DOI:10.1145/3626952

To view the accompanying Technical Perspective,
visit doi.acm.org/10.1145/3655634 tp

LaMPost: AI Writing Assistance for Adults with Dyslexia Using Large Language Models

By Steven M. Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N. Horne, Michal Lahav, Robert MacDonald, Rain Breaw Michaels, Ajit Narayanan, Mahima Pushkarna, Joel Riley, Alex Santana, Lei Shi, Rachel Sweeney, Phil Weaver, Ann Yuan, and Meredith Ringel Morris

Abstract

The natural language capabilities demonstrated by large language models (LLMs) highlight an opportunity for new writing-support tools that address the varied needs of people with dyslexia. We present LaMPost, a prototype email editor that draws upon our understanding of these needs to motivate AI-powered writing features, such as outlining main ideas, generating a subject line, suggesting changes, and rewriting a selection. We evaluated LaMPost with 19 adults with dyslexia, identifying promising routes for further exploration (such as the popular "rewrite" and "subject line" features), while also finding that the current generation of LLMs may not yet meet the accuracy and quality thresholds to be useful for writers with dyslexia. In addition, knowledge of the AI did not alter participants' perception of the system nor their feelings of autonomy, expression, and self-efficacy when writing email messages. Our findings provide insight into the benefits and drawbacks of LLMs as writing support for adults with dyslexia, and they offer a foundation to build upon in future research.

1. INTRODUCTION

Dyslexia refers to a cluster of symptoms that result in challenges with word recognition, reading fluency, spelling, and writing population.[8] While people with dyslexia may find compensatory strategies to lessen their reading difficulties by adulthood, the combination of reading, comprehension, and planning skills needed for writing may lead to ongoing difficulties.[14] Writers with dyslexia describe a variety of challenges, including spelling, grammar, organization, matching a desired tone, and expressing themselves with clarity and precision.[3,14,15] To overcome these obstacles, they may try various strategies—speech-to-text, templates, help from friends and family—but these can add complexity to an already protracted writing process.[3]

Prior work has explored various accessibility approaches to overcome the reading challenges associated with dyslexia, such as experimenting with text presentation[19] and synonym substitution for complex words.[17] However, efforts to address writing challenges associated with dyslexia have primarily focused on lower-level interventions, such as automatic word suggestions[11] and specialized spellcheck tools.[18] AI-based efforts, when present, have continued this thread; for example, Wu et al. evaluated a dyslexia-tuned Neural Machine Translation model for spelling and grammar support on social media posts.[23] However, automatic tools that can support people with dyslexia with sentence or paragraph-level writing difficulties—including organization, expression, and voice—are absent from accessibility literature. This gap presents an opportunity to explore new writing-support tools that leverage the generative capabilities of neural language models.

The original version of this paper was published in *Proceedings of the 24th Intern. ACM SIGACCESS Conf. on Computers and Accessibility, 2022].*

Neural language models are neural networks trained to predict the next word in a given sequence of words. We use "large language models," or LLMs, to refer to the recent class of neural language models (for example, GPT-3)[1] that are capable of generating long passages of text that human evaluators perceive as human-written.[4] LLMs can also be steered to complete text-generation tasks with *few-shot* learning,[1] underscoring their potential as writing-enhancement tools.[24] This functionality could prove immensely valuable for writers with dyslexia, but the correct approach for implementation is an open question. For example, although automatic text generation may help some writers with dyslexia to conquer their "fear of the blank page,"[15] machine-powered writing may also raise concerns over each author's autonomy in the writing process.[6]

We introduce LaMPost, an LLM-enabled prototype to support adults with dyslexia with writing email messages. LaMPost implements LaMDA,[20] an LLM for dialog applications, to augment a standard email editor with automatic outlining, subject generation, suggestion, and rewriting features. We evaluated LaMPost with 19 adult participants with dyslexia, finding enthusiasm among users for AI-powered support features, including rewriting passages in a particular style or tone (for example, "more formal," "more concise") and generating summative content based on the email's body (subject lines). However, we also found that accuracy and quality issues in LLMs may preclude a reliable and trustworthy writing-support experience. Further, applications to support writers with dyslexia using LLMs may require tradeoffs, such as autonomy *vs.* cognitive load, and personalization *vs.* privacy. Finally, knowledge that our writing-support tool contained AI did not have a significant effect on participants' feelings toward the system.

This paper is an abbreviated version of our ASSETS publication[7] accepted in June 2022. Since then, much has changed: our work has helped to guide subsequent literature (for example, Valencia et al.[21]), open-ended LLM applications are now available to everyday users (for example, OpenAI's ChatGPT, Google's Bard), and LLMs themselves have advanced significantly (for example, GPT-4). While our prototype may no longer reflect the state-of-the-art for LLMs, many of the performance issues highlighted by our study's participants still persist within the latest models (for example, Zhang et al.[25]). As the capabilities of generative language models—and our understanding of their risks and trade-offs—continue to mature, our work's characterization of the broader needs, preferences, and concerns among users with dyslexia toward AI writing support can provide a foundation for future work.

2. RELATED WORK

Our research is informed by and builds upon: work on dyslexia and associated writing challenges, accessibility research with this population, and AI-assisted writing tools.

2.1. Writers with Dyslexia

Dyslexia is a multifaceted condition characterized by difficulties with word recognition, reading fluency, spelling, and/or writing.[14] Dyslexia impacts up to 20% of the population, but structural disparities including gender, class, and race cause it to remain undiagnosed for many individuals.[14,22] Through a medical lens, dyslexia is defined as a cognitive deficiency associated with persistent difficulties with reading, spelling, short-term/working memory, and day-to-day organization.[3] Through a lens of neurodiversity, however, dyslexia includes heightened spatial and perceptual abilities, interconnected and dynamic reasoning, and narrative and holistic thinking[22]—alongside commonly defined deficits.

Adults with dyslexia describe wide-ranging difficulties with writing tasks:[3,14,15] on a high level (paragraph and document structure), these may include expressing one's ideas in writing, structuring and ordering topics, and concision. On a lower level, challenges can include word retrieval, appropriate language, sentence composition, grammar, spelling, punctuation, and proofreading. Writers with dyslexia may find compensatory strategies (for example, preferred spellcheckers, dictation software, support from friends and family), but these can add complexity and time to their writing process.[3] We contribute insights into the needs and challenges of adults with dyslexia in email-writing, and we explore AI interventions to address these issues.

2.2. Dyslexia and accessible technology design

Research to improve readability for users with dyslexia has explored text-presentation options, such as increased font size and margin space.[19] Rello et al.[17] explored text simplification to support reading comprehension, finding promising results displaying basic synonyms alongside complex words. A further study of Web searchers with dyslexia[13] showed users preferred multimedia Web pages with headings and bullets instead of long blocks of text. We use this body of work to improve usability in our prototype interface, and we explore text simplification as a form of writing support.

Work targeting writing for users with dyslexia has focused on word-level interventions, such as specialized spellcheckers to detect "real-word" errors (for example, "hear" and "here"),[18] and suggestions to overcome word retrieval difficulties.[11] Wu et al. examined the experience of writers with dyslexia on social media, developing a dyslexia-tuned Neural Machine Translation model to check spelling and grammar in Facebook posts.[23] Researchers have not yet explored AI-powered support for the high-level writing challenges associated with dyslexia—a gap that is addressed by our work.

2.3. Large language models

The recent class of LLMs popularized by GPT-3[1] demonstrates significant advances in natural language generation. At their core, these models have a simple API: Given a string of text (the *prompt*), they return a plausible continuation for that text. For example, the prompt *"A healthy lunch includes"* may return *"fruits, vegetables, grains, and protein."* This call-and-response paradigm makes LLMs well-suited for dialogue-based applications, such as OpenAI's ChatGPT and Google's Bard.

Prompts may also begin with exemplary calls and responses, priming the model to generate a specific type of response. In this way, LLMs are capable of *few-shot learning,*[1] which is more accurate than the *zero-shot* example here.[26] The example below prompts the model for one item of clothing and one accessory for the weather:

prompt: When it's sunny, I need:
shorts and sunscreen
When it's raining, I need:
rain boots and an umbrella
When it's snowing, I need:
language model: mittens and a shovel

LLMs have been investigated for many applications, including code generation[10] and creative writing[24] Accessibility researchers have also begun exploring LLM applications; for example, Valencia et al. looked at phrase generation in augmentative and alternative communication devices.[21]

Despite impressive performance on many tasks, LLMs also have drawbacks. Since they are trained on content from the Internet, they risk generating incorrect, offensive, or stereotyped text.[5] "Memorization" (for example, regurgitating existing text rather than producing novel content) presents another risk.[2] The risk of erroneous or inappropriate output may carry further ethical challenges in generative systems used by audiences that experience challenges assessing the quality of that output[12]—such as text generators for users with dyslexia. Mitigating the risks associated with LLMs is an active area of research; for example, LaMDA (the LLM in our study) uses fine-tuning to improve safety.[20]

2.4. AI-assisted writing

Human-AI co-creation in writing has been widely studied, with applications like Gmail's Smart Reply feature deployed to massive audiences. Gero et al.[6] studied automatic metaphor generation, finding enhancements to self-expression led to trad-eoffs in users' feelings of autonomy. How to present the algorithm in human-AI systems is an open question, as these choices can impact perceptions of the system and desires to use it.[9] Wordcraft[24] explored LLMs as creative writing support: Users can select from a variety of operations—including infilling, elaboration, rewriting, and an open-ended dialog—to collaborate with the model to write a story. The system's prompting methods also enable users to build their own custom controls, such as "*rewrite the text to be more melodramatic*." In this paper, we adapt Wordcraft's approach to explore LLMs as practical writing support for users with dyslexia.

3. THE LAMPOST SYSTEM

We engaged in over a year of formative research with the dyslexia community to motivate an AI writing support system. This included participatory design sessions, workshops with organizations with expertise in reading disabilities,[a] and brainstorming among team members with lived experience of dyslexia and other visual processing and reading disabilities. As a key step in this work, we recruited seven adults with dyslexia for (1) individual interviews on writing practices and challenges, and (2) a group assessment of AI writing-support ideas. The individual interviews highlighted a variety of writing challenges, including organization, clarity, concision, appropriate tone, and proofreading. The group interviews revealed a common interest in AI writing support: Visual outlines could help to order and structure ideas, summarization could reveal the writing's meaning, and revising feedback could improve clarity, verbosity, and tone. However, we also noted concerns over users' autonomy and control over the final work, their confidence addressing AI-provided feedback, and their privacy.

Informed by our formative inquiries and prior work,[14,15] we built LaMPost, a Web application and LLM-enabled prototype for email-writing support for users with dyslexia. Emailing provided a constrained yet practical use case to demonstrate different approaches to infuse a generative language model within a text editor. LaMPost's main panel resembled a typical email editor, with the email's recipients, subject, and body near the top. A smaller panel on the right is reserved for three LLM features: identifying main ideas (with the option to generate an email subject), suggesting possible changes, and rewriting a selection.

3.1. LLM features to support email writing

LaMPost is powered by LaMDA, a neural language model,[20] and it adapts the few-shot prompting methods introduced by the Wordcraft system[24] (a LaMDA-powered tool for creative story-writing by general audiences). To enable LaMPost's LLM operations, we built a workflow for custom *few-shot learning* prompts that combine a user's current text with several exemplars of each desired writing operation. Prompt performance can be highly sensitive to word choice, formatting, and the exemplars' content,[26] and we experimented with several different prompting methods before settling on the three in our system; our full paper describes our iterative development process and lessons we learned.[7]

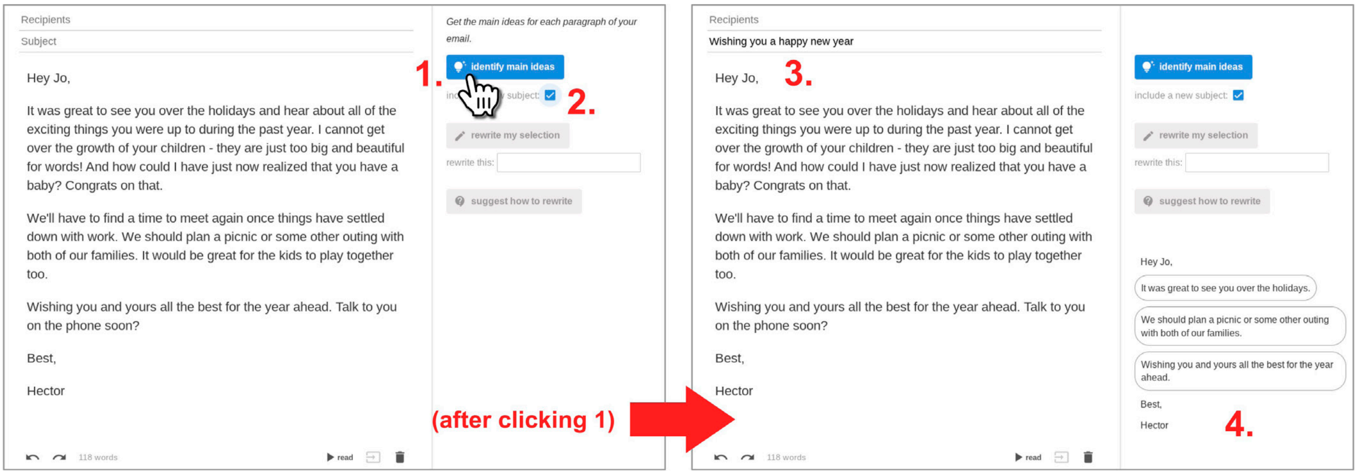
3.1.1. LLM feature 1: Identify Main Ideas

Users can ask the AI to generate a visual outline for their email containing each paragraph's main idea (Figure 1). Users also have the option to generate a new subject line based on the main ideas. The feature was motivated by feedback from our formative study with dyslexic adults, which highlighted difficulties with organizing ideas and clarity in writing, along with users' interest in automatic summarization and visual organization. By displaying the AI's interpretation of the email's most salient points, we imagined this feature could reveal how other readers might interpret (or misinterpret) that email.

3.1.2. LLM feature 2: Suggest Possible Changes

Users can select a word, phrase, or paragraph and ask the AI

Figure 1. The Identify Main Ideas feature allows users to click '*identify main ideas*' (1) and 'include a new subject' (2) to generate a visual outline of their email based on each paragraphs' main idea (4), followed by a subject line based on these ideas (3).



Hey Jo,

It was great to see you over the holidays and hear about all of the exciting things you were up to during the past year. I cannot get over the growth of your children - they are just too big and beautiful for words! And how could I have just now realized that you have a baby? Congrats on that.

We'll have to find a time to meet again once things have settled down with work. We should plan a picnic or some other outing with both of our families. It would be great for the kids to play together too.

Wishing you and yours all the best for the year ahead. Talk to you on the phone soon?

Best,

Hector

Wishing you a happy new year

Hey Jo,

It was great to see you over the holidays and hear about all of the exciting things you were up to during the past year. I cannot get over the growth of your children - they are just too big and beautiful for words! And how could I have just now realized that you have a baby? Congrats on that.

We'll have to find a time to meet again once things have settled down with work. We should plan a picnic or some other outing with both of our families. It would be great for the kids to play together too.

Wishing you and yours all the best for the year ahead. Talk to you on the phone soon?

Best,

Hector

Hey Jo,

It was great to see you over the holidays.

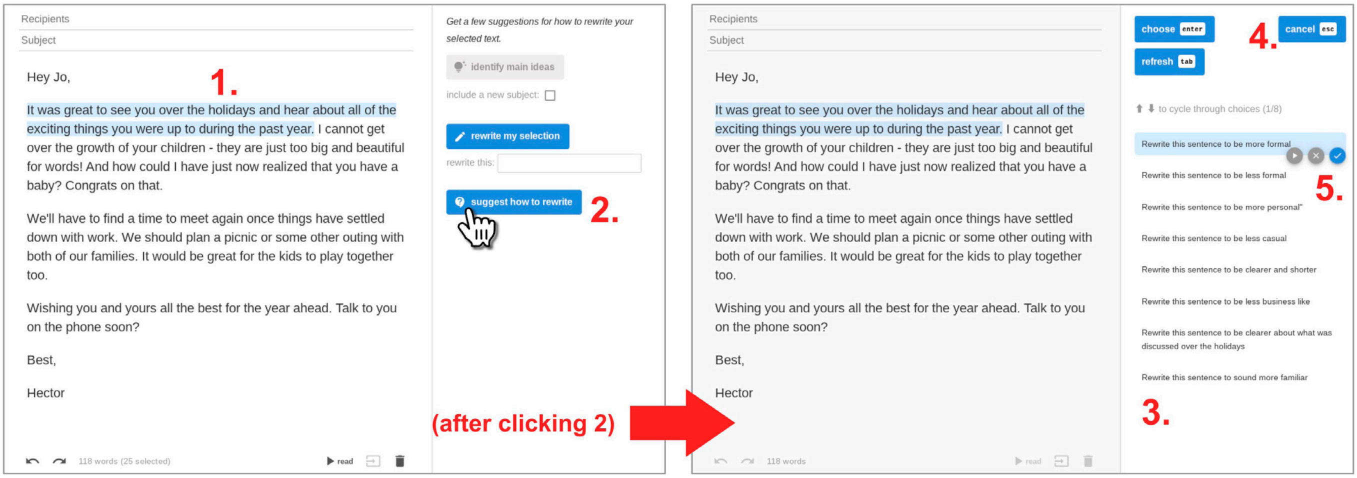
We should plan a picnic or some other outing with both of our families.

Wishing you and yours all the best for the year ahead.

Best,

Hector

Figure 2. The Suggest Possible Changes feature allows users to select a passage of text (1) and click 'suggest how to rewrite' (2) to populate the right-hand panel with several suggestions for changing the passage (3). Users can then exit the operation and rewrite the text themselves (4), choose to have individual suggestions 'read aloud', discarded, and used in the 'Rewrite My Selection' feature (5).



Hey Jo,

It was great to see you over the holidays and hear about all of the exciting things you were up to during the past year. I cannot get over the growth of your children - they are just too big and beautiful for words! And how could I have just now realized that you have a baby? Congrats on that.

We'll have to find a time to meet again once things have settled down with work. We should plan a picnic or some other outing with both of our families. It would be great for the kids to play together too.

Wishing you and yours all the best for the year ahead. Talk to you on the phone soon?

Best,

Hector

Get a few suggestions for how to rewrite your selected text.

Hey Jo,

It was great to see you over the holidays and hear about all of the exciting things you were up to during the past year. I cannot get over the growth of your children - they are just too big and beautiful for words! And how could I have just now realized that you have a baby? Congrats on that.

We'll have to find a time to meet again once things have settled down with work. We should plan a picnic or some other outing with both of our families. It would be great for the kids to play together too.

Wishing you and yours all the best for the year ahead. Talk to you on the phone soon?

Best,

Hector

⬆ ⬇ to cycle through choices (1/8)

Rewrite this sentence to be more formal

Rewrite this sentence to be less formal

Rewrite this sentence to be more personal"

Rewrite this sentence to be less casual

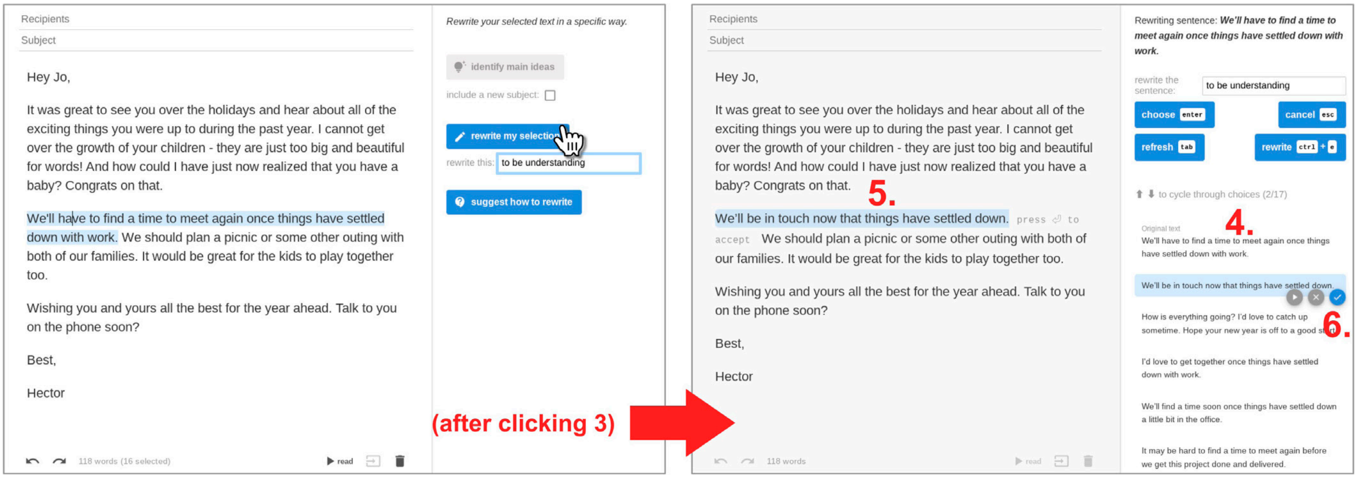
Rewrite this sentence to be clearer and shorter

Rewrite this sentence to be less business like

Rewrite this sentence to be clearer about what was discussed over the holidays

Rewrite this sentence to sound more familiar

Figure 3. The Rewrite My Selection feature allows users to select a piece of text (1), enter a custom instruction for rewriting it (2), and click *'rewrite my selection'* (3) to populate the right-hand panel with several rewritten choices from the AI (4). Hovering on a choice will show a preview in the editor (5). Each choice can be 'read aloud', discarded, or applied over the original passage (6).



Rewrite your selected text in a specific way.

rewrite this: to be understanding

Hey Jo,

It was great to see you over the holidays and hear about all of the exciting things you were up to during the past year. I cannot get over the growth of your children - they are just too big and beautiful for words! And how could I have just now realized that you have a baby? Congrats on that.

We'll have to find a time to meet again once things have settled down with work. We should plan a picnic or some other outing with both of our families. It would be great for the kids to play together too.

Wishing you and yours all the best for the year ahead. Talk to you on the phone soon?

Best,

Hector

Hey Jo,

It was great to see you over the holidays and hear about all of the exciting things you were up to during the past year. I cannot get over the growth of your children - they are just too big and beautiful for words! And how could I have just now realized that you have a baby? Congrats on that.

We'll be in touch now that things have settled down. press ⏎ to accept We should plan a picnic or some other outing with both of our families. It would be great for the kids to play together too.

Wishing you and yours all the best for the year ahead. Talk to you on the phone soon?

Best,

Hector

Rewriting sentence: ***We'll have to find a time to meet again once things have settled down with work.***

rewrite the sentence: to be understanding

⬆ ⬇ to cycle through choices (2/17)

Original text
We'll have to find a time to meet again once things have settled down with work.

We'll be in touch now that things have settled down.

How is everything going? I'd love to catch up sometime. Hope your new year is off to a good start

I'd love to get together once things have settled down with work.

We'll find a time soon once things have settled down a little bit in the office.

It may be hard to find a time to meet again before we get this project done and delivered.

for suggestions on how to rewrite it (Figure 2). Participants in our formative study described uncertainty about when and how to improve their writing, and they were interested in automatic suggestions for language characteristics such as tone and clarity. When users press the *'suggest how to rewrite'* button, the selected text is appended to the end of a custom prompt and sent to the LLM. To facilitate *few-shot* learning, the prompt begins with several exemplars, each showing an example email, a passage of selected text, and suggested improvements to that selection (that is, a best-case response). The model's responses are displayed to the user as choices; for example, *"Rewrite this sentence to be less business-like."* Users can make revisions themselves or use a preferred choice as a precursor prompt (that is, *meta-prompting*)[24] to generate rewritten passages in the operation described below.

3.1.3. LLM feature 3: Rewrite My Selection

Users can select a word, phrase, or paragraph and enter an instruction to rewrite the text (Figure 3); for example, 'rewrite this: *to be shorter*'. Participants in our formative study described issues with choosing language to fit an appropriate tone and style, and prior work shows that people with dyslexia may rely on a thesaurus[15] or templates[3] to achieve desired phrasing. When users enter a custom instruction and press the *'rewrite my selection'* button, their text and instruction is appended to a custom prompt with a handful of exemplars for rewriting tasks. Each exemplar includes an example email, a passage of selected text, a rewriting instruction, and a rewritten version of the selected passage. The prompt adapts *related example prompting*:[24] We anticipated that exemplars of the most relevant instructions would be sufficient for the model to complete a variety of unseen rewriting tasks. This included conciseness (*'to be simpler'*), tone (*'to be more polite'*), audience (*'to be more formal'*), and precision (*'to be more clear'*). After the prompt is sent to the LLM, the model responds with several rewritten passages displayed to users as choices to retain control over the final document.

3.2. Accessibility considerations

While our primary goal was to demonstrate the functionality of LLMs for writing, we recognized that other usability issues may impact the ability of users with dyslexia to evaluate the system. We made design choices to maximize usability for this population, and we tested iterations of our design with members of our team who identify as having dyslexia. For readability, we incorporated text-presentation recommendations for users with dyslexia for font, size, and line spacing.[19] To support visual referencing, we paired most buttons with an icon and added highlighting to the sentence surrounding the insertion point cursor (visible in Figure 3-1). Finally, for users who preferred to listen to on-screen text rather than visual parsing, we included a "read aloud" option for the email's body, the generated outline, and each choice returned by the LLM.

4. LAMPOST EVALUATION

We evaluated the LaMPost prototype in a hands-on demonstration and practical email writing exercise. Our primary goals were to explore the potential ways that LLMs can be incorporated into the email-writing process of writers with dyslexia and to assess users' perception of each of LaMPost's writing support features. Our secondary goals were to evaluate users' feelings of satisfaction, self-expression, self-efficacy, autonomy, and control while writing with LLMs, and to assess how their exposure to AI terminology and metaphors impacted these feelings.

For the secondary goal, we segmented the system evaluation into two between-subjects conditions: (A) **with AI metaphors**, introducing LaMPost as an *"AI-powered"* email editor and describing it as a personified AI agent throughout the session (for example,*"Hang on, the AI is thinking"*); and (B) **without AI metaphors**, introducing LaMPost as an *"enhanced"* email editor and using language to obscure the presence of an AI/LLM (for example,*"Hang on while the system loads"*). The interface itself did not reference the underlying LLM and was left unchanged for both conditions.

4.1. Participants

We recruited 32 participants via a survey shared with a large sampling pool maintained by our institution; 19 completed the study. All said English was their preferred writing language; 16 were based in the U.S. and three in Canada. We screened for experience writing emails (at least one per year) and for self-reported challenges associated with dyslexia. Fourteen participants reported having a formal dyslexia diagnosis and four reported discovering their dyslexia on their own; one participant did not specify. We aimed for balanced representation across gender and age categories, but we attained neither due to cancellations. Four participants identified as female, 14 as male, and one as non-binary. One participant was 18-24 years old, seven were 25-34 years old, and 11 were 35-54 years old. Nine participants were randomly assigned to the *with AI metaphors* study condition, and ten to the *without* condition. Participants received a $100 gift card for their time.

4.2. Procedure

The evaluation procedure was conducted remotely during early 2022 and lasted up to 75 minutes. Participants were asked to prepare two realistic email topics that they felt comfortable sharing with us for the session. They received assurance that the writing exercises during the session were only to provide a realistic experience of using the system and that their performance would not be evaluated. The session followed a three-part protocol:

4.2.1. Part 1: Background and demo (25 min)

First, to understand each participant's current writing workflow, we asked a few background questions and collected ratings of their confidence, ability to express themselves, and satisfaction with email-writing. Next, the researcher led a hands-on demonstration to introduce the functionality of the LaMPost system. Participants shared their screen as the researcher walked them through each element of the interface. The researcher explained each of the three LLM features (*Identify Main Ideas*, *Rewrite My Selection*, *Suggest Possible Changes*) and then asked participants to perform each operation on a placeholder email. After introducing each feature, we asked participants to share their immediate thoughts, concerns, and if they might use it when writing email messages.

4.2.2. Part 2: Writing exercise (25 min)

Next, the researcher conducted an informal writing exercise for participants to freely use the LaMPost system to write an email on a topic they had brought to the session. We asked participants to "think aloud" throughout the writing exercise and to voice any questions, observations, suggestions, or concerns about the system. The researcher provided limited answers to questions they had about the system's functionality, respecting the AI-framing condition for each participant. To ensure that the participants had experience using each LLM feature (*Main Ideas*, *Suggest*, *Rewrite*), the researcher allowed them to write for five minutes before suggesting they try an unused LLM feature, repeating until all features had been used at least once. Participants were told to write as much as they were able in the time provided; if they finished, they could try writing a second email.

4.2.3. Part 3: Follow-up interview (25 min)

After the writing exercise, the researcher discussed the experience with participants via semi-structured interview questions. Questions targeted their opinions about the LaMPost system—including what they liked, what needed improvement, and additional features to assist with writing—as well as how the system compared to their typical experience writing emails. To conclude, the researcher administered post-use rating scales to evaluate the system's usefulness, consistency, and participants' feelings of satisfaction, self-expression, self-efficacy, autonomy, and control. Each rating included a positive statement with a 7-point scale from *"Completely disagree"* to *"Completely agree,"* and a follow-up question to capture their reasoning.

4.3. Analysis

We analyzed our qualitative data using a thematic coding process. Before the study, we developed a set of deductive codes for: existing email-writing practices; system positives, negatives, and changes; and feelings of self-efficacy, self-expression, autonomy, and control. During data collection, three researchers generated a set of inductive codes through analytic connection across participants. Our final codebook included: level-1 codes for writing practices, the overall system, and each LLM feature; level-2 codes for positives, negatives, and concerns for each; and level-3 codes for our inductive and deductive themes. One researcher coded transcripts for each of the 19 sessions; we organized resulting themes into subsections to form our narrative. For quantitative data, we used pre-use ratings to characterize participants' existing feelings about writing emails. We used Mann-Whitney U tests (two-tailed) to compare how our between-subjects manipulation of AI metaphors shaped users' perception of the system, testing for significance in post-use ratings for usefulness, consistency, satisfaction, self-expression, self-efficacy, autonomy, and control.

4.4. Findings

The following sections describe participants' experience writing emails, reactions to the system and LLM features, and the (lack of) effects with the AI-metaphors condition.

4.4.1. Current email-writing experience

We explored participants' thoughts on writing emails to understand opportunities and challenges specific to this genre of writing. Most participants said they wrote new emails and replies daily; our full paper has a complete breakdown of their email-writing habits.[7] Participants primarily wrote emails for work communication; some also wrote to connect with family and friends or conduct service inquiries (for example, P8: *"doctor's appointments"*). Participants generally felt a strong sense of self-efficacy when emailing (Figure 4, left) and most felt confident in their email-writing ability (*avg.*=5.05, *SD*=1.22). A majority also felt satisfied with their emails (*avg.*=4.89, *SD*=1.10) and capable of expressing their ideas (*avg.*=4.79, *SD*=1.36), although agreement toward each of these statements was more mixed.

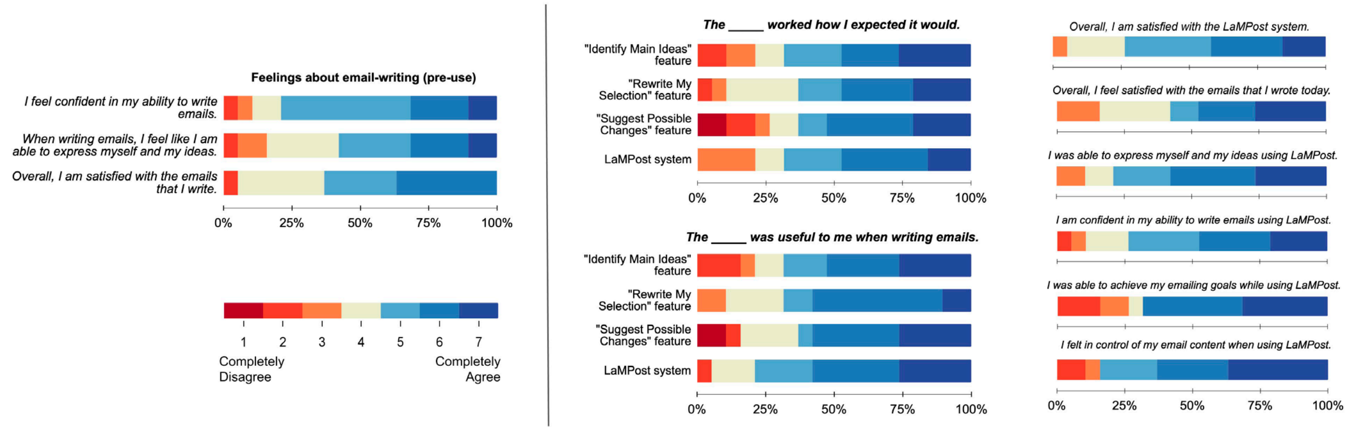
When discussing their experiences writing emails, participants described similar challenges and mitigation strategies to those in prior work[3,14,15] and our formative study. Over half of participants (*N*=10) said they preferred to draft messages outside of an emailing platform (for example, Microsoft Word, pen and paper) due to *"habit"* (P7), *"personal preference"* (P6), or *"to separate it from the anxiety of having to respond"* (P2). Within their preferred platform, participants described challenges turning their ideas into writing—or, *"Putting what I'm trying to say in my brain into words"* (P13). Common drafting approaches included bullet points (3) and a "word faucet" strategy (5): *"I get all of my thoughts out so [...] I've got this blob of text"* (P10).

After drafting, additional challenges emerged during revising and proofreading. Spelling and grammar were the most common issues, and 10 participants said they relied on a trusted spell checker. Cutting verbose, or *"wordy"* (P16), drafts down was another struggle (*N*=7). Participants recalled instances when their writing was misinterpreted by the email's recipient, either due to a lack of clarity (7) or misunderstanding of their intended tone (6): *"They thought that I was just writing to them all mad and pissed off, when in reality, I was just explaining myself"* (P8). Several participants (7) mentioned they struggled to find time to revise and proofread; some (4) asked others to read the email; and a few (3) listened for issues using text-to-speech.

4.4.2. Reactions to the LaMPost prototype

The following section describes participants' responses to our three LLM-powered features and the LaMPost system based on qualitative responses and usefulness ratings (Figure 4, bottom-center). We used a two-tailed Mann-Whitney U test to compare usefulness ratings between the study's two AI-framing conditions—*with* (*N*=9) and *without* (10) AI-related metaphors—but we did not find a significant difference for any of the ratings (p>0.05). Our full paper shows

Figure 4. Before using LaMPost, participants rated their self-efficacy, self-expression, and satisfaction during email-writing (left); most were confident about writing emails, with slightly more mixed feelings about self-expression and satisfaction. After using LaMPost, they rated the consistency (top-center) and usefulness (bottom-center) of each LLM feature and the overall system, along with other feelings (right). In general, most found the LaMPost system useful while writing emails, appreciating the "Rewrite my selection" feature most of all.



ratings for each condition.[7] We return to implications of our AI-metaphors experiment in Section 5.4.

Of LaMPost's three features, the **Rewrite My Selection** feature was rated highest for usefulness on average (*avg.*=5.26, *SD*=1.26); 13 participants agreed that it was useful for writing email messages (that is, ratings ≥ 4 on 7-point agreement scale) and nine selected it as LaMPost's most useful feature. Participants said the primary benefit of the *Rewrite* feature was iteratively finding satisfying wording for an intended idea: *"You're able to get a start on what you're going to say and you can tweak your writing from there"* (P18). They highlighted conciseness (*N*=7) and tone (4) as key support areas. For example, P1 saw the feature helping to *"explain something simply,"* while P13 saw potential to *"sculpt the email to the audience."* Despite a positive response, many participants voiced concerns over inaccurate and noisy results (*N*=13) and overly numerous choices (11).

The **Identify Main Ideas** feature was rated the second highest for usefulness on average (*avg.*=5.11, *SD*=1.82); 13 participants agreed that it was useful, but only four participants selected it as the most useful of the three. Participants said the primary value of the visual outline was in validating that the writing matched their intended idea. P4 called it the system's *"selling point,"* providing *"an independent verification that I'm hitting the points that I want to hit."* Some participants (*N*=8) voiced concerns about emotionless or *"sterile"* (P17) summarizations, while others (4) found key details missing. Nine participants selected the *Main Ideas* feature as the least useful, and four said they would never use it (for example, P6: *"It seems very obvious"*).

Notably, the option to **generate a subject line** received a very positive response. Because LaMPost generated the subject line based on the *Main Ideas* outline, we chose to group both into one feature and did not ask for separate usefulness ratings. However, several participants—who otherwise felt tepid about the *Main Ideas* outline—saw value in the automatic subject line: *"I always leave the subject line blank. [...] Nothing fits"* (P19). However, a few (*N*=3) took issue with the subject's framing; P6 voiced disappointment at an inappropriate subject line for a sensitive email: *"The idea I'm trying to get across is quite the opposite."*

The **Suggest Possible Changes** feature received mixed reactions for usefulness (*avg.*=5.05, *SD*=2.03); six participants said it was the most useful feature, and six said it was the least. Twelve participants agreed that the feature was somewhat useful, identifying its primary benefit as support for fixing a detected but ambiguous issue (P15: *"Times where I'm like, 'Something doesn't sound right'"*). A few participants enjoyed using the *Suggest* feature with the *Rewrite* feature to *"narrow down"* (P12) possible instructions. However, some suggestions lacked an explanation: *"It's saying, 'Rewrite this sentence to be less wordy'. Is it telling me that the sentence is wordy?"* (P10). Nine participants were concerned about accuracy and noise, while five mentioned the *"overwhelming"* (P19) quantity of choices.

Despite varied responses to individual features, most participants identified at least one useful feature, and they rated the usefulness of **the LaMPost system overall** fairly high (*avg.*=5.53, *SD*=1.38). Participants liked LaMPost's automatic, content-aware support, which could be used at the scale of their choosing (P10: *"sentence-by-sentence or paragraph-by-paragraph"*). Despite the positive response, participants were divided on its value for day-to-day emailing. For example, P14 only saw value for certain contexts, such as *"emails to a VIP"* or to *"convey a complex topic."* Eight participants mentioned time-saving (P4: *"I get five minutes back"*), but five thought it would lengthen their time writing email messages.

4.4.3. Concerns over accuracy and noise

One of the most common issues highlighted by participants throughout the evaluation was unhelpful, inaccurate, or *"nonsensical"* (P8, P11) results. We anticipated instability in our few-shot learning prompts on unseen tasks,[26] and we included a *consistency* rating for the system (Figure 4, top-center). A two-tailed Mann-Whitney U test to compare consistency ratings between the AI metaphor conditions did not yield a significant difference (p>0.05).

Unlike the relatively constrained *Main Ideas* summarization task, the *Rewrite* and *Suggest* features included few-shot prompts that tasked the LLM with open-ended generation. This sometimes produced "hallucinations", or seemingly relevant but factually incorrect content.[16,25] For the *Suggest* feature, this included generating irrelevant suggestions for the text, such as *"Rewrite this sentence to not use so many commas"* for a sentence without any commas (P11). In contrast, hallucinations from the *Rewrite* feature involved seemingly relevant, yet imaginary details added to the rewritten passage. For example, when a rewritten option included the phrase *"that nice patio,"* P12 noted, *"The original text doesn't mention a patio."* After discarding the option, he asked to know the source of the *"wrong info"* and was concerned about detecting others like it in the future.

The *Rewrite* feature included other accuracy issues in addition to hallucinations. Four participants commented on the LLM failing to satisfy a given instruction; for example, after concise results for the instruction *"Rewrite this text: to sound more detailed"*, P2 noted, *"To me, this implies more text than I wrote."* Three participants noted overly similar results, while two others noticed key details missing from the rewritten passage. They said the process of sifting through poor results was both *"time wasteful"* (P7) and cognitively demanding.

4.4.4. "The Paradox of Choice"

The LaMPost system included two extremes with regard to choice: the *Rewrite* and *Suggest* features displayed numerous choices to users (15 responses from the model, minus duplicates), while the *Main Ideas* and subject line only returned the model's first response. For the first extreme, 12 participants voiced concerns over the sheer volume of choices. They described the issue as *"the paradox of choice"* (P5, P6), referring to the psychological concept that large sets of choices can feel overwhelming and lead to unsatisfying final selections. Most participants (*N*=12) agreed that the number of options should be reduced, suggesting *"three to four"* (P19), or *"four to five"* (P2). The other choice extreme—the single result of the *Main Ideas* and subject line feature—also raised concerns, though less often: three participants questioned why it did not return several results, and two participants wanted to select or *"highlight"* (P1) the main ideas themselves.

Additional tensions around user choice were apparent in the *Rewrite* feature's free-form instructions. While six participants spoke of the immense value of writing their own instructions (P10: *"It's like having a thesaurus for someone's thought"*), others voiced concerns about complex inputs (*N*=3) and misspellings (P8, P13). This led six participants to request pre-written instructions in addition to the open-ended input: *"Basic ones: [...] 'make it shorter,' 'make it longer,' 'more professional,' 'more casual'"'* (P13).

4.4.5. Feelings about email-writing with LLMs

The secondary goal of our evaluation was to understand how writing with LLMs can impact personal feelings of satisfaction, self-expression, autonomy, and control among writers with dyslexia (Figure 4, right). A two-tailed Mann-Whitney U test to compare the ratings of each feeling between the evaluation's AI-framing conditions did not yield a significant difference (p>0.05) for any rating.

Participants generally felt satisfied with the system (*avg.*=5.26, *SD*=1.14), but they voiced concerns over accuracy and our implementation of choices (as described above). They also largely felt satisfied with the email messages they wrote (*avg.*=5.16, *SD*=1.52) and gave high ratings for personal autonomy (*avg.*=5.26, *SD*=1.88); however, these two ratings had greater variance due to seven participants not finishing their intended emailing tasks within the allotted time.

Participants had a strong sense of self-efficacy using LaMPost (*avg.*=5.26, *SD*=1.41); 14 participants felt confident about using the system to write future email messages. Those less confident repeated their concerns about the system's performance: *"The AI would start breaking down if it really had to compute more"* (P1). LaMPost facilitated a strong sense of self-expression (*avg.*=5.53, *SD*=1.34), particularly due to the *Rewrite* feature: *"I was able to look and say, 'Well, this is more of how I would speak'"'* (P18). However, four participants said LaMPost was not helpful for expressing ideas: *"It cannot explain much of the email if I don't give it [information], [...] and I don't always know what to write"* (P8).

AI-assisted writing systems raise questions over whether the user or agent is ultimately in control of the work.[6] Participants largely felt in control over their email content while using LaMPost (*avg.*=5.58, *SD*=1.56): *"The system gave suggestions, [but] in the end, I'm the one that's deciding"* (P15). However, three participants disagreed due to unclear suggestions (P17) and being unable to troubleshoot undesirable results (P1, P5).

4.4.6. System improvements and additional features

Discussing improvements to the system surfaced strong feelings around privacy and trust. Six participants thought the system should include personalization: either *"learning through prior conversations"* (P6), or learning from users' choices. For P17, personalization was crucial to capture each writer's voice, *"rather than it being a canned response from the computer."* However, seven participants voiced concerns about the system reading and storing their personal data; P9 desired *"the option to not have it"* while P4 requested a clear explanation of the system's data storage policy. A few participants desired mechanisms to strengthen their trust in the system, such as an *"explanation"* (P10, P16) button for each result. P13 used his own comprehension challenges to describe the issue: *"If [the system] is saying, 'I can take this sentence and make it more businesslike', I'm going to accept whatever you're offering me."*

For additional features, four participants requested full messages generated from a bulleted outline (P9: *"I write the three main topics that I want and the system writes [the email]"*), while three said the visual *Main Ideas* outline could also save time when reading email (for example, P8: *"If I'm late to a meeting"*). Finally, five participants requested features to track the system's results and their progress during each writing session, including: a *"changelog"* (P6), saving results for later (P18), persistent action items for *Suggest* results (P2), and generating suggestions in real time (P4).

5. Discussion

Our results indicate that LLMs with optimal output hold potential to assist with email-writing tasks, and our evaluation highlighted several promising routes for future exploration, such as the controllable *Rewrite* and subject line features. However, our features as-is did not surpass participants' accuracy and quality thresholds, and as a result, we conclude that LLMs (as of early 2022) are not ready to support the real-world needs of writers with dyslexia. Surprisingly, we found the use (or non-use) of AI metaphors had no effects on users' perceptions of the system, nor on their feelings of autonomy, expression, and self-efficacy during use. Here, we discuss implications of our findings and opportunities for future work.

5.1. Improving support with new datasets

Our evaluation highlighted users' limited tolerance for inaccurate or unhelpful LLM results when seeking writing support. We chose to use a pre-trained model as a generalized base for the LaMPost prototype, and we implemented each LLM feature with a few-shot learning prompt[24] that demonstrated the desired writing operation with excerpts from real email messages on a variety of topics. However, these email messages were sent by writers of unknown lexical ability, and they may not have reflected the characteristics of early drafts produced by our study's participants (for example, bullet-point outlining *vs.* "word faucet" with dictation tools). To to our knowledge, however, a public corpus of writing samples produced by adults with dyslexia does not exist. A small dataset of this type could be used to construct few-shot learning prompts for pre-trained LLMs (similar to our own work), while a sufficiently large corpus could be used to train specialized models for high performance on constrained tasks (for example, summarizing main ideas). Future work should explore the feasibility of constructing a dataset of writing by people with dyslexia and examine how LLMs (and LaMPost's LLM features) can assist email writers from the general population.

5.2. Safeguarding from risks

Assessing the results of an AI-enabled assistive technology can present obstacles for primary users when the data is inherently not accessible;[12] likewise, assessing the quality of text-based results from an LLM-powered system presents challenges for users with dyslexia. Additional assessment mechanisms may be needed for this population: LLMs have been shown to inherit biases present in their training data (for example, Internet posts)[5] and can produce factually incorrect "hallucinations."[16,25] While we did not encounter offensive results in our work, certain *"nonsensical"* results led to requests for explanations to assist with determining the suitability of results for their writing. User options to flag incorrect or unacceptable language is another promising solution for developers to implement targeted filters or fixes. Transparency and high performance from an automated system can foster trust among users and drive continued use; however, our work also suggests that with sufficient trust and confidence in the system, users with dyslexia may feel less inclined to spend time and energy assessing results—thereby increasing risk for future harm. Future work should explore potential safeguarding methods; for example, before sending an email message, a system could automatically perform a final check on all machine-written text, then ask the user to review any questionable passages.

5.3. Pitfalls with control-enabling mechanisms

Our findings suggest that users with dyslexia value control while writing with an AI—confirming and extending prior work[6]—and our participants felt strongly in control while using LaMPost. However, our evaluation also highlighted usability issues with the *Rewrite* and *Suggest* features' lists of options, which could be overwhelming and time-consuming to navigate. To mitigate these issues, future AI writing assistants should display fewer options, add more variation, sort options by length, and allow users to save options for later. In addition, the *Rewrite* feature's open-ended instruction was a valuable mechanism for users to clarify their intentions, but it also led to usability issues for dyslexic users, such as difficulty finding the language to express ideas, and spelling and grammatical issues. To overcome these issues, future AI writing assistants could provide pre-written instructions, dictation options, or probing questions to help users narrow down their revising goals. Overall, our work highlights how control-enabling mechanisms in AI writing assistants may introduce new challenges for users with dyslexia, and future work should seek to balance each user's sense of control with their ability to fully leverage the system's features.

5.4. Implications of AI metaphors results

Based on prior work showing how different conceptual metaphors can affect users' perceptions of automatic systems,[9] we hypothesized that our use (or non-use) of AI metaphors would impact users' perceptions of the LaMPost system, but we did not find statistical significance. This result may be a positive for public attitudes toward AI in writing-support tools, as knowledge of the AI did not change users' feelings about the system. However, we cannot rule out systematic error due to small sample sizes for each condition (N=9 *vs.* 10) or study fatigue (ratings were administered near the end of a 75-minute session). We also did not measure participants' prior experience with AI, though all came from outside of the technology industry per our institution's recruiting guidelines. A few participants in the *without AI* group used AI terminology while discussing the system; this may reflect increasing public awareness of AI and its presence in many new products. Future research should attempt a deeper comparison of AI-forward *vs.* obscured writing tools to characterize how each impacts user attitudes—particularly end users with dyslexia—and to provide clear design guidelines for off-the-shelf systems that incorporate LLMs.

5.5. Limitations

This work was conducted in early 2022 as a remote lab evaluation due to the ongoing COVID-19 pandemic, and we encountered several issues as a result. First, we recruited 32 participants but only 19 completed the evaluation, resulting in an unbalanced demographic representation and a smaller-than-planned sample size for our between-subjects experiment, potentially limiting the generalizability of our results. Second, the remote nature of the study limited our ability to control the testing environment and provide technical support. Many participants required support with setup and troubleshooting, which both limited their time using the system and added variation to the average time spent writing with LaMPost; two participants could not access the system at all and had to dictate the email's content to the researcher via screenshare. Third, participants only wrote one email message and did not get to experiment with different topics, email message lengths, or audiences (for example, work vs. personal), which potentially skewed their responses according to the complexity of their email-writing task.

6. Conclusion

In this paper, we introduced LaMPost, an email-writing interface to explore the potential for large language models to power writing support tools for people with dyslexia. Inspired by the varied needs of this population, LaMPost introduced LLM features including *rewrite my selection*, *identify main ideas* (with subject line generation), and *suggest possible changes*. We evaluated LaMPost with 19 adults with dyslexia, identifying many promising routes for further exploration—including the popularity of the "rewrite" and "subject line" features—but also found that LLMs do not meet the acceptable accuracy and quality thresholds of our participants. Surprisingly, we found no effect in the use (or non-use) of AI metaphors on users' perceptions of the system, nor on feelings of autonomy, expression, and self-efficacy when writing email messages. Our findings yield further insight into the benefits and drawbacks of using LLMs as writing support for adults with dyslexia and provide a foundation to build upon in future research.

References

1. Brown, T. et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*, (2020), 1877–1901.
2. Carlini, N. et al. *Extracting training data from large language models*, Dec. 2020.
3. Carter, C. and Sellman, E. A view of dyslexia in context: Implications for understanding differences in essay writing experience amongst higher education students identified as

dyslexic. *Dyslexia 19*, 3 (Aug. 2013), 149–164.

4. Clark, E. et al. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Intern. Joint Conf. on Natural Language Processing 1*,. Association for Computational Linguistics (Aug. 2021), 7282–7296.
5. Dhamala, J. et al. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conf. on Fairness, Accountability, and Transparency* (2021), 862–872.
6. Gero, K.I. and Chilton, L.B. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI Conf. on Human Factors in Computing Systems, Paper 296 in CHI '19*. Association for Computing Machinery (May 2019), 1–12.
7. Goodman, S.M. et al. LaMPost: Design and evaluation of an AI-assisted email writing prototype for adults with dyslexia. In *the 24th Intern. ACM SIGACCESS Conf. on Computers and Accessibility*. ACM (Oct. 2022), 1–18.
8. International Dyslexia Association Editorial Contributors. *Dyslexia Basics*, 2020.
9. Khadpe, P. et al. Conceptual metaphors impact perceptions of Human-AI collaboration. In *Proc. ACM Hum.-Comput. Interact. 4*, CSCW2, (Oct. 2020), 1–26.
10. Kim, T.S., Choi, D., Choi, Y., and Kim, J. Stylette: Styling the Web with natural language. In *Proceedings of the 2022 CHI Conf. on Human Factors in Computing Systems, CHI '22*. Association for Computing Machinery, (2022).
11. Li, A.Q., Sbattella, L., and Tedesco, R. PoliSpell: An adaptive spellchecker and predictor for people with dyslexia. In ***Proceedings of the Intern. Conf. on User Modeling, Adaptation, and Personalization*** (2013).
12. Morris, M.R. AI and accessibility. *Commun. ACM 63*, 6 (May 2020), 35–37.
13. Morris, M.R., Fourney, A., Ali, A., and Vonessen, L. Understanding the needs of searchers with dyslexia. In *Proceedings of the 2018 CHI Conf. on Human Factors in Computing Systems, Paper 35*, Association for Computing Machinery, (Apr. 2018), 1–12.
14. Mortimore, T. and Crozier, W.R. Dyslexia and difficulties with study skills in higher education. *Studies in Higher Education 31*, 2 (Apr. 2006), 235–251.
15. Price, G.A. Creative solutions to making the technology work: Three case studies of dyslexic writers in higher education. *ALT-J 14*, 1 (Mar. 2006), 21–38.
16. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog 1*, 8, (2019), 9.
17. Rello, L., Baeza-Yates, R., Bott, S., and Saggion, H. Simplify or help? Text simplification strategies for people with dyslexia. In *Proceedings of the 10th Intern. Cross-Disciplinary Conf. on Web Accessibility, Article 15 in W4A '13*. Association for Computing Machinery, New York, NY, USA, (May 2013), 1–10.
18. Rello, L., Ballesteros, M., and Bigham, J.P. A spellchecker for dyslexia. In *Proceedings of the 17th Intern. ACM SIGACCESS Conf. on Computers & Accessibility, ASSETS '15*. Association for Computing Machinery (Oct. 2015), 39–47.
19. Rello, L., Pielot, M., and Marcos, M.-C. Make it big! the effect of font size and line spacing on online readability. In *Proceedings of the 2016 CHI Conf. on Human Factors in Computing Systems, CHI '16*. Association for Computing Machinery (May 2016), 3637–3648.
20. Thoppilan, R. et al. *LaMDA: Language models for dialog applications*, (Jan. 2022).
21. Valencia, S. et al. The less i type, the better: How ai language models can enhance or impede communication for aac users. In *Proceedings of the 2023 CHI Conf. on Human Factors in Computing Systems, CHI '23*. Association for Computing Machinery (2023).
22. van Schaik, M. *"Accept the Idea That Neurodiverse Kids Exist": Dyslexic Narratives and Neurodiversity Paradigm Visions*. Wilfrid Laurier University, 2021, PhD thesis.
23. Wu, S., Reynolds, L., Li, X., and Guzan, F. Design and evaluation of a social media writing support tool for people with dyslexia. In *Proceedings of the 2019 CHI Conf. on Human Factors in Computing Systems, number Paper 516 in CHI '19*. Association for Computing Machinery (May 2019), 1–14.
24. Yuan, A., Coenen, A., Reif, E., and Ippolito, D. Wordcraft: Story writing with large language models. In *27th Intern. Conf. on Intelligent User Interfaces, IUI '22*. Association for Computing Machinery (Mar. 2022), 841–852.
25. Zhang, M. et al. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.
26. Zhao, Z. et al. Calibrate before use: Improving few-shot performance of language models. In *Intern. Conf. on Machine Learning*. PMLR, (2021), 12697–12706.

Steven M. Goodman (smgoodmn@uw.edu), University of Washington, Google Research, Seattle, WA, USA.

Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N. Horne, Michal Lahav, Robert MacDonald, Rain Breaw Michaels, Ajit Narayanan, Mahima Pushkarna, Joel Riley, Alex Santana, Lei Shi, Rachel Sweeney, Phil Weaver, Ann Yuan, Google Research, Mountain View, CA, USA.

Meredith Ringel Morris, Google Research, Seattle, WA, USA.