A Design Space for Intelligent and Interactive Writing Assistants

Mina Lee[1], Katy Ilonka Gero[1], John Joon Young Chung[1], Simon Buckingham Shum[2], Vipul Raheja[2]
Hua Shen[2], Subhashini Venugopalan[2], Thiemo Wambsganss[2], David Zhou[2], Emad A. Alghamdi[3]
Tal August[3], Avinash Bhat[3], Madiha Zahrah Choksi[3], Senjuti Dutta[3], Jin L.C. Guo[3], Md Naimul
Hoque[3], Yewon Kim[3], Simon Knight[3], Seyed Parsa Neshaei[3], Agnia Sergeyuk[3], Antonette Shibani[3],
Disha Shrivastava[3], Lila Shroff[3], Jessi Stark[3], Sarah Sterman[3], Sitong Wang[3], Antoine Bosselut[4]
Daniel Buschek[4], Joseph Chee Chang[4], Sherol Chen[4], Max Kreminski[4], Joonsuk Park[4]
Roy Pea[4], Eugenia H. Rho[4], Shannon Zejiang Shen[4], Pao Siangliulue[4*]

ABSTRACT

In our era of rapid technological advancement, the research landscape for writing assistants has become increasingly fragmented across various research communities. We seek to address this challenge by proposing a *design space* as a structured way to examine and explore the multidimensional space of intelligent and interactive writing assistants. Through a large community collaboration, we explore five aspects of writing assistants: task, user, technology, interaction, and ecosystem. Within each aspect, we define dimensions (i.e., fundamental components of an aspect) and codes (i.e., potential options for each dimension) by systematically reviewing 115 papers. Our design space aims to offer researchers and designers a practical tool to navigate, comprehend, and compare the various possibilities of writing assistants, and aid in the envisioning and design of new writing assistants.

*Corresponding author: Mina Lee at the University of Chicago (mnlee@uchicago.edu) and Microsoft Research (minal@microsoft.com). We denote each author's self-assigned role with the following superscripts: 1 for project leads, 2 for team leads (alphabetical), 3 for team members (alphabetical), and 4 for advisors (alphabetical). Please see Appendix A for the full author list with their roles, affiliations, and contributions.

1 INTRODUCTION

A *writing assistant* is a computational system that assists users with improving the quality and effectiveness of their writing, from grammar and spelling checks to idea generation, text restructuring, and stylistic improvement. In our current era of rapid technological advancement, however, the research landscape for writing assistants is becoming increasingly fragmented across various communities. While numerous writing assistants have emerged in recent years, quite disparate research communities like Natural Language Processing (NLP), Human-Computer Interaction (HCI), and Computational Social Science (CSS) study writing assistants with different emphases, such as model performance, user interaction, and social phenomena. There are even more specific areas like creativity support, second language acquisition, and disability studies, each of which may struggle to stay up to date with work happening across other communities. This splintering poses a significant challenge for researchers and designers seeking a holistic view, making it essential to bridge these gaps for effectively navigating the complexities of sociotechnical systems [145, 245].

In this paper, we contribute a *design space* to provide a structured way to explore the multidimensional space of intelligent and interactive writing assistants. Design spaces are taxonomies that present critical aspects of a design [32, 33, 166, 178, 214] with three main uses. The first is to establish a shared vocabulary that can help streamline communication and collaboration between researchers, designers, and other stakeholders. The second use case is to provide support for understanding existing designs. They allow one to reflect on why certain design choices were made, and why a given design may succeed in certain ways and fail in others. The third is to support envisioning new designs. By thinking about regions of a design space that may be "empty" (i.e., given a set of dimensions, no design considers all of them), we can think about what a design in that space would look like, and whether it might be worth pursuing.

Through a large community collaboration, we create a design space based on five key *aspects* of writing assistants—task, user, technology, interaction, and ecosystem (Figure 1)—based on the sociotechnical systems perspective. Within each aspect, we identify *dimensions* (i.e., fundamental components of an aspect) and *codes* (i.e., potential options for each dimension) by systematically reviewing 115 papers and employing an iterative coding process. As a result, our design space contains 35 dimensions and 143 codes (Table 1–7), which includes writing contexts (e.g., academic, creative, and journalistic) and users' relationships to a system (e.g., agency, ownership, trust, and privacy) that are closely related to interaction

metaphors (e.g., agent and tool). It also includes diverse learning problems for technology (e.g., classification, regression, and generation) as well as wider ecosystem considerations, such as digital infrastructure (e.g., usability consistency and technical interoperability), norms and rules (e.g., laws and conventions), and change over time (e.g., written artifacts, information environment). We provide two illustrative scenarios that demonstrate how our design space can be utilized for a range of stakeholders (e.g., researchers and policymakers) in Section 5.1.

Our main contribution is the development of the design space by systematically reviewing existing work, identifying important components within each aspect, and connecting those elements under one framework. While we recognize that this design space may not be exhaustive or permanent, we believe it provides researchers and designers with a practical means for exploring, understanding, and comparing the diverse potential of writing assistants beyond their immediate fields. We anticipate that this work will foster dialogue and research in the realm of writing assistants, ultimately aiding in the creation of innovative, ethical writing assistant designs. We publicly release our annotated papers as a living artifact at https://writing-assistant.github.io to promote community involvement in refining and extending the design space.

2 BACKGROUND

Technology has had a significant impact on the way we write. Some technologies long precede the appearance of computers, such as clay tablets known as cuneiform writing [184, 219], and in the 19th and 20th century, typewriter technologies [194], each of which dramatically changed the way we produce written artifacts. Technical advances introduced computer-powered text-entering systems, which further developed into word processors. Word processors allowed users to flexibly edit texts with functions like deletion, copy-paste, and find-and-replace [194]. These systems further evolved to support the writer's *cognitive process*, or the distinctive thought processes the writers go through when composing writings [73, 101]. Examples of such support include providing text analysis [78, 164] and text-planning support [50] (see Appendix C for more examples).

While supporting the cognitive process of writing can help people improve their writing proficiency and efficiency, how to do so consistently and persistently remains a challenge. For example, what kinds of support should we provide to the human writing process and with which interactions? How can we technically enable such support? And when should we "fade" this support as a temporary "scaffolding" [193] to encourage independent writing? One thread of work is understanding the human writing process, such as the cognitive process theory of writing [73, 100]. Another thread has been on extending the "intelligence" of writing assistants, so that these assistants can provide a wider range of support, such as those that require "understanding" of long and complicated texts. For example, for a writing assistant that provides suggestions on filling in between already written story events [62, 86], the system would need to grasp what happened in the preceding and subsequent events in order to coherently connect them. Fortunately, relevant technologies have advanced rapidly within the fields of NLP, where most recently, language models (LMs) [22, 185, 186] and their prompt- and example-based usage showed impressive capabilities in generating coherent text.

Integrating scientific knowledge of human writing and technical advances, researchers designed and built many intelligent and interactive writing assistants in recent years. Some of these aimed to support existing human writing tasks and domains, such as help request writing in professional settings [115], instant message writing in affectionate relationships [135], or tweetorial writing for scientific communication [85]. Researchers also studied how writers interact with these new writing assistants and how they influence human writing. For example, Buschek et al. [27] studied how generative writing suggestions can influence email writing behaviors. Similarly, Lee et al. [148] collected a dataset of human-LM interactions during creative and argumentative writing to analyze how people interact with and get support from these technologies. Some work studied how writers leverage different prompting strategies for LMs [58, 269] and introduced novel interaction paradigms to steer LM outputs, such as visual sketching [45].

As new writing assistants are being introduced at an increasingly fast pace, there are growing concerns that we lack a comprehensive understanding of when and in what ways it is desirable to use these writing assistants in order to avoid potential unforeseen consequences or ethical issues [121, 171]. For instance, in the academic community, there are concerns about students' use of AI during writing and homework [20, 55, 74, 170]. Moreover, while there are many design considerations for these assistants (e.g., related to task, user, interaction, ecosystem), often only a few are considered (e.g., technology), which can lead to a model-centric approach to building systems [132]. Reflecting upon these concerns, some researchers strive to deepen our understanding about intelligent and interactive writing assistants as a whole. For example, Wan et al. [253] studied interaction patterns of recent human-AI writing assistants. Similarly, Gero et al. [82] analyzed existing writing assistants along the dimensions of writing goals and technologies. Another effort involved understanding user perspectives in getting support for their writing processes [86]. However, we are still far from having a comprehensive picture of the landscape. Outside of writing assistants, researchers studied various aspects of creativity support tools with the goal of comprehensive understanding of the space [44, 79]. However, such a study has not yet been conducted specifically on writing assistants.

3 APPROACH

To develop a design space for intelligent and interactive writing assistants, we first decide on the scope (Section 3.1) and core aspects of writing assistants (Section 3.2). Then, we perform a systematic literature review while employing an iterative coding process and create the design space (Section 3.3, Figure 2). We highlight that this process was a collaborative effort consisting of a large team of researchers from a variety of disciplines, including HCI, NLP, Information Systems (IS), and Education.

3.1 Scope

We define the scope of our work by specifying what we consider as *intelligent* and *interactive writing* assistants. Note that these are working definitions that align with the objectives of our study, and are not intended to be imposed as a universally accepted definition.

Task
- **Writing stage** (e.g., planning, drafting, revision)
- **Writing context** (e.g., academic, journalistic, technical)
- **Purpose** (e.g., expository, narrative, descriptive)
- **Specificity** (e.g., general direction, detailed requirements)
- **Audience** (e.g., specified, implied)

User
- **Demographic profile** (e.g., age, language & culture)
- **User capabilities** (e.g., writing expertise, efficiency)
- **Relationship to system** (e.g., agency, ownership, trust)
- **System output preferences** (e.g., coherence, diversity)

- **User - Steering the system** (e.g., explicit, implicit, no control)
- **User - Integrating system output** (e.g., selection, inspiration)

Interaction
- **UI - Interaction metaphor** (e.g., agent, tool, hybrid)
- **UI - Layout** (e.g., writing area, separated, input UI)
- **UI - Interface paradigm** (e.g., text editor, chatbot)
- **UI - Visual differentiation** (e.g., formatting, location)
- **UI - Initiation** (e.g., user-initiated, system-initiated)

Technology
- **Data - Source** (e.g., experts, users)
- **Data - Size** (e.g., small, medium, large)
- **Model - Type** (e.g., rule-based, foundation model)
- **Model - External resource access** (e.g., tool, data)
- **Learning - Problem** (e.g., classification, generation)
- **Learning - Algorithm** (e.g., supervised, unsupervised)
- **Learning - Training and adaptation** (e.g., fine-tuning, prompting)
- **Evaluation - Evaluator** (e.g., automatic, machine-learned)
- **Evaluation - Focus** (e.g., linguistic quality, controllability)
- **Scalability** (e.g., cost, latency)

- **System - Output type** (e.g., analysis, generation)
- **System - Curation type** (e.g., deterministic, curated options)
- **System - User data access** (e.g., input text, additional data)

Ecosystem
- **Digital infrastructure** (e.g., usability consistency, technical interoperability)
- **Social factors** (e.g., designing with stakeholders, designing for social writing)
- **Locale** (e.g., local writing, remote writing)
- **Access model** (e.g., free and/or open-source software, commercial software)
- **Norms & Rules** (e.g., laws, conventions)
- **Change over time** (e.g., authors, readers, writing, information environment)
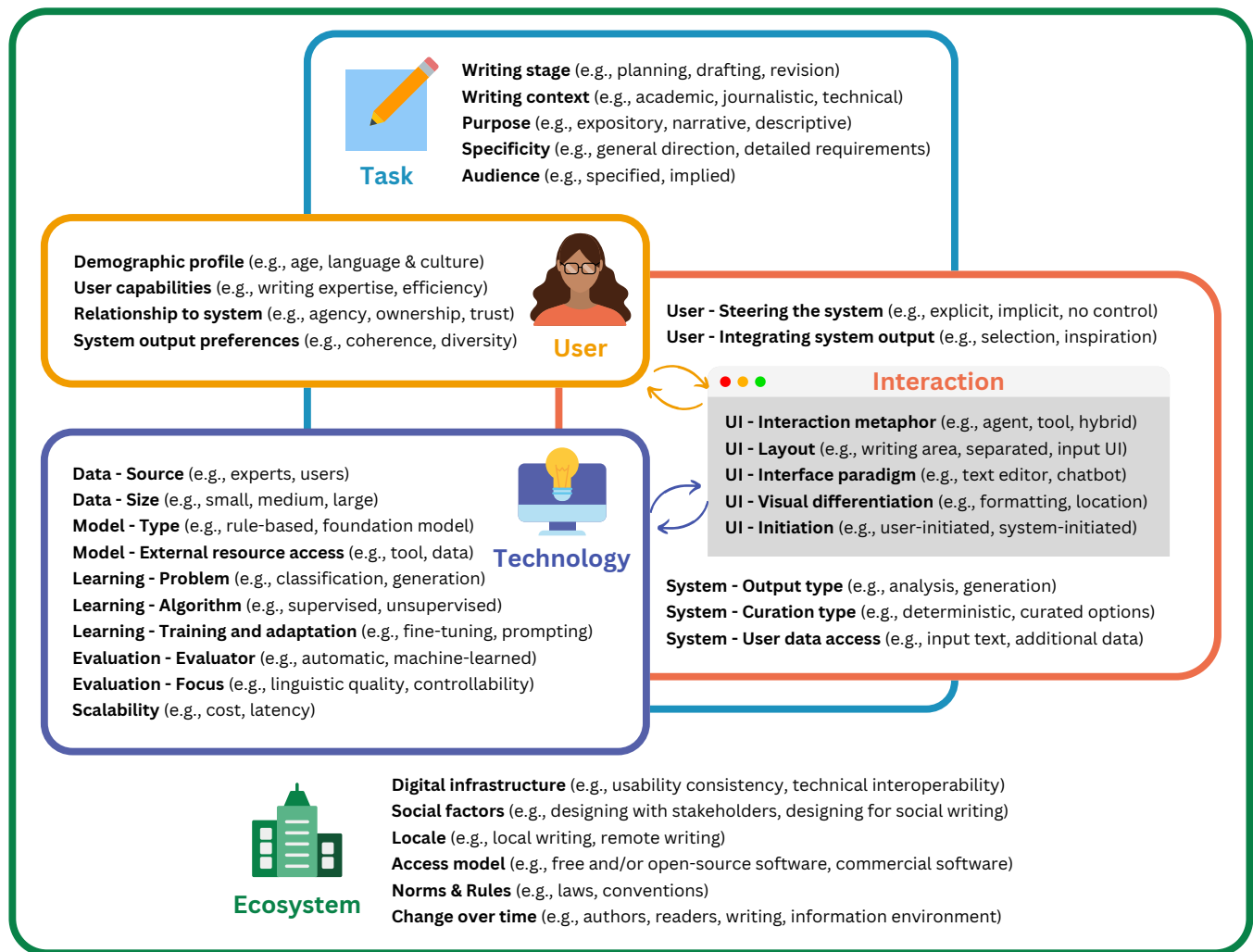
Figure 1: Our design space for intelligent and interactive writing assistants consists of five key *aspects*—task, user, technology, interaction, and ecosystem—that are interconnected and interdependent. Within each aspect, we define *dimensions* (bold texts) that represent fundamental components of the aspect and *codes* (examples associated with bold texts) that represent possible options for each dimension. When necessary, we group semantically relevant dimensions together within each aspect and use a prefix to denote the group name; the interaction dimensions are grouped by user, user interface (UI), and system; likewise, the technology dimensions are grouped by data, model, learning, and evaluation.

- **Intelligent**: We consider systems to be intelligent if they are capable of autonomous decision-making and/or text generation, with a special focus on modern AI, such as language models (LMs).
- **Interactive**: We consider systems to be interactive if they reflect human input and/or output (as opposed to generating text without human involvement) and facilitate an iterative process to produce a written artifact.
- **Writing**: We consider systems to be relevant to writing if at some point human users translate their thoughts into written language, perhaps via system support, and produce text in a natural language (e.g., English) as a final artifact.

To have a stronger and tighter focus on intelligent and interactive writing assistants, we exclude the following types of work: programming tools (to focus on writing as a means to communicate with humans as opposed to machines), collaborative tools (to focus on the dynamics between a user and a system as opposed to multiple users), text-entry tools (to focus on tools that facilitate the cognitive process of writing, rather than input methods, such as gestural, hand-writing, or speech-to-text). Throughout the paper, we interchangeably use "interactive and intelligent writing assistants" and "writing assistants" to refer to the writing assistants that fall within our scope (see Appendix B for our definitions of similar terms like "technology," "system," and "model").

3.2 Five Aspects of Writing Assistants

Writing assistants are *sociotechnical systems* that involve interaction between both technical (e.g., AI and hardware infrastructure) and social (e.g., user behaviors and societal norms) components. The sociotechnical systems perspective is an interdisciplinary approach [19, 145, 217], widely used in various research fields like HCI [140, 151] and IS [39, 259] to analyze complex interactions involving technology, workflow design, and organizational changes.

A core assumption of the sociotechnical systems perspective is that technology does not exist in a vacuum and is always interconnected with people and other systems. The perspective considers "task," "user," "technology," and "structure" as integral and interdependent parts of a complex system [145]. We adopt this point of view with the following modifications: We first split "technology" into two—"technology" and "interaction"—to be more aligned with current research in NLP and HCI, respectively, and rename "structure" as "ecosystem" to account for the broader context in which writing assistants operate. As a result, we designate these five key *aspects* of writing assistants.

- **Task**: Writing stages, contexts, and purposes that writing assistants aim to support. This involves understanding the purpose of writing, the constraints imposed on writing, and the intended audience for the written content.
- **User**: Characteristics and preferences of users of writing assistants, providing insights into how different user groups may prioritize different attributes in their interactions with these systems.
- **Technology**: Building blocks of developing underlying models that power writing assistants, involving the quality and quantity of training data, modeling problems and techniques, and evaluation methods.
- **Interaction**: Diverse interaction paradigms and essential user interface components of writing assistants, that contribute to the dynamic interplay between the user, the interface, and technology powering the interface.
- **Ecosystem**: Issues stemming from the broader context in which writing assistants operate. This involves economic, social, and regulatory considerations that impact how these systems are developed, used, and evolved.

3.3 Systematic Literature Review

Figure 2 shows the overall process of our systematic literature review and the design space creation.

3.3.1 Paper Selection. To understand the current research literature on writing assistants, we performed a systematic literature review. We first identified fields closely related to intelligent and interactive writing assistants. We selected the fields of HCI and NLP as our core fields[1] and their relevant associated venues listed in Appendix D.1. From the ACM Digital Library and ACL Anthology, we retrieved 419 candidate papers (as of August 2023) that included the participles of the verb "write" and its variations either in the title

---

[1] While there are other relevant fields (e.g., Machine Learning, Cognitive Science, Writing, and Education), we excluded them as they usually do not consider all of the three elements (i.e., intelligent, interactive, and writing) that define our scope (Section 3.1). Nevertheless, we note that many of the selected papers tend to be interdisciplinary and cover various subcommunities.

or keywords, if provided by authors (see Appendix 3.3 for details). Then, three of the authors filtered candidates encompassed by the scope of the project (Section 3.1). These authors first independently annotated the relevance of candidate papers, discussed disagreements, and then refined the scope. As a result, we selected 115 out of 419 papers that fulfilled the inclusion criteria for our review.

Note that the above sampling method targeted a specific set of *existing* writing assistants, as opposed to all possible options for *future* writing assistants. Concretely, papers under our review explicitly included a specific technology, a human user, and interaction between the user and the technology through an interface in the context of writing. If a paper did not present a strong connection to all of these components, we excluded it from the review. For instance, we excluded a technical paper that rewrites content in a target style even though it could be used in writing assistants (i.e., it did not explicitly consider user and interaction aspects, but studied style transfer in isolation). In the following section, we describe how we complement the strict exclusion of these works by leveraging the expertise of our paper's authors by adding flexibility to the coding process.

3.3.2 Code Development. With the 115 papers, the authors were split into five teams based on the five aspects (Section 3.2) to develop codes for each aspect. Then, every team sub-selected the papers that were relevant to their aspect. For instance, some papers might allude to potential users but lack descriptions of exact target users; the team focusing on the user aspect (henceforth "the user team") excluded those papers from their consideration. At least two authors from each team read each paper to decide the relevance while resolving any disagreement through discussions.

To go beyond the understanding of *existing* writing assistants and think about possibilities for *future* writing assistants, we intentionally allowed flexibility in the code development process in the following ways. First, while some teams (task and user) derived the initial codes mostly from their selected papers (i.e., inductive coding), other teams (technology, interaction, and ecosystem) created the initial codes based on external knowledge (i.e., deductive coding). Second, to provide insights on recent advancements in technology that are not yet leveraged for writing assistants, the technology team brought in 25 additional papers (outside of the 115 papers) based on relevance and importance (see Appendix D.4 for more details) and referenced them while iteratively refining the codes. Third, most teams considered relevant works that are not necessarily about writing assistants (e.g., papers solely about LMs) to inform their codes. Note that we used these external references to aid our code development but did not code them as part of our review process.

With the subset of 115 papers selected by each team, the authors in the team iteratively coded papers while refining the codes. The iterative coding process began with the distribution of papers to members of each team. We distributed papers so that each paper was read by at least two members, while maximizing the number of combinations of readers to facilitate discussion among team members. With distributed papers, each team member first independently read and coded papers with the up-to-date version of the codes. Then, team members had a discussion session to share and agree on how they coded each paper. If team members decided that
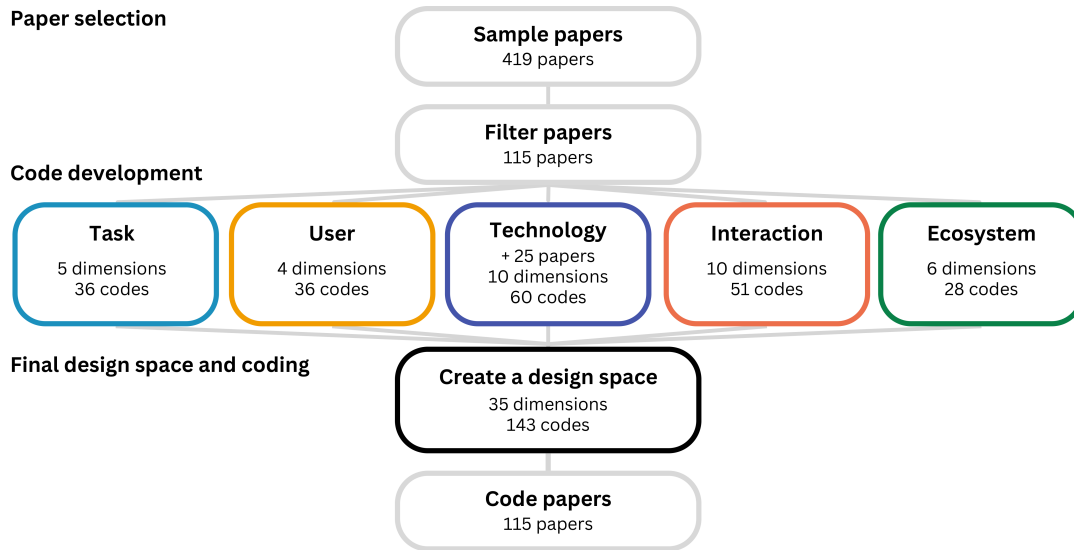
Figure 2: Our systematic literature review has three stages. First, we sampled and filtered papers relevant to writing assistants from HCI and NLP venues. Second, the authors split into five teams based on the five key aspects of writing assistants and developed codes by reviewing papers and employing an iterative coding process. In the final stage, all teams gathered to create the final design space by combining all dimensions and codes and coded all papers.

a specific paper could not be adequately coded with the existing code structure, they updated the code structure either by revising, merging, splitting, adding, or removing codes. As codes were updated, team members returned to already-reviewed papers to ensure that papers were coded with the up-to-date code structure. The process of coding papers repeated until the team reviewed all papers under the final set of codes.

3.3.3 Final Design Space and Coding. After each team developed their codes, all teams gathered to create the final design space by combining the dimensions and codes from all teams. During the process, we removed overlapping dimensions and codes, improved their consistency across the aspects (e.g., inclusion criteria and granularity), and revised their names and definitions for clarity. Once we finalized the design space, we repeated the iterative coding process, where each paper was read by two authors (same as the initial iterative coding in Section 3.3.2). This time, all 115 papers were coded with the full set of 35 dimensions and 143 codes from the five aspects. We briefly analyze trends and gaps based on this final coding of the papers (Section 5.2) and report two metrics for inter-coder reliability: Percent agreement (mean: 0.93, std: 0.06) and Krippendorff's alpha (mean: 0.69, std, 0.17).[2] We release our coded papers as a living artifact at https://writing-assistant.github.io and encourage others to contribute by adding papers beyond those we covered in this work to track future developments in this space.

---

[2]Krippendorff's alpha is applicable when multiple coders see only partial instances, which is our case as many coders contributed to coding all the papers. However, Krippendorff [141] suggest that this metric is only reliable for binary classification (in our case, codes either occur or do not) when certain conditions (e.g., minimum code occurrence) are met; for this reason we excluded codes that occur less than 50 times and considered 72 codes for calculating Krippendorff's alpha.

4 DESIGN SPACE

In this section, we present a *design space* as a structured way to examine and explore the multidimensional space of writing assistants (Figure 1). Our design space encompasses five key *aspects*—task, user, technology, interaction, and ecosystem (Section 3.2)—which are interconnected and play vital roles in the realm of writing assistants. Within each aspect, we define *dimensions* that represent the fundamental components of that aspect and *codes* that denote potential options for each dimension, based on our systematic literature review (Section 3.3). In the following sections, we provide detailed explanations regarding the dimensions and codes specific to each aspect along with concrete examples of research papers associated with each code. When doing so, we formulate our dimensions as questions and present codes as possible answers to the questions, similar to questions and options in MacLean et al. [166].

4.1 Task

In the design space of writing assistants, we define a *task* as a rhetorical purpose of a written artifact identified by a user (e.g., to persuade) at a specific stage of writing (e.g., revision) and in a particular context (e.g., academic). This task can be articulated to a writing assistant with varying degrees of requirements (from no explicit guidelines to detailed instructions) and a particular audience. With this definition, the task aspect embodies user-driven objectives, facilitating a nuanced, user-centered approach.

4.1.1 Dimensions and Codes. We describe dimensions and codes for the task aspect. Figure 1 ("Task") shows task dimensions in a broad context, while Table 1 lists all dimensions, codes, and definitions.

	Code	Definition
Writing Stage	*At what point in the writing process is the task taking place?*	
	Idea generation	Brainstorming and developing concepts or content themes
	Planning	Organizing the structure and content outline
	Drafting	Composing the written content
	Revision	Reviewing and refining the written material
Writing Context	*What combination of stylistic norms, audience expectations, and domain-specific conventions characterize the approach to the task?*	
	Academic	Focuses on research, analysis, and formal presentation of knowledge within educational contexts
	Creative	Focuses on imagination, narrative, artistic elements, and original storytelling
	Journalistic	Focuses on factual reporting, news coverage, and conveying information to the public
	Technical	Focuses on complex information, instructions, or explanations in specialized fields
	Professional	Focuses on precise and formal writing like reports and documents
	Personal	Focuses on individual thoughts, experiences, and emotions; can be private or communicative
Purpose	*What is the purpose of the written artifact?*	
	Expository	Intending to convey factual and informative content to the audience
	Narrative	Intending to convey a story
	Descriptive	Intending to provide expressive details
	Persuasive	Intending to convince or influence opinions or actions
	Educational	Intending to teach and help people learn
	Entertainment	Intending to engage and amuse for leisure or enjoyment
	Analytical	Intending to provide in-depth examination, analysis, or evaluation
	Accessibility	Intending to support individuals with health conditions or impairments
	Translation	Intending to convert content from one language to another
	Feedback	Intending to provide evaluative comments or responses to content
Specificity	*How detailed are the task requirements?*	
	Nonspecific	Having no explicit guidelines, instructions, or objectives
	General direction	Having a broad indication of the desired outcome without specifying precise steps
	Specific objectives	Having some specific objectives contributing to the task
	Detailed requirements	Having detailed and measurable instructions
Audience	*Who is the intended recipient of the written artifact?*	
	Specified	Audience is clearly identified or explicitly stated
	Implied	Audience is assumed or inferred without explicit mention

Table 1: Task dimensions, codes, and definitions.

Writing Stage: **At what point in the writing process is the task taking place?** In the intricate and iterative process of writing, a writing assistant can be designed to support specific writing stages[3] to provide more targeted and relevant support. For example,

during the idea generation stage, writers brainstorm and develop concepts. In this stage, writing assistants can offer ideas for potential

---

[3]Note that our writing stages differ from the cognitive processes of writing proposed by Flower and Hayes [73], despite the similarity in terminology (they use "planning," "translating," and "reviewing" to describe writing subprocesses). Rather, our writing stages resemble (yet are not the same as) the stage models of writing, such as Britton et al. [21], Rohman [213], that model the growth of the written artifact rather than the inner workings of the writer. We choose this approach because we find it more

intuitive to design writing assistants to support a writing stage as a high-level cluster of relevant cognitive processes, compared to designing one writing assistant for each cognitive process. To illustrate the distinction, consider writing an outline of a paper as an example of a task in the planning stage. There are multiple cognitive processes involved in the task, such as figuring out what to write ("planning"), jotting these ideas down as bullet points ("translating"), and reorganizing these points to enhance the overall flow ("reviewing"). Therefore, there is a natural one-to-many relationship between our writing stages and the cognitive processes of writing.

topics and content [85, 220]. In the planning phase, writers focus on organizing the structure and content outline. Writing assistants can play a crucial role in aiding writers during this phase, for instance, by reducing planning time [10, 130]. During the drafting stage, writers may require assistance with thought facilitation or initial drafts [180, 191]. The revision stage involves identifying and rectifying mistakes, making necessary corrections, and enhancing the content for its purpose(s). Writing assistants at this stage might be used to provide feedback on various aspects of the written text [2, 252].

Writing Context: **What combination of stylistic norms, audience expectations, and domain-specific conventions characterize the approach to the task?** This dimension delves into the diverse contexts of writing that shape the task's approach based on community and domain-specific norms. The academic context is characterized by rigorous research, thorough analysis, and the formal presentation of knowledge. Within this context, writers may benefit from assistance that upholds formal practices in genres such as scientific writing [9] and theses [206]. In the creative context, the focus lies on imagination, artistic expression, and original storytelling. An intelligent writing assistant within this context can provide support for diverse creative endeavors, including writing lyrics [257], crafting metaphors [84], and offering inspiration and assistance for storytelling [187, 220]. The journalistic context centers around factual reporting, news coverage, and effective communication of information to the public. Writing assistants can play supportive roles in this process by assisting with science writing [133] and generating suggestions based on keywords [47]. In the technical context, authors convey complex information, provide instructions, and explain specialized topics, such as by writing figure captions [188]. In the professional context, precise and formal writing is essential. Writing assistants in this context can be tailored to aid in producing reports and documents related to companies, trade, and professional work [114, 250]. In the personal context, individuals engage in private or public reflections on their thoughts, experiences, and emotions. Writing assistants tailored for this context can provide support for tasks that involve emotional expression in writing [191, 255], as well as fostering empathetic relations with readers [195].

Purpose: **What is the purpose of the written artifact?** A writing task may be motivated by a broad range of writing goals that guide the writing process. One fundamental goal is to explain; expository applies to writing tasks whose purpose is to convey relevant facts and knowledge [138, 206]. Another common purpose is telling a story; narrative applies to tasks whose goal is to convey an account of real [90] and imagined experiences [88, 109, 211]. Some texts are written to be descriptive in order to convey emotion [255] and provide expressive details, particularly with evocative language [83, 84, 160]. Meanwhile, persuasive writing aims to convince or influence the audience through text [2, 121], requiring high readability [129] and conciseness [115]. Sometimes, the purpose of the written artifact can be educational (e.g., readers learning about a topic by reading the written artifact). In this context, writing assistance could assist writers in conveying educational ideas in an informative and clear manner [10, 85]. With entertainment as its purpose, a writing task can aim to engage and amuse readers through text, requiring coherent storytelling [133, 174] to engage

the audience and may use the structure of the writing itself to convey surprise [80] and pleasure [146, 257]. Analytical writing provides in-depth examination, which benefits from reflection and iteration [57, 113]. Accessibility describes tasks that aim to support inclusivity; for example helping neurodivergent individuals explain and write about common social situations [134]. Likewise, second language writing can be difficult due to unfamiliar spelling and grammar rules [157] as well as vocabulary [112], and writing assistants can help by aiding in translation to ease the difficulty of writing in an unfamiliar language [112, 112, 157, 277]. Lastly, feedback describes tasks that provide evaluative comments, such as customer reviews [13, 63] and constructive responses [220].

Specificity: **How detailed are the task requirements?** To answer this question, we examine design choices that add specification to the writing task. Nonspecific is applied to tasks that do not have specific objectives. Systems supporting these tasks often include generic drafting environments that bundle many functionalities [108, 146] or general-purpose systems that do not specify a particular objective [195, 211]. Tasks that provide a broad indication of the desired outcome without specifying the precise steps needed to achieve it are considered to have general direction as their specificity. These tasks may suggest the direction of an outcome through writing conventions [192], writing reflection variations [80, 134], or writing format [47, 64], but will not define how the desired outcome should be realized. On the other hand, some writing tasks have specific objectives that directly connect to the overall goal. For example, while a general direction might be a target writing format, specific objectives may suggest the inclusion of certain subsections [54, 115] or information [188] that contribute to that overall goal. Finally, tasks with detailed requirements come with detailed and measurable instructions for the task. Weber et al. [258], for example, outline a task for supporting legal writing based on the case solution's major claim, definition, subsumption, and conclusion, as well as elements and relations in the subsumption, which are highly specific requirements about what should be done in the task.

Audience: **Who is the intended recipient of the written artifact?** This dimension describes the audiences for whom the output of the task is intended. Specified audiences are specifically mentioned in the paper, such as the academic community [85], people on the autistic spectrum [134], people with limited vision [180], and writers themselves [12, 216]. An implied audience is inferred by the system design or characterization of the textual artifact, for example online shoppers for a tool designed to help writers review products [64], business stakeholders for a system that aids with writing introductory help requests [115], and writers themselves for a tool that assists with personal reflection [255].

4.2 User

In this section, we ask who stands to benefit from these assistants and illuminate the varied needs and preferences of users that might influence design considerations for their systems. These distinctions remind us that the effectiveness of a writing assistant often lies in its attunement to a user's unique requirements.

	Code	Definition
Demographic Profile	*What are the demographic details of the users considered?*	
	Gender	User's gender
	Race	User's race or ethnicity
	Socioeconomic status	User's socioeconomic status
	Language & Culture	User's primary language and cultural background
	Age	User's age
	Education	User's educational background
	Profession	User's profession
User Capabilities	*What user attributes associated with the writing process can be influenced by writing assistants?*	
	Writing expertise	User's writing expertise in terms of writing quality or genre specializations
	Efficiency	User's writing efficiency, often measured as number of words written, time spent, or effort expended
	Technical proficiency	User's understanding of and comfort level with the underlying technology
	Confidence	User's confidence in the writing process and final product, such as self-efficacy or perceived skill level
	Creativity	User's engagement in creative exploration, including fostering curiosity and innovative thinking
	Emotion	User's emotional state before, during, or after using the writing assistant
	Empathy	User's ability to emotionally and cognitively empathize within the context of writing
	Cognition	User's cognitive aspects, such as focus, sense of immersion, cognitive load, and writer's block
	Neurodiversity	User's neurological profiles (both neurotypical and neurodivergent)
Relationship to System	*What aspects characterize users' perceptions and their relationship with the system?*	
	Agency	User's sense of control or autonomy in their interactions with the writing assistant
	Ownership	User's sense of ownership or authenticity over the written artifact when using the writing assistant
	Integrity	User's concerns about plagiarism and sense of integrity when using the writing assistant
	Trust	User's sense of trust in the writing assistant's ability or perception of its suitability for a task
	Availability	User's expectation of the writing assistant being at hand when one needs or wants to use it
	Privacy	User's concerns about how their data is handled by the writing assistant
	Transparency	User's understanding of the writing assistant's mechanism, capabilities, and limitations
System Output Preferences	*What influences users' perceptions of system outputs?*	
	Textual coherence	Outputs that are coherent in terms of grammar, content, and tone
	Textual diversity	Outputs that are novel and diverse, providing inspiration or surprises to the user
	Explainability	Additional information for outputs that explains the rationale behind system outputs
	Bias	Outputs that exhibit various forms of bias, such as skewed perspectives on topics and societal stereotypes
	Personalization	Outputs that are personalized based on user preferences

Table 2: User dimensions, codes, and definitions.

4.2.1 Dimensions and Codes. Figure 1 ("User") shows user dimensions in a broad context, while Table 2 lists all dimensions, codes, and definitions.

Demographic Profile: **What are the demographic details of the users considered?** The demographic profile of users reflects a broad spectrum of their characteristics, highlighting the diversity inherent across various groups of users. Some systems

ensure equal representation of minority and marginalized groups by incorporating gender and race into their design [106]. Others assist the writing process of users with diverse socioeconomic status, as it influences the availability and usage of technology, with consequences for user experiences and needs [89]. Language and culture, reflecting users' fluent languages and cultural backgrounds, uniquely shape their writing experiences and outcomes. A number

of studies focus on non-native English writers [27, 37, 277] in English writing contexts to address challenges in non-native language writing and to enhance language learning experiences. Understanding age is important when tailoring systems to accommodate the developmental and cognitive characteristics of users across a wide age spectrum, ranging from young, pre-literate children [270] to adolescents [89, 216]. Similarly, education level and background of users, whether they are high school students [216] or university graduates [2, 251], indicates varying cognitive and learning competencies, which in turn may influence how users interact with and critically engage in systems. Finally, some studies emphasize tailoring system designs to a specific profession, such as professional creative writers [80, 174], ensuring that the systems meet the unique writing needs of different professions.

User Capabilities: **What user attributes associated with the writing process can be influenced by writing assistants?** User capabilities represent a distinct category from demographic attributes, which remain largely unchanged while interacting with writing assistants; user capabilities are often targeted for improvement by writing assistants. Writing expertise captures the user's writing proficiency or expertise (e.g., amateur vs. professional writer) in terms of writing quality [187, 188] or genre specializations (e.g., science writing vs. email writing) [85, 115]. Efficiency, or how efficiently the user can complete specific writing tasks, was also considered in many papers. Several studies have utilized metrics such as word count, time, and effort expended by the user to quantify improvements in the user's writing efficiency before, during, and after using a writing assistant [7, 27]. Technical proficiency relates to the extent to which a user is knowledgeable about the underlying technology of a writing assistant. Understanding how a LM functions, for instance, has been shown to influence how effectively the user engages with a writing assistant [187]. Several papers in the literature focus on enhancing user capabilities related to emotional and cognitive aspects. For example, several studies capture the user's confidence in both the writing process and the final product, possibly expressed as self-efficacy or perceived skill level [115, 262]. Creativity examines how the writing assistant can foster the user's creative exploration or curiosity when performing a writing task, for instance, by supporting idea generation [45, 85, 268]. Emotion refers to the user's emotional state before, during, and after interacting with the writing assistant [16, 191, 255]. Empathy focuses on the user's ability to emotionally and cognitively empathize with others in the writing process. This empathetic focus was observed within educational contexts, where students are instructed to write more empathetic peer reviews [251, 252]. Cognition looks at cognitive aspects like the user's focus, sense of immersion, and cognitive load, as systems can increase cognitive engagement by tackling phenomena like writer's block [12, 45, 220]. Finally, neurodiversity encompasses considerations for users with diverse neurological profiles, such as aphasia [181] and dyslexia [71].

Relationship to System: **What aspects characterize users' perceptions and their relationship with the system?** As users engage with a system, they gradually develop a mental model of its functioning, which subsequently shapes their interaction and engagement with the system. First, agency refers to users' sense of control over the system or the writing process. It is typically facilitated by providing users with options to steer the model outputs, either through adjusting model parameters [47, 88] or by using customized prompts [58, 268]. The ownership or authenticity of a final product can be influenced by system design. A writer's sense of ownership may diminish as the proportion of system-generated text increases [148], yet this issue could be mitigated by personalizing writing assistants to mimic a writer's unique style [86], or by designing assistants with greater agency [181]. Similarly, maintaining a sense of integrity is an important factor when assisted by AI. This encompasses worries about unintentional plagiarism and the moral implications of using writing assistants [15, 85]. Trust is another critical facet of users' perceptions, referring to their perception of the system's capabilities and their sentiments toward the technology itself. The level of trust users hold towards AI could influence the human-AI collaborative writing experience [15, 158]. A system's availability emerged in the context of comparing human support to computer support, where human writers (e.g., friends) are not always readily available, while computer programs are typically perceived as constantly accessible [86]. Privacy highlights users' concerns regarding how their data is handled by a system, including a sense of surveillance over their writing process [12, 191, 262]. Lastly, users are concerned with the transparency [155] of writing assistants as they seek clarity on how systems operate [13, 195], how data is used [195], and AI's role in these systems [13, 34, 191].

System Output Preferences: **What influences users' perceptions of system outputs?** Understanding how writers evaluate system outputs, such as writing suggestions, is crucial as it can influence their interaction and engagement with the system. One common consideration is textual coherence, which underlines the need for grammatically and contextually coherent outputs [85, 148]. Another significant dimension is textual diversity, which emphasizes the importance of offering varied system outputs to foster creativity in writing [84, 85, 229]. The explainability of the system can also influence users' perceptions of its outputs. Providing additional information to explain the rationale behind system outputs may enhance user understanding and engagement [192]. Additionally, system-generated content may exhibit various forms of bias, ranging from skewed perspectives on topics [121, 197] to societal stereotypes [106]. Lastly, the personalization of system outputs, which involves adapting to and reflecting an individual's unique writing style [80], may enhance the user's writing experience.

4.3 Technology

The technology aspect of writing assistants considers the advancements that underpin their intelligence and capabilities. In this section, we describe core building blocks of the end-to-end development process of models that power writing assistants. In particular, we consider data, models, learning problems and techniques, and evaluation, all of which play a crucial role in determining the quality and degree of intelligence in the writing assistants.

4.3.1 Dimensions and Codes. Figure 1 ("Technology") shows technology dimensions in a broad context, while Tables 3 and 4 list all dimensions, codes, and definitions.

	Code	Definition
Data Source	*Who is the creator of the data used to train or adapt a model?*	
	Experts	Experts of the task or domain of interest
	Users	Current or target users of a downstream application
	Crowdworkers	Crowdworkers from various platforms, such as Amazon Mechanical Turk and Prolific
	Authors	Authors of the research paper
	Machine	Another model that generates synthetic data
	Other	Other creators, such as non-experts and unspecified individuals for web crawled data
Data Size	*What is the size of the dataset that used to train or adapt a model?*	
	Small (<100)	Uses one to a hundred data points
	Medium (<10k)	Uses a hundred to a few thousand data points
	Large (<1M)	Uses tens to hundreds of thousands of data points
	Extremely large (>1M)	Uses millions of data points or more
	Unknown	Unknown data or dataset size
Model Type	*What is the type of the underlying model?*	
	Rule-based model	Model relying on heuristics or deterministic approaches (e.g., lookup tables and regular expressions)
	Statistical ML model	Machine learning (ML) model typically trained for a specific task (e.g., logistic regression)
	Deep neural network	ML model that uses multiple layers between the input and output layers (e.g., RNNs and LSTMs)
	Foundation model	Pre-trained deep neural network that can be adapted for downstream tasks (e.g., BERT, GPT-4)
Model External Res.	*What additional access does a model have at inference time?*	
	Tool	External software or third party APIs the model might rely on to perform its task
	Data	External datasets or resources, such as external knowledge repositories
Learning Problem	*How is the writing assistance task being formulated as a learning problem?*	
	Classification	Assigns inputs to predefined categories or classes
	Regression	Predicts continuous numeric values for a given input
	Structured prediction	Focuses on capturing dependencies, relationships, and patterns in language data
	Rewriting	Translates text from one form to another, while preserving its meaning and information content
	Generation	Creates new, coherent, and contextually relevant text
	Retrieval	Finds and optionally ranks relevant instances for a given input
Learning Algorithm	*How is the underlying model being trained?*	
	Supervised learning	Model is trained on labeled data where each input is associated with the correct output
	Unsupervised learning	Model is trained on unlabeled data to discover patterns and structures within the data
	Self-supervised learning	Model creates a supervisory signal from the data itself, without human-annotated labels
	Reinforcement learning	Model learns by interacting with an environment and receiving feedback in the form of rewards

Table 3: Technology dimensions, codes, and definitions (continued in the next table).

Data - Source: **Who is the creator of the data used to train or adapt a model?** The source of the data used to develop a system or train a model can have a direct effect on the system's overall performance and reliability. A dataset can be authored by experts who have domain knowledge of the specific downstream task [3, 129, 258, 276], or users of the system, during their interaction with the writing assistant [9, 27, 179, 255]. However, due to the difficulty of recruiting real experts and users, many researchers resort to crowdworkers to create data or annotate data entries [35, 131, 254, 275]. Sometimes, authors themselves participate in the preparation and annotation of the dataset [126, 195, 197, 272]. Recently, we see more datasets that are generated by a machine [106, 124, 199, 230], which has the advantage of being relatively cheap and fast to generate at scale compared to human-generated datasets. Finally,

	Code	Definition
Learning *Training & adaptation*	*How is the underlying model being trained or adapted for a specific task at hand?*	
	Training from scratch	A new model is trained from scratch, or a foundation model is used without adaptation
	Fine-tuning	A foundation model is further trained on a specific dataset
	Prompt engineering	A foundation model is further adapted via prompts (a.k.a. in-context learning)
	Tuning decoding parameters	A model's decoding parameters are adjusted to stir model outputs (e.g., temperature and logit bias)
Evaluation *Evaluator*	*Who evaluates the quality of model outputs?*	
	Automatic evaluation	Quality is evaluated based on simple aggregate statistics or on similarity metrics
	Machine-learned evaluation	Quality is evaluated by another model to produce ratings or scores (e.g., BERTScore)
	Human evaluation	Quality is evaluated by human annotators
	Human-machine evaluation	Quality is evaluated by both human annotators and another model, often by having the model filter a subset of outputs to be annotated by humans
Evaluation *Focus*	*What is the focus of evaluation when evaluating individual model outputs?*	
	Linguistic quality	Evaluation focuses on qualities such as grammaticality, readability, clarity, and accuracy
	Controllability	Evaluation focuses on the reflectiveness to controls or constraints specified by users or designers
	Style & Adequacy	Evaluation focuses on the alignment between model ouputs and their surrounding texts
	Ethics	Evaluation focuses on social norms and ethics, such as bias, toxicity, factuality, and transparency
Scalability	*What are the economic and computational considerations for developing and deploying a model?*	
	Cost	Cost of deploying the model
	Latency	Delay between when the model receives an input and generates a corresponding output

Table 4: Technology dimensions, codes, and definitions (continued from the previous table).

there are other types of creators such as non-experts, unspecified individuals, or a broad set of creators (e.g., in the case of web crawled data) [13, 228, 268, 277].

Data - Size: **What is the size of the dataset**[4] **used to train or adapt a model?** Depending on the size of a dataset required to train or adapt (e.g., fine-tuning or prompting) a model, there can be a huge overhead in terms of data collection. While some models can be developed using very small data (between 1 to 100 examples) [10, 85, 158], the others require much larger data. If the training needs more data (around 100 to few thousands of examples) which is often the case for fine-tuned models, we categorize them as medium [37, 97, 255, 256, 271]. For larger datasets (around tens of 1000s of examples) we denote this as large [56, 207, 257]. For models that undergo extensive large-scale pre-training, we categorized data used in this process as extremely large to indicate a dataset of millions of examples [43, 223, 228, 276] or more. We also included an unknown if the paper did not explicitly mention the dataset used for training [180, 192, 238].

Model - Type: **What is the type of the underlying model?** Advancements in AI accelerators and the availability of large amounts of data have led to an evolution in model architectures,[5] which

we capture as the following four types. First, rule-based models rely on pre-defined logic, lookup tables, regular expressions, or other similar heuristic approaches that are deterministic in nature [10, 29, 90, 221]. For statistical machine learning (ML) models, we consider models that are trained from scratch on historical data, are not necessarily "deep" (as in deep neural networks), and are used to make future predictions (e.g., support vector machines and logistic regression) [112, 122, 195, 271]. Over the past decade, deep neural networks have been the popular models of choice for writing assistants, including recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) [7, 47, 211, 262]. Finally, recent works have increasingly utilized foundation models, such as BERT [61], RoBERTa [159], GPT [23, 201, 202], and T5 [203], to name a few. A foundation model is "any model that is trained on broad data that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" [17]. These models can perform a wide range of tasks out of the box, learn from a few examples to provide tailored support to users, and be further fine-tuned for specific downstream task(s) [148, 223, 254, 258].

Model - External Resource Access: **What additional access does a model have at inference time?** Recently, models have been developed with access to additional tools or data at inference time to make them capable of providing assistance beyond the knowledge encoded in their parameters. In the case of tool, a model may access external software or third-party APIs to perform tasks like search, translation, or calculator, or even setting calendar events

---

[4]Note that we defined the size of a dataset with respect to the number of examples, not the total number of words in the dataset. As examples can vary in length, ranging from a single word to the length of a book, the actual size of the dataset doesn't necessarily correlate with the number of examples.

[5]Bommasani et al. [17, §1] provides an overview on AI research over the last 30 years.

on behalf of users [43, 180, 267, 272]. Data refers to external datasets or resources, such as information stored in a database, external knowledge repositories, or any other structured/unstructured data sources that the models might leverage to provide writing assistance [133, 223, 230, 276].

Learning - Problem: **How is the writing assistance task being formulated as a learning problem?** How exactly writing assistants support their users usually varies based on the learning problem their underlying models are designed to solve. Classification refers to the class of problems that require categorizing data into predefined classes based on their attributes. It is one of the most widely formulated classes of problems in writing assistants, applicable to tasks such as detecting errors in writing [77, 246] and detecting the purpose of writing revisions [124], among other tasks. In contrast, regression problems involve the prediction of a continuous numerical value or quantity as the output instead of categorical labels or classes. This includes problems such as the prediction of the writer's sentiments [252], the readability [129], or the emotional intensity [240] of written text as numerical ratings or scores. Structured prediction refers to a class of learning problems that involve predicting structured outputs or sequences (e.g., sequences, trees, and graphs) rather than single, isolated labels or values. Numerous works have focused on developing these approaches to make edits to improve the quality of written text during the revision stage [137, 167–169, 233]. Rewriting problems involve sequence transduction tasks, where texts from one form are transformed to another while improving the quality by making them fluent, clear, readable, and coherent. These tasks are essential in various writing assistance applications, such as grammatical error correction [37, 43, 277], paraphrasing [267], or general-purpose text editing [66, 70, 204, 226] to name a few. Generation refers primarily to problems that involve the creation of new, contextually relevant, coherent, and readable text from relatively limited inputs, such as autocomplete, paraphrasing, and story generation [7, 45, 117, 220]. Retrieval problems take the input from a user as a query (e.g., keywords) and search in a knowledge base or dataset for relevant information. Such problems may involve ranking the available data based on its relevance and similarity to the input but do not necessarily include the generation of new text beyond what is available in the knowledge base [37, 223, 232].

Learning - Algorithm: **How is the underlying model trained?** The models used as back-bones of writing assistants incorporate different training mechanisms based on the type of the available data, as well as the specific downstream tasks. In supervised learning, models are trained on a labeled dataset where each input is associated with the correct output. Some of the commonly used methods include Logistic Regression, Random Forests, and Naive Bayes [16, 26, 97, 256]. Supervised learning also includes approaches such as Transfer Learning, which involves training a model on a large dataset and then fine-tuning it for a specific task or domain using a smaller dataset [68, 277]. In unsupervised learning, models are trained on unlabeled data to learn patterns and structures within the data. This approach includes techniques such as representation learning and clustering methods, to name a few [187, 257]. Self-supervised learning approaches train models on unlabeled data with a supervisory signal [81]. These approaches leverage the benefits of both supervised and unsupervised learning, especially in scenarios where obtaining a large amount of labeled data is challenging. This includes pre-training objectives for large language models such as Causal Language Modeling [202] and Masked Language Modeling [61]. In reinforcement learning (RL), models learn by interacting with an environment and receiving feedback in the form of rewards. This approach is useful for tasks requiring action sequences, such as language generation and dialogue systems [228].

Learning - Training and Adaptation: **How is the underlying model being trained or adapted for a specific task at hand?** The training and adaptation process is integral part of developing an intelligent model that can perform tasks at hand and support user needs. Before foundation models, many models used to be trained from scratch [83, 98, 160, 238]. On the other hand, with the advance of foundation models, the common learning paradigm has been shifted to "pre-training" a large model on broad data and then "adapting" the model to a wide range of downstream tasks. One way to adapt a model is fine-tuning, where the pre-trained model is further trained on a specific dataset [13, 197, 229, 275]. Note that there are numerous variants of fine-tuning, such as *transfer learning*, *instruction tuning*, *alignment tuning*, *prompt tuning*, *prefix tuning*, and *adapter tuning*, among others. Another way to adapt a model is prompt engineering (or "prompting"), where one can simply provide a natural language description of the task (or "prompt") [22] to guide model outputs [58, 121, 148, 174]. A prompt may include a few examples for a model to learn from ("few-shot learning" or "in-context learning"). Lastly, it is possible to tune decoding parameters of a model to influence model outputs (e.g., changing temperature to make outputs more or less predictable, manipulating logit bias to prevent some words from being generated) [88, 148, 221].

Evaluation - Evaluator: **Who evaluates the quality of model outputs?** A core aspect of model development is its evaluation. We consider four common types of evaluators who can review and evaluate various qualities of model outputs (as opposed to writing assistants or user interactions). Automatic evaluation compares machine-generated outputs with human-generated labels or texts using aggregate statistics or syntactic and semantic measures. These include metrics like precision, recall, F-measure, and accuracy, as well as ones commonly used in generation tasks, such as BLEU [190], METEOR [144], and ROUGE [266] to name a few [3, 37, 238]. Machine-learned evaluation uses automated metrics, which are themselves produced by a machine-learned system. These are typically classification or regression models that are trained to evaluate the quality of model outputs [124, 199, 222, 252, 273]. In contrast, human evaluation corresponds to evaluating the system with human annotators either directly interacting with, or evaluating the output of a writing assistant. Some evaluations may require the judging of task-specific criteria (e.g., evaluating that certain entity names appear correctly in the text [175]), while others can be generalized for most text generation tasks (e.g., evaluating the fluency or grammar of the generated text [158, 165, 250, 268]). Human-machine evaluation captures cases where both machine-learned metrics or models and human judges are involved in the evaluation of the outputs. This hybrid evaluation is particularly relevant in co-creative, mixed-initiative writing assistance settings. Such

studies often involve expert users and participatory methodologies [45, 148, 156, 174].

Evaluation - Focus: **What is the focus of evaluation when evaluating individual model outputs?** Evaluating (or benchmarking) models has been a long standing challenge in NLP [154]. In particular, as we increasingly use foundation models (e.g., GPT-4) for a wide range of downstream tasks, it is difficult to evaluate the quality of model outputs across all tasks, let alone the difficulty of evaluating open-ended generation. Here, we highlight four common evaluation foci relevant to writing assistants in the literature. Linguistic quality focuses on the grammatical correctness, readability, clarity, and accuracy of the model's outputs. This aspect ensures that the outputs are not only correct in terms of language use but also easily understandable and precise in conveying the intended message [58, 129, 254, 276]. Controllability assesses how well the model's outputs reflect constraints (or control inputs) specified by users or designers (e.g., how effectively the model adheres to any specific level of formality or writing style) [121, 197, 220, 241]. Furthermore, it is crucial that the model's responses not only make sense in isolation, but also fit seamlessly within the broader context of the text. Style and adequacy pertain to the alignment between the model's outputs and their surrounding texts or contexts. This includes evaluating the stylistic and semantic coherence, relevance, and consistency of the outputs with the given context [84, 162, 180, 268]. Finally, ethics encompasses a broad range of crucial considerations such as bias, toxicity, factuality, and transparency. Ethics focuses on the adherence of the model's outputs to social norms and ethical standards, and seeks to avoid generating outputs containing harmful biases, misinformation, and other unethical elements [15, 106, 195, 230]. This aspect of evaluation is particularly critical in maintaining the trustworthiness and societal acceptance of the model.

Scalability: **What are the economic and computational considerations for developing and deploying a model?** Recent models, especially LMs, have demonstrated exceptional performance across various tasks [25, 40, 186]. However, the significantly large size of these models has substantially increased the cost of their development [128]. In this regard, directly utilizing pre-trained LMs via prompting [23, 260], employing efficient fine-tuning methods like low-rank adaptation [110], or prefix-tuning [153] can help avoid the cost of full fine-tuning. During deployment, this affects not only the inference costs but also system latency, which often degrades user experience [30, 149]. Techniques such as quantization [87] and knowledge distillation [104] have shown promising results in addressing these issues.

4.4 Interaction

The interaction between a user and a writing assistant primarily involves three key components: User, user interface (UI; frontend), and system (backend). The UI acts as a mediator, facilitating interaction between the user and the system, as illustrated in Figure 3.

4.4.1 Dimensions and codes. Figure 1 ("Interaction") shows interaction dimensions in a broad context, while Figure 3 shows a detailed visualization. Tables 5 and 6 list all dimensions, codes, and definitions.

UI - Interaction Metaphor: **What is the interaction metaphor for the system?** Interaction metaphors shape how the user relates to the system. We identify three primary metaphors. First, systems designed as agents employ designs meant to evoke human-like interaction, including roles like "collaborator" or "co-writer" [36, 148], "dialog partner" [191, 249], "assistant" [80], and "companion" [84]. Techniques for evoking the agent metaphor include character-by-character text rendering to simulate typing [121], avatars [29], implicit and explicit conversational interaction [191], and first-person conversational styles [195]. In contrast, other systems present as tools, where there is no sense of interacting with another agent. These systems tend to avoid conversational styles and rather present feedback in imperative or factual style [114, 250, 258]. They provide traditional GUI elements (e.g., checkboxes, buttons) that spread out system capabilities rather than centralizing them into one "agent." Hybrid systems blend "agent" and "tool" metaphors in their design or their authors' descriptions [84, 148].

UI - Layout: **Where are the system interactions situated in the UI?** Often, interaction with the underlying system takes place in the writing area where users create text. This supports the selection of the existing writing [57, 223] and/or seamless integration of output from the system [58, 148]. Alternatively, a design might choose to isolate the interaction with AI as a separated UI element, such as through a sidebar. A separated design puts clear boundaries between the users' writing and the system output. For example, this design style is used to display information related to text diagnosis [246], provide inspiration [45, 133], or support language learning [37]. Interaction with the system can also be embedded with the user's text input UI, such as text suggestions on touchscreen keyboards [7]. Finally, there are custom designs: for example, a tangible UI that triggers the system by lifting a coffee cup [12], or an exploratory visualization for referencing information which has no user text input [131].

UI - Interface Paradigm: **What is the general platform of the interface?** The text editor is the prevalent interface for AI-assisted writing [58, 223], providing writers with a traditional blank canvas while incorporating a variety of system-driven functionalities such as feedback [114], automated checklists [80], or completion suggestions [121]. In contrast, chatbot interfaces use turn-taking interactions for motivation [191] or suggestions [35]. Conversational exchanges serve to progressively and iteratively achieve a goal, like fiction writing [15, 48, 220]. Finally, other interfaces cater to specific needs, introducing novel [131, 216], sometimes multimodal [45, 229], interactions. For example, such custom interfaces take lyrical structure into account for music generation [257], and emphasize the conceptual and figurative expressions to generate metaphors for scientific writing [133].

UI - Visual Differentiation: **How is system output differentiated from user output?** This dimension identifies different visual designs for separating user-written text and system outputs. A system must present its outputs to the user in some way, or the interaction loop between the user and the system is incomplete. The most common mechanism is to use text formatting such
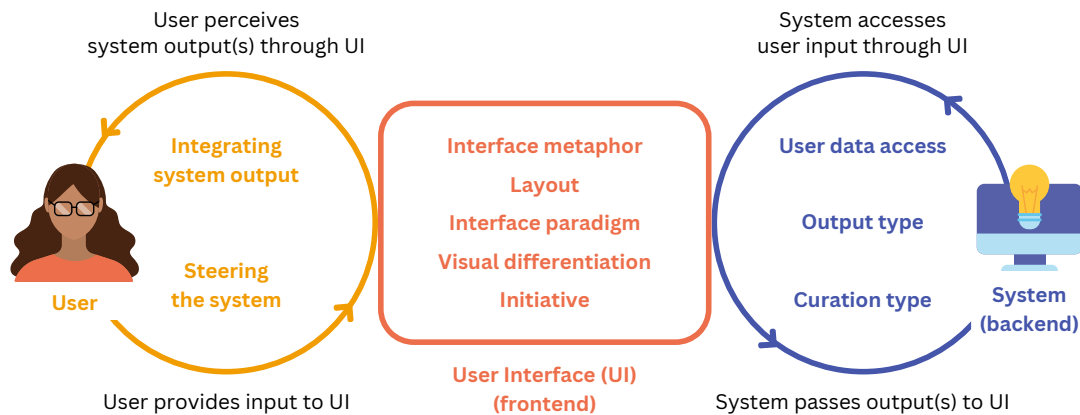
Figure 3: A visualization of the three main components for the interaction aspect in our design space. The visualization outlines the relationships between a user, user interface (frontend), and system (backend). Users interact with the system through the interface by perceiving system outputs and providing input to the system. The system interacts with the user through the interface by accessing user input and data and generating outputs that get rendered in the interface.

as colors and underlines [58, 267, 268]. Another common mechanism is to keep system outputs in a separate location like a wizard or panel [57, 195, 252]. A system may not use any formatting when the system output has a different media type (e.g., meta-analysis [114, 258] or audio [12]) than the text written by the user. Alternatively, the differentiation is "none": System output is presented identically to user output after it is added into the text [7, 148].

UI - Initiation: **How is system output triggered?** System outputs can be triggered in two main ways: User-initiated triggers give the user control over when system output is created or presented in the UI. For example, phrase suggestions might be displayed whenever the user presses the "tab" key [148] or clicks on a "generate" button [85]. In contrast, with system-initiated triggers, the system has control over when it brings in its output. This might be a rule, heuristic, or a dedicated trigger decision model. For example, the user and system might take turns such that the system output is triggered when the user submits a chat message [191] or sentence [48], or pauses writing for a certain amount of time [13, 27, 121]. For some systems, this initiative is almost real-time, or "live." For example, several systems update suggestions and feedback in a dedicated panel while the user is typing [7, 37, 246].

User - Steering the System: **How can the user steer the system?** Users need to communicate their intentions and goals to a system to steer its behavior. In the implicit paradigm, users directly compose the artifact (typically, text in a workspace) as a way to communicate their intentions to the system which in turn provides support based on this information. Such systems often provide support in the form of text suggestions [148, 268], reflection [57, 191], or inspiration and ideation [84]. Alternatively, users have explicit controls over the system behavior. For example, users give thumbs up or down on shown text suggestions to steer future output towards user preferences [49], control the diversity of generated text through numeric parameters [88], or steer story arcs via sketching

[45]. The two paradigms have implications for the users' workflow [58]: implicit input can be less disruptive to the writing process, while explicit input offers more ways for users to express intentions. Finally, some systems offer no control to the user, for example, if the LM is prompted in the background to respond in a particular way regardless of the user's prior text [121].

User - Integrating System Output: **How does the user integrate system output?** After the system creates an output, the user can choose how to integrate it into the overall goal. For material intended to be included in the output, the user may take a selection action, such as accepting a suggestion using a button or key press [7, 133, 148]. When the system does not provide material for direct inclusion, the user may engage with the output as inspiration and choose to type their own text [191, 255]. If the system text is inserted into the final text with no explicit user interaction, the user must then choose whether to keep, edit, or remove the text, through editing [45, 174]. Finally, the system may provide outputs that require no integration, if they are not meant to be added to an output or to inform direct contributions, as in analysis for future improvements [114, 258].

System - User Data Access: **What user data can the system access through the UI?** Systems can access different types of user data to produce the desired outputs. Typically, the user's input text is the primary source of data; the current writing progress can be used to generate completions [7, 148] or provide feedback [29, 77]. It is also important to allow users to have additional controls over the text generations, which we characterize as additional data. For example, some systems allow users to specify the task for generation via explicit instruction [35, 268] or via sketching [45].

System - Output Type: **What type of output does the system create?** Writing assistants can generate different types of outputs to support users' writing. System output might serve as an analysis of the user's writing, e.g., how clear text is in an administrative context [77], how empathetic a peer review is [252], or how structured

	Code	Definition
UI — Interaction metaphor	*What is the interaction metaphor for the system?*	
	Agent	System is meant to evoke a sense of another agent acting on the interface or data
	Tool	System is presented as tools, where there is not a sense of interacting with another agent
	Hybrid	System draws from both "agent" and "tool" metaphors
UI — Layout	*Where are the system interactions situated in the UI?*	
	Writing area	In-situ interactions take place in the same place as user output text
	Separated	Interface isolates the system outputs and controls in a separate layout panel
	Input UI	Interactions are situated in the input device
	Custom	System has a dedicated custom interface
UI — Interface paradigm	*What is the general platform of the interface?*	
	Text editor	The main interface is a text editor
	Chatbot	The main interface is a chat client
	Other	Other UI paradigms, including those developed custom for the application at hand
UI — Visual differentiation	*How is system output differentiated from user output?*	
	None	System output is not differentiated from user output after it is accepted into the text
	Formatting	System output is included with user output, but is differentiated by color or other formatting
	Location	System output is in a separate location or UI panel from user output
	Media	System creates a different type of output than the user, such as images
UI — Initiation	*How is system output triggered?*	
	User-initiated	Users initiate a request or prompt for system to generate an output
	System-initiated	System provides output based on internal rules
User — Steering the system	*How can the user steer the system?*	
	Explicit	User can control system behavior by selecting buttons, checkmarks, etc.
	Implicit	User updates user text and the system takes it as input to generate output
	No control	User cannot control system; the only controls were pre-decided by the designer

Table 5: Interaction dimensions, codes, and definitions (continued in the next table).

		Code	Definition
User *Integrating system output*		*How does the user integrate system output?*	
		Selection	Users select a system output, such as accepting a suggestion from multiple suggestions using a button or key press
		Inspiration	Users do not keep system output, but may be inspired to create new text on their own
		Editing	Users keep, edit, or remove system output
		No integration	System may provide outputs that are not meant to be added or inform direct contributions
System *User Data Access*		*What user data can the system access through the UI?*	
		Input text	The text the user is working on
		Additional data	Extra data that are not intended to be part of the input text, such as random seeds, control labels, or prompts
System *Output Type*		*What type of output does the system create?*	
		Analysis	Feedback, analytics, or context based on automatic analysis of the user's text
		Generation	New content that is intended to be incorporated into the final product
		Proposal	New content that is meant to be referenced but not directly incorporated into the final product
System *Curation Type*		*How are system outputs curated?*	
		Model	Model generates outputs which are directly used by the system
		Curated	System designers curates a list of outputs in advance, and the system picks one for the user
		Customized	A response from a curated list is selected and then further customized by the system for the current context
		Deterministic	User input automatically determines outputs

Table 6: Interaction dimensions, codes, and definitions (continued from the previous table).

and formal writing is in a professional context [114]. Systems can provide this analysis in the form of annotations or highlights on the text [114, 252], analytics in a separate UI [129, 133], or high-level statements about the text [29, 206]. In contrast, generation is usually realized as completions [7, 13, 257], although it can also encompass longer sections (e.g., creating a paragraph of text [45, 68]). Finally, some systems' outputs only provide proposals: Content meant to be used as reference but not directly incorporated into the document. This content can be provided in the form of planning and outlining support [206, 216], or questions about the user's writing to encourage reflection [29] or summaries of the text so far [57].

System - Curation Type: **How are system outputs curated?** The predominant way that intelligent writing support systems curate outputs is to generate a custom response from an NLP model, especially with the recent dramatic increase in fluency and capability of LMs [121, 148, 211, 257]. However, providing the natural language outputs of such models requires the designer to give up a great deal of control over the possible outputs. In sensitive contexts such as topics like mental health [191, 195, 255] or to leverage the

benefit of high-quality source material, designers may choose to trade off the flexibility of an NLP model for the control of a pre-made, curated list of responses [232], from which an output is selected by the system [29, 37, 220]. To increase personalization, these pre-curated responses may also be dynamically customized [191]. Alternatively, the output may be generated from user input in a deterministic way, in writing systems that do not use AI and/or prefer rule-based automation [12, 138, 216].

4.5 Ecosystem

As writing assistants become embedded in authentic contexts, we must consider the embodied, material, sociotechnical environment (see Appendix C for discussion on micro- vs. macro-HCI). Following Guggenberger et al. [92], we consider the ecosystem as *"the overarching sociotechnical context in which the writer and the tool are situated, encompassing a range of complex, interdependent actors that frequently play a role in the functioning of the writing assistant."* This aligns with the writing model proposed by Hayes [99], updating the cognitive processes of writing proposed in 1981 [73] to add the *social* and *physical* environments. Although much of the ecosystem

	Code	Definition
Digital Infrastructure	*What compatibility issues are considered?*	
	Usability consistency	Alignment of user experience with other systems that the user is familiar with
	Technical interoperability	Ability to communicate and work together with other systems, applications, or devices seamlessly
Access Model	*How does the openness of data, models, and products influence design decisions?*	
	Free and open-source software	Factors related to free access, derivation of work, and redistribution
	Commercial software	Factors related to the use of or integration as part of a commercial product
Social Factors	*Who affects the design and use of writing assistants?*	
	Design with stakeholders	Accounting for stakeholders' perspectives and behaviors
	Design for social writing	Accounting for writers' social context of writing, such as co-authors
Locale	*Does the writing assistant's design take into account features of a physical locale?*	
	Local writing	Design for writing at "home base" with full sociotechnical networking
	Remote writing	Design for writing remotely from "home base"
Norms and Rules	*What norms and rules affect the design and use of a writing assistant?*	
	Laws	Legal requirements, such as privacy, copyright, and age-appropriate content
	Conventions	Cultural, professional, or organizational norms & standards
Change Over Time	*What are the key temporal considerations when designing a writing assistant?*	
	Authors	Changes in authors' perception, knowledge, and skills regarding writing assistants
	Readers	Changes in readers' perception, knowledge, and skills regarding writing assistants
	Writing	Changes in written artifacts due to the use of writing assistants
	Information environment	Changes in the existing text and knowledge landscape due to the use of writing assistants
	Technologies	Changes in the technologies powering writing assistants
	Regulation	Changes in the laws and conventions that govern the use of writing assistants in the ecosystem

Table 7: Ecosystem dimensions, codes, and definitions.

is beyond the immediate control of writing assistant designers, this aspect draws attention to how one might design in anticipation of its influences.

4.5.1 Dimensions and codes. Figure 1 ("Ecosystem") shows ecosystem dimensions in a broad context, while Table 7 lists all dimensions, codes, and definitions.

***Digital Infrastructure:* What compatibility issues are considered?** A key compatibility issue to consider is usability consistency, i.e., the extent to which the writing assistant's user experience intentionally aligns with other systems in the writer's ecosystem. Examples in the corpus include integrating AI language technologies into everyday writing apps [96], extending Google Docs [247], and using familiar-looking BibTeX style codes to enable writers to invoke remote bibliographic searches [10]. This example also illustrates technical interoperability with external services, which designers may conceive in many ways. For instance, another project integrated tangible media with their writing assistant, such

that lifting a mug on a digital coaster triggered text-to-speech replay of recent sentences to assist reflection [12].

***Access Model:* How does the openness of data, models, and products influence design decisions?** This dimension taps into how the openness and licensing of data, models, and products may influence design decisions and dissemination of artifacts. First, when a new writing assistant is developed to be embedded within existing commercial software, products, or platforms, the aesthetic and display space of these products exerts a strong influence over the UX and UI design of the writing assistant, as illustrated by the examples of Airbnb [122] and Facebook [191]. This influence could further affect the choice of data (e.g., proprietary data) and models (e.g., closed models). On the other hand, for many researchers, it is a common and often desired practice to build and work with free and open-source software (as well as open-source data and models [242, 244, 261]) given their transparency and free availability. Lastly, researchers in turn can open source their writing

assistants [148, 181] as well as other associated artifacts, such as interaction traces for human-LM collaborative writing [148] and programming languages for generating text [108], to promote collaboration and reproducibility.

Social Factors: **Who affects the design and use of writing assistants?** This dimension draws attention to how *designers* engage with stakeholders, and the social support around *authors*. First, designing with stakeholders concerns human-centered and participatory design methods to meaningfully involve target authors, and other groups such as educators, coaches, or subject-matter experts. This was by far the most prevalent ecosystem dimension in the corpus, attributable to design concept interviews, co-design sessions, and usability studies with target authors [174, 188, 229]. While not represented in the corpus, other research focuses on co-designing tools with educators to build trust in their design [215, 224]. Second, designing for social writing covers the formal and informal support network which writers may call on, including co-authors, peers, mentors, friends and family. For instance, children writing at home may involve informal social support from their parents [234, 236].

Locale: **Does the writing assistant's design take into account features of a physical locale?** Digital systems are used in physical environments, whose affordances could aid writing (e.g., whiteboard used to plan a document, sticky notes on screens, and opportunistic conversations with passing colleagues) or could impede writing (e.g., a noisy environment may permit notetaking but disrupt the attention needed for detailed writing; fieldwork with limited Internet access may prevent certain writing until back in the office). This dimension draws attention to whether the writer is engaged in local writing (i.e., at what is considered to be "home base" with full social and technical networking), or remote writing (i.e., away from home base, with different affordances and constraints) such as fieldwork or commuting. Much of the research to date did not attend to this explicitly, but examples include designing an intentionally calm interface to help reduce classroom distractions for vulnerable young people [89], and using a messaging platform to encourage more reflective writing for well-being [191, 255].

Norms & Rules: **What norms and rules affect the design and use of a writing assistant?** As writing is embedded in countless societal processes, it will inevitably be governed by various norms and rules, both formal and informal. First, there can be alignments or misalignments with laws. While none of the corpus papers addressed legal issues, this is of course an active field of scholarship [95]. Furthermore, U.S. and E.U. legislation is changing [53, 107] with intellectual property legal cases under way [75, 209]. Consequently, this code draws designers' attention to legal changes, which could conceivably shift market preference to LMs trained on ethically sourced data, or passing a particular algorithmic impact assessment. Second, writing assistants could account for societal conventions. User-centered design builds on concepts and practices familiar to users, such as writing conventions in job application letters [114] and a system based on an established typology of expository phrases for science writing that readers and writers would recognize [85]. Other types of conventions guiding design decisions include features of good metaphors [84], the social acceptability

of automating emails, and clinical principles to support writers with aphasia [181] or their mental health [191]. Emerging evidence [31, 60] may raise expectations around employee productivity, with the Writers Guild of America strike in the U.S. demonstrating the conflict that the proliferation of AI writing is now provoking [183]. While there were no corpus papers documenting the embedding of writing assistants into established work practices and conventions, we see this beginning to happen in K-12 [215] and higher education [52, 139].

Change Over Time: **What are the key temporal considerations when designing a writing assistant?** This last dimension recognizes that people, technology, regulation, and the broader information environment are in motion, not static. Like the 5,000 year history of writing systems, we anticipate writing assistants and the ecosystem influencing each other over different timescales, from instantaneous to longer-term change. As detailed in the user and interaction dimensions, writing assistants can be shown to have demonstrable effects on authors and their writing outcomes, with empirical studies documenting immediate effects on product reviews [7, 121], stories [229], screenplays [174] and business pitches [250], to name a few. However, while we found no empirical studies beyond a single writing session in our corpus, some researchers anticipate the longer-term risks of homogenization in creative writing [5, 84, 189], professional deskilling in written communication [27], and loss of author agency simply through fatigue in reviewing AI suggestions [13]. Designers should consider the risk that AI-generated text is ingested as training data by other AI projects ("model collapse" [227, 243]), an example of change in information environment [7].

We can also anticipate changes in readers. As LMs can (co-)author complex writing that is hard to distinguish from human-authored text [46, 123], designers should consider different readers' criteria for trustworthiness [122]. However, as with authors, there were no longitudinal studies of readers or reading practices with writing assistants. Given the current pace of technological development, frequent changes in technology will be the norm, much more so than what we observed in the past [94, 163]. An HCI perspective might ask how systems can be designed to gracefully adapt to the evolution of technology, for instance, by assisting authors who dislike an updated LM to roll back to their older, more personally-tuned version (cf. user protests about companion chatbot updates [198] resulting in version management [143]). Finally, as regulation catches up with technological advances, it could affect model performance (e.g., models trained on copyright material could be banned), procurement (e.g., corporate AI governance restricts products meeting new ethical standards), or subscription models (e.g., more "ethical" LMs might cost more).

5 DISCUSSION

In this section, we share use cases for our design space, our reflections on the current literature and ethical implications, the challenges and limitations we encountered in creating the design space, and our plans for future work.
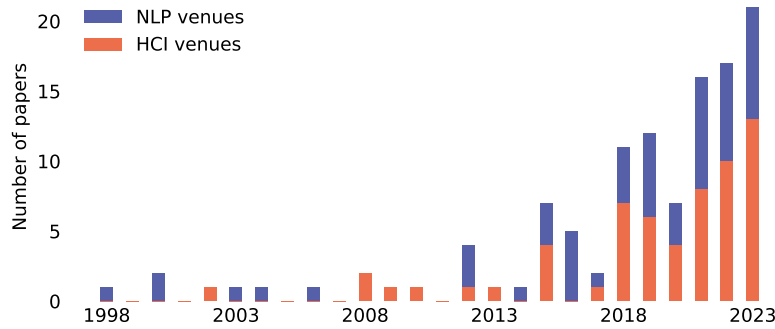
Figure 4: The number of papers in our corpus over the years from NLP and HCI venues (retrieved in August 2023). We see a sharp increase in papers on writing assistants starting in the mid-2010s, with roughly equal increase from NLP and HCI venues.

5.1 Use Case Scenarios for the Design Space

In this section, we present two use case scenarios for our design space. These scenarios illustrate how the design space can be utilized (e.g., generative and analytical) and demonstrate its value for a range of stakeholders (e.g., researchers and policymakers). They also underscore the interdependencies and trade-offs between different dimensions and codes.

Generative scenario. Suppose a group of researchers aims to create a writing assistant, specifically designed to aid non-native English writers in choosing the best paraphrase among multiple paraphrases generated by AI.[6] Motivated by previous work, the researchers plan to build a prototype of the writing assistant to improve the writing quality of the target user group by offering support features. Upon examining the design space, the researchers find they have already factored in numerous dimensions, such as the *writing process* [revision], *writing context* [academic], *demographic profile* [language & culture], *system output preferences* [explainability], and *digital infrastructure* [usability consistency]. However, they realize that they overlooked certain key dimensions like *data - source* and *evaluation - focus* of the foundation model they intend to use, as well as dimensions like *relationship to system* [trust, transparency] and *interaction metaphor* that could influence how the user perceives the system. These insights prompt them to do investigations into their options for foundation models that they may not have previously considered, take into account user concerns of trust and transparency, and think about various ways to frame and present the system to users. Overall, the researchers find that the design space ensures they do not overlook important design decisions, resulting in a richer and more thoughtful design.

Analytic scenario. Suppose a group of policymakers is concerned about the unintended consequences of AI-powered writing assistants swaying public opinions. This worry stems from a research report suggesting that co-writing with opinionated language models (LMs) can influence writers' views.[7] To gain a comprehensive understanding of the context from which these findings originate, the policymakers refer to the research paper as well as our design space. By mapping the writing assistant used in the paper to the design space, they gain a nuanced understanding of the experimental context (e.g., *writing stage* [drafting], *specificity* [general direction], and *model - type* [foundation model]). More importantly, they identify several factors that could potentially alter the findings. For example, the writing assistant in the study automatically provided suggestions to users (*UI - initiative* [system-initiated]), rather than allowing users to request suggestions when needed. Furthermore, users had no way to *explicitly* control or guide the system's output, and had limited *implicit* control; even though the underlying model took user text as input to account for the user's writing style and existing contents, it was consistently prompted by the system to output text in favor of a pre-determined stance on the topic that the user was asked to write about (*user - steering the system* [implicit]). Recognizing these factors, the policymakers realize they could potentially introduce regulations to make user-initiated interactions mandatory and to allow users to explicitly steer the system's output. They could also recommend that designers visually differentiate between user-generated and AI-generated text (*UI - visual differentiation*). Finally, the design space draws attention to a state regulatory proposal to categorize as "high risk" any AI system that could subtly bias voting behavior through nudging (*norms & rules* [laws]).

5.2 Trends and Gaps in the Literature

Based on an analysis of our corpus of papers, we see that there is a sharp increase in papers about writing assistants starting in the mid-2010s (Figure 4). This increase is roughly equal among HCI and NLP venues, though there are slightly more HCI than NLP papers in our corpus. It is this increase that spurred our interest in creating a design space to support the increasing number of researchers and designers working in this space.

Based on our final coding of all papers, we observe that certain dimensions in the design space are over- or under-represented. Figure 5 shows the number of papers per dimension which were coded as relevant. To highlight a few notable trends, we see *audience* is under-represented compared to the other task dimensions, suggesting future work may want to more explicitly consider who is the audience of a piece of writing. *Scalability* is quite under-represented

---

[6]Note that this is a hypothetical scenario based on Kim et al. [136].
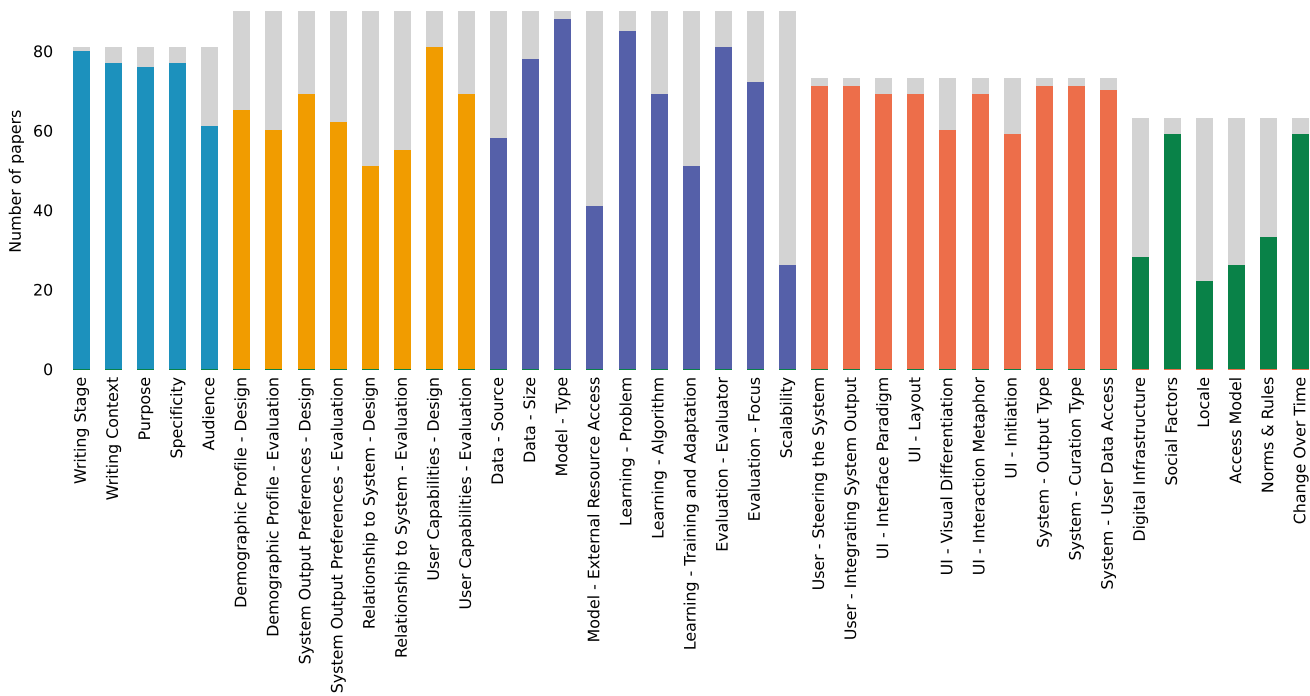[7]Note that this is a hypothetical scenario based on Jakesch et al. [121].

Figure 5: The number of papers for each dimension in the design space. Out of the papers that are relevant to each aspect (i.e., gray bars), we show the number of relevant papers for each dimension (i.e., colored bars). We observe that certain dimensions are over- or under-represented in the current literature and highlight them in Section 5.

overall, as well as relative to other technology dimensions, suggesting that there may currently not be enough consideration of the economic and computational costs of training and using recent large models in the context of writing assistants. Finally, most ecosystem dimensions are, as previously noted, under-represented, representing a rich area of future study as writing assistants become more widely adopted and the circumstance and context of their adoption becomes increasingly important. Longitudinal studies should illuminate if, and in what ways, writers' *relationship to system* changes through extended use, and how it will affect not only the writers, but also readers and information environment (*change over time*).

We also note that technological advances are driving changes in writing assistants. The use of foundation models has rapidly increased in just the past few years; we see 13 papers with this code in 2023 versus 1 in 2020 in our corpus. We expect this number to grow substantially in the coming years. However, we have not yet seen a corresponding increase in codes that seem relevant to their increased usage, such as user concerns of trust and transparency, or technological evaluations of controllability or issues of ethics. We hope that the provision of our design space can help researchers and designers think about these issues as they become increasingly important with rapid technological advances.

5.3 Ethical Implications of Writing Assistants

Writing assistants, while beneficial, hold a potential for serious risks, particularly when intentionally designed or misused by individuals

or organizations to plagiarize content [147, 196], generate deceptive content [116], or systematically sway opinions [121], thus requiring careful scrutiny and ethical considerations. Additionally, uses of writing assistants have begun to affect labor markets [24, 69, 182, 183], signifying substantial societal consequences and evidencing the need for monitoring such effects. As these systems integrate into various industries, a comprehensive, multi-faceted approach to ethical governance is essential to tackle sector-specific concerns.

Another ethical consideration is accommodating the needs and preferences of diverse users, including those from differently-abled, under-represented, and marginalized communities [1, 14, 103]. Beyond traits inherent to users, such as primary language and culture, such accommodations should consider user preferences or capabilities like writing expertise and technical proficiency, which may widely vary across educational, socioeconomic and neurocognitive backgrounds. Failure to address these factors can lead to misalignment of expectations or biased outputs that further perpetuate inequalities and stereotypes. Future work could consider, for instance, *value-sensitive design*, which acknowledges the importance of understanding the needs, preferences, and concerns of different user groups. This is especially important in sensitive contexts, such as education, counseling, or healthcare.

5.4 Challenges in Developing a Design Space

We underscore that the five aspects within the design space have blurry boundaries, as some dimensions may straddle multiple aspects. When defining dimensions, we sought to increase coverage and make them as mutually exclusive as possible. However, in some cases (e.g., the relationship between *writing context* and *purpose*), this cleaving was simply impossible. Defining dimensions and codes that were not frequently mentioned or implicitly addressed in research papers posed additional challenges. For instance, many dimensions for ecosystem (e.g., *digital infrastructure*, *locale*, and *access model*) were sometimes possible to infer from papers but were often not explicitly mentioned.

In addition, we noticed that the dimensions of some aspects have inherently different natures. For instance, the user and ecosystem aspects focus on the very *existence* of codes in work (e.g., whether a paper takes *demographic profile* [age] or *digital infrastructure* [usability consistency] into account when designing a writing assistant). On the other hand, other aspects (e.g., interaction) presume the existence, and focus on the *classification* (e.g., whether *UI - interface metaphor* is close to agent, tool, or hybrid). Furthermore, we find that it is possible for user dimensions to be not only design choices, but reported properties from user studies. For example, researchers might use a general-purpose writing assistant in their user studies, but focus their evaluation and analysis on the users' *relationship to system* [agency]. To handle this contingency, we duplicated user dimensions and coded for both design choices and reported properties, while keeping the official set of user dimensions without duplicates.

Some codes intrinsically have continuous values, and converting these into discrete codes remained challenging (e.g., *specificity* and *data - size*). Even when codes are discrete, the space of possible codes can be vast; in this case, we focus on the elements that are explicitly mentioned and frequently observed in the coding process (e.g., *system output preferences* and *evaluation - focus*), leaving room for future extension. When applicable, we abstracted codes to increase their coverage and generalizability, while trading off their specificity (e.g., generation instead of "dialogue," "story generation," and "question answering"). During the coding process, we were able to select multiple codes for a dimension, to account for a writing assistant's various functionalities and purposes.

5.5 Limitations and Future Work

One limitation of our work is the coverage of sampled and filtered papers, restricted by the search criteria. Using "write" in titles or keywords may not be particularly suitable for NLP papers. While expanding the search to include abstracts would increase the corpus, this approach could also lead to a diminishing return rate. Consequently, the focus remained on developing useful dimensions and codes, rather than striving for an exhaustive collection of papers. Another limitation was not explicitly including commercial writing assistants (without research papers) in our search. In our coding process, it is possible that we may have misinterpreted authors' intentions, which may have resulted in errors in our annotation. Despite these limitations, we argue that this design space serves as a reference for examining existing assistants and developing new ones, while preventing implicit assumptions or overlooked

considerations, thereby facilitating a holistic understanding of the factors that drive design considerations and choices. For future work, we hope to continue to refine our dimensions and codes and code commercial writing assistants to further understand the gaps and opportunities in the current research landscape and suggest possible directions for future research and development.

5.6 Lessons for Future Writing Assistants

To promote creativity and exploration in this emerging area, we intentionally avoid imposing subjective views, prescribing how to design writing assistants, or when and where writing assistants should be used. Instead, we share lessons learned along the way as helpful for our fellow researchers and designers. We believe it is important to recognize the interconnection of dimensions and codes across the five aspects and their trade-offs, and utilize them as a reference when designing writing assistants. As technology continues to evolve, we anticipate the emergence of new capabilities, interaction designs, and insights about user preferences and behaviors. Therefore, it is essential to remain adaptive to these changes and heed them when designing and analyzing writing assistants. Lastly, there are many under-represented or under-explored dimensions in the design space (Section 5.2). We encourage researchers and designers to venture into these overlooked dimensions and offer their innovative ideas and unique insights, contributing to the holistic development of better writing assistants.

6 CONCLUSION

In this work, we present a design space as a structured way to examine and explore the multidimensional space of intelligent and interactive writing assistants. Through community collaboration and systematic literature review across multiple disciplines, we define 35 dimensions and 143 codes exploring five key aspects of writing assistants. We hope that this design space provides researchers and designers with a practical tool to navigate, comprehend, and compare the various possibilities of writing assistants, and aids them in the design of new writing assistants.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. *arXiv preprint arXiv:2101.05783* (2021).

[2] Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective Interfaces for Student-Driven Revision Sessions for Argumentative Writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 58, 13 pages.

[3] 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*. Association for Computational Linguistics, Dubrovnik, Croatia, 2550–2561.

[4] Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6470–6484.

[5] Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. arXiv:2402.01536 [cs.HC]

[6] Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A Resource and Analyses on Edits in Instructional Texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 5721–5729.

[7] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive Text Encourages Predictable Writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 128–138.

[8] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, Tokyo Japan, 603–608.

[9] Tal August, Lauren Kim, Katharina Reinecke, and Noah A. Smith. 2020. Writing Strategies for Science Communication: Data and Computational Analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5327–5344.

[10] Tamara Babaian, Barbara J. Grosz, and Stuart M. Shieber. 2002. A Writer's Collaborative Assistant. In *Proceedings of the 7th International Conference on Intelligent User Interfaces* (San Francisco, California, USA) *(IUI '02)*. Association for Computing Machinery, New York, NY, USA, 7–14.

[11] Beata Beigman Klebanov and Nitin Madnani. 2020. Automated Evaluation of Writing – 50 Years and Counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7796–7810.

[12] Jekaterina Belakova and Wendy E. Mackay. 2021. SonAmi: A Tangible Creativity Support Tool for Productive Procrastination. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy) *(C&C '21)*. Association for Computing Machinery, New York, NY, USA, Article 7, 10 pages.

[13] 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* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 436–452.

[14] Tingting Bi, Xin Xia, David Lo, John Grundy, Thomas Zimmermann, and Denae Ford. 2022. Accessibility in software practice: A practitioner's perspective. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 4 (2022), 1–26.

[15] Oloff C. Biermann, Ning F. Ma, and Dongwook Yoon. 2022. From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1209–1227.

[16] Robert Bixler and Sidney D'Mello. 2013. Detecting Boredom and Engagement during Writing with Keystroke Analysis, Task Appraisals, and Stable Traits. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (Santa Monica, California, USA) *(IUI '13)*. Association for Computing Machinery, New York, NY, USA, 225–234.

[17] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* (2021).

[18] Kyle Booten and Katy Ilonka Gero. 2021. Poetry Machines: Eliciting Designs for Interactive Writing Tools from Poets. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy) *(C&C '21)*. Association for Computing Machinery, New York, NY, USA, Article 51, 5 pages.

[19] Robert P Bostrom and J Stephen Heinen. 1977. MIS problems and failures: A socio-technical perspective. Part I: The causes. *MIS quarterly* (1977), 17–32.

[20] Emma Bowman. 2022. A new AI chatbot might do your homework for you. But it's still not an A+ student. Accessed: Jan 26, 2024.

[21] James Britton et al. 1975. The Development of Writing Abilities. (1975).

[22] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[23] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165* (2020).

[24] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. 2023. *Generative AI at work*. Technical Report. National Bureau of Economic Research.

[25] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

[26] Jill Burstein and Magdalena Wolska. 2003. Toward Evaluation of Writing Style: Overly Repetitious Word Use. In *10th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Budapest, Hungary.

[27] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages.

[28] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Conference on Human Factors in Computing Systems (CHI)*.

[29] Aoife Cahill, James Bruno, James Ramey, Gilmar Ayala Meneses, Ian Blood, Florencia Tolentino, Tamar Lavee, and Slava Andreyev. 2021. Supporting Spanish Writers using Automated Feedback. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*. Association for Computational Linguistics, Online, 116–124.

[30] Shanqing Cai, Subhashini Venugopalan, Katrin Tomanek, Ajit Narayanan, Meredith Ringel Morris, and Michael P Brenner. 2022. Context-Aware Abbreviation Expansion Using Large Language Models. *arXiv preprint arXiv:2205.03767* (2022).

[31] Alexia Cambon, Brent Hecht, Benjamin Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, Margarita Bermejo-Cano, Eric Knudsen, James Bono, Hardik Sanghavi, Sofia Spatharioti, David Rothschild, Daniel G. Goldstein, Eirini Kalliamvakou, Peter Cihon, Mert Demirer, Michael Schwarz, and Jaime Teevan. 2023. *Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity*. Technical Report MSR-TR-2023-43. Microsoft.

[32] Stuart K. Card, Jock D. Mackinlay, and George G. Robertson. 1990. The Design Space of Input Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) (CHI '90). Association for Computing Machinery, New York, NY, USA, 117–124.

[33] Stuart K Card, Jock D Mackinlay, and George G Robertson. 1991. A morphological analysis of the design space of input devices. *ACM Transactions on Information Systems (TOIS)* 9, 2 (1991), 99–122.

[34] Dashiel Carrera and Sang Won Lee. 2022. Watch Me Write: Exploring the Effects of Revealing Creative Writing Process through Writing Replay. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) (C&C '22). Association for Computing Machinery, New York, NY, USA, 146–160.

[35] Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a Poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6848–6863.

[36] Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[37] Jim Chang and Jason Chang. 2015. WriteAhead2: Mining Lexical Grammar Patterns for Assisted Writing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, Denver, Colorado, 106–110.

[38] Yung-Chun Chang, Cen-Chieh Chen, Yu-Lun Hsieh, Chien Chin Chen, and Wen-Lian Hsu. 2015. Linguistic Template Extraction for Recognizing Reader-Emotion and Emotional Resonance Writing Assistance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, 775–780.

[39] Leida Chen and Ravi Nath. 2008. A Socio-Technical Perspective of Mobile Work. *Inf. Knowl. Syst. Manag.* 7, 1,2 (apr 2008), 41–60.

[40] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374* (2021).

[41] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) (KDD '19). Association for Computing Machinery, New York, NY, USA, 2287–2295.

[42] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[43] Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. 2016. Adapting Grammatical Error Correction Based on the Native Language of Writers with Neural Network Joint Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1901–1911.

[44] John Joon Young Chung, Shiqing He, and Eytan Adar. 2021. The intersection of users, roles, interactions, and technologies in creativity support tools. In

[45] *Designing Interactive Systems Conference 2021*. 1817–1833.

[45] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages.

[46] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 7282–7296.

[47] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) (IUI '18). Association for Computing Machinery, New York, NY, USA, 329–340.

[48] Elizabeth Clark and Noah A. Smith. 2021. Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 3566–3575.

[49] Elizabeth Clark and Noah A. Smith. 2021. Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models. In *Association for Computational Linguistics (ACL)*.

[50] Allan Collins and John Seely Brown. 1988. *The Computer as a Tool for Learning Through Reflection*. Springer US, New York, NY, 1–18.

[51] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning (ICML)*. 160–167.

[52] Elena Cotos, Sarah Huffman, and Stephanie Link. 2020. Understanding graduate writers' interaction with and impact of the Research Writing Tutor during revision. *Journal of Writing Research* 12, 1 (2020), 187–232.

[53] European Council. 2023. Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world. Accessed: Dec 12, 2023.

[54] Iria da Cunha, M. Amor Montané, and Luis Hysa. 2017. The arText prototype: An automatic system for writing specialized texts. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, 57–60.

[55] Susan D'Agostino. 2023. ChatGPT Advice Academics Can Use Now. Accessed: Jan 26, 2024.

[56] Xianjun Dai, Yuanchao Liu, Xiaolong Wang, and Bingquan Liu. 2014. WINGS:Writing with Intelligent Guidance and Suggestions. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, 25–30.

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

[58] 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* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 408, 17 pages.

[59] Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews. arXiv:2306.12587 [cs.CL]

[60] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Report. Harvard Business School.

[61] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Association for Computational Linguistics (ACL)*. 4171–4186.

[62] Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling Language Models to Fill in the Blanks. In *Association for Computational Linguistics (ACL)*.

[63] Ruihai Dong, Kevin McCarthy, Michael O'Mahony, Markus Schaal, and Barry Smyth. 2012. First Demonstration of the Intelligent Reviewer's Assistant. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (Lisbon, Portugal) (IUI '12). Association for Computing Machinery, New York, NY, USA, 337–338.

[64] Ruihai Dong, Kevin McCarthy, Michael O'Mahony, Markus Schaal, and Barry Smyth. 2012. Towards an Intelligent Reviewer's Assistant: Recommending

Topics to Help Users to Write Better Product Reviews. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (Lisbon, Portugal) *(IUI '12)*. Association for Computing Machinery, New York, NY, USA, 159–168.

[65] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Association for Computational Linguistics, Dublin, Ireland, 96–108.

[66] Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding Iterative Revision from Human-Written Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3573–3590.

[67] Mark Dunlop and John Levine. 2012. Multidimensional Pareto Optimization of Touchscreen Keyboards for Speed, Familiarity and Improved Spell Checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2669–2678.

[68] Alexandre Duval, Thomas Lamson, Gaël de Léséleuc de Kérouara, and Matthias Gallé. 2021. Breaking Writer's Block: Low-cost Fine-tuning of Natural Language Generation Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 278–287.

[69] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv:2303.10130 [econ.GN]

[70] Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text Editing by Command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5259–5274.

[71] Min Fan, Jianyu Fan, Alissa N. Antle, Sheng Jin, Dongxu Yin, and Philippe Pasquier. 2019. Character Alive: A Tangible Reading and Writing System for Chinese Children At-Risk for Dyslexia. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6.

[72] Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A Multilingual Corpus of Wikipedia Edits for Modeling Language and Discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 305–315.

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

[74] Center for Teaching Exellence at The University of Kansas. 2023. Using AI ethically in writing assignments. Accessed: Jan 26, 2024.

[75] Forbes. 2023. OpenAI, Microsoft hit with new US consumer privacy class action. Accessed: Dec 12, 2023.

[76] Andrew Fowler, Kurt Partridge, Ciprian Chelba, Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2015. Effects of Language Modeling and Its Personalization on Touchscreen Typing Performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 649–658.

[77] Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. 2020. AMesure: A Web Platform to Assist the Clear Writing of Administrative Texts. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Suzhou, China, 1–7.

[78] L. T. Frase. 1983. Human factors and behavioral science: The UNIX™ Writer'S workbench software: Philosophy. *The Bell System Technical Journal* 62, 6 (1983), 1883–1890.

[79] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the Landscape of Creativity Support Tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–18.

[80] Richard P. Gabriel, Jilin Chen, and Jeffrey Nichols. 2015. InkWell: A Creative Writer's Creative Assistant. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition* (Glasgow, United Kingdom) *(C&C '15)*. Association for Computing Machinery, New York, NY, USA, 93–102.

[81] Michael Gamon. 2010. Using Mostly Native Data to Correct Errors in Learners' Writing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, 163–171.

[82] Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. 2022. A Design Space for Writing Support Tools Using a Cognitive Process Model of Writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Association for Computational Linguistics, Dublin, Ireland, 11–24.

[83] Katy Ilonka Gero and Lydia B. Chilton. 2019. How a Stylistic, Machine-Generated Thesaurus Impacts a Writer's Process. In *Proceedings of the 2019 Conference on Creativity and Cognition* (San Diego, CA, USA) *(C&C '19)*. Association for Computing Machinery, New York, NY, USA, 597–603.

[84] Katy Ilonka Gero and Lydia B.Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12.

[85] 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* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019.

[86] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social Dynamics of AI Support in Creative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 245, 15 pages.

[87] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. 2022. A Survey of Quantization Methods for Efficient Neural Network Inference. In *Low-Power Computer Vision*. Chapman and Hall/CRC, 291–326.

[88] Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, Write, and Revise: an Interactive System for Open-Domain Story Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 89–97.

[89] Frederica Gonçalves, Pedro Campos, Julian Hanna, and Simone Ashby. 2015. You're the Voice: Evaluating User Interfaces for Encouraging Underserved Youths to Express Themselves through Creative Writing. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition* (Glasgow, United Kingdom) *(C&C '15)*. Association for Computing Machinery, New York, NY, USA, 63–72.

[90] Amy L. Gonzales, Tiffany Y. Ng, OJ Zhao, and Geri Gay. 2010. Motivating Expressive Writing with a Text-to-Sound Application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1937–1940.

[91] Mitchell Gordon, Tom Ouyang, and Shumin Zhai. 2016. WatchWriter: Tap and Gesture Typing on a Smartwatch Miniature Keyboard with Statistical Decoding. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3817–3821.

[92] Tobias Moritz Guggenberger, Frederik Möller, Tim Haarhaus, Inan Gür, and Boris Otto. 2020. Ecosystem Types in Information Systems. In *Twenty-Eighth European Conference on Information Systems (ECIS2020)*.

[93] Olga Gurevich and Paul Deane. 2007. Document Similarity Measures to Distinguish Native vs. Non-Native Essay Writers. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Association for Computational Linguistics, Rochester, New York, 49–52.

[94] Christina Haas. 2013. *Writing technology: Studies on the materiality of literacy*. Routledge.

[95] Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and Other Large Generative AI Models. (2023), 1112–1123.

[96] Masato Hagiwara, Takumi Ito, Tatsuki Kuribayashi, Jun Suzuki, and Kentaro Inui. 2019. TEASPN: Framework and Protocol for Integrated Writing Assistance Environments. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Association for Computational Linguistics, Hong Kong, China, 229–234.

[97] Harry Halpin, Johanna D. Moore, and Judy Robertson. 2004. Automatic Analysis of Plot for Story Rewriting. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 127–133.

[98] Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring Methods for Generating Feedback Comments for Writing Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 9719–9730.

[99] J.R. Hayes. 1996. A new framework for understanding cognition and affect in writing. In *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, C.M. Levy and S. Randall (Eds.). Lawrence Erlbaum Associates, Mahwah, NJ, 6–44.

[100] John R. Hayes. 2012. Modeling and Remodeling Writing. *Written Communication* 29, 3 (2012), 369–388.

[101] John R Hayes and Linda S Flower. 1986. Writing research and the writer. *American psychologist* 41, 10 (1986), 1106–1113.

[102] D. Jeffery Higginbotham. 1992. Evaluation of keystroke savings across five assistive communication technologies. *Augmentative and Alternative Communication* 8, 4 (Jan. 1992), 258–272.

[103] Charles Hill, Maren Haag, Alannah Oleson, Christopher J. Mendez, Nicola Marsden, Anita Sarma, and Margaret M. Burnett. 2017. Gender-Inclusiveness Personas vs. Stereotyping: Can We Have it Both Ways? *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017).

[104] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*.

[105] Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences. *Transactions of the Association for Computational Linguistics* 11 (2023), 565–581.

[106] Md Naimul Hoque, Bhavya Ghai, and Niklas Elmqvist. 2022. DramatVis Personae: Visual Text Analytics for Identifying Social Biases in Creative Writing. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1260–1276.

[107] The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Accessed: Dec 12, 2023.

[108] Daniel C. Howe. 2009. RiTa: Creativity Support for Computational Literature. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition* (Berkeley, California, USA) *(C&C '09)*. Association for Computing Machinery, New York, NY, USA, 205–210.

[109] Ting-Yao Hsu, Yen-Chia Hsu, and Ting-Hao (Kenneth) Huang. 2019. On How Users Edit Computer-Generated Visual Stories. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6.

[110] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

[111] Chung-chi Huang, Ping-che Yang, Keh-jiann Chen, and Jason S. Chang. 2012. TransAhead: A Computer-Assisted Translation and Writing Tool. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 352–356.

[112] Chung-chi Huang, Ping-che Yang, Mei-hua Chen, Hung-ting Hsieh, Ting-hui Kao, and Jason S. Chang. 2012. TransAhead: A Writing Assistant for CAT and CALL. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, 16–19.

[113] Yi-Ching Huang, Hao-Chuan Wang, and Jane Yung-jen Hsu. 2018. Feedback Orchestration: Structuring Feedback for Facilitating Reflection and Revision in Writing. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (<conf-loc>, <city>Jersey City</city>, <state>NJ</state>, <country>USA</country>, </conf-loc>) *(CSCW '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 257–260.

[114] Julie Hui and Michelle L. Sprouse. 2023. Lettersmith: Scaffolding Written Professional Communication Among College Students. 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 703, 17 pages.

[115] Julie S. Hui, Darren Gergle, and Elizabeth M. Gerber. 2018. IntroAssist: A Tool to Support Writing Introductory Help Requests. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[116] Justin Hutchens. 2023. *The Language of Deception: Weaponizing Next Generation AI*. John Wiley & Sons.

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

[118] Tsunenori Ishioka and Masayuki Kameda. 2006. Automated Japanese Essay Scoring System based on Articles Written by Experts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, 233–240.

[119] Takumi Ito, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, and Kentaro Inui. 2020. Langsmith: An Interactive Academic Text Revision System. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 216–226.

[120] Robert Iv, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2022. FRUIT: Faithfully Reflecting Updated Information in Text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3670–3686.

[121] 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.

[122] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[123] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120.

[124] Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the Human Revision Process in Scientific Writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9420–9435.

[125] Peiling Jiang, Jude Rayan, Steven P Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. *arXiv preprint arXiv:2305.11473* (2023).

[126] Sophia Jit, Jennifer Spinney, Priyank Chandra, and Robert Soden. 2023. Semi-Automated Approach for Evaluating Severe Weather Risk Communication. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 249, 8 pages.

[127] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 955–964.

[128] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv* (2020).

[129] Jakob Karolus, Sebastian S. Feger, Albrecht Schmidt, and Paweł W. Woźniak. 2023. Your Text Is Hard to Read: Facilitating Readability Awareness to Support Writing Proficiency in Text Production. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 147–160.

[130] Harmanpreet Kaur, Alex C. Williams, Anne Loomis Thompson, Walter S. Lasecki, Shamsi Iqbal, and Jaime Teevan. 2018. Using Vocabularies to Collaboratively Create Better Plans for Writing Tasks. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6.

[131] Chris Kim, Uta Hinrichs, Saif M. Mohammad, and Christopher Collins. 2020. Lexichrome: Text Construction and Lexical Discovery with Word-Color Associations Using Interactive Visualization. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 477–488.

[132] Juho Kim. 2022. Interaction-Centric AI (NeurIPS 2022 Keynote). Accessed: Jan 26, 2024.

[133] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 115–135.

[134] Kyunghee Kim, Rosalind W. Picard, and Henry Lieberman. 2008. Common Sense Assistant for Writing Stories That Teach Social Skills. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (Florence, Italy) *(CHI EA '08)*. Association for Computing Machinery, New York, NY, USA, 2805–2810.

[135] Taewook Kim, Jung Soo Lee, Zhenhui Peng, and Xiaojuan Ma. 2019. Love in Lyrics: An Exploration of Supporting Textual Manifestation of Affection in Social Messaging. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 79 (nov 2019), 27 pages.

[136] Yewon Kim, Mina Lee, Donghwi Kim, and Sung-Ju Lee. 2023. Towards Explainable AI Writing Assistants for Non-native English Speakers. arXiv:2304.02625 [cs.CL]

[137] Zae Myung Kim, Wanyu Du, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Improving Iterative Text Revision by Learning Where to Edit from Other Revision Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9986–9999.

[138] Tomi Kinnunen, Henri Leisma, Monika Machunik, Tuomo Kakkonen, and Jean-Luc LeBrun. 2012. SWAN - Scientific Writing AssistaNt. A Tool for Helping Scholars to Write Reader-Friendly Manuscripts. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon,

France, 20–24.

[139] Simon Knight, Antonette Shibani, Sophie Abel, Andrew Gibson, Philippa Ryan, Nicole Sutton, Raechel Wight, Cherie Lucas, Agnes Sandor, Kirsty Kitto, Ming Liu, Radhika Vijay Mogarkar, and Simon Buckingham Shum. 2020. AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research* 12, 1 (2020), 141–186.

[140] Megan Knittel, Shelby Pitts, and Rick Wash. 2019. "The Most Trustworthy Coin": How Ideological Tensions Drive Trust in Bitcoin. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 36 (nov 2019), 23 pages.

[141] Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication methods and measures* 5, 2 (2011), 93–112.

[142] Per Ola Kristensson and Keith Vertanen. 2014. The inviscid text entry rate and its application as a grand goal for mobile text entry. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services (MobileHCI '14)*. Association for Computing Machinery, Toronto, ON, Canada, 335–338.

[143] Eugenia Kuyda. 2023. Replika Updates. Accessed: Feb. 24, 2024.

[144] Alon Lavie and Michael Denkowski. 2009. The Meteor Metric for Automatic Evaluation of Machine Translation. *Machine Translation* 23 (2009).

[145] Harold J Leavitt. 1965. Applied organizational change in industry: Structural, technological and humanistic approaches. In *Handbook of Organizations (RLE: Organizations)*. Routledge, 1144–1170.

[146] Hsin-Pei Lee, Jhih-Sheng Fang, and Wei-Yun Ma. 2019. iComposer: An Automatic Songwriting System for Chinese Popular Music. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 84–88.

[147] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize?. In *Proceedings of the ACM Web Conference 2023*. 3637–3647.

[148] 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages.

[149] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2022. Evaluating Human-Language Model Interaction. *arXiv preprint arXiv:2212.09746* (2022).

[150] Florian Lehmann. 2023. Mixed-Initiative Interaction with Computational Generative Systems. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 501, 6 pages.

[151] Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". 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 135, 17 pages.

[152] Jingjing Li, Zichao Li, Tao Ge, Irwin King, and Michael Lyu. 2022. Text Revision by On-the-Fly Representation Optimization. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, Ting-Hao 'Kenneth' Huang, Vipul Raheja, Dongyeop Kang, John Joon Young Chung, Daniel Gissin, Mina Lee, and Katy Ilonka Gero (Eds.). Association for Computational Linguistics, Dublin, Ireland, 58–59.

[153] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4582–4597.

[154] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). Featured Certification, Expert Certification.

[155] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. arXiv:2306.01941 [cs.HC]

[156] Zhiyu Lin, Upol Ehsan, Rohan Agarwal, Samihan Dani, Vidushi Vashishth, and Mark Riedl. 2023. Beyond Prompts: Exploring the Design Space of Mixed-Initiative Co-Creativity Systems. *arXiv preprint arXiv:2305.07465* (2023).

[157] Ting Liu, Ming Zhou, Jianfeng Gao, Endong Xun, and Changning Huang. 2000. PENS: A Machine-Aided English Writing System for Chinese Users. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (Hong Kong) *(ACL '00)*. Association for Computational Linguistics, USA, 529–536.

[158] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI Console Me When I Lose My Pet? Understanding Perceptions of AI-Mediated Email Writing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 474, 13 pages.

[159] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[160] Yuanchao Liu, Bo Pang, and Bingquan Liu. 2019. Neural-based Chinese Idiom Recommendation for Enhancing Elegance in Essay Writing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5522–5526.

[161] Annie Louis and Ani Nenkova. 2013. What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of the Association for Computational Linguistics* 1 (2013), 341–352.

[162] Michal Lukasik and Richard Zens. 2018. Content Explorer: Recommending Novel Entities for a Document Writer. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3371–3380.

[163] Charles A MacArthur. 2006. The effects of new technologies on writing and writing processes. *Handbook of writing research* (2006), 248–262.

[164] N. H. Macdonald. 1983. Human factors and behavioral science: The UNIX™ Writer's Workbench software: Rationale and design. *The Bell System Technical Journal* 62, 6 (1983), 1891–1908.

[165] I Scott MacKenzie and Steven J Castellucci. 2016. Empirical research methods for human-computer interaction. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 996–999.

[166] Allan MacLean, Richard Young, Victoria Bellotti, and Thomas Moran. 1991. Questions, Options, and Criteria: Elements of Design Space Analysis. *Human-Computer Interaction* 6 (09 1991), 201–250.

[167] Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. EdiT5: Semi-Autoregressive Text Editing with T5 Warm-Start. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2126–2138.

[168] Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. FELIX: Flexible Text Editing Through Tagging and Insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1244–1255.

[169] Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, Tag, Realize: High-Precision Text Editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5054–5065.

[170] Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining* 16, 1 (2023), 20.

[171] Hannah Mieczkowski and Jeffrey Hancock. 2023. Examining Agency, Expertise, and Roles of AI Systems in AI-Mediated Communication. In *Computer-Supported Cooperative Work And Social Computing (CSCW)*.

[172] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781* (2013).

[173] Eleni Miltsakaki and Karen Kukich. 2000. The Role of Centering Theory's Rough-Shift in the Teaching and Evaluation of Writing Skills. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (Hong Kong) *(ACL '00)*. Association for Computational Linguistics, USA, 408–415.

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

[175] Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: a systematic review of recent research. *Journal of biomedical informatics* 52 (2014), 457–467.

[176] Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Towards Automated Document Revision: Grammatical Error Correction, Fluency Edits, and Beyond. arXiv:2205.11484 [cs.CL]

[177] Yusuke Mori, Hiroaki Yamane, Ryohei Shimizu, and Tatsuya Harada. 2022. Plug-and-Play Controller for Story Completion: A Pilot Study toward Emotion-aware Story Writing Assistance. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, Ting-Hao 'Kenneth' Huang, Vipul Raheja, Dongyeop Kang, John Joon Young Chung, Daniel Gissin, Mina Lee, and Katy Ilonka Gero (Eds.). Association for Computational Linguistics, Dublin, Ireland, 46–57.

[178] Meredith Ringel Morris, Carrie J Cai, Jess Holbrook, Chinmay Kulkarni, and Michael Terry. 2023. The design space of generative models. *arXiv preprint arXiv:2304.10547* (2023).

[179] Ryo Nagata. 2019. Toward a Task of Feedback Comment Generation for Writing Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3206–3215.

[180] Rosiana Natalie, Joshua Tseng, Hernisa Kacorri, and Kotaro Hara. 2023. Supporting Novices Author Audio Descriptions via Automatic Feedback. 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 77, 18 pages.

[181] Timothy Neate, Abi Roper, Stephanie Wilson, and Jane Marshall. 2019. Empowering Expression for Users with Aphasia through Constrained Creativity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12.

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

[183] Writers Guild of America. 2023. What We Won. Accessed: Dec. 12, 2023.

[184] David R Olson. 2002. What writing does to the mind. *Language, literacy, and cognitive development: The development and consequences of symbolic communication* (2002), 153–166.

[185] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

[186] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[187] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 19, 10 pages.

[188] Vishakh Padmakumar and He He. 2022. Machine-in-the-Loop Rewriting for Creative Image Captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 573–586.

[189] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity?. In *The Twelfth International Conference on Learning Representations*.

[190] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Association for Computational Linguistics (ACL)*.

[191] SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. 2021. "I Wrote as If I Were Telling a Story to Someone I Knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 926–941.

[192] Taehyun Park, Edward Lank, Pascal Poupart, and Michael Terry. 2008. Is the Sky Pure Today? AwkChecker: An Assistive Tool for Detecting and Correcting Collocation Errors. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) *(UIST '08)*. Association for Computing Machinery, New York, NY, USA, 121–130.

[193] Roy D Pea. 2018. The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. In *Scaffolding*. Psychology Press, 423–451.

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

[195] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15.

[196] Mike Perkins. 2023. Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice* 20, 2 (2023), 07.

[197] Ritika Poddar, Rashmi Sinha, Mor Naaman, and Maurice Jakesch. 2023. AI Writing Assistants Influence Topic Choice in Self-Presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 29, 6 pages.

[198] Washington Post. 2023. They fell in love with AI bots. A software update broke their hearts. Accessed: Feb. 24, 2024.

[199] Fanchao Qi, Yanhui Yang, Jing Yi, Zhili Cheng, Zhiyuan Liu, and Maosong Sun. 2022. QuoteR: A Benchmark of Quote Recommendation for Writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 336–348.

[200] Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 83–88.

[201] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving language understanding by generative pre-training*. Technical Report. OpenAI.

[202] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[203] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[204] Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text Editing by Task-Specific Instruction Tuning. (2023). arXiv:2305.09857 [cs.CL]

[205] Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One Document, Many Revisions: A Dataset for Classification and Description of Edit Intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 5517–5524.

[206] Christian Rapp, Otto Kruse, Jennifer Erlemann, and Jakob Ott. 2015. Thesis Writer: A System for Supporting Academic Writing. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW'15 Companion)*. Association for Computing Machinery, New York, NY, USA, 57–60.

[207] Marek Rei and Helen Yannakoudakis. 2016. Compositional Sequence Labeling Models for Error Detection in Learner Writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1181–1191.

[208] Machel Reid and Graham Neubig. 2022. Learning to Model Editing Processes. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3822–3832.

[209] Reuters. 2023. OpenAI, Microsoft hit with new US consumer privacy class action. Accessed: Dec 12, 2023.

[210] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 724, 18 pages.

[211] Melissa Roemmele and Andrew S. Gordon. 2018. Automated Assistance for Creative Writing with an RNN Language Model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (Tokyo, Japan) *(IUI '18 Companion)*. Association for Computing Machinery, New York, NY, USA, Article 21, 2 pages.

[212] Yvonne Rogers. 2012. *HCI Theory: Classical, Modern, and Contemporary*. Morgan & Claypool Publishers.

[213] D Gordon Rohman. 1965. Pre-writing the stage of discovery in the writing process. *College composition and communication* 16, 2 (1965), 106–112.

[214] K. Romer and F. Mattern. 2004. The design space of wireless sensor networks. *IEEE Wireless Communications* 11, 6 (2004), 54–61.

[215] Rod D Roscoe, Laura K Allen, Jennifer L Weston, Scott A Crossley, and Danielle S McNamara. 2014. The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition* 34 (2014), 39–59.

[216] John Sadauskas, Daragh Byrne, and Robert K. Atkinson. 2015. Mining Memories: Designing a Platform to Support Social Media Based Writing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3691–3700.

[217] Steve Sawyer and Mohammad Hossein Jarrahi. 2014. Sociotechnical approaches to the study of information systems. In *Computing handbook, third edition: Information systems and information technology*. CRC Press, 5–1.

[218] Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. PEER: A Collaborative Language Model. arXiv:2208.11663 [cs.CL]

[219] Denise Schmandt-Besserat. 1992. *Before writing, vol. I: from counting to cuneiform.* Vol. 1. University of Texas press.

[220] Oliver Schmitt and Daniel Buschek. 2021. CharacterChat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy) *(C&C '21)*. Association for Computing Machinery, New York, NY, USA, Article 10, 10 pages.

[221] David Schneider and Kathleen F. McCoy. 1998. Recognizing Syntactic Errors in the Writing of Second Language Learners. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2* (Montreal, Quebec, Canada) *(ACL '98/COLING '98)*. Association for Computational Linguistics, USA, 1198–1204.

[222] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).

[223] Shuming Shi, Enbo Zhao, Wei Bi, Deng Cai, Leyang Cui, Xinting Huang, Haiyun Jiang, Duyu Tang, Kaiqiang Song, Longyue Wang, Chenyan Huang, Guoping Huang, Yan Wang, and Piji Li. 2023. Effidit: An Assistant for Improving Writing Efficiency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations).* Association for Computational Linguistics, Toronto, Canada, 508–515.

[224] Antonette Shibani, Simon Knight, and Simon Buckingham Shum. 2020. Educator perspectives on learning analytics in classroom practice. *The Internet and Higher Education* 46 (2020), 100730.

[225] Ben Shneiderman. 2011. Claiming success, charting the future. *Interactions* 18, 5 (2011), 10–11.

[226] Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Canoee Liu, Simon Tong, Jindong Chen, and Lei Meng. 2023. RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting. arXiv:2305.15685 [cs.CL]

[227] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv:2305.17493 [cs.LG]

[228] Hrituraj Singh, Gaurav Verma, Aparna Garimella, and Balaji Vasan Srinivasan. 2021. DRAG: Director-Generator Language Modelling Framework for Non-Parallel Author Stylized Rewriting. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* Association for Computational Linguistics, Online, 863–873.

[229] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* (feb 2022). Just Accepted.

[230] Gabriella Skitalinskaya and Henning Wachsmuth. 2023. To Revise or Not to Revise: Learning to Detect Improvable Claims for Argumentative Writing Support. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Toronto, Canada, 15799–15816.

[231] Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards Evaluating Narrative Quality In Student Writing. *Transactions of the Association for Computational Linguistics* 6 (2018), 91–106.

[232] Hubert Soyer, Goran Topić, Pontus Stenetorp, and Akiko Aizawa. 2015. CroVeWA: Crosslingual Vector-Based Writing Assistance. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.* Association for Computational Linguistics, Denver, Colorado, 91–95.

[233] Felix Stahlberg and Shankar Kumar. 2020. Seq2Edits: Sequence Transduction Using Span-level Edit Operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Online, 5147–5159.

[234] Dorothy S Strickland and Lesley Mandel Morrow. 1989. *Emerging literacy: Young children learn to read and write.* ERIC.

[235] Xiaotian Su, Thiemo Wambsganss, Roman Rietsche, Seyed Parsa Neshaei, and Tanja Käser. 2023. Reviewriter: AI-Generated Instructions For Peer Review Writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.). Association for Computational Linguistics, Toronto, Canada, 57–71.

[236] Elizabeth Sulzby, William H Teale, and George Kamberelis. 1989. Emergent writing in the classroom: Home and school connections. *Emerging literacy: Young children learn to read and write* (1989), 63–79.

[237] Simeng Sun, Wenlong Zhao, Varun Manjunatha, Rajiv Jain, Vlad Morariu, Franck Dernoncourt, Balaji Vasan Srinivasan, and Mohit Iyyer. 2021. IGA: An Intent-Guided Authoring Assistant. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5972–5985.

[238] Yusen Sun, Liangyou Li, Qun Liu, and Dit-Yan Yeung. 2023. SongRewriter: A Chinese Song Rewriting System with Controllable Content and Rhyme Scheme. In *Findings of the Association for Computational Linguistics: ACL 2023.* Association for Computational Linguistics, Toronto, Canada, 12863–12880.

[239] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS).* 3104–3112.

[240] Haruya Suzuki, Sora Tarumoto, Tomoyuki Kajiwara, Takashi Ninomiya, Yuta Nakashima, and Hajime Nagahara. 2022. Emotional Intensity Estimation based on Writer's Personality. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop.* Association for Computational Linguistics, Online, 1–7.

[241] Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story Centaur: Large Language Model Few Shot Learning as a Creative Writing Tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics, Online, 244–256.

[242] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

[243] Rohan Taori and Tatsunori Hashimoto. 2023. Data feedback loops: Model-driven amplification of dataset biases. In *International Conference on Machine Learning.* PMLR, 33883–33920.

[244] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv* (2023).

[245] Eric L Trist. 1981. *The evolution of socio-technical systems.* Vol. 2. Ontario Quality of Working Life Centre Toronto.

[246] Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S. Chang. 2020. LinggleWrite: a Coaching System for Essay Writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics, Online, 127–133.

[247] Selen Türkay, Daniel Seaton, and Andrew M. Ang. 2018. Itero: A Revision History Analytics Tool for Exploring Writing Behavior and Reflection. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6.

[248] Keith Vertanen, Haythem Memmi, Justin Emge, Shyam Reyal, and Per Ola Kristensson. 2015. VelociTap: Investigating Fast Mobile Text Entry using Sentence-Based Decoding of Touchscreen Keyboard Input. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* Association for Computing Machinery, Seoul, Republic of Korea, 659–668.

[249] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 683, 13 pages.

[250] Thiemo Wambsganss and Christina Niklaus. 2022. Modeling Persuasive Discourse to Adaptively Support Students' Argumentative Writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Dublin, Ireland, 8748–8760.

[251] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting Cognitive and Emotional Empathic Writing of Students. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Association for Computational Linguistics, Online, 4063–4077.

[252] Thiemo Wambsganss, Matthias Soellner, Kenneth R Koedinger, and Jan Marco Leimeister. 2022. Adaptive Empathy Learning Support in Peer Review Scenarios. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 227, 17 pages.

[253] Ruyuan Wan, Naome Etori, Karla Badillo-urquiola, and Dongyeop Kang. 2022. User or Labor: An Interaction Framework for Human-Machine Relationships in NLP. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances).* Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 112–121.

[254] Chenshuo Wang, Shaoguang Mao, Tao Ge, Wenshan Wu, Xun Wang, Yan Xia, Jonathan Tien, and Dongyan Zhao. 2023. Smart Word Suggestions for Writing Assistance. In *Findings of the Association for Computational Linguistics: ACL 2023.* Association for Computational Linguistics, Toronto, Canada, 11212–11225.

[255] Liuping Wang, Xiangmin Fan, Feng Tian, Lingjia Deng, Shuai Ma, Jin Huang, and Hongan Wang. 2018. MirrorU: Scaffolding Emotional Reflection via In-Situ

Assessment and Interactive Feedback. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6.

[256] Weibo Wang, Abidalrahman Moh'd, Aminul Islam, Axel Soto, and Evangelos Milios. 2016. Non-uniform Language Detection in Technical Writing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1892–1900.

[257] Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama, and Masataka Goto. 2017. LyriSys: An Interactive Support System for Writing Lyrics Based on Topic Transition. In *Proceedings of the 22nnd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 559–563.

[258] Florian Weber, Thiemo Wambsganss, Seyed Parsa Neshaei, and Matthias Soellner. 2023. Structured Persuasive Writing Support in Legal Education: A Model and Tool for German Legal Case Solutions. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 2296–2313.

[259] Florian Weber, Thiemo Wambsganss, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical agents for interactive learning: A taxonomy of conversational agents in education. In *Forty-Second International Conference on Information Systems*. Austin, Texas.

[260] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903* (2022).

[261] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, El-lie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Gi-ada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdul-mumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joy-deep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Ville-gas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muham-mad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, De-bajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Sai-ful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jeka-terina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav

Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Are-zoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Ben-jamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Ed-ward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shub-ber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémen-tine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjava-cas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchan-dani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shub-hanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100 [cs.CL]

[262] Shaomei Wu, Lindsay Reynolds, Xian Li, and Francisco Guzmán. 2019. Design and Evaluation of a Social Media Writing Support Tool for People with Dyslexia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[263] Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, and Bill Dolan. 2021. Automatic Document Sketching: Generating Drafts from Analogous Texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2102–2113.

[264] Qiongkai Xu, Chenchen Xu, and Lizhen Qu. 2019. ALTER: Auxiliary Text Rewriting Tool for Natural Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Association for Computational Linguistics, Hong Kong, China, 13–18.

[265] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying Semantic Edit Intentions from Revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Lin-guistics, Copenhagen, Denmark, 2000–2010.

[266] Chin yew Lin and Marina Rey. 2004. Looking for a Few Good Metrics: ROUGE and its Evaluation. In *NTCIR Workshop*.

[267] Seid Muhie Yimam and Chris Biemann. 2018. Demonstrating Par4Sem - A Semantic Writing Aid with Adaptive Paraphrasing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demon-strations*. Association for Computational Linguistics, Brussels, Belgium, 48–53.

[268] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852.

[269] 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 Hu-man Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <coun-try>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages.

[270] Niloofar Zarei, Sharon Lynn Chu, Francis Quek, Nanjie'Jimmy' Rao, and Sarah Anne Brown. 2020. Investigating the Effects of Self-Avatars and Story-Relevant Avatars on Children's Creative Storytelling. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[271] Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016. Ar-gRewrite: A Web-based Revision Assistant for Argumentative Writings. In *Pro-ceedings of the 2016 Conference of the North American Chapter of the Association*

for Computational Linguistics: Demonstrations. Association for Computational Linguistics, San Diego, California, 37–41.

[272] Fan Zhang and Diane Litman. 2016. Using Context to Predict the Purpose of Argumentative Writing Revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, San Diego, California, 1424–1430.

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

[274] Xuchao Zhang, Dheeraj Rajagopal, Michael Gamon, Sujay Kumar Jauhar, and ChangTien Lu. 2019. Modeling the Relationship between User Comments and Edits in Document Revision. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5002–5011.

[275] Wenjie Zhong, Jason Naradowsky, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2023. Fiction-Writing Mode: An Effective Control for Human-Machine Collaborative Writing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, Dubrovnik, Croatia, 1752–1765.

[276] Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2023. Visualize Before You Write: Imagination-Guided Open-Ended Text Generation. In *Findings of the Association for Computational Linguistics: EACL 2023.* Association for Computational Linguistics, Dubrovnik, Croatia, 78–92.

[277] Gustavo Zomer and Ana Frankenberg-Garcia. 2021. Beyond Grammatical Error Correction: Improving L1-influenced research writing in English using pretrained encoder-decoder models. In *Findings of the Association for Computational Linguistics: EMNLP 2021.* Association for Computational Linguistics, Punta Cana, Dominican Republic, 2534–2540.

A   AUTHOR CONTRIBUTIONS

This project was a large collaboration with 36 researchers across 27 institutions. This team effort was built on countless contributions from everyone involved. To acknowledge individual authors' contributions and enable future inquiries to be directed appropriately, we listed authors in three different ways in our paper and artifact.

A.1   Overall Author List

In the beginning of the project, each author signed up for one of the four roles in the project. Project leads oversaw the entire project, supporting team leads and members. Team leads kept team members on track, provided feedback on literature review and writing, and maintained alignment with the project's direction. Team members contributed to decision making, conducted extensive literature review, and wrote the paper. Advisors, although sometimes not directly involved in literature review and writing, provided additional guidance and feedback. Some authors took on two roles, occasionally blurring the role distinctions. The following list contains each author's name, affiliation, and contributions, grouped by their main self-assigned role.

Project leads

- **Mina Lee** (University of Chicago & Microsoft Research): Led and managed the overall project, prepared weekly project meetings, filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), participated in writing (initial, revision), designed figures, and open-sourced the artifact
- **Katy Ilonka Gero** (Harvard University): Led and managed the overall project, led the user team, created and managed resources for coding papers, filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), and participated in writing (initial, revision)
- **John Joon Young Chung** (Midjourney): Led the systematic literature review process, sampled papers, filtered papers, designed dimensions and codes (initial), coded papers (initial), participated in writing (initial), and analyzed annotations

Team leads (alphabetical)

- **Simon Buckingham Shum** (University of Technology Sydney): Led the ecosystem team, prepared weekly team meetings, filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), and participated in writing (initial, revision)
- **Vipul Raheja** (Grammarly): Led the technology team, prepared weekly team meetings, sampled extra papers, filtered papers, designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial, revision)
- **Hua Shen** (University of Michigan): Led the interaction team, prepared weekly team meetings, filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), and participated in writing (initial, revision)
- **Subhashini Venugopalan** (Google): Led the technology team, sampled extra papers, filtered papers, designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial)

- **Thiemo Wambsganss** (Bern University of Applied Sciences): Led the interaction team, prepared weekly team meetings, filtered papers, designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial)
- **David Zhou** (University of Illinois, Urbana-Champaign): Led the task team, prepared weekly team meetings, filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), and participated in writing (initial, revision)

Team members (alphabetical)

- **Emad A. Alghamdi** (King Abdulaziz University): Filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), and participated in writing (initial)
- **Tal August** (Allen Institute for AI): Designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial)
- **Avinash Bhat** (McGill University): Filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), identified issues in papers and coding results, and participated in writing (initial)
- **Madiha Zahrah Choksi** (Cornell Tech): Filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), and participated in writing (initial)
- **Senjuti Dutta** (University of Tennessee, Knoxville): Filtered papers, designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial)
- **Jin L.C. Guo** (McGill University): Designed dimensions and codes (initial, revision), and coded papers (initial, revision), and participated in writing (initial)
- **Md Naimul Hoque** (University of Maryland, College Park): Filtered papers, designed dimensions and codes (initial, revision), and coded papers (initial, revision)
- **Yewon Kim** (KAIST): Filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), participated in writing (initial, revision), and identified issues in papers and coding results
- **Simon Knight** (University of Technology Sydney): Designed dimensions and codes (initial, revision)
- **Seyed Parsa Neshaei** (EPFL): Filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), and participated in writing (initial, revision)
- **Agnia Sergeyuk** (JetBrains Research): Filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), participated in writing (initial, revision), and identified issues in papers and coding results
- **Antonette Shibani** (University of Technology Sydney): Designed dimensions and codes (initial, revision) and coded papers (revision)
- **Disha Shrivastava** (Google DeepMind): Designed dimensions and codes (initial) and coded papers (initial)
- **Lila Shroff** (Stanford University): Filtered papers, designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial)

- **Jessi Stark** (University of Toronto): Filtered papers, designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial)
- **Sarah Sterman** (University of Illinois, Urbana-Champaign): Filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), participated in writing (initial, revision), designed interaction framework, and analyzed annotations
- **Sitong Wang** (Columbia University): Filtered papers, designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial)

Advisors (alphabetical)

- **Antoine Bosselut** (EPFL): Filtered papers, coded papers (initial), participated in writing (initial)
- **Daniel Buschek** (University of Bayreuth): Filtered papers, designed dimensions and codes (initial, revision), coded papers (initial, revision), and participated in writing (initial, revision)
- **Joseph Chee Chang** (Allen Institute for AI): Filtered papers, designed dimensions and codes (initial), coded papers (initial), and participated in writing (initial)
- **Sherol Chen** (Google): Designed dimensions and codes (initial)
- **Max Kreminski** (Midjourney): Filtered papers and designed dimensions and codes (initial)
- **Joonsuk Park** (University of Richmond): Sampled extra papers, designed dimensions and codes (initial), and participated in writing (initial)
- **Roy Pea** (Stanford University): Designed dimensions and codes (initial) and participated in writing (initial, revision)
- **Eugenia H. Rho** (Virginia Tech): Designed dimensions and codes (initial) and participated in writing (initial, revision)
- **Shannon Zejiang Shen** (Massachusetts Institute of Technology): Designed dimensions and codes (initial), participated in writing (initial), designed tables, and ideated and open-sourced the artifact
- **Pao Siangliulue** (B12): Filtered papers, designed dimensions and codes (initial), and participated in writing (initial)

A.2 Team-Specific Author Lists

As described in Section 3.3.2, the authors were split into five teams to develop team-specific dimensions and codes based on the five aspects. Each team operated as its own project group in that most teams had separate weekly meetings. Below, we specify a team-specific author list for each team. The ordering follows the convention in Computer Science (e.g., the first person in the team member group has the most contribution and the last person in the advisor group has the most contribution within each group).

- **Task**: David Zhou (team lead), Agnia Sergeyuk (team member), Jessi Stark (team member), Emad A. Alghamdi (team member), Sitong Wang (team member), Roy Pea (advisor)
- **User**: Katy Ilonka Gero (team lead), Yewon Kim (team member), John Joon Young Chung (team member), Senjuti Dutta (team member), Lila Shroff (team member), Disha Shrivastava (team member), Eugenia H. Rho (advisor)

- **Technology**: Vipul Raheja (team lead), Subhashini Venugopalan (team lead), Seyed Parsa Neshaei (team member), Disha Shrivastava (team member), Antoine Bosselut (advisor), Sherol Chen (advisor), Joonsuk Park (advisor)
- **Interaction**: Hua Shen (team lead), Thiemo Wambsganss (team lead), Sarah Sterman (team member), Tal August (team member), Avinash Bhat (team member), Md Naimul Hoque (team member), Jin L.C. Guo (team member), Pao Siangliulue (advisor), Joseph Chee Chang (advisor), Max Kreminski (advisor), Shannon Zejiang Shen (advisor), Daniel Buschek (advisor)
- **Ecosystem**: Simon Buckingham Shum (team lead), Madiha Zahrah Choksi (team member), Antonette Shibani (team member), Simon Knight (team member)

A.3   Core Group of Annotators

Most authors coded papers as part of designing and refining dimensions and codes. During this process, they focused on the team-specific dimensions and codes and looked at a subset of the papers that were relevant to their teams. After several iterations within each team, we created an initial version of the design space. Then, a subset of the authors volunteered to form the core group of annotators and coded all papers for all dimensions and codes (beyond their own teams). This relatively small group allowed us to be more efficient and reduced communication overhead. Here, we list the authors who spearheaded and annotated the papers as part of this core group in the order of their contributions (i.e., the first person annotated the highest number of papers) as well as the authors who helped with creating the living artifact.

- **Annotators**: Avinash Bhat, Simon Buckingham Shum, Agnia Sergeyuk, Yewon Kim, David Zhou, Emad A. Alghamdi, Jin L.C. Guo, Seyed Parsa Neshaei, Hua Shen, Md Naimul Hoque, Madiha Zahrah Choksi, Katy Ilonka Gero, Sarah Sterman, Antonette Shibani, Mina Lee
- **Artifact designers**: Shannon Zejiang Shen, Mina Lee

B   TERMINOLOGY

Throughout the paper, we use "intelligent and interactive writing assistants" and "writing assistants" interchangeably. Here, we describe the distinctions we make between "writing assistants" and other related terms: "models," "systems," and "technology". Firstly, we use "writing assistants" to refer to computational systems that assist users with their writing. These writing assistants must have the frontend that can interface with users, whereas the other three terms do not. We use "model" to refer to a specific model (e.g., a specific instance of GPT-3.5, such as gpt-3.5-turbo). We use "system" quite broadly in the paper. Beyond using it as a concise way to refer to writing assistants, it can refer to 1) a model (but in this case, we prefer to say "model" to be more specific), 2) a model + alpha (e.g., ChatGPT which is GPT-3.5 with extra safety filters on top, a tool-augmented LM where a model has access to external resources), and 3) anything that is not a model-based system (e.g., rule-based system). The use of "system" in Section 4.4 is an example of the second case. We use "technology" as a much broader concept than "model" or "system" in that it incorporates data, model, learning, evaluation, and beyond.

C   ADDITIONAL BACKGROUND

C.1   Technological Evolution in Writing Assistants

Since the model architectures used to learn patterns have evolved hand-in-hand to consume and capture patterns in increasing amounts of data, we discuss these two aspects together here. A number of works on writing assistants in the early 2000s [26, 157, 192, 221] used purely human-labeled data to develop rule-based methods [26] or train statistical models [10, 192], which were often used to detect errors [26, 192] or suggest corrections. These models were trained on much smaller datasets consisting of hundreds or a few thousands of examples. With developments in model architecture such as statistical ML models and deep neural networks [51, 172], models behind writing tools could take advantage of larger sources of unlabeled data. For instance, some work started using learned word embeddings or trained embeddings specifically for a task [56, 84, 267]. They were then able to bootstrap off this to use smaller human-labeled datasets for further modeling and evaluation. With deep sequence to sequence models [42, 239] writing assistants were able to use a combination of human-labeled and machine-labeled data with several thousand sentences or examples [160, 162]. Several recent works [35, 45, 58, 65, 174, 188, 268] take advantage of the large language models that are already pre-trained on corpora of millions of sentences. They are then able to use small amounts of human-labeled data to tune the model (e.g., instruction tune [204, 218, 226]) and in many cases simply prompt the model in a zero-shot manner to help with the writing task.

C.2   User Interaction & Interface Evolution in Writing Assistants

The origin of generating text to support writers lies in augmentative and alternative communication (AAC) research. The goal of these text entry methods is to reduce manual typing, in particular, for people with motor impairments. This was typically realized by predicting next words, and showing them in the user interface for people to select directly, saving letter-by-letter input efforts [76, 102]. These ideas were later applied more generally to improve input efficiency [142]. Here we see two interlinked developments: On the technical side, systems evolved from using simple n-gram models to deep learning with today's LMs. This enabled coherent, longer generation as well as the recent prompting paradigm. This in turn shaped the interaction and UI design.

Generated text shown in the UI can now be longer, evolving from single words [67, 76, 91, 200], to phrases [8, 27, 41, 148, 248], to whole drafts (e.g., based on keywords or an incoming email [127]), depending on use cases. Relatedly, interactions for controlling text generation have become much more varied. As one key distinction [58], users can implicitly steer text completion with their preceding draft text, or write explicit instructions to the system. This instruction style is currently often combined with a user interface design—a chat history with the system (e.g., ChatGPT [185])—that differs from the traditional page view. Other emerging UI patterns in this context include sidebars for suggestions [268] or annotations [57], as well as other separate views and tools for engaging with AI text [84, 85]. In other designs, AI-generated text is directly

integrated into the user's writing area (e.g., previewed in a light grey font color [41]) or can be selected from a pop-up list at the cursor [27, 58, 148]. Beyond linear text, further UI concepts include sketching [45] and views that arrange (generated) text on a 2D canvas, for example, as a graph [125] or post-it notes.[8]

C.3 Ecosystem: Going from Micro-HCI to Macro-HCI

Historical accounts of HCI [212, 225] document how the definition of "the system" was first enlarged from the computer to include the human user, starting with attempts to apply psychological models of the individual human using a computer. This frame then expanded to include the ways in which the wider sociotechnical system impinged on (and was shaped by) how interacting groups of people appropriated technology, requiring theory and methods from many more disciplines such as design, ecological psychology, sociology, anthropology, linguistics, critical theory and more, to create today's HCI landscape. Borrowing an art history metaphor, Rogers [212] characterizes this as the evolution of HCI theory from classical, to modern to contemporary, while Shneiderman [225] uses the language of micro-HCI and macro-HCI:

"Micro-HCI researchers and developers design and build innovative interfaces and deliver validated guidelines for use across the range of desktop, Web, mobile, and ubiquitous devices. The challenges for micro-HCI are to deal with rapidly changing technologies while accommodating the wide range of users." [...] "Macro-HCI researchers and developers design and build interfaces in expanding areas, such as affective experience, aesthetics, motivation, social participation, trust, empathy, responsibility, and privacy. [...] Macro-HCI researchers have to face the challenge of more open tasks, unanticipated user goals, new measures of system efficacy, and even conflicts among users in large communities."

In these terms, the design space relating to the task, user, technology, and interaction aspects describes the *micro-HCI* level, but as writing assistants become embedded in broader sociotechnical contexts, this must extend to *macro-HCI* concerns beyond the individual writer and software, to what we term the *ecosystem*.

D SYSTEMATIC LITERATURE REVIEW

D.1 Venues

To keep the number of papers reasonable, we decided to focus on papers from the following venues. We included all their paper tracks (e.g., CHI Late-Breaking Work and ACL Findings), but excluded workshops.

- **HCI**: CHI, CSCW, UIST, IUI, C&C, DIS, and ToCHI
- **NLP**: ACL, NAACL, EMNLP, EACL, and TACL

D.2 Keywords

When retrieving candidate papers from ACM DL, we simply used "write" as our keyword, since ACM DL supports automatic matching of variations. Any paper that has the keyword in its title or its keywords was retrieved. When retrieving candidate papers from ACL Anthology, we used "writ" and "wrote" as our keywords to manually account for the verb "write"'s variations ("write," "writes,"

"writing," "wrote," and "written"). Because papers on ACL Anthology do not have associated keywords, we retrieved any paper that has at least one of the keywords in its title. Despite the differences, we retrieved the similar number of papers: 60 papers in HCI and 55 papers in NLP.

D.3 Collected Papers

- **HCI**: [2, 7, 10, 12, 13, 15, 16, 18, 28, 34, 45, 47, 57, 58, 64, 71, 80, 83–86, 89, 90, 106, 108, 109, 114, 115, 121, 122, 126, 129, 131, 133, 134, 148, 150, 158, 174, 180, 181, 187, 191, 192, 195, 197, 206, 210, 211, 216, 220, 229, 247, 252, 255, 257, 262, 268]
- **NLP**: [3, 9, 11, 26, 29, 35, 37, 38, 43, 48, 54, 56, 68, 77, 77, 88, 93, 96–98, 105, 111, 112, 118, 124, 138, 146, 157, 160–162, 173, 179, 188, 199, 207, 221, 223, 228, 230–232, 238, 241, 246, 250, 251, 254, 258, 264, 267, 271, 272, 275–277]

D.4 Additional References for Technology

As described in Section 3.3.2, the technology team selected 25 additional papers to ensure a broader, deeper, and more relevant coverage of recent technologies. The paper selection process was identical to the one in Section 3.3, but with an expanded set of search keywords, such as "text revision" and "text editing," that are more likely to appear in NLP papers. Furthermore, the team retrieved papers from an expanded set of venues. The initial set was 80 papers, which were then deduplicated based on the common pool, and then filtered and adjudicated by two authors based on their relevance to the technology aspect. Finally, the team referenced the papers chosen by both the authors as relevant to the aspect, leading to a set of 25 papers. Note that some of these papers were considered as out of scope based on the criteria in Section 3.1, but were still relevant as they are concerned with AI models built for writing-related tasks (e.g., LMs fine-tuned for specific writing tasks, such as text composition or revision). The team used these papers to develop their codes, but did not include them as part of their literature review.

- **Technology**: [4, 6, 59, 66, 70, 72, 119, 120, 137, 152, 167–169, 176, 177, 204, 205, 208, 218, 226, 233, 235, 237, 263, 265, 274]

---
[8]https://fermat.app/