Prototypical Human-AI Collaboration Behaviors from LLM-Assisted Writing in the Wild

Sheshera Mysore[1][△]   **Debarati Das**[2]   **Hancheng Cao**[1]   **Bahar Sarrafzadeh**[1][△]

[1]Microsoft, WA, USA [2]University Of Minnesota, MN, USA

[△]Corresponding authors:

smysore@iesl.cs.umass.edu, bahar.sarrafzadeh@microsoft.com

Abstract

As large language models (LLMs) are used in complex writing workflows, users engage in multi-turn interactions to steer generations to better fit their needs. Rather than passively accepting output, users actively refine, explore, and co-construct text. We conduct a large-scale analysis of this collaborative behavior for users engaged in writing tasks in the wild with two popular AI assistants, Bing Copilot and WildChat. Our analysis goes beyond simple task classification or satisfaction estimation common in prior work and instead characterizes how users interact with LLMs through the course of a session. We identify prototypical behaviors in how users interact with LLMs in prompts following their original request. We refer to these as Prototypical Human-AI Collaboration Behaviors (PATHs) and find that a small group of PATHs explain a majority of the variation seen in user-LLM interaction. These PATHs span users revising intents, exploring texts, posing questions, adjusting style or injecting new content. Next, we find statistically significant correlations between specific writing intents and PATHs, revealing how users' intents shape their collaboration behaviors. We conclude by discussing the implications of our findings on LLM alignment.[1]

1 Introduction

LLMs' generalization ability and natural language interfaces have made powerful AI models accessible to a range of users engaged in diverse tasks (Ouyang et al., 2022). The logged natural language interactions from LLM-powered AI assistants have emerged as a rich data source for understanding user-AI interaction (Zhu et al., 2025). Leveraging this data, recent studies have explored the high-level tasks users engage in (e.g., search, coding, writing, etc.) (Tamkin et al., 2024), their variation across occupations (Handa et al., 2025), and
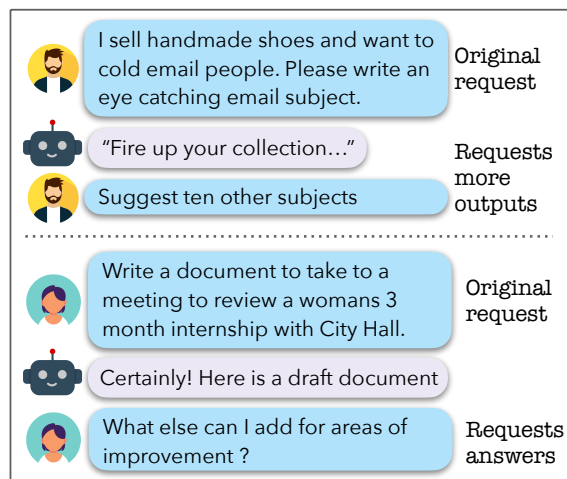

Figure 1: Users follow up their original requests to collaborate with LLMs in writing sessions. We identify prototypical human-AI collaboration behaviors (PATHs), and find statistically significant correlations between users' writing intents and PATHs.

measured user satisfaction based on interaction patterns (Lin et al., 2024). However, little prior work have examined how users collaborate with conversational LLMs in real-world deployments.

The conversational nature of current AI assistants enables users to engage in follow-up interactions after stating their original requests. The follow-ups allow users to articulate their needs better and obtain more helpful responses from the LLM (Figure 1). Analyzing these follow-up interactions promises to provide a rich characterization of human-AI collaboration and guide research on LLM alignment based on realistic in-the-wild interactions. We contribute such an analysis.

We focus on LLM-assisted writing for our analysis, a important and increasingly prevalent use case for conversational AI assistants (Tamkin et al., 2024; Suri et al., 2024). Recent studies have found that LLM-assisted writing is now common in impactful domains such as press releases, job post-

---

[1]Artifacts: https://prototypical-hai-collaborations.github.io/

ings, and peer reviews, among others (Liang et al., 2025, 2024a). Despite this, no prior research has systematically analyzed how users collaborate with AI assistants in the wild. We address this gap by formulating two key research questions: **(RQ1)** What high-level collaboration behaviors emerge from user interactions in AI-assisted writing? And **(RQ2)** How do these collaboration behaviors differ across writing intents?

To address these questions, we conduct a large-scale analysis of writing sessions from two AI assistants: Bing Copilot (Mehdi, 2023) and WildChat (Zhao et al., 2024a). Our datasets span 20.5M and 800k English user-LLM conversation sessions over seven and thirteen months of global Bing Copilot and WildChat usage, respectively. The two datasets enable us to identify shared collaboration behaviors across distinct AI assistants, and the public WildChat logs support reproducibility and future research. To answer RQ1, we use GPT-4o to classify users' follow-up utterances into high-level types and cluster them using Principal Component Analysis (PCA) (Bengio et al., 2013). Each cluster represents sessions with consistent collaboration behavior, which we term Prototypical Human-AI Collaboration Behaviors (PATH). To address RQ2, we use GPT-4o to identify writing intents from the users original requests and conduct regression analysis to correlate them with PATHs. This lets us detect statistically significant relationships between writing intents and collaboration behaviors.

Takeaways: We identify seven PATHs that capture 80-85% of variance across datasets, with shared behaviors like revising intents, exploring texts, asking questions, or modifying generations despite differences in deployments. Correlating writing intents and PATHs enables us to uncover intent-specific alignment needs. For instance, users sought to explore diverse generations in follow-ups for brainstorming eye-catching texts, indicating users need for LLMs aligned for brainstorming applications. Users generated long texts by staging generation and interactively providing feedback with different levels of specificity, indicating the need for session-level alignment from under-specific feedback. And in generating professional or technical texts, users followed up with questions aimed at learning about a domain's norms or to seek feedback, indicating the need to align LLMs for promoting learning in users. By analyzing collaborative writing behaviors in the wild, we offer insights to guide future research on LLM alignment.

2   Related Work

The rise of conversational LLMs has seen the emergence of user-LLM conversation log datasets (Kirk et al., 2024; Zhao et al., 2024a) and analysis, building on rich traditions of log analysis in HCI and Information Retrieval (Jansen and Spink, 2006; Dumais et al., 2014). This work has developed an understanding of how users interact with intelligent systems in-the-wild aiming to inform future system and model development. Such work has analyzed user-LLM interactions in the context of information seeking (Trippas et al., 2024), theorem proving (Collins et al., 2024a), image generation (Palmini et al., 2024; Vodrahalli and Zou, 2024), and writing assistance (Lee et al., 2022). Most relevant is the prior work on log analysis for UI based writing assistance (Lee et al., 2022; Sarrafzadeh et al., 2021). Respectively, they analyze model outputs, collaboration patterns, and the ability of interactions to predict the different stages of writing in UI based applications. Our work differs in its examination of conversation logs to uncover collaboration behaviors. Our focus on user-LLM conversations ties to recent large scale analysis of such logs to infer high level task intents for conversations (Tamkin et al., 2024; Suri et al., 2024). We extend this beyond analyzing task intents alone by focusing on collaboration behaviors captured in users' follow-up utterances. Further, correlating collaboration behaviors with intents enables us to uncover meaningful implications for LLM alignment (§7). Finally, Collins et al. (2024a) present a notable exception in conducting a small-scale qualitative analysis of user-LLM collaboration behaviors in mathematical theorem proving. Similar to our findings (§6.1,6.2), they find users to engage in exploration and question asking behaviors. Our work differs in its focus on writing, at scale analysis of in-the-wild interactions and, correlating writing intents with collaboration behaviors. We review related work on LLM-assisted writing with a human-centered focus and user satisfaction estimation from logs in Appendix A, and further discuss relevant studies in the context of our results in Section 6.

3   Analysis Setup

Our analysis is based on user-LLM conversational logs from Bing Copilot (Mehdi, 2023, BCP) and WildChat-1M (Zhao et al., 2024a, WC). The two systems vary in their base LLMs, interfaces, and

	Users	Countries	Top 5 countries
BCP_{Wr}	202k	219	US 30%, IN 15%, GB 6% PH 6%, AU 6%
WC_{Wr}	22k	166	US 25%, GB 10%, RU 9% IN 5%, PH 4%

Table 1: The number of users, countries, and countries where sessions originate (in %) in BCP_{Wr} and WC_{Wr}. They have 250k and 68k sessions respectively.

user bases and allow our analysis to identify shared behaviors likely to hold beyond deployments. Further, the public WildChat-1M dataset enables reproducibility and future work based on our analysis. For our analysis, we focus on English sessions engaged in writing tasks, excluding sessions focused on tasks like search or software development. We conceptualize writing sessions broadly to be the ones where users generated inter-personal or public communicative texts, technical texts, creative texts, and those focused on summarization. We treat complete generations and rewrites of whole or parts of texts as writing. We operationalize our definition in an iteratively developed and manually validated GPT-4o based multi-label Task Classifier (f_{CoarseT}) and use it to identify sessions focused on writing. We refer to the writing log datasets as BCP_{Wr} and WC_{Wr}. Appendix B details both datasets, their filtering, and f_{CoarseT}.

Bing Copilot - BCP_{Wr} We construct BCP_{Wr} from a daily random sample of sessions from Bing Copilot gathered from April-Oct 2024, resulting in 20.5M sessions. To enable the study of user-LLM collaboration, we ensure that sessions contain users' follow-up utterances and retain sessions with at least 2 user utterances and those in English. On the resulting 2.8M sessions, we run f_{CoarseT} to identify writing sessions and retain 250k sessions in BCP_{Wr}. GPT-4 powered all interactions in BCP.

WildChat - WC_{Wr} We follow a similar procedure to construct WC_{Wr} from the public WildChat-1M. Zhao et al. (2024a) gathered in a research study from April 2023 to May 2024. We retain sessions with at least 2 user utterances and English sessions. This results in 160k sessions of which 68k are identified as writing sessions by f_{CoarseT}. GPT-4 and GPT-3.5-Turbo powered the interactions in WC.

In Table 1, we see that the resulting BCP_{Wr} and WC_{Wr} contain sessions from a large group of users from over 150 countries. While many sessions originate in English-speaking countries, we observe a long tail of countries. In Table 5, we present ses-
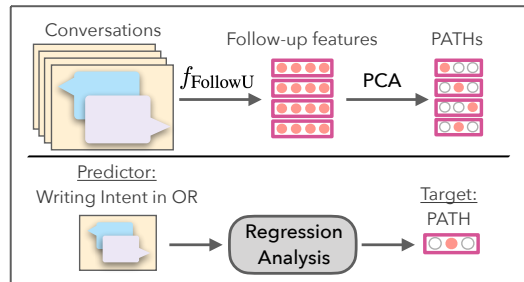

Figure 2: Our analysis methods aim to discover PATHs (above) and identify how PATHs differ across writing intents in original requests (ORs) (below).

sion length characteristics, and find that sessions had 2 follow-ups on average and that half of all sessions contain only one writing intent.

4 Log Analysis with PATHs

Our analysis centers around two research questions: (RQ1) What high-level collaboration behaviors emerge from user interactions in LLM-assisted writing? And (RQ2) How do these collaboration behaviors differ across writing intents? We overview our analysis method here (see Figure 2) and include detailed descriptions in Appendix C.

Identifying PATHs. To answer RQ1 we classify user utterances into "Original Requests" (OR) and set of follow-up types (Table 2) using a GPT-4o based follow-up classifier (f_{FollowU}). ORs represent the primary writing goal of a user in a session, and follow-up types categorize users' follow-ups into higher-level behaviors (e.g., Figure 1). f_{FollowU} was developed iteratively and manually validated to ensure its accuracy for subsequent analysis (see Appendix C.1). Then, we identify co-occurring patterns of follow-up types using Principal Component Analysis (PCA) (Bengio et al., 2013). We take each principal component to represent a PATH.

Specifically, we represent a dataset's sessions $\mathcal{S}$ with a "tf-idf" representation of its follow-up types ($\mathbf{F}$) and run PCA on $\mathbf{F}$. This transforms it as $\mathbf{P} = \mathbf{FW}$ where each dimension of $\mathbf{P}$ represents a mutually co-occurring set of follow-up types and segments $\mathcal{S}$ into subsets of sessions with a consistent PATH. We retain the first l dimensions of $\mathbf{P}$ that explain 80-85% variance in $\mathbf{F}$, treating the rest as noise. Our use of PCA follows its standard use for exploratory data analysis (Eagle and Pentland, 2009; Reades et al., 2009), its linear nature enables us to easily visualize how follow-up types combine to form PATHs (through $\mathbf{W}$), and its closed form

User utterance type	Description
ORIGINAL REQUEST	User makes a new request
RESTATES REQUEST	Reformulates their request
ELABORATES REQUEST	Expands on their request
REQUESTS ANSWERS	Question related to output
REQUESTS MORE OUTPUTS	Asks for additional output
CHANGE STYLE	Changes style of output
ADDS CONTENT	Adds content to output
REMOVES CONTENT	Remove content from output
COURTESY RESPONSE	A courtesy or pleasantry
RESPONDS POSITIVELY	Explicitly pleased with output
RESPONDS NEGATIVELY	Explicitly unhappy with output
UNDEFINED RESPONSE	No defined label applies

Table 2: User utterances are classified into ORIGINAL REQUESTS and high-level follow-up types.

Writing Intent Types	
IMPROVE TEXT	GENERATE PROFESSIONAL DOC
GENERATE MESSAGE	GENERATE CATCHY TEXT
GENERATE BIO	GENERATE STORY
GENERATE SUMMARY	GENERATE TECHNICAL TEXT
GENERATE SCRIPT	GENERATE CHARACTER
GENERATE ESSAY	GENERATE POEM
GET REFERENCES	GENERATE SONG
GENERATE ONLINE POST	GENERATE JOKE
QUESTION ABOUT WRITING	UNDEFINED REQUEST

Table 3: Users ORIGINAL REQUESTS are classified into the above writing intents.

solution ensures the consistency of results across re-runs (Greene et al., 2014).

Correlating Intents and PATHs. To answer RQ2, we classify Original Requests into a finer-grained set of writing intents (Table 3) with a GPT-4o based multi-label intent classifier (f_{WritingI}). f_{WritingI} was also developed iteratively and manually validated (see Appendix C.2). Then we run a logistic regression correlating intents (predictors) from f_{WritingI} with PATHs (targets). Analyzing the learned coefficients of the regression models allows us to identify statistically significant correlations between writing intents and PATHs in a principled manner. Logistic regressions also enable easy interpretation and follow on a large body of prior work (Gujarati, 2021). Finally, to gain a deeper understanding of user behaviors in a correlated intent-PATH pair, two authors conducted a qualitative analysis of the pairs, which showed statistically significant correlations and were repeated in BCP$_{\text{Wr}}$ and WC$_{\text{Wr}}$. Author-driven manual analysis was conducted to overcome the lack of access to Bing Copilot or WildChat users, a fundamental challenge in all log-based studies (Dumais et al., 2014). In §6 we include example conversations examined by both authors to illustrate our findings.
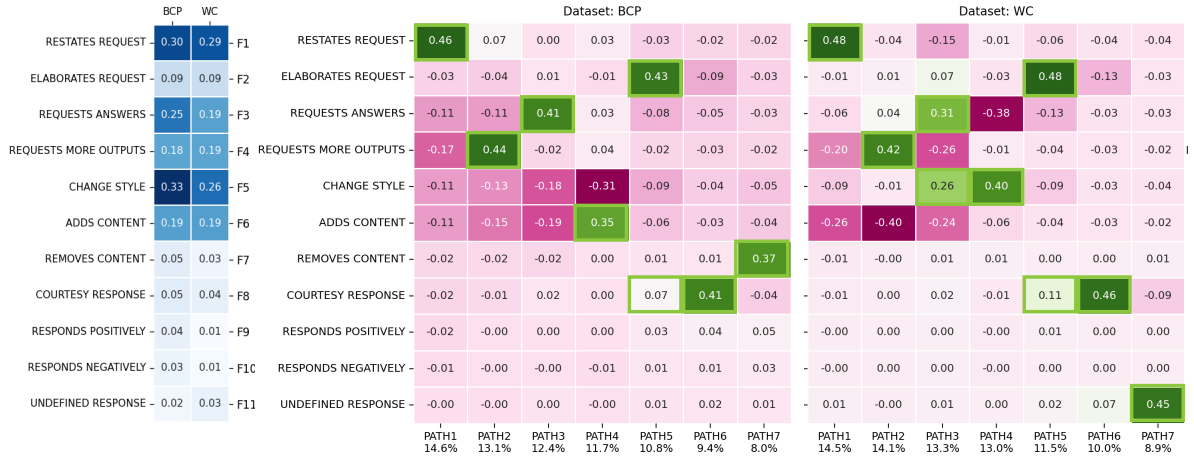
5 Results – Exploring PATHs

We start with (RQ1): What high-level collaboration behaviors emerge from user interactions in LLM-assisted writing? We do this by visualizing the frequency of follow-up types identified by f_{FollowU} (Figure 3a) and the correlations between the follow-up types and PATHs identified by PCA (Figure 3b). We discuss specific PATHs alongside writing intents and examples in §6.

Follow-up Trends. In Figure 3a, we see that the most frequent follow-up types across BCP$_{\text{Wr}}$

and WC$_{\text{Wr}}$ are similar (Table 9 contains examples). Across datasets, users frequently (18-30% sessions) follow up by revising or elaborating on their Original Request (F1, F2 in Fig. 3a), ask questions about the generation (F3), explore additional outputs (F4), or modify the generations (F5, F6). Follow-ups with explicit positive/negative feedback or courtesy responses indicating satisfaction (F8-F10) are rare and occur only in 1-5% of the sessions.

High-level trends in PATHs. In Figure 3b we visualize how follow-up types form PATHs (**W** from PCA) and the variance explained by each PATH/PC. Seven PATHs explain 80-85% of the variance, showing that a small set of collaboration behaviors explains the bulk of variance in follow-up behaviors in BCP$_{\text{Wr}}$ and WC$_{\text{Wr}}$. Further, each PATH accounts for a similar and small percentage of variance (8-14%), with most PATHs corresponding to a single follow-up type. This suggests that each follow-up type captures a distinct form of collaboration, with multiple co-occurring follow-up types being less frequent. When follow-up types do co-occur, this is more common in rare follow-up types (e.g., ELABORATES REQUEST and COURTESY RESPONSE co-occur in PATH5).

PATHs in BCP$_{\text{Wr}}$ vs WC$_{\text{Wr}}$. In Figure 3b, we also see that the discovered PATHs share significant similarities despite differences in Bing Copilot and WildChat deployments (e.g. base LLMs, system prompts, interfaces, and user bases). Specifically, we find RESTATES REQUEST (F1) accounts for similar amounts (14.5-14.6%) and the maximum variance in both datasets. Revising requests is similar to query reformulations in search, such as Google or Bing Search. This is a dominant mode of interaction that is familiar to users (Alaofi et al., 2022, Sec 7.4), and they continue to engage in it. Further, exploring additional outputs (PATH2) explains the

	BCP	WC	
RESTATES REQUEST	0.30	0.29	F1
ELABORATES REQUEST	0.09	0.09	F2
REQUESTS ANSWERS	0.25	0.19	F3
REQUESTS MORE OUTPUTS	0.18	0.19	F4
CHANGE STYLE	0.33	0.26	F5
ADDS CONTENT	0.19	0.19	F6
REMOVES CONTENT	0.05	0.03	F7
COURTESY RESPONSE	0.05	0.04	F8
RESPONDS POSITIVELY	0.04	0.01	F9
RESPONDS NEGATIVELY	0.03	0.01	F10
UNDEFINED RESPONSE	0.02	0.03	F11

(a) Frequencies.

Dataset: BCP

	PATH1 14.6%	PATH2 13.1%	PATH3 12.4%	PATH4 11.7%	PATH5 10.8%	PATH6 9.4%	PATH7 8.0%
RESTATES REQUEST	0.46	0.07	0.00	0.03	-0.03	-0.02	-0.02
ELABORATES REQUEST	-0.03	-0.04	0.01	-0.01	0.43	-0.09	-0.03
REQUESTS ANSWERS	-0.11	-0.11	0.41	0.03	-0.08	-0.05	-0.03
REQUESTS MORE OUTPUTS	-0.17	0.44	-0.02	0.04	-0.02	-0.03	-0.02
CHANGE STYLE	-0.11	-0.13	-0.18	-0.31	-0.09	-0.04	-0.05
ADDS CONTENT	-0.11	-0.15	-0.19	0.35	-0.06	-0.03	-0.04
REMOVES CONTENT	-0.02	-0.02	-0.02	0.00	0.01	0.01	0.37
COURTESY RESPONSE	-0.02	-0.01	0.02	0.00	0.07	0.41	-0.04
RESPONDS POSITIVELY	-0.02	-0.00	0.00	0.00	0.03	0.04	0.05
RESPONDS NEGATIVELY	-0.01	-0.00	-0.00	-0.01	0.01	0.01	0.03
UNDEFINED RESPONSE	-0.00	-0.00	0.00	-0.00	0.01	0.02	0.01

Dataset: WC

	PATH1 14.5%	PATH2 14.1%	PATH3 13.3%	PATH4 13.0%	PATH5 11.5%	PATH6 10.0%	PATH7 8.9%
RESTATES REQUEST	0.48	-0.04	-0.15	-0.01	-0.06	-0.04	-0.04
ELABORATES REQUEST	-0.01	0.01	0.07	-0.03	0.48	-0.13	-0.03
REQUESTS ANSWERS	-0.06	0.04	0.31	-0.38	-0.13	-0.03	-0.03
REQUESTS MORE OUTPUTS	-0.20	0.42	-0.26	-0.01	-0.04	-0.03	-0.02
CHANGE STYLE	-0.09	-0.01	0.26	0.40	-0.09	-0.03	-0.04
ADDS CONTENT	-0.26	-0.40	-0.24	-0.06	-0.04	-0.03	-0.02
REMOVES CONTENT	-0.01	-0.00	0.01	0.01	0.00	0.00	0.01
COURTESY RESPONSE	-0.01	0.00	0.02	-0.01	0.11	0.46	-0.09
RESPONDS POSITIVELY	-0.00	0.00	0.00	-0.00	0.01	0.00	0.00
RESPONDS NEGATIVELY	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00
UNDEFINED RESPONSE	0.01	-0.00	0.01	0.00	0.02	0.07	0.45

(b) Correlations between follow-up types and PCs in BCP$_{Wr}$ and WC$_{Wr}$.

Figure 3: (a) The fraction of sessions which contain a follow-up type. (b) The correlation between follow-up types and principal components (PC) inferred by PCA. Each PC represents a PATH. Large positive values (green boxes) indicate stronger correlations. The percentages (bottom) depict the variance explained by each PATH.

2nd largest amount of variance in both datasets. Similarly, PATH5 and PATH6 are also shared across both datasets, though these constitute less frequent follow-up types. These trends suggest that users collaborate with LLMs in very similar ways across both systems we examined. However, there are also some notable differences. PATH3 while correlating with REQUESTS ANSWERS in both datasets, also correlated with CHANGE STYLE in WC$_{Wr}$. Similarly, PATH4 correlated with ADDS CONTENT and CHANGE STYLE respectively. Despite the differences in PATH3 and 4, note that they aim to modify LLM generations in different ways. We hypothesize that this difference is due to varying writing intent mixes across BCP$_{Wr}$ and WC$_{Wr}$ (Figure 8). As we see in §6, writing intents correlate with different PATHs and may result in different behaviors at the dataset level when intent proportions vary. Our finding that explicit feedback is rare and the consistent presence of request revisions suggest the following implications for future work:

Implications: • Leverage implicit feedback in users' follow-ups for LLM alignment. • Understand why users revise requests and leverage it for LLM alignment.

6 Results – Correlating Intents and PATHs

Here we answer (RQ2): How do users' collaboration behaviors differ across writing intents? We do this by correlating writing intents and PATHs in regressions. When PATHs aren't shared across BCP$_{Wr}$ and WC$_{Wr}$, we use follow-up types as tar-

Target in BCP$_{Wr}$, WC$_{Wr}$	Description	Section
PATH2, PATH2	Requesting more output.	§6.1
PATH3, REQUESTS ANSWERS	Requesting answers.	§6.2
PATH4, ADDS CONTENT	Adding content.	§6.3
CHANGE STYLE, PATH4	Changing style.	§D.1
PATH1, PATH1	Revising requests.	§D.2
PATH6, PATH6	Elaborating on requests.	§D.2

Table 4: Overview of PATHs in our regression analysis.

gets. Table 4 summarizes the behaviors examined in our analysis. To ensure generalization of our findings, we highlight statistically significant correlations shared across BCP$_{Wr}$ and WC$_{Wr}$, and examine frequent writing intents (Figure 8). We present the results of our mixed-methods analysis with regression coefficient plots, example conversations (truncated and rephrased) examined in qualitative analysis, and discuss the implications of our findings for future work.

6.1 Requesting more outputs to brainstorm or stage long generations

Figure 4a depicts writing intents correlated with PATH2 where users REQUEST MORE OUTPUTS. We start by noting that GEN CATCHY TEXT (I10) shows a strong positive correlation with PATH2. Analysis of sessions revealed that when users aimed to generate eye-catching texts such as product names, document titles, email subjects, etc, users requested more outputs aiming to brainstorm more creative, catchy texts. Figure 4b depicts examples (Ex 1 and 2) from BCP$_{Wr}$ and WC$_{Wr}$. This behavior finds precedent in prior work on creativity

(a) Logistic regression coefficients for intents vs PATH2 (requesting more outputs) in WC_{Wr}.

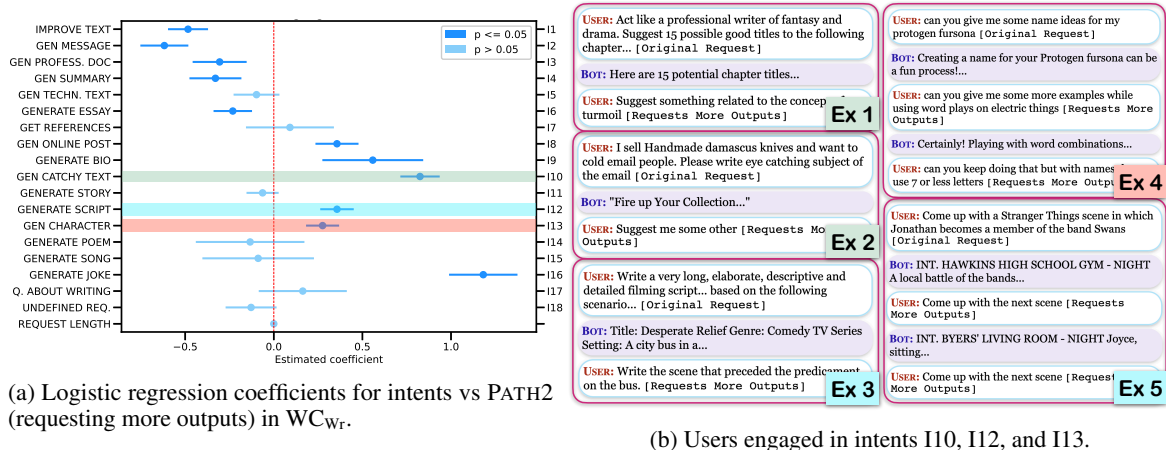(b) Users engaged in intents I10, I12, and I13.

Figure 4: (a) Large positive values in coefficient plots indicate strong correlations. The intents discussed in §6.1 are highlighted in color. Coefficients for BCP_{Wr} are plotted in Figure 12. (b) Example conversations from the intents highlighted in (a) – intent and example colors are matched.

support, who note the value of diverse ideas during brainstorming (Frich et al., 2019). PATH2 may also be seen as a form of pluralistic alignment, i.e., eliciting overtone alignment (Sorensen et al., 2024) – exploring overtone alignment for brainstorming represents meaningful future work.

Next, the intents GENERATE SCRIPT (I12) and GEN CHARACTER (I13) show a weaker positive correlation with PATH2. Here, users attempted to generate media scripts (e.g., YouTube videos) or fictional characters (Ex 4, 3, and 5 in Figure 4b). Analysis of sessions revealed that while some GENERATE CHARACTER sessions engaged in brainstorming (Ex 4), users primarily engaged in staged generation of long texts (Ex 3 and 5). Here, users' follow-ups also varied in specificity from simply asking for more output (Ex 5) to being more specific (Ex 3). To our knowledge, this represents the first evidence of multi-turn construction of creative narratives in the wild. While prior work has explored interactive generation of long narratives through plans (Yao et al., 2019) interactions (Brahman et al., 2020), and complex instructions (Pham et al., 2024), multi-turn construction of creative narratives remains under-explored.

Implications: • Develop and evaluate overtone-aligned LLMs for brainstorming. • Develop resources and models to generate creative narratives with under-specific multi-turn interaction.

6.2 Asking follow-up questions to learn or stage long generations

Figure 5a shows writing intents correlated with PATH3, where users ask questions in response to model generations. Users tended to ask questions when they generated professional documents (I3), summaries (I4), technical texts (I5), and fictional character narratives (I13). Our analysis revealed that users' questions differed across these intents, from those asking about domain-specific knowledge and norms (Ex 1, 2, 4, 5) to those focused on the LLMs prior generation (Ex 3, 6). We demonstrate this empirically in Figure 13.

When users generated technical texts or summaries, they asked knowledge-rich questions – needing the LLM to be grounded in domain-specific knowledge (Ex 2) or the task context (Ex 5). While prior work notes that users engage in information-seeking during knowledge-rich writing (Shen et al., 2023) – we provide evidence for this behavior in the wild. Information seeking during knowledge-rich writing and the use of these questions for LLM alignment remain under-studied and may be explored in future work.

We also uncover novel and emerging behaviors in users asking questions to learn about professional norms or in generating fictional narratives. When users generated professional documents such as cover letters, their questions sought to learn about professional norms and expectations (Ex 1, 4) – needing the LLM to be grounded in these norms. While some prior work has explored writers' feedback-seeking behavior (Gero et al., 2023) and LLMs' potential to provide writing feedback (Li et al., 2024; Liang et al., 2024b), aligning LLMs to provide feedback in writing sessions represents an emerging problem.

Finally, when users generated characters for their
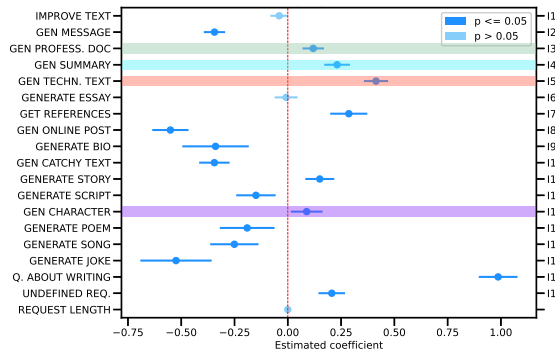
(a) Logistic regression coefficients for intents vs PATH3 (requesting answers) in BCP_{Wr}.
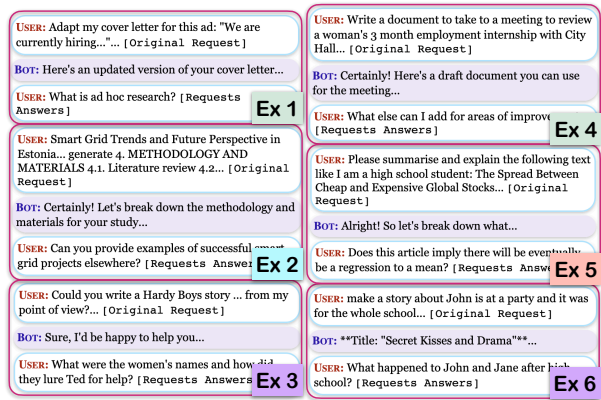
(b) Users engaged in intents I3, I4, I5, and I13.

Figure 5: (a) The intents discussed in §6.2 are highlighted in color. Coefficients for WC_{Wr} are plotted in Figure 12. (b) Example conversations from the intents highlighted in (a) – intent and example colors are matched.

stories (I13), their questions followed up on the generated stories. Analysis of sessions revealed that users sought to direct the generation of long-form fictional narratives (Ex 3, 6), mirroring §6.1. While some prior work explores question answering for fictional narratives (Xu et al., 2022), a detailed understanding of users' motivations for this behavior and the use of questions to interactively build narratives remains under-explored.

Implications: • Investigate question asking behaviors in knowledge-rich and creative writing and explore its use for session-level alignment. • Evaluate and develop methods to use LLMs for providing writing feedback in high-stakes document writing.

6.3 Adding to generations when they lacked content known only to the user

Figure 6a shows writing intents correlated with PATH4 where users add content to model generations. This behavior correlates with generating professional documents (I3), messages (I2), and creative narratives (I11-I13). Our analysis of sessions found that users overwhelmingly added content when model generations were missing information likely to be known only to the users. When users generated professional documents such as resumes, cover letters, clinical notes, etc. they added information about their personal encounters or skills (Ex 1, 4). Similarly, when they generated communicative texts such as emails, letters, or speeches, they added personal stories (Ex 2, 5). This suggests that these intents may benefit from personalization from users' historical chats or through proactive question asking to obtain missing information. While a large body of recent work has

explored personalization of LLMs (Mysore et al., 2024; Magister et al., 2024) or proactive interaction (Deng et al., 2025), we highlight intents where personalization or proactive interaction may be most meaningful in the wild. Finally, in generating fictional texts such as stories, scripts, or character descriptions (I11-I13), users added plots which tailored generations to their fictional visions (Ex 3, 6). This behavior overlaps with PATHs where users requested more outputs (§6.1), and asked questions to stage long generations (§6.2), and remains understudied.

Implications: • Develop methods to infer when long communicative texts are missing users' personal knowledge, if appropriate, obtain this information proactively, or personalize generations based on users' historical documents. • Develop resources and methods to incorporate long-form feedback into creative narratives generated in multi-turn interactions.

In Appendix D.1 we show how users change generation style to better match readers' expectations, and find dataset-dependent trends in revising (PATH1) and elaborating (PATH5) on requests (Appendix D.2).

7 Discussion

Taken together, our results reveal that users interact with LLMs in highly collaborative and goal-driven ways — they co-construct creative texts, probe LLMs for domain knowledge as they write, and shape text to meet long-term communicative goals. These PATHs highlight behaviors not captured by simpler measures of satisfaction and classification into tasks, and highlight the importance of model-
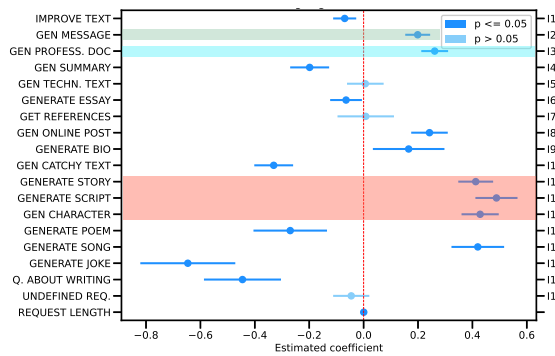
(a) Logistic regression coefficients for intents vs PATH4 (adding content) in BCP_Wr.
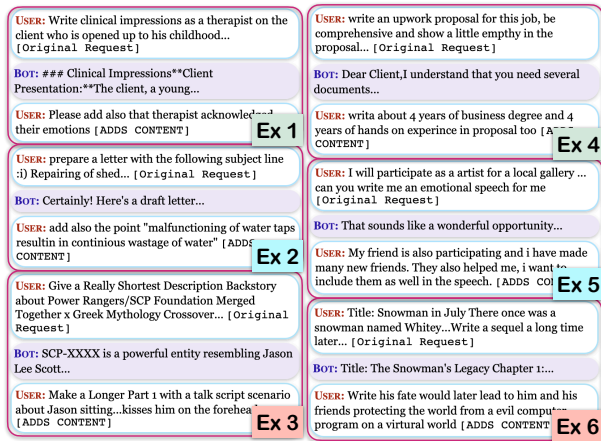
(b) Users engaged in intents I2, I3, I11-I13.

Figure 6: (a) The intents discussed in §6.3 are highlighted in color. Coefficients for WC_Wr are plotted in Figure 12. (b) Example conversations from the intents highlighted in (a) – intent and example colors are matched.

ing user behaviors to align LLMs with long-term user needs. Next, we synthesize the implications of our results and highlight meaningful areas of future work suggested by our results.

Toward Session-Level Alignment from Natural Feedback. We showed how user follow-ups rarely included explicit positive or negative feedback in §5 and engage in open-ended collaboration behaviors in §6. However, a majority of present work on LLM alignment remains restricted to single-turn alignment from explicit preference feedback (Chaudhari et al., 2024). Our results suggest that aligning LLMs from natural feedback over multi-turn interactions presents an increasingly meaningful area of future work. Early work has begun to explore alignment with single-turn natural language feedback (Don-Yehiya et al., 2024; Shi et al., 2024) or simulated multi-turn feedback (Gao et al., 2024; Wu et al., 2025). Future work may explore building and learning from simulators that follow realistic user behaviors captured in existing log datasets or the design of reward models trained on implicit user feedback.

Further, while some PATHs are task-oriented and aim to directly tune LLM generations (e.g. changing style or adding content), others guide an exploratory session (e.g. requesting more outputs or answers). While task-oriented sessions have received significant attention, alignment and evaluation for exploratory sessions remain largely understudied in the LLM-era, representing exciting future work.

Aligning to Under-Specified Feedback and Goals. User follow-ups vary in their level of speci-

ficity, this is clearest in the staged generation of long texts. Here, users follow-ups increased in specificity from REQUESTS MORE OUTPUT (§6.1), to REQUESTS ANSWERS (§6.2), and to ADDS CONTENT (§6.3). We show this empirically in Figure 14. This highlights a line of future work focused on LLM-alignment from under-specified natural language feedback.

Most current work assumes user interactions (ORs and PATHs) to be fully specified. While a sizable body of work has focused on uncertainty estimation for LLM outputs (Geng et al., 2024), limited work has explored uncertainty in user interactions. This early work has sought to resolve under-specificity through task context (Malaviya et al., 2024; Sarkar et al., 2025), proactive question asking (Kobalczyk et al., 2025; Li et al., 2025), and coverage-based methods (Hou et al., 2024). Exploring these and other strategies is a valuable direction for future work.

LLMs as Learning Partners: Aligning for User Growth. Across a writing intent and PATHs, we find users to leverage LLMs not merely as task assistants, but as partners in their learning and development. Users brainstormed with models to generate more compelling or creative text (§6.1), asked questions to understand professional norms while drafting documents (§6.2), requested feedback to improve fictional narratives (§6.2), and iterated on text to better suit their audiences (§D.1). These behaviors reflect a desire not only to produce better outputs, but to use LLMs for longer-term learning.

Our findings suggest that aligning LLMs to support users' growth and learning represents a com-

pelling and underexplored design goal. Recent work has echoed this vision (Collins et al., 2024b), and prototype systems are beginning to emerge (Li et al., 2024; Chamoun et al., 2024). However, realizing this potential will require addressing critical challenges: mitigating the risk of cultural bias in "coaching," preventing homogenization of users' voice and ideas (Agarwal et al., 2025; Drosos et al., 2025), and designing interactions that preserve the "productive struggle" known to facilitate deeper learning (Rus and Kendeou, 2025).

Ultimately, building LLM systems for learning will demand rich interdisciplinary work, drawing on learning sciences, HCI, and NLP to understand how LLM systems can guide and challenge users, rather than merely serve their short-term needs.

8 Conclusion

In this paper[2], we analyzed a large sample of Bing Copilot and WildChat user-LLM interaction logs spanning seven and thirteen months of global usage. We focused on user-LLM collaboration behaviors in English writing sessions. Our analysis took a mixed-methods approach, using PCA to identify prototypical human-AI collaboration behaviors (PATHs) and regression analysis to identify statistically significant correlations between writing intents and PATHs. We paired this with lightweight qualitative analysis to gain further insight into automatically highlighted behaviors. We ensure the reliability of our analysis through manual validation of all automated components. We contribute the first understanding of in-the-wild user-LLM collaboration behaviors and discuss their implications for LLM alignment.

9 Limitations and Future Work

Now, we highlight some limitations of our analysis, meaningful future work, and comment on the broader impacts of our work.

Data and Analysis. Our analysis was limited to English sessions. Therefore, it is likely that some collaboration behaviors specific to non-English users were missed in our analysis. We made this choice to ensure the reliability of LLM classifiers and to ensure that log text was intelligible to all authors during annotation and analysis. Extending our analysis to non-English logs represents a meaningful area of future work. Next, our analysis (§4)

does not model the order of follow-ups in a session and uses linear PCA to identify PATHs. We rely on simpler modeling to ensure the interpretability of results in a first analysis, future work may explore richer modeling of follow-ups to discover more complex PATHs. In similar vein, we don't analyze when in a conversation specific follow-ups occur since conversations had 2 follow-ups on average (§3). Future work focusing on long-running conversations might more effectively analyze follow-up positions. Finally, we rely on a light-weight qualitative analysis to understand users' goals in correlated intent-PATH pairs. Future work may explore these behaviors in controlled human-centered studies with direct access to the actual users.

Utility of PATHs. While our analysis explored the use of PATHs to study user-LLM collaboration it does not explore their use for model training or evaluation. However, their direct use in model development may be explored. E.g. PATHs segment a dataset of sessions into clusters with a consistent user behavior. Such clusters may serve as a targeted evaluation set for LLMs. Similarly the clusters may serve as a post-training dataset to encourage specific behaviors in models. We leave such explorations to future work.

Broader Impacts. We highlight LLM alignment from user-LLM interactions to be a meaningful area of future work in §6. However, we note that learning from implicit interactions may pose safety risks if user interactions are adversarial (e.g. Zhao et al. (2024b)) or pose privacy risks (e.g. Xin et al. (2024)) when implicit interactions encode private user information. This may result in user interactions being leaked from aligned LLMs. Therefore, ensuring safety and privacy must be an important aspect of learning from implicit interactions.

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. *Preprint*, arXiv:2409.11360.

Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where do queries come from? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2850–2862, New York, NY, USA. Association for Computing Machinery.

Aparna Ananthasubramaniam, Hong Chen, Jason Yan,

---

[2] LLM assistants were used to improve the phrasing of some sentences in this paper and fix grammatical errors.

Kenan Alkiek, Jiaxin Pei, Agrima Seth, Lavinia Dunagan, Minje Choi, Benjamin Litterer, and David Jurgens. 2023. Exploring linguistic style matching in online communities: The role of social context and conversation dynamics. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 64–74, Toronto, Canada. Association for Computational Linguistics.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi. 2023. Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 436–452, New York, NY, USA. Association for Computing Machinery.

Ananya Bhattacharjee, Jina Suh, Mahsa Ershadi, Shamsi T. Iqbal, Andrew D. Wilson, and Javier Hernandez. 2024. Understanding communication preferences of information workers in engagement with text-based conversational agents. *Preprint*, arXiv:2410.20468.

Param Biyani, Yasharth Bajpai, Arjun Radhakrishna, Gustavo Soares, and Sumit Gulwani. 2024. Rubicon: Rubric-based evaluation of domain-specific human ai conversations. In *Proceedings of the 1st ACM International Conference on AI-Powered Software*, AIware 2024, page 161–169, New York, NY, USA. Association for Computing Machinery.

Faeze Brahman, Alexandru Petrusca, and Snigdha Chaturvedi. 2020. Cue me in: Content-inducing approaches to interactive story generation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 588–597, Suzhou, China. Association for Computational Linguistics.

Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno, Franziska Roesner, and Niloofar Mireshghallah. 2024. Developing story: Case studies of generative ai's use in journalism. *Preprint*, arXiv:2406.13706.

Daniel Buschek. 2024. Collage is the new writing: Exploring the fragmentation of text and user interfaces in ai tools. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 2719–2737, New York, NY, USA. Association for Computing Machinery.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving emerging writers. *Preprint*, arXiv:2309.12570.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *arXiv preprint arXiv:2404.08555*.

Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the Web Conference 2021*, WWW '21, page 743–755, New York, NY, USA. Association for Computing Machinery.

Katherine M. Collins, Albert Q. Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B. Tenenbaum, William Hart, Timothy Gowers, Wenda Li, Adrian Weller, and Mateja Jamnik. 2024a. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121.

Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, and 1 others. 2024b. Building machines that learn and think with people. *Nature human behaviour*, 8(10):1851–1863.

Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Trans. Inf. Syst.*, 43(3).

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 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2024. Naturally Occurring Feedback is Common, Extractable and Useful. *Preprint*, arXiv:2407.10944.

Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The ai ghostwriter effect: When users do not perceive ownership of ai-generated text but self-declare as authors. *ACM Trans. Comput.-Hum. Interact.*, 31(2).

16828

Ian Drosos, Advait Sarkar, Neil Toronto, and 1 others. 2025. " it makes you think": Provocations help restore critical thinking to ai-assisted knowledge work. *arXiv preprint arXiv:2501.17247*.

Susan Dumais, Robin Jeffries, Daniel M. Russell, Diane Tang, and Jaime Teevan. 2014. *Understanding User Behavior Through Log Data and Analysis*, pages 349–372. Springer New York, New York, NY.

Nathan Eagle and Alex Sandy Pentland. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral ecology and sociobiology*, 63:1057–1066.

Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the landscape of creativity support tools in hci. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–18, New York, NY, USA. Association for Computing Machinery.

Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. 2024. From text to self: Users' perception of aimc tools on interpersonal communication and self. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning LLM agents by learning latent preference from user edits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social dynamics of ai support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 498–513. Springer.

Damodar N Gujarati. 2021. *Essentials of econometrics*. Sage Publications.

Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, and 1 others. 2025. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19023–19042. PMLR.

Xi Yu Huang, Krishnapriya Vishnubhotla, and Frank Rudzicz. 2024. The gpt-writingprompts dataset: A comparative analysis of character portrayal in short stories. *Preprint*, arXiv:2406.16767.

Angel Hsing-Chi Hwang, Q. Vera Liao, Su Lin Blodgett, Alexandra Olteanu, and Adam Trischler. 2024. "It was 80% me, 20% AI": Seeking Authenticity in Co-Writing with Large Language Models. *Preprint*, arXiv:2411.13032.

Bernard J. Jansen and Amanda Spink. 2006. How are we searching the world wide web? a comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248–263. Formal Methods for Information Retrieval.

Maryam Kamvar and Shumeet Baluja. 2006. A large scale study of wireless search behavior: Google mobile search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 701–709, New York, NY, USA. Association for Computing Machinery.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Preprint*, arXiv:2404.16019.

Katarzyna Kobalczyk, Nicolas Astorga, Tennison Liu, and Mihaela van der Schaar. 2025. Active task disambiguation with llms. *arXiv preprint arXiv:2502.04485*.

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, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, and 17 others. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Ge Li, Danai Vachtsevanou, Jérémy Lemée, Simon Mayer, and Jannis Strecker. 2024. Reader-aware writing assistance through reader profiles. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, HT '24, page 344–350, New York, NY, USA. Association for Computing Machinery.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. 2025. Aligning llms to ask good questions a case study in clinical reasoning. *arXiv preprint arXiv:2502.14860*.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *ICML*.

Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025. The widespread adoption of large language model-assisted writing across society. *Preprint*, arXiv:2502.09747.

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

Ying-Chun Lin, Jennifer Neville, Jack Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. 2024. Interpretable user satisfaction estimation for conversational systems with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11100–11115, Bangkok, Thailand. Association for Computational Linguistics.

Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje ter Hoeve. 2024. On the way to llm personalization: Learning to remember user conversations. *arXiv preprint arXiv:2411.13405*.

Moushumi Mahato, Avinash Kumar, Kartikey Singh, Javaid Nabi, Debojyoti Saha, and Krishna Singh. 2024. Exploring user dissatisfaction: Taxonomy of implicit negative feedback in virtual assistants. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 230–242, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, and Kyle Lo. 2024. Contextualized evaluations: Taking the guesswork out of language model evaluations. *arXiv preprint arXiv:2411.07237*.

Yusuf Mehdi. 2023. Reinventing search with a new ai-powered microsoft bing and edge, your copilot for the web. Accessed: 3 March 2025.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 198–219, Miami, Florida, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Maria-Teresa De Rosa Palmini, Laura Wagner, and Eva Cetinic. 2024. Civiverse: A dataset for analyzing user engagement with open-source text-to-image models. *Preprint*, arXiv:2408.15261.

Shramay Palta, Nirupama Chandrasekaran, Rachel Rudinger, and Scott Counts. 2025. Speaking the right language: The impact of expertise alignment in user-ai interactions. *Preprint*, arXiv:2502.18685.

Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1722–1753, Miami, Florida, USA. Association for Computational Linguistics.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 989–992, New York, NY, USA. Association for Computing Machinery.

Jonathan Reades, Francesco Calabrese, and Carlo Ratti. 2009. Eigenplaces: analysing cities using the space–time structure of the mobile phone network. *Environment and planning B: Planning and design*, 36(5):824–836.

Vasile Rus and Panayiota Kendeou. 2025. Are llms actually good for learning? *AI & SOCIETY*, pages 1–2.

Rupak Sarkar, Bahareh Sarrafzadeh, Nirupama Chandrasekaran, Nagu Rangan, Philip Resnik, Longqi Yang, and Sujay Kumar Jauhar. 2025. Conversational user-ai intervention: A study on prompt rewriting for improved llm response generation. *Preprint*, arXiv:2503.16789.

Bahareh Sarrafzadeh, Sujay Kumar Jauhar, Michael Gamon, Edward Lank, and Ryen W. White. 2021. Characterizing stage-aware writing assistance for collaborative document authoring. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3).

Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond summarization: Designing ai support for real-world expository writing tasks. *arXiv preprint arXiv:2304.02623*.

Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Jauhar, Sihao Chen, Shan Xia, and 1 others. 2024. Wildfeedback: Aligning llms with in-situ user interactions and feedback. *arXiv preprint arXiv:2408.15549*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *ICML*.

Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. 2024. The use of generative search engines for knowledge work and complex tasks. *Preprint*, arXiv:2404.04268.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, and 2 others. 2024. Clio: Privacy-preserving insights into real-world ai use. *Preprint*, arXiv:2412.13678.

Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, Texas. Association for Computational Linguistics.

Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do users really ask large language models? an initial log analysis of google bard interactions in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2703–2707, New York, NY, USA. Association for Computing Machinery.

Kailas Vodrahalli and James Zou. 2024. ArtWhisperer: A dataset for characterizing human-AI interactions in artistic creations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 49627–49654. PMLR.

Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "it felt like having a second mind": Investigating human-ai co-creativity in prewriting with large language models. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Sterling Williams-Ceci, Maurice Jakesch, Advait Bhat, Kowe Kadoma, Lior Zalmanson, and Mor Naaman. 2024. Bias in ai autocomplete suggestions leads to attitude shift on societal issues. *PsyArXiv*.

Shujin Wu, Yi R. Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025. Aligning LLMs with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662, Abu Dhabi, UAE. Association for Computational Linguistics.

Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2024. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. In *Neurips Safe Generative AI Workshop 2024*.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7378–7385.

16831

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024a. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Wenting Zhao, Alexander M Rush, and Tanya Goyal. 2024b. Challenges in trustworthy human evaluation of chatbots. *arXiv preprint arXiv:2412.04363*.

Zixin Zhao, Damien Masson, Young-Ho Kim, Gerald Penn, and Fanny Chevalier. 2025. Making the write connections: Linking writing support tools with writer's needs. *arXiv preprint arXiv:2502.13320*.

Shengqi Zhu, Jeffrey M. Rzeszotarski, and David Mimno. 2025. Data paradigms in the era of llms: On the opportunities and challenges of qualitative data in the wild. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

A Extended Related Work

Having discussed prior work on user-LLM log analysis in §2 we discuss additional related work here.

Measuring user satisfaction. Our analysis of users' follow-up behaviors also ties it to prior work on predicting binary signals of dis/satisfaction from conversations (Lin et al., 2024; Biyani et al., 2024; Mahato et al., 2024). In our work, we go beyond binary notions of dis/satisfaction and provide a richer characterization of users' collaboration behaviors. Further, we find explicit satisfaction signals to be rare and highlight a need to understand and use implicit behaviors for LLM alignment (see §5). Concurrent analysis on Bing Copilot and WildChat-1M logs is also relevant to our work. Here, Palta et al. (2025) correlates the difference between expertise in human and LLM utterances with binary metrics of user satisfaction. And Brigham et al. (2024) conduct a small-scale analysis of journalistic writing sessions in WildChat-1M and discuss its implications for responsible AI use in journalism. While these works differ in focus, together with our work, they paint a more complete picture of human-LLM interactions in the wild.

Human-centered Studies on Writing Assistance A sizable body of work in HCI has focused on writing assistance. Recent surveys review this body of work more completely (Lee et al., 2024; Zhao et al., 2025). Prior work examining interactions with LLM writing assistants is most relevant to our work. In contrast to user-LLM log analysis, this research has had more direct interaction with users – enabling studies to probe user perceptions, behaviors, and impacts more deeply while forgoing an understanding of in-the-wild user behaviors outside study settings. Broadly, this work has focused on the impacts of LLM-assisted writing (Hancock et al., 2020), studied user-LLM co-creation behaviors, and designed novel interaction paradigms (Buschek, 2024). Work on impacts has examined aspects such as writers perceived ownership over texts co-created with personalized LLMs (Draxler et al., 2024; Hwang et al., 2024), convergence toward normative stances, stereotypes, and styles in LLM assisted writing (Williams-Ceci et al., 2024; Huang et al., 2024; Agarwal et al., 2025), and the perception of fit between LLM outputs and tasks contexts (Fu et al., 2024; Bhattacharjee et al., 2024), among others. Studies focusing on the dynamics of human-AI co-creation examine the differences in how writers seek feedback from humans vs LLMs (Gero et al., 2023), or how they seek assistance at different stages of writing (Wan et al., 2024; Chakrabarty et al., 2024). Still others have examined the impact of specific interaction behaviors on productivity and decision making (Bhat et al., 2023; Dhillon et al., 2024). Our analysis complements this line of work by supporting laboratory findings through large-scale analysis where applicable, and uncovers novel user behaviors likely to be missed because of smaller samples or constrained study designs.

B Analysis Setup - Details

In §3 we presented an overview of our writing session datasets BCP_{Wr} and WC_{Wr} filtered from Bing Copilot and WildChat-1M logs. Both datasets are obtained after two broad stages of filtering, we describe this filtering here:

- Appendix B.1: Preliminary filtering, i.e. Bing Copilot → BCP_{All} and WildChat-1M → WC_{All}.
- Appendix B.2: Writing session filtering i.e. BCP_{All} → BCP_{Wr} and WC_{All} → WC_{Wr}.
- Appendix B.3: Validating Task Classifiers used to identify writing sessions.

B.1 Dataset Description and Preliminary Filtering

Bing Copilot → BCP_{All}. We start with a daily random sample of Bing Copilot logs from a seven-month period spanning 1 April to 31 October 2024, amounting to 20.5M conversational sessions. GPT-4 powered Bing Copilot during this period. We perform an initial filtering based on logged metadata. Since we aim to study users' follow-up behaviors, we retain sessions with 2 or more user utterances, sessions that are likely to be in English (based on the users system language) to ensure familiarity of data to all authors, and sessions conducted on a personal computer – this results in 3.6M sessions. We choose to retain sessions on personal computers to limit the impact of behavioral differences across devices, e.g., prior work notes limited diversity of queries in search on mobile phones compared to personal computers (Kamvar and Baluja, 2006).

Next, we perform a second stage of language filtering based on the first 1500 characters (about 300 words) of a session text using a Language ID model. [3] Metadata-based filtering alone proved to be insufficient to filter non-English sessions. We re-
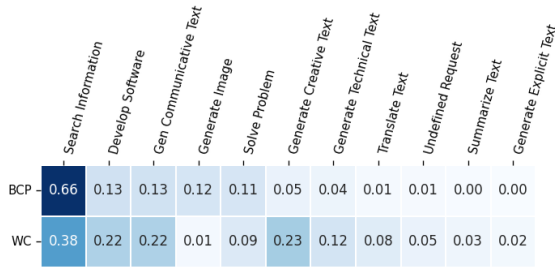
---

[3] FastText Language ID model: `lid.176.bin`

Figure 7: The proportion of coarse tasks in BCP_{All} (2.8M sessions) and WC_{All} (160k sessions). The sessions are in English and likely to have at least 1 user follow-up. A session can have multiple coarse tasks. On average, BCP_{All} and WC_{All} had 1.25 and 1.43 tasks per session.

tain 2.8M English sessions, and we refer to this as BCP_{All}. This dataset was labeled with a coarse Task Classifier ($f_{CoarseT}$) to identify writing sessions. Note that all our Bing Copilot data was scrubbed for personally identifying information, held on secure servers, labeled with privacy-compliant LLM deployments, and our analysis was conducted in accordance with organizational privacy policies and data handling practices.

WildChat-1M → WC_{All}. We start with the public[4] WildChat-1M dataset available under an AI2 ImpACT license. Zhao et al. (2024a) gathered this data over thirteen months from 9 April 2023 to 1 May 2024 in a HuggingFace Hub space. The sessions contain interactions with GPT-4 and GPT-3.5-Turbo. Of the 840k publicly available sessions, we retain sessions labeled as English in the metadata and with 2 or more user utterances – this results in 160k sessions, we refer to this as WC_{All}. Then WC_{All} sessions are labeled with $f_{CoarseT}$ to identify writing sessions.

Dataset Characteristics In Table 1 we show the number of users and countries of origin for BCP_{All} and WC_{All}. In Figure 7 we show the proportion of tasks identified by $f_{CoarseT}$ before filtering them down to writing sessions alone. We see that Bing Copilot sessions are focused on search ("Search Information"), whereas WildChat has a more even proportion of tasks. A likely explanation is that Bing Copilot was available alongside the Bing search engine, whereas WildChat was framed as a general-purpose conversational assistant in a research study. Further, $f_{CoarseT}$, identified 1.25 and 1.43 tasks per session on average in BCP_{All} and WC_{All}, respectively. The median was 1 task

---

	Utterances	Original requests	Follow ups
BCP_{Wr}	7.5 / 6	1.6 / 1	2.1 / 2
WC_{Wr}	7.1 / 6	1.6 / 1	1.9 / 1

Table 5: Session level characteristics for our datasets (mean / median). We classify session utterances from the user into Original Requests and Follow ups, where ORs may have 1 or more fine-grained intents.

– indicating that most sessions focus on a single coarse task.

B.2 Identifying Writing Sessions

Task Classifier. We follow prior work and conceptualize writing sessions to be the ones where users translate their thoughts into writing (Lee et al., 2024). We interpret this broadly and treat sessions where users write or re-write, whole or parts of texts to be a writing session. We implement this in GPT-4o based $f_{CoarseT}$ and retain sessions labeled as generating communicative, creative, technical, and summarizing text as writing tasks – Figure 7 depicts their proportions.

Prompt 1 was used for $f_{CoarseT}$. We generate our task labels and label definitions by collating tasks identified in Suri et al. (2024), and through additional manual analysis of 200 BCP_{All} conversations. We ensure the label coverage by manually applying them to 200 sessions in WC_{All}. In both data sets, $f_{CoarseT}$ inputs the first 10 utterances to predict a session's coarse task. Given the larger size of BCP_{All} and the expense (time and dollar amounts) of GPT-4o calls, we trained an embedding-based scalable classifier on labels obtained from GPT-4o for 100k BCP_{All} sessions. Our scalable classifier used a frozen `E5-large-v2` (Wang et al., 2024) model for embedding the conversation text and fine-tuned logistic regression classifiers in a one-vs-all setup to obtain a $f_{CoarseT}$ for BCP_{All}. Since WC_{All} is a smaller dataset, we used GPT-4o directly for it.

BCP_{All} → BCP_{Wr} To obtain writing sessions, we label BCP_{All} sessions with $f_{CoarseT}$ and retain sessions labeled with the labels Gen Communicative Text, Gen Creative Text, Gen Technical Text, and Summarize Text as writing sessions. Out of 2.8M sessions, we identify 500k as writing sessions. These made up 17-19% sessions in each month. Finally, we randomly subsample 250k of 500k, resulting in BCP_{Wr}. Our sub-sampling was done to lower the costs of subsequent analysis, which relies on LLM classifiers while retaining a large enough

	Coarse Tasks	
	Acc. (%)	Agr. (κ)
BCP$_{\text{All}}$	89.09	1.00
WC$_{\text{All}}$	80.00	0.87
	Writing Tasks	
BCP$_{\text{All}}$	85.71	1.00
WC$_{\text{All}}$	91.07	0.87

Table 6: Accuracy (Acc.) of GPT-4o based Task Classifier (f_{CoarseT}) in BCP$_{\text{All}}$ and WC$_{\text{All}}$. Prompt 1 contains label definitions for the eleven coarse task classes. Four of the eleven classes are treated as writing sessions – accuracy and agreement for these classes alone are also included above. Agreements (Agr.) are computed from judgments made by two independent annotators.

sample for analysis.

WC$_{\text{All}}$ → WC$_{\text{Wr}}$ Out of 160k sessions, 68k (42%) were identified as writing sessions by f_{CoarseT} to form WC$_{\text{Wr}}$. We follow the same filtering process as for BCP$_{\text{All}}$.

Session length characteristics of BCP$_{\text{Wr}}$ and WC$_{\text{Wr}}$ are shown in Table 5. We see that sessions had 2 follow-ups on average, enabling our analysis to focus on collaboration behaviors.

B.3 Validating Task Classifiers

Setup. Table 6 presents the results of manually validating f_{CoarseT}. Our validation aims to establish the accuracy of identifying coarse tasks so that subsequent analysis is not affected by noisy predictions. Two annotators judged the correctness of f_{CoarseT} predictions for 55 sessions from BCP$_{\text{All}}$ and WC$_{\text{All}}$– 110 sessions in all. We sample an equal number of sessions for each label. The two annotators' judgments were used to compute agreements with Cohen's Kappa, κ. In Table 6 we report metrics across all labels and for the labels considered as writing (i.e. the labels Gen Communicative Text, Gen Creative Text, Gen Technical Text, and Summarize Text).

Results. From Table 6 we see that f_{CoarseT} is able to identify writing sessions in BCP$_{\text{All}}$ and WC$_{\text{All}}$ with a sufficiently high accuracy and substantial agreement. We also note that BCP$_{\text{All}}$ has a slightly lower accuracy than WC$_{\text{All}}$ for identifying writing tasks. We attribute this to the difference in task frequencies in the two datasets (see Figure 7). Bing Copilot was used most frequently for search, and most errors occurred because of mislabeling search sessions as a different task. On the other hand, WildChat was used in nearly equal propor-

tion for various tasks, with errors resulting from different types of writing tasks confused for each other (e.g., communicative vs creative). Since different writing tasks are collapsed into one writing class, the accuracy of identifying writing sessions remains high.

C Log Analysis with PATHs - Details

In §4 we describe our analysis method focused on identifying PATHs and on analyzing how PATHs vary across writing intents (Figure 2). We provide the details of the methods here:

- Appendix C.1: Details for the Follow-up Classifier and Identifying PATHs with PCA.
- Appendix C.2: Details for with Intent Classifier, Regression Analysis, and Qualitative Analysis.
- Appendix C.3: Details about manually validating the Follow-up and Intent Classifier.

C.1 Identifying PATHs

We identify PATHs by first classifying users' follow-up utterances into high-level follow-up types and clustering these into PATHs using PCA.

Follow-up Classifier. We use a GPT-4o based follow-up classifier (f_{FollowU}) to classify user utterances in BCP$_{\text{Wr}}$ and WC$_{\text{Wr}}$ into original requests or one of eleven follow-up types $\mathcal{F}$ (Table 2). Table 9 shows example follow-up utterances for each label.

Prompt 2 was used for f_{FollowU} and contains labels and their definitions. Our follow-up type labels were derived by clustering user utterances embedded with a pretrained language model and manually generating labels for each cluster. We iteratively repeat this process until no new labels are generated. Then the labels were merged to remove redundancies. We confirm their coverage by manually applying the labels to 100 randomly sampled BCP$_{\text{Wr}}$ sessions. Our labels (Table 2) closely mirror labels generated in prior conversational interaction datasets (Qu et al., 2018) – indicating their generality. We validate the accuracy of f_{FollowU} through manual annotation in Appendix C.3.

PATHs with PCA. After labeling user utterances with f_{FollowU} we identify PATHs. Given a dataset of sessions $\mathcal{S}$, we represent each session $S \in \mathcal{S}$ using a "`tf-idf`" representation of follow-up types, $\mathbf{F} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{F}|}$. Each feature vector per session in $\mathbf{F}$ contains the follow-up type counts normalized by the total number of user utterances in the session ("`tf`"). These frequencies are then multiplied by

the log inverse of the follow-up frequency in the entire dataset ("idf"). This ensures that more frequent follow-ups don't dominate the variance of our dataset.

Next, we run PCA on $\mathbf{F}$ and treat each identified principal component as a PATH. PCA is a linear dimensionality reduction technique which maps $\mathbf{F}$ to a new space of variables as $\mathbf{P} = \mathbf{FW}$. The dimensions of $\mathbf{P}$ are uncorrelated with each other, and $\mathbf{W}$ ensures that the variance of $\mathbf{F}$ is retained in $\mathbf{P}$ (Bengio et al., 2013). Further, the dimensions of $\mathbf{P}$ capture decreasing amounts of the variance in $\mathbf{F}$. In our work, we retain the first l $(< |\mathcal{F}|)$ dimensions of $\mathbf{P}$ that explains 80-85% variance in $\mathbf{F}$ while rejecting the rest as noise, therefore, $\mathbf{P}_l \in \mathbb{R}^{|\mathcal{S}| \times l}$. Because each dimension of $\mathbf{P}_l$ represents a mutually co-occurring set of follow-ups, they segment $\mathcal{S}$ into sessions with a consistent PATH. We visualize a normalized $\mathbf{W}_l$, referred to as the "loading matrix", in §5 to show the correlations between follow-up types and PATHs.

C.2 Correlating intents and PATHs

Intent Classifier. To correlate writing intents with PATHs, we classify 'Original Requests' identified by f_{FollowU} into fine-grained writing intents (Table 3) using a GPT-4o based classifier (f_{WritingI}). We set up f_{WritingI} as a multi-label classifier since requests can express more than one intent.

Prompt 3 was used for f_{WritingI} and contains labels and their definitions. Mirroring label generation for follow-ups, we derive writing intents iteratively. We cluster user utterances in BCP_{Wr} and manually generate intent labels per cluster while merging redundant intents. We repeat this procedure until new labels aren't generated. We manually apply the labels to 200 randomly sampled BCP_{Wr} sessions to ensure their coverage. Table 10 shows example original requests for each writing intent. We validate the accuracy of f_{WritingI} through manual annotation in Appendix C.3.

Regression Analysis. We run logistic regression analysis[5] by treating fine-grained intents of a session as predictors and the sessions PATH as a target variable. We rely on logistic regressions due to their ease of interpretation and follow a large body of prior work (Gujarati, 2021). In our regression, we represent the fine-grained intents identified by f_{WritingI} as one-hot features, $\mathbf{I}$, of the sessions – these serve as predictors. Next, we obtain the ses-
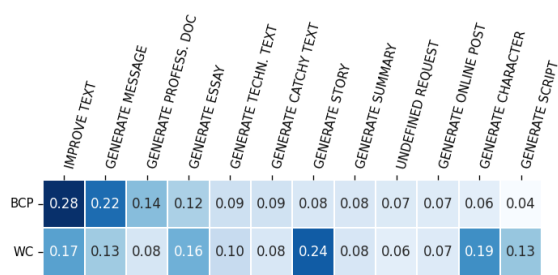
---

[5] `statsmodels.Logit`

Figure 8: Frequencies of writing intents in BCP_{Wr} and WC_{Wr} identified by multi-label f_{WritingI}. From the top-3 intents, we see that Bing Copilot was used for creating communication or professional texts, while WildChat was used for creative texts such as stories, scripts, or characters. Intents occurring in fewer than 3% sessions are excluded for space.

sions membership in l different PATHs, $\mathbf{M}_l$, by thresholding the columns of $\mathbf{P}_l$. Then, we correlate the intents $\mathbf{I}$ with PATHs of $\mathbf{M}_l$. This results in l independent logistic regressions: $\text{sigmoid}(b + \mathbf{Ic})$. In thresholding $\mathbf{P}_l$, we select a score to retain 15-20% of the sessions with the highest scores per dimension in $\mathbf{P}_l$. Our thresholds were selected to ensure the accuracy of our logistic regressions. In our results, we examine the coefficients $\mathbf{c}$ – correlating the intents with each one of the l PATHs. Finally, note that the presence of a follow-up type in a session can be used directly as a binary target variable instead of PCA-based PATHs – we pursue both paths in our analysis.

Note that we also use the length of original requests as predictors and find that they are uncorrelated with PATHs (§6). Similarly, early analysis also found the deployment LLM be uncorrelated with PATHs – likely due to PATHs capturing high-level behaviors. Therefore they were dropped as features in subsequent analysis. To indicate the goodness of classifier fit, Table 7 presents the accuracy of classifiers used to identify correlations between writing intents and PATHs/follow-up types and shows significantly better accuracy than a random classifier (50% accuracy).

Qualitative Analysis. Our automatic approach highlights correlated writing intents and PATHs. To gain a deeper understanding of user behaviors in a correlated intent-PATH pair, two authors conducted a qualitative analysis of the pairs, which showed statistically significant correlations and were repeated in BCP_{Wr} and WC_{Wr}. For each pair, both authors independently examined the top 300 conversations with the highest principal component scores, aim-

Target in BCP_{Wr}, WC_{Wr}	Description	Accuracy (%) BCP_{Wr}, WC_{Wr}
PATH2, PATH2	Requesting more output.	79, 79
PATH3, REQ. ANS.	Requesting answers.	77, 74
PATH4, ADDS CON.	Adding content.	76, 74
CH. STY., PATH4	Changing style.	61, 78
PATH1, PATH1	Revising requests.	73, 74
PATH6, PATH6	Elaborating on requests.	84, 84

Table 7: The accuracies for logistic regressions used to find correlations between writing intents and PATHs or follow-up types in our regression analysis.

	Follow-ups		Intents	
	Acc. (%)	Agr. (κ)	Acc. (%)	Agr. (κ)
BCP_{Wr}	79.09	0.78	81.58	0.79
WC_{Wr}	84.55	0.74	78.70	0.82

Table 8: Accuracy (Acc.) of GPT-4o classifiers for labeling follow-up types ($f_{FollowU}$) and writing intents ($f_{WritingI}$) in BCP_{Wr} and WC_{Wr}. Agreements (Agr.) are computed from judgments made by two independent annotators.

ing to understand the user goals for engaging in a PATH. Then both authors met and resolved the differences in their understanding. In our analysis, the two authors differed in the interpretation of one of the six intent-PATHs presented in our results. Our analysis was conducted to overcome the lack of access to Bing Copilot or WildChat users who could be probed about their intents and behaviors – a fundamental challenge in all log-based studies (Dumais et al., 2014).

C.3 Validating Follow-up and Intent Classifiers

Setup. Table 8 presents the results of manually validating $f_{FollowU}$ and $f_{WritingI}$. Similar to $f_{CoarseT}$, our annotation validates the accuracy of $f_{FollowU}$ and $f_{WritingI}$ to establish that the subsequent analysis is error-free. Validation was conducted by two co-authors serving as annotators, they manually judged the correctness of $f_{FollowU}$ predictions for follow-ups and $f_{WritingI}$ predictions for original requests. Both annotators judged the correctness of the labels independently and on the basis of a shared annotation guideline. Predictions were considered incorrect if a different label was applicable. The two annotators' judgments were used to compute agreements with Cohen's Kappa, κ. We constructed our validation data by selecting follow-ups from 110 sessions and original requests from a different set of 110 sessions – 220 sessions in all. We ensured that every follow-up and writing intent label was uniformly present in the validation data. BCP_{Wr} and WC_{Wr} had 110 sessions each in the validation data. For the multi-label $f_{WritingI}$ annotators independently judged the correctness of all predicted labels.

Results. From Table 8 we note that the follow-up and writing intent classifiers have a reasonably high accuracy and substantial agreement between the annotators. Our annotators noted that most writing intent errors were the result of hard-to-

distinguish boundaries between creative writing tasks (i.e. GENERATE STORY, GENERATE CHARACTER, GENERATE SCRIPT). Similarly, follow-up type errors were the result of hard-to-distinguish boundaries between different ways to update a generation (i.e. CHANGE STYLE, ADD CONTENT, REMOVE CONTENT).

D Extended Results

We discuss how writing intents correlate with PATHs in §6. We include additional results here.

- Appendix D.1: Discusses writing intents correlated with CHANGE STYLE/PATH4 – we include it here for a shortage of space.
- Appendix D.2: Includes dataset dependent trends for PATH1 and 5 – we include it here for completeness.
- Figures 12, 13, and 14 supplement the result (§6) and discussion (§7).

D.1 Changing style to align generations with readers

Figure 9a shows the writing intents correlated with CHANGE STYLE where users modified the style of LLM generations. This PATH correlated with intents to improve texts across domains (I1), and generate messages, summaries, and online posts (I2, I4, I8). Analysis of sessions revealed that users frequently changed the style of generations to better match the assumed norms for a communication (Ex 2, 3, 4) or to the likely preference of readers (Ex 1, 5). Specifically, when users sought to IMPROVE TEXT (I1), their follow-ups changed the style to better match the preferences of specific readers or the business context of a communication (Ex 1, 2). This remained true for communicative texts intended for groups or individuals (Ex 3, 4). And when users generated summaries, they sought to change their format for readability (Ex 5), focusing on readers. Our findings lend support
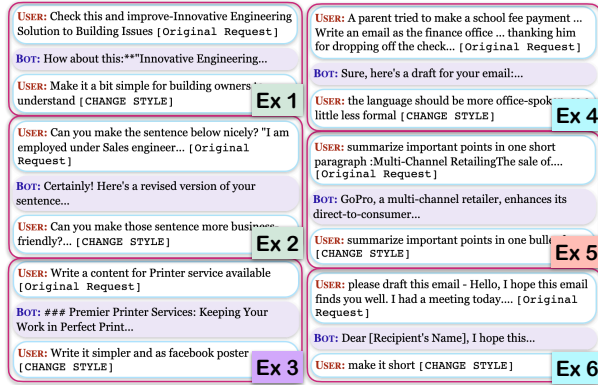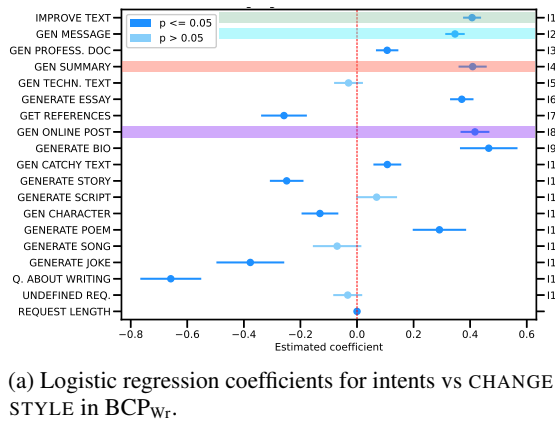
16837

(a) Logistic regression coefficients for intents vs CHANGE STYLE in BCP$_{Wr}$.

(b) Users engaged in intents I1, I2, I4, and I8.

Figure 9: (a) The intents discussed in §D.1 are highlighted in color. Coefficients for WC$_{Wr}$ are plotted in Figure 12. (b) Example conversations from the intents highlighted in (a) – intent and example colors are matched.

to prior work on style matching in communicative settings. This work notes that style matching is crucial for engagement and community identity on online platforms (Ananthasubramaniam et al., 2023; Tran and Ostendorf, 2016), and an important element of effective communication (Pickering and Garrod, 2013). Our work extends style matching to the case of user-LLM co-creation. Inferring and aligning LLMs to aid co-creation for specific audiences remains under-explored – this may be seen as a form of reader personalization. Finally, across intents we also noted CHANGE STYLE to request more concise outputs (Ex 6) – this may be attributed to RLHF aligned models' generating lengthy outputs (Singhal et al., 2024).

Implications: • Infer intended readers and their preferred style, enabling LLMs to customize the generation style to them.

D.2 Dataset dependent trends in revising and elaborating on requests

Figure 10 and 11 show writing intents correlated with PATH1 and PATH5. We avoid discussing them in detail and include them for completeness. Respectively, they represent users revising original requests and elaborating on requests in follow-ups. For both PATHs, we don't see positively correlated and statistically significant intents across BCP$_{Wr}$ and WC$_{Wr}$. We hypothesize that the dataset-dependent correlation with writing intents in PATH1 may be because users provide many different forms of feedback in a revised prompt (Chen et al., 2021), making correlation with specific writing intents less likely. Further, we also hypothesize that both PATH1 and 5 may be closely tied to user satisfaction, which is likely to be system dependent. In this

regard, recent work (Sarkar et al., 2025) leverages the feedback contained in revised requests to improve LLM performance. Future work may attempt to identify more distinguished PATHs through a finer-grained analysis of sessions with revised or elaborated requests.
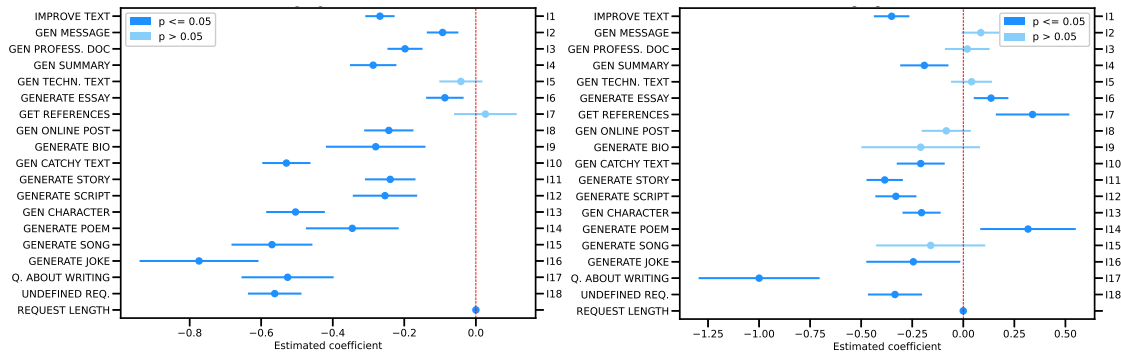
(a) Writing intents vs PATH1 in BCP_Wr.
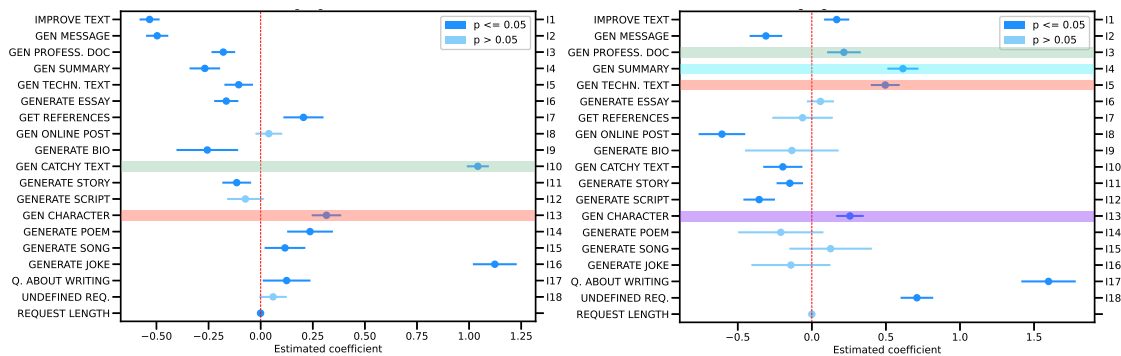
(b) Writing intents vs PATH1 in WC_Wr.

Figure 10: Coefficient plots for writing intents vs PATH1 (revising requests) showing dataset specific trends.
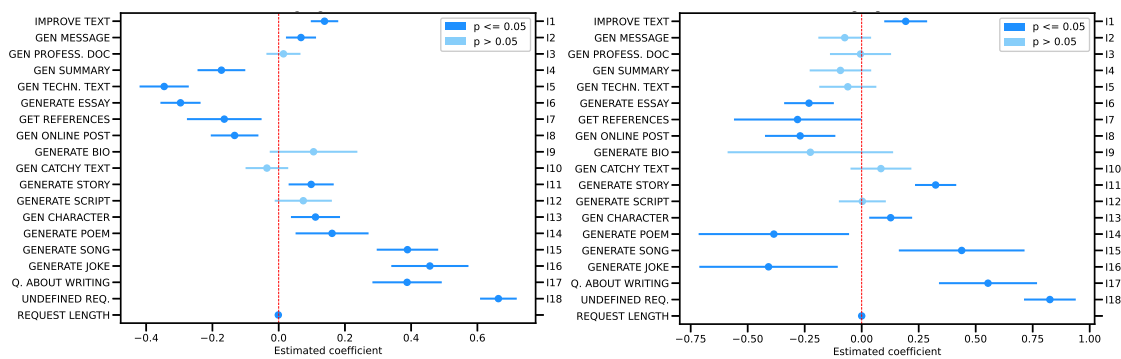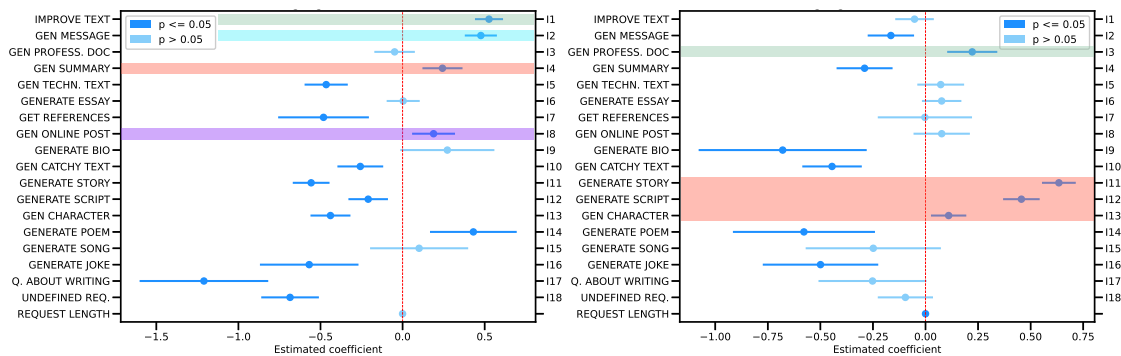

(a) Writing intents vs PATH5 in BCP_Wr.

(b) Writing intents vs PATH5 in WC_Wr.

Figure 11: Coefficient plots for writing intents vs PATH5 (Elaborating on requests) showing dataset specific trends.

(a) Writing intents vs PATH2 in BCP_{Wr}.

(b) Writing intents vs PATH3 in WC_{Wr}.

(c) Writing intents vs PATH4 in WC_{Wr}.

(d) Writing intents vs ADDS CONTENT in WC_{Wr}.

Figure 12: Supplementary coefficient plots for Figure 4a, Figure 5a, Figure 9a, and Figure 6a. Colored intents are discussed in §6 and match the colors of their corresponding coefficient plots and examples in §6.
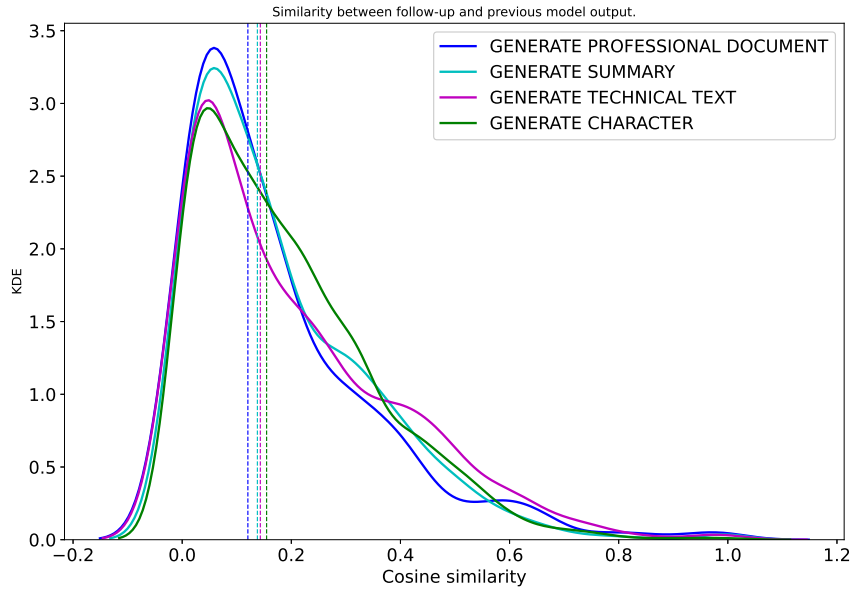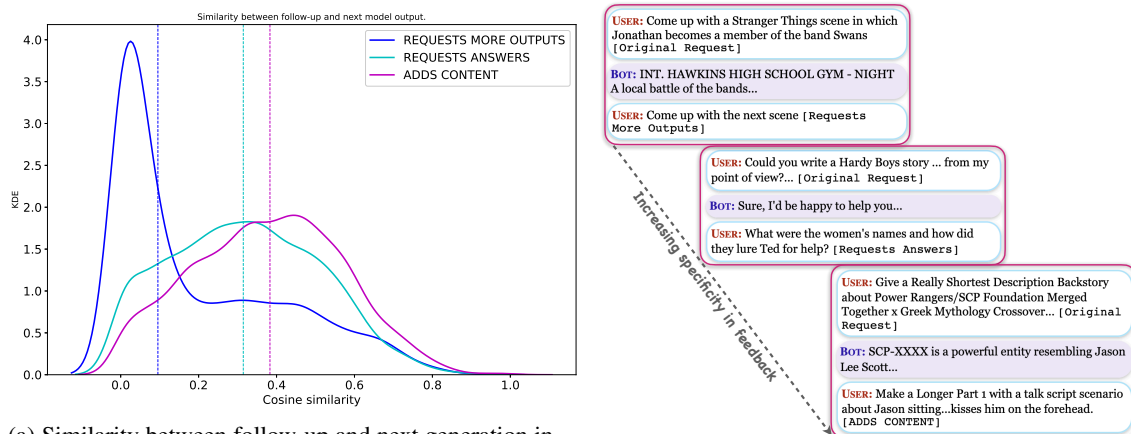
Figure 13: Increasing similarity between the follow-up utterance and the previous model generation in WC_{Wr} for writing intents correlated with PATH3 (asking follow-up questions). We discuss this in §6.2. Similarities are increase from users co-creating professional documents, to summaries, and technical texts. But when users co-create fictional character narratives, we see the highest similarity, indicating that they tend to ask closed-domain questions grounded in the model generation. Similarity is based on a `tf-idf` representation of texts. Vertical lines indicate the median similarity.

(a) Similarity between follow-up and next generation in WC_{Wr} for creative writing intents.

(b) Increasing specificity of feedback in user follow-ups.

Figure 14: (a) Increasing similarity between user follow-up and the next model generation across follow-up types REQUESTS MORE OUTPUTS, REQUESTS ANSWERS, and ADDS CONTENT. This indicates an increasing amount of specificity in user feedback across the three kinds of follow-up types. We discuss this in §7. (b) Example conversations from WC_{Wr} that illustrate the increasingly specific follow-up types for the writing intent GENERATE CHARACTER. Similarity is based on a `tf-idf` representation of texts. Vertical lines indicate the median similarity.

Prompt 1 The prompt for GPT-4o based multi-label Task Classifier (f_{CoarseT}) used to obtain the writing task sessions from BCP$_{\text{All}}$ and WC$_{\text{All}}$. This is discussed in Section 3 and Appendix B. The text in green maps the labels used in the prompt to the labels used in Figure 5 and the rest of the paper. **Bold text** is replaced with raw conversation data.

```
# Task Instructions
You an an experienced linguist who helps analyze conversations. Given a CONVERSATION
 between a user and an assistant, classify the CONVERSATION into one or more of the
following labels.
INFORMATION SEARCH: The user is asking a question about a specific document or could
 have searched for this information on the internet. ← Search Information
GENERATING COMMUNICATIVE TEXT: The user is trying to generate text for communication
 with an individual, a group, or an online platform. ← Gen Communicative Text
SOFTWARE DEVELOPMENT: The user is seeking assistance in software development. Often
phrased as a how to question with accompanying code. ← Develop Software
HOMEWORK PROBLEM: The user is posing a question or problem which is likely to be
from an exam or homework. ← Solve Problem
GENERATING CREATIVE TEXT: The user is seeking to generate creative stories,
characters, titles, slogans and other creative texts. ← Generate Creative Text
GENERATING TECHNICAL TEXT: The user is working with the assistant to generate
technical or research texts. ← Generate Technical Text
SUMMARIZATION: The user seeks to summarize a text that they provide. ← Summarize Text
TRANSLATION: The user is seeking to translate a text or understand text in a
language that isnt English. ← Translate Text
IMAGE GENERATION: The user asks to generate an image or other forms of visual art.
 ↖ Generate Image
EXPLICIT CONTENT: The user or bot text contains explicit or pornographic content.
 ↖ Generate Explicit Text
UNDEFINED: A request for which none of the defined labels are applicable or there
isn't an explicit request by the USER. ← Undefined Request

For your response use the following instructions:
1. Output one of more of the correct labels for the CONVERSATION
2. Output the labels in decreasing order of their relevance to the CONVERSATION

CONVERSATION:
{{conversation_text}}
Output a explanation and then one or more of the labels for the CONVERSATION.
```

Prompt 2 The prompt for GPT-4o used to label user utterances as original requests and or one among eleven follow-up types in BCP$_{\text{Wr}}$ and WC$_{\text{Wr}}$. This is discussed in Section 4 and Appendix C.1. The text in green maps the labels used in the prompt to the labels used in Table 2 and the rest of the paper if they differ. **Bold text** is replaced with raw conversation data.

```
# Task Instructions
You an an experienced linguist who helps analyze conversations. Given a conversation
 between a USER and an ASSISTANT, label the USER UTTERENCES with the below labels.
Use the following labels:
NEW REQUEST: The user is making a new request to the ASSISTANT for a new topic and
for the first time. ← ORIGINAL REQUEST
RESTATES REQUEST: The user is restating the NEW REQUEST with a modification.
ELABORATES REQUEST: The user is elaborating on the NEW REQUEST without restating
the request.
COURTESY RESPONSE: The user is responding as a courtesy to the ASSISTANT (e.g. "yes
please", "go ahead" etc) or is exchanging pleasantries with the ASSISTANT.
CHANGE RESPONSE: The user is requesting the ASSISTANT to change the stylistic
elements (e.g. length, tone, formality etc) of its previous response. ← CHANGE STYLE
ADDS CONTENT: The user is requesting new content to be added to the ASSISTANTs
response.
REMOVES CONTENT: The user is requesting specific content to be removed from the
ASSISTANTs response.
REQUESTS ANSWERS: The user is requesting new or additional information related to
the ASSISTANTs response.
REQUESTS MORE OUTPUTS: The user is requesting additional output from the ASSISTANT.
RESPONDS POSITIVELY: The user is responding positively to the ASSISTANTs response
without providing additional detail.
RESPONDS NEGATIVELY: The user is responding negatively to the ASSISTANTs response
without providing additional detail.
UNDEFINED RESPONSE: None of the defined labels describe the users response.

Use these rules when you respond:
1. Output one best label for every USER UTTERENCE.
2. Output the labels in the order of the USER UTTERENCES.

Here is the conversation:
{{conversation_text}}
Output a label for each USER UTTERENCE.
```

Follow-up Type	Description	Examples (original request and follow-up)
RESTATES REQUEST	Reformulates their request	"Msitu Africa is requesting for funding through crowdfunding websites. Write for me a small compelling write up requesting for funds." → "Fundraiser Story. Msitu Africa is requesting for funding through crowdfunding websites. Write for me a small compelling write up requesting for funds to help us actualize our initiative." "wrire a detail script for animation video about Pan Am Flight 103" → "write a detail script for animation video about Pan Am Flight 103, video will be 20 minutes in length"
ELABORATES REQUEST	Expands on their request; often after being asked by the LLM	Can you help me with writing the discussion for my article? → "The article objectives are to explore and analyze the various impacts of D-penicillamine (DPCA) treatment..." "i want to provide you with 4 separate inputs as follows story1 chapter1, story1 chapter7, story2 chapter7, i would like you to write story2..." → "Chapter 1: Shadows of Rebellion..."
REQUESTS ANSWERS	Question related to the output	"My potential client has just returned from a trip last week. How should I greet him in an email" → "What should I write in the subject line in my email" "I want you to write a story for me. There are two girl who hate each other. They are going to wrestling in front of their friends..." → "can you describe me last pin position ?"
REQUESTS MORE OUTPUTS	Asks for additional output	"Write 10 funny and short answers to this comment: ..." → "Write more" "you are a novelist. settings - late medieval england like kingdom. kingdoms A and B. ... list 25 approaches A can benefit from" → "Give 25 more"
CHANGE STYLE	Changes style of output	"rewrite this Hotel A is an excellent choice for tourists seeking a memorable and enjoyable stay..." → "rewrite it in 240 words" " write an email to change the name of the HR manager, which due to mistake on my part was spelled incorrectly" → "write something instead of Dear"
ADDS CONTENT	Adds content to output	"write in bullet points my responsibilities as an intern during a neurology rotation" → "mention discussions" "script about cincinnati blowing a 21 point lead vs johnny manziel" → "the next week seattle blew out cincinnaty by more than 30"
REMOVES CONTENT	Remove content from output	"Make this into a episode and give it More life" → "Remove panel 1-4 and re number the rest" "Today is my senior patrol election and I want to win. I need to provide a 2-3 minute speech..." → "Remove the details about having less experience and emphasize how I want to use my creativity..."
COURTESY RESPONSE	A courtesy or pleasantry	"write a short 100 word Youtube channel description for a channel..." → "thank you" "Can you write a 300 pages English novel if I give you the plot" → "Ok so can we start"
RESPONDS POSITIVELY	Explicitly pleased with output	"In my book there is a central plot with the main characters who influence global events..." → "excellent, keep going" "i want to create 4 different kinds of postings. one should be Good morning" → "i like them plase continue with 10 for each type"
RESPONDS NEGATIVELY	Explicitly unhappy with output	"Could you come up with some two syllables long feminine robot names? They need to be based..." → "Most of those where three syllables long" "Write a brief man to man heartfelt reply to this comment..." → "Do you not understand brief? No yapping"

Table 9: Example utterances for each follow-up type selected from WC_Wr. We include the original requests to better illustrate the follow-up utterance by the user. We omit the LLM response for space.

Prompt 3 The prompt for GPT-4o used to label original requests with finer-grained writing intents BCP_{Wr} and WC_{Wr}. This is discussed in Section 4 and Appendix C.2. The text in green maps the labels used in the prompt to the labels used in Table 3 and the rest of the paper if they differ. Bold text is replaced with raw conversation data.

```
# Task Instructions
You an an experienced linguist who helps analyze conversations.
Given a USER REQUEST to an assistant, classify the USER REQUEST into one or more of
the following labels:
GENERATE SONG: Request to write or re-write a song.
GENERATE JOKE: Request to write or re-write a joke.
GENERATE POEM: Request to write or re-write a poem, haiku or a similar other verse.
GENERATE STORY: Request to write or re-write a story.
GENERATE SCRIPT: Request to write or re-write a script for a video, drama, movie or
similar other media.
GENERATE CHARACTER: Request to generate a fictional character and their story.
GENERATE CASUAL BIO: Request to write or re-write a bio for an online or mobile app
profile. ← GENERATE BIO
GENERATE PROFESSIONAL DOCUMENT: Request to write or re-write a resume,
recommendation letter, or other professional document.
GENERATE MESSAGE: Request to write or re-write a interpersonal message, cover letter
, email, letter or other interpersonal communication.
GENERATE ONLINE POST: Request to write or re-write a text for an online platform
like a social media post, review etc.
GENERATE TITLE: Request to write a title, slogan, or other eye catching text.
 ↖ GENERATE CATCHY TEXT
GENERATE SUMMARY: Request to summarize a text provided by the user.
GENERATE ESSAY: Request to write an essay, blog, article or other long text on any
topic.
GENERATE TECHNICAL TEXT: Request to write or re-write a scientific or technical text
QUESTION ABOUT WRITING: A question about literary works, writing or publishing.
IMPROVE TEXT: Request to improve grammar, style, tone or other aspect in the
provided text.
GET REFERENCES: Request to add, format or generate references.
UNDEFINED REQUEST: A request for which none of the labels are applicable, or there
isn't an explicit request, or a request that a writing assistant cannot fulfill.

For your response use the following instructions:
1. Output one of more of the correct labels for the USER REQUEST.
2. Output the labels in decreasing order of their relevance to the USER REQUEST.

USER REQUEST: {{user_request}}
Output a explanation and then one or more of the labels for the USER REQUEST.
```

Writing intent type	Example original requests
IMPROVE TEXT	"say better and more professional: Beckey, really appreciate your advocacy during" "Can you improve and make a variation of this sentence?: Is to be expected of any aspect..."
GENERATE MESSAGE	"Hi, May I have your updated delivery schedule for face plate ... How to reply to customer and we have stock 250pcs can arrange delivery on 16 jan 2024." "write a one sentence valentine card message for a girl who bought a new pontiac..."
GENERATE PROFESSIONAL DOCUMENT	"create a school policy regarding parents using a school account ..." "I am writing an annual performance appraisal for my direct report who is a sales engineer..."
GENERATE SUMMARY	"give introduction to my chapter 5 (conclusion and future research) based on this text: Discussion: This study utilized..." "Make a summary of this including tools and what they are used for: A version control system..."
GENERATE TECHNICAL TEXT	"write me a report about miss feeding amc ... in 300 words" "I am writing a project proposal for seed funding for an engineering project ... for my project "Waveguard" for about 1000 words using ... if there are extra information needed, please let me know."
GENERATE ESSAY	"Write me a 500 word essay about zodiac serial killer" "Write a blog post where Pop in France has picked up several new shows..."
GET REFERENCES	"Please explain, with examples, and in detail, why Western cultures tend to value a sun tan ... Please cite sources, APA style." "as a postgraduate student I want a literature review to my scientific paper ... do the references."
GENERATE ONLINE POST	"Heart Shaped Pendant Necklaces ... This is the title for a jewelry I am selling on etsy. Write a short and descriptive description that is intended for a younger teen audience..." "Write 10 short tweets about today's rainy morning at the beach from one of the most windy places"
GENERATE BIO	"Write me a bio for me 450 charters of someone who's loves to read books" "i want to write a short biography about myself saying that i am an audio engineer that graduated from..."
GENERATE CATCHY TEXT	"generate 5 new original desalination company names" "suggest me 10 thumbnail texts for this youtube video Young Boy Transfers..."
GENERATE STORY	"Can you write a hypothetical what if alternate history scenario, what if Russia never sold Alaska to the US..." "Once upon a time, in the heart of a vast national park, there lived a dedicated and diligent park worker ... throw a party to celebrate it's recovery"
GENERATE SCRIPT	"Write a funny dialogue where /co/ anon asks Valerie to go to a cottage..." "I have a youtube short Idea I need you to write a script. It is legendary 1v1s..."
GENERATE CHARACTER	"Make a Really Shortest Description Backstory about SpongeBob SquarePants..." "Describe rogue gas giant named Silicodum, it has frozen dense hazes of silicate..."
GENERATE POEM	"Write me a funny poem about my brother malachi..." "Turn this into an unrhyming poem..."
GENERATE SONG	"make a rap song to pony by geniune about amy oppong" "Write a music-hall song called "The Parlor Upstairs" ..."
GENERATE JOKE	"You are a Radio DJ at 99.7 Now Radio. Write me a funny radio break about presenting the next radio program..." "tell me some jokes with a naval or military theme which can be used in a farewell speech"
QUESTION ABOUT WRITING	"could a person with potential with lots of friends ... be an effective Villain backstory? How can it be written well, and how can it be written poorly?" "The following excerpt is from an alternate history novel. Can you critique the system of government described in the excerpt?..."

Table 10: Example original requests from WC_{Wr} for each writing intent type.