POLITECNICO DI TORINO
Repository ISTITUZIONALE

Investigating How Computer Science Researchers Design Their Co-Writing Experiences With AI

Original
Investigating How Computer Science Researchers Design Their Co-Writing Experiences With AI / Monge Roffarello, A., Calò, T., Scibetta, L., De Russis, L.. - ELETTRONICO. - (2025), pp. 1-17. (CHI '25: CHI Conference on Human Factors in Computing Systems Yokohama (JPN) April 26 - May 1 2025) [10.1145/3706598.3713205].

Availability:
This version is available at: 11583/2996669 since: 2025-04-22T08:07:51Z

Publisher:
Association for Computing Machinery

Published
DOI:10.1145/3706598.3713205

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

28 November 2026

Investigating How Computer Science Researchers Design Their Co-Writing Experiences With AI

Alberto Monge Roffarello
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy
alberto.monge@polito.it

Tommaso Calò
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy
tommaso.calo@polito.it

Luca Scibetta
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy
luca.scibetta@polito.it

Luigi De Russis
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy
luigi.derussis@polito.it

Abstract

Recent advancements in AI have significantly enhanced collaboration between humans and writing assistants. However, empirical evidence is still lacking on how this collaboration unfolds in scientific writing, especially considering the variety of tools researchers can use nowadays. We conducted observations and retrospective interviews to investigate how 19 computer science researchers collaborated with intelligent writing assistants while working on their ongoing projects. We adopted a design-in-use lens to analyze the collected data, exploring how researchers adapt writing assistants during their use to overcome challenges and meet their specific needs and preferences. Our findings identify issues such as workflow disruptions and over-reliance on AI, and reveal five distinct design-in-use styles—*teaching*, *resisting*, *repurposing*, *orchestrating*, and *complying*—each consisting of different practices used by researchers. This study contributes to understanding the evolving landscape of human-AI co-writing in scientific research and offers insights for designing more effective writing assistants.

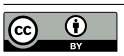
CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Interactive systems and tools.*

Keywords

Generative AI, scientific writing, writing assistants, design-in-use

ACM Reference Format:
Alberto Monge Roffarello, Tommaso Calò, Luca Scibetta, and Luigi De Russis. 2025. Investigating How Computer Science Researchers Design Their Co-Writing Experiences With AI. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3706598.3713205



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713205

1 Introduction

Recent advancements in AI—including Large Language Models (LLMs)—have opened up new possibilities for human-AI collaborative writing, enabling richer interactions compared to traditional word-level auto-completion. This has reached the point where writers can now create entire paragraphs or sections with AI assistance in a collaborative fashion [9, 14, 36]. Such a collaboration is supported by a variety of *intelligent writing assistants* that *"assist users with improving the quality and effectiveness of their writing, from grammar and spelling checks to idea generation, text restructuring, and stylistic improvement"* (Lee et al. [35], p. 2). This definition includes any tools that are a) intelligent (i.e., powered by AI); b) interactive (i.e., reflecting human input and/or output through an iterative process); and c) focused on generating text in natural language.

While the potential of co-writing with intelligent writing assistants is increasingly recognized, the specifics of how this collaboration unfolds in the context of *scientific writing* are still being explored [14, 64]. Scientific writing is an iterative and non-linear process that involves multiple phases and levels of abstraction, from ideation to evidence gathering, involving unresolved tensions such as concerns about agency and ownership [27, 44], plagiarism [34], and the generation of factually incorrect information [45]. It can encompass various writing forms—from argumentative to descriptive—demanding special attention due to its heightened stakes, as the societal impact of research amplifies the need for accuracy, integrity, and ethical rigor. Furthermore, writing assistants for scientific writing—including those originated as a research artifact—often focus on specific writing phases, exposing functionality that are best suited for particular use cases [64]. Recent works in Human-AI Interaction suggest that users often struggle to "teach" everyday intelligent systems with their preferences and current usage goals [16, 32], leading them to repurpose these tools in creative ways [32]. Given the inherent complexity of scientific writing and the varied capabilities of contemporary writing assistants, we hypothesize that similar misalignments may occur in this domain, where, for instance, chat-based systems like ChatGPT might overlook implicit writing context and user intent [64].

Overall, these gaps motivated us in exploring more closely how reseachers take advantage of intelligent writing assistants to support their scientific writing in their everyday working activities. Specifically, we adopted a *design-in-use* lens [32, 46], investigating **how researchers 'design' their own co-writing experiences with writing assistants**, adapting them to overcome challenges and meet their specific needs and preferences during use. As explained by Kim and Lim [32], the design-in-use construct underscores the importance and the role of end users in designing their own experiences with intelligent systems, embracing the open-ended and uncertain nature of AI rather than confining it to fixed design choices.

To run our investigation, we conducted remote observations on how 19 computer science researchers from diverse backgrounds, academic roles, and native languages interacted with intelligent writing assistants while working on their ongoing projects. We focused on computer science researchers recognizing them as early adopters of AI who bring critical perspectives and informed strategies to its integration in scientific writing, acknowledging potential biases as a baseline for future studies with broader user groups. Specifically, we considered computer science researchers at public universities with at least one year of research experience, including Ph.D. students in their second year or beyond, whose expertise—from HCI to software engineering—position them to engage deeply with AI tools in scientific writing. Each writing session was then followed by a retrospective semi-structured interview through which we integrated what we observed with participants' qualitative feedback.

Our findings reveal a scenario where researchers employ a diverse range of AI-powered tools, including general-purpose LLMs like ChatGPT and AI grammar checkers like Grammarly, across different stages of the writing process. Their motivations vary widely, from summarizing literature and generating new ideas to refining text. Nevertheless, none of the participants relied solely on AI to generate text, and observations and participants' feedback revealed several challenges and concerns, including disruptions to writing workflows and worries about over-reliance on AI, which could lead to a loss of writing skills. We identified five design-in-use styles, each consisting of different practices that, depending on the underlying AI technology, researchers employ to adapt their writing assistants to their needs. Similar to findings in other domains [32], our participants engaged in i) *teaching* information to a writing assistant, particularly when interacting with general-purpose LLMs; ii) *resisting* the use of writing assistants and their suggestions; and, in a minor number of cases, iii) *repurposing* writing assistants to redefine their potential uses. Within these design-in-use styles, we identified practices that are unique to the scientific writing domain, such as (re)aligning an assistant with the status of previous writing and cross-referencing AI-generated outputs with external sources to ensure accuracy and reliability. Additionally, we observed two novel design-in-use styles: iv) ***orchestrating*** different tools and approaches to optimize writing, and v) ***complying*** with the suggestions of a writing assistant—especially grammar checkers—to achieve a perfect score or anticipate the assistant's outcomes.

To summarize, our work makes the following contribution:

- Remote ethnographic observations into how researchers collaborate with intelligent writing assistants to support their scientific writing in their daily work activities. To our knowledge, no prior study has observed how researchers use these tools while working on real tasks.
- Five design-in-use styles and 14 related practices highlighting how researchers adapt their writing assistants to their needs and design their own co-writing experiences with AI.
- A discussion of the design implications of our findings, illustrating how supporting the identified design-in-use practices could enhance collaboration with AI in scientific writing.

2 Related Work

2.1 AI in the Research Workflow

The increasing adoption of AI tools in research workflows has been well-documented across numerous scientific disciplines. For example, Morris et al. [47] studied how generative AI might transform scientific work through interviews with twenty scientists across physical, life, and social sciences. They found that scientists view AI primarily as a tool to complement and accelerate their work, such as literature synthesis and experimental design, rather than automate it entirely. Alongside these benefits, participants voiced concerns over issues like hallucination, trustworthiness, and potential downstream effects on scientific training and publication norms. Similarly, Fecher et al. [15] conducted a Delphi study with 72 AI researchers and identified significant apprehensions regarding bias, misinformation, and quality assurance, despite widespread recognition of LLMs' transformative potential. These concerns are amplified by findings from a large-scale survey [37], which revealed that researchers from multiple domains are now adopting LLMs at numerous stages of the scholarly workflow. For example, researchers now employ AI to assist with ideation [26, 38, 49], literature reviews [3], data creation [57], cleaning and analysis [42], and the writing and drafting of research manuscripts [25, 47].

Our emphasis on writing, particularly writing and drafting research papers, is crucial for several reasons. First, writing represents the primary means of communicating scientific knowledge and ensuring research impact. Second, writing involves complex cognitive processes that require maintaining coherence, accuracy, and scientific rigor—aspects that current AI tools may inadvertently disrupt [37, 47]. Third, writing directly connects to core issues of authorship, originality, and research integrity that become increasingly complex with AI involvement [19]. Critics have argued that while AI can democratize access to advanced writing tools, it risks homogenizing creative expression and deepening environmental costs associated with large-scale model deployment [6, 25]. These critiques are particularly salient in scientific writing, where precision, originality, and ethical rigor are crucial [33, 56]. For example, Jakesch et al. [28] investigated how co-writing with opinionated language models can subtly influence users' views, highlighting potential inadvertent bias.

In the writing domain, recent developments of large language models such as ChatGPT [51], Claude [4], and Gemini [20] have allowed the flourishing of fine-tuned variations specialized on writing tasks. Commercial products like Grammarly [23], Ref-N-Write [48],

Smoodin [53], and Writefull [61] exemplify the application of advanced Natural Language Processing (NLP) techniques in widely-used writing tools. This shift towards neural-based models has enabled more sophisticated and context-aware assistance, but also raised new challenges. A primary concern is that LLMs can produce *hallucinated* outputs that are either contradictory to their input or contain unverifiable information, as demonstrated by Ji et al.'s systematic categorization of these phenomena in natural language generation systems [30]. These accuracy issues extend into broader *ethical implications*, with Sun et al.'s analysis revealing how well-articulated but incorrect AI responses can systematically mislead users [56]. Beyond accuracy concerns, the *risk of plagiarism* emerges from these models' unsupervised training on vast datasets that may contain licensed content, as highlighted in Lee et al.'s investigation into collaborative human-AI writing systems and their inherent attribution challenges [34]. In addition, these challenges are compounded by a tendency toward *over-reliance* on AI systems, as studies have shown how AI's comprehensive responses can create a false sense of reliability even when incorrect [31].

Mitigation strategies have been introduced, for example Hoque et al. [27] proposed HaLLMark, a tool to capture the provenance of interaction with an LLM to help writers retain their agency, conform to policies, and communicate their use of AI to publishers and readers transparently. Understanding how researchers navigate these challenges in their writing process emerges as a critical imperative for developing more effective AI writing support. By documenting researchers' actual practices and interaction patterns with AI writing tools, we can better understand opportunities for improved design while identifying potential risks of misuse, meeting scientists' need for tools that enhance rather than hinder their research outputs [47].

2.2 AI-Assisted Scientific Writing

Opportunities and issues around using AI in scientific writing have made it a rapidly evolving research area, prompting scholars to explore diverse human-AI collaborative strategies, authoring paradigms, and design innovations. We can observe several key research directions emerging in this space, primarily stemming from studies that focus on science communication rather than academic writing. A first significant body of work has focused on making scientific writing more accessible and engaging for broader audiences. Gero et al. [21] investigated how LLMs could assist STEM graduate students in creating "tweetorials"—500-word technical explanations for general audiences on Twitter. Their findings revealed that LLMs were particularly valuable for generating initial inspiration, translating complex ideas into clearer language, and providing external perspectives on the writing. Building on this work, Long et al. [39] specifically examined how LLMs could help create engaging "hooks" for tweetorials, demonstrating that enhanced prompts incorporating common experiences significantly improved both relevance and creativity.

Another research direction explores how AI can enhance scientific explanation and communication. Kim et al.'s Metaphorian [33] investigated LLMs' potential in helping writers create extended metaphors for complex scientific concepts, achieving more original and understandable explanations while maintaining coherence. Similarly, Radensky et al. [50] developed tools to help researchers translate their academic papers into blog posts, achieving higher writer satisfaction and more extensive revisions without increasing cognitive load. Researchers have also investigated AI's role in improving writing quality and comprehension. For example, Wang et al. [58] demonstrated that LLM tools could significantly reduce the time required for paper comprehension tasks while improving performance.

While not exclusively focused on scientific writing, AI-assisted argumentative writing research has revealed approaches that could benefit scientific writing tools. Both domains require logical structure, evidence-based reasoning, and clear articulation of ideas. In the argumentative writing domain, Afrin and Litman [2] developed models for analyzing evidence and reasoning—an approach that could help researchers strengthen their scientific arguments and methodology descriptions. Lee et al.'s CoAuthor dataset [36] focused on argumentative writing tasks revealed patterns in how writers integrate AI suggestions that could inform effective human-AI collaboration strategies for scientific manuscript development. Similarly, Zhang et al.'s VISAR [64], though designed for argumentative writing, demonstrated how visual programming and rapid prototyping could support the systematic organization of scientific arguments and iterative refinement of research narratives.

Despite significant academic efforts to develop specialized solutions for researchers utilizing LLMs, neither existing research prototypes nor commercial implementations have successfully supplanted general-purpose language models in scholarly writing tasks [47]. In the absence of domain-specific affordances, researchers have developed their own workarounds and interaction patterns to bridge the gap between these powerful but general-purpose tools and their specialized academic writing needs [25]. Users adapting systems beyond their intended design has recently been investigated by Kim and Lim [32] in the context of recommender systems. Although existing research has contributed valuable insights regarding human-AI collaboration in writing tasks, an empirical approach based on naturalistic observation of researchers' AI-assisted writing practices has not been reported. This research gap assumes particular significance given the inherently iterative and non-linear character of scientific writing, wherein researchers must simultaneously manage multiple cognitive demands while interfacing with an diverse ecosystem of tools. We analyze how researchers adapt and interact with writing assistants during their scientific writing process. The analysis of the collected data revealed the presence of Kim and Lim's identified styles in researchers' practices and uncovered two additional design-in-use styles unique to scientific writing with AI.

3 Methodology

To explore current practices in human-AI interaction for scientific writing, we conducted remote ethnographic observations followed by retrospective interviews. Remote ethnographic studies are well-suited for observing certain tasks, particularly in office environments, with minimal limitations [24, 43]. This method allowed us to derive design-oriented insights from actual human practices, minimizing the risk of biases associated with relying solely on self-reported behaviors or opinions, e.g., as in Kim and Lim [32].

Table 1: An overview of the 19 computer science researchers who participated in our observations and retrospective interviews, including the academic role, area of expertise, years of research experience (YoE), native language, english proficiency, writing experience, and writing assistants frequency of use (FoU).

ID (Gender, Age)	Academic Role	Area of Expertise	YoE	Native Language	English Proficiency*	Writing Experience**	Assistants FoU***
P1 (M, 26)	Ph.D. student	Intelligent tutoring systems	6-10	English	●●●●●	●●●○○	●●●●○
P2 (M, 32)	Postdoc researcher	Human-Computer Interaction	3-5	Hindi	●●●●○	●●●●○	●●●●●
P3 (M, 44)	Associate Professor	Human-Computer Interaction	11-15	Italian	●●●●○	●●●●○	●●●●●
P4 (M, 27)	Ph.D. student	Computational Interaction	3-5	German	●●●●●	●●○○○	●●●●○
P5 (M, 28)	Assistant Professor	Machine Learning	3-5	Italian	●●●●○	●●●●○	●●●●●
P6 (M, 35)	Assistant Professor	Mixed Reality	6-10	Italian	●●●○○	●●●○○	●●●●○
P7 (M, 27)	Ph.D. student	Computer Graphics	3-5	Italian	●●●●○	●●●○○	●●●●●
P8 (M, 28)	Ph.D. student	Intelligent tutoring systems	3-5	English	●●●●●	●●●●○	●●○○○
P9 (W, 36)	Postdoc researcher	Learning Analytics	3-5	Spanish	●●●○○	●●●○○	●●●●●
P10 (W, 32)	Ph.D. student	Learning Science	1-2	English	●●●●●	●●●○○	●●○○○
P11 (M, 33)	Assistant Professor	Software Engineering	11-20	Spanish	●●●○○	●●●○○	●●●●●
P12 (M, 32)	Postdoc researcher	Cybersecurity	3-5	Italian	●●●●○	●●●○○	●●●●●
P13 (M, 35)	Ph.D. student	Machine learning	3-5	Spanish	●●●○○	●●○○○	●●●●○
P14 (M, 36)	Postdoc researcher	Internet of Things	6-10	Italian	●●●○○	●●●○○	●●●●●
P15 (M, 28)	Ph.D. student	Human-AI Interaction	1-2	Italian	●●●●○	●●●○○	●●●●●
P16 (M, 27)	Ph.D. student	Software Engineering	3-5	Italian	●●●●○	●●●○○	●●○○○
P17 (M, 31)	Ph.D. student	Artificial Intelligence	1-2	Spanish	●●●●○	●●○○○	●●●●●
P18 (M, 31)	Ph.D. student	Computer Vision	6-10	Hindi	●●●●○	●●●○○	●●●●○
P19 (M, 38)	Ph.D. student	Cybersecurity	3-5	Portuguese	●●●○○	●●●○○	●●○○○

* ●○○○○ = elementary ●●○○○ = limited working ●●●○○ = professional ●●●●○ = full professional ●●●●● = native or bilingual.
** ●○○○○ = novice ●●○○○ = developing ●●●○○ = competent ●●●●○ = advanced ●●●●● = expert
*** ●○○○○ = never used ●●○○○ = occasionally ●●●○○ = monthly ●●●●○ = weekly ●●●●● = daily

3.1 Participants

We recruited participants for our study using a combination of snowball and convenience sampling, focusing our recruitment efforts on computer science research communities, including those in HCI. We made this choice deliberately, as we considered researchers in computer science to be early adopters of AI and potentially more aware of the risks and benefits of integrating AI into scientific writing workflows. We acknowledge that a deeper technical understanding of AI tools compared to researchers in other disciplines may introduce biases, such as overestimating tool capabilities or undervaluing traditional writing methods. Yet, our aim was to observe, in a naturalistic way, how skilled users adopt intelligent writing assistants, providing a baseline for future studies involving less experienced populations. Prospective participants were asked to complete an initial questionnaire, which covered the following aspects:

- *Demographic and background:* age, gender, nationality, current academic role, years of research experience, and area of expertise.
- *Language:* native language and self-assessed level of English proficiency (*native or bilingual*, *full professional*, *professional*, *limited working*, *elementary*).
- *Writing:* ongoing writing projects (*not currently writing*, *one project*, *two to five projects*, *more than five projects*) and self-assessed level of writing expertise (*novice*, *developing*, *competent*, *advanced*, *expert*).
- *Experience with writing assistants:* used tools and frequency of use (*never used*, *occasionally*, *monthly*, *weekly*, *daily*).

Participants qualified for the study if they met the following minimum selection criteria: *a)* working as a computer science researcher at a university with at least 1 year of experience, including Ph.D. students in their second year or beyond; *b)* currently writing a scientific document in English, such as a research paper or grant proposal; and *c)* regularly using at least one intelligent writing

assistant—as previously defined according to Lee et al. [35]—in the writing process. We stopped the recruiting process after reaching saturation, i.e., when no new information was being generated by new observations and interviews.

Our final sample consisted of 19 participants (17 self-identifying as men, 2 as women) with a mean age of 32 years (SD = 4.5), representing a diverse range of backgrounds, native languages, and experiences with scientific writing and AI. While the sample resulted to be overwhelmingly male, this reflects the well-documented gender gap in computer science, where women remain unfortunately underrepresented across various roles and research contributions [11, 59]. Table 1 provides an overview of the sample. Our participants included 11 Ph.D. students, 4 postdoc researchers, and 4 professors, with a median experience of 3-5 years. Participants had expertise in various disciplines, such as Human-Computer Interaction, Machine Learning, and Cybersecurity. Native languages included Italian (8), Spanish (4), English (3), Hindi (2), German (1), and Portuguese (1). Except for 4 participants who reported using intelligent writing assistants occasionally, all the other participants stated that their writing assistants frequency of use was weekly (5) or daily (10). English proficiency was generally rated high, ranging from competent (6) to advanced (9) and expert (4). Writing experience was more diverse, with 3 participants rating themselves as developing, 11 as competent, 4 as advanced, and none as expert.

3.2 Procedure

We conducted the study via Zoom's calls. At the beginning, participants were introduced to the study and signed online an informed consent form approved by the Institutional Review Board of our university. We then structured the study in two main phases:

Observation. First, we asked participants to screen-share their computers and perform a writing session as they normally would for 30 minutes. Session duration was defined taking into account prior studies on human-AI writing collaboration [64]. To ensure participants completed the study in their natural environment and followed their usual workflow, we asked them to work on one of their ongoing scientific writing projects, be it a research paper, a grant proposal, or another relevant document. Although the lack of a specific writing task for the participant may limit detailed analysis of individual writing phases, it allowed us to capture a broad range of activities within the iterative scientific writing process, providing valuable insights into diverse writing practices and tool interactions, as detailed in Section 4.3. Participants were asked to use a single screen in the session, so that everything could be captured through screen sharing.

Interview. After the observation phase, a semi-structured interview was conducted to uncover participants' perspectives on the observed sessions and their use of intelligent writing assistants for scientific writing in general. The interview was audio recorded for later analysis.

To facilitate further analysis, all study sessions were recorded following participants' explicit consent. Observations were video-recorded, while interviews were audio-recorded. The study took approximately 1 hour per participant (M = 58 minutes, SD = 5 minutes).

3.3 Measurements

During the observation phase, we recorded the type of writing task performed by the participants, the writing assistants they used, and all the interactions they had with these tools. Specifically, we defined an interaction as a distinct sequence of actions or communications between the participant and the used writing assistant, aimed at achieving a specific goal or outcome. This includes the steps taken by the participant to initiate the interaction, e.g., asking ChatGPT to generate a paragraph on a given topic, the tool's responses or actions, any adjustments or refinements made by the participant, e.g., asking ChatGPT to add a concept to the initially generated paragraph, and the ultimate result of the interaction, e.g., copying and pasting the refined paragraph into the manuscript.

For each interaction, we tried to capture the motivation for using the writing assistant(s), the outcome of the human-AI collaboration, and any noticeable challenge encountered during the interaction.

During the semi-structured interview phase, we first conducted a retrospective analysis of the writing session. We asked participants to reflect on their general usage and understanding of the assistant(s) they used, as well as their perception of the degree of collaboration they had with these systems. Next, we delved deeper into the personalization of human-AI collaboration in scientific writing with questions such as:

- *How do the writing assistants you are currently using impact your writing experiences?*
- *Do the features of the writing assistants you use and their intended use match the goals you have for your writing?*
- *Have you taken any actions to make the experience better fit your needs? What were your intentional and thoughtful actions, and why did you do so?*
- *Do you find yourself adapting your writing style or process to better align with the capabilities of these tools? If so, in what ways?*

3.4 Data Analysis

We reviewed all the recordings to consolidate the notes taken during the experiment, transcribing the interviews verbatim. Then, we employed a mixed deductive-inductive coding approach [17] on the available data, using the observations as the primary source and the interviews to complement and enrich what we found.

The first author developed an initial code manual based on Kim and Lim's [32] work. Three broad code categories formed the code manual, inspired by the three design-in-use styles discussed in the paper: (i) *teaching*, where users provide useful information to refine an intelligent system; (ii) *resisting*, which involves rejecting or counteracting a system's learning models; and (iii) *repurposing*, where users redefine a system's intended use to better suit their needs. Each main code included sub-codes that characterized Kim and Lim's styles, as shown in Table 2. For example, *onboarding* procedures were associated with teaching, *reorienting* a system to 'break' a loop of personalization was linked to resisting, and *misalignments* between users' habits and the services provided by a system were categorized under repurposing.

Three authors of this paper used the initial code manual to deductively analyze the collected findings and extract initial themes. To

Table 2: The initial code manual, inspired by the work of Kim and Lim [32], was used deductively to analyze the collected findings.

Main codes	Sub-codes
Teaching	Onboarding, implicit feedback, proactive preferences
Resisting	Disengagement, over-personalization, reorienting
Repurposing	Misalignments, unintended use, alternative utility

ensure consistency and reliability, we calculated inter-coder agreement after independently coding a subset of the data, achieving a Cohen's Kappa score of 0.82, which indicates substantial agreement. The initial themes were refined iteratively through regular meetings where the research team reviewed initial coding outcomes and resolved discrepancies. For inductive coding, new themes not encompassed by the initial code manual were incorporated and validated through team discussion. Such an iterative process allowed us to refine and adapt the coding to our specific context, identifying design-in-use styles, challenges, and motivations that are unique to collaborating with intelligent writing assistants in scientific writing.

4 Findings

To present the findings of the study, we begin with an overview of the observed writing sessions, followed by a discussion of the themes that emerged from our analysis. Themes encompass the impacts of intelligent writing assistants on writing processes and skills (Section 4.2), as well as the design-in-use styles and practices adopted by researchers (Section 4.3). These themes are often interconnected, with some reported practices serving as strategies to mitigate the negative effects of writing assistants perceived by our participants. Throughout the rest of the section, themes are highlighted in bold and accompanied by their prevalence, expressed as the number of participants who originated them.

4.1 Overview of the Observations

Figure 1 provides an overview of the intelligent writing assistants used by the participants to support their scientific writing. One of the key findings emerging from the observations is that participants used and combined assistants with significantly varying characteristics, leveraging different AI technologies to support their scientific writing, from general-purpose LLMs to AI-powered grammar checkers and translators. ChatGPT was by far the most commonly used tool (16 participants), with researchers using either the paid (10) or free (6) version. Two participants also used another general-purpose tool powered by an LLM, Claude, although they always used it in support of ChatGPT. A significant number of participants used AI grammar checkers such as Grammarly (8) and Writefull (1). Interestingly, although these tools also offer the capability to (re)generate entire sentences or paragraphs, none of the participants utilized such a generative AI feature, instead delegating this task to general-purpose LLMs. Two participants used tools powered by domain-specific LLMs, Elicit and Consensus, whose primary goal is to help users find relevant papers and obtain answers grounded in the literature. Similarly, two other participants leveraged domain-specific LLMs embedded in VSCode extensions designed to provide specialized writing assistance: one of them developed custom-made extension to support writing exploiting OpenAI APIs, while the other one used an integration with Copilot to receive inline writing suggestions. Finally, a participant used an AI translator (DeepL) when translating sentences in another language.

Overall, 13 participants used a combination of two writing assistants during their sessions (see the *orchestrating* practices described in Section 4.3), while the remaining 6 relied on a single assistant. The most common combination of tools was ChatGPT and Grammarly (6 participants), followed by ChatGPT and Claude (2 participants).

The observed sessions reflected various writing tasks, including drafting a new section or an abstract of a research article (9 participants), continuing (6 participants) or revising a section (2 participants). Additionally, one participant used the session to draft a review, while another wrote a research summary for a website.

4.2 Impact on the Writing Process and Skills

4.2.1 Advantages. Our findings indicate that collaborating with intelligent writing assistants can **positively impact the efficiency of the writing process (N = 13).** Indeed, these tools help researchers complete their writing tasks more quickly by streamlining processes such as summarizing related works (P5, P10, P14, and P19), checking grammar (P3, P4, P7, P9, P17, P18, and P19), or shortening one's work (P1, P3, P15). Furthermore, general purpose LLMs are particularly useful for addressing writer's block (P2, P3, P4, P6, P7, and P14)—a temporary inability to write that can last from minutes to weeks [22]. As P3 explained:

> *"Sometimes the problem is the blank page. With these tools, I can start from quick ideas organized in bullet points and ask them to generate a draft of a paragraph. Although it requires modifications and adjustments on my part, the result is like a prototype that serves as a foundation for the final work."*

This collaborative approach reduces the intimidation of starting from scratch and fosters a more fluid and productive writing experience, simplifying brainstorming and contributing to new ideas.

Besides efficiency, intelligent writing assistants help researchers in **producing high-quality text (N = 8).** By collaborating with these tools, researchers can refine their text, enhancing its clarity and cohesion (P2, P6, P8, P9, P11, P15, and P16). This was particularly highlighted and acknowledged by some of our non-native English-speaking participants, who stated that intelligent writing assistants are essential tools for overcoming language barriers (P2, P3, P6, P9, P13, and P16). P2, for example, now depends on writing assistants to enhance the quality of their English, almost forgetting how challenging this task was before the advent of these tools:

> *"Given that most of the papers and journals are in English, I definitely feel there is a requirement of having a certain level of English on your papers and some good knowledge of vocabulary and all of that, which I think would be a very challenging task, or at least for me, it was."*

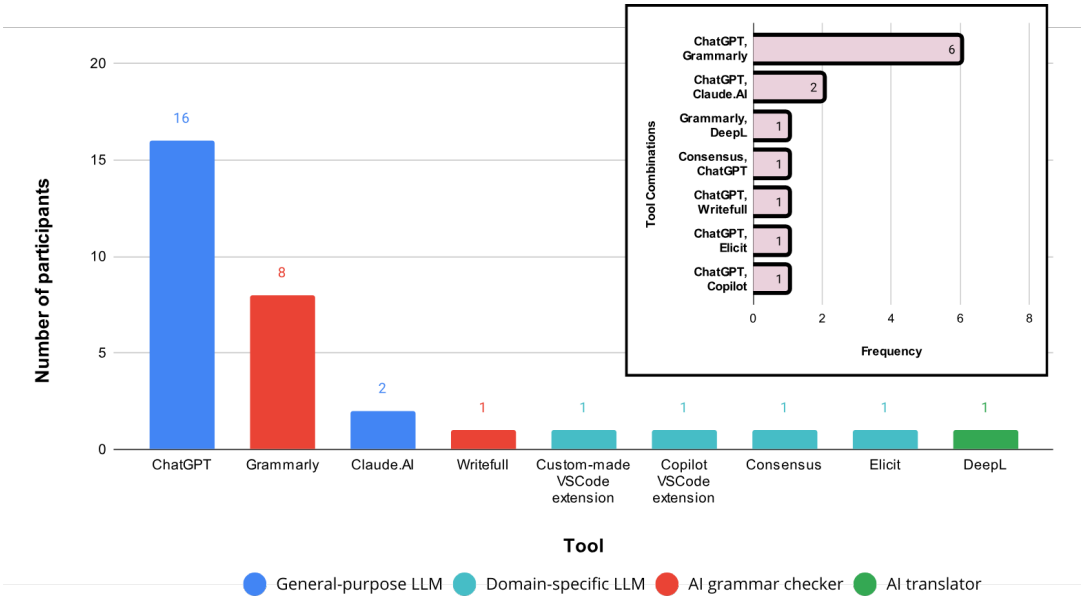


Figure 1: A summary of the writing assistants used by the participants during the observed writing sessions, with a focus on the observed tool combinations, i.e., participants using a combination of two assistants to support their writing.

4.2.2 Workflow Disruptions. Although participants generally viewed their collaboration with intelligent writing assistants positively, our findings indicate that using writing assistants sometimes disrupt the writing process.

We observed that **frequent switching between different browser tabs and tools can fragment the writing process (N = 8)**, making it harder to maintain focus and continuity. Managing multiple tools, browsers, and accounts, in particular, complicates the workflow, creating a chaotic environment that may hinder productivity. Observations revealed suggestion loops where different tools were not aligned with the output, such as Writefull correcting texts generated by ChatGPT (P13). Additionally, we noticed that participants were concerned about the disruption caused by copying and pasting text between various writing assistants (P2, P3, P11, P14, and P15).

Overall, **heavy reliance on writing assistants has been observed to slow progress (N = 9)**, as researchers may find themselves waiting for or managing AI-generated content rather than advancing their writing. For example, excessive suggestions displayed by tools integrated into text editors, e.g., Grammarly in Overleaf, can interrupt the writing flow , as these suggestions often require additional time and attention to incorporate smoothly (P3, P7, P11, P12, P17). As P12 said in the interview:

> *"The only thing that really bothers me, especially with Grammarly integrated into Overleaf, is seeing all these sentences marked as errors simply because the tool dislikes the passive voice."*

Similarly, in explaining how is custom-built VSCode extension works, P3 remarked:

> *"With my extension, I do not have something so pervasive within the page that highlights every single misspelling. This selective approach helps me maintain focus and efficiency."*

In our study, other slowdowns occurred when some participants had difficulty interpreting the AI's suggestions, such as understanding how and why certain changes were made by an assistant (P3, P11). Also the restrictions of the free versions of writing assistants used during the observed writing sessions often impeded progress, forcing users to spend extra time finding workarounds or resorting to manual processes (P5, P12, P14, and P16).

The final workflow disruption that emerged from our study is that **adapting and integrating AI-generated content into one's writing can be challenging (N = 9)**, particularly when it comes to aligning the content with individual writing styles, specific paper requirements, and overall coherence. Our participants were concerned about introducing plagiarism by directly copying content from a writing assistant (P8), and some expressed a general desire to avoid any evidence that AI had been used (P3, P4, P5, and P10). Furthermore, suggestions from some writing assistants interfered with LaTeX commands, particularly when they were used to revise or rephrase text (P11, P17). P11, for example, accepted some suggestions from Grammarly that broke LaTeX commands and then had to spend time restoring those commands in the manuscript. Another major issue that researchers encounter when incorporating suggestions from writing assistants is handling AI-generated hallucinations, e.g., in citation creation (P2 and P8). In our observations, these hallucinations resulted in inaccurate or nonexistent references, thereby increasing the complexity of the integration process and contributing to some of the design-in-use practices described in Section 4.3. Overall, these integration difficulties are

compounded by the lack of context provided by the AI, which can lead to suboptimal results and necessitates additional effort to refine and fit the generated content effectively (P5, P16). For instance, P16 attempted to use DeepL to find the appropriate English phrasing for a concept, only to realize that the tool produced a literal translation due to insufficient context.

4.2.3 Forecasts of Future (Negative) Impacts. As we have outlined in the previous sections, the current impact of intelligent writing assistants on scientific writing is both promising and challenging, according to our participants. Through the behaviors we observed in their writing sessions and the qualitative feedback collected in the interviews, participants delved deeper into these impacts, forecasting future trends and implications.

Our study suggests that computer science researchers consider **AI as a tool to augment human intelligence and creativity rather than replacing them (N = 11)**. This was reflected not only in how participants interacted with their intelligent writing assistants most of the time (see the *resisting* design-in-use style, Section 4.3), but also in what they explicitly stated during the interviews. They argued that prioritizing automation over augmentation, e.g., by making LLMs write entire paragraphs or sections, generates ethical concerns (P1) and, at the same time, would lead to lower-quality writing (P5, P6, P7, P8). In essence, these researchers recognize that AI-generated text still lags behind high-quality human writing, especially in terms of creativity [10].

Another potential negative consequence of interacting with intelligent writing assistants, as identified in our findings, is the **fear of becoming overly reliant on AI and consequently losing writing skills (N = 8)**. By allowing AI to take on tasks traditionally performed by humans, such as structuring arguments or refining language, there is a risk that users may gradually lose the ability to engage deeply with the writing process, e.g., in terms of creativity and critical thinking (P2, P7, P15, and P16), agency (P9 and P12), and efficiency (P6). In parallel, relying too much on writing assistants may result in a diminished enjoyment of the writing task itself (P2, P6, and P13). P6 clearly explained this point in the interview:

> *"The back-and-forth with ChatGPT has been so time-consuming that I've started to wonder if writing it myself wouldn't have been more efficient. I feel like I'm losing some skill by delegating so much to AI, and it's also less enjoyable sometimes. Drawing a parallel with I don't always consult documentation; I enjoy the challenge, even if I know there are faster ways. It's the process that's rewarding."*

4.3 Design-in-use Styles and Practices

In addition to exploring the impacts of using intelligent writing assistants on the writing process and skills, our hybrid coding approach revealed five design-in-use styles, each consisting of multiple practices (Table 3). Similar to Kim and Lin's findings on everyday recommender systems [32], participants engaged in *teaching* information to their writing assistants to enhance co-writing experiences, *resisting* the assistants' outputs to make co-writing less personalized and more authentic, and *repurposing* writing assistants to redefine their potential applications. However, the diversity of the observed practices within these design-in-use styles largely exceeds those observed in the usage of everyday recommender systems, reflecting the more complex nature of collaborating with AI in scientific writing. Furthermore, we observed two novel design-in-use styles and related practices that are specific to the collaboration with AI in scientific writing: *orchestrating* different tools and approaches, and *complying* with the suggestions of a writing assistants.

4.3.1 Teaching. According to Kim and Lim [32], teaching occurs when users have a clear "design objective" and actively shape AI systems to match their personal preferences. With everyday recommender systems—such as those investigated by the authors—teaching primarily involves proactively providing feedback on desired and undesired outcomes. In the context of collaborating with intelligent writing assistants, we identified various related practices associated with this design-in-use style, all originating from interactions with general-purpose and domain-specific LLMs.

In some cases, teaching occurred before the collaboration actually started, as a way to **(re)align the writing assistants—particularly those powered by an LLM—with the user's specific goals and preferences, or even the status of previous writing (N = 3)**. For example, before starting to write, P15 opened a browser tab with Claude, uploaded a PDF with the current version of his research paper, and prompted the LLM with the following: *"Help me continue writing this paper; wait for instruction."* Similarly, P9 started the writing session by uploading a collection of PDFs of research papers to ChatGPT to provide the LLM with information to be analyzed and summarized, while P14 tried to upload several PDFs to Elicit to create a library of related works to be analyzed with the help of the AI. Unfortunately, this initial teaching was far from successful, either due to misunderstandings by the LLM or limitations of the account on the given platform. Claude, for example, responded to P15 with a summary of the uploaded research paper without waiting for further instructions. P14, on the other hand, was unable to upload PDFs to Elicit due to the limitations of his free account plan. In the interview, he acknowledged this by saying:

> *"I actually think the problem is related to the fact that I don't want to pay."*

Teaching also occurred repeatedly within a writing session, with the aim of directly **generating an output, e.g., an individual paragraphs or sentences (N = 4)**. This included asking writing assistants such as Claude and ChatGPT to generate a paragraph by specifying the key points to be included (P5, P6, and P15), or requesting them to generate or rephrase text with a specific tone (P3, P15).

A specific teaching practice that we consistently observed when participants focused on generating an output involved the process of **improving upon initial results by refining the instructions communicated to the writing assistant (N = 7)**. These participants refined prompts in ChatGPT or Claude with more instruction and examples to achieve the expected output (P5, P8, P10, P17), corrected ChatGPT when it provided technical information they knew was wrong (P3), or repeatedly refined and resubmitted text to ChatGPT for further elaboration in order to obtain a final version of the text (P6 and P7). As P8 said in the interview:

Table 3: The design-in-use styles and related practices we observed in the collaborations between participants and writing assistants.

Style	Definition	Technology	Practices	Example
Teaching	Teaching useful information to a LLM to enrich personalized co-writing experiences.	General-purpose LLMs Domain-specific LLMs	*(Re)aligning* *Generating* *Refining* *Implicit teaching*	Uploading PDFs to Claude to provide the LLM with some context and information before starting the writing session (*(Re)aligning*, P15)
Resisting	Resisting learning models in a system to make co-writing experiences less personalized, more authentic, and reliable.	General-purpose LLMs AI grammar checkers	*Less personalization* *Text resistance* *Validating*	Turning off the data training option in ChatGPT to ensure that the sessions remain private (*Less personalization*, P2)
Repurposing	Repurposing a system to redefine its potential usage and [co-writing] experiences.	General-purpose LLMs	*Wrong use* *Developing*	Developing an extension for Visual Studio Code that utilizes GPT to support writing (*Developing*, P3)
Orchestrating	Integrating different tools and approaches to optimize writing processes and achieve desired outcomes efficiently.	General-purpose LLMs Domain-specific LLMs AI grammar checkers	*Pipelining* *Parallelizing* *Assigning*	Using ChatGPT to draft text and Grammarly to refine it as a pipeline (*Pipelining*, P6)
Complying	Adhering to all suggestions made by a writing assistant, including questionable ones, to achieve a perfect score or to anticipate the tool's outcomes.	AI grammar checkers	*Perfect score* *Anticipating*	Anticipating Grammarly detections and modifying personal writing conventions accordingly (*Anticipating*, P12)

> *"I take an iterative approach with the tool [ChatGPT]. I typically start with the most basic form of my question (how I might ask another person) because sometimes it can work with this. If the answer is not focused enough, I might add extra context to my question, add constraints on formatting of the output or what it can include in its response, or an example response."*

Finally, we also observed a form of **implicit teaching (N = 3)**, somewhat resembling the design-in-use practices adopted with recommender systems [32]. P11, in particular, leveraged the communication context of ChatGPT to repeat the same task—checking if a paragraph is well written—without having to specify the prompt each time. P4, instead, reopened an old chat with ChatGPT to leverage the existing context of a previous conversation, aiming to build upon it and maintain a coherent flow with the background information previously established. However, maintaining such a communication context is challenging, as popular tools like ChatGPT and Gemini do not offer a structured way to do so, aside from providing the history of past conversations with the underlying LLM. To overcome this limitation, P13 began saving frequently used prompts in a file and using them whenever necessary.

4.3.2 Resisting. By analyzing how users design their co-writing experiences in practice, Kim and Lim [32] found that some users tend to resist learning models as they are concerned about the potential harms of excessive personalization. In the context of recommender systems, they do this by intentionally slowing the system's learning process, such as by avoiding interactions that could reveal their preferences. According to our observations, resistance is a design-in-use style that characterize the domain of human-AI collaboration in scientific writing as well. The practices that we identified, in particular, were adopted with general-purpose LLMs and AI grammar checkers.

Similarly to the work of Kim and Lim [32], we found resisting practices related to the **need of having less personalized systems (N = 2)**. P2, for example, turned off the data training option in ChatGPT to ensure that the session remain private. Later in the interview, the participant explained:

> *"For me, unless the paper is published, I don't want the latest language model to train on my data. I don't know, maybe it shouldn't be an issue, but I tend to turn off the training option. On ChatGPT, for example, I usually make sure that option is off so that it remains a private session between me and ChatGPT, and my data isn't used for training."*

Another similar resistance practice was employed by P4, who started a new chat with ChatGPT when messages exchanged in the previous chat began to negatively influence the collaboration outcomes. This approach allowed the participant to restart the collaboration with a fresh—less personalized—context.

In the domain of human-AI co-writing, resisting can also involve **refusing to incorporate a text that has been generated or modified after interacting with a writing assistant (text resistance, N = 8)**. In particular, we observed several instances where participants refused to incorporate text generated by ChatGPT or

Grammarly as-is, instead using it as a starting point (P4, P7, P8, P9, P10, P11, P13, and P18). These participants copied and pasted the generated text into their articles but then immediately began modifying and expanding it. During the interview, when asked to reflect on this behavior, P18 said:

> *"I just use the ideas and the words from ChatGPT just to improve my sentences. So, it's giving me ideas, but I don't entirely depend on them.*

As P10 clearly explained in the interview, such resistance is needed to maintain authentic authorship of the text:

> *"I am very conscious of my writing not sounding like it's been generated by an LLM. So, I would never use something it gives me verbatim. I probably wouldn't discuss word choice, transition sentences, or that kind of stuff with it. I haven't been impressed in the past by the way it summarizes and writes conclusions. Its word choices are distinctive, so I want to be careful that it doesn't sound like somebody else."*

In the interview, P3, P8, and P11 expressed similar concerns, with P3 saying that *"there are certain words or phrases that seem very AI-generated, like 'delve into,' which has been featured in several humorous memes."*

In addition to text, we noticed a specific type of resistance practice when general purpose writing assistants were employed to gather (technical) information through questioning. Most of the time, these interactions involved **cross-referencing and validating AI-generated outputs with external sources to ensure accuracy and reliability (N = 8)**. P9 and P13 sought confirmation of ChatGPT's responses on a technical issue by conducting a web search. Similarly, P13 sought confirmation in internal documents, while P2 and P8 had to manually verify the existence of papers suggested by ChatGPT and Claude through a web search. In general, all these participants emphasized that familiarity with the topic is crucial in mediating the AI's suggestions, as P9 said in the interview:

> *"The only way you can balance the use of ChatGPT and what you're writing is if you know about the topic."*

Likewise, P13 explained:

> *"It's part of our job as a researcher to just to be aware of the possible errors that AI can make. Everything you read, you need to be critic and you need to go to the sources and seek as much as sources you can."*

The significance of this practice was demonstrated by participants who consciously refused to use general purpose assistants like ChatGPT for seeking evidence and citations (P11, P12, and P15).

4.3.3 Repurposing. Repurposing is a design-in-use style where users adapt a system's original design to serve new, unintended purposes [32]. This happens when there are misalignments between users' goals and the intended uses of an AI-based system, leading users to creatively redesign their experiences with that system.

In our study, this design-in-use style was not prevalent. The only participant who repurposed a tool with an unintended but potentially **wrong use (N = 1)** was P8, who used a general-purpose writing assistant, Claude, to find citations. Unfortunately, this practice is risky, as these tools frequently hallucinate by fabricating false citations [8]: the participant ended up searching for the suggested papers online, only to discover that the references did not exist.

Overall, we attribute this lack of repurposing practices to the abundance of AI-enabled tools available for use as writing assistants, as well as the general-purpose nature of the most commonly observed tools, such as ChatGPT. Indeed, we observed that rather than repurposing a given system, our participants tended to use multiple tools to achieve their goals (see the *orchestrating* style).

Interestingly, P3 refused to adopt such an approach, recognizing the drawbacks of using multiple tools discussed in Section 4.1. Instead, the participant employed a repurposing strategy by **developing a custom-made support for writing (N = 1)**. This support—shown in Figure 2—is an extension for Visual Studio Code that utilizes OpenAI GPT's APIs. As observed in the writing session, the extension allows the researcher to write content and then apply various modifications such as grammar correction, text manipulation, summarization, or paraphrasing (Figure 2, A). It supports multiple languages and tones and offers flexibility in selecting the underlying LLM from the OpenAI family. Furthermore, by clicking on the 'See Difference' button, the researcher can compare the newly generated text with the old one (Figure 2, B). As explained by the participant later in the interview:

> *"The tool is selective and integrated, meaning that I can select the specific text I want to work on—down to a single sentence—directly within my writing document. This means I don't have to copy and paste text into another platform, which is very annoying."*

4.3.4 Orchestrating. We observed a majority of participants who integrated different tools and approaches to optimize writing processes. Given the iterative and non-linear nature of scientific writing, which involves multiple phases and levels of abstraction—from ideation to evidence gathering—and considering the diverse characteristics of the contemporary writing assistants, these participants demonstrated remarkable skill in effectively orchestrating these tools, including general-purpose and domain-specific LLMs, as well as AI grammar checkers.

Figure 3 illustrates the complex orchestration of various AI writing tools and assistants observed in our study. Several users demonstrated this design-in-use style by **employing different tools in a pipeline, sequentially feeding the output of one tool into another (N = 5)**. The most common pipelining practice that we observed, in particular, was using ChatGPT to refine an initial version of a draft text, e.g., a sentence or a paragraph, and then using Grammarly to check the newly generated text (P6, P7, P14, and P19). P7 clearly explained this practice later in the interview:

> *"So, I use them [ChatGPT and Grammarly] like a pipeline. First, I write a rough draft without focusing too much on style or form. Then, I use ChatGPT to improve the structure, explore ideas more deeply, and even find better ways to connect my thoughts. It's definitely helpful for refining the overall form. Finally, I run it through Grammarly as a final check, because I'm a bit particular about details."*

Similarly, P14 said:

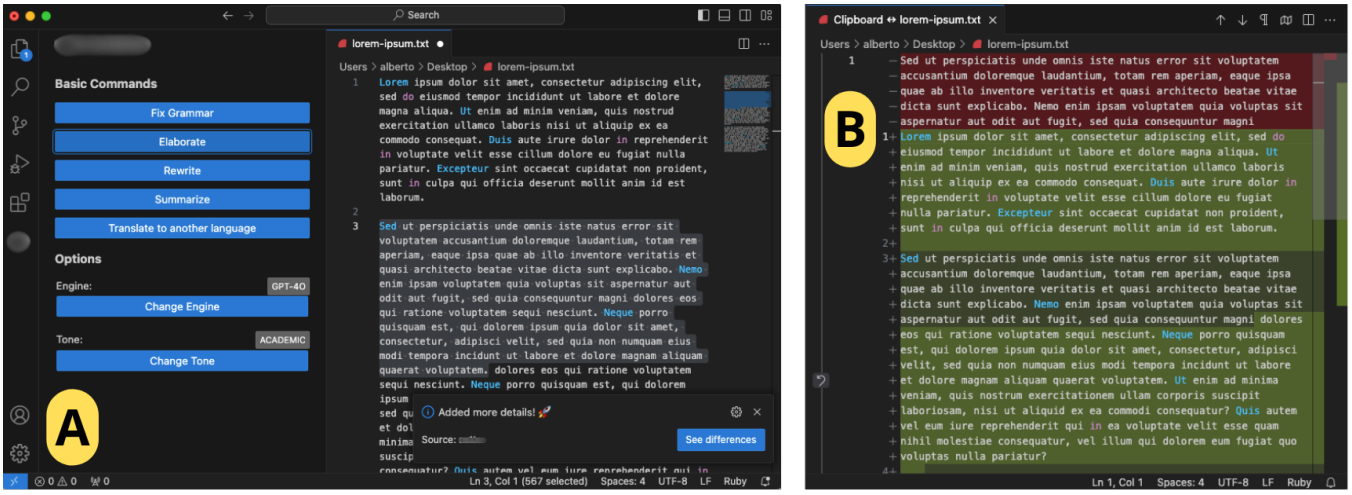


Figure 2: The Visual Studio Code extension developed by one of our participant to avoid using multiple writing assistants. The extension leverages OpenAI GPT's APIs to support writing and perform tasks like grammar correction, text manipulation, summarization, and paraphrasing.

> *"I use ChatGPT to generate some text when I just don't know where to start; then I edit it in my own way, and finally I use Grammarly to check that I have written it correctly."*

P18, on the other hand, consistently used a pipeline that involved initial drafting in Grammarly, refining the text in ChatGPT, and then performing a final check in Grammarly again.

Another common orchestrating design-in-use practice that we observed, which we called 'parallelizing', involved **using multiple tools or approaches simultaneously to make comparisons, allowing for exploration of different alternatives and fostering greater self-judgment (N = 5)**. This practice involved asking ChatGPT (P11) or Claude (P8) to regenerate an output for comparison, pasting different versions of an LLM-generated text into the article to compare them with the original sentence (P11, P13, and P18), and using different tools to perform the same task, e.g., using both Consensus and ChatGPT to get inspiration for a given topic (P2). In all these cases, the participants 'borrowed' from the various LLM-generated outputs, creating a final result that combined their own ideas and writing with the suggestions they considered most useful.

Through the observed pipelining and parallelizing practices, participants also demonstrated the ability to **allocate specific tasks to different intelligent writing assistants based on their capabilities (N = 8)**. While some allocations may seem obvious—like using ChatGPT for text generation and Grammarly for grammar checking (N = 5, P6, P7, P11, P14, and P18)—others involved more nuanced decisions. P17 used the Copilot VSCode extension for inline recommendations and ChatGPT to analyze text retrospectively when additional context was needed. P14 repeated the same task with another tool, ChatGPT, when the output obtained with the first tool, Elicit, was not satisfactory. Finally, P15 used similar assistants for different tasks, e.g., Claude for text generation and ChatGPT for seeking advice on how to format LaTeX tables. As the participant explained in the interview, these different usages were deliberate:

> *"When you ask ChatGPT to write parts of a document, it can sometimes seem overly elaborate, as if aiming for a Nobel Prize. On the other hand, Claude writes in a more neutral tone, although ChatGPT's reasoning is better. So I use Claude for supporting my writing, but when I need something more related to programming tasks—like asking how to do something in LaTeX—I tend to switch to ChatGPT."*

Overall, while orchestrating workflows is not exclusive to intelligent writing assistants, the integration of AI tools like ChatGPT has streamlined processes that previously relied on multiple specialized tools with limited capabilities. At the same time, our findings show that these advancements enabled the emergence of new practices, such as dynamic switching between AI functions and iterative refinement, which reflect the evolving role of AI in enhancing writing workflows.

4.3.5 Complying. We observed a form of *complying* with AI suggestions that involved adhering to all the recommendations received from a writing assistant, including questionable ones (N = 3). Such a design-in-use style characterized the use of a specific AI grammar checker, i.e., Grammarly. The first practice to implement it was adopted in **an attempt to achieve a perfect score (N = 2).** The platform, indeed, dynamically evaluate the quality of the submitted text, providing the user with an overall score (ranging from 0 to 100) that reflects the writing quality in the document, which can be improved by following Grammarly's recommendations. As P11 explained in the interview:

> *"With Grammarly, I find myself wanting to make everything, let's say, compliant with Grammarly, you know? It's more of a personal thing, being so meticulous and*

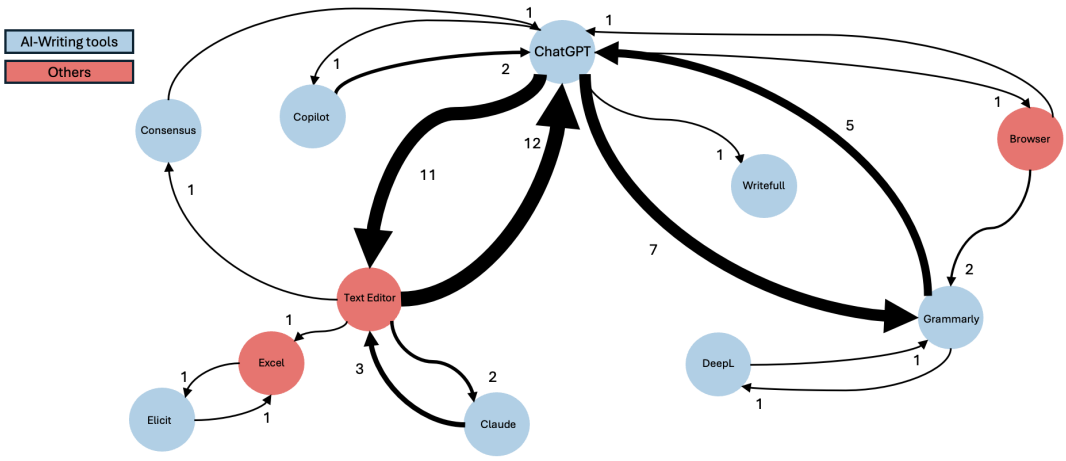


Figure 3: Diagram illustrating the orchestration of various AI writing assistants in the scientific writing process. Light-blue nodes represent AI tools, while red nodes represent non-AI applications. Numbers and edge thickness represent the absolute number of tool transitions observed in our study.

> *thinking that when everything is aligned with Grammarly, it works better. So, I don't know if you noticed, but for example, it suggested some changes to the figure references...it told me to leave a space between the colon and the text. I knew it didn't make sense, but I still did it because I like making sure that everything is perfectly in line with its suggestions, even when I know they're not necessary."*

Adhering to the recommendations of a writing assistant like Grammarly is also needed to **anticipate the AI's outcomes and prevent additional disruptions to the writing flow (N = 2)**. According to P12, in particular:

> *"If you dismiss a suggestion [on Grammarly], it still reappears the next time, so I started saying, 'Alright, fine, if you want me to write it this way, I'll write it this way, at least it won't bother me anymore'."*

As such, the participant felt he was adapting himself to the tool, for example, when writing table and image references that he knew Grammarly would not detect as errors.

In general, our findings link the complying style exclusively with AI grammar checkers, as we did not observe instances of participants using AI-generated text as-is (see the resisting practices described above). We argue that one reason for such behaviors and differences in compliance may lie in the familiarity of our participants with AI-based systems and LLMs, most of whom explicitly demonstrated awareness of issues such as potential plagiarism and hallucinations. However, we recognize that less experienced users, such as those unfamiliar with AI, may be at a higher risk of over-relying on and passively complying with the outputs of other intelligent writing assistants, including those based on general-purpose and domain-specific LLMs. This highlights the need for further research on how diverse user groups navigate the balance between tool guidance and their own writing autonomy.

5 Discussion

5.1 Design Implications

In this paper, we adopted a design-in-use lens to investigate how researchers navigate the landscape of intelligent writing assistants to collaborate with AI in their scientific writing. In particular, we conducted ethnographic observations to investigate how researchers engaged in their daily work activities, aiming to gather ecologically valid findings that could complement previous research in the field.

Echoing previous research on human-AI interaction [32], we observed a trend where users frequently take an active role in shaping their desired experiences, rather than passively accepting the predetermined nature of the technology. Despite the growing number of specialized writing assistants for scientific writing proposed in the literature, such as VISAR [64] and Sparks [21], and those available in the market, such as Elicit [60] and Consensus [12], our participants opted for an ensemble of diverse technologies, orchestrating the use of different tools and assigning distinct to them based on their perceived capabilities.

Understanding these behaviors is crucial for informing new research directions in human-AI co-writing for scientific writing and designing better assistants. Some of the findings in this work, such as the disruption caused by using multiple tools in the writing process, may suggest that a more integrated approach is needed, e.g., to combine the advantages of general-purpose LLMs with those of domain-specific solutions. Nevertheless, participants demonstrated that in certain cases, such as through the parallelizing and pipelining practices, the usage of multiple tools and approaches is beneficial. To address these trade-offs, we now present a series of design implications (Table 4) based on the most prominent design-in-use

styles observed in our study—teaching, resisting, orchestrating, and complying—along with their associated practices. Although not exhaustive, these examples offer a glimpse into how our findings may be used to identify opportunities for better supporting researchers' workflows and goals.

5.1.1 Support for Teaching. Kim and Lim [32] found that most users resist being passive consumers of everyday AI systems, instead desiring to actively influence the system to align with their preferences and needs. Given the non-linear and complex nature of scientific writing, it is not surprising that many of our participants engaged in teaching practices, actively collaborating with writing assistants to enhance their co-writing experiences. However, teaching information to writing assistants, especially general-purpose LLMs, is not always straightforward, particularly as these assistants are used at various stages of the writing process.

We argue that supporting this design-in-use style would first require having writing assistants that streamline and simplify the ***initial setup*** process, facilitating better initial context-setting. Users, for example, could be prompted to define their goals and preferences in detail through guided setup processes, where they are asked to provide specific instructions or examples to align the AI's outputs with their expectations from the outset.

Reflecting on what happens during the interaction, instead, another possibility to support teaching is to allow users to create and manage ***prompt libraries*** where they can save frequently used prompts and settings. This would make it easier to maintain consistency and quickly reapply successful teaching strategies across different tasks. These libraries could include options for categorization and quick access to frequently used contexts, representing the user's preferences.

5.1.2 Support for AI-Resistance. AI resistance and skepticism are phenomena that characterize interactions with intelligent systems across various domains, from the use of recommender systems for entertainment purposes [32] to applications in the medical field [40]. As prior work demonstrated, writing is not exempt from this resistance. Our observations revealed various resisting practices through which researchers sought co-writing experiences that were less personalized, more authentic, and more reliable. We view these design-in-use practices as an opportunity to design future writing assistants that could engage even those researchers who are more reluctant to use AI for writing, helping them leverage the advantages of co-writing with such tools.

Providing users with the possibility of working in ***isolated environments*** with nuanced privacy settings and refreshed contexts, for example, could allow them to control how their data is used, especially regarding the training of AI models. These controls should be easily accessible and offer clear explanations of the implications of enabling or disabling data sharing, thereby empowering users to protect their work during sensitive writing tasks.

Our findings also highlight the need for writing assistants with ***integrated validation features*** by incorporating tools that allow users to easily cross-reference AI-generated information with trusted external sources. This would be especially important for general-purpose LLMs and could include built-in search functions, citation verification tools, or links to authoritative databases, ensuring the accuracy and reliability of the information used in scientific writing.

Overall, these implications are undoubtedly shaped by the ecosystem—one of the key aspects described by Lee et al. in their design space [35]—in which writing assistants operate, as implementing them would require addressing ethical and privacy challenges. In this context, we consider regulation and policies fundamental to establishing these design opportunities.

5.1.3 Support for Orchestration Rather Than Repurposing. Rather than relying on a single writing assistant and repurposing it for different uses, we observed multiple instances of researchers using various tools and approaches to achieve their writing goals. While this approach has potential drawbacks and might seem to reflect limitations in the exploited tools, we actually observed that some of our participants exhibited a clear preference for having multiple tools at their disposal, as demonstrated by their orchestrating practices. In other words, our findings suggest design implications that support these orchestration practices, rather than aiming to develop a one-size-fits-all assistant.

Based on our observations, we envision the potential of offering researchers the capability to define ***reusable pipelines***. This would empower them to customize their workflows within a single platform or across multiple platforms, including the option to establish reusable sequences of actions (e.g., draft, refine, and check) or the ability to save and reuse specific tool chains where outputs from one tool can be seamlessly integrated into another.

Another example of support for orchestration is the development of ***multi-model interfaces***, which would allow users to run and compare outputs from multiple learning models side by side. For example, integrating multiple LLMs in a single interface where users can compare and merge the results could enhance decision-making and creativity. In that sense, we see a parallel to the concept of parallel prototyping in user experience design [41]. Through such an approach, designers can simultaneously and independently develop multiple design solutions, evaluating and comparing them to identify the strengths and weaknesses of each approach. This enables them to rapidly and efficiently explore a range of design options, leading to more informed decisions and ultimately enhancing the overall user experience. This implication echoes findings from prior work in human-AI co-writing [13], which showed that writers prefer selecting outputs from multiple suggestions generated by LLMs using a technique known as 'diegetic prompting'—integrating prompts within the narrative—rather than refining the output through iterative, non-diegetic prompts, which involve giving explicit instructions to the LLM.

5.1.4 Support for (Avoiding) Compliance. Previous research has highlighted several drawbacks of over-relying on AI-generated content in writing. Overrelying on LLMs, for example, can hinder critical thinking and creativity [62], as well as leading to the spread of misinformation and ethical challenges, including plagiarism and privacy violations [54]. In our work, we observed a form of overreliance where researchers accepted all the suggestions of a writing assistant, not necessarily because they trusted them, but rather to satisfy the tool, achieve a perfect score, or avoid annoying interruptions.

Table 4: Examples of design implications to support the design-in-use styles observed in our study.

Design implication	Description	Design-in-use style(s)
Facilitating initial setup	Facilitate better initial context-setting by guiding users through a setup process to define their goals and preferences.	Teaching
Supporting prompt libraries	Create libraries for users to save and categorize frequently used prompts and settings.	Teaching
Providing isolated environments	Provide users with isolated environments and nuanced privacy settings to control data usage and context.	Resisting
Integrating validation features	Incorporate tools for users to cross-reference AI-generated information with trusted sources.	Resisting
Enabling reusable pipelines	Enable users to customize workflows within or across platforms by setting up reusable action sequences.	Orchestrating
Developing multi-model interfaces	Develop interfaces for users to run and compare outputs from multiple learning models side-by-side.	Orchestrating
Providing context-aware assistance	Prioritize suggestions based on their impact on writing flow, deferring non-critical issues to later stages to minimize interruptions.	Complying
Supporting writing tutoring	Promote active learning to help users develop their writing skills over time.	Complying

We consider this "user-overfitting" to LLMs potentially problematic as well, as it may inadvertently compromise the quality and authenticity of the text and contribute to disruptions in the writing flow. For example, writing scores might serve as a target, potentially leading users to prioritize achieving a high score over making sound, independent decisions—a phenomenon known as Goodhart's Law in economics [55]. Furthermore, users may unconsciously align their behavior to meet the expectation of achieving higher scores—even if it means accepting suboptimal or incorrect recommendations—demonstrating a form of social conformity bias [5]. Finally, we noticed that our participants explicitly said that complying contributes to create disruptions to the writing flow. Dang et al. [13] discussed this problem in their work, highlighting that moving from diegetic writing to non-diegetic prompting—i.e., explicit instructions to refine an output—demands a cognitive shift that forces writers to move away from directly shaping their narrative and instead focus on how to best guide the system. This implies that writers must constantly assess AI suggestions, which can disrupt their typical writing routines and increase distractions [7].

As such, our findings highlight the need to offer users alternatives to this design-in-use approach. Based on our participants' feedback and behavior, we see value in tools that implement ***context-aware assistance***, prioritizing suggestions according to their potential impact on the writing flow and the current writing phase. For instance, minor issues like spacing or stylistic preferences could be bundled for later review, allowing users to maintain focus on content generation without unnecessary interruptions. This might mean implementing context-aware features that reduce such interruptions by deferring non-critical suggestions until a more suitable time in the writing process, e.g., during editing rather than drafting. This aligns with existing theories of writing, which state that focusing on grammar and style during the early stages of writing can interfere with the cognitive processes involved in composition. The Cognitive Process Theory of Writing by Flower and Hayes [18], in particular, suggests that postponing grammar and style concerns to a later revision stage may improve the outcome and reduce cognitive load.

The final design implication we extracted from our findings is based on the participants' concern that overreliance on AI could lead to a decline in their writing skills. Specifically, we suggest that designers of writing assistants prioritize ***writing tutoring***, incorporating opportunities for active learning and skill development. This involves creating features that not only suggest improvements in a passive way but also explain the reasoning behind them, helping users develop their writing skills over time.

5.2 Ethical Implications

Our research shows that AI-powered writing assistants, particularly commercial ones, are becoming increasingly integrated into scientific writing processes. Therefore, addressing ethical implications is fundamental for promoting responsible and effective human-AI collaboration in this field [1]. As previous research has clearly demonstrated, the overuse of AI-generated texts might be negatively perceived by writers, as it undermines their sense of agency and ownership [27, 44]. Other common concerns include plagiarism [34], the generation of factually incorrect information [45], as well as the generation of stereotyped text [21].

We, of course, acknowledge the importance of addressing all these ethical concerns and advocate for the development of comprehensive regulations and policies that may assist designers of AI-driven writing assistants in creating technologies that not only enhance productivity and support researchers but also uphold the

highest standards of privacy, fairness, transparency, and user autonomy. Yet, we also stand with the view of Hoque et al. [27], who posit that *"LLMs are here to stay, [...] we should just learn how to best leverage them."* Although future research is needed to confirm or challenge these findings across different academic communities, our study generally observed a conscious use of writing assistants, even those based on LLMs. Participants consistently approached AI-generated text critically, ensuring that the content was thoughtfully reviewed and adapted to meet their specific needs and standards. Moreover, most of the design implications we identified—from those related to resting practices to those related to complying practices—actually support a vision where intelligent writing assistants serve as supertools [52] that augment researchers' capabilities rather than replacing them.

5.3 Limitations

The primary limitations of our study stem from the sample of participants we observed and interviewed. Despite our efforts to recruit a diverse range of academic roles and years of experience, the sample was ultimately skewed towards early-career researchers, including Ph.D. students. Yet, early-career researchers represent an informative category as they are still developing writing skills and are typically involved in writing duties under the supervision of their mentor [29].

Our sample was also predominantly composed of men. Although this gender imbalance unfortunately reflects the well-documented gender gap in computer science [11, 59], we acknowledge that it may introduce potential biases in our findings. We recognize the importance of tackling this issue and believe that greater efforts are needed to improve gender diversity in our field. In addition, the sample was intentionally composed exclusively of computer science researchers, which means our findings may not generalize to other research communities due to potential biases associated with their technical experience and familiarity with AI tools. Researchers with less technical expertise and AI knowledge, for example, may be less critical when interacting with writing assistants and less aware of the risks associated with integrating AI into scientific writing workflows. At the same time, prior studies show that non-technical users often use prompt-based systems opportunistically and struggle to achieve systematic progress [63]. This suggests that they may adopt alternative design-in-use styles to meet their goals. Replication studies in other research communities will help to determine the generalizability of our findings, potentially leading to the development of guidelines for responsible AI writing assistants tailored to different researchers' background.

Finally, technical limitations of the study, including the requirement for participants to use a single screen during the writing session and the limited timeframe of the session, may have influenced participants' natural workflows and restricted the depth of interaction with AI tools, potentially impacting the generalizability of observed behaviors.

6 Conclusions

This paper explored how computer science researchers collaborate with intelligent writing assistants to support their scientific writing. Through observations and retrospective interviews conducted with 19 participants, we shed light on five design-in-use styles and 14 related practices that researchers use to shape their own co-writing experiences. These practices represent both a means to overcome existing challenges in interacting with AI and an opportunity to inform the development of more effective writing assistants that can further support researchers in all the iterative and nonlinear process of scientific writing.

References

[1] L. Acion, M. Rajngewerc, and G. et al. Randall. 2023. Generative AI poses ethical challenges for open science. *Nature Human Behaviour* 7 (2023), 1800–1801. https://doi.org/10.1038/s41562-023-01740-4

[2] Tazin Afrin and Diane Litman. 2023. Predicting Desirable Revisions of Evidence and Reasoning in Argumentative Writing. In *Findings of the Association for Computational Linguistics: EACL 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2550–2561. https://doi.org/10.18653/v1/2023.findings-eacl.193

[3] Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. 2024. LitLLM: A Toolkit for Scientific Literature Review. *arXiv preprint arXiv:2402.01788* (2024).

[4] Anthropic. 2024. Claude AI. https://www.anthropic.com/claude Accessed: 2024-03-13.

[5] Solomon Asch. 1956. Studies of Independence and Conformity: I. A Minority of One Against a Unanimous Majority. *Psychological Monographs: General and Applied* 70 (01 1956). https://doi.org/10.1037/h0093718

[6] Christopher A Bail. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences* 121, 21 (2024), e2314021121.

[7] 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)*. Association for Computing Machinery, New York, NY, USA, 436–452. https://doi.org/10.1145/3581641.3584060

[8] M. Bhattacharyya, V. M. Miller, D. Bhattacharyya, and L. E. Miller. 2023. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus* 15, 5 (May 2023), e39238. https://doi.org/10.7759/cureus.39238

[9] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv:2303.04226 [cs.AI]

[10] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or Artifice? Large Language Models and the False Promise ofCreativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA.

[11] J. McGrath Cohoon, Sergey Nigai, and Joseph "Jofish" Kaye. 2011. Gender and computing conference papers. *Commun. ACM* 54, 8 (Aug. 2011), 72–80. https://doi.org/10.1145/1978542.1978561

[12] Consensus. 2024. Consensus: AI-powered Academic Search Engine. https://consensus.app/ Accessed: March 13, 2024.

[13] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 408, 17 pages. https://doi.org/10.1145/3544548.3580969

[14] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel P. 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* (Honolulu, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA.

[15] Benedikt Fecher, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. Friend or foe? Exploring the implications of large language models on the science system. *Ai & Society* (2023), 1–13.

[16] K. J. Kevin Feng, Xander Koo, Lawrence Tan, Amy Bruckman, David W. McDonald, and Amy X. Zhang. 2024. Mapping the Design Space of Teachable Social Media Feed Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 733, 20 pages. https://doi.org/10.1145/3613904.3642120

[17] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (2006), 80–92. https://doi.org/10.1177/160940690600500107

[18] Linda Flower and John R. Hayes. 1981. A Cognitive Process Theory of Writing. *College Composition and Communication* 32, 4 (1981), 365–387. http://www.jstor.

org/stable/356600

[19] Association for Computing Machinery. 2024. ACM Policy on Authorship. https://www.acm.org/publications/policies/acm-policy-on-authorship Accessed: December 4, 2024.

[20] Google Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL]

[21] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533

[22] Frederica Gonçalves, Ana Caraban, Evangelos Karapanos, and Pedro Campos. 2017. What Shall I Write Next? Subliminal and Supraliminal Priming as Triggers for Creative Writing. In *Proceedings of the European Conference on Cognitive Ergonomics* (Umeå, Sweden) *(ECCE '17)*. Association for Computing Machinery, New York, NY, USA, 77–84. https://doi.org/10.1145/3121283.3121294

[23] Grammarly. 2024. Grammarly: Free AI Writing Assistance. https://www.grammarly.com/. Accessed: March 13, 2024.

[24] Nielsen Norman Group. 2022. Remote Contextual Inquiry. https://www.nngroup.com/articles/remote-contextual-inquiry/. Accessed: March 13, 2024.

[25] Dritjon Gruda. 2024. Three ways ChatGPT helps me in my academic writing. *Nature* (2024). https://www.nature.com/articles/d41586-024-01042-3

[26] Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. 2024. IdeaBench: Benchmarking Large Language Models for Research Idea Generation. *arXiv preprint arXiv:2411.02429* (2024).

[27] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA.

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

[29] Hamid R. Jamali, David Nicholas, Anthony Watkinson, Abdullah Abrizah, Blanca Rodríguez-Bravo, Cherifa Boukacem-Zeghmouri, Jie Xu, Tatiana Polezhaeva, Eti Herman, and Marzena Świgon. 2020. Early career researchers and their authorship and peer review beliefs and practices: An international study. *Learned Publishing* 33, 2 (2020), 142–152. https://doi.org/10.1002/leap.1283

[30] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. https://doi.org/10.1145/3571730

[31] Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 935, 17 pages. https://doi.org/10.1145/3613904.3642596

[32] Hankyung Kim and Youn-Kyung Lim. 2023. Investigating How Users Design Everyday Intelligent Systems in Use. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 702–711. https://doi.org/10.1145/3563657.3596039

[33] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: leveraging large language models to support extended metaphor creation for science writing. *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (2023), 115–135.

[34] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do Language Models Plagiarize?. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3637–3647. https://doi.org/10.1145/3543507.3583199

[35] 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, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1054, 35 pages. https://doi.org/10.1145/3613904.3642697

[36] 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)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. https://doi.org/10.1145/3491102.3502030

[37] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2024. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. *arXiv preprint arXiv:2411.05025* (2024).

[38] Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024. Personaflow: Boosting research ideation with llm-simulated expert personas. *arXiv preprint arXiv:2409.12538* (2024).

[39] Tao Long, Dorothy Zhang, Grace Li, Batool Taraif, Samia Menon, Kynnedy Simone Smith, Sitong Wang, Katy Ilonka Gero, and Lydia B. Chilton. 2023. Tweetorial Hooks: Generative AI Tools to Motivate Science on Social Media. In *Proceedings of the 14th Conference on Computational Creativity (ICCC '23)*. Association for Computational Creativity.

[40] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to Medical Artificial Intelligence. *Journal of Consumer Research* 46, 4 (05 2019), 629–650. https://doi.org/10.1093/jcr/ucz013

[41] Aran Lunzer and Kasper Hornbæk. 2008. Subjunctive interfaces: Extending applications to support parallel setup, viewing and control of alternative scenarios. *ACM Trans. Comput.-Hum. Interact.* 14, 4, Article 17 (jan 2008), 44 pages. https://doi.org/10.1145/1314683.1314685

[42] Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated Social Science: Language Models as Scientist and Subjects. *SSRN Electronic Journal* (2024).

[43] Nikolas Martelaro. 2022. Exploring the Future of Remote User Research. arXiv:2208.03349 [cs.HC] https://arxiv.org/abs/2208.03349

[44] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–34.

[45] Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour* 7, 11 (01 11 2023), 1830–1832. https://doi.org/10.1038/s41562-023-01744-0

[46] Thomas P Moran. 2002. Everyday adaptive design. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (London, England) *(DIS '02)*. Association for Computing Machinery, New York, NY, USA, 13–14. https://doi.org/10.1145/778712.778715

[47] Meredith Ringel Morris. 2023. Scientists' Perspectives on the Potential for Generative AI in their Fields. *arXiv preprint arXiv:2304.01420* (2023).

[48] Ref n Write. 2024. Academic Phrasebook - A Guide for Writing Research Papers. https://www.ref-n-write.com/academic-phrasebook/. Accessed: March 13, 2024.

[49] Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2024. Acceleron: A Tool to Accelerate Research Ideation. *arXiv preprint arXiv:2403.04382* (2024).

[50] Marissa Radensky, Daniel S Weld, Joseph Chee Chang, Pao Siangliulue, and Jonathan Bragg. 2024. Let's Get to the Point: LLM-Supported Planning, Drafting, and Revising of Research-Paper Blog Posts. *arXiv preprint arXiv:2406.10370* (2024).

[51] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 3 (2023), 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

[52] Ben Shneiderman. 2022. *Human-Centered AI.* Oxford University Press, Oxford, United Kingdom.

[53] Smodin. 2024. Smodin: Multi-lingual Writing Assistance. https://smodin.io. Accessed: March 13, 2024.

[54] Anna Strasser. 2024. Chapter 10 - Pitfalls (and advantages) of sophisticated large language models. In *Ethics in Online AI-based Systems*, Santi Caballé, Joan Casas-Roma, and Jordi Conesa (Eds.). Academic Press, 195–210. https://doi.org/10.1016/B978-0-443-18851-0.00007-X

[55] Marilyn Strathern. 1997. 'Improving ratings': audit in the British University system. *European Review* 5, 3 (July 1997), 305–321. https://doi.org/10.1002/(sici)1234-981x(199707)5:3<305::aid-euro184>3.0.co;2-4

[56] Lu Sun, Stone Tao, Junjie Hu, and Steven P Dow. 2024. MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–32.

[57] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. *arXiv preprint arXiv:2305.15041* (2023).

[58] Jiyao Wang, Haolong Hu, Zuyuan Wang, Song Yan, Youyu Sheng, and Dengbo He. 2024. Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 12, 18 pages. https://doi.org/10.1145/3613904.3641917

[59] Samuel F. Way, Daniel B. Larremore, and Aaron Clauset. 2016. Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1169–1179. https://doi.org/10.1145/2872427.2883073

[60] Sharon Whitfield and Melissa A. Hofmann. 2023. Elicit: AI literature review research assistant. *Public Services Quarterly* 19, 3 (2023), 201–207. https://doi.org/10.1080/15228959.2023.2224125

[61] Writefull. 2024. Academic writing is hard. Writefull's AI helps you write, paraphrase, copyedit, and more. https://www.writefull.com/. Accessed: March 13, 2024.

[62] Yuhan Wu. 2024. Exploring the Influence of Large Language Models (LLMs) on English Learners and Their Teachers. *Journal of Education, Humanities and Social Sciences* 27 (Mar. 2024), 530–535. https://doi.org/10.54097/zghke663

[63] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[64] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 5, 30 pages. https://doi.org/10.1145/3586183.3606800