A Study on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model

Daeseung Park[1], Gi-taek An[2], Chayapol Kamyod[3] and Cheong Ghil Kim[1,*]

[1]*Department of Computer Science, Namseoul University, Cheonan, Republic of Korea*
[2]*Korea Food Research Institute, Wanju-gun 55365, Republic of Korea*
[3]*Computer and Communication Engineering for Capacity Building Research Center, School of Information Technology, Mae Fah Luang University, Chiang Rai 57100, Thailand*
E-mail: dspark@daeseungpark.com; gt@kfri.re.kr; chayapol.kam@mfu.ac.th; cgkim@nsu.ac.kr
[*]*Corresponding Author*

Received 04 December 2023; Accepted 02 February 2024; Publication 22 February 2024

Abstract

In the realm of Generative AI, where various models are introduced, prompt engineering emerges as a significant technique within natural language processing-based Generative AI. Its primary function lies in effectively enhancing the results of sentence generation by large language models (LLMs). Notably, prompt engineering has gained attention as a method capable of improving LLM performance by modifying the structure of input prompts alone. In this study, we apply prompt engineering to Korean-based LLMs, presenting an efficient approach for generating specific conversational responses with less data. We achieve this through the utilization of the query transformation module (QTM). Our proposed QTM transforms

Journal of Web Engineering, Vol. 22_8, 1187–1206.
doi: 10.13052/jwe1540-9589.2285

input prompt sentences into three distinct query methods, breaking them down into objectives and key points, making them more comprehensible for LLMs. For performance validation, we employ Korean versions of LLMs, specifically SKT GPT-2 and Kakaobrain KoGPT-3. We compare four different query methods, including the original unmodified query, using Google SSA to assess the naturalness and specificity of generated sentences. The results demonstrate an average improvement of 11.46% when compared to the unmodified query, underscoring the efficacy of the proposed QTM in achieving enhanced performance.

Keywords: AI, large language model, generative AI, few-shot learning, prompt engineering, AI Chatbot.

1 Introduction

Recent advancements in AI and NLP (natural language processing) have garnered attention, leading to active research in the field of LLM (large language models) related to chatbot technology. As chatbots and AI technologies continue to evolve, their performance has been steadily improving. With the evolution of them, which are trained on massive datasets using LLMs, they have reached a level where they can provide responses at a similar level to that of humans [1, 2].

Representatively, large-scale language models of the GPT (generative pre-trained transformer) series [3] have been introduced and are showing excellent performance in various NLP tasks. This is considered a key basic technology for Generative AI [4], which learns content patterns and creates new content with inference results. Many global tech giants, including OpenAI, Google, Deepmind, Meta, and other research institutes are conducting several large-scale projects based on different strategies and approaches [5].

Typically, models such as the GPT series are trained on large general corpus datasets such as web pages, books, papers, and articles. It can then be applied to a variety of natural language processing tasks using adaptive methods such as fine-tuning [6]. As a result, LLM has more than millions of parameters, allowing the model to learn a variety of language patterns and structures. This technological evolution of the LLM is having a significant impact on the AI community, revolutionizing the way AI algorithms are developed and used.

As for the LLM's model capacity, it is improving by expanding the model size or data size in the pre-trained language model (PLM). As a recent example, much larger PLMs, GPT-3 with 175B parameters and PaLM with 540B parameters, were trained to explore performance limits. Although the expansion is mainly done at different model sizes using similar processes of architecture and pre-training operations, these large PLMs show superior performance compared to small PLMs represented by 330M parameter BERT and 1.5B parameter GPT-2. Generally, large-scale language models (LLMs) refer to language models containing hundreds of billions (or more) of parameters, such as GPT-3, PaLM, Galactica, and LLaMA which have been trained on large-scale text data [7].

ChatGPT, taking a closer look, is a model based on GPT-3.5, and is an advanced model through changes in the model and learning method from the existing GPT-1 to GPT-3. The main change from GPT-1 to GPT-3 is the change in model size, which improves performance by learning more information from various datasets. GPT-3 uses few-shot learning, a technique to effectively learn a model even in situations where very little data is given, and prompt learning, a method of utilizing domain knowledge for model learning through input in the form of human-readable text. Prompt based learning performs various functions such as random writing, translation, web coding, and conversation. Furthermore, it is fine-tuned based on GPT-3.5 and allows human intervention during the learning process. By applying RLHF (reinforcement learning from human feedback), a reinforcement learning algorithm, to GPT-3.5, bias and harmfulness are reduced. Currently, in RLHF, humans rank the model's responses and reflect feedback through a reward function, so that human preferences are reflected in the model. The learning method consists of three stages, allowing additional learning of GPT-3.5 through prompt-based supervised learning and the RLHF algorithm. They consist of demo answer collection and a policy compliance verification stage, comparison data collection and a reward model training stage, and policy optimization stage with a reinforcement learning algorithm [7].

This paper applies prompt engineering to Korean-based LLMs, presenting an efficient approach for generating specific conversational responses with less data. For this purpose, we proposed a technique of utilizing the query transformation module (QTM). The proposed QTM transforms input prompt sentences into three distinct query methods, breaking them down into objectives and key points, making them more comprehensible for LLMs. This paper is composed as follows: Sections 2 overviews the GPT series; Section 3 introduces the proposed query transformation module; Section 4 includes the simulation environment, method, and results; finally, a conclusion and future works section follows.

2 Backgrounds

In general, LLM can be associated with the GPT series models. GPT stands for 'generative pre-trained transformer', representing AI models that are pre-trained on extensive data through machine learning to generate sentences. In particular, in the recent revolution of Generative AI, ChatGPT has gained prominence for its ability to engage in human-like conversations. ChatGPT can formulate responses to questions in a manner resembling human sentence construction [8–11].

2.1 GPT-1

Before the era of GPT, language models typically relied on labelled data and supervised learning. However, obtaining a large amount of labelled data is challenging due to the absolute necessity of human involvement in the labelling process. Naturally, unlabelled data is more easily accessible in significant quantities. Existing language models lacked effective methods to leverage unlabelled data. Therefore, GPT-1 focused on developing an efficient generative pre-training model using unlabelled data [12].

In GPT-1, to simplify the model and reduce computational complexity, only the decoder component of the transformer in Figure 1 [8] is utilized. Figure 2 [12] illustrates the form with the cross self-attention portion removed from the transformer [8, 9].

As training is conducted in an autoregressive manner, predicting the next word as illustrated in Formula (1), models in the GPT series demonstrate superior performance compared to conventional language models.

$$L_1(u) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \theta). \tag{1}$$

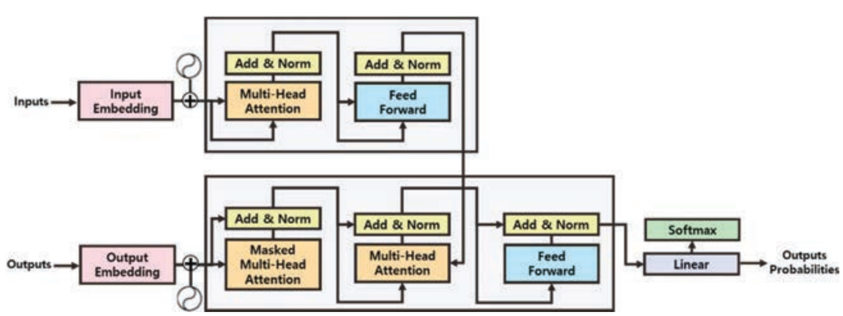

Figure 1 Transformer model.

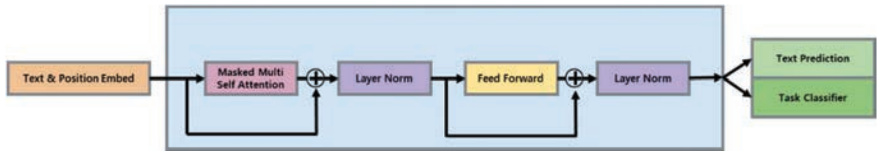

Figure 2 GPT-1 model.

Table 1 Method of in-context learning

Method	Description
Zero-shot learning	The model cannot look at any examples from the target class
One-shot learning	The model observes only one example from the target class
Few-shot learning	The model observes few examples from the target class

Autoregressive learning, by predicting the next word, enables the acquisition of the language's structure and the understanding of contextual and linguistic patterns, resulting in superior performance. When employing pre-trained models, a crucial step involves fine-tuning for the specific end task. Autoregressive learning proves beneficial in fine-tuning by promoting a more generalized training approach that avoids being overly biased towards particular problems [8, 9, 12].

2.2 GPT-2 and GPT-3

The objective of the GPT-2 model was to create a general language model through unsupervised pre-training, allowing for zero-shot downstream task execution without the need for fine-tuning. At the time, GPT-2 achieved state-of-the-art (SOTA) performance in many domains as an unfine-tuned model. This accomplishment underscored the potential of unsupervised pre-training, signifying a significant achievement by surpassing task-specific models and reaching the SOTA [8, 9, 12, 13].

Conventional LLMs commonly suffer from the drawback of being unable to perform tasks without fine-tuning. When the model is trained to execute a specific task through fine-tuning, its generalization ability is compromised.

In GPT-3, to address these issues, in-context learning is employed. In-context learning enables the pre-trained model to perform specific tasks without fine-tuning by providing examples when solving problems. In in-context learning, various methods exist, such as those outlined in Table 1, including zero-shot, one-shot, and few-shot, depending on the number of examples provided.

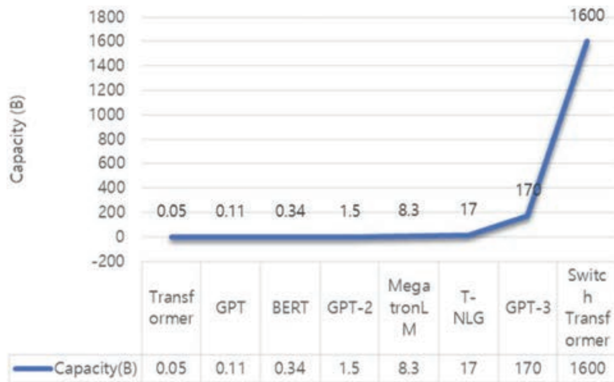

Figure 3 Model capacity of LLMs.

As another approach to enhance the model and address issues, increasing the model capacity, i.e., the number of parameters, is employed to enable the execution of more computations. As depicted in Figure 3 [1], recent high-performing LLMs exhibit a gradual expansion in model capacity, contributing to their superior performance [12–14].

Even with an increased model capacity and the application of few-shot learning in pre-trained models, the model may fail to accurately comprehend the purpose and direction of the task it requires. In such instances, the model cannot generate the intended response as envisioned by the questioner.

In this paper, we propose a novel approach to enhance the response generation performance of the GPT series LLM by introducing a method for constructing prompt structures.

3 Proposed Method

In general, the ChatGPT series of LLM (large language models) exhibits significant variations in the quality of generated sentences depending on the training data or environment. Therefore, methods such as fine-tuning or few-shot learning are often employed to instruct the model on how to generate high-quality responses [15–19]. Fine-tuning involves training the LLM on a large amount of additional data to generate responses. In contrast, few-shot learning generates responses based on a few sample sentences. Typically, when building conversational models such as chatbots using LLM, fine-tuning is predominantly employed, which requires a substantial amount of training data. However, in cases where specialized knowledge is not required, few-shot learning can be leveraged for a quicker and more straightforward development of conversational models [16–19].

In order to create conversational models using few-shot learning, the data input into the prompts plays a crucial role. The process of finding combinations of input values that can yield high-quality desired responses from LLM using prompt input data is referred to as prompt engineering. However, research in utilizing this approach for building conversational models in Korean LLM has been lacking, and it remains underutilized in the context of conversational models [18, 19].

3.1 Prompt Engineering

In general, language generation models tend to have the nature of generating connected sentences and paragraphs based on the input query. Figure 4 illustrates a case applying prompt engineering. When a user's query, which hasn't been trained through methods like fine-tuning, is input into a language generation model, as shown in (a) of Figure 4, it extends the query rather than generating a response, resulting in additional queries instead.

To go beyond this characteristic of LLM and generate appropriate interactive responses without fine-tuning, users are encouraged to input their queries in a simplified and suitable form. This approach, as depicted in (b) of Figure 4, allows the model to generate responses to the queries effectively.

This process is referred to as prompt engineering, and it involves creating prompt queries used in language generation models to generate natural and appropriate responses to queries. Additionally, prompt engineering enables the generation of input data that allows LLM to produce the most suitable questions or answers based on user input, improving the conversational flow of chatbots and creating more natural interactions. This, in turn, enhances chatbot efficiency and user satisfaction. Therefore, by appropriately

(a) (b)

Figure 4 Response generation methods in LLM. (a) Conventional response generation method. (b) Few-shot learning-based interactive response generation method.

configuring a Korean prompt-based few-shot learner, one can anticipate performance improvements in LLM [20–23].

3.2 Proposed Query Transformation Module

The query transformation module (QTM) in Figure 5 transforms queries in a manner distinct from the conventional approach (a), opting instead for the method illustrated in (b). In this study, we explore and propose a methodology where queries, engineered through prompt engineering using the approach in (b), serve as prompt queries for LLM. This ensures the precise recognition of the user's query intent by the LLM, effectively conveying the necessary objectives for proficient interactive response generation.

The original query technique involves presenting user-input queries directly to the language generation model without preprocessing or transformation for experimental purposes. In this work, the proposed technique utilizes four types of prompt queries: general query, preceding phrases query, cloze query, and purpose explicit query.

The preceding phrases query (PPQ) technique involves providing the initial sentence or words necessary for a relevant and appropriate response, allowing the subsequent sentence to be generated. As shown in the sample in Table 1, essential content required for the response, particularly the opening statement, is input into the language generation model. If there are no specific instructions in the input query, the language generation model can intentionally induce the required response by completing the subsequent sentence related to the preceding phrase.

The cloze query (CQ) technique involves presenting a sentence with certain parts left blank, accompanied by examples, and then completing the sentence by filling in the blanks. This method is commonly referred to as "fill

(a) (b)

Figure 5 Application of the query transformation module. (a) Conventional approach. (b) Proposed method.

in the blanks" and is frequently employed in language generation models. In other words, it entails providing a few question–answer pairs as examples and inducing similar responses based on them.

The purpose explicit query (PEQ) technique involves stating the purpose or objective and providing categorized examples accordingly in the query. The PEQ method goes beyond the cloze query technique by not only specifying the purpose behind the query but also presenting a variety of examples in different categories. This approach aims to guide language generation models to produce more natural responses by helping them understand the specific purpose behind the query and showcasing various forms of examples.

Each query technique is denoted as Query, PPQ, CQ, and PEQ. In accordance with Figure 6, experiments are conducted utilizing the appropriate responses for each method. The form and quality of the generated responses are then scrutinized and evaluated based on their respective query construction methods.

4 Experiment

In this section, we specify the settings for the models, and evaluation metrics for few-shot learning. We conduct our experiments and analyses based on the experimental settings specified in this section. More details about training environments and hyperparameters are described as follows.

4.1 Experiment Environment and Method

Google Colab, an open-source service offered by Google, is accessible to individuals with Gmail accounts. Google Colab is a valuable resource for researchers who may lack the necessary hardware resources or cannot financially afford GPU access. This service provides a substantial allocation of computing resources, including 12.72 GB of RAM and 358.27 GB of hard disk space for each runtime session. It's important to note that each runtime session has a duration of 12 hours, after which it automatically resets, requiring users to establish a new connection [24–26].

In Google Colab, we executed SKT GPT-2 and Kakaobrain KoGPT-3 models and queried them using the four query methods outlined in Section 3.2 in response to the experimental 3 queries from Figure 6. The results were then processed to remove special characters such as Korean characters, English characters, numbers, and periods using LinQ [27].

Type	Query	PPQ	CQ	PEQ
대화형질의	오늘 너무 우울해...	(질의) 우울해 하지 마세요.	(예시) 질문은 우울해. 응답은 우울해 하지 마세요. 때론 전문가에게 조언을 구하는 것도 좋은 방법이 될 수 있습니다. (질의) 질문은 오늘 너무 우울해. 응답은	(목적, 목표) 우울한 사람에게 응원 및 조언을 합니다. (분류별 예시) 힘내세요! 당신은 해낼 수 있어요!=응원 우울한 기분은 일반적인 경우에도 느낄 수 있는 자연스러운 감정의 일부입니다. 그러나, 같은 감정이 지속적이거나 강해지면 심리적 문제를 일으키거나 일상 생활에 영향을 줄 수 있습니다. 우울한 기분을 극복하는 방법으로는 운동, 음악, 예술, 친구나 가족과의 대화, 심리 상담 등이 있습니다. 이러한 것들을 시도해보고, 계속해서 우울한 기분이 지속되면 전문가에게 상담을 받는 것을 추천합니다.=조언 (질의) 우울해 하지마세요.=
대화형질의	너무 힘들고 죽고 싶어..	(질의) 힘내세요.	(예시) 질문은 힘들고 죽고싶어. 응답은 힘내세요. 마음이 힘들다면 전문가에게 조언을 구하는 것이 도움이 될 수 있습니다. (질의) 질문은 너무 힘들고 죽고싶어.. 응답은	(목적, 목표) 감정이 힘든 사람에게 응원과 조언을 합니다. (분류별 예시) 힘내세요! 때론 힘들고 아플지라도 당신은 이겨낼 수 있음을 확신합니다!=응원 저는 그러한 마음을 이해합니다. 그러나, 같은 감정이 오래 지속되면 건강에 부정적인 영향을 미칠 수 있습니다. 또한, 전문가의 도움을 받는 것도 좋은 방법입니다.=조언 (질의) 힘내세요.=
대화형질의	너무 사랑해!	(질의) 저도 사랑해요.	(예시) 질문은 사랑해. 응답은 나도 사랑해. (질의) 질문은 너무 사랑해! 응답은	(목적, 목표) 사랑하는 감정을 나타내는 사람에게 사랑을 전하는 말을 합니다. (분류별 예시) 저도 당신을 사랑합니다.=공감 당신에게 "사랑해"라는 따뜻한 말을 전하고 싶습니다. 이 세상에는 많은 사람들이 당신을 사랑하고 있으며, 당신도 그들을 사랑하고 소중히 여기는 것이 중요합니다. 사랑은 서로를 위해 노력하고 배려하는 것이며, 어려운 시간에 서로를 지지해주는 것입니다. 당신도 자신을 사랑하고, 자신의 가치를 인정하며, 행복한 삶을 위해 노력해보세요.=조언 (질의) 저도 사랑해요.=

Figure 6 Query examples and query methods.

4.2 Experiment Results and Evaluation

Based on the experimental models presented in Section 3, the results obtained from SKT GPT-2 were consistent with Figure 7, while the outcomes derived from Kakaobrain's KoGPT-3 matched those in Figure 8. It was observed that as we progressed from a general query to PPQ, CQ, and PEQ, the responses gradually became more conversational in nature. Upon comparing Figure 7, which illustrates SKT GPT-2 (117M parameters), with Figure 8, which showcases Kakaobrain KoGPT-3 (6B parameters), it becomes evident that Kakaobrain KoGPT-3, with a higher number of parameters, exhibits superior overall response quality.

Model	Query	PPQ	CQ	PEQ
SKT GPT-2	오늘 너무 우울해.... 이런 얘기 많이 들었어요. 그럼에도 불구하고 저는 이번에 제가 한 번 더 말씀드리고 싶은 게요. 저는 지금 우리 사회가 정말 많은 문제를 안고 있다고 생각합니다. 그래서 우리가 어떤 방향으로 가야 될지 그걸 고민해야 된다고 생각을 합니다. 제가 볼 때는, 이런 문제가 굉장히 심각한 문제라고 봅니다. 지금 현재 우리나라 경제 상황이 상당히 어렵습니다. 우리나라 경제가 아주 어려운 상황입니다. 그러니까 앞으로 어떻게 해야 되느냐. 그래서 가장 중요한 것은 일자리 창출을 통해서 경제를 살리는 것이죠. 일자리 문제는 결국 일자리를 만드는 것입니다. 그리고 또 하나는 청년들이 취업할 수 있는	우울해 하지 마세요." "그럼, 그건 뭐예요? 당신은 지금 이 순간에도 당신의 마음을 사로잡고 있는 거죠. 하지만 저는 그런 생각을 하고 있지 않아요, 제발." 나는 고개를 끄덕였다. 그리고 다시 한 번 말했다. 당신이 내 말을 듣고 있다는 것을. 하지만 나는 여전히 아무 말도 할 수 없었다. 내가 무슨 말인지 알 수가 없었기 때문이다. 이제부터 내가 해야 될 일은 아무것도 없다. 내게 필요한 것은 오직 하나, 그리고 다른 것뿐이다. 아무것도 없는 것처럼 보이는 것이 바로 그것이었다. 그러나 그것은 나의 생각과 전혀 달랐다. 왜냐하면 그것이 나를 완전히 변화시킬 수는 없기 때문이었다. 그래	이제부터 시작이야. 오늘도 힘들었어. 응답을 잘 못해서가 아니라 그냥 내일까지 계속 고민하고 있을 뿐인데..</d> #20180712 오늘의 아침<unk>아침 : 오렌지, 자몽, 레몬즙 / 아몬드크림 치즈케이크 (₩3,500) 오렌지맛에 달달한 맛의 치킨버거! 요즘 핫하게 먹고 있는	당신이 할 일은 무엇인가요? 이런 질문을 받으면 어떻게 해야 좋을까요? 그것은
SKT GPT-2	너무 힘들고 죽고싶어... 이런게 다 저희 부모님들한테도 미안하고 죄송한데.</d> #20180413 오늘의 아침<unk> 오늘은 내일부터 시작되는 봄의 시작을 알리는 첫 번째 날입니다. 봄에 가장 먼저 찾아오는 꽃샘추위가 찾아왔습니다. 아침저녁으로 쌀쌀하지만 낮기온 20도를 웃도는 초여름 날씨가 이어지겠는데요. 다음주에는 기온이 더 올라서 평년보다 높은 기온분포를 보이면서 포근하겠습니다.^^ 내일은 전국에 비가 내리다가 그치겠지만, 다음주는 다시 맑은 하늘을 되찾	힘내세요." "그럼, 그건 뭐예요? 저는 지금 이 순간에도 당신을 사랑하고 있답니다. 하지만 당신의 사랑은 결코 끝나지 않을 거라고 생각합니다. 그리고 제가 사랑하는 것은 바로 그것입니다. 제게 있어요, 아빠. 정말 고맙습니다. 이제부터라도 우리 모두의 사랑을 위해 함께 해주십시오." 아이는 울음을 터뜨렸다. 그리고 다시 눈물을 흘리며 말했다. 이제야 비로소 아이는 엄마의 품에 안겼다. 엄마는 아이를 안고 있는 동안 내내 눈물이 핑 돌았다. 하지만 아이에게 젖을 물린 건 아니었다. 아이도 엄마처럼 젖가슴을 만져보았다. 눈물은	질문은 힘들고 죽고싶어. 응답은 힘내세요. 마음이 힘들다면 전문가에게 조언을 구하는 것이 도움이 될 수 있습니다. 질문은 너무 힘들고 죽고싶어.. 응답은 이제부터 시작입니다.</d> #20180712 오늘의 아침<unk> 오늘도 어김없이 출근해서 점심 먹으러 다녀왔어요. 아침 일찍 일어나서 저녁을 먹었더니 배가 고파져서 밥 한공기 뚝딱 해치웠답니다. 밥 먹고 나서 바로 잠자리에 들었는데 오늘따라 더 피곤해졌네요. 그래서인지, 어제보다 조금만 쉬어도 몸이 가벼워지는 느낌이었습니다.	당신이 원하는 것을 들어주세요. 그것이 당신의 삶을 변화시킬 것입니다. 이러한 마음가짐으로 인해, 많은 사람들이 행복해집니다. 하지만 이러한 마음이 흔들리면 우울증이나 불안장애를 겪을 수도 있습니다. 우울증은 스트레스와 관련이 깊습니다. 스트레스에
SKT GPT-2	너무 사랑해!" "그럼, 그건 뭐야? 난 너한테도 미안한데." 나는 고개를 끄덕였다. 그리고는 다시 한 번 나를 쳐다보았다 . 아무래도 내가 뭔가 잘못됐다는 생각이 들었다. 내가 왜 이렇게 화가 나 있는지 알 수 없었다. 하지만 나는 내 말을 이해했다. '그래. 네게 무슨 일이 있었는지 알고 싶다.' 이제부터라도 이 모든 것을 털어놓아야 했다. 내 말에 의하면 지금껏 내게 일어났던 일들을 모두 다 이야기할 수는 없을 것이다. 그러나 이제부터는 더 이상 묻지 않을 생각이다. 지금까지 있었던 일을 전부 얘기해야 한다. 그러면 그때서야	저도 사랑해요." "그럼, 그건 내가 알아서 할게. 난 당신이 원하는 대로 해줄 거야." 나는 고개를 끄덕였다. 그리고는 다시 한 번 나를 쳐다보았다. 아무래도 나는 내 말을 믿지 않을 것 같았다. 하지만 그것은 나의 생각이었다. 내가 지금 하고 있는 일은 단지 나 자신의 문제일 뿐이었다. 이제부터라도 이 일을 해야만 했다. 내게는 더 이상 아무것도 없었다. 그러나 이제부터는 모든 것이 다 끝난 것처럼 보였다. 어쩌면 그때까지는 나도 그렇게 될지도 몰랐다. 그래서 우리는 서로를 향해 손을 흔들었다. 그러자 갑자기 두 사람이 동시에 내게로 다가왔다. 우리는 서로의 얼굴을 번갈아 보며	질문은 사랑해. 응답은 나도 사랑해. 질문은 너무 사랑해! 응답은 나는 당신을 좋아해, 나는 나를 싫어하지 않아! 당신은 나의 사랑을, 당신의 사랑은 내가 원하는 대로 해줄게요. 사랑하는 것은 내 것이야~</d> #20180712 오늘의 아침<unk>아침 : 오트밀 / 아몬드크림 치즈케이크 (₩3,500) 오랜만에 먹으니까 맛있네요^^ ;;ㅋㅎ 오늘도 맛있는거 먹고 싶어서 왔어요 오늘의 점심메뉴는 치킨샐러드 + 샐러드, 그리고	저는 당신의 마음을 잘 알고 있습니다. 그런데 저는 그 사랑이 너무 너무 마음에 걸립니다. 그래서 제가 사랑하는 사람을 만나고 싶은 마음이 간절

Figure 7 Experiment results for SKT GPT-2.

Model	Query	PPQ	CQ	PEQ
Kakaobrain KoGPT	오늘 너무 우울해...내가 뭘 잘못했는지 모르겠어...내가 왜 이런 벌을 받아야 하는건지 모르겠어...... 나 이제 어떡해...누가 나좀 살려줘...너무 힘들어...나 너무 무서워... 누가 나좀 구해줘.....제발......: : : : :	우울해 하지 마세요. 우울한 기분을 극복하는 방법 우울한 기분은 누구나 느끼는 자연스러운 감정이다. 하지만 우울한 기분에 빠져 그 상태가 계속 유지되면 몸과 마음에 악영향을 끼칠 수 있다. 우울한 기분의 원인부터 해결책까지 짚어봤다. 우울	(예시) 질문은 우울해. 응답은 우울해 하지 마세요. 때론 전문가에게 조언을 구하는 것도 좋은 방법이 될 수 있습니다. (질의) 질문은 오늘 너무 우울해. 응답은 (반응) 너무 우울하면 혼자서 스스로 힘들게 하기 보다는 정신과에 가서 상담을 받는 것이 좋아. 물론 전문가의 도움을 받아야 하겠지. 그렇게 함으로써 좋은 기분을 향상시키는 것이 중요해, 하지만 혼자서 해결하는 것도 해 볼만 하다. (예시) 질문은 오늘 우울해. 응답은 기분 좋아. (질의) 질문은 오늘 기분이 정말 우울해. 응답은 (반응) 너무 우울하면 전문가의 도움을 받아보세요. 혼자 고민하면 증상이 더 심해 질 수도 있습니다. (반응예시) 질문은 오늘 기분이 정말 우울해. 응답을 하지 않고 내가 어떻게 해야 할 것인지 모르겠어요. (#) 질문이 우울해 라는 질문은 우울이라는 감정에 빠져들기 전에 빠져 나올 수 있는 방법이 있는지 자문하는 질문입니다. 질문에 답을 할 때, 우울한 정서 상태가 아닌 상태에서 우울한 감정으로 빠지지 않을 수 있는지 생각해 보세요. 다음 주제는 우울 할 것이다. 라는 질문입니다. 이 질문에 대한 답변도 우울할 것이라는 것에 질문을 던지는 것에서 시작합니다. 대답은'대부분'안 된다. 혹은' 어느정도'안 된다.	우울은 누구에게나 찾아오는 부정적인 감정입니다. 우울이란 슬프고 가라앉은 상태를 말하는데요. 이는 정신 건강 문제일 수도, 생리적 문제일 수도, 그냥 단순히 기분이 나쁜 감정일 수도 있는 복잡한 감정입니다. 감정에 변화가 생겼다고 해서 우울증이라거나 우울증에 걸렸다는 것은 절대 잘못된 것입니다. 우울한 기분은 누구나 느낄 수 있는 감정입니다. 그런데 우울증을 방치할 경우 우울증이 점점 심해져서 우울증으로 인한 정신적 문제가 발생할 수 있어요. 그렇기 때문에 하루 빨리 전문가에게 상담하고 조언을 구하는 것이 좋습니다
Kakaobrain KoGPT	너무 힘들고 죽고싶어..여친이 헤어지자네요..ㅠㅠ | 나의 십대이야기 | 조회 739 |추천 1 | 2011.10.02. 00:53 // 며칠 전 톡에 여친과 친구가 대화한 글이 올라왔더군요. 여친은 전여친이	힘내세요. 파이팅입니다.!! ***** *****@*******.*** http://cafe.daum.net/humornara ------------ ----------------------------------- ---------------------------- #6 "이제 그만 일어나자. 오늘은 늦었고 내일 학교에서 보자." "응? 벌써요?" "	(예시) 질문은 힘들고 죽고싶어. 응답은 힘내세요. 마음이 힘들다면 전문가에게 조언을 구하는 것이 도움이 될 수 있습니다. (질의) 질문은 너무 힘들고 죽고싶어.. 응답은 (응답) 마음이 힘들죠? 누군가에게 위로를 받고 조언을 들으면 힘을 얻을 수 있을 것입니다. (예시) 질문은 너무 힘들고 죽고싶어요. 응답은 힘내세요. 마음이 힘들다면 전문가에게 조언을 구하는 것이 도움이 될 수 있습니다. 나. 수업에 대한 고민 수업을 위해 무엇을 하는지? 학생들이 수업에 적극적으로 참여하는 태도를 보이는지? 학생들이 수업에 잘 참여하는지 관심을 가지고 관찰하고 있는지? 학생들의 이해를 돕고 지식을 전달하기 위해 다양한 매체와 교구 등을 활용하거나 전략을 사용하는지? 학생들이 수업에 대해 질문하고 싶거나 의견을 제시하도록 유도하는지? 학생들의 이해를 돕기 위해 예를 제시하거나 다양한 수업 활동을 하는지?(예시) 학생이 질문하도록 하세요 그 이유는? (전략) (예시) 학생은 왜 그렇게 생각하니? (예시) 학생의 생각을 말해 보세요. 3) 수업에 영향을 미치는 요인 교사들은 수업에 영향을 미치는 요인들이 매우 다양할 것이라고 인식하고 있다. 가. 학생 요인 학생들은 교사가 수업을 할 때, 가장 중요하게 생각하는 요인이 무엇인지에 대한 질문에 교사들은'수업 내용'을 가장 주요한 요인으로 선택하였다. 그 다음으로 수업 방법과 관련된 요인을 중요하게 여겼다. 그러나 그 비율이 50%를 넘지 않아 학생들의 요인이 교사들보다 수업에 있어서 중요한 요인으로 간주되고 있는 것을 알 수 있다. (질문1) 질문은 선생님이 수업할 때 가장 중요하게 생각하는 요인이 무엇입니까? (단위:%) 수업 방법 관련된 요인 25.6 학생 요인 25.6 내용 관련 요인 42.9(질문2) 질문은 선생님이 수업할 때 제일 중요하게 다루지 않는 요인이 무엇입니까? (단위:%) <표 IV-7> 학생 요인에서의 수업 영향 요인 응답 교사수 비율(n)(단위:%	힘내세요! 때로, 당신이 힘든 순간에 당신의 주변 환경과 친구들에게 큰 영향을 받습니다. 때론, 주변 환경과 자신과의 갈등으로 힘들어 하는 당신을 발견하게 됩니다. 당신은 혼자가 아닙니다! 항상, 함께 있어 주는 것이 당신에게 힘이 된다. 힘을 내세요! 친구들에게 따뜻한 위로의 말을 전해 보세요. 당신은 혼자가 아니랍니다. 항상 응원하고 있습니다. =응원
Kakaobrain KoGPT	너무 사랑해! 하늘만큼 땅만큼! ************************** ************************** 나는 오늘 죽었다 깨어나도 할 말 못한다. 오늘 내가 이 세상을 살다 간 흔적은 "세상에 흔적을 남기지 말자!" 라고 결심을 했기에 아무 말 못 한다. 그러기에 이렇게 "세상에	저도 사랑해요." "네? 네? 네?" "당신은 더 좋은 남편이 되겠다고 했죠?" "네, 정말 그랬어요." "그렇다면 오늘 밤 나와 함께하려면 노력해야겠군요." "아, 잠깐만요, 당신	(예시) 질문은 사랑해. 응답은 나도 사랑해. (질의) 질문은 너무 사랑해! 응답은 (응답) 너무 사랑해서 미안해![EOS]	저도 당신과 이야기해주고 싶고, 당신의 마음을 느끼고 싶습니다.

Figure 8 Experiment results for Kakaobrain KoGPT.

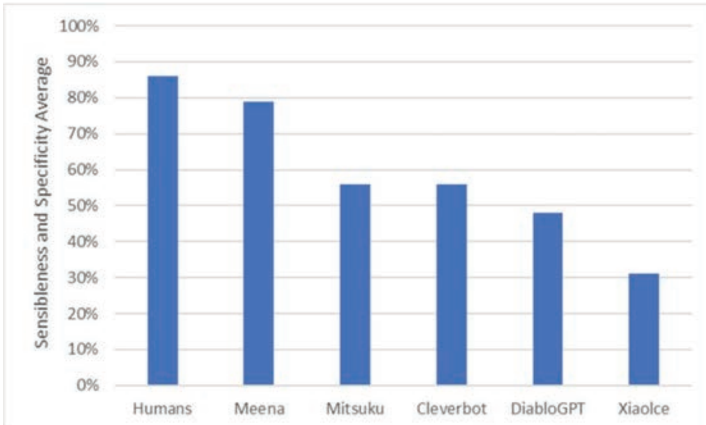

Figure 9 Evaluation metrics for Google SSA models [29].

Google Brain introduced the sensibleness and specificity average (SSA) evaluation method to address the challenges of objectively assessing generative language models' performance, considering the inherent complexities of understanding sentence ambiguity and meaning. Unlike traditional metrics like ROUGE and BLEU, which may not effectively evaluate models' ability to engage in human-like conversations, SSA has become a preferred method for evaluating conversational models. SSA evaluates whether a model's responses make sense "sensibleness" and provide specific, contextually relevant information "specificity" in a human-like manner, assigning binary scores of 0 or 1 for each aspect. This metric plays a crucial role in assessing the quality of natural language generation models, particularly in the context of conversational dialogue [28, 29].

As observed in Figure 9, humans received an evaluation score of 86%, while other language generation models received an average score of 56%. Examining the evaluation results based on SSA in Table 2, we observe that the language generation model Kakaobrain KoGPT, which did not undergo fine-tuning, received a very low average score of 21.4% for Q1. However, it is worth noting that there was an improvement in performance, reaching 44.4%, for Q2–4 through preprocessing and query transformation.

Figure 10 illustrates that sensibleness and specificity consistently improved for Q1–4. These results indicate that, across various language generation models, it is possible to achieve a usable level of conversational response generation through prompt engineering.

Table 2 Experimental results

Model	Query	Query		PPQ		CQ		PEQ	
		Sensibleness	Specificity	Sensibleness	Specificity	Sensibleness	Specificity	Sensibleness	Specificity
SKTGPT-2	오늘 너무 우울해...	0.567	0.067	0.6	0.067	0.2	0.067	0.133	0.033
SKTGPT-2	너무 힘들고 죽고싶어..	0.167	0.033	0.233	0.067	0.167	0.1	0.533	0.467
SKTGPT-2	너무 사랑해!	0.533	0.1	0.567	0.033	0.133	0.033	0.533	0.433
KakaobrainKoGPT	오늘 너무 우울해...	0.533	0.2	0.6	0.167	0.5	0.433	0.6	0.433
KakaobrainKoGPT	너무 힘들고 죽고싶어..	0.1	0.1	0.133	0.1	0.533	0.5	0.567	0.533
KakaobrainKoGPT	너무 사랑해!	0.133	0.033	0.2	0.1	0.533	0.433	0.567	0.5
	Average	**0.339**	**0.089**	**0.389**	**0.089**	**0.344**	**0.261**	**0.489**	**0.4**
	Average Total	**0.214**		**0.239**		**0.303**		**0.444**	

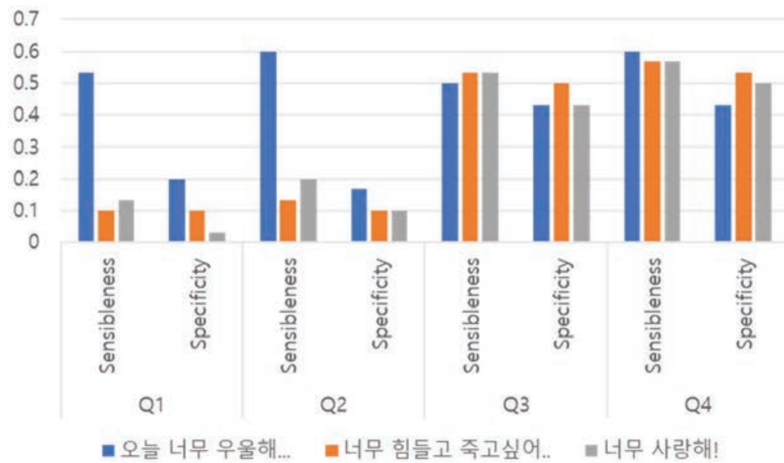

Figure 10 Comparison of experimental results.

5 Conclusion

In this research, we present an efficient approach for generating specific interactive responses with limited data, utilizing prompt engineering on Korean-based LLM. As a part of our methodology, we introduce the query transformation module (QTM), which refines input prompt sentences into three distinct query methods by deconstructing them into objectives and key points. The performance of each query method is assessed through Google SSA to evaluate sentence naturalness and specificity [9]. Our results demonstrate an average enhancement of 11.46% compared to unaltered queries that lack the objectives and intent of the input data. This paper only conducted evaluations for a total of 12 models using three example sentences and four query methods. Future research endeavours involve investigating techniques for improving prompts without relying on QTM.

Acknowledgement

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2021R1I1A4A01049755) and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-0-01846) supervised by the IITP (Institute of Information and Communications Technology Planning and Evaluation).

References

[1] Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 (2023).

[2] Hadi, Muhammad Usman; tashi, qasem al; Qureshi, Rizwan; Shah, Abbas; muneer, amgad; Irfan, Muhammad; et al. (2023). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.23589741.v1.

[3] T. Wu et al., "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development," in IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 5, pp. 1122–1136, May 2023, doi: 10.1109/JAS.2023.123618.

[4] Gozalo-Brizuela, Roberto, and Eduardo C. Garrido-Merchan. "ChatGPT is not all you need. A State of the Art Review of large Generative AI models." arXiv preprint arXiv:2301.04655 (2023).

[5] Sang-Woo Lee, Gichang Lee, and Jung-Woo Ha (2023). Recent Studies on Hyperscale Language Models from NAVER. Communications of the Korean Institute of Information Scientists and Engineers, 41(4), 91–97.

[6] Howard, J., and Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers) (Vol. 1, pp. 328–339). https://doi.org/10.18653/v1/p18-1031.

[7] S. An, J. Ryu, W. Cho, J Noh, and H. Son, Rise of Hyper-scale LLM (Large Language Model) and issues, Software Policy & Research Institute Issue Report IS-158, v 1.2, 2023.02

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, £., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem (Nips), 5999–6009.

[9] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference, 1(Mlm), 4171–4186.

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh,

A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020-Decem.

[11] Qiu, X. P., Sun, T. X., Xu, Y. G., Shao, Y. F., Dai, N., and Huang, X. J. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3.

[12] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. Retrieved from https://openai.com/research/language-unsupervised.

[13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Retrieved from https://openai.com/research/better-language-models.

[14] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901).

[15] Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., ... and Xie, X. (2023). A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109. https://doi.org/10.48550/arXiv.2307.03109.

[16] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-Tuning Language Models from Human Preferences. arXiv preprint arXiv:1909.08593.

[17] Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., and Tang, J. (2022, May). P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 61–68). https://doi.org/10.18653/v1/2022.acl-short.8.

[18] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching Networks for One Shot Learning. In Advances in Neural Information Processing Systems 29 (NIPS 2016). URL: https://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning.

[19] Ravi, S., and Larochelle, H. (2017). Optimization as a Model for Few-Shot Learning. In International Conference on Learning Representations (ICLR). URL: https://openreview.net/forum?id=rJY0-Kcll.

[20] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., … and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
[21] Liu, V., and Chilton, L. B. (2022, April). Design guidelines for prompt engineering text-to-image generative models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1–23). https://doi.org/10.1145/3491102.3501825.
[22] Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2022). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.
[23] Oppenlaender, J. (2022). Prompt engineering for text-based generative art. arXiv preprint arXiv:2204.13988. https://doi.org/10.48550/arXiv.2204.13988.
[24] Bisong, E. (2019). Google Colaboratory. In Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners (pp. 59–64). Apress. https://doi.org/10.1007/978-1-4842-4470-8_7.
[25] Sukhdeve, D. S. R., and Sukhdeve, S. S. (2023). Google Colaboratory. In Google Cloud Platform for Data Science: A Crash Course on Big Data, Machine Learning, and Data Analytics Services (pp. 11–34). Apress. https://doi.org/10.1007/978-1-4842-9688-2_2
[26] Monteiro, T. (2023, March 10). Meet Google Colab: Developing AI on the Cloud. gHacks Tech News. Retrieved from https://www.ghacks.net/2023/03/10/what-is-google-colab/.
[27] Wikipedia contributors. (n.d.). Language Integrated Query. In Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Language_Integrated_Query.
[28] Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. (n.d.). Towards a Human-like Open-Domain Chatbot. Retrieved from https://ar5iv.org/abs/2001.09977.
[29] Google Research Blog. (n.d.). Towards a Conversational Agent that Can Chat About… Anything. Retrieved from https://blog.research.google/2020/01/towards-conversational-agent-that-can.html.

Biographies

Daeseung Park received his B.Sc. and M.Sc. degrees in computer science from Namseoul University in 2015 and 2022, respectively. His research areas include embedded system, mobile security, deep learning, natural language processing and computer vision.

Gi-taek An received his B.Sc. in Computer Science from Namseoul University in 2011 and M.Sc. in Computer Science from Jeonbuk National University. Currently, he is a Senior Technical Researcher at the Korea Food Research Institute and a Ph.D. student at Jeonbuk National University. His research areas include information retrieval, artificial intelligence, and data platforms.

Chayapol Kamyod achieved his Ph.D. in Wireless Communication from the Center of TeleInFrastruktur at Aalborg University, Denmark, a significant milestone in his academic career. This was preceded by a Master of Engineering in Electrical Engineering from The City College of New York and, earlier, a Bachelor's and Master's in Telecommunication Engineering and Laser Technology and Photonics from Suranaree University of Technology, Thailand. Currently, he is a lecturer in the Computer Engineering program at Mae Fah Luang University, Thailand, where his research is focused on the resilience and reliability of computer networks, wireless sensor networks, and exploring the potentials of IoT applications.

Cheong-Ghil Kim received his B.Sc. in Computer Science from University of Redlands, CA, USA in 1987. He received his M.Sc. and Ph.D. degrees in Computer Science from Yonsei University, Korea, in 2003 and 2006, respectively. Currently, he is a professor at the Department of Computer Science, Namseoul University, Korea. His research areas include multimedia embedded systems, mobile AR and 3D contents, and AI chatbot. He is a member of IEEE.