



SYNthia: An Interface Concept for Writing With Large Language Models

Citation

Liang, Ivy. 2025. SYNthia: An Interface Concept for Writing With Large Language Models. Bachelors Thesis, Harvard University Engineering and Applied Sciences.

Link

<https://dash.harvard.edu/handle/1/42719272>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

SYNthia: An Interface Concept for Writing With Large Language Models

A thesis submitted to the
Department of Computer Science
in partial fulfillment of the requirements
for the degree of
Bachelor of Arts with Honors
in the subject of
Computer Science

by

Ivy Liang

Supervisor: Elena Leah Glassman
Co-supervisor: Katy Ilonka Gero

Harvard College
Cambridge, Massachusetts

March 2025

ACKNOWLEDGMENTS

This thesis is a cumulation of my entire Harvard College experience, and so I thank everyone who has supported me through this experience.

First, I would like to thank Professor Elena Leah Glassman and Dr. Katy Ilonka Gero for serving as my supervisors for this thesis. Taking and teaching CS279R / CS2790R with you both introduced me to the field of HCI, and the final project for that course ultimately became this thesis project. Your influences and teachings have truly changed my research trajectory and made me a more responsible programmer.

Second, I would like to thank my family for their constant reassurance, support, and presence in my life. Not only did you support me through this thesis, but through all of undergraduate (and my life leading up to this point). None of my accomplishments today would have been possible without your presence, and this thesis is only one example of this.

Finally, I would like to thank Alexander Joaquín Estevez. Alex was there from the beginning of this thesis, back when it was just a class project, helping me prototype the system and run the pilot studies. Thank you for trusting me with this project and I hope you like how the system ended up.

TABLE OF CONTENTS

Acknowledgments	i
Table of Contents	ii
List of Figures	v
List of Tables	vi
Abstract	vii
1 Introduction	1
2 Literature Review	5
2.1 Importance of Human Voice in Writing	5
2.1.1 Writing as a Vessel for Authentic Self Expression	5
2.1.2 The Impact of Revision	6
2.2 Novice vs. Expert Writers	6
2.2.1 Cognitive Processes	6
2.2.2 Text Production, Integration, and Fluency	8
2.2.3 Metacognition	8
2.3 Large Language Models	9
2.4 System-Assisted Writing	9
2.5 The Good, the Bad, and the Ugly of System-Assisted Writing	11
2.6 Efficient Error Correction	12
3 Prototype and Pilot	14
3.1 SYNthia's Prototype Description	14
3.1.1 Invoking and supplying suggestions	14
3.1.2 Adjusting number of suggestions provided per generation	15
3.2 Design Space	16
3.3 Pilot User Study	17
3.3.1 Trial 1: Control	18
3.3.2 Trial 2: Experimental	18

3.4	Pilot Findings	18
3.4.1	Trial 1: Control	18
3.4.2	Trial 2: Experimental	19
3.4.3	Additional Findings	20
4	System Overview	22
4.1	Design Goals	22
4.2	SYNthia	23
4.2.1	Changes from Prototype	24
4.3	System Architecture	25
5	User Study	26
5.1	Participants	26
5.1.1	Recruitment	26
5.1.2	Background Information	27
5.2	Methodology	28
5.2.1	Pre-Interview	29
5.2.2	Play Test with SYNthia	30
5.2.3	Post-Interview	30
5.3	Results	30
5.3.1	Interactions with SYNthia	30
5.3.2	Opinions of AI in Writing	32
5.3.3	Qualitative Analysis: Overview of SYNthia	35
5.3.4	Qualitative Analysis: Context Feature	36
6	Discussion	39
6.1	Research Questions	39
6.1.1	Usage of LLM Powered Thesaurus (RQ1)	39
6.1.2	Accepting Suggestions (RQ2)	40
6.1.3	Thought Process (RQ3)	41
6.2	Design Goals	43
6.2.1	Coherence to Context (DG1)	43
6.2.2	Steerability (DG2)	43
6.2.3	Divergent Outcomes (DG3)	44
6.3	Additional Factors that Contribute to AI-Writer Interactions	44
6.3.1	Contextual Framing	44
6.3.2	Task Specific Needs	45
6.4	Limitations	46
6.4.1	Experimental Design	46
6.4.2	Overreliance on Pretrained GPT-4 Model	46

6.5	Future Works	47
6.5.1	Additional Interface Features	47
6.5.2	Manual vs. Automatic Triggers	48
6.5.3	Integration with Existing Document Editors	48
7	Conclusion	49
	Appendix A - Pilot Prompts and User Study	51
	Prompts	51
	Trial 1: Control	51
	Trial 2: Experimental	51
	GPT 4 Prompt	51
	User Study	51
	Pre-Interview Questions - Required	51
	Post-Interview Questions - Required	52
	Post-Interview Question - Optional Conversations	52
	Appendix B - Formal Prompts and User Study	53
	Prompts	53
	Writing Prompt	53
	GPT 4 Prompt	53
	User Study	53
	Participant Recruitment Survey Questions - Required	53
	Pre-Interview Questions - Required	54
	Post-Interview Questions - Required	54
	Bibliography	56

LIST OF FIGURES

FIGURE	Page
2.1 Flowchart of how user prompts help the LLM answer specific questions or demands.	10
3.1 Prototype Interface of SYNthia. (1) User input textbox. (2) Word marking with '@' symbol. (3) Suggestion box. (4) Generated list of suggestions. (5) Input for additional context from user. (6) Slider determining number of suggestions provided per generation. (7) Word counter.	15
3.2 A QOC representation of the design space for SYNthia, including a map of options and corresponding criteria. Boxed options indicate choices made in the design of SYNthia.	16
4.1 Updated interface of SYNthia. On the left, users can create/edit their writing in the text box. A word counter is provided on the bottom left. SYNthia can be prompted with the @ symbol, which triggers the suggestion box to pop up on the right. Users can toggle the number of suggestions with the slider on the top right corner, as well as interact with the suggestion boxes by providing additional context for re-prompts. Additional measurements are collected via a dual timer system (top left corner) and a live tracker of prompted words, context given, and word chosen through a JSON file (top right corner).	23
4.2 SYNthia system architecture	25
5.1 Participant editing style, need for alternative words, and current writing aids . . .	27
5.2 Distribution of participant AI skills (self-rated on a scale from one to ten)	28
5.3 Words prompts and percent of words accepted per participant, aggregated by writing skill (novices vs. experts)	31
5.4 Time spent writing and time spent engaging with SYNthia per participant, aggregated by writing skill (novices vs. experts)	32

LIST OF TABLES

TABLE	Page
2.1 Comparison of Novice and Expert Characteristics	7
2.3 Strengths of ChatGPT in Writing	12
2.4 Weaknesses of ChatGPT in Writing	12
5.1 Participant Background on Writing and AI Skills	29
5.2 Representative sample of contexts given by participants	36

ABSTRACT

Artificial intelligence (AI)-infused systems can offer valuable assistance to writers, but they may also produce imperfect or unsatisfactory suggestions that require efficient correction. Word choice presents a challenge for writers that can be addressed by several tools, but these systems typically require users to switch browser tabs or tools and break their flow of thinking, or otherwise fail to incorporate the context associated with users' writing or their intentions, leaving them with subpar or unrelated suggestions. We present SYNthia, a word-suggestion interface that allows users to be directly involved in the suggestion generation process by providing natural language feedback. We performed two pilot qualitative studies, finding that SYNthia provided users with a more practical interface that (1) allowed them to receive their target word more efficiently, (2) eliminated the need to for users switch contexts (e.g. switching tabs or devices), and (3) improved users' perceived quality of writing. In addition, we performed a formal user study comparing how novice and expert writers interact with SYNthia differently, ultimately concluding that the writing level had no quantitatively significant impact on interactions with the tool, raising more questions for further study. However, the qualitative study surfaced several interesting observations regarding how writers interact with an AI-powered thesaurus, making progress towards the greater goal of integrating AI in the writing process while maintaining human agency and ownership. All code for this project can be found at the Github repository: <https://github.com/AEst2002/word-suggester/tree/thesis>.

INTRODUCTION

In 2023, Canadian novelist, essayist and "enemy of boredom", Stephen Marche published the first AI generated novel (using ChatGPT, Sudowrite and Cohere) reviewed in The New York Times, appropriately titled *Death of an Author*. With the novel being 95% AI generated, the opinions on this novel-writing process are quite controversial. While some call this product a "groundbreaking experiment in the meta world of man meets machine" [1], others label the usage of AI in this novel "a tumor through the spine of their other abiding freakouts" [2].

But was it a good novel? In The New York Times review, staff writer Dwight Garner observed that the AI attempted to incorporate figurative language into the novel, though it was "a hit-and-mostly-miss. ("The smell of coffee was like a fog burning off a field.')." In addition, he noted the concerning difficulty to discern between human and AI contributions, highlighting the threats of ownership to authors. While Garner was mostly ambivalent towards the usage of AI in writing, Marche seemingly had a positive impression of AI, "convincingly [making] the case that writers will manipulate A.I. the way that hip-hop producers dig up and arrange samples" [2]. Marche brings up a valid comparison: if singers like Ava Max can release songs that heavily sample or remake other pieces of music [3], why can't authors sample from AI generated works to help their own writing process?

While large language models (LLMs) demonstrate remarkable versatility across diverse domains, they are not reliable for all tasks – ultimately, humans are at the center of all AI interactions. First, we visit the capabilities of AI. Whether it be graduate school examinations or chatbots that are able to answer high stakes medical questions [4, 5], LLMs have proven themselves a fierce competitor to their human counterparts, a battle seen across different fields, including writing. Modern models are able to process complex inputs spanning text, code [6], and visual data [4]. However, while these models excel at pattern recognition and information synthesis, they face limitations in open-ended creativity and contextual judg-

ment, necessitating hybrid human-AI workflows for optimal results. Although some writers believe AI can enhance individual creativity, research indicates that its outputs become thematically repetitive over time, ultimately reducing diversity [7]. Furthermore, AI models inherit biases present in human data [8]. These limitations highlight the need for human oversight, reinforcing the importance of human-AI collaboration.

This project focuses on the integration of LLMs and AI in writing tools, which has also become increasingly prevalent, with major platforms like Grammarly [9, 10], Google, and Power Thesaurus incorporating AI-powered features to enhance user experience and productivity. This trend has been further accelerated by the development of LLMs, which has significantly expanded AI's writing capabilities [11]. LLMs are able to remember a subset of its past interactions with the user [12], giving them additional context to adapt to user prompts given within a specific context window. Recent studies have also shown that AI writing assistants can significantly improve writing efficiency and quality. For instance, workers using AI tools could complete writing tasks more quickly, reducing time spent on synthesizing material from 60% to just 10% of their workload. In that same study, researchers found that documents produced with AI support received significantly higher quality ratings (4.5 vs 3.8 on a 7-point scale) compared to non-assisted writing [13].

More specifically, AI has increasingly become a tool in creative writing, sparking debates on its role in the artistic process. While AI can assist with brainstorming, refining language, and enhancing stylistic choices [14], its use also raises concerns about originality [15], authorship [16], and over-reliance [17] on machine-generated content. Some are worried that novice writers may lean on AI to overcome writer's block or enhance their vocabulary, while experienced writers might use it more selectively to refine and experiment with language. The distinction between AI as a collaborative tool and AI as a crutch could vary based on the writer's skill level, creative confidence, and personal philosophy on authorship. In 2023, a paper was published detailed the advantages and disadvantages of using AI in writing: AI enhances writing by overcoming writer's block, generating creative analogies, and refining key sections like titles and conclusions. It aids in discovering overlooked references, structuring complex topics, ensuring thorough coverage, and presenting unfamiliar concepts in an accessible way. However, AI can also encourage over-reliance, replacing critical thinking and thorough research. It risks producing generic, uninspired work, enabling users to bypass reading essential sources, and may generate inaccurate or outdated information.

The line between using AI as a tool and having AI take over the writing can be easily crossed. Although writers (degree holders [18] and students [19]) agree that AI should be incorporated with human oversight, the disagreement arises about where to draw the line between human contribution and AI assistance. In our project, we investigate the question of how writers of different skill levels would interact with AI differently, and how their opinions differ on the usage of AI in their works. To make sense of this gray area, we present SYNthia, an AI-powered thesaurus that enables users to refine suggestions by adding con-

text (much like the chat functionality in ChatGPT or Claude), while maintaining confidence of writer ownership, even after using the tool. In the process of designing, building, and testing SYNthia, we aim to explore the following research questions:

- RQ1: How do novice and experienced writers use a large language model (LLM) powered thesaurus in creative writing differently?
- RQ2: How open are novice and experienced writers to accepting LLM generated word suggestions in their writing?
- RQ3: How do LLM generated word suggestions affect the thought processes of novice and experienced writers?

SYNthia is a browser-based AI-powered thesaurus that combines a traditional writing interface, thesaurus, and OpenAI’s GPT-4 model. The system allows users to write at their own pace, while manually prompting the AI thesaurus on the same page, allowing users to easily integrate AI suggestions back into their draft. SYNthia also allows users to provide context to the thesaurus as a way to refine the list of synonym suggestions.

We evaluated SYNthia with a between subjects user study, comparing novice and expert writers. Participants were asked to write to a creative writing prompt with the assistance of SYNthia and answer questions about their interactions. On average, participants requested alternative suggestions for 9 words during their writing process (average of 228 words written during the 15 minute time block), with an average 44% acceptance rate. When evaluating their draft (written with the assistance of SYNthia), 74% of participants were adamant that their draft had absolutely no AI influence, despite their prior interactions with the system. Finally, almost all participants stated that despite not accepting every suggestion, they were inspired by the tool’s outputs, which ultimately played a role in shaping their direction, tone, and structure.

In summary, this thesis covers the following material:

- SYNthia, a novel large language model-powered thesaurus and interactive web interface that allows users to write and prompt for personalized word suggestions without the additional transition of context switching. The tool and its features are described in more detail in chapters 3 and 4.
- A list of design goals and a description of the design space is introduced in chapter 3 (pilot study), and referenced repeatedly in the following chapters.
- Results from the quantitative evaluation yielded no statistically significant differences in how users interacted with SYNthia, aggregated by both writing and AI expertise. Summarized in chapter 5 and discussed more in depth in chapter 6, these assessments provided an objective metric to compare novice and expert writers’ experiences with the system.

- Results from the qualitative evaluation yielded common themes in the participants' overall view of AI in writing, the efficiency of the system, and how context (novel component) affected them differently. The results are summarized in chapter 5 and discussed more in depth in chapter 6.
- A detailed discussion on the implications of results from this project on the future of AI in creative writing, detailed in chapter 6.

LITERATURE REVIEW

2.1 Importance of Human Voice in Writing

Writing has long served as a primary mode of self-expression, enabling individuals to cumulate their thoughts, emotions, and perspectives into a structured form. Writing allows for refinement, revision, and careful selection of words, contributing to the development of a unique authorial voice. This section explores the significance of human voice in writing and the role of rewriting in shaping meaning.

2.1.1 Writing as a Vessel for Authentic Self Expression

Writing has been a fundamental vessel for self-expression throughout history, providing a space for individuals to record personal narratives, develop arguments, and communicate ideas in a healthy, comprehensive manner [20]. Written records offer a level of permanence and intentionality, encouraging individuals to craft their thoughts with precision, often reflecting deeply on word choice, structure, and meaning. This deliberative process allows for a clearer and more intentional articulation of personal voice.

Writers often emphasize the importance of authenticity in their work. Readers often seek out writing that feels genuine and reflective of an individual's unique perspective [21]. Authenticity is the connective thread between writers and readers, acting as both a measure of a work's value [22] and a reflection of the writer's sense of ownership [23]. For many writers, the act of writing is deeply fulfilling, serving as both a cognitive and emotional process; each writer has a unique perspective of their "own history, values, and point of view" that naturally comes out in their writing, which serves as the drawing point for the readers [24].

2.1.2 The Impact of Revision

Writing does not just include the physical process of writing or typing words on a page. The process of editing and rewriting also plays a crucial role in the overall quality of the final piece; rewriting involves the writers repeatedly involving themselves in their own writing, a cycle that transforms their drafts into polished works that resonate more effectively with their intended audience. Writers themselves also claim that the editing phase of their process not only polish their work, but also are the most rewarding due to the (usually) long hours spent in it [25]. In support of this, many studies have found that feelings of authenticity and ownership are linked to the challenge of brainstorming creative ideas [26] and the dedication of effort, attention, and time to the their work [27].

The relationship between rewriting and authenticity and ownership is particularly significant. When writers feel a sense of agency over their work, such as through topic selection or stylistic choices, they are more likely to produce writing that feels meaningful and personal [28]. The revision process allows writers to reconnect with their original intent and ensure that their voice remains intact throughout the piece. These efforts also translate through the page to readers of their works, transferring the same emotions the writers felt to their audience.

2.2 Novice vs. Expert Writers

The distinction between novice and expert writers has been a subject of significant research in cognitive psychology and writing studies. Table 2.1 provides a cumulative comparison between novices and experts in any domain [29]. As this section dives deeper into the key differences between them in their writing processes and strategies, we will start to see the themes in this table apply in the context of writing.

2.2.1 Cognitive Processes

Expert and novice writers employ the cognitive process of writing differently. The cognitive processes in writing includes: planning, translating, and reviewing [30]. Assuming that writers are writing in one language (e.g. English), translating is not applicable. We will examine the processes of planning and reviewing in depth below, drawing on the work of Anna Becker, who did extensive research on the cognitive progress in writing [31].

2.2.1.1 Planning and Brain Storming

The brainstorming stage is fundamental for establishing a clear idea and outline for the final product. Expert writers typically engage in more extensive planning before and during the writing process. They construct complex representations of their goals, with frequent

Table 2.1: Comparison of Novice and Expert Characteristics

Novice Characteristics	Expert Characteristics
1. Follows the rules and plans. Acquires information as a prerequisite to learning. 2. Does not feel responsible except for following the rules. 3. Has no discretionary judgment. 4. Spends time remembering information. 5. Attempts to conform behavior to the rules. 6. Learning is context dependent.	1. Thinks intuitively. 2. No longer relies on the rules, guidelines, or principles. 3. Analytical approaches used only in novel situations or when problems occur. 4. Has responsibility for self, others, and the environment.

check-ins during writing. Experts spend more time analyzing the writing task, considering their audience, and developing a mental model of the problem before trying to solve it. While novice writers also participate in planning and brain storming, their processes are often simpler, driven by the goal of completing the task at hand. They do not consider the same depth of information and how their writing could influence potential readers as much as expert writers [31].

2.2.1.2 Reviewing and Revising

Reviewing and revising is fundamental for polishing the first drafts, transitioning the piece from initial ideas to a publishable work. Expert writers demonstrate a smooth and efficient integration of conscious and automatic processes during revision, drawing on a large repository of past writing experiences stored in their extensive past experiences. This allows them to make more fitting decisions about specific revisions without overloading their working memory capacity. On the other hand, novice writers often struggle with managing multiple cognitive demands during editing. The lack of experience combined with the struggle of multitasking often cause cognitive overloads, leading to lackluster results and the writers "cutting corners" in their work.

In addition, the term "revision" carries a different definition to expert writers and novice writers. Expert writers engage in more substantive revision, often reorganizing or rewriting entire sections to better communicate their ideas, not being afraid of "throwing away" hours

of work if the results do not reflect the efforts. They are more likely to consider global issues in their writing, such as overall coherence and argument structure. However, novice writers tend to think of revision as purely syntax, often focusing on surface-level revisions, such as grammar and spelling without significantly altering the content or structure of their work [31].

2.2.2 Text Production, Integration, and Fluency

Two main components in the development of writing expertise are: extensive knowledge relevant to writing and fluent language generation processes [32]. This section breaks down how expert and novice writers differ in those two categories.

Expert writers are more skilled at integrating domain knowledge into all processes of their writing, drawing on a broader range of relevant information and experiences. They can organize and structure this knowledge more effectively within their texts, purely based on their prior experiences and long-term memories. Novice writers may struggle to select and organize relevant information, failing to prioritize more important information. This results in novice writers over relying on external tools, such as generative AI, to assist them in coming up with the most effective way of knowledge integration into their writing [33].

Fluent language generation processes correspond to the usage of short-term vs. long-term memory in writing. Expert writers more often use long-term memory, as their prior experiences and expertise allow them to almost write by instinct. Word associations are a common example. Expert writers, when presented with a prompt or scenario, can quickly list the words they would like to use, while novice writers (who lack this fluency) will need more time to grapple with the words they want. Because novices use more short-term working memory, they are more likely to get overwhelmed during the writing process and limit their ability to engage in more complex writing tasks, as they have not developed the sophisticated writing habits and instincts expert writers have [33].

2.2.3 Metacognition

Metacognition describes thinking about one's thinking as a means of reflection. In writing, it means spending more time thinking about the process, barriers, and strategies for writing, rather than the final product [34]. Experienced writers demonstrate stronger metacognitive skills, allowing them to monitor their writing process more effectively and identify areas for improvement. They are more adept at critically evaluating their own work and making strategic decisions about how to improve it. Novice writers often lack this level of self-awareness and may struggle to critically assess their own writing, thus influencing them to seek out external help (e.g. generative AI) [29].

2.3 Large Language Models

The technical aspect of this project relies on large language models (LLMs). In recent years, LLMs have emerged as groundbreaking technology in the field of artificial intelligence (AI) and natural language processing (NLP), being capable of understanding and generating human-like text across a wide range of applications.

LLMs are large-scale neural networks, trained on large amounts of textual data in an attempt to mimic the human brain and its inner workings. Containing billions of parameters [35], these models are designed to capture the nuances of the human language and communicate with us in a seamless and natural manner. With the growth of LLM prevalence came the growth in the number of parameters (amount of learning) each model contains. OpenAI's GPT-1 model (2018), the first ever LLM released to the public, had only 117 million parameters [36]. To put that into scale, GPT-4 (2023) is estimated to have at least 1.8 trillion parameters, which is comprised of eight smaller models, with each internal model made up of 220 billion parameters [37].

LLMs have several key capabilities that make them more attractive than existing language models, at least for the general public. First, LLMs can apply transfer learning, essentially transferring their knowledge of how to perform one task to another, allowing them to perform well on a wide range of tasks with minimal model training. In addition, LLMs have different use cases, ranging from generating text to document summarization and question answering [38]. All of the LLMs' generative features rely on the user's "prompts", essentially learning through examples or guidelines that the users provide. An example of how users prompts are used by LLMs can be seen in Figure 2.1 [39]. When the user initiates a session with a prompt, it goes into the LLM's context window (history or memory). The LLM uses the prompt to generate an output, which is also added to the model's context window. This way, the LLM is able to remember past details from the session and use them to address any incoming user prompts.

However, while LLMs give off the feeling that they are able to "think on their own", it is important to acknowledge the possibilities of hallucinations, biases, and other forms of incorrect outputs [40]. Even though LLMs perform extraordinarily well with data they have been trained on (billions and trillions of parameters), its chances of under performing emerge when an user prompt does not match its inner dictionary of information. Models like ChatGPT also have a tendency to please the users, which can be seen by its over willingness to change a response at the slightest nudge from the user.

2.4 System-Assisted Writing

Writing assistance tools come in several forms and continue to be actively developed and researched, with a history ranging from the dictionary and thesaurus to the common uses

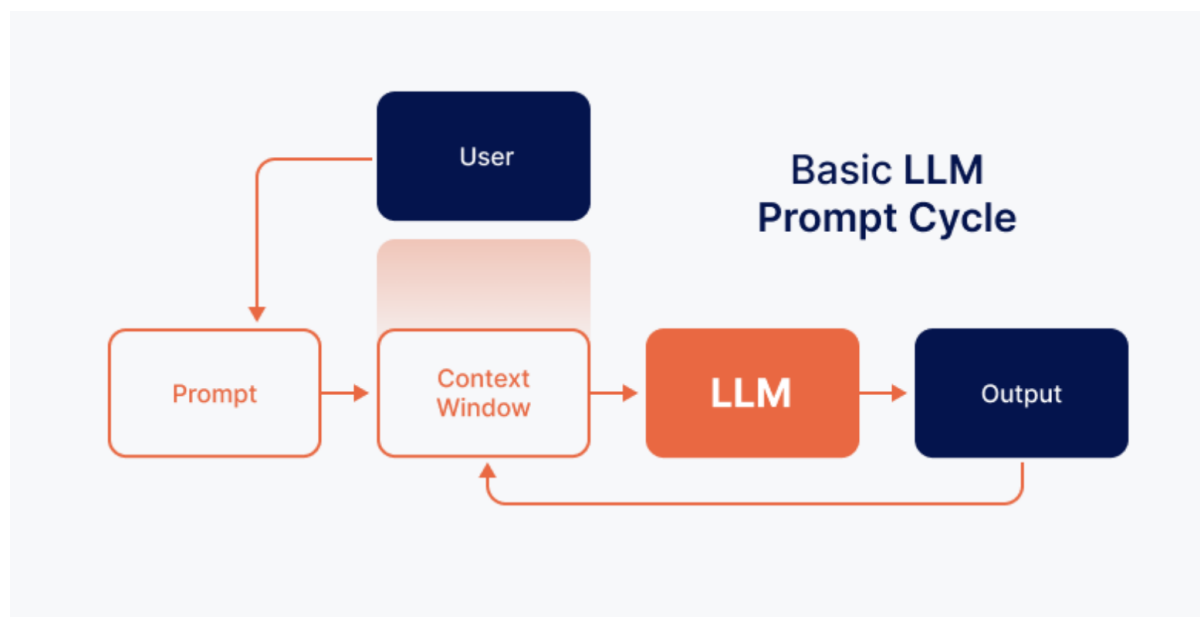


Figure 2.1: Flowchart of how user prompts help the LLM answer specific questions or demands.

of spellcheck and grammatical suggestion tools like Grammarly. With the increase in the use of generative AI in writing came the increase in their use in the creative writing space, where word choice, colorful descriptions [41], and general world building [42] are of the utmost importance. Prior work has also indicated that at different stages of the writing process, different types of assistance are generally sought after by users [43], further granulating the space of tool-assisted writing by assistance type as well as by the stage of writing targeted by that assistance. Several interfaces target the space of suggestions and predictions, focusing on completion or addition of a sentence or word [44, 45, 46, 47], or larger-scale contributions like generation of entire poems or other bodies of text [48, 49, 50]. Systems like Metaphoria [46] and Clark et. al’s slogan writing interface [45] highlight the importance of providing suggestions that accurately reflect users’ intentions or context of their writing so far in order to provide contributions that are most likely to be useful and accepted by users. This idea is referred to as ‘coherence to context’ in Metaphoria [46], and we pursue a similar goal for suggestions provided by SYNthia. Another implicit goal within these works is steerability – the ability for users to provide meaningful input around which the system adapts in order to achieve coherence to context, whether lines in a story, selection of a vessel for a metaphor, or keywords describing a slogan. Divergent outcomes, explicitly highlighted and defined, describe the supply of several valid options for users when receiving suggestions, avoiding funneling users into a single frame or style of writing [46].

Other interfaces such as Soylent [51], ABscribe [52], and RevTogether [53], rather than focusing on completion or addition of writing elements, lend focus to revision of user-written text in an effort to improve or lend input as the user develops their work. These studies

proved that users generally enjoyed having AI serve as a revision assist, almost as a second pair of eyes, though participants also wanted transparency in any revision suggestions from the tools. With the rise in accessibility of large language model (LLM) technology, however, future approaches can make use of a model's training on human writing as an effective replacement for crowd sourcing, using a model's suggestions for revision as opposed to directly requesting human input.

2.5 The Good, the Bad, and the Ugly of System-Assisted Writing

As AI gets increasingly integrated into the writing process (mostly through the form of tools and interfaces), questions start to arise regarding how technology affects the authenticity and human quality of the resulting piece.

In general, the usage of generative AI in writing can increase writer productivity and boost confidence in their writing [54]. A seamless integration of AI in the writing process can help writers diversify their ideas, while keeping the writing itself clear and concise [55]. This is particularly valuable in creative writing and brainstorming processes, where AI can serve as a source of inspiration and help overcome writer's block. On the other hand, AI has also been proven to assist in improving the overall quality of writing by reducing the time it takes for manual, syntax corrections [54]. This time-saving aspect allows writers to focus more on strategic thinking and relationship-building activities, leading to more innovative and impactful work.

However, the increase in the integration of generative AI in writing has also been a source of controversy. In the context of publishing, the use of AI-generated content without proper credit has given rise to a new form of plagiarism, dubbed "AI-giarism". While AI may increase individual creativity, it diminishes the collective creativity of the public, since all models are trained on relatively similar datasets. In addition, with the rapidly increasing sophistication of AI, it becomes harder to distinguish between AI generated writing and authentic human writing [56]. Aside from readers, writers often feel a loss of ownership over their own work [50, 57], and readers lose trust in works because they do not know how much of it is generated by a "souless machine" [57]. This loss of ownership can lead to ethical dilemmas, particularly when it comes to declaring authorship. Studies have shown that people are less likely to disclose AI involvement in their writing compared to human collaboration, even when they don't consider themselves the sole authors [58]. A more detailed list of the strengths and weaknesses of AI in writing are detailed in Tables 2.3 and 2.4 [59].

Table 2.3: Strengths of ChatGPT in Writing

Strengths of ChatGPT in Writing
1. Help to break mental log jams when writing, or when struggling to write those first words 2. Make interesting analogies, when properly prompted, and generate seemingly creative links between disparate concepts and ideas, although these require a reality check to ensure that they are reasonable or plausible 3. Be used effectively to improve the title, abstract, and conclusion of your manuscript and to tailor it to match the journal parameters and better match its scope or readership 4. Identify references for a specific topic that might be missed by conventional literature searches 5. Provide a guidance on writing structure by breaking up a difficult topic into smaller pieces 6. Help a writer be more thorough when covering a topic by reminding them of aspects they had not considered 7. Provide knowledge in an area in which one has little familiarity, in a structured, easy-to-digest manner

Table 2.4: Weaknesses of ChatGPT in Writing

Weaknesses of ChatGPT in Writing
1. Too easy if one fails to use it responsibly and with care 2. Be used to write and replace critical thinking and thorough literature reviews to the detriment of the user 3. Lead to banal, cookie-cutter and uninteresting science if not used as only a jumping-off point for creative [writing] 4. Be used without reading the actual papers [and books] that support claims made by the author 5. Generate output that is incorrect or recently shown to be false

SYNthia aims to be a creative writing tool that serves as a level below the human writer, playing more of an assistance role rather than a co-writer role, by allowing users to have control over the number of suggestions, when the AI is triggered, and the prompt it works with. More about SYNthia's interface design is discussed in Section 3.

2.6 Efficient Error Correction

Implementing efficient error correction has long been a subject of discussion. With algorithms such as handwriting, speech, and text recognition models increasing in popularity [60, 61, 62, 63, 64], their susceptibility to error was quickly uncovered [63]. While such recognition models could only achieve a certain threshold of accuracy, user satisfaction could still

be achieved through the process of error correction, whether it be automatic or manual. Most prior works that have incorporated error correction in their models do so in a multimodal format, mostly targeting some combination of text, handwriting, and speech. For example, a prior Chinese handwriting recognition model utilized a crossmodal error correction technique, allowing users to feed in an initial handwritten Chinese phrase, as well as an audio file of the spoken sentence. The model would then create a fusion of the text and audio data, increasing the overall accuracy of the final output [61].

However, it is important to note that most error correction papers concern themselves with the scope of error correction in transcription/recognition contexts, and not necessarily in the context of writing or word choice. Speech is used as a common form of error correct due to its easy, fast, and hands-free method of input [60, 63]. Like any sort of automatic detection, automatic speech recognition (ASR) still requires a layer of text modification, which serves as a form of efficient user error correction. While speech inputs are common, they are often insufficient in specifying error locations, something that the added text inputs could take into account [62]. This line of logic also explains why multimodal methods are an appropriate fit for most types of error correction.

Although our project does not focus on the fusion of text and voice inputs, we incorporate additionally provided textual context by the user as a form of error correction, a process described by Davis as human-computer co-creativity [65]. SYNthia enhances the collaboration between between the human and computer, with the computer taking an adaptive role in accordance to human input (additional context). However, a challenge of having both the initial user input and the error correction input be in textual form is the possibility of users getting distracted during the prompting process [66]. During the writing process, writers likely have to juggle multiple ideas or tasks [43], one of which could be finding alternative word choice while not interrupting their current flow of thinking. To address this issue, our stylistic choice of having the context box on the side of the main text editor ensures that users did not get distracted during their writing/drafting process, nor during their process of providing context and reprompting words.

PROTOTYPE AND PILOT

In fall of 2023, we created the first prototype of SYNthia. Consideration for SYNthia’s first design incorporates influence from both prior works and the observed use of OpenAI’s ChatGPT platform and Grammarly, another AI-infused grammar correction application. As a prompt-based LLM platform, ChatGPT provides users with an extremely flexible interface in which writing (among other) questions and requests can be fulfilled, though users may need to explore and experiment with prompt types in order to refine the output given. However, use of ChatGPT requires the presence of the interface in a separate tab or window on a user’s device, resulting in users switching tabs and focus in order to make use of its capabilities, as opposed to more readily available suggestions or help that are built into interfaces or devices [44]. Conversely, Grammarly provides suggestions for words, sentences, or other grammatical features within its own interface or as a background application that reads user-written text wherever it is written. However, these suggestions, particularly in reference to suggestions for revised word choice, are broad in nature and Grammarly does not provide the means for users to explicitly express their intentions or additional context for the suggestions generated. This section provides an overview of prototype SYNthia’s design, prompting, experiments, and findings.

3.1 SYNthia’s Prototype Description

3.1.1 Invoking and supplying suggestions

With reference to Figure 3.1, SYNthia’s interface allows users to type free-form text into a text box on the lefthand side of a split view (1). To request an initial generation of suggestions for a written word, users can append a ‘@’ character directly to the end of a word, illustrated in (2). Doing so results in a suggestion box 3 to appear on the righthand side of

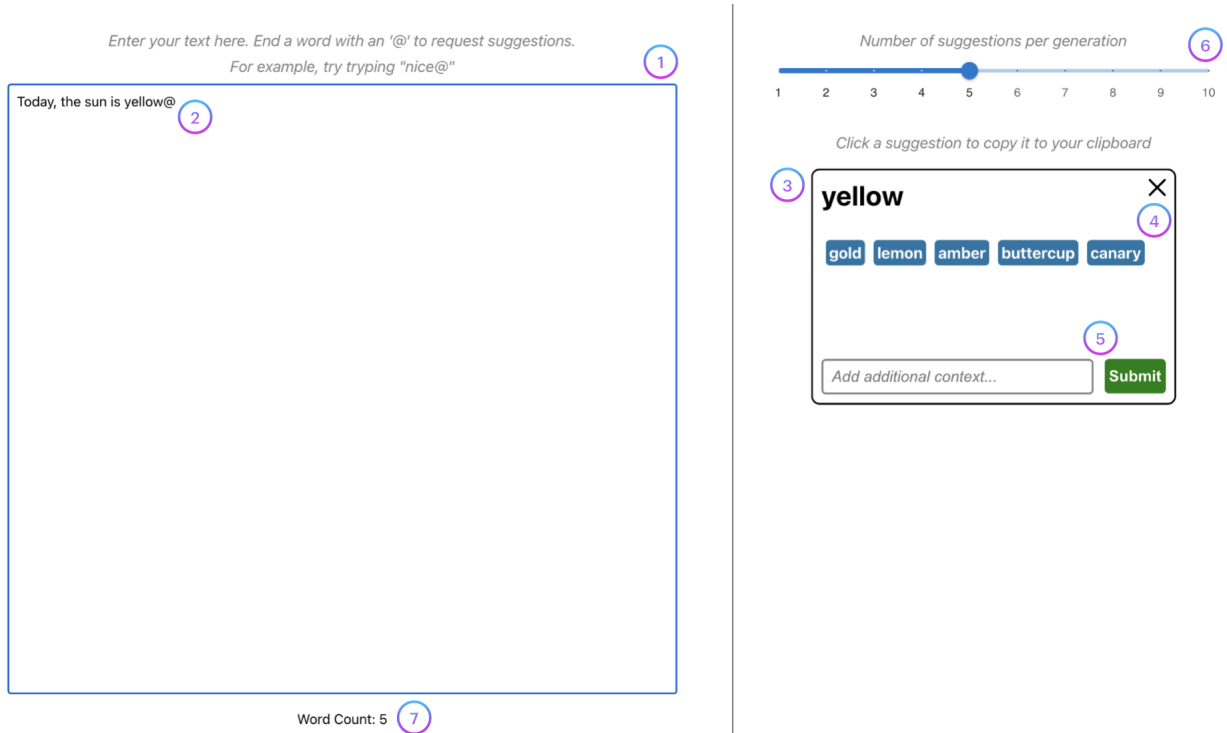


Figure 3.1: Prototype Interface of SYNthia. (1) User input textbox. (2) Word marking with '@' symbol. (3) Suggestion box. (4) Generated list of suggestions. (5) Input for additional context from user. (6) Slider determining number of suggestions provided per generation. (7) Word counter.

the split view, supplying an initial list of suggestions (4). Using a provided input (5), users then have the ability to provide any context or other clarifications or requirements for the suggestions provided. Users can click 'Submit' without any input to trigger a 'clean' regeneration of suggestions that use the same context as the previous generation. Generation of these suggestions uses a prompt with a fixed structure sent to an instance of GPT-4 (separate for each individual word), available in Appendix 7. The prompt requests a series of suggestions for a given word that is interpretable via JSON parsing, and additional messages are sent to the conversation instance if the user chooses to provide additional context, resulting in a regeneration of suggestions by GPT-4, taking any additional messages into account. Users can click a suggested word to copy it to their clipboard. In the case of a lack of valid suggestions or other error, users are notified by a replacement of the list of suggestions with a message reading "No suggestions available".

3.1.2 Adjusting number of suggestions provided per generation

The slider in the top right corner of SYNthia's interface (6) specifies the number of suggestions to be provided per generation, set to a default of 5 when the interface is first launched. Any change to the number of suggestions provided apply to the next generation triggered, whether through initial generation by marking a word with '@', or regeneration through sub-

mission of context (or 'clean' regeneration). The range of possible counts is from 1 to 10, with 5 as a 'middle ground' default value.

3.2 Design Space

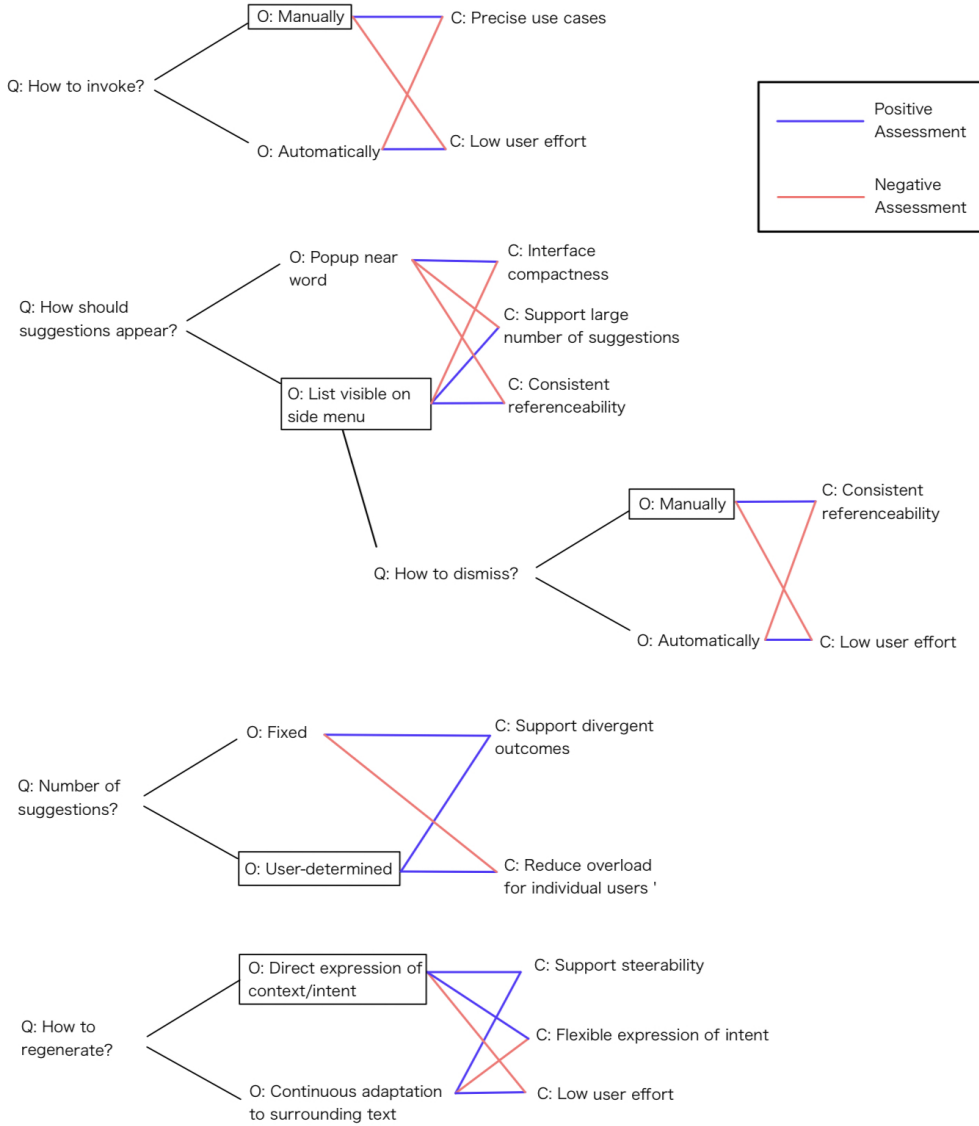


Figure 3.2: A QOC representation of the design space for SYNthia, including a map of options and corresponding criteria. Boxed options indicate choices made in the design of SYNthia.

As a supplement to the description of SYNthia's design goals and features, we include in Figure 3.2, a Questions, Options, and Criteria (QOC) representation of the design space explored when designing SYNthia. Inclusion of this diagram takes inspiration from MacLean et. al.'s proposal of the diagram for analyzing design spaces [67].

We include a set of additional criteria considered while designing SYNthia, such as consistent referenceability and level of user effort. We define consistent referenceability as the

ability for users to look back onto suggestions generated for a given word in the case that they wish to use multiple suggestions to replace multiple instances of that word. This idea underlies the adoption of manual invocation and dismissal, among other considerations for SYNthia’s usability and ability to test our decided research questions.

Additionally, we highlight the presence of steerability (which leads to coherence to context) and divergent outcomes among the listed criteria. While exploring the design space that resulted in the creation of SYNthia, we observed multiple methods of supporting steerability and divergent outcomes, though prioritizing methods that entailed additional benefits such as flexibility in expression of context at the expense of raising user effort in doing so. Our design of SYNthia is of course not the only possible means of exploring the impact of efficient correction when writing with LLMs. We present our QOC diagram as evidence of alternative possibilities and an expression of our priorities and associated choices during the design process.

3.3 Pilot User Study

For the following two studies, we recruited 10 Harvard College students, 6 males and 4 females, with an average age of 20 ($\sigma^2 = 1$). All participants majored in some sort of scientific field of study. Each participant performed two writing tasks (one for Trial 1 and another for Trial 2) and took part in a semi-structured interview. At the end of the experiment, each participant received a \$10 gift card as compensation for their time.

Before the two trials, each participant was asked about how often they write, both within and outside of schoolwork, as well as their prior experience in using AI tools as a writing aid. In addition, we asked each participant about their tendencies to seek out alternative words while writing. Before the trials, participants were introduced to SYNthia and each of its relevant features, including a brief demonstration of the word marking process by the interviewer. Participants were invited to ask any questions they had relating to how to use the interface before the beginning of either trial.

Participant pre- and post-interviews were recorded and transcribed, and interviewers screen-recorded and took notes on participant behavior during both trials. Interview questions are included in Appendix 7.

After the two trials, we asked each participant what they perceived as the most difficult or frustrating parts of SYNthia to use, noting any specific instances of frustration they experienced, ranging from general usage to their tolerance for re-prompting in Trial 2. Next, we asked each participant whether they believed using the interface helped improve their writing as a way to further gauge the usefulness of SYNthia. Then, we asked each participant whether they would regret not having the opportunity to use this interface.

After the structured interview questions, the scope of the interview was expanded to be more conversational via participant-specific followup questions and additional feedback

and suggestions. For all participants, we began by asking about their revision process, followed by a discussion of any feedback they may have regarding the interface. Additionally, we asked each participant to compare SYNthia to other tools they have used for finding synonyms, if any, including AI-infused tools or a thesaurus, noting any significant distinctions. We ended this section with a question about how novel each participant found the interface. Participants were also invited to share any additional thoughts, with some discussion afterward.

3.3.1 Trial 1: Control

In the first trial, the feedback mechanism within the suggestion box ((5) in Figure 3.1) created for marked words was disabled, such that participants could only change the number of suggestions generated, with no steerability of the suggestions themselves. Participants were asked to respond to a writing prompt (included in Appendix 7), ideally within 100 to 150 words. Participants were invited to think out loud whenever interacting with the interface. Interviewers took notes on the words being marked for suggestions, as well as on any additional behavior relevant to the interface, including changes to the number of suggestions per generation.

3.3.2 Trial 2: Experimental

In the second trial, the feedback mechanism within the suggestion box created for marked words was enabled, allowing participants to provide additional context in return for a new batch of suggestions. Participants were asked to respond to a different writing prompt (included in Appendix 7), ideally within 100 to 150 words. Participants were once again invited to think out loud when interacting with the interface. Interviewers took notes on the words being marked for suggestions, the context, if any, each participant provided SYNthia, as well as other statistics on their interactions with the interface (e.g. the number of times participants submitted context for the same word).

3.4 Pilot Findings

3.4.1 Trial 1: Control

In the absence of steerability, users may seek out breadth: Without the ability to explicitly provide correction via prompting, a majority of users fell into two categories during Trial 1. The first category consisted of users who effectively ignored the "Number of suggestions per generation" slider, seemingly satisfied with the default of 5 suggestions or otherwise uninterested in receiving more. Participants in the control typically chose a word from the initial list of suggestions or quickly decided to move on instead, and so participants who remained

at 5 suggestions may have also found satisfying words each time by chance, missing any motivation to deviate. The second category of users appeared to seek as much material to sort through as possible, setting the slider to 10 (the maximum supported within SYNthia) suggestions either at the beginning or at some intermediary point during Trial 1 and choosing to leave it there for the duration of the trial. Participant 2 noted that they would "always want like, as many synonyms as possible" before choosing the one they felt was best. Similarly, participant 7 saw the slider at the beginning of the experiment and asked "Why wouldn't I just keep it at ten?", again suggesting that for some users breadth is an important factor in the absence of steerability.

When breadth is not enough, a lack of steerability leads to abandonment: Inexpressible subtleties of context caused several users to abandon their search for suggestions or attempt to regenerate suggestions (by deleting and replacing the '@' symbol to generate a new set). While this resulted in some variations in the suggestions provided, the context of such suggestions often remained inadequate for participants' purposes, resulting in them deciding on a word themselves or abandoning the suggestions altogether and leaving their word unchanged. This is seen most clearly in observations of Participant 3, who received suggestions for 14 different words, replacing only 4 with suggestions from SYNthia, leaving their words unchanged otherwise. Similar behavior was observed with Participant 5, who replaced 3 out of 8 words they received suggestions for, and Participant 6, who replaced 1 out of 4 words, both deciding to leave their word choice unchanged when dissatisfied. Participant 3 and 5 raised the number of suggestions to 10 during Trial 1, while Participant 6 left the number at the default of 5.

3.4.2 Trial 2: Experimental

Coherence to context and divergent outcomes: Although it was difficult to observe the diversity in SYNthia's outputs, as different participants typically prompted for different words, we did observe that there were no duplicates in outputs when users gave the interface additional context. In general, the outputs users received also seemed to correspond appropriately to the additional details given to the interface, causing participants that chose to regenerate words through providing additional context to always pause to consider SYNthia's outputs. Several participants expressed surprise or satisfaction with the degree to which SYNthia obeyed their provided context. We saw a connection between SYNthia's coherence to context and some participants' ability to trigger their "tip of the tongue" memory, perhaps as a result of seeing their context reflected back to them, most notably in the case of Participant 9. From these examples, we identify the implication that systems that allow users to express their intent should have an additional effect of helping users produce their target words, discussed here further in relation to cognitive engagement.

Cognitive engagement: Multiple participants were observed displaying elevated cogni-

tive engagement when providing additional context to SYNthia. Notably, Participant 7 was able to trigger a recall of the word "trudge" by prompting for the word "walk" and engaging with the suggestions returned when adding the context "more intense walk, like a hike without an incline". Though the context they provided did not result in SYNthia explicitly suggesting the word "trudge," it appeared to facilitate recall of the word by stimulating them with contextually similar words like 'trek', 'tramp', and 'ramble', among others. Similarly, Participant 9 was able to think of their target word, "silhouetted", by seeing contextually similar words provided by SYNthia on their prompted word "dark", with suggestions including 'murky' and 'shadowy'. These participants claimed that either by writing out a description of their target word, or by seeing contextually similar words, that they were able to overcome some mental blockage, allowing them to recall the word they wanted, even if SYNthia did not explicitly suggest it to them, described by Participant 7 as "forcing myself to write out what I was looking for helped me come up with the word I was thinking of in my head" and by Participant 9 as "some sort of lock until I was able to find like, contextually similar words".

Steerability: With all 10 participants, providing SYNthia with additional context always resulted in distinct suggestions, with no overlap with the previous set. However, according to observed participant behavior and responses, the added steerability did not guarantee a word that would satisfy the user. This may imply that despite giving users the ability to directly express their intention via natural language, the limitations of that natural language and participants' ability to precisely express their intentions may have limited the quality of outputs provided by SYNthia. Succinctly, SYNthia's suggestions are limited by the quality of user-provided context descriptions. A potential solution to increasing satisfaction may include providing examples of potential prompts or devising an alternative means by which users can express their intent, though this remains to be explored.

3.4.3 Additional Findings

Stifling creativity: Participant 4 expressed concern regarding a tradeoff between using SYNthia to write and their own feeling of creativity and ownership over their work. As several participants stated, SYNthia has potential as a time-saver and is likely to be called upon out of a user's 'laziness'. The intended purpose of SYNthia was to enhance human-computer co-creativity, having the AI be tailored towards some given human input [46]. However, some participants prioritized saving time over increasing creativity by prompting the AI without an explicit goal in mind. Participant 4 in particular, noted that SYNthia could be used in two different ways: either as a tool to save time or a tool to enhance creativity, depending on a given user's intention. While Participant 4 claimed that the interface "can save quite a bit of time and energy", they also compared using SYNthia as a time saver to having a crutch, as the user would then be prompting SYNthia out of necessity, becoming an increasingly dominant source of creativity. On the other hand, they compared using SYNthia as a tool for

enhancing creativity to a hiker using hiking sticks; the hiker would still be able to accomplish their goal of climbing a mountain without their sticks, but using hiking sticks would only enhance their abilities, without absolute dependence on such a tool.

Usage: No matter what participants indicated their revision style was (mid-writing vs. post-writing), while using SYNthia, all 10 of them prompted the AI mid-writing. We suspect that there was some sort of novelty effect, with some participants claiming to prompt the interface simply "for the sake of using it" as described by Participant 6.

SYSTEM OVERVIEW

The interface name SYNthia is derived from the word "synonyms", as it is an AI-powered thesaurus aimed at giving users alternative word suggestions. This section provides an overview of SYNthia's design goals and how limitations from the prototype/pilot are addressed in SYNthia's final form.

4.1 Design Goals

SYNthia's design goals are informed by both gaps in existing tools (Chapter 2) and the observations obtained from interviews with 10 Harvard College students in the pilot study (Chapter 3). SYNthia attempts to fulfill three primary objectives: coherence to context, steerability of suggestions, and divergent outcomes, consistent with the objectives studied in the pilot. Coherence to context, as defined by Gero and Chilton, states that systems should "provide suggestions relevant to the task at hand" [46]. Users who use words in particular contexts may require suggestions to reflect that context, or to respect a particular voice or tone, among other possible considerations. In order to fulfill this coherency to context, systems accepting user input should demonstrate steerability. Systems that are steerable adapt their outputs to the input of their users, creating an environment of co-creativity [65]. Systems supporting divergent outcomes provide several valid candidates for suggestions, attempting to support variation in writers' work [46].

To support coherency to context, SYNthia always incorporates all context provided by users through the context box to produce the most relevant suggestions possible when given one or more statements of clarification. By allowing users to directly express the types of suggestions they are seeking, SYNthia is able to effectively remain coherent to their context by providing suggestions that are relevant to the description of context provided.

To support steerability, as related to coherency to context, SYNthia's suggestion outputs change accordingly with each statement of context provided by users, as instructed by the fixed LLM prompt, available in Appendix 7. By seeing the effects of the context provided on suggestions, users can continuously "steer" or "nudge" SYNthia to providing suggestions that best reflect their standards. Effectively, this steerability is the mechanism that allows SYNthia to be coherent to context. We specifically chose to look at the influence of system steerability on the coherence to context because of the lack of efficient steerability in prior works, attempting to fill this gap.

To support divergent outcomes, SYNthia provides users with the option to adjust the number of suggestions provided, supporting up to 10 different word suggestions at a time. The LLM prompt provided to SYNthia specifies that these suggestions should be distinct, and this is reflected in the lack of in-list duplicates and variety of words provided in every generation of suggestions, giving users a concentrated set of candidate words from which to choose from.

4.2 SYNthia

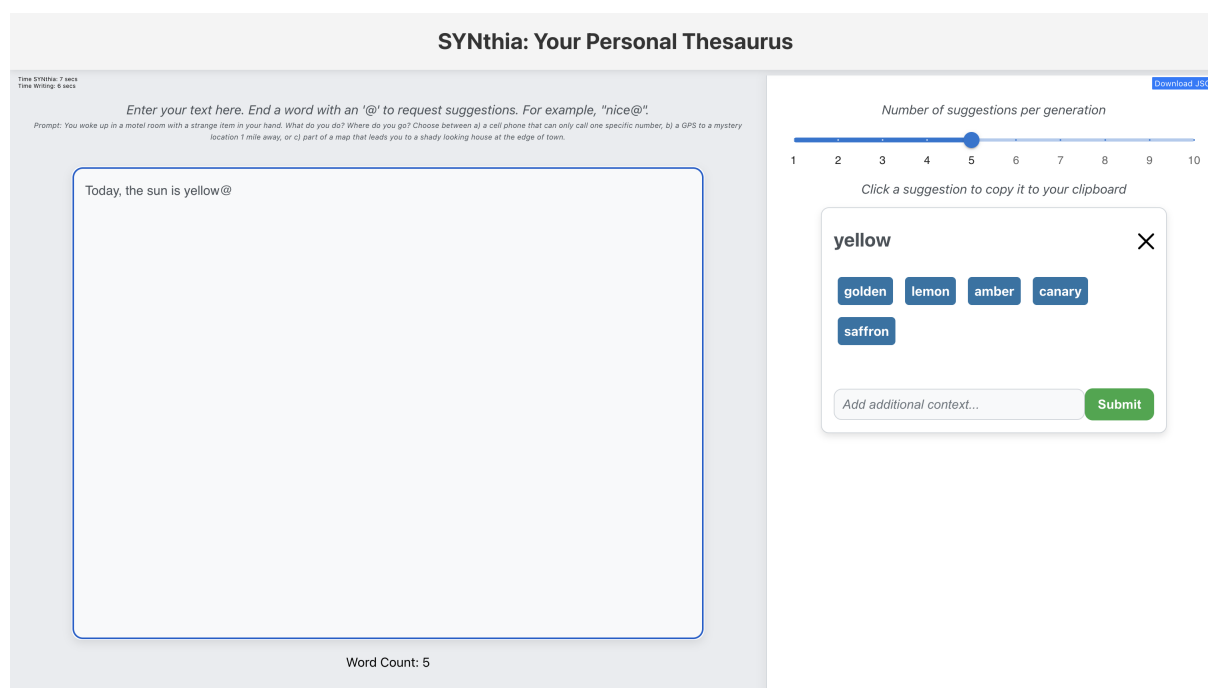


Figure 4.1: Updated interface of SYNthia. On the left, users can create/edit their writing in the text box. A word counter is provided on the bottom left. SYNthia can be prompted with the @ symbol, which triggers the suggestion box to pop up on the right. Users can toggle the number of suggestions with the slider on the top right corner, as well as interact with the suggestion boxes by providing additional context for reprompts. Additional measurements are collected via a dual timer system (top left corner) and a live tracker of prompted words, context given, and word chosen through a JSON file (top right corner).

SYNthia's interface provides a clean and concise way for writers to engage with an interactive thesaurus during their writing process. We specifically chosen to have this "side-by-side" design, with both the writing and word suggestions, to ensure that users have minimal context switching. Unlike a traditional thesaurus (online or a physical copy), SYNthia gives writers complete control over the number of suggestions, although we cap it at 10 to avoid overwhelming users, as well as offer an interactive feature for users to cognitively engage with the system (and indirectly with their writing through the ability to manually provide additional context) to find the perfect word for their piece.

A video demonstration of SYNthia can be found here: <https://youtu.be/oMu6PbfYXMo>.

4.2.1 Changes from Prototype

Comparing the prototype (Figure 3.1) to the final iteration of SYNthia (Figure 4.1), the most striking updates are the aesthetic improvements of the front end. Leveraging principles of color theory and the strategic use of shadows, we crafted a cleaner, more visually appealing interface designed to enhance the overall user experience. Beyond aesthetic updates, the final iteration of SYNthia also incorporates significant back-end enhancements, including bug fixes and system output improvements. Previously, the prototype experienced errors after 1-2 rounds of re-prompting, a critical issue that has now been resolved. In addition, we introduced tracking features to monitor user interactions with the interface. These include timers and a live JSON tracker that logs user inputs. Importantly, all backend tracking mechanisms are transparently disclosed to users and designed to avoid interfering with the user testing experience.

4.2.1.1 Dual Timers

One of the key data points we wish to collect and analyze is the time a user spends writing versus the time a user spends interacting with SYNthia (i.e. looking for alternative words and deviating from making progress on the writing). To do so, we have implemented a dual timer system, located in small text at the top left corner of the interface to avoid users being distracted by the moving numbers – ultimately, these times are for developers to analyze, and not for users to stress over.

Time spent engaging with SYNthia is defined as any time when: (1) the user has an @ symbol present in the text box or (2) the user gives context to the suggestion box at any point, even if there is no active @ symbol present, and they are just going back to a previous suggestion box. All other forms of engagement / writing in the textbox count towards time spent writing.

4.2.1.2 JSON Logging

Another set of key data points we wish to collect and analyze are specific prompted words, additional contextual inputs given to the system, and the ultimate chosen words (if any). Thus, we created the ability for SYNthia to add such information to a live JSON file that corresponds to the current user. The @ symbol that triggers SYNthia will also add the corresponding word to the JSON; the "enter" key or "submit" button when providing additional context will now add that context to the JSON; clicking a suggestion will now add that choice to the JSON. At the end of the session, we will download the JSON file by clicking the "Download JSON" button, located at the top right corner of the screen, again, to prevent users from being distracted by its presence.

4.3 System Architecture

The interface starts with an empty text box on the left side and no suggestion boxes on the right side. The user can type in the text box and trigger SYNthia by appending an @ symbol to the end of the word for which they want alternative suggestions.

The word associated with the @ symbol is then sent through a GPT-4 prompt, which outputs an initial list of synonyms in a list format. That list is then parsed and displayed on the HTML page. Users also have the ability to provide additional context to update the list of synonyms within the suggestion box. After pressing the "enter" key or the "submit" button, the user provided context is appended to the original GPT-4 prompt that regenerates a new list of synonyms. Ultimately, users can click on one of the synonyms (blue box) to copy that word to their clipboard, which indicates to the backend that the user has "chosen" that particular SYNthia suggestion. In addition to being shown on the user-facing interface, all words that trigger SYNthia, corresponding context, and chosen suggestions are added to a JSON file for further analysis.

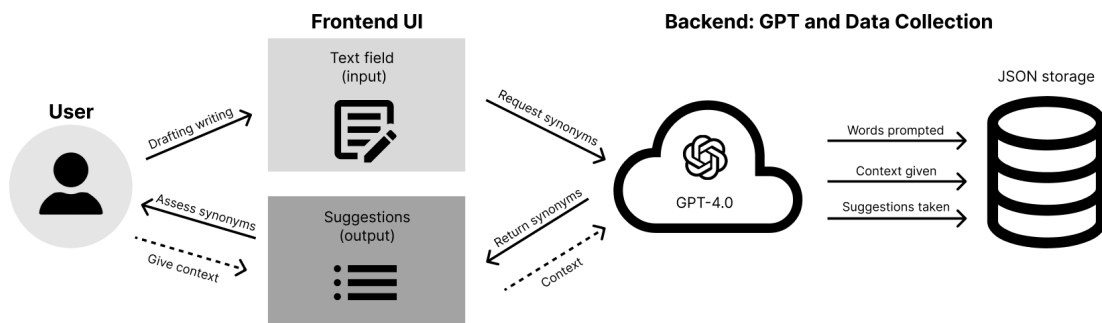


Figure 4.2: SYNthia system architecture

USER STUDY

This study targets both novice and experienced creative writers. We aim to analyze how writers of different skill levels and experience interact with an AI-powered thesaurus (writing aid) differently. In addition, we hope to analyze how writers with differing levels of AI experience (i.e. usage in their every day lives) can or cannot transfer those existing skills in their interactions with SYNthia. This section provides a description of the participant pool, detailed outline of the user study, and study results.

5.1 Participants

5.1.1 Recruitment

In this study, we interviewed 19 participants: 12 novice writers and 7 expert writers. Participants were recruited through the Harvard College student body, surrounding MFA programs, faculty members of the Harvard College English department, and Prolific. Potential participants were asked to fill out a survey that included relevant information, such as:

1. Editing style
2. Frequency in which they think of alternative words
3. Current writing aids they use
4. Prior writing experiences
5. Prior AI experiences
6. Self-rated writing skills (1-10)

7. Self-rated AI prompting skills (1-10)

The entire participant recruitment survey can be found in Appendix 7.

Depending on the survey responses, users were then labeled as "novice" or "expert" writers. Expert writers are defined as writers who have non-research published works and/or degrees in writing (e.g. MFA). All other participants who do not meet the previously described standards are labeled as novice writers.

As an incentive to volunteer, each participant was financially compensated \$25 for taking part in the interview process.

5.1.2 Background Information

Figure 5.1 and Table 5.1 provide a holistic summary of the participant backgrounds, all self-reported through an initial Google form and follow-up interview questions.

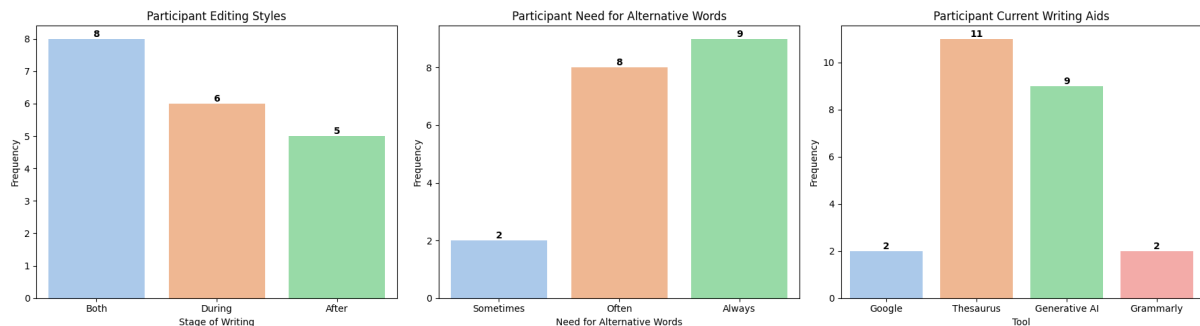


Figure 5.1: Participant editing style, need for alternative words, and current writing aids

13 participants have published works. The novices who have published works are all academic or research papers in their area of study, including but not limited to: computer science, accounting, physics, and neuroscience.

14 participants have experience editing during their writing, which was the original target audience for SYNthia. Due to SYNthia's heavily integration of a thesaurus, AI, and the writing text box, it was originally designed for those who need to access a smart thesaurus during their writing process. However, we realized that it was also suitable for participants who would write, and come back to edit, as the tool allowed for prompting during any part of the writing stage. Thus, we recruited participants with a diverse spread of editing styles.

17 participants either often or always have the thought of alternative word choice on their mind when writing, and 2 said that they sometimes do. This was the one soft requirement we had for all participants: they needed to have a history of wanting alternative words in their writing processes, at least sometimes.

9 participants use generative AI in their everyday writing process, including ChatGPT, QuillBot, and Perplexity. No participants admitted to using generative AI to copy and paste writing and pass it off as their own. Most participants used AI platforms for doing research

(Perplexity), editing or catching grammar mistakes (QuillBot), and providing alternative phrasing of their writing (ChatGPT). In regards to editing, 2 participants mention using Grammarly as their tool of choice. Similarly, with research, 2 participants mentioned using Google as their tool of choice. 11 participants use a regular thesaurus (either a physical book, thesaurus.com, or Power Thesaurus) in their writing. Many participants listed several tools for this survey question. However, the trends are clear: the overwhelming number of participants who currently use some model of generative AI and thesaurus in their writing process supports the need for an AI-powered thesaurus like SYNthia.

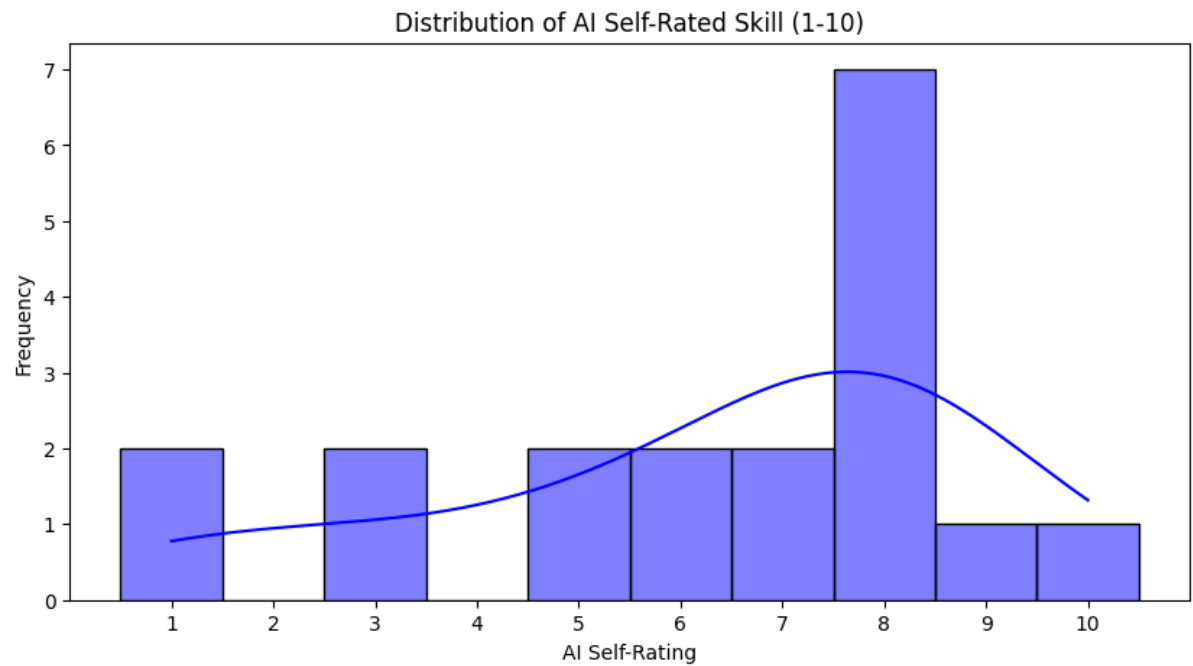


Figure 5.2: Distribution of participant AI skills (self-rated on a scale from one to ten)

Due to how common AI usage is in writing, we also collected a self-rating from each participant regarding their AI skills, ranging from one ("I can never get the AI to output what I want") to ten ("I can always get the AI to output what I want"). The specific scores corresponding to each participant is listed in the right-most column of Table 5.1, and the distribution can be seen in Figure 5.2. The most common score is eight, symbolizing that most participants were comfortable using AI in their writing, and had sufficient prior experiences with prompting and refining LLMs.

5.2 Methodology

The user interview lasted around 45 minutes per participant, and was divided into three sections: pre-interview questions, play test with SYNthia, and post-interview questions. Interviews were conducted over zoom, which were recorded and transcribed using the built-in zoom transcription feature. Before recording, each participant was made aware of their

Participant	Age Range	Writing Level	Creative Writing Experience (Years)	Writing Hours/Week	Published?	AI (Self-Rated)	Skill
1	18-24	Novice	0	2	Yes	3	
2	35-44	Expert	10	30	Yes	1	
3	35-44	Expert	10	25	Yes	1	
4	18-24	Novice	0	4	No	6	
5	18-24	Novice	8	25	No	8	
6	18-24	Novice	2	8	Yes	7	
7	18-24	Novice	0	7	Yes	8	
8	18-24	Novice	8	5	No	8	
9	18-24	Expert	4	3	Yes	5	
10	45-54	Expert	10	10	Yes	8	
11	25-34	Expert	8	3	Yes	3	
12	25-34	Novice	10	5	No	8	
13	18-24	Novice	4	6	Yes	9	
14	18-24	Novice	2	4	No	6	
15	35-44	Expert	10	5	Yes	7	
16	25-34	Expert	7	20	No	8	
17	55+	Novice	0	6	Yes	8	
18	45-54	Novice	10	8	Yes	10	
19	18-24	Novice	5	1	Yes	5	

Table 5.1: Participant Background on Writing and AI Skills

rights to stop the recording at any point, and recordings were only taken with the informed verbal consent of the participant.

5.2.1 Pre-Interview

Due to the existence of the participant recruitment survey, the pre-interview portion was relatively short, only containing three questions:

1. How many hours per week do you write on average?
2. Do you have any published works and/or manuscripts?
3. What are your opinions on using AI in creative writing?

These questions aimed to get more information on the participant's writing levels and any existing opinions / thoughts towards the intersection of AI in creative writing.

5.2.2 Play Test with SYNthia

The interview continued with a brief demonstration of SYNthia’s interface, followed by a 15 minute writing session with a universal creative writing prompt. During the session, participants were encouraged to try out all of SYNthia’s capabilities, and to voice their thoughts during the process. Quantitative data, such as words prompted, context given, word chosen, and interaction times with the system were recorded on the backend. The writing prompt can be found in Appendix 7.

5.2.3 Post-Interview

The post play test interview included 8 questions that covered the following categories:

1. Motivation behind user giving additional context for certain words
2. Comparison of SYNthia to existing resources participants use to find alternative words
3. Satisfaction with SYNthia’s suggestions
4. Satisfaction with the writing produced from the play test with the assistance of SYNthia
5. Feeling of ownership of the writing produced from play test with the assistance of SYNthia

A full list of all questions asked can be found in Appendix 7.

5.3 Results

On average, participants requested nine synonyms per session, provided additional context for two words, and ultimately accepting 43% of SYNthia’s suggestions. In this section, we first report the quantitative results regarding how users interacted with SYNthia (words prompted, context given, etc.), comparing them across novice and expert groups. Then, we report on common themes found during our qualitative analysis, such as the conflicting opinions of AI usage in writing (pre-interview), SYNthia’s novelty, and the usage of context during each user’s interview session.

5.3.1 Interactions with SYNthia

As mentioned in Chapter 4, we collected backend data on words prompted, context given, words accepted, time spent writing, and time spent with SYNthia for each participant. We performed independent T-Tests for both words prompted and times spent, though neither yielded statistically significant results when comparing between novice and expert writers.

However, there were still differences in how the two groups of writers interacted with SYNthia, as detailed in the following subsections.

5.3.1.1 Words Prompted and Accepted

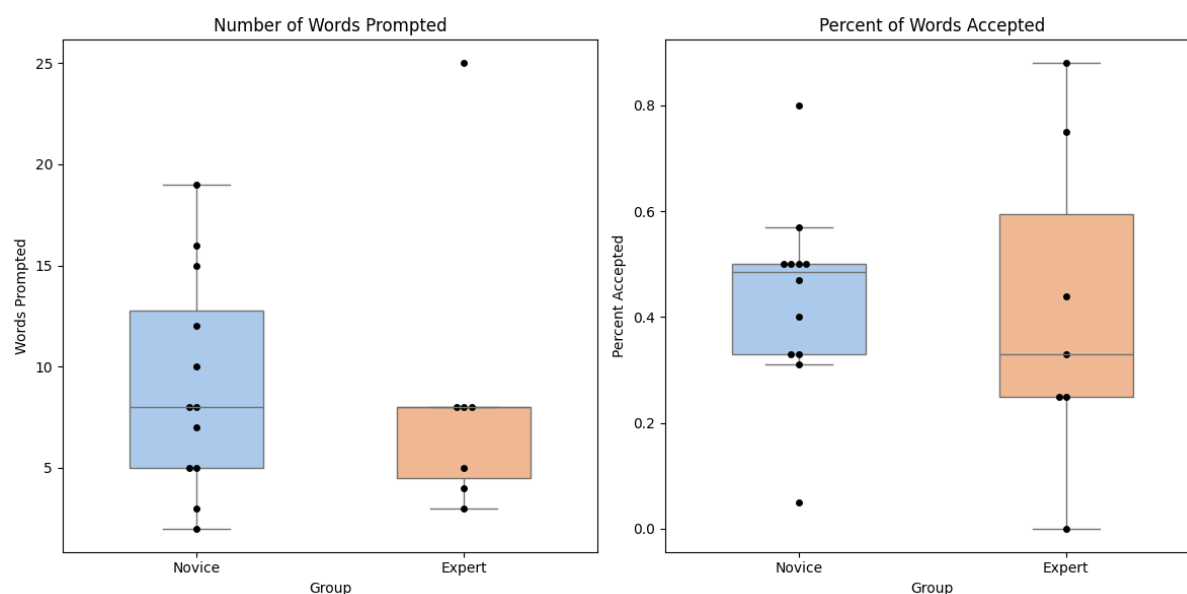


Figure 5.3: Words prompts and percent of words accepted per participant, aggregated by writing skill (novices vs. experts)

From Figure 5.3, the distribution of the number of words prompted by novices and experts shows a clear distinction in writing behavior. On average, novice writers prompted the system for more words, as evident by the larger interquartile range (IQR). In general, novice writers had greater variation in the number of words they prompted, ranging from <5 to almost 20. For expert writers, there was only one outlier, with a participant prompting 25 words. However, that can be explained by his occupation as a poet, requiring him to look for alternative words more than most.

The percentage of words accepted also varied between novice and expert writers. On average, novice writers tended to accept more words (median of 0.5), while expert writers accepted around 0.38. However, both groups exhibited large distributions, with expert writers having more variation, as demonstrated by the larger IQR. Some expert writers accept no suggestions, while some accepted close to every suggestion. The higher presence of extreme values in the expert group suggests more individualized writing strategies—some relying more heavily on AI suggestions while others rejecting most of them.

5.3.1.2 Time Spent Writing and With SYNthia

From Figure 5.4, the distribution of writing times for both novice and expert writers exhibits a similar range, with both groups spending an average of 600 to 800 seconds on writing.

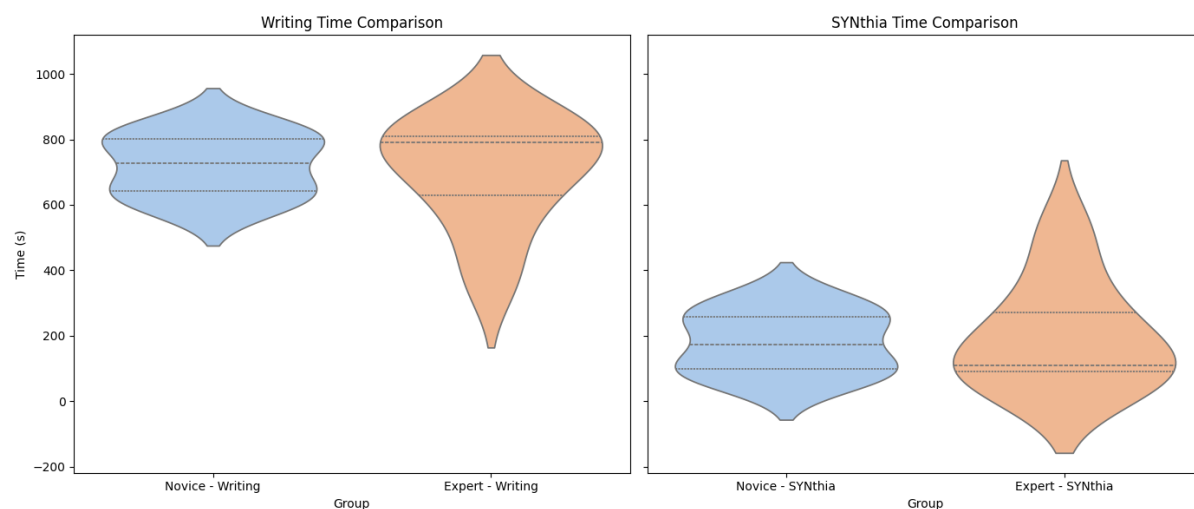


Figure 5.4: Time spent writing and time spent engaging with SYNthia per participant, aggregated by writing skill (novices vs. experts)

However, experts exhibited more variation in writing times, which could potentially be explained by some expert outliers in Figure 5.3, one expert prompting 25 words, greatly lowering their writing time. In addition, some experts exhibited greater curiosity towards SYNthia as a tool (P3, P10), while others expressed severe disinterest (P2), which explains the greater variation in engagement with SYNthia as well. Overall, novices were more focused on completing the writing task, and did not have much curiosity towards engaging with the tool outside of getting the writing done.

5.3.2 Opinions of AI in Writing

Four participants came into the interview with negative outlooks on using AI in their writing. The remaining were more open-minded, though each participant who was open to AI had their own reservations.

5.3.2.1 Current Problems with AI in Writing

Concerns about overreliance, ownership, authenticity, and quality shape a broader skepticism toward AI in writing. While AI may offer convenience, many writers see it as more of a hindrance than a helpful tool, raising ethical, creative, and practical concerns about its role in the writing process.

One major concern surrounding AI in writing is **overreliance**, as users worry that using AI tools might diminish their own skills and critical thinking. P8 noted, "My worry would be that people would rely on that too much." Similarly, some participants believe that using AI removes the intellectual labor of forming arguments and structuring ideas, as P7 explained, "I still hold on to the value of actually doing the writing yourself, because I find I do a lot of the thinking through the arguments and analysis, and you lose all that if you just plug things

into AI." These concerns suggest that while AI can assist with writing, excessive dependence on it might weaken a writer's ability to develop and refine their own skills.

Another key issue is the **loss of ownership** over the product when AI is involved. Writing is often deeply personal, and some fear that AI-generated content blurs the boundaries between their own intellectual property and that of the model. P6 expressed this concern, saying, "If I ask a tool for help on my writing, that sense of ownership gets lost, and I don't know what is my intellectual property versus a model's intellectual property." This uncertainty can be particularly troubling for writers who take pride in their original work and feel uncomfortable incorporating AI-generated suggestions.

A related and widely expressed concern is that AI **removes the "human-ness"**, or human voice, from writing. Many writers view the human voice as "special and reflective" (P2), and AI-generated text often "sounds very mechanical" (P13). Other participants echoed this sentiment, arguing that AI lacks creative integrity and fails to capture the individuality of a writer's style.

Lastly, many users found AI-generated content to be **poor in quality**, making it unreliable for serious writing tasks and diminishing the trust between the user and AI. One common issue is hallucination, where AI fabricates information that is inaccurate or misleading. As P2 observed, "They (AI) actually hallucinates. I mean, that was like striking to me how bad it was." Others have dismissed AI-generated content entirely due to its lack of coherence and originality, with P3 claiming that "no one's made anything good." This skepticism is reinforced by the presence of AI-generated books that many see as uninspired and poorly written, as P12 noted, "There are so many books that are just written completely by AI, and they're all horrible."

5.3.2.2 Desired and Acceptable Use Cases of AI

While 4 participants were strictly against AI, the remaining participants were open-minded in acknowledging the benefits of AI, and when it was acceptable to use it.

Many users acknowledge that AI can be a **valuable research tool** when used appropriately. Many participants voiced that they have previously used AI in a way similar to search engines and can be seen as beneficial for creative brainstorming, particularly in genre writing, where "professional writers...are really interested in using it for idea generation" (P3). While some have concerns about AI's accuracy, many view it as a tool that can support—but not replace—traditional research methods.

Generally, most participants voiced their support in using AI **to enhance their writing, but not to generate**. Many users found AI helpful for word choice and stylistic improvements, particularly when "you don't know exactly what word it is you want to use in a particular instance, or if you feel like you've been using a word too many times" (P8). Many emphasized that AI should function purely as a tool, and not as a way to generate "the sub-

stantial content of what you're going to deliver" (P18). Specifically, many participants label AI as a tool to **edit human writing**. Many writers appreciate its ability to correct "grammar stuff, spelling errors, or places where a sentence could use more clarity" (P4). Participants also claimed that AI provides a good **starting point for their work**. Participants often found themselves asking AI formatting questions, such as "how to phrase some parts [of the writing]" (P13) or how to "come up with a better way to say something" (P12). However, no matter what type of prompts users ask AI, the common consensus was that they would always "edit [the response] at least" (P17) or "rewrite it in [their] own words" (P12), effectively adding their own voice and flare back into the AI generate template.

Efficiency is also a major benefit of AI in writing. Many participants appreciate how AI "saves a lot of time" (P15) by speeding up the editing and proofreading process, by "proofreading and editing much faster" (P16), helping to "get things published quicker" (P16). This time-saving aspect makes AI a practical tool, especially for professionals working on tight deadlines.

5.3.2.3 Broader Impacts of AI

Zooming out, the current problems of AI in writing also have broader impacts on the world, drawing both skepticism and hope.

One major concern regarding AI's broader impact is its **breach of copyright**. It is a well known "contentious topic when it comes to training AI models on copyrighted works" (P19). Many writers worry that AI models have been "trained on their books" (P2) and "without compensation" (P3), raising ethical and legal questions about data usage. These concerns extend beyond individual cases, as some participants believed AI's training methods will never be "completely ethical for the stuff that it's trained on" (P14).

Another often overlooked issue is "**AI's environmental issues**" (P5). Training large-scale AI models requires immense computational power, which consumes significant energy and contributes to carbon emissions. The environmental footprint of AI development raises concerns about sustainability, especially as demand for AI-generated content continues to grow.

Despite these concerns, some participants viewed AI's rapid advancements with optimism. As P10 stated, "People are scared of change. But I think overall, it's an exciting time to see all of this [development] happening." This perspective highlights the dual nature of AI's impact; while AI has harmful aspects, it also opens doors to new possibilities for human-computer collaboration in various fields.

5.3.3 Qualitative Analysis: Overview of SYNthia

5.3.3.1 SYNthia's vs. Thesaurus

Like a traditional thesaurus, SYNthia **helps with repeating words** by making users more aware of their word choices. P8 noted, "Without the [tool], I would be just repeating words," while P13 shared, "I realized I was using some words too many times when I was writing." Even when users didn't immediately select a suggested word, they found value in seeing alternatives by "saving it for later" (P17). This highlights SYNthia's role in helping writers break habits of repetition and refine their language.

In addition, SYNthia was described by various participants as a **one-stop-shop**, where the writing and thesaurus are in one interface, removing the need for users to context switch during their editing / writing. Specifically, SYNthia was able to help participants "streamlining [their writing] process" (P4), "very clear cut" (P5), and "faster to have right next to [users]" (P12). The tool's integration within an existing app made it even more convenient, removing the need of "switching tabs" (P10). SYNthia's one-stop-shop quality also **sped up actions that are traditionally done manually**. Users found the tool "faster than thumbing through a thesaurus" (P11), and could "very quickly generate a list of words" (P19) for users to consider in their writing.

However, SYNthia is not always the preferred option when put head-to-head with the traditional thesaurus. Many participant found that SYNthia has **less breadth than a real thesaurus**, as it provides fewer word suggestions, with a maximum of 10 suggestions at a time. P1 claims that "one of the reasons thesaurus work is cause they give you so many suggestions," unlike SYNthia. In addition, sites like Power Thesaurus let users "click on words and explore more" (P17), allowing them to see more potential word replacements.

5.3.3.2 SYNthia's Preservation of Writer Ownership

13 out of 19 participants rated their response during the writing session as purely their own, with no AI influence (score of 10 out of 10). The remaining 6 participants rated their writing with minimal influence from AI (scores of 7-9 out of 10).

Most participants felt that **SYNthia is just a tool**, simply another representation of a "thesaurus that augments [their] thoughts" (P4), rather than an AI that generates significant amounts of content. In addition, **SYNthia's scope of influence is small**, meaning that it does not significantly alter the writing process. Existing generative AI systems create full sentences or paragraphs, with P11 making the following comparison: "ChatGPT kind of wants to do it for you. It's not really set up to be used as a tool, rather than a purely generative thing." On the other hand, SYNthia "does not edit the entirety of the work [and] just replaces words" (P9). Because SYNthia only offers word suggestions rather than dictating sentence structure or creative direction, most participants felt that their writing remained their own, with P17 stating: "the fact that it's AI was almost irrelevant."

Another key feature that contributed to the overwhelming support of writer ownership is **SYNthia's manual component**, which allows the most user control over each decision. While a handful of participants wanted parts of SYNthia to be automated (e.g. word replacement or context reading), most participants were grateful that "nothing was being automated" (P6). While SYNthia presented options, it was always "[the user's] decision as to what [word they] would choose" (P4). Because SYNthia required users to actively select words rather than generate text for them, they felt a greater sense of agency. P7 phrased it best: "I had a lot of control over not just when I prompted it, but also what words I was picking."

5.3.4 Qualitative Analysis: Context Feature

One of SYNthia's novel contributions is its ability for user input and feedback on its suggestions by allowing them to provide additional context. In the user study, participants gave a wide variety of contexts, from specifying syntax (tense or part of speech) to refining the overall tone of the phrase. A representative sample of the types of contexts can be found in Table 5.2. This section further breaks down the utility of the context-providing feature, as well as its cognitive impacts on the writers.

Writing Phrase	Context Given	Word Chosen
Across the street is a dense forest. It looks dark, foreboding .	Adjective	Ominous
Where am I? How did I get here? Am I in SquidGame ?	Shows similar to the Netflix show Squid Game	Money Heist
He threw off the ratty sheets and slid his feet into disintegrating flip-flops.	To describe sheets	Threadbare
A small motel room. Too dark to really see, but it's not somewhere I've been before.	Size of a room	Cramped

Table 5.2: Representative sample of contexts given by participants

5.3.4.1 Utility of Context

Sometimes, the initial set of suggestions generated by SYNthia provided **unsatisfactory results**, where participants "needed a word that will fit the specific context" (P13). Allowing participants to provide additional context plays a crucial role in ensuring that SYNthia provides accurate and relevant word suggestions. All but 2 participants attempted to use the context feature of the tool, mostly when they **saw potential in SYNthia in going in the right**

direction, prompting a nudge. P5 gave an example of how they use context to arrive at their desired word: "Say, I can't think of the word melancholy. So I start with sad, and then there's something in the synonyms of sad, and I'm like, 'Oh, that's closer to what I need', and then I click that, and then I look through the list. And then I eventually find melancholy." Others compared this feature to ChatGPT, explaining that it "tweaked the results a bit" (P19) and helped to "specify exactly what [they] wanted" (P12).

There are four scenarios where participants gave additional context to SYNthia. Sometimes, the initial suggestions were "using [words] as a **wrong part of speech** or it's a word that can be used in multiple ways" (P14). An example of a participant specifying the part of speech in their context can be seen with P1, who prompted SYNthia on "place" and gave the context of "esoteric noun." Similarly, other participants found that SYNthia sometimes gave them the "**wrong tense**" (P17). An example can be seen with P14, who prompted SYNthia on "woke," and provided the additional context of "past tense verb" (specifying both tense and part of speech) to ultimately get the satisfactory suggestion of "awakened." Beyond grammar, **words with multiple meanings** posed another challenge. Some users noted that SYNthia occasionally misunderstood their intent, offering synonyms based on an unintended definition of a word. P18 gave a good example: "I wanted 'lead' as the metal instead of the act of leading", with P19 similarly noting that "some words have different synonyms based on what you're describing." For many, context was also essential in "achieve a **particular feeling or tone**" (P16) in their writing. An example can be seen with P15, who prompted SYNthia on "random," gave the context "more old fashioned," and ultimately accepted the suggestion "arbitrary." In addition, P12 described how giving specific context improved the quality of suggestions, saying, "On the last word, I did [like the suggestions] where the context I gave it was 'the size of a room,' and it put out really good suggestions, like 'compact'."

The ability to specify context and *only* get suggestions in that context is yet another differentiating quality between SYNthia and a traditional thesaurus. A traditional thesaurus, as mentioned before, gives breadth, providing a masterlist of synonyms for the prompted word in all four context scenarios. On the other hand, SYNthia provides depth on one context specified by the user to avoid overwhelming them.

5.3.4.2 Cognitive Impacts of Context

While most participants found the context feature helpful, a handful of participants found the act of giving context **energy-intensive and sometimes confusing**. Whether it be phrasing or wording, some participants found it "hard to completely describe the context" (P14) and did not "feel sure that the context would make sense to the system" (P10). The effort required to frame context correctly sometimes deterred participants from fully engaging with the feature.

However, despite the difficulty of giving context, the act of giving context and seeing

the resulting suggestions **triggered critical thinking about plot and word choice**, helping writers refine their own ideas. P6 claimed that seeing the refined suggestions "prompted [them] to think about different directions the story could go in." Others found that even though they rejected certain suggestions, "the refusal to take any of the suggestions steered [them] in a particular way" (P11) and "helped [them] try to be more descriptive" (P12). The suggestions "challenged [their] thinking" (P16) in a way that helped participants "write with their instinct" (P3).

Sometimes, giving context ultimately resulted in unsatisfactory results. However, although participants may sometimes be dissatisfied with the suggestions, seeing them offered **confirmation of the effectiveness of the original word**, boosting their confidence as they continue writing. For writers who "constantly trying to second guess [themselves]" (P15), seeing the unsatisfactory suggestions "confirmed that the [original] word was probably right (the best one)" (P7). Even when users were initially uncertain about their word choice, SYNthia helped validate their instincts, with P16 explaining, "When I saw these were the more aligned words to a particular word, but it wasn't what I wanted, it kind of made me feel a little bit more confident that I had chosen the best word."

DISCUSSION

In this chapter, we discuss how well the system design and experiment adhered to the three design goals and answered the three research questions. We also discuss observations from both the quantitative and qualitative results, limitations of the experiment, and future directions this project can expand into. Overall, this section zooms out, situating observations from previous chapters into a broader scope for the future.

6.1 Research Questions

6.1.1 Usage of LLM Powered Thesaurus (RQ1)

RQ1 asked about how novice and expert writers use a LLM-powered thesaurus, like SYNthia, in the context of creative writing differently. We attempted to quantify this difference by observing the following five factors: words prompted, context given (type of context and number of context), words accepted, time spent writing, and time spent interacting with SYNthia. While there were visible differences of those factors when compared across novice and expert writers, a 2-variable independent T-test did not yield any statistical significance across the two groups. In an effort to see if grouping by writing skills was the issue, we ran the same tests again, but grouping by AI skills instead (self-rated, in Table 5.1). Unfortunately, those tests again yielded no statistically significant results.

However, despite the quantitative results not holding as much statistically significant value, there were many observations made about how the two groups (novice and expert writers) differed in their interactions with SYNthia during the writing process. Most experts prompted the tool sparsely, with a few over-prompting due to curiosity. Experts wrote in a more fluid manner, already having established habits that SYNthia might not always fit in best with. Generally, experts were either open-minded to using SYNthia (and found it

useful) or heavily against it (prompted a few times out of curiosity but accepting no suggestions). This difference may be explained by the timeline of AI development. All expert writers in this study have been writing for at least 10 years, meaning AI was not available to them during their learning stages. The recent integration of AI into their world could cause a polarizing divide in opinion and willingness to use a tool that carries many negative connotations with their colleagues. Expert writers during the user study have specifically voiced their discomfort on AI "being trained on some of [their works]" (P2), and external studies have shown that writers have expressed anger in their work being taken for LLM training without consent, credit, and compensation [68]. On the other hand, many novice writers are also currently students, surrounded by AI in both their academic and extracurricular lives, potentially influencing them to have a more positive view on SYNthia, a tool that is not at all foreign to them in the age of rapid generative AI development. We do not currently have evidence to support this hypothesis; although prior works support the conclusion that novice writers are more open minded to AI compared to expert writers [69], the influence of age and environment needs to be further studied.

The difference in formal training and experience could also be seen in how expert and novice writers utilized the context-giving feature. Experts generally knew the type of word they wanted, even if they did not know the specific word, prompting them to give definitive pieces of context to draw out a potential fit. For example, P3 (expert) wanted a medieval-style motel, prompting the word "motel" and giving the context of "19th century, portal to hell". On the other hand, novice writers often "struggled to think of what context to give" (P4), not knowing what exact type of word they are looking for, other than a better one.

6.1.2 Accepting Suggestions (RQ2)

RQ2 asked about how novice and expert writers compare in their openness to accepting LLM-generated word suggestions. From Figure 5.3, both novice and expert writers varied in their willingness to accept SYNthia's suggestions and incorporate them in their writing.

However, a participant's openness to accepting suggestions appeared to be influenced more by their attitude toward AI and writing style than by their level of expertise. For example, the novice writer with the lowest percent of accepted words was P8, who prompted SYNthia 19 times and only took one suggestion. While they were relatively open minded about AI's capabilities of "generating ideas or being used as a tool" (P8), they stressed that their normal process of coming up with suggestion was to text a friend and ask "Hey, what's that word that has like this very specific energy?" (P8). They stated that often, they have a specific word in mind, but don't know how to reach it without the external input of a friend. The friend served as a guide to them, majorly influencing their opinion. However, SYNthia is just a tool, and requires the user to understand what type of word they want in order to be the most useful. Thus, P8 did not end up accepting many of SYNthia's suggestions, likely due

to their lack of knowledge of how to best give context to the system. Similarly, P2 is the expert writer with the lowest percent of accepted words, accepting a grand total of 0 suggestions. They were also the most vocal with their distaste of AI in writing, in almost any capacity aside from research; even with research, P2 severely doubted AI capabilities and emphasized on its likelihood of hallucinating results. Being a published author and a well-established journalist, they had personal reasons (a threat to job security and distaste towards AI being trained on their works without their consent) that justified their hostility towards any sort of AI influencing their writing. Thus, they made it clear from the beginning of the interview, that they were happy to try out the system, but was ultimately influenced by prior viewpoints and confidence in their own capabilities, resulting in the 0% acceptance rate. On the other hand, P10 was an expert writer that was arguable the most open towards incorporating AI in their writing process. Despite being an English professor and understanding the malicious capabilities of AI, they proclaimed their excitement for the future of AI, stating that "it's here to stay, and I want to work with it instead of against it" (P10). Their openness to incorporating such a tool into their writing made them prompt and use the system more, ultimately taking two out of eight suggestions (25%).

Another factor that contributed to the two group's openness to accepting SYNthia's suggestions were how satisfied each participant was with them. While this was not immediately obvious in the quantitative analysis, the qualitative and interview analysis yielded a difference between how novice and expert writers ultimately decide on whether to accept a suggestion or not. Novice writers often accept suggestions if they were simply better than the word they prompted on, even if it might not have been the best. Many novices stated that they were simply prompting the system and giving context because they "did not know if there was a better word [that fix the specific context]" (P19). P19 also attempted to prompt SYNthia on its own suggestions to see if they could get even better words, though when asked, they stated that it was more out of curiosity and not because they had a clear direction. On the other hand, experts were more picky with their suggestions, often having a clear type of word they wanted. This difference in confidence could also be seen in the types of context novice writers gave. Most of the time, the contexts were some syntax change, such as tense ("make it past tense") or part of speech ("verb"), rather than tone or voice changes. Experts however, made more stylistic changes with their context, changing both the tone ("in the style of Louise Gluck") and conceptual meaning ("humanism") of the words beyond semantics.

6.1.3 Thought Process (RQ3)

RQ3 asked about how the system's suggestions affected the thought processes of novice and expert writers. In Chapter 5, we discussed the findings of the cognitive impacts of giving context and observing the resulting suggestions, although the impact of such interactions

differed between novice and expert writers.

In general, SYNthia's suggestions had less impact on the thought process of expert writers, and more impact on the thought process of novice writers. In interviews, four out of the seven experts stated that the seeing the suggestions did not affect their writing process, and three out of seven experts stated that the context feature did not help them. On the other hand, only three out of twelve novice writers stated that seeing the suggestion did not affect their writing process, and only four out of twelve novices stated that the context feature did not help them.

This difference in the influence of SYNthia's suggestions can be explained by how novice and expert writers view the SYNthia's role as. Novice writers tend to treat AI tools as authoritative, while experts treat the same tools as more of a "junior collaborator" [69] – establishing a difference in power dynamics and trust when it comes to the two groups of writers. Novices often place their own judgment below that of the system, while experts tend to be more confident in making the final decision. Due to this power dynamic, studies show that novice writers who are less self-efficient, tend to over-delegate responsibilities in their writing process to the AI [70], corresponding to a weaker stylistic confidence. In addition, because experts view their knowledge above the AI's, it highlights the metacognitive deficits of novice writers, who struggle with identifying tonal mismatches in AI suggestions with their own writing, often resulting in their own work converging to the default AI voice [22]. On the other hand, experts have developed their personal linguistic intuition, allowing them to actively reject AI suggestions that clash with their style, hence why the "junior collaborator" relationship with AI suggestions make sense. Expert writers also prioritize maintaining authentic creative processes over output optimization, with one study showing that 74% of their interviewed professional writers rejected AI-generated suggestions that disrupted their established drafting/editing rituals, even when outputs improved efficiency [22].

In the age of rapid AI development, it was important to us to create a system that encouraged critical thinking and protected writer cognition [71], which was the initial motivation behind creating SYNthia. Ultimately, we found success with this goal, with the majority of participants stating that seeing the LLM-generated suggestions and engaging with the system through providing additional context encouraged them to cognitively engage more with their writing. Prior studies have shown that AI provocations can help induce critical thinking in writers [72], particularly student writers who actively try to seek out new ideas or directions with AI assistance [73]. At the end, the participants got out however much effort they put in. For those who "got lazy" (P6) with their engagement, the results were not as influential as those who actively used the system with the goal of improving their story. This conclusion proves that ultimately, with any AI-powered tool, the interactions between user and tool have the greatest effect on the final product. One cannot simply blame the tool for affecting the product, and must look into the interactions with it as well.

6.2 Design Goals

SYNthia was built with the design goals from Chapters 3 and 4 in mind. In this section, we revisit these design goals and reflect on how SYNthia was able to address each one.

6.2.1 Coherence to Context (DG1)

DG1 specifies that SYNthia's suggestions should be coherent and relevant to the context of the user's writing. SYNthia was able to meet this goal, with 14 out of 19 participants voicing their satisfaction with the overall suggestions they received. Participants stated that the tool would always output suggestions that aligned with the additional context they provided, which matched their mental models of how an AI-powered thesaurus would function.

However, some participants wanted SYNthia to also take "the context of the text before it, like ChatGPT" (P14) when they were dissatisfied with the initial suggestions. We discuss this more in Section 6.5.2, but we wanted the first set of generated suggestions to mimic those by a traditional thesaurus, which does not take into account any context from the user's personal writing. We made the purposeful decision to encourage participants to engage with the context-giving feature.

6.2.2 Steerability (DG2)

DG2 specifies that SYNthia should be steerable, adapting their outputs based on the inputs (context) provided by the users. In our project, the context-giving feature is able to address both DG1 and DG2, with SYNthia's steerability causing its coherence to context. Steerable systems foster a healthy environment for co-creativity while maintaining human agency, a goal that SYNthia was able to meet. Across all 19 participants, the average rating of their ownership during the writing task was 9.6 out of 10, indicating that the majority of participants viewed the final product as (essentially) fully their own, despite the minor influences from SYNthia. In addition, participants found SYNthia easy to use, allowing them to manually trigger words and provide context to, as one participant phrased it, "nudge the system in the right direction" (P7).

Giving participants the ability to steer, or nudge, the suggestions did result in overall user satisfaction. 12 out of 19 participants indicated that giving context to the tool changed the suggestions towards the direction they were hoping. However, P17 observed that certain types of context, such as specifying tense or part of speech, led to more significant changes than when participants aimed to fine-tune or adjust a word's tone.

6.2.3 Divergent Outcomes (DG3)

DG3 specifies that SYNthia will generate a diverse range of outcomes that all fit the user prompt / context. From Chapter 5, participants were overall satisfied with the range of suggestions they received. However, some compared the outputs of SYNthia to those of a traditional thesaurus, noting that a traditional thesaurus outputs a significantly larger number of suggestions, even though only a handful of them might be a good fit for the current user context.

There is a tradeoff between a tool like SYNthia that relies on AI API tokens and breadth. The more suggestions that SYNthia outputs, the more resources (physical and monetary) are needed, which is usually not a problem for a normal thesaurus. However, a normal thesaurus does not have the additional option of personalizing the suggestions through giving additional context, effectively fine-tuning the results. There is also a tradeoff between breadth and depth. Too much breadth can potentially overwhelm users, but too much depth might not allow users to see as many suggestions as they want.

6.3 Additional Factors that Contribute to AI-Writer Interactions

While writing expertise and technical proficiency with AI tools are often considered primary determinants of successful human-AI collaboration, emerging evidence throughout the user study suggests contextual framing [69] and task specific needs play an equally critical role in shaping interactions. These additional axes—the ability to structure prompts with domain-specific context and task specific awareness—potentially explain the significant variance in outcomes that writing and AI skills alone cannot.

6.3.1 Contextual Framing

No matter the writer's level of expertise, both novice and expert writers voiced their confusion in phrasing context at times. Writers who can effectively frame their queries with relevant contextual information tend to receive more accurate and useful suggestions. While expert writers had a better understand of the specific word they wanted, some lacked the distinct skill of framing the context in a way that would be optimal for the LLM tool. At first, we suspected that contextual framing would be similar to prompt engineering, but looking at a division by AI skill, again, did not yield significant differences. Therefore, we suspect that the skill of phrasing context transcends both writing and AI expertise, suggesting a distinct "contextual intelligence" axis [74]. Examples of experts in contextual intelligence are those who sufficiently embody:

- **Iterative refinement:** Progressively narrowing context across multiple AI interactions

- **Metacognitive framing:** Explicitly stating intent (e.g., "Suggest metaphors for grief, avoiding clichés")

A prior study found that mastering these techniques correlates with 42% higher output relevance compared to single-step prompting [69], which would be an appropriate axis to study in the future.

6.3.2 Task Specific Needs

Throughout the user study, we found that the writer's primary domain, profession, and goals heavily impacted how they interacted with SYNthia.

First, the writer's preferred domain or profession influence the user's need for alternative words. In our user study, we interviewed writers across different domains, including poets, journalists, and academic writers. P11, a poet, noted that his profession required him to be in constant need for alternative word suggestions, more than other writing professions, which match his behavior, prompting SYNthia 25 times during the 15 minute writing task. Poets approach AI-powered thesauruses as a "procedural palette" for experimental writing, often seeking unusual or context-specific language [75]. Poets value AI-powered thesauruses for their ability to generate neologisms and provide unexpected phrasings, enhancing their artistic exploration. P2, a journalist, noted that they used external sources only for research purposes, being careful to fact check all outputs for potential hallucinations or factual inaccuracies. While P2 did not specifically need alternative words for their work, journalists could benefit from AI-powered thesauruses to enhance the clarity and readability of their work for a more general audience. P13, a computer science graduate student with several publications, noted that academic writing is "more boring," with the focus being on communicating highly technical information instead of creating an attention-grabbing narrative. Academic writers tend to follow certain guidelines and target similar audiences (varies depending on the field) [76], requiring a particular type of language. However, there is growing concern about the impact of AI on academic language, with studies noting a detectable increase in "flowery" language in papers written with AI assistance [77], taking away from the real purpose of conveying clear, scientific research.

Another potential axis that falls under the category of "task specific needs" is the difference in intent for using an AI-powered thesaurus like SYNthia, specifically the intent of exploration vs. precision. Some writers use thesauruses for exploratory purposes, allowing unexpected words to influence their creative direction [78]. Multiple participants indicated that they used SYNthia's suggestions as a way to "help shape the direction" (P15) as way to inspire new ideas and shift the trajectory of their work during the drafting process. On the other hand, some writers seeking precision use thesauruses to refine language and select words that convey exact meanings. For example, academic writers and journalists prioritize conciseness, clarity, and accuracy in their writing over flowery language [79].

6.4 Limitations

Although the system design and user study yielded several novel observations, there are also limitations that prevented the study from reaching its full potential.

6.4.1 Experimental Design

During the user study, participants were given a universal writing prompt and 15 minutes to respond to it with the assistance of SYNthia. The initial reasoning behind having the same prompt across all users was to eliminate any discrepancies if we chose to analyze the content of the writing. However, throughout the interview process, some participants pointed out the 1) potential absence of the necessity for descriptive words and 2) overemphasis focus on writing instead of editing due to the new prompt.

A handful of participants (P2 being a strong advocate) noted that the writing prompt did not carry much need for descriptive words. Instead, the prompt focused on world building and story telling, which influenced participants to write a moving plot line rather than a descriptive piece. While the prompt itself did say to "describe the world around you," the remaining parts of it focused more on describing actions and movements of the character.

In addition, requiring participants to write to a completely new prompt with a limited time (15 minutes) might have shifted the focus away from engaging with the system for editing, and instead placed the main focus on coming up with a cohesive piece before the time ran out. When asked to rate their own writing on a scale of one to ten, most participants rated themselves a five or below, indicating a common theme of dissatisfaction with their work under time pressure. In addition, participants came from various writing backgrounds, including journalism, creative writing, and poetry to name a few. Even though I indicated that the prompt was loose and just a suggestion, it was unfair to force participants to write in a form that might not have been their preferred choice. If this study were to be run again, I would encourage participants to bring their own writing. This way, participants can focus on the editing stage (what SYNthia is best at) and also be more motivated to use the system to edit a piece of work that they are more invested in.

6.4.2 Overreliance on Pretrained GPT-4 Model

While the integration of backend and frontend flowed seamlessly, the backend generation of suggestion and processing of context ultimately relied completely on GPT-4. P5, in particular, mentioned how AI models have a black box nature that makes the results less interpretable, potentially leading to lower user satisfaction with the suggestions. Their concerns could be addressed if we had trained a model from scratch that specifically targeted the task of finding synonyms, incorporating a way for the model to output its thought process along the way (or at the very least, gain a better understanding of the model training done). How-

ever, given the strict time constraints and lack of training resources (large datasets, GPU hosts, etc.), it would have been unwise to train a completely new LLM from scratch, which ultimately resulted in SYNthia overrelying on GPT-4 for all the suggestion generations, despite its black-box nature. This choice limited our understanding of the model's training data, and how it was able to process each user input.

6.5 Future Works

While some future possibilities regarding the axis of division have been discussed in this chapter, this section offers a few more directions this project can go.

6.5.1 Additional Interface Features

The version of SYNthia described in this thesis builds upon a 2023 prototype. While there has been drastic improvement in the accuracy of suggestions and overall efficiency, there is still room for more improvement. Participants generally viewed SYNthia as a better version of a thesaurus, which prompted intense comparisons between this tool and a traditional thesaurus. Below, we discuss some future improvements that surfaced during the user study.

1. **History of prior prompts:** Some participants mentioned that seeing suggestions sometimes "gave them ideas for later" (P17). Thus, a history of all prior prompts, context given, and resulting suggestions could prove useful for those users. Because each suggestion comes in its own little box, a possible way to represent user prompting history could be to have a collapsed list that could be opened or expanded with a manual click.
2. **More breadth of suggestions:** While most participants enjoyed the concentrated list of suggestions provided by SYNthia, a handful of participants voiced their dislike towards the smaller list, claiming that "one of the reasons thesaurus work is cause they give [users] so many suggestions, [unlike SYNthia]" (P1). However, an initial motivation behind the 1-10 toggle for suggestion length was to avoid overwhelming the users. A potential solution to satisfy both users who want a more concentrated list and those who prefer more depth is to implement a "show more" option within each suggestion box. With this solution, users who want a more diverse, longer list of results could manually trigger for that, and those who do not want to be overwhelmed can choose not to.
3. **Built-in dictionary:** During the writing exercise, some participants stated that they sometimes encountered words they have never seen before. Creating a built-in dictionary to provide definitions for each term when manually triggered (e.g. via a click) could help users who might not have seen all those words before.

4. **Phrases:** Some participants stated that a traditional thesaurus interface allows them to "type phrases, whereas [SYNthia] only lets you hit the last word" (P17). A future implementation could be adding a grouping method (via parentheses or brackets) to allow for phrase prompting as well.
5. **Alignment scale:** Some participants stated that a traditional thesaurus provides information on how close each suggestion aligns with the user's criteria. To avoid overwhelming the user with too many functionalities, we can implement a colored bar (green for good, yellow for okay, and red for bad) with the predicted alignment to that suggestion and how well it meets to user's context.

6.5.2 Manual vs. Automatic Triggers

There were conflicting opinions on whether some features of the tool should remain manual, or instead be automatic. As a functionality, most participants did not complain about the click to copy+paste feature. However, a handful of participants wanted the click of a suggestion to automatically replace the original word in the text box. This functionality change could absolutely be implemented in further studies. However, changing SYNthia's word replacement method from manual to automatic could also affect ownership, a relationship that would be interesting to study in a future project.

In addition, some participants wanted "the rest of the text was being read" (P1) automatically as context when they initially trigger SYNthia on a word. The original motive behind *not* taking any context during the initial trigger was to have the original trigger yield results that were similar to the initial results from a traditional thesaurus. However, the effects of automatically reading everything in the text box as context could be a potential future direction of study for this project.

6.5.3 Integration with Existing Document Editors

Many participants wished they had more "time to play around with the tool" (P10), and many participants found themselves prompting the system more out of curiosity. Because this was a new interface, it was understandable that participants had a stronger urge to explore and learn the tool, rather than immediately finding usefulness for it. This was also briefly touched on in the pilot study, but a potential solution to the strangeness of a new interface is to integrate SYNthia as an extension to an existing document editor, such as Google Docs or Microsoft Word. Integrating SYNthia as an extension makes the tool more familiar, since it is now building upon an interface that the user already uses, rather than being a completely new entity.

CONCLUSION

AI has slowly been worming its way into multiple fields, posing as a threat and competitor to the existing human workers, many of which fear that their role will eventually be replaced by a machine, writers especially. *Death of an Author* is a reasonable baseline of how AI, specifically LLMs, write, and by no means is it good. The story makes strange comparisons / metaphors and overall lacks a human voice. Currently, creativity is sacred to the human race, something that AI simply does not have [80]. LLMs require a human overseer to prompt them, a distinct feature that characterizes them as tools.

With the prominence of AI in our everyday lives, as computer developers, we bear the responsibility of creating AI-powered tools solely for the purpose of assisting, and not replacing, genuine human efforts. SYNthia was developed with this responsibility in mind, aiming to combine two fields (AI and creative writing) that are at odds with each other by conflicts of interest. While SYNthia's impact is not significant enough to marry the two conflicting fields, it is able to show writers the capabilities of AI to assist them, and show developers how to create useful tools that do not pose a threat for writers. The user study yielded results that show word level edits from a tool do not interfere with writer ownership at all, with most writers voicing their support of such an AI-powered thesaurus being on the product market in the future. The results of this user study also open doors to further study on what writers value in their work, as well as what makes the human voice prominent and attractive in a piece of writing.

The trust of writers in LLM-powered tools is hard to earn, particularly those who also have a background in AI. AI hallucinations are not uncommon, with each inaccurate output pushing writers further away from seeing them as a valuable tool. If used irresponsibly, AI can dominate a project, producing low-quality results and undermining human effort, ultimately creating a toxic environment where every piece of writing is scrutinized for AI involvement. These concerns are addressed by SYNthia's two main contributions:

1. SYNthia is a starting point to repairing the distrust between writers and AI tools by showing writers that AI-powered tools can generate satisfactory results, while not taking away from human agency.
2. SYNthia provides an example of how developers can create AI tools with the writers' best interest in mind.

The introduction of SYNthia into the world will continue to foster trust between writers and AI, making progress towards a world where the two fields live more harmoniously without taking away from either.

APPENDIX A - PILOT PROMPTS AND USER STUDY

Prompts

Trial 1: Control

Describe your ideal habitat. If you stepped outside right now, what weather, season, geographic location, etc. would make you happiest? Describe these as best you can.

Trial 2: Experimental

You, the main character of your own fantasy story, just left town and sold everything you own. Where would you go, what would you do? Why? Describe your journey and destination as best you can.

GPT 4 Prompt

"From now on, give me **numSuggestions** distinct synonyms for the word **word**. Format your response as an array, for example: ["word1", "word2", "word3", "word4"]. Do not include any other information in your response. If you cannot come up with **numSuggestions**, provide as many as you can. All of my future messages will provide extra context for the word, you should incorporate them into your suggestions. The words you respond with should ALWAYS be synonyms for **word**."

numSuggestions = Number of suggestions as selected in slider

word = Word for which synonyms are to be generated

User Study

Pre-Interview Questions - Required

1. Age, gender (or choose not to disclose), affiliation (college, etc.)
2. How often do you produce pieces of writing, academic or otherwise? What type of writing do you usually engage in? (e.g. poetry, short story, research, etc.)

3. Do you typically use AI assistance for writing or other tasks? If so, which tasks, and how?
4. How often do you find yourself looking for alternative words to use during your writing process?

Post-Interview Questions - Required

1. What was the most difficult part of the interface to work with?
2. Were there any significant moments of frustration or confusion you encountered when using our interface, if any?
3. Were you ever frustrated/annoyed with the number of times you needed to provide additional context for your word suggestions, if at all?
4. Do you think the use of this interface improved the quality of your writing, compared to if you were to write without any external help?
5. Now having used this interface, how upset would you be if you never had the opportunity to use it, or something similar, again?

Post-Interview Question - Optional Conversations

1. Do you usually revise during/after writing? Would this change your process at all?
2. What would you add or change about the interface if you had the opportunity?
3. How does this compare with other interfaces you've used, if any? (i.e. GPT, Grammarly, manual lookup of synonyms)
4. How disappointed would you be if an interface like this never existed?

APPENDIX B - FORMAL PROMPTS AND USER STUDY

Prompts

Writing Prompt

You woke up in a motel room with a strange item in your hand. What do you do? Where do you go? Describe the world around you and your actions. Choose between a) a cell phone that can only call one specific number, b) a GPS to a mystery location 1 mile away, or c) a sticky note with a list of seemingly random items, written in dried blood.

GPT 4 Prompt

From now on, give me **numSuggestions** distinct synonyms for the word **word**. You MUST format your response as an array, for example: ["word1", "word2", "word3", "word4"]. Do not include any other information in your response. If you cannot come up with **numSuggestions**, provide as many as you can. All of my future messages will provide extra context for the word, you should incorporate them into your suggestions. The words you respond with should ALWAYS be synonyms for **word** and match the additional context.

numSuggestions = Number of suggestions as selected in slider

word = Word for which synonyms are to be generated

User Study

Participant Recruitment Survey Questions - Required

1. First Name, Last Name (optional)
2. Email
3. Age (optional)
4. Gender (optional)
5. If you are a student, what do you study?

6. What is your editing style? Do you edit as you write, or do you edit after drafting the full piece?
7. How often do you find yourself thinking about alternative words during writing?
8. Do you use any writing aids while you write (thesaurus / generative AI / Grammarly)? Please list them below.
9. On a scale of 1 (worst) to 10 (best), how would you rank your creative writing skills? To help guide you: a 5 is equivalent to the skills of someone who is coming out of a MFA program (masters degree in writing).
10. How many years of creative writing experience do you have?
11. Have you received formal education in creative writing (e.g. degrees, classes, etc.)? If so, please list them here.
12. How often do you use AI?
 - a) If your answer was not "Never" to the previous question, please list your prior AI experiences.
13. How comfortable / confident do you feel with using generative AI and prompting them to get satisfactory results (1 - least comfortable; 10 - most comfortable)?

Pre-Interview Questions - Required

1. How many hours per week do you write on average?
2. Do you have any published works and/or manuscripts?
3. What are your opinions on using AI in creative writing?

Post-Interview Questions - Required

1. How did you decide when to give extra context?
2. Did you think giving context changed the suggestions significantly?
3. Did you feel like giving context to SYNthia was similar to
4. prompting a generative AI (e.g. ChatGPT)?
5. Were there any instances where you didn't take the specific word suggestion, but drew inspiration from the given suggestions?
 - a) Were you satisfied with the word suggestions given by SYNthia?

6. How satisfied are you with the final product, from a scale of 1 to 10, with 10 being most satisfied (no notes) and 1 being least satisfied? Explain.
 - a) Do you think the use of SYNthia improved the quality of your writing, compared to if you were to write without any external help?
7. Do you feel like this is your piece of writing, even though you used AI assistance? Rate your ownership level from 1 to 10, with 10 being “I feel like this writing was purely produced by me”, and 1 being “I feel like this writing is completely AI generated”.

BIBLIOGRAPHY

- [1] Stephen Marche.
Death of an author, 2025.
Accessed: 12 March 2025.
- [2] Alexandra Alter.
Aidan marchine’s “death of an author” and the future of ai-generated books, May 2023.
The New York Times, Accessed: 12 March 2025.
- [3] Rachel Guo.
Music review: To the critics, whatever, 2024.
Granite Bay Today, Accessed: 12 March 2025.
- [4] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood.
A visual-language foundation model for computational pathology.
Nature Medicine, 30:863–874, 2024.
- [5] Shan Zhou, Xiao Luo, Chan Chen, Hong Jiang, Chun Yang, Guanghui Ran, Juan Yu, and Chengliang Yin.
The performance of large language model-powered chatbots compared to oncology physicians on colorectal cancer queries.
International Journal of Surgery, 110(10):6509–6517, October 2024.
- [6] Md Sultanul Islam Ovi, Nafisa Anjum, Tasmina Haque Bithe, Md. Mahabubur Rahman, and Mst. Shahnaj Akter Smrity.
Benchmarking chatgpt, codeium, and github copilot: A comparative study of ai-driven programming and debugging assistants.
arXiv, 2024.
- [7] Shota Shiiku, Raja Marjieh, Manuel Anglada-Tort, and Nori Jacoby.
The dynamics of collective creativity in human-ai social networks.
2025.
- [8] Alejandro Hernandez, Levin Brinkmann, Ignacio Serna, Nasim Rahaman, Hassan Abu Alhaija, Hiromu Yakura, Mar Canet Sola, Bernhard Schölkopf, and Iyad Rahwan.

- Alien recombination: Exploring concept blends beyond human cognitive availability in visual art.
2024.
- [9] Grammarly.
Grammarly ai, 2025.
Accessed: 13 March 2025.
- [10] Grammarly.
Grammarly ai writer, 2025.
- [11] OpenAI.
Gpt-4 technical report.
Technical report, OpenAI, 2023.
- [12] Yiran Ding, Li Lina Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang.
Longrope: Extending llm context window beyond 2 million tokens.
2024.
- [13] Daniel Felix.
Best ai writing assistants in 2025: Which one should you use?, March 2024.
Accessed: 14 March 2025.
- [14] Alicia Guo, Shreya Sathyanarayanan, Leijie Wang, Jeffrey Heer, and Amy Zhang.
From pen to prompt: How creative writers integrate ai into their writing practice.
11 2024.
- [15] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski.
Homogenization effects of large language models on human creative ideation.
In *Creativity and Cognition*, CC '24, page 413–425. ACM, June 2024.
- [16] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch.
The ai ghostwriter effect: When users do not perceive ownership of ai-generated text but self-declare as authors.
ACM Transactions on Computer-Human Interaction, 31(2):1–40, February 2024.
- [17] Samir Passi and Mihaela Vorvoreanu.
Overreliance on ai: Literature review.
Technical Report MSR-TR-2022-12, Microsoft Corporation, 2022.
- [18] Shoumik Saha and Soheil Feizi.
Almost ai, almost human: The challenge of detecting ai-polished writing.
2025.
- [19] David James Woo, Hengky Susanto, Chi Ho Yeung, and Kai Guo.

- Approaching the limits to efl writing enhancement with ai-generated text and diverse learners.
2025.
- [20] Double Care ABA.
Supporting emotional expression through storytelling and writing.
- [21] Christopher Candlin and Ken Hyland.
Writing: Texts, Processes, and Practices.
Longman, London; New York, 1999.
- [22] Angel Hsing-Chi Hwang, Q. Vera Liao, Su Lin Blodgett, Alexandra Olteanu, and Adam Trischler.
"it was 80% me, 20% ai": Seeking authenticity in co-writing with large language models.
arXiv:2411.13032 [cs.HC], Nov 2024.
Submitted on 20 Nov 2024.
- [23] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch.
The ai ghostwriter effect: When users do not perceive ownership of ai-generated text but self-declare as authors.
In *ACM Transactions on Computer-Human Interaction (TOCHI)*, volume 30, pages 1–20, Dec 2023.
- [24] Lise Saffran, Sisi Hu, Amanda Hinnant, Laura D Scherer, and Susan C Nagel.
Constructing and influencing perceived authenticity in science communication: Experimenting with narrative.
PLoS One, 15(1):e0226711, 2020.
- [25] David Zhou and Sarah Stermann.
Ai.llude: Encouraging rewriting ai-generated text to support creative expression.
May 2024.
Submitted on 28 May 2024.
- [26] Russell W. Belk.
Possessions and the extended self.
Journal of Consumer Research, 15(2):139–168, Sep 1988.
- [27] Mihaly Csikszentmihalyi, Eugene Rochberg-Halton, and Ellen Halton.
The Meaning of Things: Domestic Symbols and the Self.
Cambridge University Press, 1981.
- [28] Yang A and McDonnell L.
Student definitions of ownership and perceived ways ownership influences writing in a biology laboratory class.
Journal of Microbiology Biology Education, 25:e00197–23, 2024.

-
- [29] Adam M. Persky and John D. Robinson.
Moving from novice to expertise and its implications for instruction.
American Journal of Pharmaceutical Education, 81(9):6065, 2017.
 - [30] Anni Holila Pulungan.
The cognitive process of writing in english: Developing the cognitive-based learning model.
2006.
 - [31] Anna Becker.
A review of writing model research based on cognitive processes.
Education, Linguistics, 2006.
 - [32] Deborah Mccutchen.
From novice to expert: Implications of language skills and writing- relevant knowledge for memory during the development of writing from novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill.
Journal of Writing Research, 3, 07 2011.
 - [33] April Nelms and Miriam Segura-Totten.
Expert-novice comparison reveals pedagogical implications for students’ analysis of primary literature.
CBE life sciences education, 18:ar56, 12 2019.
 - [34] Susan Barber.
Making metacognition part of student writing, November 2022.
 - [35] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen.
A survey of large language models.
2025.
 - [36] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu.
Harnessing the power of llms in practice: A survey on chatgpt and beyond.
ACM Trans. Knowl. Discov. Data, 18(6), April 2024.
 - [37] Exploding Topics.
Gpt parameters: Everything you need to know, 2025.
 - [38] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli.
Understanding the capabilities, limitations, and societal impact of large language models.

- 2021.
- [39] Index.dev.
What is llm?, 2024.
- [40] Walid S. Saba.
Stochastic llms do not understand language: Towards symbolic, explainable and ontologically based llms.
In *Conceptual Modeling: 42nd International Conference, ER 2023, Lisbon, Portugal, November 6–9, 2023, Proceedings*, page 3–19, Berlin, Heidelberg, 2023. Springer-Verlag.
- [41] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan.
Ai-augmented brainwriting: Investigating the use of llms in group ideation.
In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.
- [42] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui.
Charactermeet: Supporting creative writers' entire story character construction processes through conversation with llm-powered chatbot avatars.
In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.
- [43] Bahareh Sarrafzadeh, Sujay Kumar Jauhar, Michael Gamon, Edward Lank, and Ryen W. White.
Characterizing stage-aware writing assistance for collaborative document authoring.
Proc. ACM Hum.-Comput. Interact., 4(CSCW3), jan 2021.
- [44] Quentin Roy, Sébastien Berlioux, Géry Casiez, and Daniel Vogel.
Typing efficiency and suggestion accuracy influence the benefits and adoption of word suggestions.
In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.
- [45] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith.
Creative writing with a machine in the loop: Case studies on slogans and stories.
In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 329–340, New York, NY, USA, 2018. Association for Computing Machinery.
- [46] Katy Ilonka Gero and Lydia B. Chilton.
Metaphoria: An algorithmic companion for metaphor creation.
In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.
- [47] Melissa Roemmele and Andrew S. Gordon.
Automated assistance for creative writing with an rnn language model.

- In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, IUI '18 Companion, New York, NY, USA, 2018. Association for Computing Machinery.
- [48] Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont.
Synthetic literature: Writing science fiction in a co-creative process.
In Hugo Gonalo Oliveira, Ben Burtenshaw, Mike Kestemont, and Tom De Smedt, editors, *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, pages 29–37, Santiago de Compostela, Spain, September 2017. Association for Computational Linguistics.
- [49] Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight.
Generating topical poetry.
In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191, Austin, Texas, November 2016. Association for Computational Linguistics.
- [50] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert.
Shaping human-ai collaboration: Varied scaffolding levels in co-writing with language models.
In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.
- [51] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich.
Soylent: A word processor with a crowd inside.
Commun. ACM, 58(8):85–94, jul 2015.
- [52] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan “Michael” Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams.
Abscribe: Rapid exploration and organization of multiple writing variations in human-ai co-writing tasks using large language models.
page 1–18, May 2024.
- [53] Yu Zhang, Kexue Fu, and Zhicong Lu.
Revtogether: Supporting science story revision with multiple ai agents.
03 2025.
- [54] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin.
The value, benefits, and concerns of generative ai-powered assistance in writing.
2024.
- [55] Carlos Alves Pereira, Tanay Komarlu, and Wael Mobeirek.

- The future of ai-assisted writing.
arXiv, 2023.
- [56] Cecilia Chan.
Is ai changing the rules of academic misconduct? an in-depth look at students' perceptions of 'ai-giarism'.
06 2023.
- [57] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist.
The hallmark effect: Supporting provenance and transparent use of large language models in writing with interactive visualization.
CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.
- [58] Jessica He, Stephanie Houde, and Justin Weisz.
Which contributions deserve credit? perceptions of attribution in human-ai co-creation.
02 2025.
- [59] et. al. Jillian M. Buriak, Deji Akinwande.
Best practices for using ai when writing scientific manuscripts: Caution, care, and consideration: Creative science depends on it.
Journal of Scientific Writing and Communication, 2024.
- [60] Ryotaro Toba, Tsuneo Kato, and Seiichi Yamamoto.
Efficient speech-recognition error-correction interface for japanese text entry on smartwatches.
In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA, 2019. Association for Computing Machinery.
- [61] Xiang Ao, Xugang Wang, Feng Tian, Guozhong Dai, and Hongan Wang.
Crossmodal error correction of continuous handwriting recognition by speech.
In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, page 243–250, New York, NY, USA, 2007. Association for Computing Machinery.
- [62] Maozheng Zhao, Wenzhe Cui, IV Ramakrishnan, Shumin Zhai, and Xiaojun Bi.
Voice and touch based error-tolerant multimodal text editing and correction for smartphones.
In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, page 162–178, New York, NY, USA, 2021. Association for Computing Machinery.
- [63] Xugang Wang, Junfeng Li, Xiang Ao, Gang Wang, and Guozhong Dai.
Multimodal error correction for continuous handwriting recognition in pen-based user interfaces.

- In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, IUI '06, page 324–326, New York, NY, USA, 2006. Association for Computing Machinery.
- [64] Jun Ogata and Masataka Goto.
Speech repair: quick error correction just by using selection operation for speech input interfaces.
In *Proc. Interspeech 2005*, pages 133–136, 2005.
- [65] Nicholas Davis.
Human-computer co-creativity: Blending human and computational creativity.
01 2013.
- [66] Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi.
Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing.
In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 436–452, New York, NY, USA, 2023. Association for Computing Machinery.
- [67] Allan MacLean, Richard M. Young, Victoria M. E. Bellotti, and Thomas P. Moran.
Questions, options, and criteria: Elements of design space analysis.
Hum.-Comput. Interact., 6(3):201–250, sep 1991.
- [68] Katy Gero, Meera Desai, Carly Schnitzler, Nayun Eom, Jack Cushman, and Elena Glassman.
Creative writers' attitudes on writing as training data for large language models.
09 2024.
- [69] Max Kreminski.
The dearth of the author in ai-supported writing.
2024.
- [70] Philipp Spitzer, Joshua Holstein, Patrick Hemmer, Michael Vössing, Niklas Köhl, Dominik Martin, and Gerhard Satzger.
Human delegation behavior in human-ai collaboration: The effect of contextual information.
2025.
- [71] Anjali Singh, Karan Taneja, Zhitong Guan, and Avijit Ghosh.
Protecting human cognition in the age of ai.
2025.
- [72] Ian Drosos, Advait Sarkar, Xiaotong Xu, and Neil Toronto.
"it makes you think": Provocations help restore critical thinking to ai-assisted knowledge work.
2025.

- [73] Khonzoda Umarova, Talia Wise, Zhuoer Lyu, Mina Lee, and Qian Yang.
How problematic writer-ai interactions (rather than problematic ai) hinder writers' idea generation.
2025.
- [74] Alicia Guo, Pat Pataranutaporn, and Pattie Maes.
Exploring the impact of ai value alignment in collaborative ideation: Effects on perception, ownership, and output.
2024.
- [75] Alex Calderwood, John Joon Young Chung, Yuqian Sun, Melissa Roemmele, and Max Kreminski.
Phraselette: A poet's procedural palette.
2025.
- [76] Hai Dang, Chelse Swoopes, Daniel Buschek, and Elena Glassman.
Corpusstudio: Surfacing emergent patterns in a corpus of prior work while writing, 03
2025.
- [77] Javier Conde, Pedro Reviriego, Joaquín Salvachúa, Gonzalo Martínez, José Alberto Hernández, and Fabrizio Lombardi.
Understanding the impact of artificial intelligence in academic writing: Metadata to the rescue.
Computer, 57(1):105–109, January 2024.
- [78] Katy Gero and Lydia Chilton.
How a stylistic, machine-generated thesaurus impacts a writer's process.
pages 597–603, 06 2019.
- [79] English Education Lab.
Seven ways using a thesaurus can transform your writing and boost your vocabulary,
February 2024.
- [80] Wired.
Death of an author: Ai book review, 2023.