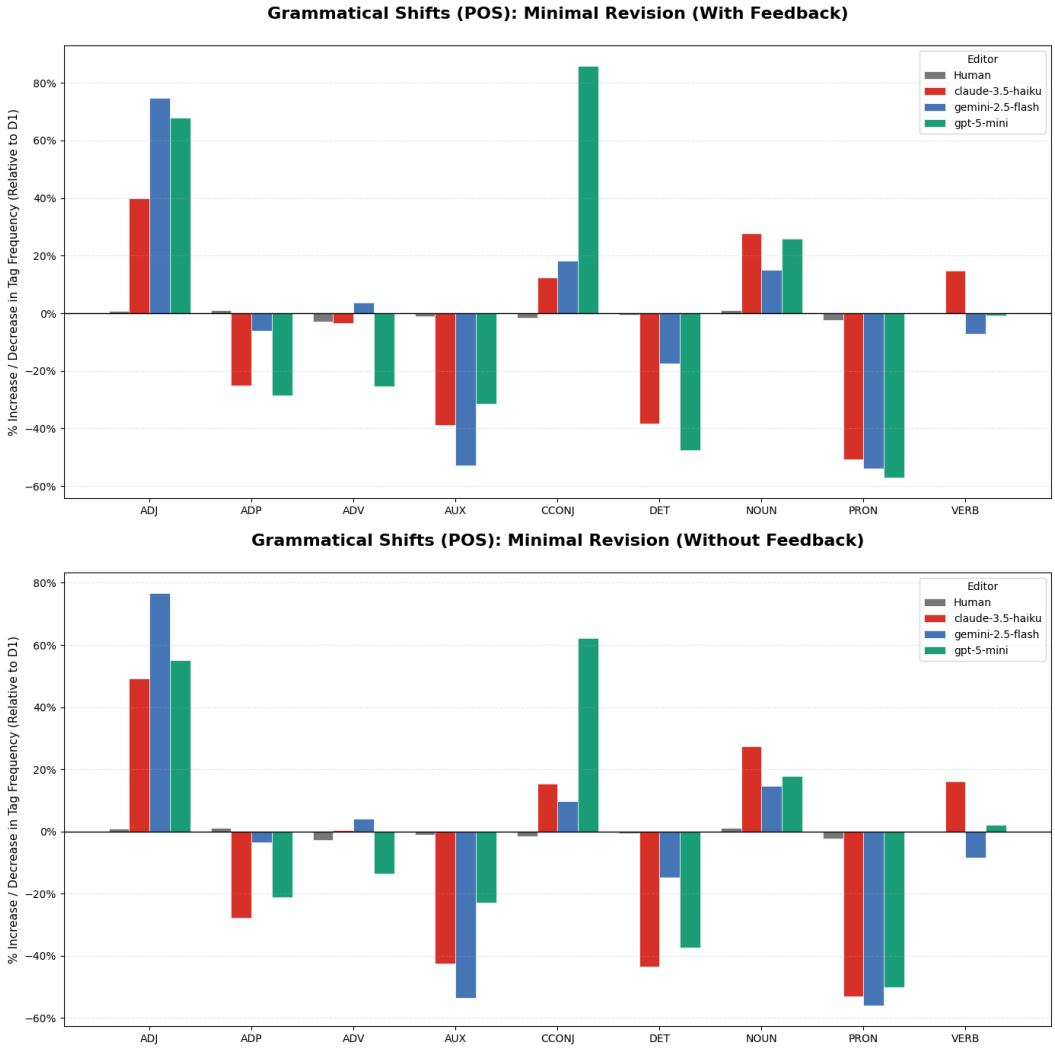


Figure 36: Grammatical shifts (POS) for **Minimal revisions**. Top: with expert feedback. Bottom: without feedback.

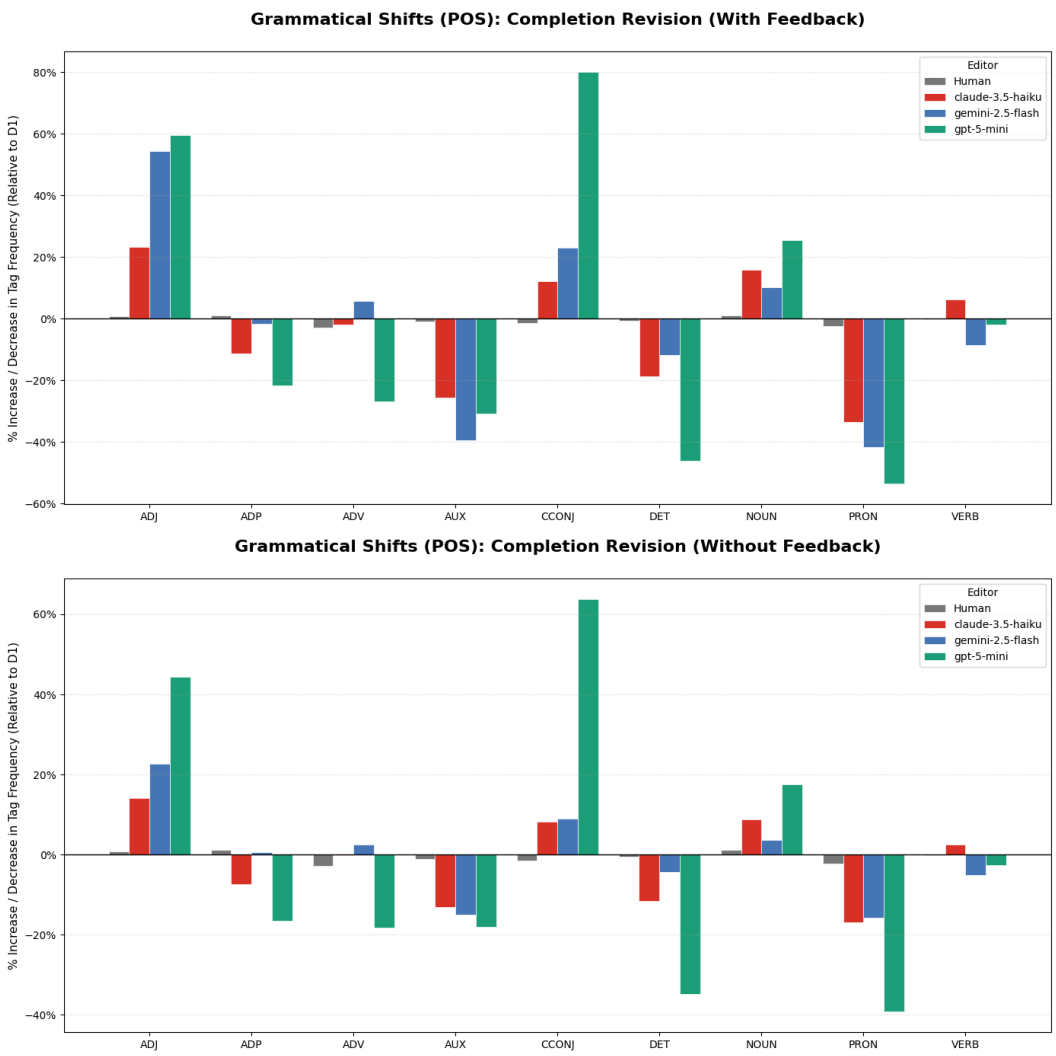


Figure 37: Grammatical shifts (POS) for **Completion revisions**. Top: with expert feedback. Bottom: without expert feedback.

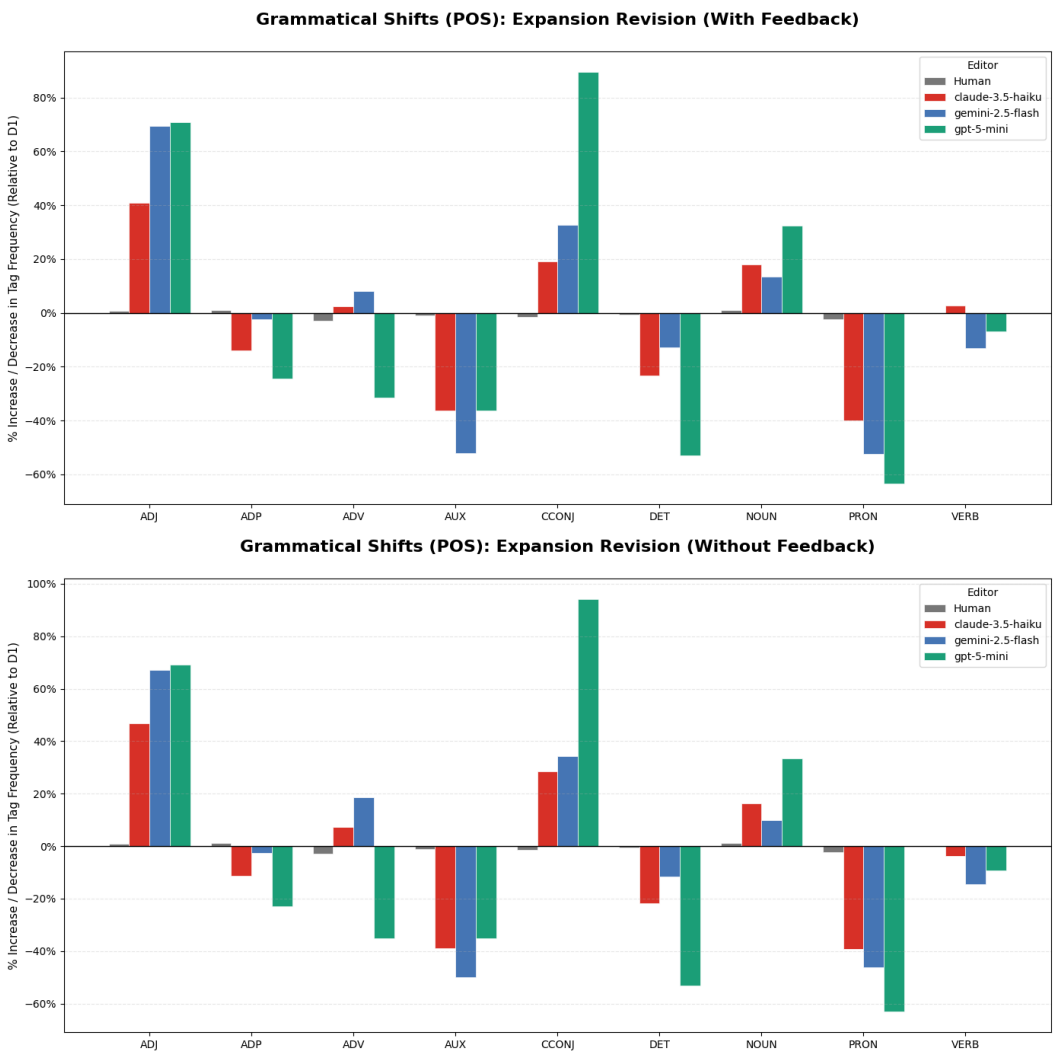


Figure 38: Grammatical shifts (POS) for **Expansion revisions**. Top: with expert feedback. Bottom: without expert feedback.

C.7 Emotional Shift

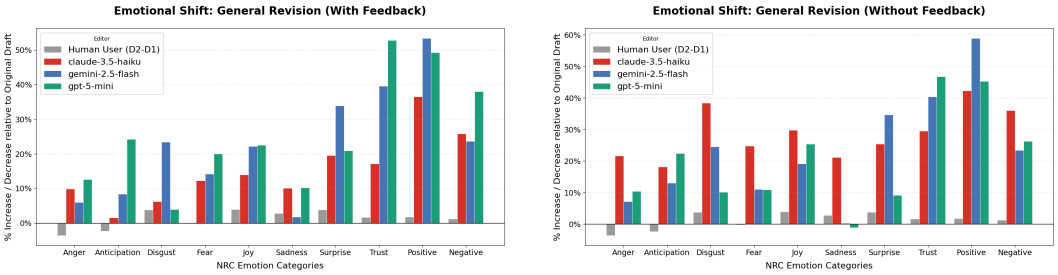


Figure 39: Emotional shifts from D1 to D2 for **General revisions**. Left: with expert feedback. Right: without expert feedback.

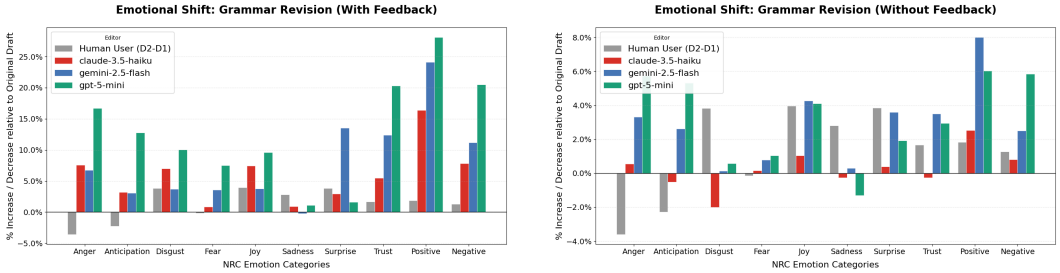


Figure 40: Emotional shifts from D1 to D2 for **Grammar revisions**. Left: with expert feedback. Right: without expert feedback.

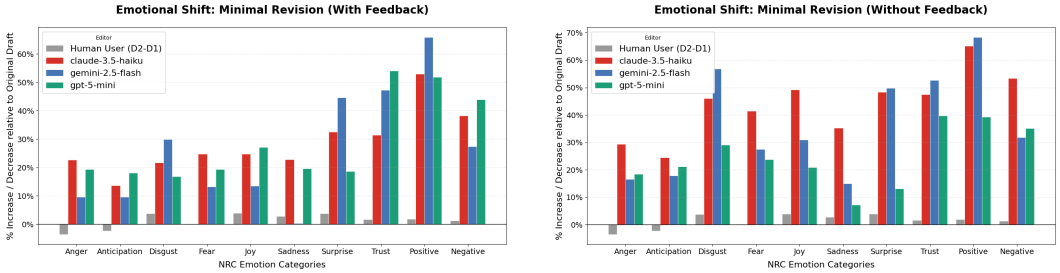


Figure 41: Emotional shifts from D1 to D2 for **Minimal revisions**. Left: with expert feedback. Right: without expert feedback.

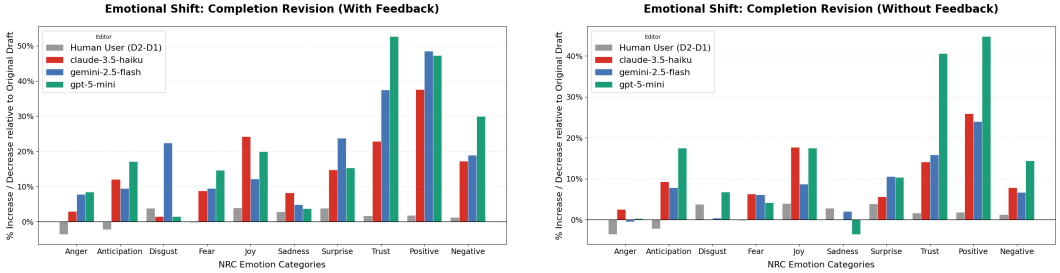


Figure 42: Emotional shifts from D1 to D2 for **Completion revisions**. Left: with expert feedback. Right: without expert feedback.

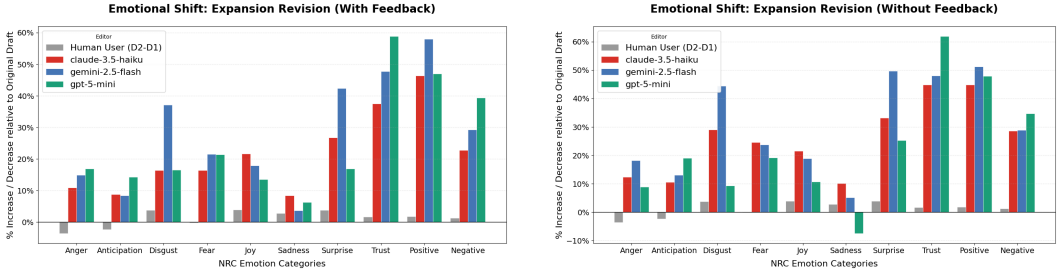


Figure 43: Emotional shifts from D1 to D2 for **Expansion revisions**. Left: with expert feedback. Right: without expert feedback.

C.8 ICLR Review Analyses

For ICLR reviews, the possible scores for ICLR papers are: 0: strong reject, 2: reject, 4: borderline reject, 6: borderline accept, 8: accept, 10: strong accept. With each score, there is a confidence level from 1 to 5 on how confident the reviewer is, where 1 represents low confidence and 5 represents absolute certainty.

D Arguments Analysis

In Figure 44, we use LLM-as-a-Judge to determine the arguments used by the human users and find that LLMs are far more likely to use statistical and expert opinion arguments.

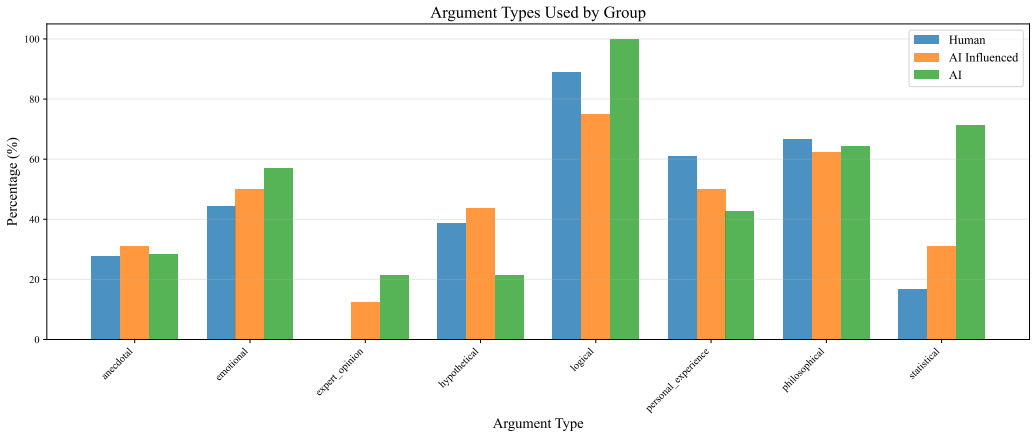


Figure 44: We display the results of using LLM-as-a-Judge to categorize the human study essays based on the argument type. Humans who interact with LLMs use expert opinion and statistical arguments significantly more frequently than humans who write the essay on their own.

E LLM-AS-A-JUDGE PROMPTS

E.1 ICLR REVIEW ANALYSIS PROMPTS

This prompts the LLM-as-a-Judge to create categories and definitions for the different categories.

```
Analyze the following {len(review_texts)} peer reviews and create a
    set of consistent categories for strengths and weaknesses.

Reviews (one per line, separated by "---"):
{chr(10).join([f"REVIEW {i+1}:{chr(10)}{text}{chr(10)}---"
for i, text in enumerate(review_texts)])}

Return your response as a JSON object with two arrays: "
   strength_categories" and "weakness_categories".

Example:
{
   "strength_categories": ["Clarity", "Novelty", "Strong Results"],
   "weakness_categories": ["Lack of Clarity", "Insufficient
       Experiments", "Outdated Methods"],
   "definitions": {
      "Clarity": "The paper is well-written and easy to understand
          .",
      "Lack of Clarity": "The paper is difficult to follow and
          poorly organized.",
      "Novelty": "The paper presents new and original ideas.",
      "Insufficient Experiments": "The experiments do not
          adequately support the claims made in the paper.",
      "Strong Results": "The results are compelling and demonstrate
           the effectiveness of the proposed method.",
      "Outdated Methods": "The methods used are not state-of-the-
          art."
   }
}
```

This prompts the LLM to reduce the extracted categories to reduce redundant categories.

Type B - Logic/conditional bug

```
    f"""Given the following list of categories, reduce redundancy
        by merging similar ones. Return a JSON array of unique
        categories.
    with a maximum of 15 categories. Do NOT provide more than 15.

    Give a description of the categories and make sure they are
        distinct, e.g. don't include both "Clarity" and "Motivation
        Clarity".
Categories:
{json.dumps(categories, indent=4)}
Example:
{{
    "reduced_categories": ["Clarity", "Novelty", "Strong Results",
        "Insufficient Experiments", "Outdated Methods"]
    "definitions": {{
        "Clarity": "How clear and understandable the paper is.",
        "Novelty": "The originality and innovativeness of the
            research.",
        "Strong Results": "The robustness and significance of the
            experimental results.",
        "Insufficient Experiments": "Lack of adequate experimental
            validation.",
        "Outdated Methods": "Use of methods that are no longer
            state-of-the-art."
    }}
}}
    """
```

This prompt is for extracting the top three strengths and weaknesses from the previously extracted strengths and weaknesses from the previous two prompts.

Type B - Logic/conditional bug

```
    f"""Analyze the following peer review and extract the key
        points.

    Review:
    {review_text}

    Please identify:
    1. Top 3 strengths mentioned in the review
    2. Top 3 weaknesses mentioned in the review

    Return your response as a JSON object with the following
        structure:
    {{
        "strengths": ["strength 1", "strength 2", "strength 3"],
        "weaknesses": ["weakness 1", "weakness 2", "weakness 3"]
    }}

    You make a pick from these strengths and weaknesses categories:

    Strengths:
    {json.dumps(strength_categories, indent=4)}

    Weaknesses:
    {json.dumps(weakness_categories, indent=4)}
```

```
    Do NOT invent strengths or weaknesses that are not in the above
        categories, and do not mention anything not in the review.

    Only return the JSON object, no additional text."""
```

F ICLR Additional Categories

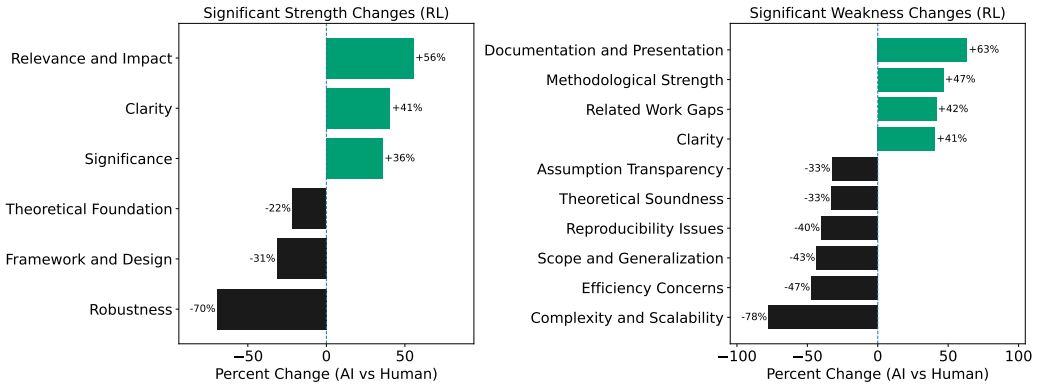


Figure 45: Analyzing reviews corresponding to 447 papers in the RL category, each with one review generated by an LLM and the other by a human.

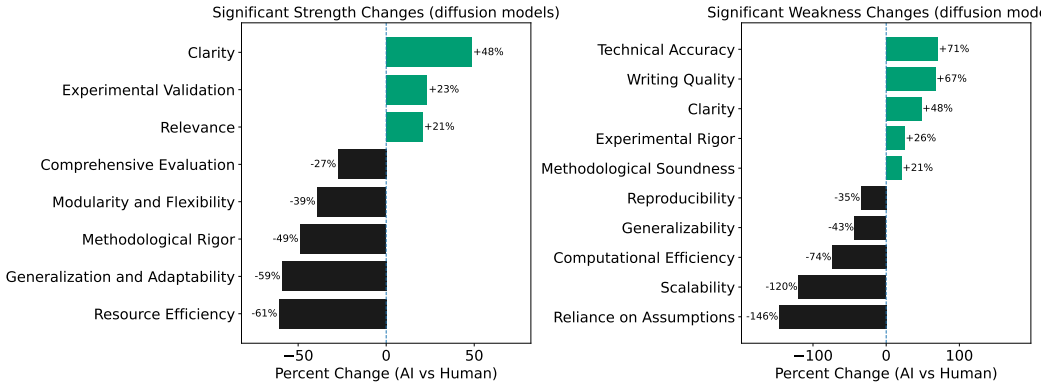


Figure 46: Analyzing reviews corresponding to 336 papers in the diffusion models category, each with one review generated by an LLM and the other by a human.