Towards Explainable AI Writing Assistants for Non-native English Speakers

Yewon Kim
KAIST
Republic of Korea
yewon.e.kim@kaist.ac.kr

Mina Lee
Stanford University
United States
minalee@cs.stanford.edu

Donghwi Kim
KAIST
Republic of Korea
dhkim09@kaist.ac.kr

Sung-Ju Lee
KAIST
Republic of Korea
profsj@kaist.ac.kr

ABSTRACT

We highlight the challenges faced by non-native speakers when using AI writing assistants to paraphrase text. Through an interview study with 15 non-native English speakers (NNESs) with varying levels of English proficiency, we observe that they face difficulties in assessing paraphrased texts generated by AI writing assistants, largely due to the lack of explanations accompanying the suggested paraphrases. Furthermore, we examine their strategies to assess AI-generated texts in the absence of such explanations. Drawing on the needs of NNESs identified in our interview, we propose four potential user interfaces to enhance the writing experience of NNESs using AI writing assistants. The proposed designs focus on incorporating explanations to better support NNESs in understanding and evaluating the AI-generated paraphrasing suggestions.

Figure 1: AI tools for paraphrasing often generate multiple suggestions given a user text as input [44, 53]. In our interview, we observe that non-native English speakers (NNESs) struggle to assess and compare the quality of suggestions, particularly when suggestions lack explanations.

1 INTRODUCTION

Effective written communication skill is an integral part of academic and professional success [7, 24]. However, non-native English speakers (NNESs) often struggle to achieve the desired level of linguistic accuracy and complexity when writing in English [16, 36]. Despite their considerable efforts, they are prone to making errors and face difficulties in using diverse vocabularies [40, 43]. Moreover, they have a hard time selecting appropriate words or phrases in various contexts due to different cultural backgrounds [50, 52]. To improve their writing, many NNESs resort to AI writing assistants, such as grammar and style checkers [20], translators [19, 35], and paraphrasers [44, 53]. However, NNESs face unique challenges when using AI writing assistants, particularly paraphrasing tools, which often generate multiple paraphrasing suggestions without explanations (Figure 1). This lack of explainability can be especially critical for NNESs who might lack the ability to independently assess the appropriateness of these suggestions.

In this paper, we present findings from our preliminary interview study with 15 NNESs with varying levels of English proficiency and experiences using AI writing assistants. Through our study, we highlight the specific challenges faced by NNE speakers when paraphrasing with AI writing assistants, and suggest potential design implications for enhancing the explainability of these tools. Although our study encompassed all AI writing assistants, our findings and suggestions are primarily concerned with the challenges associated with the paraphrasing tasks (see Figure 3 and 4 in the Appendix for the illustration of paraphrasing capabilities). For the rest of the paper, we refer to these AI writing assistants used in paraphrasing tasks as AI tools for simplicity.

2 FINDINGS FROM THE INTERVIEW

To understand the challenges faced by NNESs when using AI tools, we conducted a semi-structured interview with 15 NNESs who frequently write English emails and use AI tools. We first asked participants to choose one of the email scenarios used in the prior work [32] and compose an email with the writing tools they use on a daily basis in a think-aloud manner. We then asked participants about the challenges they faced while using the tools and their solutions to those challenges. To analyze the qualitative data gathered from the interviews, we transcribed all audio and screen recordings and conducted open coding and thematic analysis. See Appendix A for a detailed description of the interview protocol and participants' backgrounds.

To improve their writing, many participants turned to the paraphrasing capabilities of AI tools and wished these tools could solve the challenges. Specifically, we observed that 12 participants used Grammarly [20], four used Wordtune [53], two used Quillbot [44], and one used Writefull [54] (see Appendix B.1 for the illustration of the paraphrasing features of each tool). Seven participants also used translators such as Google Translate [19] and Naver Papago [35] to paraphase the translated English outputs by perturbing their inputs in first language (see Appendix B.2 for the illustration).

2.1 Challenges Faced by NNESs in Writing

The thematic analysis revealed that NNESs encounter unique challenges when writing in English, which often leads them to rely on AI tools to address these difficulties. Out of 15 participants, 11 expressed difficulties in **accurately conveying their intention** in English sentences. Although NNESs knew the literal meaning of words, they often struggled with using those words in the appropriate context. For instance, P12 said: *"In Korean, the words 'ratio' and 'proportion' are translated the same, so I wrote like, 'The ratio of participants who are X is high.' My advisor got really angry at me and asked who uses 'ratio' in this case."* Similarly, they had difficulties in **controlling the nuances and tones** of sentences, as P8 stated: *"When it comes to emails or messages, the content is usually not too challenging. Nonetheless, I frequently find myself concerned about whether the nuances of my message align with my intended meaning. I worry that the message might come across as excessively polite or that it may be out of place with the given context."* Five participants also commented that their **writing often sounds unnatural**. P11 said: *"When I try to write something in English, I can put together a series of words that make sense, but it doesn't always sound natural. Instead, it may sound like the way sentences are structured in Korean, which I am not happy with."*

2.2 Challenges in Accepting Paraphrased Suggestions from AI Tools

In our interview, NNESs had mixed feelings about the effectiveness of AI tools, noting that while these tools were helpful in addressing certain difficulties, they were not always effective in meeting their specific writing needs. The main challenge was that when presented with multiple paraphrased suggestions from AI tools, NNESs **could not confidently choose (or "accept") the best suggestion**. In addition, it appeared that NNESs **did not fully trust AI tools**, as seven participants noted. Two participants, who were familiar with AI, were aware that the generative capabilities of AI are not guaranteed to be perfect, while the others learned through their experiences that AI tools often generated contextually inappropriate expressions as well as unnatural sentences with broken grammar and altered meanings. Four participants were also cautious towards accepting suggestions because they **did not know the rationale behind the suggestions provided by AI tools**. P6 questioned the reason for the changes in the paraphrased sentences: *"I am not sure why Quillbot suggested changing the sentence like that."* P8 also questioned what information the system is considering when generating suggestions: *"I am suspicious whether the tool knows if I am writing an email right now and giving me suggestions that fit the context."* As a result, they viewed these tools as compromise solutions that were at least better than themselves. P9 noted: *"When I use writing tools, I doubt the quality of my writing, but it is still better than writing without them. This is the best option for me."*

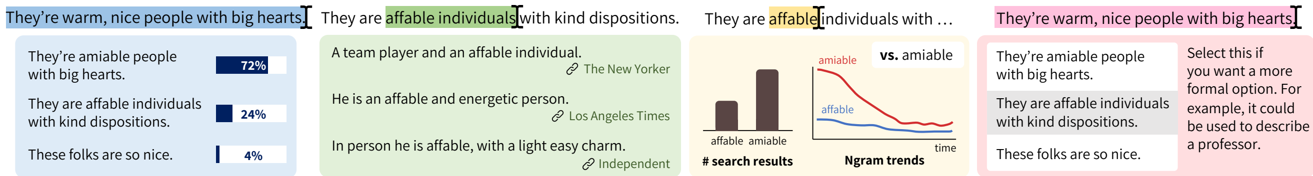
NNESs, particularly those with lower levels of English proficiency, seemed to face greater difficulties in accepting paraphrased suggestions compared to those with high proficiency. Six participants mentioned that they found accepting the paraphrased suggestion challenging, especially because they **did not know the tones and nuances** of texts presented to them. P4 noted: *"Even if it is grammatically correct, I often feel uncertain whether this suggestion* is polite. For example, I am not sure if the question I have written in this email to my advisor sounds polite or rude." Moreover, they were **worried that the suggestion might have changed the meaning of their original text**. P8 noted, *"I often doubt whether this suggestion is accurate. Take, for instance, the phrase 'apply for leave' - while 'leave' can refer to a vacation, I am concerned it could be interpreted in a different way, like quitting my job. I am worried that my boss might misunderstand my intentions."* In contrast, participants with higher levels of English proficiency reported less overhead in selecting the best suggestion. P10 stated: *"I write about eight emails every day and read a bunch of messages from native speakers, so I know which expressions are commonly used and have no trouble picking the right suggestion."*

2.3 NNESs' Strategies to Overcome Challenges

When selecting the suggestion, participants used various strategies to aid their decision-making. The most common strategy used by eight participants was referring to **human-authored texts from credible sources**. They referred to example sentences provided in web sources such as dictionaries, Ludwig [33], and Thesaurus [49] to understand in what context and how the suggested words or expressions are used. P11 believed the example sentences are written by native English speakers: *"I often refer to the Longman dictionary because it provides accurate information on how words are used in specific contexts. Its examples are taken from the native English corpus, so you can be confident that the words are used in the way that natives use them."* Three participants, who are academic researchers, mentioned that they refer to relevant papers to see if particular phrases are frequently used in the literature. Two even found YouTube useful, as P5 stated, *"I would search 'I wish you are doing well' in YouTube to watch educational videos that explain the specific contexts this expression could be used. The fact that I can see the instructor's face, check their subscribers count, and read all the positive comments makes those videos much more trustworthy than a dictionary."*

Four participants appreciated **textual explanations** of suggestions, with three specifically noting Grammarly's explanation features, which not only suggest paraphrases but also explain why the suggested changes are necessary. P2 stated: *"When you read Grammarly's explanations, you can learn why you should make the corrections. It is helpful because it helps you become more careful when using similar expressions in the future."* P5, who did not use Grammarly, wished for textual indicators that explain *"the most appropriate contexts"* to use the expression as well as *"the comparison between two expressions with similar meanings."* P4 also wished to know the reasoning behind the suggestions from AI tools such as translators, as he noted: *"I wish the tool also generates the reason it is paraphrasing the original sentence like this."*

Three participants found **statistical evidence** helpful. Specifically, two participants searched Google to refer to the number of search results. P11 explained: *"One way to find out if an expression is commonly used by native speakers is to search it on Google and see how many results you get."* Moreover, P8 referred to Google's Ngram



(a) Model confidence **(b) Examples from credible sources** **(c) Data-driven usage trends** **(d) Textual explanations**

Figure 2: Example user interfaces portraying design implications for non-native English speakers based on our findings. Explanations can be presented in the form of (a) confidence scores of model outputs, (b) example sentences from credible sources, (c) comparison of usage trends of original and paraphrased suggestions, and (d) textual explanation of the suggestions.

viewer [18] to see if the word or expression is trendy. Lastly, two participants mentioned model confidence scores. For instance, P9 appreciated Writefull [54] for displaying confidence scores alongside multiple suggestions, where the confidence score, represented as a percentage, indicates the likelihood that an AI's suggestion is accurate [56] (see Figure 3d for an illustration of confidence scores). He stated: *"If something has a really high confidence score, like 100%, it is probably a good idea to pay attention to it and give it some extra consideration. But if it is more like 50%, it is worth considering, while also keeping in mind that there might be other options you can explore."* On the other hand, P8 stopped using Wordtune [53] because *"there are so many suggestions, but I have no clue which one is the winner,"* which made him feel like *"the suggestions are made carelessly,"* wishing for *"a ranking system based on model confidence."*

Seven participants mentioned they have given up on obtaining satisfactory results. P12 noted: *"Grammarly does not fix everything, but that does not mean I am trying to improve the sentence any further. I just acknowledge that it might not be perfect, but it is the best I can do. Trying to fix all the errors on my own would cost me too much. Plus, getting the nuances of sentences right is something I mostly give up on. There are parts I do not know at all, so there is not much I can do about it."* Furthermore, NNESs appeared to perceive the feedback provided by native English speakers to be the ultimate form of assistance. For important tasks such as paper writing or professional emails, three participants opted to seek human resources, such as native friends or proofreading services. P5 commented: *"Whenever I ask my native friends for feedback, they usually say something like, 'It is not wrong, but we usually say it like this.' So, when I am writing something important, I prefer to seek their help rather than relying on writing tools."*

3 DESIGN IMPLICATIONS

By drawing inspiration from explainable AI (XAI) literature [21] and language acquisition theory [28], we explore the design implications of AI tools and offer recommendations for addressing the challenges NNESs encounter identified in our interview. Figure 2 presents four potential explanation features of AI tools.

3.1 Revealing AI Models' Internal Mechanisms

Our study suggests that NNESs need additional support in accepting suggestions as they struggle to evaluate paraphrased suggestions of AI tools and question the rationale behind these suggestions. A viable solution could involve elucidating the behavior of AI models by revealing their internal mechanisms. One such approach could involve presenting the **confidence scores** of model outputs (Figure 2a), which indicate the probability of each model output and potentially provide users with a sense of when to trust or distrust the model [56]. Previous research on prediction tasks has demonstrated that user performance improves with high accuracy indicators [30] and user trust increases in high-confidence cases [56]. However, caution should be exercised, as model confidence could be misleading [30] and might result in low human trust in the model [55]. For instance, high confidence score does not necessarily imply high accuracy and models can be overconfident in their predictions [55]. Additionally, confidence scores might be calibrated using temperature scaling [22], which could affect the presentation of the confidence scores and accordingly the users' decisions.

In addition to confidence scores, prior studies showed that visualizing **attention scores** [3] of language models can help improve explainability [1, 23]. Attention scores provide a measure of the importance of different input tokens when generating a specific output token. For example, visualizing the strength of dependencies between input-output tokens could help users understand which input tokens significantly impact the model's output, revealing the relationship between tokens and their influence on the generation [1]. Such an approach could enhance NNESs' understanding of contextual associations in English by providing insights into the relationship between words in original and paraphrased texts. For example, a user might learn polite expressions by analyzing paraphrased words with high attention scores that correspond to polite expressions they are already familiar with.

3.2 Presenting Real-world Language Use Cases

In this study, participants assessed the appropriateness of paraphrased suggestions by seeking real-world evidence, such as example sentences from credible sources or data-driven trends such as Google search results. This behavior aligns with the findings of prior research that suggested creating a corpus of situation-specific facts is essential before generalizing across situations in learning tasks [9]. Another line of research in the Natural Language Processing (NLP) field, which aims to correct language learners' errors by retrieving example sentences [27] also complies with this finding. In line with this, we hypothesize that NNESs' tendency to assess suggestions using real-world examples reflects *learning with understanding*. Specifically, NNESs engage in an *implicit* language learning process [12], whereby individuals acquire language through exposure to language inputs and communicative interactions in real-life contexts.

As such, we propose that effective presentation of **example sentences** (Figure 2b) along with paraphrased suggestions could enhance NNESs' decision-making and language learning processes. A critical aspect to consider when selecting credible sources of sentences would be the use of corpora written by native English speakers. This consideration stems from the findings of our interview study, in which participants perceived texts written by native English speakers (e.g., Longman Corpus Network [42]) as credible. Another interesting approach could be presenting use cases from **educational videos**, as two participants found YouTube educational videos credible based on subscriber counts, positive comments, and the instructor's appearance. Supporting NNESs' decision-making could also be achieved by visualizing **data-driven usage trends** (Figure 2c). Deriving quantitative evidence from big data could offer a straightforward approach to representing word prevalence. However, it is essential to carefully contextualize the information, as naïve search results might not be a reliable indicator of appropriateness. Various factors, including search algorithms and paid advertising, could influence these results [41].

3.3 Generating Textual Explanations

Integrating textual explanations (Figure 2d) into AI tools could foster natural and effective AI-powered support for NNESs. These explanations could take the form of local explanations to **provide the rationale for a single suggestion** and help users determine whether to trust a model on a case-by-case basis [45]. Alternatively, the model could be given the context of writing, such as professional email writing or academic writing, and **generate feedback comments** [34] on inappropriate user texts, explaining why certain expressions are unsuitable within the underlying context. Another approach would involve **providing general explanations** that instruct users on the typical use cases of certain words or expressions. For example, when generating feedback comments on inappropriate usage of expressions, AI tools could also explain the contexts in which those expressions can be used. This approach can offer users an *explicit* guidance [12] on how language is used to convey meaning in different situations [2, 47]. In addition, the effective presentation of textual explanations is a crucial consideration, particularly for NNESs who might struggle to comprehend the generated explanations due to limited English proficiency. Possible solutions include translating explanations into the user's first language or simplifying the language used in the explanations.

4 DISCUSSION

4.1 Trade-off Between Efficiency and Quality

In general, considering the trade-off between efficiency and quality is important in designing explanations [5]. Providing complex explanations (e.g., Figure 2b) may result in better suggestion selection and facilitate learning; however, it could take longer to select suggestions than simpler features (e.g., Figure 2a). Conversely, while simple features may help users quickly select suggestions, they might be difficult to comprehend, resulting in lower quality and learning outcomes. Moreover, users with varying levels of English proficiency and needs may perceive efficiency differently. Prior study [5] have suggested that the cost of efficiency in accepting suggestions might be more burdensome for native speakers than NNESs. Similarly, we observed in our interview that participants with a higher level of English proficiency experienced less overhead when selecting suggestions compared to those with low proficiency. In such cases, providing explanations may not be necessary or may even hurt overall efficiency [5].

4.2 Learning Effects of Explanations

Providing explanations to NNESs may facilitate their language learning [12]; however, it is crucial to carefully design the explanations to maximize the learning effects of NNESs. A key design consideration involves balancing explicit and implicit learning guidance to promote efficient language acquisition [14]. Explicit learning guidance, such as textual instructions, could accelerate the learning process, but it may not always translate to flexible language application in diverse real-life contexts [11, 13]. On the other hand, implicit learning, such as providing example sentences, can be effective in helping NNESs generalize to new contexts [15] but may result in over-reliance on such examples [9]. Additionally, an imbalanced amount and abstraction level of information could result in unsatisfactory learning experiences. Oversimplification of the explanations is known to increase the mental load on users and decrease their trust in the explanations [29], as well as causing over-reliance on AI [51]. Conversely, presenting an excessive amount of information could overwhelm users [46].

4.3 Dialogue-based User Interaction

In addition to the conventional one-way interaction where human users accept or reject suggestions from AI tools, these assistants could also engage in a dialogue with users to better understand their needs and preferences. Recent advances in language models, such as ChatGPT [39] and GPT-4 [38], have already demonstrated the potential for dialogue-based user interaction. These models allow users to freely communicate with AI, providing them with greater control and flexibility to steer the model's output. However, this increased freedom also introduces new challenges, as NNESs might face challenges effectively communicating their intentions through English prompts, which could lead to suboptimal outputs or miscommunication between the user and AI. Future work can focus on assisting NNESs to effectively query the model. This may involve creating more NNE user-friendly interfaces or providing users with suggested prompts based on their specific needs. Additionally, research aimed at enhancing the robustness of AI models to variations in input can help mitigate the impact of poorly constructed prompts.

4.4 Potential Bias AI Tools May Cause

The design decisions involved in presenting explanations can unintentionally introduce further bias into AI tools. For instance, the manner in which an explanation is visualized or presented to the user could influence their decision-making process [4, 10]. This could lead to a biased text outcome, as recent work [25] suggests that collaborative writing with AI affects users' views, possibly affecting the contents of the written text. Moreover, NNESs might be particularly susceptible to this issue, as their lack of expertise or familiarity with the language might make them more vulnerable to the biases introduced by the AI outputs, leading to over-reliance [6].

ACKNOWLEDGMENTS

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00064, Development of Human Digital Twin Technologies for Prediction and Management of Emotion Workers' Mental Health Risks).

REFERENCES

[1] David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 412–421. https://doi.org/10.18653/v1/D17-1042

[2] Fauzul Aufa. 2016. Explicit Pragmatic Instruction In Teaching English As A Foreign Language. *Journal of English and Education* 5, 1 (Apr. 2016), 37–44. https://journal.uii.ac.id/JEE/article/view/4458

[3] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

[4] Joeran Beel and Haley Dixon. 2021. The 'unreasonable'effectiveness of graphical user interfaces for recommender systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 22–28.

[5] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. https://doi.org/10.1145/3411764.3445372

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. 160–169. https://doi.org/10.1109/ICHI.2015.26

[7] Jenny Cameron, Karen Nairn, and Jane Higgins. 2009. Demystifying Academic Writing: Reflections on Emotions, Know-How and Academic Identity. *Journal of Geography in Higher Education* 33, 2 (2009), 269–284. https://doi.org/10.1080/03098260902734943 arXiv:https://doi.org/10.1080/03098260902734943

[8] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2287–2295. https://doi.org/10.1145/3292500.3330723

[9] Michelene T.H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science* 13, 2 (1989), 145–182. https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1302_1

[10] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing? How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 585–592.

[11] Robert DeKeyser. 1998. Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. *Focus on form in classroom second language acquisition* 28 (1998), 42–63.

[12] Robert DeKeyser. 2003. Implicit and explicit learning. *The handbook of second language acquisition* (2003), 312–348.

[13] Robert DeKeyser. 2020. *Skill Acquisition Theory*. 83–104. https://doi.org/10.4324/9780429503986-5

[14] Catherine Doughty and Jessica Williams. 1998. Pedagogical choices in focus on form. *Focus on form in classroom second language acquisition* 3 (1998), 197–262.

[15] Nick Ellis, Ute Römer, and Matthew O'Donnell. 2016. *Usage-based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*.

[16] Muhammad Fareed, Almas Ashraf, and Muhammad Bilal. 2016. ESL learners' writing skills: Problems, factors and suggestions. *Journal of education and social sciences* 4, 2 (2016), 81–92.

[17] Google. 2023. Google Docs AutoCorrect. https://docs.google.com/document/u/0/.

[18] Google. 2023. Google Ngram Viewer. https://books.google.com/ngrams/.

[19] Google. 2023. Google Translate. https://translate.google.com/.

[20] Grammarly. 2023. Grammarly: Free Online Writing Assistant. https://www.grammarly.com.

[21] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine* 40, 2 (Jun. 2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[23] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 187–196. https://doi.org/10.18653/v1/2020.acl-demos.22

[24] Julie S. Hui, Darren Gergle, and Elizabeth M. Gerber. 2018. IntroAssist: A Tool to Support Writing Introductory Help Requests. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173596

[25] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. *arXiv preprint arXiv:2302.00560* (2023).

[26] Kakao. 2023. Kakao i Translate. https://translate.kakao.com/.

[27] Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for Language Learners Using Example-Based Grammatical Error Correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 7176–7187. https://doi.org/10.18653/v1/2022.acl-long.496

[28] Stephen Krashen. 1981. Second language acquisition. *Second Language Learning* 3, 7 (1981), 19–39.

[29] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[30] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[31] LanguageTool. 2023. LanguageTool - Online Grammar, Style & Spell Checker. https://languagetool.org/.

[32] Hajin Lim, Dan Cosley, and Susan R Fussell. 2022. Understanding Cross-lingual Pragmatic Misunderstandings in Email Communication. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–32.

[33] Ludwig. 2023. Ludwig • Find your English sentence. https://ludwig.guru/.

[34] Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared Task on Feedback Comment Generation for Language Learners. In *Proceedings of the 14th International Conference on Natural Language Generation*. Association for Computational Linguistics, Aberdeen, Scotland, UK, 320–324. https://aclanthology.org/2021.inlg-1.35

[35] Naver. 2023. Papago. https://papago.naver.com/.

[36] David Nunan. 1989. *Designing tasks for the communicative classroom*. Cambridge university press.

[37] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

[38] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[39] OpenAI. 2023. Introducing ChatGPT. https://openai.com/blog/chatgpt.

[40] Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics* 24, 4 (2003), 492–518.

[41] Eli Pariser. 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books, USA.

[42] Pearson. 2023. Longman Dictionaries: Corpus Network. http://www.pearsonlongman.com/dictionaries/corpus/.

[43] Charlene G Polio. 1997. Measures of linguistic accuracy in second language writing research. *Language learning* 47, 1 (1997), 101–143.

[44] Quillbot. 2023. Paraphrasing Tool - QuillBot AI. https://quillbot.com.

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[46] Peter Gordon Roetzel. 2018. Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research* (2018), 1–44.

[47] Kenneth R. Rose. 2005. On the effects of instruction in second language pragmatics. *System* 33, 3 (2005), 385–399. https://doi.org/10.1016/j.system.2005.06.003

Pragmatics in Instructed Language Learning.
[48] Tabnine. 2023. Tabnine: AI assistant for software developers. https://www.tabnine.com/.
[49] Thesaurus.com. 2023. Thesaurus.com: Synonyms and Antonyms of Words. https://www.thesaurus.com/.
[50] Evangeline Marlos Varonis and Susan M Gass. 1985. Miscommunication in native/nonnative conversation. *Language in society* 14, 3 (1985), 327–343.
[51] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. arXiv:2212.06823 [cs.HC]
[52] Jane A Vignovic and Lori Foster Thompson. 2010. Computer-mediated cross-cultural collaboration: Attributing communication errors to the person versus the situation. *Journal of Applied Psychology* 95, 2 (2010), 265.
[53] Wordtune. 2023. Wordtune - Rewrite Text in Seconds. https://www.wordtune.com.
[54] Writefull. 2023. Writefull. https://www.writefull.com/.
[55] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509
[56] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

A INTERVIEW DETAILS

A.1 Interview Protocol

We employed a semi-structured interview protocol to gather insights from participants regarding their email writing experiences. During the interview, the participants were first asked to select one of the provided email scenarios introduced in the prior work [32] (see Table 1 for the full list of scenarios) that they found most relevant to their personal or professional context. Participants were encouraged to write in their naturalistic settings, which may involve the use of AI writing assistants (e.g., Grammarly and Wordtune) or online searches (e.g., Thesaurus and dictionaries) to aid their writing process. Throughout the task, participants were asked to verbalize their thought processes in a think-aloud manner. Upon completion, we inquired the participants about their writing experiences, focusing on the challenges encountered during the session, strategies employed to overcome these difficulties, and suggestions for enhancing the writing tools they utilized.

A.2 Participant Details

We recruited 15 NNE speakers (five females, four males, six preferred not to say) who engage in English writing activities at least once a week. The participants included six undergraduate students, seven graduate students, one research assistant, and one information technologist. Participants were recruited through social networking services. Each participant received a compensation of 35,000 KRW (approximately 27 USD) for participating in the study. The study was approved by the Institutional Review Board of the first author's institution. Table 2 shows the detailed background of the interview participants we recruited. When recruiting participants, we asked the participants to self-assess their English proficiency according to the Common European Framework of Reference for Languages (CEFR) [37] measurement. CEFR levels are defined as basic (beginner and elementary levels; A1 and A2), independent (intermediate and upper intermediate levels; B1 and B2), and proficient (advanced and proficiency levels; C1 and C2). We provided the following rubrics to the participants for rating one's English proficiency. As a result, we recruited two basic-level, seven independent-level, and six proficient-level participants.

- A1 (Beginner): You can understand and use basic phrases and expressions. You can communicate in simple ways when people speak slowly to you.
- A2 (Elementary): You can take part in simple exchanges in familiar topics. You can understand and communicate routine information.
- B1 (Intermediate): You can communicate in situations and use simple language to communicate feeling, opinions, plans and experiences.
- B2 (Upper Intermediate): You can communicate easily with native English speakers. You can understand and express some complex ideas and topics.
- C1 (Advanced): You can understand and use a wide range of language. You can use English flexibly and effectively for social and academic purposes.
- C2 (Proficiency): You can understand almost everything you hear or read. You can communicate very fluently and precisely in complex situations.

1	You ask your professor you took a class with a while ago to introduce you to someone who may be hiring in your chosen career path.
2	You would like to take a week off from work to attend the wedding of a friend who lives abroad. You are emailing your boss to ask for a week off.
3	You have an interview for a new job next week, but you realize you cannot make it on the scheduled date. You are emailing the hiring manager to reschedule.
4	You are interested in working at a particular company and heard that an alumnus of your college currently works there. You are emailing this person to ask if they would speak with you about their company in a video call.
5	You are in charge of a fundraising campaign at work to help needy children. You are emailing your colleagues to ask them to donate to this campaign.
6	You received an email from your apartment management office fining you for not recycling properly. You believe the fine is an error and are emailing the office to ask them to cancel it.

Table 1: The six email scenarios we used in the interview study. We borrowed the scenarios from the previous literature [32].

B AI WRITING ASSISTANTS DETAILS

B.1 Paraphrasing Functionalities

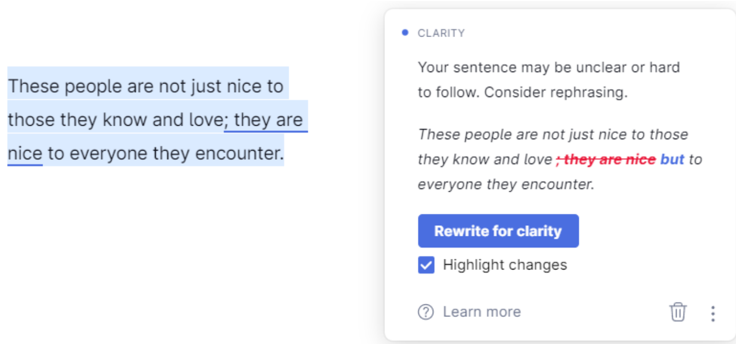
Figure 3 provides the screenshots of the paraphrasing features in the AI writing tools (Grammarly [20], Wordtune [53], Quillbot [44], and Writefull [54]) that the participants mentioned during the interview.

B.2 Using Translators as Paraphrasers

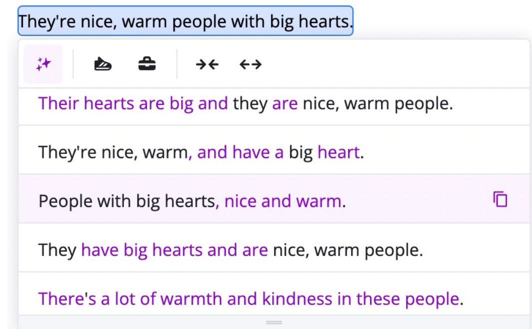
Figure 4 illustrates how users used translators as paraphrasing tools. Users employed two strategies to use translators as paraphrasing tools, either by perturbing their inputs in first language within a single translator or by translating the same L1 input with different systems. In our interview, participants used among Google Translate [19], Naver Papago [35], and Kakao i Translate [26].

	English Proficiency	First Language	Residence	AI Writing Assistants	Profession (Major)
P1	A1	Korean	South Korea	Grammarly, Naver Papago	Undergraduate student (Electrical Engineering)
P2	A2	Korean	South Korea	Grammarly, Naver Papago	Undergraduate student (Computer Science)
P3	B1	Korean	South Korea	Naver Papago, Google Docs AutoCorrect	Undergraduate student (Medicine)
P4	B1	Korean	United States	Grammarly, Google Translate, Naver Papago, Google SmartCompose, Tabnine	Graduate student (Information Science)
P5	B1	Korean	Sweden	Google Translate, Google Docs AutoCorrect, Google SmartCompose	Undergraduate student (Education)
P6	B2	Chinese	United States	Grammarly, Google Translate, Quillbot	Research assistant (Human-Computer Interaction)
P7	B2	French	United States	Grammarly, Google Translate, Wordtune	Information technologist
P8	B2	Korean	South Korea	Grammarly, Google Translate, Naver Papago, Kakao i Translate, Wordtune	Graduate student (Human-Computer Interaction)
P9	B2	Korean	South Korea	Grammarly, Google Translate, Naver Papago, Kakao i Translate, Google SmartCompose, Writefull	Graduate student (Computer Science)
P10	C1	Korean	United States	Grammarly, Wordtune, Google SmartCompose	Graduate student (Social Science)
P11	C1	Korean	United Kingdom	Wordtune	Graduate student (Psychology)
P12	C1	Korean	South Korea	Grammarly, Google Translate, Google SmartCompose	Graduate student (Computer Science)
P13	C2	Korean	South Korea	Grammarly	Undergraduate student (Business)
P14	C2	Filipino	South Korea	Grammarly, Google Translate, Naver Papago	Graduate student (Electrical Engineering)
P15	C2	Hindi	India	Grammarly, Quillbot, LanguageTool	Undergraduate student (Information Technology)

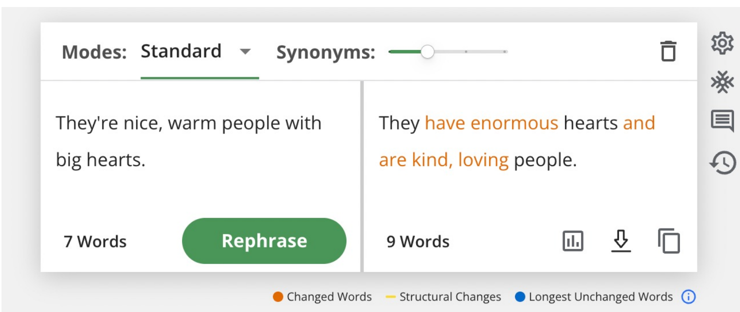
Table 2: Detailed background information of the interview participants. The participants used various sets of AI writing assistants, which include Grammarly [20], Google Translate [19], Naver Papago [35], Google SmartCompose [8], Wordtune [53], Google Docs AutoCorrect [17], Kakao i Translate [26], Quillbot [44], Writefull [54], LanguageTool [31], and Tabnine [48]. Refer to Section A.2 in Appendix for the descriptions of english proficiency levels.



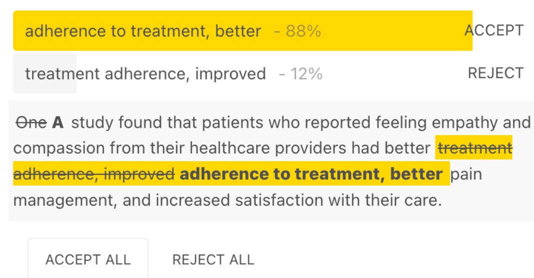
(a) Grammarly [20]



(b) Wordtune [53]

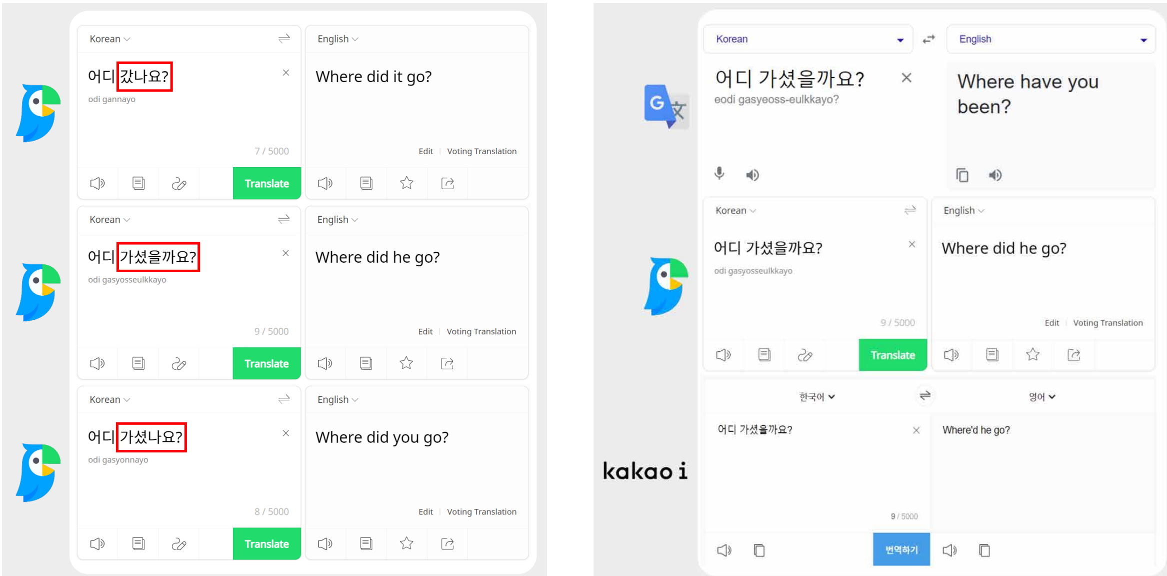


(c) Quillbot [44]



(d) Writefull [54]

Figure 3: Screenshot of the paraphrasing features provided in each AI writing tools. (a) When a user clicks underlined words or phrases, Grammarly [20] displays an explanation of the issue and offers suggestions for improvement. Users can either accept the suggestion by clicking "Rewrite for clarify" button or reject it by clicking the trashcan icon on the bottom right corner of the display. (b) Wordtune [53] suggests multiple paraphrased suggestions for each input. Users can accept or copy one of the suggestions, or dismiss the suggestions. (c) Quillbot [44] suggests a single paraphrased suggestion, where users can accept it or regenerate the suggestion by clicking "Rephrase" button. (d) When a user clicks underlined words or phrases, Writefull [54] shows a paraphrased suggestion and compares the original and paraphrased versions by showing confidence scores in percentage.



(a) Perturbing L1 inputs.

(b) Using multiple translators.

Figure 4: Illustration of how interview participants used translators as paraphrasing tools. In (a), only the honorific forms of texts in first language (L1; in this example, Korean) are perturbed (perturbations are indicated by red boxes) and translated using Naver Papago [35]. In (b), the same L1 input was translated using Google Translate [19], Naver Papago [35], and Kakao i Translate [26].