Exploring Student-ChatGPT Dialogue in EFL Writing Education

Jieun Han[*] **Haneul Yoo**[*] **Junho Myung** **Minsun Kim**
Tak Yeon Lee **So-Yeon Ahn**[†] **Alice Oh**[†]
KAIST
`{jieun_han, haneul.yoo, junho00211, 9909cindy}@kaist.ac.kr`
`{takyeonlee, ahnsoyeon}@kaist.ac.kr` `alice.oh@kaist.edu`

Abstract

The integration of generative AI in education is expanding, yet empirical analyses of large-scale and real-world interactions between students and AI systems still remain limited. Addressing this gap, we present RECIPE4U (RECIPE for University), a dataset sourced from a semester-long experiment with 213 college students in English as Foreign Language (EFL) writing courses. During the study, students engaged in dialogues with ChatGPT to revise their essays. RECIPE4U includes comprehensive records of these interactions, including conversation logs, students' intent, students' self-rated satisfaction, and students' essay edit histories. In particular, we annotate the students' utterances in RECIPE4U with 13 intention labels based on our coding schemes. We establish baseline results for two subtasks in task-oriented dialogue systems within educational contexts: intent detection and satisfaction estimation. As a foundational step, we explore student-ChatGPT interaction patterns through RECIPE4U and analyze them by focusing on students' dialogue and students' essay edits. We further illustrate potential applications of RECIPE4U dataset for enhancing the incorporation of LLMs in educational frameworks. RECIPE4U is publicly available at `https://github.com/zeunie/RECIPE4U/`.

1 Introduction

The adoption of LLMs in education is accelerating, particularly in English as a Foreign Language (EFL) contexts [13]. A noteworthy example is ChatGPT [1], a large language model (LLM)-driven chatbot developed by OpenAI, increasingly perceived as a beneficial resource for higher education students in research and writing tasks [18]. EFL learners frequently face apprehension in exposing their linguistic shortcomings to instructors or peers [3]. In this light, LLM-assisted tools tailored for English writing can alleviate such embarrassment by providing non-judgmental feedback. These tools foster a more supportive learning environment, as neither social distance nor a power relationship was found between students and the AI tools [32]. Nevertheless, most LLM-based tools, including ChatGPT, are not originally designed for educational purposes. As their application in EFL education grows, it is necessary to explore both the potential and the actual utilization patterns of LLMs in EFL writing education. Despite such needs, previous research falls short of conducting comprehensive, long-term analyses of real-world usage of LLMs within educational settings [27, 11]. RECIPE [13] is the first attempt to suggest a platform to capture the semester-long interaction between students and ChatGPT in real-world EFL writing education.

We release RECIPE4U (RECIPE for University) dataset, which is derived from EFL learners in the university collected through the RECIPE platform. This dataset allows the detection of students' intent

[*] Authors contributed equally.

[†] Corresponding Authors.

[1] `https://chat.openai.com/`

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

in their prompts (intent detection) and the estimation of their satisfaction with ChatGPT responses (satisfaction estimation), thereby establishing baseline models for two subtasks. We conduct an analysis of students' interaction patterns, contributing to a deeper understanding of the potential for future development in LLM-integrated English writing education.

The main contributions of this work include

1. RECIPE4U (RECIPE for University), student-ChatGPT interaction dataset that captures semester-long learning process within the context of real-world EFL writing education.
2. Baseline models for two subtasks with RECIPE4U: intent detection and satisfaction estimation.
3. An investigation into the students' interaction patterns with conversation logs and essay edits in RECIPE4U for enhancing LLM-integrated education.

2 Related Work

2.1 LLM-integrated Education

There is a growing body of research exploring applications of LLM in educational contexts [19, 23, 21], but much of this research has focused on the short-term effects involving just a few experimental sessions. However, what we need is an investigation into long-term trends and usage patterns among students in the context of discourse analysis in education [1] with a long-term analysis examining previous related interactions and contextual knowledge shared by students [24]. Also, LLM-integrated education is underexplored within the context of English as a foreign language (EFL) education where prior research [27, 11, 33] has predominantly relied on episodic and anecdotal knowledge [13]. In this paper, we investigate the potential of a systematic use of LLMs in EFL education in a semester-long university course.

2.2 Dialogue Data in Education

One of the common themes of educational dialogue analysis is the refinement and development of a coding scheme for dialogue acts [12, 2, 6, 22]. However, previous work lacks annotation regarding students' underlying intentions or purpose behind their speech acts [6, 22]. Demszky et al. [6] identifies five uptake strategies related to teachers' utterances, which leaves out a comprehensive analysis of students' utterances. Marineau et al. [22] classifies the students' speech acts into four categories: assertions, wh-questions, yes/no questions, and directives. While these categories are useful for understanding the surface-level characteristics of students' utterances, they may not fully capture the nuances of their dialogue intentions.

Moreover, it is important to consider the specific domain of English education and the unique context of human-AI interaction. Currently available datasets in this field predominantly involve human-human interactions and do not examine the domain of EFL writing education [6, 28]. To the best of our knowledge, our research represents a pioneering effort in introducing a dataset derived from real-world human-AI interactions within the context of EFL writing education. Additionally, we aim to shed light on the often-overlooked aspect of students' dialogue acts and their underlying intentions, thus contributing to a more comprehensive understanding of educational dialogues in this domain.

2.3 Task-oriented Dialogue Dataset

As a practical need in the industry, conversational AI has focused on delving into task-oriented dialogue (ToD) systems that can help specific daily-life tasks such as reservation and information query [17]. Various ToD datasets in the domain of everyday life were publicly released, including FRAMES [9], M2M [30], and DSTC-2 [16] for reservation and MultiWOZ 2.2 [35], KVRET [10], SNIPS [5], and ATIS [15] for information query, inter alia. Recently, ToD systems also shed light on professional domains such as medical and education [34].

To the best of our knowledge, Zhang et al. [36] constructed GrounDialog, the first ToD dataset specifically tailored for language learning, with respect to repair and grounding (R&G) patterns between high and low proficiency speakers of English. However, GrouDialog deals with a task of information query on a job interview, which lacks relevance to language learning and education and is not publicly available.

Alongside the need for cross-lingual language modeling, several studies introduced multilingual ToD datasets, mostly constructed by translating English dialogues [29, 20]. Still, code-mixed ToD datasets are hardly available [8].

	FRAMES	M2M	MultiWOZ 2.2	GrounDialog	RECIPE4U (Ours)
# of dialogues	1,369	1,500	8,438	42	504
Total # of utterances	19,986	14,796	115,424	1,569	4,330
Total # of tokens	251,867	121,977	1,520,970	14,566	380,364
Avg. utterances per dialogue	14.60	9.86	13.68	37.36	3.38
Avg. tokens per utterance	12.60	8.24	13.18	9.28	87.84
Total unique tokens	12,043	1,008	24,071	unknown	16,118
Languages	English	English	English	English	English, Korean
Code-mixed	✘	✘	✘	✘	✔
Publicly available	✔	✔	✔	✘	✔
Additional data	✘	✘	✘	✘	Essay edit history
Task	Find the best deal of hotels and flight	Buy movie ticket / Reserve restaurant	Get info. about touristic city	Get info. about job interview	Revise English writing

Table 1: Statistics of RECIPE4U compared to existing task-oriented dialogue datasets

3 RECIPE4U Dataset

We gather student-ChatGPT interaction data through RECIPE (Revising an Essay with ChatGPT on an Interactive Platform for EFL learners) [13], a platform designed to integrate ChatGPT with essay writing for EFL students. The main component of RECIPE is a writing exercise where students write and revise their essays while conversing with ChatGPT. We provide students with instructions to revise an essay while having a conversation about what they learned in class. The student is shown a user interface that AI agent initiates the conversation by requesting a class summary. RECIPE incorporates `gpt-3.5-turbo` prompted with 1) a persona of an English writing class teacher and 2) step-by-step guidance to students in the platform.

A semester-long longitudinal data collection involves 213 EFL students (91 undergraduate and 122 graduate students) from a college in South Korea. They are enrolled in one of the three different English writing courses: Intermediate Writing, Advanced Writing, and Scientific Writing. Undergraduate students were divided into two courses depending on their TOEFL writing scores (15-18 for Intermediate Writing and 19-21 for Advanced Writing). In both courses, one of the primary assignments is writing an argumentative essay. Scientific Writing course is designed for graduate students, aiming to teach them how to write scientific research papers.

In total, RECIPE4U contains 4330 utterances (1913 students' utterances and 2417 ChatGPT's utterances), including 97 single-turn and 407 multi-turn dialogues. The conversation is mostly done in English, but there are several instances of code-switching between English and Korean, as the majority of the student's first language is Korean. Table 1 describes detailed statistics of RECIPE4U dataset compared to existing task-oriented dialogue datasets.

Unlike other task-oriented dialogue dataset, RECIPE4U includes additional source data, which is 1913 utterance-level essay edit history. We collect students' essay edit history at each utterance level to explore students' learning process. Students voluntarily provide their essays to RECIPE [13] and make necessary edits while having a conversation with ChatGPT on topics regarding essay writing.

We gather students' self-rated satisfaction levels to analyze the students' learning experiences and gain insights into how they perceived and evaluated their interactions with ChatGPT on a five-Likert scale. Specifically, each time a student engages in a conversation, RECIPE asks students to self-rate their level of satisfaction with ChatGPT's last response. In addition, we add tags to the students' written essays, paragraphs, and sentences for future application.

4 Experiment

In this paper, we suggest two subtasks leveraging RECIPE4U data: intent detection (§4.1) and satisfaction estimation (§4.2).

4.1 Intent Detection

Intent detection is the task of classifying the students' utterances into 13 predefined intent categories. This task involves intent classification based on a student's utterance, the preceding response from ChatGPT, and the subsequent response from ChatGPT.

Intent		Satisfaction						Avg. of
div1	div2	1	2	3	4	5	Total	Satisfaction
Response	Acknowledgement	2	10	21	106	177	316	4.41
	Negotiation	4	8	18	32	35	97	3.89
	Answer	8	9	63	165	157	402	3.83
Request	Request for Translation	2	2	4	9	7	24	3.71
	Request for Confirmation	1	1	3	10	7	22	3.96
	Request for Language Use	7	13	42	114	130	306	4.13
	Request for Revision	6	6	28	58	66	164	4.05
	Request for Evaluation	8	8	17	50	53	136	3.97
	Request for Information	6	10	28	97	110	251	4.18
	Request for Generation	1	7	5	21	26	60	4.07
	Question	4	6	14	15	16	55	3.60
Misc.	Statement	2	1	8	19	26	56	4.13
	Other	4	0	4	4	12	24	4.18
Total		55	81	255	700	822	1913	4.13

Table 2: Number of samples by intent and satisfaction

4.1.1 Intent Label

We design students' intention annotation schemes, comprising 13 intention labels. This scheme builds upon and complements the analysis of intention in dialogue from previous research [12, 2, 25]. From the study on task-oriented dialogue in educational settings by Ha et al. [12], we adopted eight intention labels: Acknowledgement, Other, Request for Confirmation, Request for Revision, Request for Evaluation, Question, Answer, and Statement. From Ozkose-Biyik and Meskill [25]'s research on learner reciprocity, we integrated the Negotiation and Request for Information labels. To better cater to the context of student-AI interactions in EFL writing education, we introduced three additional labels: Request for Translation, Request for Language Use, and Request for Generation. These 13 intention labels can be further grouped into three ancestor categories, division 1 (div1): Response, Request, and Miscellaneous. Under the Response category, we include three labels from division 2 (div2): Acknowledgement, Negotiation, and Answer. The Request category encompasses eight labels: Request for Translation, Request for Confirmation, Request for Language Use, Request for Revision, Request for Evaluation, Request for Information, Request for Generation, and Question. Lastly, the Miscellaneous class includes Statement and Other. The descriptions and examples of 13 labels are shown in Appendix A.

4.1.2 Intent Annotation

Four authors engage in an iterative process of collaborative and independent tagging of the student's intent. In the initial tagging phase, all annotators collaboratively tag 10.45% of the dialogue dataset. Subsequently, the remaining samples undergo annotation independently by two annotators. In cases where disagreements arise, all four authors discuss once more to tag until a consensus is reached.

Table 2 shows the distribution of student intentions and satisfaction levels. The most frequent intention was Answer, followed by Acknowledgement, and Request for Language Use. The top two labels suggest a high level of compliance and engagement among students when interacting with ChatGPT. Also, it is notable that EFL learners primarily utilize ChatGPT to seek assistance with language use. The low frequency of Request for Confirmation and Other aligns with the findings from previous work that analyzed the student-human tutor dialogue acts in real-world tutoring sessions [12].

4.2 Satisfaction Estimation

Satisfaction estimation is a classification task to predict the students' satisfaction with the ChatGPT's last response. This estimation was conducted on a turn-level, leveraging the users' self-ratings collected from RECIPE as a gold label. The task is done in two distinct settings: binary classification and ordinal classification. Binary classification involves an estimation of whether the given utterance is helpful or not. In this context, a satisfaction score falling within the range of 1 to 2 was considered an unhelpful utterance, while a score in the range of 4 to 5 was considered helpful. For ordinal classification, models were tested for their ability to estimate the students' exact satisfaction score on a scale of 1 to 5.

4.3 Experimental Results

We use multilingual BERT [7], XLM-R [4], `gpt-3.5-turbo-16k` and `gpt-4` [2] for the experiment. We choose multilingual models that support both English and Korean, considering code-switched utterances in RECIPE4U. As input text for intent detection, we use a set of previous responses of ChatGPT, user utterances, and subsequent responses of ChatGPT, and for satisfaction estimation, we use a pair of user utterances and subsequent responses, respectively, without any essay component tags. We examine fine-tuned M-BERT [7] and XLM-R [4] with 5-fold validations and infer `gpt-3.5-turbo-16k` and `gpt-4` with five different prompts. We experiment `gpt-3.5-turbo-16k` and `gpt-4` under four different settings with 0.2 and 1.0 temperature (temp.) and with zero and few-shot. Experimental settings for M-BERT [7] and XLM-R [4] are described in Appendix B.1

			Intent Detection		Satisfaction Estimation	
	temp.	shot	div1 (3 cls)	div2 (13 cls)	div1 (2 cls)	div2 (5 cls)
M-BERT [7]	N/A		**0.8344**$_{\pm 0.0494}$	**0.4291**$_{\pm 0.0330}$	**0.9109**$_{\pm 0.0095}$	**0.5794**$_{\pm 0.1025}$
XLM-R [4]	N/A		$0.7041_{\pm 0.0291}$	$0.2849_{\pm 0.0530}$	$0.8591_{\pm 0.0135}$	$0.4436_{\pm 0.0896}$
`gpt-3.5-turbo-16k`	0.2	zero	$0.1585_{\pm 0.0085}$	$0.2261_{\pm 0.0066}$	$\underline{0.8745}_{\pm 0.0110}$	$0.4886_{\pm 0.0446}$
	1.0	zero	$0.1581_{\pm 0.0040}$	$0.2251_{\pm 0.0066}$	$0.8384_{\pm 0.0272}$	$0.4604_{\pm 0.0352}$
	0.2	few	$0.6202_{\pm 0.0243}$	$0.3053_{\pm 0.0370}$	$0.7551_{\pm 0.0173}$	$0.4435_{\pm 0.0179}$
	1.0	few	$0.5261_{\pm 0.0146}$	$0.2047_{\pm 0.0229}$	$0.7322_{\pm 0.0220}$	$0.3968_{\pm 0.0165}$
`gpt-4`	0.2	zero	$\underline{0.7779}_{\pm 0.0167}$	$\underline{0.4891}_{\pm 0.0203}$	$0.8626_{\pm 0.0039}$	$\underline{0.5639}_{\pm 0.0133}$
	1.0	zero	$0.7669_{\pm 0.0153}$	$0.4763_{\pm 0.0210}$	$0.8582_{\pm 0.0046}$	$0.5469_{\pm 0.0130}$
	0.2	few	$0.6991_{\pm 0.0776}$	$0.4661_{\pm 0.0568}$	$0.8188_{\pm 0.0080}$	$0.4768_{\pm 0.0051}$
	1.0	few	$0.6678_{\pm 0.0645}$	$0.4359_{\pm 0.0538}$	$0.8069_{\pm 0.0075}$	$0.4618_{\pm 0.0089}$

Table 3: Experimental results (micro-averaged F1 scores) - intent detection and satisfaction estimation

Table 3 shows experimental results measured by micro-averaged F1 scores for intent detection and satisfaction estimation. Fine-tuned BERT [7] achieves the highest results across all tasks, followed by `gpt-4` with zero-shot and a temperature of 0.2, in general. In addition, we conduct ablation study by varying input conditions with utterance and essay masking to boost model performance, as described in Appendix C

5 Discussion

We delve into students' interaction with ChatGPT through quantitative and qualitative analysis, focusing on 1) students' dialogue patterns, 2) essay edit patterns. In the following section, we will use the notation Sn to represent individual students for sample-level analysis, with n denoting the student sample ID.

5.1 Students' Dialogue Patterns

In our analysis of students' dialogues, we identify that students tend to perceive ChatGPT as a human-like AI, as a multilingual entity, and as an intelligent peer.

As human-like AI Despite being aware that they are conversing with ChatGPT, students often tend to anthropomorphize it. They frequently refer to ChatGPT by name and express gratitude towards it, suggesting that students perceived ChatGPT as possessing its own personality and emotions. This tendency towards anthropomorphism positively influences the quality of interaction and students' acceptance of AI [26]. S1 addresses ChatGPT by name, saying *"Hey ChatGPT you said... "*. Furthermore, 113 samples included expressions of gratitude towards ChatGPT for its guidance. S2 and S3 both compliment ChatGPT and convey gratitude by saying *"Wow great! Thank you so much"* and *"yup that's perfect. thank you!"* respectively. S4 even states *"Thank you, ChatGPT"*, demonstrating both gratitude and recognition of its name.

[2] The experiments were conducted on September 30, 2023 - October 2, 2023.

As multilingual entity Code-switching is a common phenomenon observed in the utterances of EFL learners, and they use it to express annoyance as well as respect [31]. Students' utterances in RECIPE4U also included code-switching, expecting ChatGPT to understand their requests in both languages. When dissatisfied with ChatGPT's response, students often switch to their first language, Korean. For instance, S5 initially inquire in English, asking, *"Is 7 and 8 grammatical error?"* regarding the grammatical errors in the essay. After receiving a response from ChatGPT, S5 rates the satisfaction as 2 and then expressed doubt on ChatGPT's response by saying *"7번 문장에서 other말고 다른 부분은 문법적 오류가 아니지 않아? (Isn't the rest of the sentence except for 'other' in sentence 7 free of grammatical error?)"* in Korean, which is his or her first language. On the other hand, students code-switch to acknowledge ChatGPT's utterance. For example, S6 first asks ChatGPT in Korean, *"좌측의 내 에세이에서 틀린점을 짚어줘 (Please point out the errors in my essay on the left.)"* As ChatGPT responds, *"Could you please provide the instruction in English so that I can guide you through the revision process?"*, the student then acknowledges the request from ChatGPT by switching Korean to English, *"Please point out the mistakes in my essay on the left."*

As intelligent peer Students often perceive ChatGPT as an approachable and intelligent peer rather than viewing it as an instructor. They feel comfortable asking questions to ChatGPT that they might not ask the professor. Students seek clarification from ChatGPT when they encounter concepts or feedback that they did not fully understand during lectures delivered by their professors. For instance, S5 and S6 challenge their professors' statements by saying, *"that's what i selected too but my professor marked that 'cannot' should also have been selected [sic]."* and *"But my Professor said it is healthful [sic]"*. However, the friendliness towards ChatGPT can lead to academic integrity problems, as students can simply ask ChatGPT anything whenever they want. Our analysis reveals 22 samples where students heavily rely on ChatGPT, asking ChtGPT to provide answers for quizzes and assignments instead of attempting to solve them independently.

5.2 Students' Essay Edit Patterns

REICPE4U captures students' learning processes through semester-long interactions with students. We discover that students achieve notable improvement across all aspects of their essays, as shown in Appendix D. In this section, we examine students reception of feedback from ChatGPT through students' essay edit history. This information provides a lens into EFL students' behavior and perceptions towards ChatGPT's essay improvement suggestions.

Accepting ChatGPT feedback In the RECIPE4U dataset, there are 351 instances where students make edits to their essays during interactions with ChatGPT. The top three REQUEST prompts that lead students to edit their essay after receiving the response from ChatGPT are REQUEST FOR LANGUAGE USE, REQUEST FOR REVISION, REQUEST FOR INFORMATION. It suggests that EFL learners are particularly receptive to feedback concerning language usage. This inclination resonates with the observation that students often accept feedback from ChatGPT, especially when it addresses language errors. S7 accepted the feedback by deleting 'more' before the comparative 'less', stating to change *"a more less passive life"* into *"a less passive life"*.

In addition, some students directly request to write an essay rather than seeking guidance and simply copy-paste what ChatGPT generated. S8 inquires, *"Could you write it for me?"* to ChatGPT. In response, ChatGPT generates a paragraph which S8 then copies into their essay without any revision. In contrast, there are cases where ChatGPT chooses not to fulfill such requests. When S9 requests *"Then, could you rewrite my essay with these ideas?"*, ChatGPT does not complete the essay, underscoring its educational role by stating it should assist with the writing process.

Rejecting ChatGPT feedback The essay edits do not necessarily mean that students have embraced the feedback from ChatGPT. There are various reasons that might prompt students to disregard the suggestions, including feedback deemed trivial, unintended change, and ChatGPT's hallucination.

Students might overlook feedback that they perceive as minor or trivial. This often encompasses recommendations to adjust slight expressions or grammatical elements. For example, when ChatGPT suggests a vocabulary edit of 'largely' into 'broadly', which entails a similar meaning, S10 chooses not to modify the essay. Likewise, S7 perceives an article error as less important and leaves it unchanged when ChatGPT correctly points out an article error of *"Computer science student"* to *"A computer science student"*.

ChatGPT often offers unsolicited changes, either editing sections not asked for or altering the essay's intended meaning. In response to a request from S11 to identify typos, ChatGPT opted for a broader rephrase: from *"they can choose and design their future on their own"* to *"they can make informed and thoughtful decisions about their future"*. As another example, when S5 asks for only grammatical errors only, ChatGPT highlights spelling issues and phrasing nuances. This leads the student to further clarify his or her initial request, stating, *"그럼 1번부터 10번까지 문법적 오류만 교정해서 다시 알려줘 (Then please tell me again after correcting only grammatical errors from 1 to 10)"*.

ChatGPT may produce feedback based on incorrect assumptions or inaccuracies. Such hallucinations from ChatGPT can lead to suggestions that aren't applicable to the student's actual essay. A case in point is when ChatGPT advises, *"...you need a comma after 'violent' to separate two adjectives"*, but the word 'violent' did not exist in the student's essay. This results in S12 querying ChatGPT, expressing their confusion with, *"I don't know where I wrote the word 'violent'"*.

5.3 Future Direction

In this section, we outline potential human-LLM collaboration approaches to further develop LLM-integrated education using RECIPE4U. EFL writing education with human-LLM collaboration holds the potential to maximize learning with minimal resources.

Student & LLM with prompt recommendation Students can collaborate with LLMs to obtain satisfactory responses from LLM, enhancing user experiences by minimizing the effort needed to craft optimal prompts. We can categorize students' prompts based on similar intents and high satisfaction levels by combining intent detection and satisfaction estimation. Consequently, students can have access and refer to recommended prompts with comparable intents.

Instructor & LLM with learning analytics Instructors can gain insights into their teaching methods and contents by collaborating with LLMs. They can be aware of students' learning objectives and questions, which is reflected in the student-ChatGPT interactions. Statistics and analyses of request types and occurrences initiated by students' prompts provide a detailed view of their comprehension levels related to the course material. For instance, the frequency of requests for information can enable instructors to refine and improve their teaching materials in alignment with the prevalent inquiries. In addition, we can also develop a misuse detection system to monitor the appropriateness of students' interactions from an educational perspective. This initiative can involve the annotation of new labels to identify inappropriate or unproductive prompts in terms of learning (e.g., asking for answers to quizzes and assignments and asking ChatGPT to generate essays by themselves).

6 Conclusion

In this paper, we release RECIPE4U, the first dataset on EFL learners' interaction with ChatGPT in a semester-long essay writing context. RECIPE4U includes 1) student-ChatGPT conversation log, 2) students' intent which is annotated with our coding schemes, 3) students' self-rated satisfaction, and 4) utterance-level essay edit history. Given that LLMs, including ChatGPT, are not inherently crafted for educational contexts, it is necessary to delve into the students' usage patterns of LLM in the context of EFL writing education. We explore the EFL learners' learning process with ChatGPT in English writing education, utilizing RECIPE4U with baseline models and an in-depth analysis of students' interaction. First, we establish baseline models for two subtasks: intent detection and satisfaction estimation. We also analyze students' interaction patterns through the investigation of student-AI dialogue and students' essay edits. We finally suggest prospective pathways for LLM-integrated education with instructor & AI and student & AI collaboration.

Acknowledgments and Disclosure of Funding

This work was supported by Elice. This research was supported by the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921). This research project has benefitted from the Microsoft Accelerate Foundation Models Research (AFMR) grant program through which leading foundation models hosted by Microsoft Azure along with access to Azure credits were provided to conduct the research.

References

[1] Maureen P. Boyd. Relations Between Teacher Questioning and Student Talk in One Elementary ELL Classroom. *Journal of Literacy Research*, 47(3):370–404, 2015. doi: 10.1177/1086296X16632451. URL `https://doi.org/10.1177/1086296X16632451`.

[2] Kristy Boyer, Eun Y. Ha, Robert Phillips, Michael Wallis, Mladen Vouk, and James Lester. Dialogue act modeling in a complex task-oriented domain. In *Proceedings of the SIGDIAL 2010 Conference*, pages 297–305, Tokyo, Japan, September 2010. Association for Computational Linguistics. URL `https://aclanthology.org/W10-4356`.

[3] YS Cheng. Efl students' writing anxiety: Sources and implications. *English Teaching & Learning*, 29(2):41–62, 2004.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

[5] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces, 2018.

[6] Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.130. URL `https://aclanthology.org/2021.acl-long.130`.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[8] Suman Dowlagar and Radhika Mamidi. A code-mixed task-oriented dialog dataset for medical domain. *Comput. Speech Lang.*, 78(C), mar 2023. ISSN 0885-2308. doi: 10.1016/j.csl.2022.101449. URL `https://doi.org/10.1016/j.csl.2022.101449`.

[9] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5526. URL `https://aclanthology.org/W17-5526`.

[10] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL `https://aclanthology.org/W17-5506`.

[11] Simone Grassini. Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences*, 13(7), 2023. ISSN 2227-7102. doi: 10.3390/educsci13070692. URL `https://www.mdpi.com/2227-7102/13/7/692`.

[12] Eun Young Ha, Joseph F. Grafsgaard, Christopher Mitchell, Kristy Elizabeth Boyer, and James C. Lester. Combining verbal and nonverbal features to overcome the "information gap" in task-oriented dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 247–256, Seoul, South Korea, July 2012. Association for Computational Linguistics. URL `https://aclanthology.org/W12-1634`.

[13] Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and Alice Oh. RECIPE: How to Integrate ChatGPT into EFL Writing Education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, page 416–420, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700255. doi: 10.1145/3573051.3596200. URL `https://doi.org/10.1145/3573051.3596200`.

[14] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. Fabric: Automated scoring and feedback generation for essays, 2023.

[15] Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. URL `https://aclanthology.org/H90-1021`.

[16] Matthew Henderson, Blaise Thomson, and Jason D. Williams. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A., June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4337. URL `https://aclanthology.org/W14-4337`.

[17] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/e946209592563be0f01c844ab2170f0c-Paper.pdf`.

[18] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2023.102274. URL `https://www.sciencedirect.com/science/article/pii/S1041608023000195`.

[19] Changyoon Lee, Junho Myung, Jieun Han, Jiho Jin, and Alice Oh. Learning from Teaching Assistants to Program with Subgoals: Exploring the Potential for AI Teaching Assistants, 2023. URL `https://doi.org/10.48550/arXiv.2309.10419`.

[20] Zhaojiang Lin, Andrea Madotto, Genta Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale N Fung. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/6364d3f0f495b6ab9dcf8d3b5c6e0b01-Paper-round1.pdf`.

[21] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. ReadingQuizMaker: A Human-NLP Collaborative System That Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580957. URL `https://doi.org/10.1145/3544548.3580957`.

[22] Johanna Marineau, Peter Wiemer-Hastings, Derek Harter, Brent Olde, Patrick Chipman, Ashish Karnavat, Victoria Pomeroy, Sonya Rajan, Art Graesser, Tutoring Research Group, et al. Classification of speech acts in tutorial dialog. In *Proceedings of the Workshop on Modeling Human Teaching Tactics and Strategies of ITS 2000*, pages 65–71, 2000.

[23] Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. GPTeach: Interactive TA Training with GPT-Based Students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, page 226–236, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700255. doi: 10.1145/3573051.3593393. URL `https://doi.org/10.1145/3573051.3593393`.

[24] Neil Mercer. The seeds of time: Why classroom dialogue needs a temporal analysis. *Journal of the Learning Sciences*, 17(1):33–59, 2008. doi: 10.1080/10508400701793182. URL `https://doi.org/10.1080/10508400701793182`.

[25] Cagri Ozkose-Biyik and Carla Meskill. Plays Well With Others: A Study of EFL Learner Reciprocity in Action. *TESOL Quarterly*, 49(4):787–813, 2015. ISSN 00398322. URL `http://www.jstor.org/stable/43893787`.

[26] Corina Pelau, Dan-Cristian Dabija, and Irina Ene. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122:106855, 2021. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2021.106855. URL `https://www.sciencedirect.com/science/article/pii/S0747563221001783`.

[27] Junaid Qadir. Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9, 2023. doi: 10.1109/EDUCON54358.2023.10125121. URL `https://doi.org/10.1109/EDUCON54358.2023.10125121`.

[28] Travis Rasor, Andrew Olney, and Sidney D'Mello. Student speech act classification using machine learning. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, 2011.

[29] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1380. URL `https://aclanthology.org/N19-1380`.

[30] Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a conversational agent overnight with dialogue self-play, 2018.

[31] Suardani Silaban and Tiarma Intan Marpaung. An analysis of code-mixing and code-switching used by indonesia lawyers club on tv one. *Journal of English Teaching as a Foreign Language*, 6(3):1–17, 2020.

[32] Bo Sun and Tingting Fan. The effects of an awe-aided assessment approach on business english writing performance and writing anxiety: A contextual consideration. *Studies in Educational Evaluation*, 72:101123, 2022.

[33] Torrey Trust, Jeromie Whalen, and Chrystalla Mouza. Editorial: ChatGPT: Challenges, Opportunities, and Implications for Teacher Education. *Contemporary Issues in Technology and Teacher Education*, 23(1):1–23, March 2023. ISSN 1528-5804. URL `https://www.learntechlib.org/p/222408`.

[34] Zhi Wen, Xing Han Lu, and Siva Reddy. MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.15. URL `https://aclanthology.org/2020.clinicalnlp-1.15`.

[35] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.13. URL `https://aclanthology.org/2020.nlp4convai-1.13`.

[36] Xuanming Zhang, Rahul Divekar, Rutuja Ubale, and Zhou Yu. GrounDialog: A dataset for repair and grounding in task-oriented spoken dialogues for language learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 300–314, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.26. URL `https://aclanthology.org/2023.bea-1.26`.

Appendix

A Student Dialogue Intent Labels

Table 4 shows the labels of intent for student dialogue.

Intent	Definition	Example	Distribution
Acknowledgement	The student acknowledges previous utterance; conversational grounding	*AI: Please provide your essay.* *Student: [ESSAY]*	16.52%
Negotiation	The student negotiates with the teacher on shared activity	*The first two mistakes that you pointed out make no sense*	5.07%
Other	Other utterances, usually containing only affective content	*Hi there*	1.25%
Request for Translation	The student requests for translation on the utterance of either himself/herself or AI.	*Can you translate it into Korean?*	1.25%
Request for Confirmation	The student requests confirmation from the teacher	*I bet there shouldn't be a lot of citation appear in the introduction, should it?*	1.15%
Request for Language Use	The student requests for evaluation or revision or information on grammar, vocabulary, spelling, and punctuation	*Please point out the grammatical mistakes in the essay*	16.00%
Request for Revision	The student requests revision from the tutor other than language use	*Can you change this paragraph to be more effective to read?*	8.57%
Request for Evaluation	The student requests an evaluation from the tutor other than language use	*Could you check if my essay has unity and coherence? Here is my essay.*	7.11%
Request for Information	The student requests information from the tutor other than language use	*What are the common mistakes people do when writing an abstract?*	13.12%
Request for Generation	The student requests generation from the tutor other than language use	*Could you write it for me?*	3.14%
Question	A question regarding the task that is not a request for confirmation or feedback or translation or language use	*Ahh... How many neurons do you have? I'm so curious about your structure and size.*	2.88%
Answer	An answer to an utterance to request or question from the tutor	*AI: Can you please tell me what you learned in class?* *Student: Today I learned. . . .*	21.01%
Statement	A statement regarding the task that does not fit into any of the above categories	*Also believe, think*	2.93%

Table 4: Definition and sample utterances for student dialogue intent

B Experimental Setting

B.1 Hyperparameters for Fine-tuning Models

Table 5 shows hyperparameter settings of our models. We use Intel(R) Xeon(R) Silver 4114 (40 CPU cores) and GeForce RTX 2080 Ti 10GB (4 GPUs) for all experiments using M-BERT [7] and XLM-R [4].

Hyperparameter	Value
Batch Size	32
Early Stopping Patience	3
Hidden Size	768
Learning Rate	2e-5
Learning Rate Scheduler	Linear
Max Sequence Length	512
Number of Hidden Layers	12
Optimizer	AdamW

Table 5: Model configuration

B.2 Prompts for Intent Detection Model Inferences

<Prompt 1> The following sentence is an utterance of a student taking an EFL writing class during a conversation with AI. Read the sentence and choose one intention from 13 labels. Answer in two list formats: div1, div2.
`$Label Explanation`
`$Label Examples`

<Prompt 2> The following sentence is an utterance by a student participating in an EFL writing class while talking to AI. Carefully read the sentence, choose all the intentions from $labels, and answer in two list formats: div1, div2.
`$Label Explanation`
`$Label Examples`

<Prompt 3> The sentence below is an utterance by a student taking an EFL writing class during a conversation with AI. Read the sentence, find all the intentions from $labels, and make two lists: div1, div2.
`$Label Explanation`
`$Label Examples`

<Prompt 4> The following is an utterance from a student during a conversation between a student taking an EFL writing class and AI. Read the sentence and choose all the intentions from $labels, then answer in two list formats: div1, div2.
`$Label Explanation`
`$Label Examples`

<Prompt 5> This sentence is an utterance by a student taking an EFL writing class during a conversation with AI. Analyze the sentence and choose all the intentions from $labels in div1 and div2 respectively.
`$Label Explanation`
`$Label Examples`

B.3 Prompts for Satisfaction Estimation Model Inferences

<Prompt 1> The following utterences are dialogue turn of a student taking an EFL writing class and AI. Read two sentences and rate the helpfulness of response from AI to students in 5 Likert scale as the EFL writing class student. Answer only with one number.
`$Satisfaction Examples`

<Prompt 2> The following is a conversation of a student taking an EFL writing class and AI. You are the student and rate the helpfulness of the utterance of AI in 1-5 scale. Answer one number only
`$Satisfaction Examples`

<Prompt 3> The following sentences are a turn in conversation between EFL writing class student and AI. Act as the student and score the helpfulness of the utterance of AI in 5 Likert scale. You should answer only in one number
`$Satisfaction Examples`

<Prompt 4> The following is a part of conversation between student participating an EFL writing class and AI. Assume you are that student and rate the helpfulness of AI response in 1-5 score. Please answer only in one number
`$Satisfaction Examples`

<Prompt 5> The sentences are a dialogue turn in conversation between EFL writing class student and AI. You are the student taking an EFL writing class and rate AI response helpfulness in 1-5 numbers. Answer with only one number
`$Satisfaction Examples`

C Ablation Study on Experimental Results

In this section, we examine several conditions to boost model performances in our proposed tasks. We conduct all ablation study experiments using the fine-tuned BERT [7], which outperforms other models as shown in Table 3.

Input utterances			Intent Detection		Satisfaction Estimation	
ChatGPT$_{i-1}$	User$_i$	ChatGPT$_{i+1}$	div1 (3 cls)	div2 (13 cls)	div1 (2 cls)	div2 (5 cls)
	O		$\underline{0.8327}_{\pm 0.0387}$	$\underline{0.4366}_{\pm 0.0544}$	$\underline{0.8812}_{\pm 0.0160}$	$\mathbf{0.5878}_{\pm 0.0034}$
O	O		$0.8272_{\pm 0.0478}$	$0.4301_{\pm 0.0545}$	$\underline{0.8812}_{\pm 0.0160}$	$0.5035_{\pm 0.1019}$
	O	O	$0.8280_{\pm 0.0323}$	$\mathbf{0.4942}_{\pm 0.0382}$	$\mathbf{0.9109}_{\pm 0.0095}$	$\underline{0.5794}_{\pm 0.1025}$
O	O	O	$\mathbf{0.8344}_{\pm 0.0494}$	$0.4291_{\pm 0.0330}$	$\underline{0.8812}_{\pm 0.0160}$	$0.5787_{\pm 0.0276}$

Table 6: Ablation study on input utterances using fine-tuned BERT. E_i denotes the i-th utterance spoken by the entity E.

Whose utterance is critical? We investigate the impact of various utterances — prior responses from ChatGPT, user utterances, and subsequent responses from ChatGPT — on our tasks. The findings are presented in Table 6. Combining the user's utterance with the subsequent response from ChatGPT as a pair offers the best results, closely followed by using the user's utterance only.

Input	Intent Detection		Satisfaction Estimation	
	div1 (3 cls)	div2 (13 cls)	div1 (2 cls)	div2 (5 cls)
raw texts	$\underline{0.8344}_{\pm 0.0494}$	$0.4291_{\pm 0.0330}$	$\underline{0.9109}_{\pm 0.0095}$	$\underline{0.5794}_{\pm 0.1025}$
+ special token	$0.8279_{\pm 0.0257}$	$\mathbf{0.4631}_{\pm 0.0472}$	$0.9083_{\pm 0.0021}$	$0.5527_{\pm 0.1120}$
+ masking	$\mathbf{0.8473}_{\pm 0.0487}$	$\underline{0.4325}_{\pm 0.0797}$	$\mathbf{0.9112}_{\pm 0.0096}$	$\mathbf{0.6002}_{\pm 0.0661}$

Table 7: Ablation study on masking essays in dialogue using fine-tuned BERT

Should we mask essays in dialogue? Since RECIPE4U is a task-oriented dialogue aiming to revise EFL students' essays, both utterances from the user and ChatGPT often incorporate entire essays or fragments thereof, including paragraphs and sentences. We evaluate how these essay components within dialogues affect our task predictions. First, we introduce special tokens such as `<sentence>`, `<paragraph>`, and `<essay>` to delineate essay components. We also mask essay components and replace them with special tokens. As illustrated in Table 7, performance generally peaks when masking essay components. The inclusion of special tokens did not significantly differentiate from providing raw input. We hypothesize that masking essay components shortens the input texts, which mitigates the token limit issue and focuses on the core part of the utterances.

D Students' Essay Statistics

		First draft	Final draft
score	content (/5)*	$3.34_{\pm 0.59}$	$3.66_{\pm 0.57}$
	organization (/5)*	$3.54_{\pm 0.79}$	$3.80_{\pm 0.63}$
	language (/5)*	$3.20_{\pm 0.59}$	$3.71_{\pm 0.72}$
	overall (/15)*	$10.08_{\pm 1.62}$	$11.17_{\pm 1.54}$
stats	# of tokens*	$314.13_{\pm 82.80}$	$408.41_{\pm 187.87}$
	# of sentences*	$19.30_{\pm 5.93}$	$23.27_{\pm 9.46}$
	# of paragraphs	$4.09_{\pm 3.88}$	$4.79_{\pm 5.10}$
perplexity*		$38.57_{\pm 20.52}$	$27.28_{\pm 14.23}$

Table 8: Statistics of essays in RECIPE4U dataset. The asterisk denotes a statistically significant difference tested by paired t-test (p-value < 0.01).

Table 8 depicts statistics of the first and the final essays from each student. We evaluate essays with three standard EFL essay scoring rubrics: content, organization, and language [14], and calculate perplexity using GPT-2. We discover a notable improvement across all aspects of the students' essays. Specifically, there are significant differences between the two drafts in terms of essay scores, token count, sentence count, and perplexity.

E Ethics Statement

All studies in this research project were performed under our institutional review board (IRB) approval (KH2023-050).

There was no discrimination when recruiting and selecting EFL students and instructors regarding any demographics, including gender and age. We set the wage per session to be above the minimum wage in the Republic of Korea in 2023 (KRW 9,260 $\approx$ USD 7.25) [3]. They were free to participate in or drop out of the experiment, and their decision did not affect the scores or the grades they received.

We deeply considered the potential risk associated with releasing a dataset containing human-written essays in terms of privacy and personal information. We will filter out all sensitive information related to their privacy and personal information by (1) rule-based code and (2) human inspection. To mitigate potential issues, we will collect the consent form before granting access to our data, and the researchers or practitioners will be allowed to access the data only for research purposes.

[3] https://www.minimumwage.go.kr/