

CIKMar: A Dual-Encoder Approach to Prompt-Based Reranking in Educational Dialogue Systems

Joanito Agili Lopo¹, Marina Indah Prasasti¹, Alma Permatasari¹, Yunita Sari²

Department of Computer Science and Electronics
Universitas Gadjah Mada

¹{joanitoagililopo, marinaindahprasasti, almapermatasari}@mail.ugm.ac.id

²yunita.sari@ugm.ac.id

Abstract

In this study, we introduce CIKMar¹, an efficient approach to educational dialogue systems powered by the Gemma Language model. By leveraging a Dual-Encoder ranking system that incorporates both BERT and SBERT model, we have designed CIKMar to deliver highly relevant and accurate responses, even with the constraints of a smaller language model size. Our evaluation reveals that CIKMar achieves a robust recall and F1-score of 0.70 using BERTScore metrics. However, we have identified a significant challenge: the Dual-Encoder tends to prioritize theoretical responses over practical ones. These findings underscore the potential of compact and efficient models like Gemma in democratizing access to advanced educational AI systems, ensuring effective and contextually appropriate responses.

1 Introduction

The emergence of powerful Large Language Models (LLMs) such as ChatGPT has been proven effective in various tasks, including generating text that is nearly indistinguishable from human-written text (Kasneji et al., 2023; Omidvar and An, 2023). Building on the success in text generation, LLMs have shown significant potential in various applications, especially in the educational domain.

In recent years, there have been various efforts to utilize these powerful LLMs in education. They have been deployed in teacher-student collaborations as virtual tutors, guiding students through exercises, offering personalized learning experiences, and providing intelligent tutoring (Kamalov et al., 2023). Additionally, they are used for adaptive assessments and serve as conversational partners in learning scenarios (Tan et al., 2023; Li et al., 2024).

Despite these promising opportunities, the use of generative models as a foundation for downstream tasks presents several crucial challenges

¹<https://github.com/joanitolopo/cikmar-system>

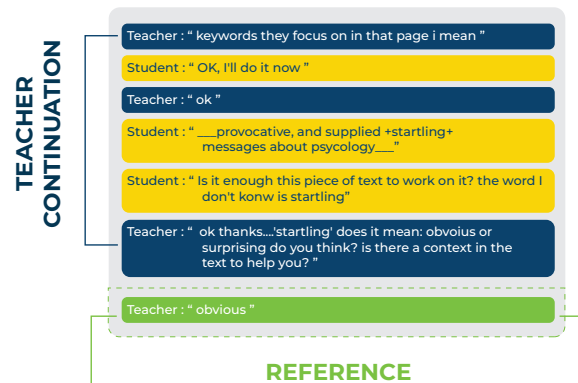


Figure 1: Teacher Continuation Data Visualization

such as inconsistently delivering accurate and contextually appropriate responses (Tack et al., 2023). Furthermore, language models in current scenarios mostly use extremely large models in terms of their parameter size, such as proprietary 175 and 137 billion-parameter GPT-3 model (Brown et al., 2020), or open source LLMs such as 70 billion-parameter LLaMA2 (Touvron et al., 2023), 14 billion-parameter Qwen (Bai et al., 2023), and 6 billion-parameter ChatGLM3 models (Zeng et al., 2023).

Language models at this scale are not practical and inaccessible for many researchers and even practitioners, due to their large memory consumption and slow generation times (Ding et al., 2024; Jimenez Gutierrez et al., 2022), data privacy, and inflexibility of customization (Sinha et al., 2024). Therefore, it is essential to determine how solid that foundation is and how it can be accessible for further development, especially in the educational domain.

According to the challenges above, we designed a simple but effective approach by leveraging Large Language Models and prompt-and-rerank approach (Suzgun et al., 2022) to build the dialogue AI system especially in educational domain. We chose to work with a smaller, pre-trained language model

called Gemma 1.1 2B (IT), which can run efficiently on less than 12 GB of RAM and a single GPU T4. This makes it suitable for real-world applications by maintaining a reasonable model size without compromising performance. Additionally, a Dual-Encoder approach strategy has been adopted to re-rank the candidate outputs generated by the model using hand-written prompts. This approach aims to increase the relevance and effectiveness of the responses generated by our system in educational dialogues.

2 Related Work

Researchers have extensively investigated the effectiveness of various approaches utilizing language models. [Sridhar et al. \(2023\)](#) enhanced LLM performance on web navigation tasks using Actor-Summarizer Hierarchical (ASH) prompting, while [Kong et al. \(2024\)](#) improved reasoning benchmarks with role-play prompting. [Kojima et al. \(2023\)](#) showed that modifying prompt structure enables LLMs to perform multi-step reasoning in zero-shot settings.

In the educational context, [Adigwe and Yuan \(2023\)](#) and [Hicke et al. \(2023\)](#) used GPT-3 and GPT-4 to generate educational dialogue responses, achieving high DialogRPT and BERTScore results with hand-written zero-shot prompts. Similarly, [Vasselli et al. \(2023\)](#) used GPT-3.5 Turbo with manual few-shot prompts based on DialogRPT selection, which contributed most to the final outputs.

Fine-tuning has also proven effective by utilizing large language models (LLMs) in educational domain. [Baladón et al. \(2023\)](#) used the LoRa method to fine-tune models like BLOOM-3B, Llama 7B ([Touvron et al., 2023](#)), and OPT 2.7B ([Zhang et al., 2022](#)). They found that even the smaller OPT 2.7B model, with careful fine-tuning, could achieve better performance. Similarly, [Huber et al. \(2022\)](#) demonstrated that GPT-2, enhanced with reinforcement learning via the NLPO algorithm ([Ramamurthy et al., 2023](#)), achieved high BERTScores.

Due to the high computational power needed for fine-tuning and domain adaptation, [Omidvar and An \(2023\)](#) introduced semantic in-context learning, using private knowledge sources for accurate answers. [Gu et al. \(2024\)](#) proposed reducing LLM sizes through knowledge distillation, training smaller models to replicate larger ones. Their experiments with distilled GPT-3 versions showed competitive performance on various benchmarks.

Our research aims to develop an educational dialogue system using Gemma 1.1 IT 2B. This system uses prompts to guide LLMs in generating outputs based on contextual understanding, relevance, engagement, clarity, and feedback. To optimize results, it employs dual encoders (BERT and SBERT) to rerank top candidates. Our objective is to democratize open model LLM in real-world scenarios, ensuring accurate, relevant responses while enhancing student engagement and understanding in educational dialogues.

3 Methods

3.1 Data

We used data from the BEA 2023 shared task, sourced from the Teacher-Student Chatroom Corpus (TSCC) ([Caines et al., 2020, 2022](#)). This corpus consists of several conversations where an English teacher interacts with a student to work on language exercises and assess the student’s English proficiency ([Tack et al., 2023](#)). Each conversation contains multiple responses and starts with either **teacher:** or **student:** prefixed. The reference text is the teacher’s response that follows the previous input dialogue. The corpus includes a training set of 2,747 conversations, a development set of 305 conversations, and a test set of 273 conversations, totaling 3,325 conversations.

Since the data were collected from real-time teacher-student conversations, turn-taking is not as consistent as in most dialogue systems ([Vasselli et al., 2023](#)). Two patterns mostly occur: conversations ending with the student (teacher reply) and conversations ending with the teacher (teacher continuation). This condition occurs in 38% of the training data and 40% of the development data. [Figure 1](#) shows an example of a conversation in the teacher continuation condition.

3.2 Prompt Ensemble

We utilized hand-written prompts from [Vasselli et al. \(2023\)](#) to build our system. The prompts include Zero-shot and Few-shot types, targeting both general and specific scenarios. We used only five main prompts available as they are already tailored for teacher responses and continuations. This selection also ensures general applicability to other datasets or conversations. Full details explanation of each prompt are described in the [Appendix A](#).

In the creation of the few-shot prompts, it requires positive and negative examples to help the

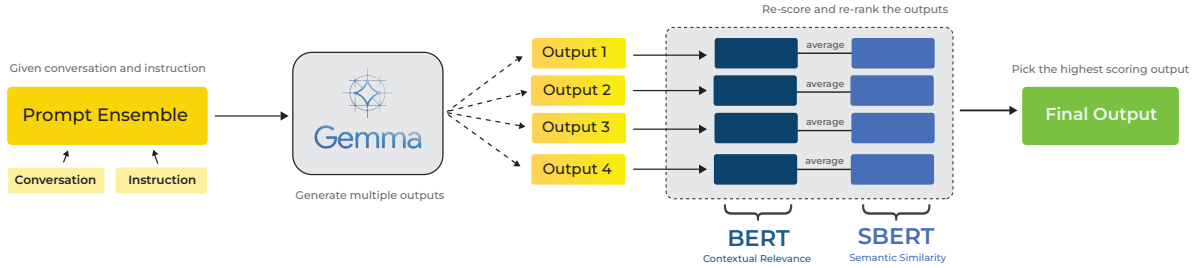


Figure 2: An illustration of the CIKMar system. Given an input conversation and instruction, we create the prompt ensemble and feed it to Gemma to generate multiple outputs. We then re-score each output by averaging BERT and SBERT scores and select the candidate with the highest re-ranked score as the final output

model avoid irrelevant responses. We adopted the method of Vasselli et al. (2023) who applied the handcrafted, generative, and iterative prompt methods. However, we modified the iterative method from the original paper. Instead of using DialogRPT, we employed the BM25 ranking function to select the highest and lowest scoring responses as positive and negative examples.

BM25 (Robertson and Walker, 1994; Robertson and Zaragoza, 2009) was chosen over DialogRPT because it reduces the computational power required for the prompting and re-ranking process, as DialogRPT needs additional memory capacity to calculate and choose the best candidate. Additionally, BM25 is known as the first-stage ranker in lexical retrieval systems (Askari et al., 2023) which ensures positive and negative examples are selected based on their lexical match with the conversation history.

3.3 Gemma Instruct-tuned Model

Our main system leverages a pretrained language model with a prompting approach rather than training one from scratch or fine-tuning it on a new dataset. We used the Gemma 1.1 IT 2B model (Team et al., 2024), 2-billion parameter open model developed by Google for efficient CPU and on-device applications. The model has shown strong performance across academic benchmarks for language understanding, reasoning, and safety, such as MMLU (Hendrycks et al., 2021), SIQA (Sap et al., 2019), HumanEval (Chen et al., 2021), and Winogrande (Sakaguchi et al., 2019). These results indicate its promising performance in educational contexts.

We followed the instruction-formatted control tokens suggested in the Gemma technical report to avoid out-of-distribution and poor generation. Table 1 shows an example dialogue with user and model control tokens. Specifically, the relevant

User:	<start_of_turn>user conversation instruction <end_of_turn> <start_of_turn>model
Model:	responses<end_of_turn>

Table 1: Example dialogue with user and model control tokens.

token user represents the role, and its content includes the conversation history followed by the prompt. Meanwhile, the model turn responds to the user dialogue.

In our experiments with the training and development sets, the Gemma model sometimes generated hallucinations in the first attempt, such as factually incorrect response, nonsensical content, overly long response, and content disconnected from the input prompt. However, performance improved on the second and third attempts. Therefore, to ensure the best response, we generated each candidate three times before selecting the final output.

We configured several parameters to control the model’s output such as set the max_length of the generated output to 512 tokens, no_repeat_ngram_size to 2 to avoid repetition, and used top_k=50 and top_p=0.95 to balance randomness and coherence. The temperature was set to 0.7 for more conservative choices. Finally, we enabled probabilistic sampling over greedy decoding.

3.4 Dual-Encoder Reranking

Inspired by previous research (Vasselli et al., 2023; Suzgun et al., 2022; Haroutunian et al., 2023), our system generates multiple candidate outputs from different manually designed prompts and then re-ranks these outputs using a heuristically defined scoring function. For the scoring function, we employed SBERT (Reimers and Gurevych, 2019)

and BERT (Devlin et al., 2019). Specifically, we used the paraphrase-MiniLM-L6-v2 version of SBERT, which maps sentences and paragraphs to a 384-dimensional dense vector space, and the bert-base-uncased model for BERT. We averaged the cosine similarity scores of their embeddings to evaluate the fine-grained semantic relevance and context-response matching in the embedding space between the conversation history and the generated responses.

In the given setup, we started with a dialog as a context ctx and a list of candidate responses $\{cand_1, cand_2, \dots, cand_m\}$. Initially, we computed SBERT and BERT embeddings for both the context and the candidate responses. For BERT embeddings we calculated by averaging token embeddings across the sequence dimension.

The cosine similarity between the context and each candidate response embedding, for both SBERT and BERT, are calculated using:

$$S_{emb}(i) = \cos(e_{ctx}^{emb}, e_{cand_i}^{emb}) = \frac{e_{ctx}^{emb} \cdot e_{cand_i}^{emb}}{\|e_{ctx}^{emb}\| \|e_{cand_i}^{emb}\|}$$

where $emb \in \{sbert, bert\}$.

To combine these similarity scores for each candidate response, we averaged the SBERT and BERT similarity scores.

Finally, the candidates are ranked based on these combined similarity scores in descending scores. The indices of the candidates are sorted according to their combined scores, and it returns the list of candidates responses ordered from most to least relevant to the given context.

3.5 Post-processing

The raw outputs from model often included inconsistent formatting, such as phrases prefixed by "***" or starting with unwanted text like Teacher: or Student:. Additionally, the model sometimes appended lengthy explanations to its responses beginning with Explanation:, adding unnecessary length. However, we observed a consistent pattern where the actual response always began with a quotation mark ".

To standardize these outputs, we implemented a post-processing step. First, we defined a regular expression pattern, `**\.*\?:**\n\n`, to identify and remove any unwanted initial phrases. This pattern effectively removed prefixes like "***", Teacher:, or Student:. Next, each response was processed to retain only the text following the

#	Precision	Recall	F1-Score
CIKMar (ours)	0.69	0.70	0.70
NAISTeacher Vasselli et al. (2023)	0.71	0.71	0.71
Adaio Adigwe and Yuan (2023)	0.72	0.69	0.71
GPT-4 Hicke et al. (2023)	0.71	0.69	0.70
S-ICL Omidvar and An (2023)	0.72	0.69	0.70
OPT-2.7B Baladón et al. (2023)	0.74	0.68	0.71
NLP-HSG Huber et al. (2022)	0.72	0.63	0.67
Alpaca Baladón et al. (2023)	0.72	0.68	0.70
DT Tack et al. (2023)	0.67	0.62	0.64

Table 2: Comparison of our proposed system with previous research based on BERTScore (Zhang et al., 2020)

first occurrence of a quotation mark, discarding any preamble or unnecessary content. Finally, we trimmed any leading or trailing whitespace.

4 Result & Analysis

4.1 Main Result

Our main result are presented in Table 2, showcasing comparisons among systems from the BEA Shared Task 2023 (Tack et al., 2023), ranked primarily by BERTScore. However, this comparison isn't fully comprehensive as the BEA Shared Task also considers human evaluations and DialogRPT (Gao et al., 2020) score. The human evaluation metric is restricted and not publicly available, and we encountered challenges with DialogRPT, which might have issues with the model, as it is return the same score for each context.

CIKMar demonstrates competitive performance against baseline systems like NAISTeacher and Adaio based on BERTScore². Specifically, we achieve a robust recall score of 0.70, slightly below NAISTeacher's 0.71. This indicates that our Dual-Encoder ranking effectively retrieves many contextually relevant responses compared to the reference answer. Furthermore, our F1-Score of 0.70 is comparable to models such as S-ICL and Alpaca, which utilize fine-tuning and larger model sizes, demonstrating our model's capability to capture similarity and produce coherent, contextually appropriate responses even using simple and small model size.

4.2 Evaluation Metrics

To ensure the reliability of our approach, we employed word overlap-based metric ROUGE

²BERTScore was calculated using Hugging Face's evaluate package with the distilbert-base-uncased model, averaging precision, recall, and F1 scores across the entire test set.

id	responses	analysis
test_0006	I understand you're feeling a bit chilly . Would you like to continue with the grammar practice or do you have any questions about the vocabulary word 'cold' ?	feeling, relevant, and emphatic
test_0004	Great point! Speed is an important aspect of your time management , so maintaining a sense of efficiency is crucial. How did you achieve your faster timeframe?	appreciation, discussion focus, and skill development
test_0007	Great point! Now, let's focus on using this new vocabulary in a meaningful way. Can you tell me some examples of how you've used this word in your own writing or conversations?	appreciation, practice of the new material, and reflection
test_0010	Following your insightful observation , could you provide an instance where speaking truthfully might be particularly effective in achieving a desired outcome?	appreciation, and critical thinking
test_0011	I appreciate your perspective on this, student. To further explore this idea , how does playing games help children cope with difficult emotions ?	relevant, emotion, motivation, critical skills and practical solutions

Table 3: Qualitative examples of model generated responses. We manually highlight in color, phrases that we find to be associated with the context and the critical aspects as mentioned in Tack and Piech (2022).

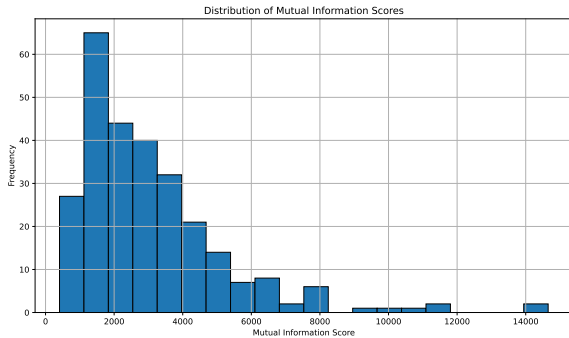


Figure 3: The distribution of mutual information scores derived from combined perplexity values

(Lin, 2004) and the neural network-based metric Dialog perplexity (Li et al., 2016)³ to further asses our system. We computed ROUGE metrics: rouge1, rouge2, rougeL, and rougeLsum resulting in scores of 0.12, 0.0047, 0.084, and 0.087, respectively.

Based on the ROUGE scores, the generated text demonstrates significant overlap with the reference text at the unigram level (ROUGE-1) and in longer common sequences (ROUGE-L and ROUGE-Lsum). This suggests the system stays on-topic and uses relevant vocabulary, beneficial for educational content. However, it shows noticeable shortcomings with exact word sequences (ROUGE-2), and discrepancies in longer sequences (ROUGE-L and ROUGE-Lsum) indicate challenges in maintaining coherence and well-structured responses.

Additionally, Figure 3 depicts the distribution of mutual information scores derived from combined perplexity values. The histogram's right-

³We used DialogGPT and its reverse model to compute perplexity

skewed shape, with scores predominantly in the lower range, suggests that the generated teacher responses are often predictable. While this indicates clarity, conciseness, and consistency in the generated text, which are advantageous for educational contexts, it also reveals a drawback: the responses lack depth and exhibit monotony, significantly reduce text engagement and the nuanced understanding required for deeper learning.

4.3 Output Analysis

We manually inspected the model's outputs and evaluated each prompt's contribution by examining 10 outputs in detail. Table 3 presents the top candidate responses selected through Dual-Encoder ranking for five examples.

To examine the impact of prompts on the best responses, we used the dialogue context test_0006, as shown in Table 4, as an example. Here, the teacher is explaining a grammar lesson when the student mentions needing 10 more minutes and feeling very cold in the room. The model's response is inconsistent, as it incorrectly associates "cold" with the grammar lesson rather than the student's condition. This suggests that the model may focus on one situation in the conversation and struggle to adapt when new contexts arise. Consequently, the context of "cold" is incorrectly forced to fit the context itself.

We also found that the model struggles with teacher continuation problems. When the dialogue ends with the teacher, the model often seems unsure about the next response, which happens frequently in the generated outputs. This aligns with research by Vasselli et al. (2023), indicating that

Embedding Visualization of Dialogues and Candidates

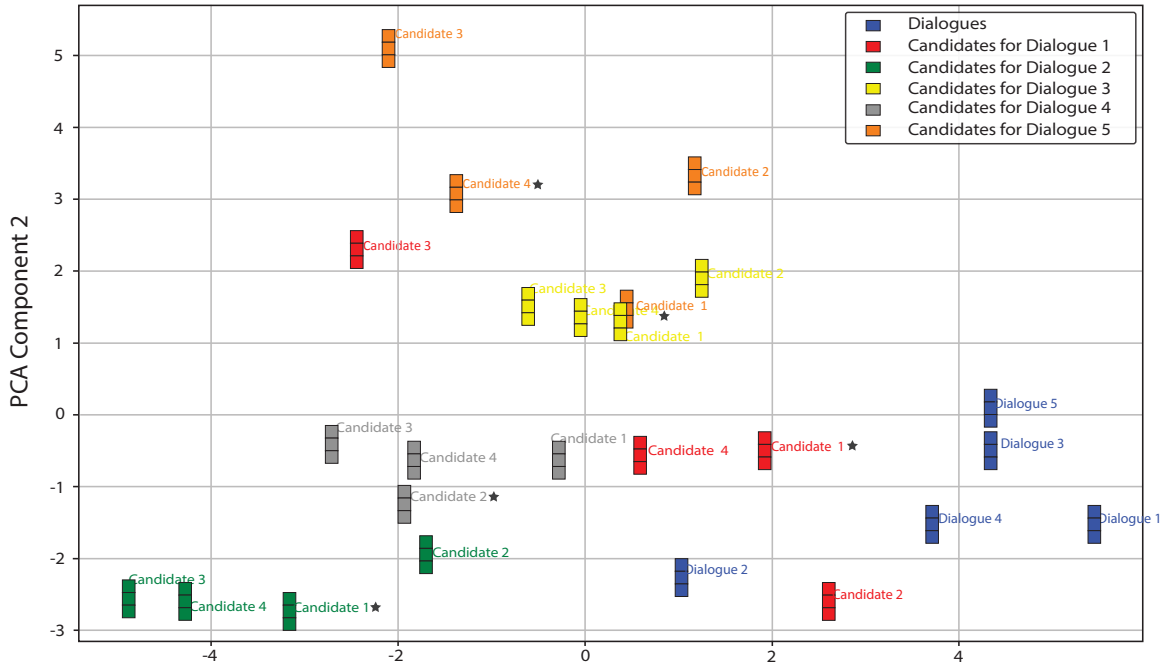


Figure 4: Embedding Space Visualization

Teacher: Which is easy, because you can use my / his / your etc. and not think about articles!
Student: Only 10 minutes left!
Teacher: I know, we can finish early if you are getting cold?
Student: I'm really cold

Table 4: Example dialogue context of test_006 between student and teacher

instruct-tuned models trained in user-assistant settings find it difficult to adapt when the setting changes abruptly. For example, in test_004, the model repeats the word "great" from the dialogue but fails to understand the context despite managing to introduce a follow-up conversation by asking the student examples.

Furthermore, we analyzed several dialogues with minimal context, some having only two exchanges. This limited context makes it difficult for the model to grasp the overall conversation and provides fewer reference words. A prime example is test_0011, which has only one turn with at least 5 words per turn. This lack of context makes it challenging for the model to generate the best response, as the context is insufficiently clear.

Lastly, we analyzed the contributions of each prompt to the final output selected by Dual-Encoder ranking for 10 data points. Prompt 1 significantly influenced the final output, being chosen in 5 examples. This is likely due to the model's strong

performance in academic tasks and the straightforward nature of these conversations, which aligned well with Prompt 1's instructions. In contrast, test_010 involved a complex, multi-turn conversation where Prompt 1 was not chosen because the teacher needed to explain the learning context in greater depth. As conversation complexity increases, the dual encoder selects Prompts 2 and 4, which are better suited to handle more intricate dialogues.

4.4 Word-Level Inspection

To explore the contextual relationships of the best candidates selected by the ranking function, we visualize the attention scores. We analyze the attention generated by the BERT model, as shown in Figure 5 for the example test_0007. The dialogue focuses on the teacher's general role, which, while informative, does not directly advance learning in the context of the vocabulary that the students have just learned or used.

The most effective response is: "Great point! Now, let's focus on using this new vocabulary in a meaningful way..." This response directly guides students to practice and apply the newly learned vocabulary in a more meaningful context, aligning more closely with the educational objectives of the dialogue.

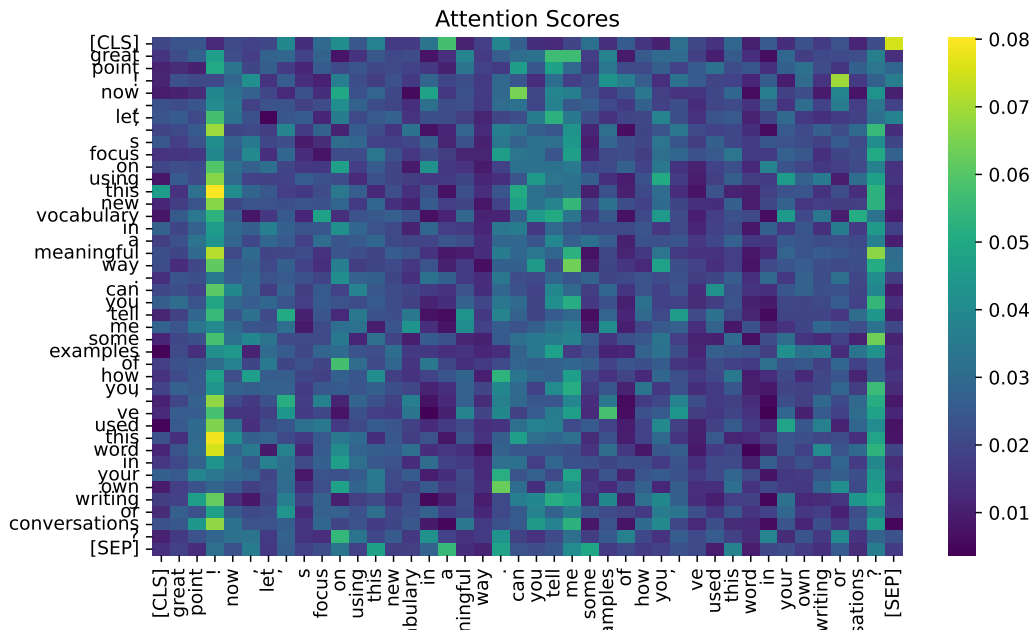


Figure 5: BERT Attention Score for example test_0007

According to the BERT Attention Score, attention is concentrated on key words in the dialogue, such as "great," "point," "focus," "using," "this," "meaningful," "way," and "vocabulary," which receive a high level of attention. These words are emphasized because they are directly connected to the higher learning goal of encouraging students to use new vocabulary in meaningful contexts. The teacher highlights the importance of guiding students not just to recognize new vocabulary but also to actively apply it in their writing or conversation. The most effective response aligns better with this learning objective, as it prompts students to consider how to use the vocabulary in practical ways. Ultimately, the attention score fosters a deeper understanding and retention of new vocabulary, contribute to the selection of the best response candidate in the ranking function.

4.5 Dual Encoder Effect

We conducted a manual investigation to assess the dual encoder's impact on selecting the best candidates. Analyzing five dialogue-response pairs' embedding spaces, as shown in Figure 4, we discovered that the Dual-Encoder can avoid the pitfalls of distance-measurement-only. Notably, in dialogue 2, candidate 2 appeared closer to its context than candidate 1 in the embedding space, yet the dual encoder ranked candidate 1 as the best candidate (denoted by *). This phenomenon occurred

across multiple dialogues, highlighting SBERT and BERT's role in enhancing the model's consideration of contextual relevance and semantic similarity between dialogues and responses, as discussed earlier.

To evaluate the dual encoder's ranking quality, we investigated the phenomenon of closely clustered embedding. Specifically, candidates for dialogue 3 exhibited dense clustering, where increasing embedding proximity indicated greater similarity, complicating candidate selection. After analyzing all candidates, candidates 1 and 4 emerged as optimal choices for this dialogue, supported by their relatedness in the embedding space. However, the Dual-Encoder prioritized candidate 4, suggesting a preference for theoretical discussion and exploration rather than practical context in its ranking criteria.

Finally, we noted a tendency for candidates within each dialogue to cluster together. This indicates that the Gemma model consistently produces similar embedding for each candidate per dialogue, demonstrating stable performance across various dialogues. However, certain candidates were positioned farther from their cluster and nearer to candidates in another cluster. This suggests that the model sometimes encounters difficulties accurately interpreting the dialogue context. We suspect that this issue may arise because SBERT dominance over BERT leads to a loss of full context. Further

investigation is required to delve deeper into this matter.

5 Conclusion & Future Work

We have shown that CIKMar, an educational dialogue generation approach using prompts and a Dual-Encoder ranking with the Gemma language model, yields promising results in educational settings. By utilizing the Gemma 2B model, we maintain high performance in response relevance and accuracy with a smaller, more accessible model.

Despite these strong performances, we have identified limitations hindering optimal results. Specifically, the Dual-Encoder often prioritizes theoretical discussion over practical contextual responses, potentially leading to irrelevant rankings. Future research should explore scenarios where either SBERT or BERT dominates ranking scores.

Additionally, crafting more specific prompts is crucial for deeper contextual understanding in educational dialogues. Lastly, refining the Gemma model to focus on educational contexts and adapt to shifting conversation dynamics is recommended.

Acknowledgments

This work was partially supported by the Department of Computer Science and Electronics, Universitas Gadjah Mada under the Publication Funding Year 2024.

References

Adaeze Adigwe and Zheng Yuan. 2023. [The ADAIO system at the BEA-2023 shared task: Shared task generating AI teacher responses in educational dialogues](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 796–804, Toronto, Canada. Association for Computational Linguistics.

Arian Askari, Amin Abolghasemi, Gabriella Pasi, Wessel Kraaij, and Suzan Verberne. 2023. [Injecting the bm25 score as text improves bert-based re-rankers](#).

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang

Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#).

Alexis Baladón, Ignacio Sastre, Luis Chiruzzo, and Aiala Rosá. 2023. [RETUYT-InCo at BEA 2023 shared task: Tuning open-source LLMs for generating teacher responses](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 756–765, Toronto, Canada. Association for Computational Linguistics.

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

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. [The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts](#). In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 23–35, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. [The teacher-student chatroom corpus](#). In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20, Gothenburg, Sweden. LiU Electronic Press.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [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. Association for Computational Linguistics.
- Tianyu Ding, Tianyi Chen, Haidong Zhu, Jiachen Jiang, Yiqi Zhong, Jinxin Zhou, Guangzhi Wang, Zhihui Zhu, Ilya Zharkov, and Luming Liang. 2024. [The efficiency spectrum of large language models: An algorithmic survey](#).
- Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. [Dialogue response ranking training with large-scale human feedback data](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. [Minillm: Knowledge distillation of large language models](#).
- Levon Haroutunian, Zhuang Li, Lucian Galescu, Philip Cohen, Raj Tumuluri, and Gholamreza Haffari. 2023. [Reranking for natural language generation from logical forms: A study based on large language models](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1067–1082, Nusa Dua, Bali. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#).
- Yann Hicke, Abhishek Masand, Wentao Guo, and Tushaar Gangavarapu. 2023. [Assessing the efficacy of large language models in generating accurate teacher responses](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 745–755, Toronto, Canada. Association for Computational Linguistics.
- Patrick Huber, Armen Aghajanyan, Barlas Oguz, Dmytro Okhonko, Scott Yih, Sonal Gupta, and Xilun Chen. 2022. [CCQA: A new web-scale question answering dataset for model pre-training](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2402–2420, Seattle, United States. Association for Computational Linguistics.
- Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. [Thinking about GPT-3 in-context learning for biomedical IE? think again](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Firuz Kamalov, David Santandreu Calong, and Ikhlās Gurrib. 2023. [New era of artificial intelligence in education: Towards a sustainable multifaceted revolution](#).
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchermann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, 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. 2023. [Chatgpt for good? on opportunities and challenges of large language models for education](#). *Learning and Individual Differences*, 103:102274.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. [Large language models are zero-shot reasoners](#).
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. [Better zero-shot reasoning with role-play prompting](#).
- Jiwei Li, Michel Galley, Chris Bockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. 2024. [Adapting large language models for education: Foundational capabilities, potentials, and challenges](#).
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Amin Omidvar and Aijun An. 2023. [Empowering conversational agents using semantic in-context learning](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 766–771, Toronto, Canada. Association for Computational Linguistics.
- Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. [Is reinforcement learning \(not\) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization](#).
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

- and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, London. Springer London.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavathula, and Yejin Choi. 2019. [Winogrande: An adversarial winograd schema challenge at scale](#).
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social IQa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Neelabh Sinha, Vinija Jain, and Aman Chadha. 2024. [Evaluating open language models across task types, application domains, and reasoning types: An in-depth experimental analysis](#).
- Abishek Sridhar, Robert Lo, Frank F. Xu, Hao Zhu, and Shuyan Zhou. 2023. [Hierarchical prompting assists large language model on web navigation](#).
- Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. [Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Anais Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. [The BEA 2023 shared task on generating AI teacher responses in educational dialogues](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.
- Anais Tack and Chris Piech. 2022. [The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues](#).
- Kehui Tan, Tianqi Pang, Chenyou Fan, and Song Yu. 2023. [Towards applying powerful large ai models in classroom teaching: Opportunities, challenges and prospects](#).
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepey, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharmar, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. [Gemma: Open models based on gemini research and technology](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Justin Vasselli, Christopher Vasselli, Adam Nohejl, and Taro Watanabe. 2023. [NAISTeacher: A prompt and](#)

rerank approach to generating teacher utterances in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 772–784, Toronto, Canada. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. *Glm-130b: An open bilingual pre-trained model*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. *Opt: Open pre-trained transformer language models*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: Evaluating text generation with bert*.

A Ensemble Prompts Explanation

Below are the prompts we are using in this research. The details explanation of each prompt can refer to [Vasselli et al. \(2023\)](#).

Zero-shot prompts consist of instructions without examples, while few-shot prompts include examples to guide the model towards relevant responses. Prompt (1) is categorized as a zero-shot prompt, refined to address issues like overly direct answers and sounding too much like an assistant. The rest of the prompts—(2), (3), (4), (5)—are few-shot prompts that require positive and negative examples to help the model avoid irrelevant responses.

Each prompt serves a specific purpose: Prompt (1) focuses on Contextual Understanding, Prompt (2) ensures Relevance, Prompt (3) aims to enhance Engagement, Prompt (4) emphasizes Clarity, and Prompt (5) is dedicated to providing Feedback. Together, these prompts tailor the model’s responses to match the student’s current learning stage and needs. By grasping the context (Contextual Understanding), the prompts direct the model to produce responses that are relevant to the student’s queries, thereby maintaining focus and relevance (Relevance). This relevance boosts student engagement (Engagement), encouraging sustained interest and participation, which is further supported by clear communication (Clarity) that makes complex concepts easier to understand and reduces confusion. Collectively, these prompts help the model generate optimal responses for educational contexts.

- (1) The following is a partial conversation between an English language learner and their teacher:

(conversation)

Can you give an example teacher follow-up to their previous message that would be helpful for the language learner? The message should be concise, and worded simply. It should either encourage the continuation of the current topic or gracefully transition to a new teacher-provided topic. Questions should be specific and not open-ended. Try to not sound like an assistant, but a teacher, in charge of the flow of the lesson.

- (2) Concatenation of prompt (1) and the following:

Good example: 'Can you make a sentence using 'within'?' Bad example: 'Do you have any questions about prepositions?'

- (3) Concatenation of prompt (1) and the following:

How does a teacher sound when responding to a student? What kinds of things would teachers say that chatbots would not? What do they not say? In your response provide an example of a response that sounds like a teacher and one that sounds like a chatbot? Respond succinctly

- (4) The following is a partial conversation between an English language learner and their teacher:

(conversation)

They are in the middle of a lesson. Can you give a possible way the teacher could respond?

Remember: A teacher typically sounds knowledgeable, authoritative, and focused on guiding and instructing students. They may use formal language and provide detailed explanations. Teachers often offer constructive feedback, encourage critical thinking, and ask probing questions to stimulate learning.

Example of a teacher-like response: "That's a great observation, but let's delve deeper into the topic. Can you provide some evidence to support your claim?"

A chatbot, on the other hand, may sound more informal and conversational. It tends to provide general information or brief responses without much elaboration.

Example of a chatbot-like response: "Interesting! Tell me more." Teachers typically avoid expressing personal opinions or biases. They also refrain from engaging in casual banter or unrelated conversations to maintain a professional and educational atmosphere.

- (5) Concatenation of prompt (1) and the following:

Here is an example of an exceptional teacher follow-up:

"Great job, student! Just a small correction, we should use the present tense verb "built"

instead of "build" since the construction has already been completed. So the correct sentence is: "The International Space Station is built by NASA." Keep up the good work! Now, let's move on to a new topic - let's talk about your favorite hobbies. Can you tell me what activities you enjoy doing in your free time?"

Here is an example of a poor teacher followup: "That's an interesting observation about poshness. Can you think of any examples of British accents that might be associated with poshness?"