Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Immediate Feedback in Computational Thinking: Generating hints using a Knowledge Graph

Nitesh Kumar JHA[a], Plaban Kumar BHOWMIK[b] & Kaushal Kumar BHAGAT[a*]
[a]*Advanced Technology and Development Centre, Indian Institute of Technology Kharagpur, India*
[b]*G.S Sanyal School of Telecommunications Indian Institute of Technology Kharagpur, India*
*kkntnu@hotmail.com

Abstract: In this study, we present an Online Inquiry-based Learning Platform for Computational Thinking (CT-ONLINQ) to develop the CT skills of students. The platform provides immediate feedback with hints to support students during problem-solving activities and encourages them to explore multiple solutions for a problem. The hints are generated using a Knowledge Graph that stores information about the solution details of a problem. A six-week study was conducted on 79 high school students to determine the effectiveness of the platform with hints on students' CT skills. The results showed that the students improved their CT skills significantly after the intervention. Findings suggest that the CT-ONLINQ platform consisting of IBL-based CT activities with immediate feedback could help school students improve their CT skills.

Keywords: Computational Thinking, Immediate feedback, Knowledge Graph, Inquiry-based learning

1. Introduction

Computational Thinking (CT) is a cognitive ability that allows individuals to develop computational solutions for a variety of problems (Wing, 2006). Inquiry-Based Learning (IBL) is an approach to learning that guides students through various phases, starting with exploring a problem, followed by collecting data to develop models, generating findings, and rigorously evaluating those findings to gain new understanding (Pedaste et al., 2015). In terms of the learning process, IBL in science and CT education are relatively similar due to the cyclic process of repeated revisions and refinement (Hoppe and Werneburg, 2019).

The current CT research focuses on teaching coding concepts to students (Jacob et al., 2020), but researchers argue that problem-solving, creativity, and algorithmic thinking should be taught, as these transferable skills hold greater value (García Peñalvo et al., 2016). In addition, studies on IBL in CT concentrate on creating logical artifacts and testing them to improve CT abilities (Hoppe and Werneburg, 2019). Moreover, researchers argue that analysis and evaluation of artefacts is critical to improving learning outcomes (Prayogi and Yuanita, 2018). To overcome the above issues, we have created CT-ONLINQ, an online platform for Computational Thinking (CT) that follows IBL-based CT activities (Jha et al., 2023). The platform provides hints and prompts during problem-solving activities, which are stored in a Knowledge Graph (KG). In KG, concepts are represented as nodes, and relationships between concepts are represented as edges. Different forms of immediate feedback have been developed by researchers to support coding and CT activities (Basu et al., 2017). However, there are lack of studies that provides hints stored in KG during IBL-based CT activities. The current study discusses the design of KG and evaluates the effect of the platform CT-ONLINQ on CT skills of high school students. The current study addresses the following research question (RQ):

- RQ: Did the online platform CT-ONLINQ with immediate feedback improved CT skills of high school students?

2. The CT-ONLINQ platform

CT-ONLINQ is an online educational platform that supports IBL-based CT activities (Jha et al., 2023). A brief description of the platform architecture and example interface tab is shown in Jha et al. (2023). The platform uses a KG to store hints and answers during IBL-based CT activities. Figure 1 shows a basic schema structure of the KG with properties for each problem. The KG schema is divided into three subschemas: Answer comparison, Goal, Hints and Answers, as described below.

- Answer comparison: It contains problem's question and the sequence of goals related to CT step.
- Goal: It contains step name, task statement, the best solution, hints and answers sequence related to the task, and a no matched answer section that contains hints when input does not match with any correct or incorrect answer.
- Hints and Answers: It contains type of answer (correct/incorrect) that matches the input, an alternate answer to compare with the input, a prompt for the hint, and three types of problem-specific hints: basic (with little elaboration), intermediate (with more elaboration), and alt hint (for alternate correct answer). The alternate correct answer (only available for the correct answer type) is a solution but not the best solution for the task.

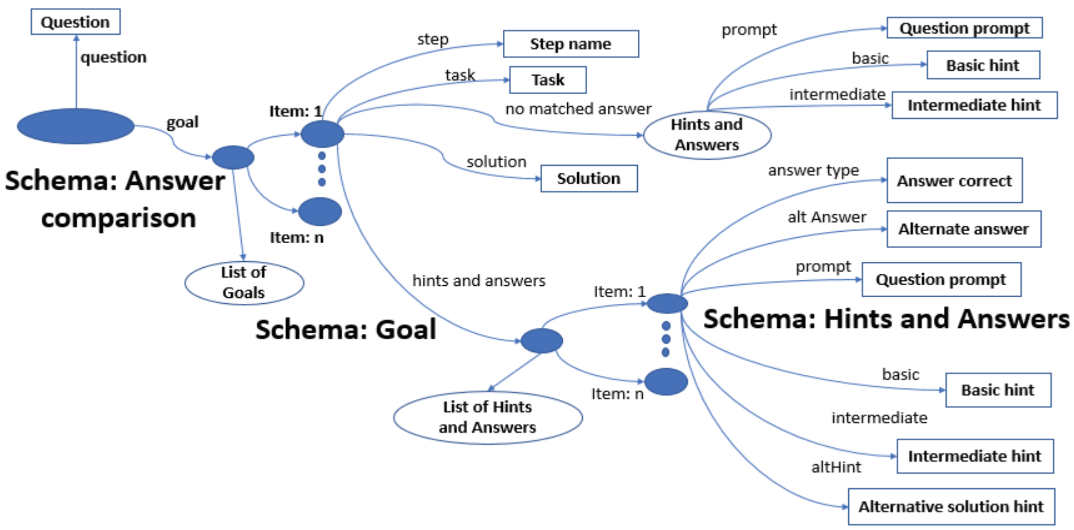

Figure 1. Schema of the Knowledge Graph

The platform parses the graph to provide relevant hints based on the number of attempts made during problem-solving activities. For example, Consider the problem statement 'Write an algorithm to find the sum of numbers 1+2+…+60'. For an incorrect answer in decomposition step- 'Find a way to add number from 1 to 60' the hints are: basic(attempts<4)- can you do better? Add number paiwise to find total sum; intermediate (attempts<6)- can you do better? Find pairs having sum 61; Advanced- (1+60)+(2+59)+…(30+31).

3. Methodology

The experiment involved 79 high school students, aged 13-15, from a public high school in India. All the students were given immediate feedback with hints during problem-solving activities. We adapted two different sets of 12 questions (max score 108, pre and post-test) from Bebras Thinking Challenge (Bebras-Ireland, 2020). Initially, the students signed up to the CT-ONLINQ platform, completed an example activity (see Jha et al., 2023), and completed bebras questions (pre-test with Cronbach's alpha 0.74). Next, the students completed six problem-solving activities on the platform and completed post-test with Cronbach's alpha 0.77.

4. Result and Discussion

The average pre-test score was 42.11 with a standard deviation (SD) of 10.69, while the post-test score was 55.43 with a SD of 14.07. We conducted a paired samples t-test to compare the pre and post-test scores to assess the improvement in CT skills. Findings showed that there was a significant difference in CT skills (t=6.697, p=.000, d=1.06), with an average difference of 13.316 (SD=1.988). The result supports the study that useful feedback and hints improved student understanding in a CT-based learning environment (Basu et al., 2017).

5. Conclusion and future work

The present study investigated the effect of CT-ONLINQ with immediate feedback on the CT skills of high school students. A total of 79 students from high school participated in the current study. The platform implemented the IBL-based CT workflow, allowing students to develop algorithmic solutions, analyze the algorithms, and compare the algorithms. The platform provided immediate feedback at each CT step. Hints and answers were stored in the KG. The findings confirmed that the intervention significantly improved the CT skills of students working on the platform.

With regards to future work, the current study examined the effect of IBL-based CT activities on the online platform but did not compare them with offline IBL-based CT activities. Future studies can compare the effects of offline versus online IBL-based CT activities on the enhancement of CT skills. The study provides a new way to teach CT skills to high school students using IBL-based CT activities without any coding practice.

References

Basu, S., Biswas, G., & Kinnebrew, J. S. Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Modeling and User-Adapted Interaction 27*, no. 1 (2017): 5-53.

Bebras-Ireland (2020). Online reference included in article [Internet document] URL https://bebras.techweek.ie/ Accessed 1/04/2023.

García Peñalvo, F. J., Reimann, D., Tuul, M., Rees, A., & Jormanainen, I. (2016). *An overview of the most relevant literature on coding and computational thinking with emphasis on the relevant issues for teachers*. Belgium. doi:10.5281/zenodo.165123

Hoppe, H. U., & Werneburg, S. (2019). Computational thinking—More than a variant of scientific inquiry. *Computational thinking education*, 13-30.

Jacob, S., Nguyen, H., Garcia, L., Richardson, D., & Warschauer, M. (2020, March). Teaching computational thinking to multilingual students through inquiry-based learning. In *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)* (Vol. 1, pp. 1-8). IEEE.

Jha, N.K., Bhowmik, P.K., & Bhagat, K. K. (2023, July). Usability Evaluation of an Online Inquiry-based Learning Platform for Computational Thinking (CT-ONLINQ). In 2023 *International Conference on Advanced Learning Technologies (ICALT)* (pp. 182-186). IEEE. https://doi.org/10.1109/ICALT58122.2023.00059.

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational research review, 14*, 47-61.

Prayogi, S., & Yuanita, L. (2018). Critical-inquiry-based-learning: Model of learning to promote critical thinking ability of pre-service teachers. In *Journal of Physics: Conference Series* (Vol. 947, No. 1, p. 012013). IOP Publishing.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33-35.

Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Composite Score for ChatGPT Prompt Efficiency: A Computational Linguistic Analysis of Engineered Chatbot Prompts

Michelle BANAWAN
Asian Institute of Management, Philippines
mbanawan@aim.edu

Abstract: The use of chatbots has become increasingly popular in the field of education. Hence, the quality of the prompts used by chatbots can greatly influence the quality of the generated responses which in turn contributes to desirable outcomes related to the overall learning experience. However, even as known criteria for efficient prompts are prevalent, there is a dearth of measurable and concrete linguistic factors that guide prompt engineering. This paper presents a composite score for ChatGPT prompts which can be made applicable to other foundational generative AI chatbots. Through a computational linguistic analysis of known efficient prompts used in learning, emergent linguistic factors point to the relationship of linguistic features and the confidence of ChatGPT responses to well-structured prompts that use the said linguistic features. The linguistic features are the average collostructural strength, collostructural ratio diversity, specificity, and academic language use. These features depict the quality of prompts that pertain to the grammatical structure, specificity, and relevance to the task at hand, and academic language use. Further, these features constitute a composite score for prompts introduced in this study that represent linguistic efficiency and subsequently, correlates to perplexity or certainty estimates of the generated responses of ChatGPT.

Keywords: ChatGPT, PCA, NLP, computational linguistics, prompt engineering

1. Introduction

Conversational agents such as chatbots are rapidly emerging as a promising technology for education and learning. With the help of advanced language models, chatbots can facilitate personalized learning (Allen, et al., 2016; Ashok, et al., 2021; Holotescu, 2016), provide instant feedback (Graesser, 2015; McCarthy, et al., 2022, and engage students in an interactive and engaging manner (Lin & Chang, 2020). However, the effectiveness of these chatbots, specifically generative AI-based chatbots, heavily relies on the quality of prompts that they use to elicit responses from the AI large language models (LLMs) (Dang, et al., 2022).

The challenge lies in engineering prompts that can generate accurate and comprehensive responses from LLMs. While many existing prompts are effective, they often rely on domain-specific knowledge, which limits their generalizability across different educational settings. Moreover, there is a lack of understanding of the linguistic features that contribute to the effectiveness of prompts, which makes it difficult to develop effective prompts systematically.

The current state of prompt engineering for educational chatbots is largely characterized by a trial-and-error approach, where developers often rely on ad-hoc strategies to generate prompts. This approach is not only time-consuming and resource-intensive but also limits the effectiveness of the chatbots in facilitating learning.

To address these challenges, this paper presents a computational linguistic analysis of prompts which can be leveraged in enhancing the learning experience. By leveraging computational methods, this work attempts to quantify prompt efficiency.

2. ChatGPT[1] and Generative AI Tools

Generative language models have revolutionized the field of natural language processing (NLP), enabling the creation of chatbots that can generate human-like responses to user inputs. Among these models, GPT (Generative Pre-trained Transformer) stands out as one of the most widely used and effective models for generating text.

The GPT model is based on the transformer architecture, a deep neural network that is capable of processing input sequences of variable length and generating output sequences of variable length. The model is trained on a large corpus of text data, and its parameters are optimized to minimize the prediction error of the next token in the sequence given the previous tokens. One of the most well-known implementations of the GPT model is ChatGPT, a conversational agent that can generate human-like responses to user inputs. ChatGPT is based on the GPT-3, GPT-3.5, and now GPT-4 architecture, which is currently one of the largest and most powerful models available for generating text.

In the field of education, ChatGPT and other generative AI tools have numerous applications. One of the most promising applications is in the area of personalized learning. Chatbots can be used to deliver customized learning experiences, adapting to the needs and preferences of individual learners. They can also provide instant feedback and support, helping learners to overcome obstacles and stay motivated.

Another application of generative AI tools in education is in the creation of educational content. Chatbots can generate text-based content such as summaries, study guides, and practice questions, which can be used to supplement traditional educational materials. Additionally, they can be used to create interactive learning experiences, such as simulations and games, that engage learners in a fun and immersive way.

ChatGPT, as well as other generative AI tools, has now become prevalent in education and across many domains and industries. These tools have the potential to revolutionize the field of education by enabling personalized learning, creating interactive content, and providing instant feedback and support.

2.1 Prompt Engineering for Education

Effective prompt engineering is critical for generating accurate and comprehensive responses from LLMs in various domains, most especially in educational settings. Prompt engineering involves crafting well-designed prompts that clearly and concisely communicate the intended learning objectives and desired outcomes. However, prompt engineering is not a straightforward process, and often requires iterations of trial and error, feedback, and back-and-forth conversations with ChatGPT as additional information is required to achieve the desired information in generated responses. By investing in prompt engineering and creating effective prompts, educational institutions can harness the power of ChatGPT for many applications including to enhance student learning and automate time-consuming tasks, such as grading and feedback, to name a few. However, it's important to note that prompt engineering requires careful planning and ongoing refinement to ensure that the prompts are effective and aligned with the intended learning outcomes. As such, concrete measures of what constitute efficient prompts should be investigated and put in place. Ongoing work related to prompt engineering of large language models (LLMs) have found that a number of strategies affect the quality of the generated outcome, e.g. repetition, providing examples, using code-like structures, or choosing the right instructions

[1] https://openai.com/product/gpt-4

(Zamfirescu-Pereira, et al., 2023), focusing on the word order of prompts (Yao Lu, 2021), using structured prompt patterns that explicitly contain elements like intent, context, motivation (White, et al., 2023), to name a few. However, there is little work on probing into the linguistic features that characterize well-structured and efficient prompts that generate confident responses from LLMs and other generative AI models, like ChatGPT.

2.2 Research Question
The central aim of the study is to seek to develop a comprehensive understanding of the linguistic factors that contribute to the effectiveness of prompts for LLMs which can benefit educational settings. The research question guiding this study is: What are the linguistic features of efficient ChatGPT prompts? By answering this question, the study aims to articulate what specific linguistic markers should be present in formulating efficient prompts which would in turn generate more confident responses from LLMs and minimize the extent of hallucinations in the generated responses.

3. Methods
3.1 Prompt Collection
The corpus used in this study was collected from publicly available effective ChatGPT prompts. The sources include effective prompts for teachers from https://www.learnprompt.org/chat-gpt-prompts-for-teachers/ and used in curriculum management, classroom management, professional development, explaining difficult concepts and providing examples, as well as other use cases. Additional effective prompts were collected from https://github.com/f/awesome-chatgpt-prompts which is a github repository of effective ChatGPT prompts with 247 pull requests, 8.9k forks, and includes 2,190 downloads from its huggingface[2] site as of March 2023. A total of 500 effective ChatGPT prompts ($mean_{word\ count}$ = 82.19; $sd_{word\ count}$ = 37.86) were gathered to comprise this study's corpus.

3.2 Prompt Efficiency Evaluation
For this study, prompt efficiency is operationalized as the ability of a prompt to elicit coherent and confident responses usually characterized by having low perplexity scores. This metric is useful when the goal is to evaluate the effectiveness of prompts in generating high-quality responses from an AI generative model. The efficiency or quality of prompts was performed using the pipeline in figure 1. A random sample of 200 prompts from the corpus was used in calculating the perplexity scores of generated responses. Perplexity scores are used to determine the quality of generated responses given the prompts that are provided to a generative AI model. The lower the perplexity score, the more likely it is generated by ChatGPT (Malinka, et al., 2023) or an AI generative model. This metric has been used in works like determining if short texts are generated by a human or by ChatGPT(Yeadon, et al., 2023; Mitrović, 2023). Using open AI's GPT3 API, the researcher generated responses to the sample corpus of 200 well-written prompts then approximated the perplexity scores of the generated answers using Hugging Face's gpt models. The resulting perplexity scores of the generated results ranged from 9.84 to 52.1 with a mean of 28.17 (sd=13.4). The relatively low approximate perplexity score mean suggests that the prompts that were randomly sampled from the collected corpora used in the study were able to elicit generated responses that were relatively more certain. This could imply that the prompts are clear and effective, as the GPT model was able to generate coherent and well-formed text in response to them with relatively high certainty and, with post-hoc qualitative inspection, were observed to be more accurate.

[2] https://huggingface.co/datasets/fka/awesome-chatgpt-prompts

Figure 1. Pipeline for Evaluating the Efficiency of Prompts

3.3 Linguistic Index Extraction

Natural language processing is performed on the entire corpus of 500 prompts (see figure2). Linguistic indices were extracted from the entire corpus of 500 prompts using the Suite of Analytic Linguistic Analysis Tools (SALAT; www.linguisticanalysistools.org/). The SALAT NLP tools extract a wide array of linguistic indices that measure syntax, readability, lexical diversity, lexical sophistication, cohesion, grammar, spelling, and other structural and mechanics constructs. A total of 515 linguistic measures were extracted from the corpus of efficient prompts. To prepare the dataset for principal component analysis (PCA), the extracted indices were further preprocessed. First, indices with zero to only 5% variance were removed, as they do not provide significant information for the analysis. Next, highly correlated indices (> 0.799) were eliminated to avoid multicollinearity issues that may arise during PCA. Finally, any indices with missing or null values were removed. The resulting dataset contained 72 linguistic indices which were used for the PCA analysis. Prior to the PCA analysis, the dataset was standardized by scaling each index to have a mean of zero and a standard deviation of one.

Figure 2. NLP and preprocessing pipeline

3.4 Principal Component Analysis

After extracting the linguistic indices, we performed a principal component analysis (PCA) to identify underlying factors that explained the variability of the linguistic indices. We used the Scikit-learn library in Python to perform PCA on the dataset. The use of PCA in this study was chosen to provide a more comprehensive understanding of the linguistic patterns present in educational prompts, and for generating more targeted recommendations for prompt engineering in educational settings.

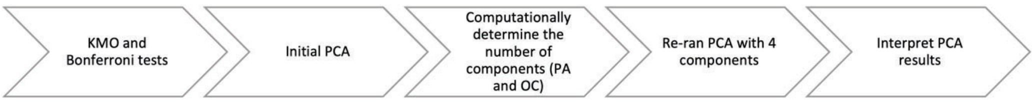

Figure 3. PCA pipeline

To ensure that the dataset was appropriate for PCA, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was conducted. The KMO measures the proportion of variance among the observed variables that can be explained by the underlying factors. In this study, the overall KMO value was 0.65 and the individual features all had a KMO above 0.50, indicating that the dataset was suitable for PCA analysis. Moreover, a Bonferroni test of sample adequacy was also conducted to determine if the sample size was sufficient for PCA. The test indicated that the sample size was sufficient for PCA analysis ($p < 0.05$). A crucial step in PCA is to determine the number of principal components (PCs) to retain. The researcher used two methods for selecting the number of components: Parallel Analysis (PA) and Optimal Coordinates (OC). PA and OC are two methods for selecting the number of components in PCA that have advantages over the elbow method and eigenvalues > 1 method. Both methods consider the variability of eigenvalues and the correlations among variables, resulting in more accurate and reliable estimates of the number of components to retain. The PCA analysis produced four principal components using computational methods (parallel analysis and optimal coordinates, see figure 4), explaining 60% of the total variance. The first component accounted for the highest proportion of the variance (22%), the second component accounted for 16% of the variance, while the last two components accounted for the lowest proportion (12% each). The components were then named and interpreted based on the linguistic attributes of the features that had the highest loadings (loadings threshold = 0.40).

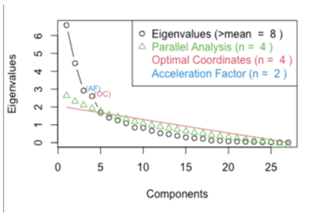

Figure 4. Computational Solutions to Finding the Optimal Number of Components.

3.5 Qualitative Analysis of PCA

The researcher conducted a qualitative review of prompts that scored high and low in each of the emergent components. The primary objective of this step was to assist in interpreting and naming the emergent components and to scrutinize the specific prompts, thereby confirming the constitutive NLP indices that loaded onto each component. In this posthoc qualitative analysis the results of the PCA are studied in a more detailed and nuanced way to gain a better understanding of what each component represents in terms of the original variables.

3.6 Linguistic Components: Interpretation of PCA Results

The PCA analysis revealed a total of 4 components that accounted for up to 60% of the cumulated variance in the linguistic properties of efficient ChatGPT prompts (see table 1).

Table 1. *Factor Loadings and Variances*

Component	**Constitutive Indices (Original Variables with loading > 0.40)**	**Proportion of Variance (%)**
Linguistic Association and Construction Strength	1. Average lemma construction combination frequency (academic, news, fiction, magazine) 2. Average approximate collostructional strength (academic, news, fiction, magazine)	22
Collostruction Ratio Diversity	1. Collostruction ratio ((academic, news, fiction, magazine)	16
Specificity	1. Sentence length 2. T-unit length 3. Count of verb phrases	12
Academic Language Use	1. Frequency of academic words 2. Combination of academic words	12

3.6.1.1. Component 1: Linguistic Association and Construction Strength (LACS)

The first component, which accounted for the largest amount of variance, 22%, is comprised of the average lemma construction combination frequency (ALCP) and the average approximate collostructional strength (AAC). The ALCP measures how frequently particular combinations of words are used together and how frequently particular combinations of words are used together. The AAC measures the strength of association between a given word and the other words that typically co-occur with it. Both measures emerged specifically for words on specific genres, such as academic, news, fiction, and magazine and how they are combined in a prompt. Collostruction or a person's collostructional preferences reveal interesting effects of the grammatical choice contributing to pragmatic phenomena, e.g., hedging or tentativeness (Rautionaho, 2021; Stefanowitsch & Gries, 2003). ALCP and AAC, refer to association strength which can be seen as an indicator of the level of lexical diversity, prototypicality, acquisition. Prototypicality allows for more natural and effective language. These features have been shown to be positively correlated with writing quality (Kyle, 2016). Thus, prompts that exhibit higher levels of association strength (ALCP and AAC) may be more effective in stimulating thoughtful and insightful responses from generative AI bots.

3.6.1.2. Component 2: Collostruction Ratio Diversity (CRD)

The collostruction ratio (CR) measures the degree of association between two words in each corpus. It is the ratio of the observed frequency of the collocate to the expected frequency of the collocate, given the frequency of the word. Component 2 focuses on the frequency of certain collocations within a prompt, namely the ratio of words that tend to co-occur with other words in specific domains, such as academic, news, magazine, fiction, and all domains. The CR is a measure of the strength of association between two words, and it has been found to be an effective indicator of semantic relatedness. An increase in collostruction ratio has been found to also depict an increase in writing quality (Mostafa & Crossley, 2020). CR Diversity has five constitutive indices, each of which is a CR for a different corpus: academic, news, magazine, fiction, and all. The CR for academic words measures the strength of association between the prompt and academic words, while the CR for news words measures the strength of association between the prompt and words commonly found in news articles, and so on. A prompt that has a diverse range of collocates (collostruction ratio) from different corpora may be more effective in eliciting a response from a chatbot, especially if the chatbot is designed to respond in a way that is relevant to the academic context which is intrinsically diverse in terms of topics and knowledge domains.

3.6.1.3. Component 3: Specificity

Component 3 is comprised of three constitutive indices: sentence length, unit length, and verb phrases. This component measures the degree of specificity and clarity of prompts. Sentence length is the average number of words per sentence in the prompt, which is a measure of syntactic complexity. Unit length is the average number of clauses per sentence, which indicates how much information is packed into each sentence. Verb phrases refer to the average number of verb phrases per sentence, which reflects the level of specificity and clarity in the prompt and are related to the writing quality (Mostafa & Crossley, 2020). Verbs play a critical role in conveying specific meaning and action in a sentence. Choosing the appropriate verb can enhance the clarity and specificity of the message being conveyed. Therefore, including verb phrases as a constituent index in the composite score can help assess the specificity and clarity of the language used in prompts, which can ultimately contribute to the effectiveness of the prompts in educational settings. Further, it is also notable that all three constitutive indices are indicators of syntactical complexity (Kyle & Crossley, 2018). While the syntax of a prompt does not necessarily determine the quality of the response generated by language models and chatbots. Both simpler and more sophisticated syntax can lead to high-quality responses, depending on the context and the specific task at hand. Ultimately, the syntax of a prompt should be chosen based on the intended audience and the specific goals of the task. Chatbots are capable of understanding and generating text with a wide range of syntactic structures.

3.6.1.4. Component 4: Academic Language Use (ALU)

Component 4, Academic Language Use, is characterized by a set of constituent indices that measure the frequency and combination of academic words in prompts. These indices include measures of average lemma frequency, construction frequency, and lemma-construction combination frequency of academic words. The use of academic language in prompts is particularly important in educational settings, as it reflects the technical terms of academic subject areas and can help generative AI bots and language models extract domain-specific responses. Academic language is often characterized by its precision, specificity, and technical terminology. In academic writing, authors are expected to use language that is clear and unambiguous, and to avoid vague or imprecise expressions. As a result, academic language tends to be more specialized and specific than everyday language. When a prompt contains academic language, it often signals that the topic or subject matter is more specialized or technical in nature. The use of specific terminology and jargon in the prompt can help to define the scope and context of the question, making the generated responses more targeted and tailored to the given topic. By incorporating a higher degree of academic language use in prompts, the chatbot can produce more specialized and technical responses that align with the academic context. This can be especially important in educational settings where the topics discussed are often complex and specialized.

In summary, the results of the PCA analysis showed that the linguistic features related to LACS, CRD, specificity, and academic language use were the most prevalent features that characterized the corpus of efficient educational prompts that were known to produce high quality responses from ChatGPT.

3.7 Composite Score: Linguistic Prompt Efficiency Score (LPES)

Using the results of the PCA, a composite score is introduced in this study and derived from the PCA loadings and constituent indices. PCA loadings represent the relative importance of each component in explaining the variance in the data, while the constituent indices represent the specific linguistic features that contribute to each component. By multiplying the loadings with the normalized scores of the constituent indices for each prompt, a composite score that takes into account the relative importance of each component and the specific linguistic features that contribute to its efficiency is introduced in this paper.

Composite score (LPES) = (LACS * 0.32) + (CRD *0.16) + (Specificity * 0.12) + (Academic Language Use * 0.12)

With this prompt composite score, a more quantitative metric is introduced in terms of how prompt efficiency can be approached through linguistic features. Prompt engineers will be able to tailor their lexical choices when crafting efficient prompts using the emergent components in this study and the LPES composite score. However, the researcher would like to highlight that the PCA done in this study only resulted to 4 components that represent up to 60% of the variance in the prompt corpus. Factors that can account for the remaining 40% need to be investigated further. To evaluate the perplexity scores of the generated responses for prompts with high LPES, the following pipeline was implemented (see figure 5).

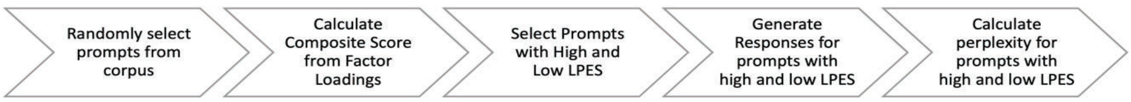

Figure 5. LPES evaluation pipeline

A prompt with high LPES is shown below:

"I want you to act as a travel guide. I will write you my location and you will suggest a place to visit near my location. In some cases, I will also give you the type of places I will visit. You will also suggest me places of similar type that are close to my first location. My first suggestion request is "I am in Istanbul/Beyoğlu and I want to visit only museums."

The prompt is specific in its request for a place to visit near the location provided, with the added specification of museums only. The prompt is clear in its instructions, with a straightforward request and a clear indication of the type of place the user is interested in. The prompt also exhibits academic language use, as it is presented in a formal and technical manner, with precise language and specific terms such as "location," "place to visit," and "museums." These elements provide a clear and unambiguous description of what the user is looking for, which helps the chatbot generate a more effective and relevant response. Overall, the prompt is well-crafted and meets the requirements for an effective educational chatbot prompt. The generated response of the aforementioned prompt is presented below with a perplexity score of 34.49.

Sure, as a travel guide, I can suggest some museums for you to visit in Istanbul/Beyoğlu. There are several museums in the area that are worth visiting, including the Istanbul Modern Museum, the Pera Museum, and the Museum of Innocence. All of these museums have a unique collection of art and artifacts that showcase the rich cultural heritage of Istanbul. Additionally, if you're interested in visiting more museums, there are several other options nearby that you might want to check out, such as the Istanbul Archaeological Museum, the Topkapi Palace Museum, and the Hagia Sophia Museum. These museums are all located within a short distance of Beyoğlu and offer a fascinating insight into the history and culture of Istanbul.

A vague prompt was written (shown below) because the corpus of prompts used in this study is comprised of known well-written and effective prompts. This ill-wirten prompt got a low LPES score and observably low presence of the constitutive elements is shown below:

"Act as a data scientist, design a hybrid recommender system approach for a real world use case."

The generated response (excerpt shown below) was detailed and informative, but the perplexity score was 59.44, i.e. showing that ChatGPT has low confidence in its generated response.

As a data scientist, I can help design a hybrid recommender system approach for a real-world use case. Let's consider a scenario where we are designing a recommender system for an online marketplace that sells a variety of products, ranging from electronics to clothing. To design a hybrid recommender system for this use case, we can combine collaborative filtering and content-based filtering. Collaborative filtering can be used to identify groups of similar users based on their purchase history, ratings, and other behaviors…

It was observed that the math prompt, even as it is well-written, had the lowest LPES and had a relatively high perplexity. Finally, a correlation analysis was performed on the sub-corpus of 200 randomly sampled prompts. The results indicate a significant negative correlation between LPES and Perplexity score, with a correlation coefficient of -0.65 ($p < 0.05$). This result indicates a meaningful relationship between LPES and Perplexity score. Additionally, the scatterplot of the data shows a clear downward trend, further supporting the negative correlation between LPES and Perplexity score. The LPES scores for the sample of 200 well written prompts ranged from -0.42 to 1.21 ($mean_{LPES} = 0.68$; $sd_{LPES} = 0.04$).

4. Conclusion

Understanding how to engineer prompts that are comprehensive and accurate can aid in the development of more effective educational materials. By leveraging these linguistic features in engineering prompts, learners and teachers can be more intentional in formulating prompts that leverage chatbots, large language models, and other generative AI tools. The four components of language in prompts reveal that efficient prompts are characterized by intentional grammatical structures (LACS and CRD), specificity, and academic language use. Prompts with high values for both LACS (component #1) and CRD (component #2) are likely to be better grammatically structured and more natural prompts compared to those with low values for both measures.

Prompts with high LACS (component #1) indicate that the target word in the prompt has a strong tendency to occur with specific collocates, which means that the prompt is more likely to be grammatically well-formed such that the message is made clearer because of the grammatical structure. Prompts with high CRD (component #2) indicate that the words in the prompt tend to occur together frequently, which suggests that they are semantically related and likely to form a coherent sentence or idea. Conversely, prompts with low LACS and CRD values suggest that the words in the prompt are not closely associated with each other, and may not form a coherent or natural-sounding sentence. In this case, the generated prompt may be less useful or effective for its intended purpose. Overall, higher values for both LACS and CRD indicate that a chat prompt is more likely to be natural, coherent, and effective for generating meaningful and engaging conversations because of the grammatical decisions when formulating the prompt. Chatbots' ability to generate better answers to prompts that are semantically coherent as a result of the well-formed grammar is due to the training on natural language patterns, also a grammatically well-formed knowledge base or training base, to generate related responses in the same context. Specificity (component 3) measures the degree of clarity of prompts through three constitutive indices: sentence length, unit length, and verb phrases. Efficient chatGPT prompts should be specific and clear to generate high-quality responses. Component 3 helps achieve this by measuring the clarity and specificity of the prompts through the three constitutive indices. By measuring sentence length, unit length, and verb phrases, the component evaluates how much information is packed into each sentence. In addition, using specific verbs can help enhance the clarity and specificity of the message being conveyed, and longer and more complex prompts may provide more context and specificity. However, the quality of a response ultimately depends on various factors beyond just the length, including the relevance and clarity of the prompt and the accuracy of the information provided. Overall, Component 3 contributes to efficient chatGPT prompts by providing a means to measure and improve the specificity and clarity of the prompts by incorporating just enough elaboration, hence having relatively longer prompts.

Lastly, academic language use is precise, specific, and technical, making it more specialized than everyday language. The presence of academic language in a prompt indicates a more technical or specialized topic. Using academic language helps define the context and scope of the question, enabling chatbots to generate targeted and tailored responses. This results in more accurate and precise answers. Component 4 is crucial in educational settings, where topics are complex and require specialized knowledge, producing more specialized and technical responses that align with academic context.

In conclusion, the linguistic attributes of efficient prompts that were identified in this study have important implications for learning. By leveraging these attributes, learners and teachers can be more intentional in their lexical choices when they leverage generative AI and similar tools in the teaching and learning process.

4.1 Limitations

The dataset used in this study was limited to publicly available corpora of known efficient prompts that teachers, researchers, and practitioners used. There was no comprehensive validation of the actual quality of the prompts aside from the perplexity estimate validation of its generated responses. Further the corpora used in this study do not represent the full range of possible prompts and topics that can be encountered in real-world applications and deployments of generative AI in education.

5. Future Work

There are several areas of future work that can build on the findings of this study. One potential avenue is to develop a tool or framework that can automatically identify high-quality prompts based on the linguistic features identified in this study. Such a tool could be used by educators and students to create more effective prompts that can lead to more accurate and comprehensive responses from LLMs. While the scope of this study is prompt engineering for chatbots, future directions in education can scale the benefits of the composite score (LPES) that is introduced in this study. LPES can be used to evaluate the linguistic efficiency of prompts used in educational chatbots or tutoring systems that are built on top of foundational models, helping to identify which prompts are most effective to facilitate optimal information retrieval.

References

Allen, L., Dascalu, M., McNamara, D. S., Crossly, S., & Trausan-Matu, S. (2016). Modeling individual differences among writers using readerbench. In *EDULearn16: 8th international conference on education and new learning technologies* (pp. 5269-5279). IATED Academy.

Ashok, M., Ramasamy, K., Snehitha, G., & Keerthi, S. R. (2021, March). A systematic survey of cognitive chatbots in personalized learning framework. In *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 241-245). IEEE.)

Holotescu, C. (2016). MOOCBuddy: a Chatbot for personalized learning with MOOCs. In *RoCHI* (pp. 91-94).)

Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 42-50.

McCarthy, K. S., Roscoe, R. D., Allen, L. K., Likens, A. D., & McNamara, D. S. (2022). Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision?. *Assessing Writing*, *52*, 100608.

Lin, M. P. C., & Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot. *Journal of Educational Technology & Society*, *23*(1), 78-92.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to Prompt? Opportunities and Challenges of Zero-and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. *arXiv preprint arXiv:2209.01390*.).

(Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023, April). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-21).

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, & Pontus Stenetorp. (2021). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. arXiv preprint arXiv:2104.08786.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Malinka, K., Perešíni, M., Firc, A., Hujňák, O., & Januš, F. (2023). On the educational impact of ChatGPT: Is Artificial Intelligence ready to obtain a university degree?. *arXiv preprint arXiv:2303.11146*.

Yeadon, W., Inyang, O. O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education*, *58*(3), 035027.

Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. arXiv preprint arXiv:2301.13852.

Rautionaho, P., & Fuchs, R. (2021). Recent change in stative progressives: a collostructional investigation of British English in 1994 and 2014. *English Language & Linguistics*, *25*(1), 35-60.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. International journal of corpus linguistics, 8(2), 209-243.

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.

Mostafa, T., & Crossley, S. A. (2020). Verb argument construction complexity indices and L2 writing quality: Effects of writing tasks and prompts. Journal of Second Language Writing, 49, 100730.

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333-349.