Journal of KIIT. Vol. 22, No. 11, pp. 167-180, Nov. 30, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.11.167

비즈니스 데이터 분석 리포트 생성을 위한 생성형 AI 프롬프트 엔지니어링 연구

박예은*, 김동훈**, 유시현***, 배윤진****, 김경외*****

Reseach on Prompmt Engineering for Generative AI to Create Business Data Analysis Reports

Yeeun Park*, DongHun Kim**, Sihyeon Yoo***, Yunjin Bae****, and Keungoui Kim*****

이 논문은 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2023S1A5A2A21086671)

요 약

본 논문에서는 퍼포먼스 마케팅을 중심으로 비즈니스 데이터 분석 리포팅을 최적화하기 위한 생성형 AI 기반 프롬프트 구조를 제안하고자 한다. 제안된 방식에는 퍼포먼스 마케팅 주요 성과 지표로 구성된 일일 리포팅 데이터셋을 전처리 과정과 테이블 형식의 데이터를 자연어 텍스트로 변환하는 'Table-to-Text' 방식이 포함되었다. 해당 도메인에서 주로 활용되는 질의에 대한 적정한 답변을 생성할 수 있도록 다양한 프롬프트 기법을 조합하고 검토하였으며 ROUGE Score와 자체 체크리스트를 활용한 평가 결과를 토대로 최적의 프롬프트를 선정하였다. 본 논문은 퍼포먼스 마케팅 분야에서의 프롬프트 엔지니어링을 연구한 점에서 기존 논문들이 다루지 않았던 분야를 탐구했다. 특히, 테이블 데이터 해석과 데이터의 주요 포인트를 반영한 코멘트를 동시에 생성하는 프롬프트 구조를 제시함으로써, 기존 연구와 차별화되는 접근 방식을 제안한다. 본 논문에서 제안한 프롬프트를 통해 퍼포먼스 리포트 작성시의 객관성을 높이고 시간 및 비용을 절약시킬 수 있을 것으로 기대된다.

Abstract

This study proposes a generative AI-based prompt structure to optimize business data analysis reporting, focusing on performance marketing. The proposed method includes preprocessing daily reporting datasets composed of key performance indicators in performance marketing and converting table-formatted data into natural language text through a 'Table-to-Text' approach. Various prompt techniques were combined and reviewed to generate appropriate answers to commonly used queries in this domain. The optimal prompts were selected based on evaluation results using ROUGE scores and an in-house checklist. This paper explores prompt engineering in the performance marketing field, an area that has not been extensively studied in previous research. It distinguishes itself by proposing a prompt structure that simultaneously interprets table data and generates comments that reflect the key points of the data, offering a novel approach compared to prior studies. It is expected that the proposed prompts will enhance objectivity in performance report writing and save time and costs.

Keywords

generative AI, prompt engineering, table-to-text, performance marketing, report generation, KPI, business data report

* 한동대학교 생명과학부 학사
- ORCID: https://orcid.org/0009-0001-2108-1845
** 한동대학교 상담심리사회복지학부 학사
- ORCID: https://orcid.org/0009-0006-4521-3757
*** 한동대학교 ICT 창업학부 학사과정
- ORCID: https://orcid.org/0009-0004-0947-2175
**** 한동대학교 경영경제학부 학사과정
- ORCID: https://orcid.org/0009-0005-8836-1568
***** 한동대학교 AI 융합교육원 조교수(교신저자)
- ORCID: https://orcid.org/0000-0002-0839-8813

· Received: Aug. 27, 2024, Revised: Oct. 15, 2024, Accepted: Oct. 18, 2024
· Corresponding Author: Keungoui Kim
School of Applied Artificial Intelligence, Handong Global University
Tel.: 82+54-260-1125, Email: awekim@handong.edu

Ⅰ. 서 론

생성형 AI 기술의 발전은 다양한 산업 분야에서 새로운 기술과 서비스의 등장을 야기하였다[1]. 특정 산업 분야에 국한되지 않는 범용적인 기술로서 특히 대량의 데이터를 정기적으로 관리하고 분석하는 작업에서 보여준 놀라운 성과는 많은 이들의 이목을 끌었다[2]. 생성형 AI를 비즈니스에 적용하고자하는 산업계의 지속적인 시도는 기존 업무 자동화 방식과 결합되어 새로운 형태의 업무 환경 조성으로 이어지고 있다. 이러한 광범위한 생성형 AI의 적용과 확장은 일반 사무 업무 환경에 적용될 경우 업무 효율성을 크게 향상시킬 것으로 예상된다[3].

확률적 관계를 기반으로 답변을 생성해내는 생성형 AI는 사용자의 질문에 가장 그럴듯한 대답을 만들어낼 수 있어 누구라도 전문 지식 없이 쉽게 사용할 수 있다는 장점이 있다[4]. 다만, 생성형 AI를 별도의 기법 없이 사용할 경우 사전에 학습된 모형이기 때문에 학습 외 영역에 대한 답을 생성해내는 것이 어렵고, 사업 환경에서 필요로 하는 적합한 결과를 일관성 있게 도출해낼 수 없기 때문에 이를 극복하기 위한 방안이 모색되어야 한다[5]. 프롬프트 엔지니어링은 생성형 AI의 출력문을 보다 효과적으로 얻기 위해 고안된 방법[6]으로 입력 프롬프트를 조정하여 사용자가 원하는 결과를 생성하도록 설계하는 것이다. 프롬프트 엔지니어링을 통해 연구자는 입력 구조에 따라 출력 구조와 그 내용을 정형화할 수 있으며, 이를 통해 생성형 AI가 만들어내는 출력의 정확성을 높일 수 있다[7]. 이처럼 생성형 AI를 효과적으로 활용하기 위해서는 사업 분야의 특성과 전문성을 반영한 정교한 입력 프롬프트 설계가 필요하다[5].

본 논문에서는 퍼포먼스 마케팅 분야에서 사용되는 KPIs(Key Performance Indicators) 데이터를 바탕으로 데이터 분석 리포팅 업무 최적화를 위한 프롬프트를 제안하고자 한다. 기존 연구에서는 주로 보편적인 프롬프트 설계 방법론을 다루고 있으나 이는 분야별 특성을 고려하지 못한다는 한계가 있다[8]. 이에 퍼포먼스 마케팅 분야의 KPI 데이터를 활용하여 해당 분야에 특화된 프롬프트를 제안한다. 제안된 프롬프트 구조에서는 KPI 데이터가 주로 테이블 형식으로 가공된다는 점에 착안하여 테이블 형식의 데이터를 자연어 텍스트로 변환하는 'Table-to-Text' 방식을 활용하고[9], 프롬프트의 순서, 예시 답변 형식, 입력 데이터 형식 등의 변수를 고려하고자 하였다[10]. 최종적으로 도출된 프롬프트의 성능은 체크리스트를 기준으로 답변의 질과 사용성을 평가하는 정성적 방식과 텍스트 유사성을 정량적으로 계산하는 ROUGE score를 사용해 성능을 종합적으로 평가하고자 하였다[11]. 본 논문을 통해 제안된 프롬프트가 퍼포먼스 마케팅 분야의 데이터 분석 리포팅 업무에 실질적인 기여를 할 수 있기를 기대한다.

Ⅱ. 이론적 배경

2.1 생성형 AI

생성형 AI란 학습 데이터로부터 텍스트, 이미지 또는 오디오와 같은 새로운 의미 있는 콘텐츠를 생성할 수 있는 인공지능 기술을 지칭한다. ChatGPT의 성공적인 등장 이후, 생성형 AI는 비즈니스, 교육, 의료 및 콘텐츠 생성을 포함한 다양한 산업에 응용되고 있다[12].

생성형 AI, 특히 언어모델 기반의 ChatGPT는 업무 효율성을 크게 향상시킬 수 있다. 많은 초기 사용자와 관리자들은 생성형 AI가 메시지 작성, 보고서 작성, 아이디어 생성 등의 일상적인 작업을 자동화하여 작업의 질과 생산성을 높인다고 보았다[13].

2.2 프롬프트 엔지니어링

기존의 생성형 AI는 사전학습된 모형이기 때문에 새로운 정보나 특정 분야에 특화된 작업을 수행하기 위한 목적으로는 주로 파인튜닝이나 프롬프트 엔지니어링 기법이 사용된다. 파인튜닝은 사전에 훈련된 모델을 기반으로 해당 작업에 특화된 데이터 세트에서 모델의 매개 변수를 조정하는 과정이다[14]. 프롬프트 엔지니어링은 매개 변수를 조정하는 대신 사용자가 LLM 모델에 자신의 의도를 효율적으로 전달할 수 있도록 돕는 프롬프트의 설계, 개선 및 최적화 과정이다[15].

기존의 연구에서는 프롬프트 엔지니어링의 용어 및 개념을 표준화하거나 성능 개선 차원에서의 프롬프트 연구가 주를 이루고 있다. 현재까지 진행된 연구를 살펴보았을 때 가장 보편적으로 많이 사용되는 프롬프트 엔지니어링 기법들은 다음과 같다. 먼저, 생성형 AI에게 특정 예시를 제공하여 답변을 유도하는 방식으로 Zero-Shot와 Few-Shot이 있다[16]. Zero-shot은 모델이 학습 과정에서 본 적 없는 새로운 클래스를 인식할 수 있도록 하는 학습 방법이다[17]. Few-Shot은 소수의 입력-출력 예시를 제공함으로써 모델이 특정 작업을 이해하도록 돕는 기법이다. 이 프롬프트 기법들은 간단히 사용할 수 있지만, 예시가 많아질수록 더 많은 토큰이 필요하다는 단점이 있다. 또한 예시의 선택과 구성에 따라 모델 성능과 편향에 영향을 줄 수 있다[18]. 따라서 복잡한 추론이나 사고를 요하는 문제에는 적합하지 않다. 복잡한 추론이 필요한 작업을 처리할 때는 주로 Chain-Of-Thought 방식이 사용된다[19]. 연속된 질문을 던지는 방식을 통해 사용자는 생성형 AI가 일관되고 단계적으로 추론할 수 있도록 유도할 수 있다[16]. 이 외에도 예제를 수동으로 만들어 제시하는 Automatic Chain-Of-Thought[20], Tree 구조를 관리하여 추론하는 Tree-of-Thought[21] 등 여러 가지 추론 프롬프트 기법들이 있다. 이들을 통해 복잡한 문제를 다룰 수는 있지만, 질문 구성 및 반복에 따라 성능 편차가 발생할 수 있으며, 자연어 질문의 모호성으로 인해 성능 변동이 생길 수 있다[17]. 이를 해결하기 위해 기호를 사용하는 CoS(Chain-of-Symbol) 기법이나[16], 컴퓨팅 언어 문법 형식으로 프롬프트를 구조화하는 시도가 이루어지고 있다[22].

2.3 프롬프트 엔지니어링 활용

LLM의 성능을 최적화하기 위한 방법 중 하나로 프롬프트 엔지니어링이 주목받고 있다. 프롬프트 엔지니어링은 사용자가 원하는 출력 결과를 얻기 위해 AI에게 제공하는 입력 형식을 설계하는 과정으로, 모델의 성능을 크게 좌우할 수 있다. 예를 들어, 인공지능 모델의 밴치마크 성능 평가에서 프롬프트 엔지니어링을 사용하여 성능을 향상시킨 연구가 있다[23]. 이 연구는 모델의 구조적 변경 없이, 단순히 입력 형식의 변화를 통해 모델의 성능을 크게 개선할 수 있음을 보여준다.

또한, 별도의 모델 학습 과정 없이도 한국어 평가 문제를 생성하는 연구가 있으며[24][25], 이는 적절한 프롬프트만으로도 데이터 생성 작업을 효율적으로 수행할 수 있다는 점을 시사한다. 이 연구는 특히 비정형 데이터를 효과적으로 처리하는 데 있어 프롬프트 엔지니어링이 유용한 도구임을 보여준다.

서술형 수학 문제의 자동 채점을 위한 연구에서도 프롬프트 엔지니어링이 중요한 역할을 하였다[26]. 이 연구는 AI 모델을 활용하여 복잡한 문제의 해석 및 평가를 수행하는데, 적절한 프롬프트 설정을 통해 그 정확성과 일관성을 높였다.

이러한 선행 연구들은 프롬프트 엔지니어링이 AI 모델의 다양한 응용 분야에서 핵심적인 역할을 한다는 점을 명확히 보여주며, 모델의 특성에 맞는 도메인 특화 프롬프트 설계가 AI의 성능을 극대화하는 중요한 요소임을 시사한다.

III. 연구 방법

3.1 연구 데이터

본 논문에서는 퍼포먼스 마케팅 기업 PTBWA에서 제공받은 데일리 리포팅 데이터를 활용하였다. 퍼포먼스 마케팅 분야에서 사용되는 데일리 리포팅 데이터의 경우 퍼포먼스 마케팅의 성과를 측정하기 위해 생성된 자료이며, 본 연구에서는 보안상의 이유로 PTBWA에서 제공한 금융 회사 마케팅 더미데이터를 사용하였다.

자료 형식은 엑셀(.xlsx) 이며, 2023년도 04월, 5월, 6월의 데일리 리포팅 데이터를 사용했고, 월에 따라 다른 파일에 저장되어 총 3개의 엑셀 파일을 사용했다. 파일의 데이터는 포털 사이트 홍보 배너(구글, 네이버, 등)와 같은 디지털 홍보 캠페인별로 기록되었으며, 캠페인의 수에 따라 총 10여 개의 데일리 리포팅 데이터 시트가 있다. 각 시트의 행은 레이블(날짜)과, 일간 주요 지표값으로 구성되며, 각 월의 1일부터 말일까지가 행의 수이다.

시트의 컬럼은 레이블(날짜)과 퍼포먼스 마케팅의 주요 지표들(24개)로 구성되며, 주요 지표는 노출, 클릭, 클릭율, CPC, 광고비, 예금+대출율, 유입수, 방문자수, 신규방분, CPS, CPU, 신규방분CPU, 접수수, 심사수, 승인수, 예금+대출, 예금, 대출, 접수CPA, 심사CPA, 승인CPA, CPA, 예금CPA, 대출CPA 의 25개 컬럼으로 구성된다. 금융 회사 마케팅 자료이므로 대출, 예금, 등의 서비스로 이어진 것이 성과 지표이므로 해당 어휘가 주요 지표에 포함된다. 주요 지표들 중 핵심 지표로는 CPC(Cost Per Click), CPS(Cost Per Sales), CPA(Cost Per Action), CPU(Cost Per User)와 같은 핵심 지표들에 대한 정보를 담고 있다. CPC는 사용자가 광고를 클릭할 때마다 광고주가 지불하는 금액을 의미하며 CPS는 매출당 비용을 의미한다. CPA는 사용자가 특성 행동을 취할 때 지불하는 비용이며 CPU는 한 명의 사용자에게 도달하기 위해 지불하는 금액을 의미한다. 디지털 마케터의 주요 업무 중 하나는 퍼포먼스 마케팅에 대한 분석을 토대로 광고 캠페인의 성과를 정량적으로 평가하거나 어떤 전략이 효과적이었는지와 어떤 개선이 필요한지를 판단하는 것이다. 실제로 디지털 마케터는 데일리 리포팅 데이터를 활용해 전일 대비 성과 비교를 통해 개선 사항을 파악하고 진행 중인 광고의 비용 효율성을 검토하는 것뿐만 아니라 알 수 없는 예산 집행이나 입찰가 조정 등과 같은 운영사항을 포함한 종합적인 평가를 수행한다.

본 논문에서는 민감 정보가 마스킹 또는 임의 조정된 3개월 치 데일리 마케팅 데이터와 실제 디지털 마케터가 작성한 리포트 샘플이 활용되었다. 제공받은 데일리 리포트 예시는 다음과 같다(표 1).

표 1. 데일리 리포트 예시
Table 1. Daily report sample

3/31(금) 대비 4/1(토) 부로 예산 -19% 감소에도 유사한 심사수 유지하면서 전체 심사CPA -13% 개선

3.2 프롬프트 설계 방법

최적의 프롬프트 구조를 선정하기 위한 선행 연구 조사 결과, 컴퓨팅 언어의 문법 구조를 적용하면 LLM의 성능이 향상될 수 있음을 확인했다[27]. 해당 연구에서는 프롬프트의 구조와 내용을 유지하면서, 문어체로 표현된 내용을 컴퓨팅 언어로 변환할 경우 LLM의 명령 이해도가 높아지고, 출력의 정확도 또한 향상된다는 결과를 도출했다. 이를 바탕으로 여러 컴퓨팅 언어 포맷을 비교한 논문에서는 마크다운 형식의 프롬프트 구조가 가장 높은 성능을 보였다[28]. 따라서 본 연구에서는 선행 연구를 바탕으로 일본 Note 사의 CXO인 후카츠 타카유키가 개발한 후카츠 프롬프트가 연구의 최적의 프롬프트 구조로 선정했다[29]. 후카츠 프롬프트는 마크다운(Markdown) 형식을 활용하여 명령의 주요 사항들을 명시하는 템플릿으로 구성된다. LLM 모델이 Transformer 기반으로 설계됨에 따라, 프롬프트 입력 시에도 모델의 Attention을 유도하는 방식이 연구되었다[30]. 후카츠 프롬프트는 기본적인 업무를 설명하는 '명령문', 작업 과정이나 출력 결과에 반영되어야 하는 '제약조건', 입력되는 작업 데이터인 '입력문', 그리고 출력 형식이 지정된 '출력문'을 각각 마크다운 문법으로 강조하는 구조로 되어 있다. 이를 통해 LLM이 사용자가 강조하고자 하는 내용을 정확하게 답변에 반영할 수 있도록 유도한다. 기본 구조에 더하여 추가적인 중요 내용들을 마크다운 문법 형식에 맞춰 작성하여 강조할 수도 있다. 후카츠 프롬프트를 기반으로 프롬프트를 설계할 경우 디지털 마케터들이 리포트 작성 시 적용해야 하는 다양한 조건들에 맞출 수 있다는 장점이 있다.

최종적으로 설계된 Optimal 프롬프트의 구조는 다음과 같다. 먼저 마크다운 형식을 활용하여 분석하고자 하는 데이터와 주요 운영사항을 명시하였다. 그런 다음, 구체적인 페르소나를 설정하여 마케팅 전문가 관점에서 데이터를 분석/요약하게 하고, 그 이후에는 구체적인 명령을 전달하게끔 하였다. 이외 부분별 세부 사항을 살펴보면, #출력문 규칙 사항에서는 마케팅 데이터 분석의 결과를 출력할 때 우선으로 지켜야 할 사항들에 대해 명시하였다. 또한 지표별 변동이 특정 범위를 넘어갈 경우를 강조하기 위해 "-3% 이상 적어진 비용 관련 지표"라는 표현을 문장마다 반복하게끔 하였다. #제약 조건에서는 GPT의 전반적인 출력 형태에 대한 사항들을 명시하였다.

#출력문은 전체적인 명령 이후 GPT의 답변을 유도하기 위한 부분이므로, 제약 조건에서 지시한 한 불렛에 한 문장 형태로 대답을 받기 위해 두 개의 '-'(대시)를 활용하였다. 입력 및 출력의 예시를 제공하는 "One-Shot", "Few-Shot"과 같은 방법은 모델이 데이터의 증감 여부와 관계 없이 주어진 예시를 단순히 반복하는 경향이 있어 제외하였다.

표 2. 전반적인 프롬프트 구조
Table 2. Overall prompt sturcture

#Data #Operational Details #Commands (Including Persona) #Cost-Related Metrics #Output Statement Rules #Constraints #Output Statement [Campaign Name - Media Name] - -

3.3 평가 방법

설계된 프롬프트를 활용해 생성된 리포트를 종합적으로 평가하기 위해 ROUGE Score를 활용한 정량적 평가와 체크리스트를 활용한 정성적 평가를 진행하였다.

3.3.1 ROUGE

ROUGE Score는 생성된 텍스트가 참조 텍스트와 얼마나 유사한지를 측정하는 방법으로 특히 자연어 처리 분야에서 텍스트 요약의 성능을 평가하기 위한 용도로 사용된다[31]. ROUGE Score는 다양한 하위 지표로 구성되어 있으며, 각각의 지표는 요약문의 특정 측면을 평가하는 데 중점을 둔다. 본 논문에서는 Lin과 Hovy가 최초에 제시한 ROUGE 평가 방법론을 코드로 구현한 ROURGE 라이브러리를 활용하였다(https://github.com/pltrdy/rouge)[32].

ROUGE score는 여러 하위 지표로 구성되어 있는데, 그중 ROUGE-N과 ROUGE-L이 가장 널리 사용된다. ROUGE-N은 n-gram의 오버랩을 기반으로 하여 자동 생성된 요약과 참조 요약 사이의 유사도를 측정하는 지표이다. 여기서 n은 연속적인 단어의 수를 나타내며, 일반적으로 ROUGE-1과 ROUGE-2가 주로 사용된다[33]. ROUGE-1은 단어 단위 일치를, ROUGE-2는 인접한 단어 쌍의 일치를 측정하여 요약문의 정확성을 평가한다[34]. ROUGE-L은 요약간의 가장 긴 공통 부분수열(LCS, Longest Common Subsequence)을 기반으로 하여 요약문의 유창성과 구조적 일관성을 평가한다[35].

ROUGE score의 예시는 다음과 같이 설명할 수 있다. 참조 요약문이 "어린 소년이 파란 공을 가지고 놀고 있다."인 경우, 첫 번째 생성된 요약문은 "소년이 파란 공을 가지고 놀고 있다."이고, 두 번째 생성된 요약문은 "어린 소년이 공을 가지고 논다."라고 가정한다. 이 두 개의 생성된 요약문에 대해 ROUGE-1과 ROUGE-2 점수를 각각 계산할 수 있다. 먼저, ROUGE-1은 단어 단위의 일치를 기반으로 계산된다. 첫 번째 생성된 요약문인 "소년이 파란 공을 가지고 놀고 있다."는 참조 요약문과 6개의 단어("소년이", "파란", "공을", "가지고", "놀고", "있다")가 일치한다. 따라서 정밀도(precision)는 생성된 요약문에서 일치하는 단어의 비율로 6/7 = 0.857이고, 재현율(recall)은 참조 요약문에서 일치하는 단어의 비율로 6/7 = 0.857이다. 이 두 값의 조화 평균인 F1-score는 0.857로 계산된다. ROUGE-2는 인접한 단어 쌍 간의 일치를 측정하는 지표이다. 첫 번째 생성된 요약문에서는 5개의 단어 쌍("소년이 파란", "파란 공을", "공을 가지고", "가지고 놀고", "놀고 있다")이 참조 요약문과 일치한다. 참조 요약문에는 총 6개의 단어 쌍이 있으므로 재현율은 5/6 = 0.833이며, 생성된 요약문에서 일치하는 5개의 단어 쌍에 대한 ㄴ정밀도는 5/6 = 0.833이다. F1-score는 0.833으로 계산된다.

두 번째 생성된 요약문 "어린 소년이 공을 가지고 논다."에서는 ROUGE-1 기준으로 참조 요약문과 4개의 단어("어린", "소년이", "공을", "가지고")가 일치하며, 정밀도는 4/6 = 0.667, 재현율은 4/7 = 0.571로 F1-score는 0.615이다. ROUGE-2 기준으로는 2개의 단어 쌍("소년이 공을", "공을 가지고")이 일치하며, 정밀도는 2/5 = 0.4, 재현율은 2/6 = 0.333으로 F1-score는 0.364로 계산된다.

이와 같은 방식으로 ROUGE-1과 ROUGE-2 점수

를 통해 생성된 요약문이 참조 요약문과 얼마나 유사한지 평가할 수 있으며, 요약문 간의 단어 일치와 문맥적 일관성을 수치화하여 분석할 수 있다.

ROUGE score는 텍스트 요약의 품질을 평가하는 데 매우 유용하다. 예를 들어, ROUGE-1 점수가 높다면, 이는 생성된 요약문이 참조 요약문과 많은 단어를 공유하고 있음을 의미한다. 반면, ROUGE-2 점수가 높다면, 생성된 요약문이 참조 요약문과 많은 이웃 단어 쌍을 공유하고 있다고 볼 수 있다.

또한 ROUGE score는 텍스트 요약 프롬프트의 성능을 비교하는 데 유용하다. 다양한 프롬프트을 사용하여 생성된 요약문을 동일한 참조 요약문과 비교함으로써, 각 프롬프트의 장단점을 분석할 수 있다. 예를 들어, 한 프롬프트의 ROUGE-1 점수는 높지만 ROUGE-2 점수가 낮다면, 이는 해당 프롬프트가 개별 단어의 일치도는 높지만, 연속된 단어 쌍의 일치도는 낮다는 것을 의미한다. 이는 해당 프롬프트가 단어 선택은 좋지만, 문맥이나 문장 구조에서 일관성이 떨어질 수 있음을 시사한다.

반면, ROUGE-2 점수가 높지만 ROUGE-1 점수가 낮은 경우, 이는 생성된 요약문이 참조 요약문과 연속된 단어 쌍은 잘 맞지만, 전체적인 단어 일치도는 낮다는 것을 의미한다. 이 경우, 프롬프트가 문맥이나 문장 구조는 잘 유지하지만, 구체적인 단어 선택에서는 다소 차이가 있음을 나타낸다.

본 논문에서는 ROUGE score는 3% 이상 감소한 중요 지표와 증감율을 얼마나 정확하게 뽑아내는지를 확인하는 지표로 사용하였으며, 이를 통해 recall, precision, 및 f1-score를 측정하였다. Recall은 참조 요약문에서 자동 생성된 요약문에 포함된 정보의 비율을 나타내며, precision은 자동 생성된 요약문에서 참조 요약문에 실제로 포함된 정보의 비율을 나타낸다. F1-score는 recall과 precision의 조화 평균으로, 두 척도의 균형을 평가한다. 이러한 척도들을 통해 요약문의 포괄성, 정확성, 그리고 전반적인 품질을 면밀하게 분석하였다.

3.3.2 체크리스트

프롬프트의 정성적인 성능 평가를 위하여 자체적으로 성능 평가 체크리스트를 개발하였다(표 3). 코멘트 정확성, 평가에 대한 분석력, 언어 유창성, 사용성 4개의 평가 영역으로 나누어 총 18개의 문항으로 구성하였다. 객관성을 높이기 위해 같은 데이터에 대해 Optimal, Case 1, Case 2, Case 3, 4개의 프롬프트를 동일 인물이 평가하도록 진행하여 일관성 있게 평가하도록 하였고, 평가는 모두 블라인드로 진행되었다.

표 3. 프롬프트 정성적 평가를 위한 체크리스트
Table 3. Checklist for qualitative evaluation of prompts

Evaluation areas	Evaluation items	Yes	No
Accuracy of comments (20%)	1) Mentions cost-related metrics with a decrease of more than 3%.		
	2) Highlights only positive indicators.		
	3) Excludes outputs unrelated to the specified output statements.		
	4) Ensures that numbers in the response are accurate (No hallucination).		
Analytical skill in evaluation (30%)	1) Selects key metrics for each date and analyzes them.		
	2) Analyzes reasons for decreases through correlation analysis.		
	3) Analyzes daily metric changes to evaluate campaign efficiency.		
	4) Identifies correlations by reflecting operational details.		
	5) Includes detailed evaluations with information beyond the prompt rules.		
Language fluency (20%)	1) Uses concise sentences in a list format.		
	2) Employs vocabulary and phrasing suitable for report comments.		
	3) Summarizes each item in one sentence, presented as bullet points.		
	4) Structures sentences clearly for better readability.		
Usability (30%)	1) Provides responses quickly.		
	2) Organizes comments with clean titles and structure.		
	3) Ensures the output does not exceed the maximum token limit.		
	4) Avoids unnecessary content.		
	5) Clearly distinguishes between increases and decreases in metrics.		

‘예(1점)', ‘아니요(0점)'로 구성된 binary 척도를 사용하였고, 코멘트 정확성 4문항 (20%), 평가에 대한 분석력 5문항 (30%), 언어 유창성 4문항 (20%), 사용성 5문항 (30%)으로 구성하여 총 100점 만점 (5*4+6*5+4*5+6*5=100)으로 점수를 환산하여 비교하였다.

코멘트 정확성 영역의 경우 프롬프트에 포함된 조건 사항들을 얼마나 잘 준수했는지를 중심으로 확인해보았다. -3% 이상 감소한 비용 관련한 지표에 대해서 언급하고 있는지, 긍정적인 지표만 언급하고 있는지, 출력문 이외의 것은 출력하지 않고 있는지, 작성된 답변에 있는 숫자가 정확한지를 확인하였고, 해당 조건을 완벽하게 준수했을 경우에만 1점을 부여하였다.

캠페인 평가에 대한 분석 능력 영역은 평가 분석 능력을 종합적으로 체크하는 영역이다. 각 일자별 중요한 지표를 선택하여 분석하는지, 연관성 분석을 통해 감소한 이유를 분석하고 있는지, 전일 대비 지표들의 변화를 분석하여 캠페인의 효율성을 검토하는지, 운영사항을 반영하여 상관관계를 평가하고 있는지, 프롬프트 규칙 이외의 내용들을 포함하여 자세하게 평가하고 있는지를 확인하였고, 증감률과 같은 단순 분석 외의 내용들을 얼마나 많이 표현하고 있는지를 중점으로 평가하였다.

언어 유창성 영역은 모델의 언어 능력을 평가하는 영역이다. 말투, 단어, 문장 선택 등의 요소가 포함된다. 개조식 문장을 사용하여 간단히 작성하고 있는지, 리포트 코멘트에 어울리는 단어와 문장을 사용하는지, 각 항목을 한 문장으로 요약하고 불렛 형태로 나타내고 있는지, 사용자가 알아듣기 쉽게 문장을 구성하고 있는지를 평가하였다.

사용성 영역은 프롬프트를 사용자들이 얼마나 편하게 사용할 수 있는가에 대한 영역이다.

대답 출력이 빠르게 진행되는지, 코멘트의 제목과 구조가 깔끔하게 정리되어 있는지, 최대 토큰 수를 초과하지 않고 출력하는지, 필요한 내용만 담고 있는지, 감소 및 증가를 헷갈리지 않게 작성하는지를 확인하였다.

앞서 언급한 체크리스트는 프롬프트의 정성적 성능 평가를 위해 전문가들의 실무 경험과 피드백을 반영하여 설계되어 내용 타당도를 확보하였다. 체크리스트 항목은 프롬프트가 사용되는 구체적인 용도, 즉 테이블 데이터를 텍스트로 요약하여 코멘트를 생성하는 작업에 맞춰 특화되었다. 기존에 유사한 프롬프트의 성능 평가 연구가 부족한 상황에서 전문가들이 실무에서 필요로 하는 기능과 성능을 기반으로 항목을 선정하고, 주간 회의를 통해 지속적인 피드백을 받아 체크리스트를 개선하였다.

이 과정에서 전문가들은 프롬프트의 사용성을 개선하기 위한 구체적인 요구사항을 제시하였으며, 각 항목이 실제 작업 환경에 적합한지 여부를 검토하였다. 이를 통해 프롬프트의 실제 사용 환경과 목적에 부합하는 평가 항목을 도출할 수 있었으며, 최종적으로 해당 체크리스트는 본 논문에서 설계한 프롬프트의 성능을 평가하는 데 적합한 도구일 것으로 판단된다.

Ⅳ. 결 과

4.1 전반적인 연구 과정

전반적인 연구 과정은 다음과 같다(그림 2). 먼저, 데이터 전처리 단계에서는 결측치 제거 및 KPI 지표 계산 및 KPI 컬럼 생성 작업뿐만 아니라 프롬프트 입력을 위한 데이터셋 전환이 포함된다. 프롬프트에 적용하는 테이블 형식은 일반 text, LaTex, Json 형식보다 Markdown 형식이 입력 토큰 수, LLM 인식 정확도 측면에서 상대적으로 낫다고 판단하여 Markdown 형식을 활용하였다.

두 번째 단계에서는 최적의 프롬프트와 여러 비교 케이스(예: Case1, Case2, Case3)를 설계하고, 이를 활용해 리포팅을 생성한다. 생성된 리포팅의 평가를 위해 리포팅 텍스트의 형식을 맞추고 불필요한 정보를 제거하는 텍스트 전처리 과정을 진행하였다(그림 1). 텍스트 전처리 단계에서는 불용어를 제거하고 %와 소수점을 제외한 구두점을 제거하였다. 이후 단어를 토큰화 시키고 지표명 사이의 띄어쓰기를 제거하였다(Text Preprocessing 1). 분석의 중심이 되는 지표, ‘증감률’, ‘증가’, ‘감소’ 외의 단어는 모두 삭제하고 ‘증가’와 ‘감소’로 통일 시켰다. 단어 배열의 순서를 지표 + 증감률 + ‘증가’ 혹은 ‘감소’ 순으로 정렬하였다(Text Preprocessing 2).

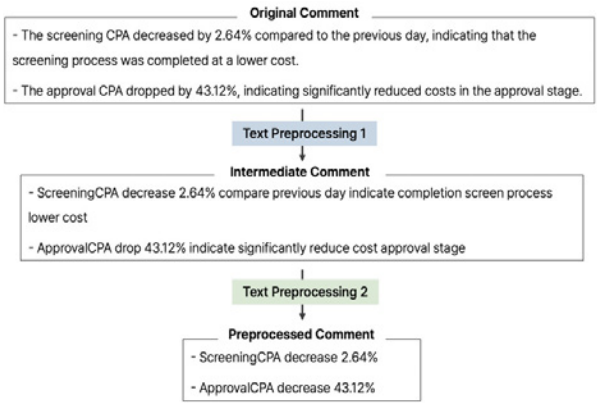

그림 1. ROUGE 평가를 위한 텍스트 전처리 과정
Fig. 1. Text preprocessing for ROUGE evaluation

마지막은 생성된 리포트를 평가하는 단계이다. 여기서는 두 가지 주요 평가 방법이 사용되는데, 첫 번째는 생성된 리포트와 제공받은 리포트 간의 유사성을 ROUGE 점수를 통해 비교하는 유사성 비교 방법이고, 두 번째는 프롬프트의 사용성을 체크리스트를 통해 평가하는 방법이다. 유사성 비교와 사용성 평가 결과를 바탕으로 최적의 프롬프트가 얼마나 효율적인지 검증하였다. 이를 통해 최적 프롬프트의 성능을 확인하고, 필요한 경우 프롬프트를 조정하는 과정을 거쳤다.

4.2 최종 프롬프트 및 대조 프롬프트 구조

후카츠 프롬프트를 활용하여 개발한 Optimal 프롬프트의 효과를 검증하기 위해, 각기 다른 구조의 프롬프트와의 성능을 비교 분석하였다. Type 1은 페르소나 기법으로 모델에게 역할을 부여하여 모델이 더 정확하게 목적과 역할을 인지할 수 있도록 하였다. Type 2는 모델에게 해야 할 일을 명시해 주는 부분이다. Type 3은 데이터 중 주요 지표인 비용 관련 지표에 대해서 명시하고 있다. Type 4 필수조건은 데이터 분석과 관련하여 반드시 있어야 하는 조건이며, Type 5 부가 조건은 프롬프트 연구 과정에서 최적의 답변을 얻기 위해 고려하였던 사항들을 정리한 조건이다. Type 6은 예시를 제공하는 기법으로, 예시를 하나만 제공하는 one-shot 기법을 사용하였다. 최종 프롬프트의 성능을 확인하기 위해 대조군 Case 1, 2, 3 프롬프트를 구성하였다. Case 1의 경우 Type 1~5의 기법들이 사용된 프롬프트이고, Case 2는 Case 1과 다르게 Type 5 부가 조건을 제외하고 Type 6 one-shot을 추가하여 구성하였다. Case 3의 경우는 Type 1~6의 기법들을 모두 사용하여 구성하였다. 최종 프롬프트인 Optimal 프롬프트는 후카츠 기법과 Type 1, 2, 3에 추가적인 규칙사항 및 제약 조건들을 추가하여 구성하였다. 연구에서 사용한 프롬프트 기법과 각 프롬프트의 구성은 (표 4)와 (표 5)을 통해 확인할 수 있다.

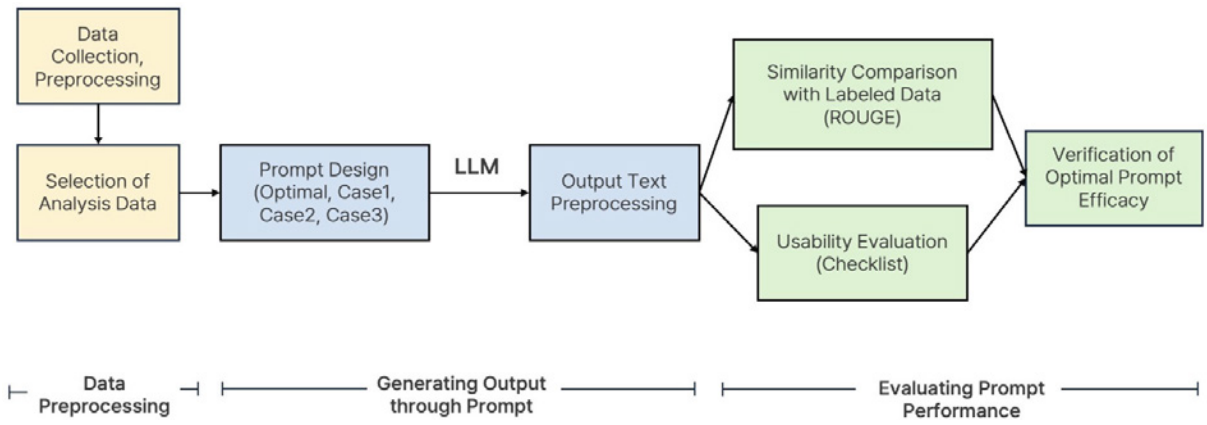

그림 2. 연구 과정 워크플로우
Fig. 2. Research process workflow

표 4. 사용한 프롬프트 기법
Table 4. Used prompt techniques

Type 1	Persona	You have 10 years of experience as a performance marketer. You are skilled at analyzing the efficiency of advertising campaigns and deriving insights from performance data. You excel in analyzing relationships between metrics and clearly explaining the causes of fluctuations in advertising cost-related metrics.
Type 2	Command	I am preparing daily report comments based on the analysis of daily performance and metric changes for advertising campaigns. I will write comments only for metrics with a decreased rate of change. The comments will strictly follow campaign management details, constraints, rules, and the specified output format. The analysis results will be provided to the client, [BANK].
Type 3	Marketing data	The cost-related metrics are as follows:CPC, CPS, CPU, New Visit CPU, Registration CPA, Screening CPA, Approval CPA, CPA, Savings CPA, Loan CPA.
Type 4	ECG (Essential Condition Giving)	- Analyze the relationships between metrics and describe how the cost-related metrics with a day-over-day change rate of -3% or more were impacted, including their unit costs. - Write comments only for cost-related metrics with a day-over-day change rate of -3% or more.
Type 5	ACG (Additional Condition Giving)	- Summarize how changes in operational details impacted cost-related metrics with a day-over-day change rate of -3% or more. - If no metric has decreased by more than 3%, provide information about the #cost-related metric with the most significant change. - Use suitable terms for the report and write concisely in bullet points with simple, structured sentences. - Provide output in the specified format only. - Summarize each point in one sentence and list them as bullets.
Type 6	One-Shot	Example response: The aaa metric decreased by -bb%, resulting in a -dd% improvement in the ccc metric.

표 5. 프롬포트 구성
Table 5. Prompt composition

Case 1	Type 1 (Persona) + Type 2 (Command) + Type 3 (Marketing data) + Type 4 (ECG) + Type 5 (ACG)
Case 2	Type 1 (Persona) + Type 2 (Command) + Type 3 (Marketing data) + Type 4 (ECG) + Type 6 (One-Shot)
Case 3	Type 1 (Persona) + Type 2 (Command) + Type 3 (Marketing data) + Type 4 (ECG) + Type 5 (ACG) + Type 6 (One-Shot)
Optimal	**#Data:** **#Operational Details:** **#Commands:** Type 1 (Persona) Type 2 (Command) **#Cost-Related Metrics:** Type 3 (Marketing Data) **#Output Statement Rules:** - Analyze the relationships between metrics and describe how the cost-related metrics with a day-over-day change rate of -3% or more were impacted, including their unit costs. - Write comments only for cost-related metrics with a day-over-day change rate of -3% or more. - Summarize how changes in operational details impacted cost-related metrics with a day-over-day change rate of -3% or more. - If no metric has decreased by more than 3%, provide information about the #cost-related metric with the most significant change. **#Constraints:** - Use suitable terms for the report and write concisely in bullet points with simple, structured sentences. - Provide output in the specified format only. - Summarize each point in one sentence and list them as bullets. **#Output Statement** [Campaign Name - Media Name] - -

4.3 ROUGE Score 및 체크리스트 평가 결과

Optimal, Case 1, Case 2, Case 3 프롬프트를 통해 생성된 출력문을 레이블 리포트와 비교하여 각 프롬프트의 성능을 ROUGE Score를 사용하여 평가하였다(그림 3, 표 6). ROUGE Score의 ROUGE-1과 ROUGE-2 점수를 사용하여 각 프롬프트의 성능을 구체적으로 분석하였다.

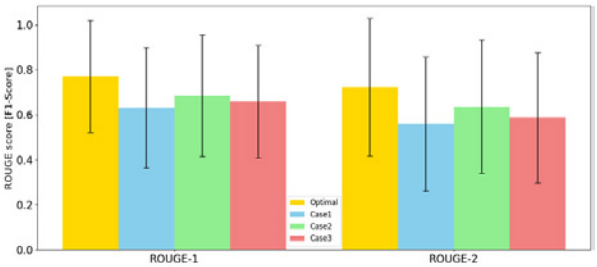

그림 3. ROUGE score (F1-score) 결과 그래프
Fig. 3. ROUGE score (F1-score) result graph

표 6. 프롬프트 타입별 ROUGE score 평가 결과
Table 6. ROUGE score result by prompt type

Prompt	Metric	Recall	Precision	F1
Optimal	1	0.92	0.70	0.77
	2	0.88	0.66	0.72
Case 1	1	0.83	0.57	0.63
	2	0.76	0.52	0.56
Case 2	1	0.77	0.70	0.68
	2	0.72	0.67	0.63
Case 3	1	0.72	0.66	0.65
	2	0.65	0.61	0.58

먼저, Optimal 프롬프트의 경우, ROUGE-1 recall 0.92, precision 0.70, f1-score 0.77을 기록하였다. ROUGE-2 점수는 recall 0.88, precision 0.66, f1-score 0.72로, 전체적으로 높은 성능을 보였다. 이는 Optimal 프롬프트가 단어 선택과 문장 구조, 문맥적 일관성 모두에서 뛰어난 성능을 나타낸다는 것을 의미한다.

Case 1 프롬프트의 성능은 상대적으로 낮았다. ROUGE-1 점수는 recall 0.83, precision 0.57, F1-score 0.63으로, ROUGE-2 점수는 recall 0.76, precision 0.52, F1-score 0.56으로 평가되었다. 이는 Case 1 프롬프트가 단어 선택과 문장 구조에서 Optimal 프롬프트보다 낮은 성능을 보인다는 것을 시사한다.

Case 2 프롬프트는 ROUGE-1 점수에서 recall 0.77, precision 0.70, F1-score 0.68을 기록하였고, ROUGE-2 점수는 recall 0.72, precision 0.67, F1-score 0.63으로 나타났다. 이는 Case 1보다 개선된 성능을 보여주었으며, 특히 문맥적 일관성에서 더 나은 결과를 나타냈다.

Case 3 프롬프트의 경우, ROUGE-1 점수는 recall 0.72, precision 0.66, F1-score 0.65였고, ROUGE-2 점수는 recall 0.65, precision 0.61, F1-score 0.58로 Case 2에 비해 낮은 성능을 보였다.

위 결과를 바탕으로, Optimal 프롬프트가 단어 선택, 문장 구조, 문맥적 일관성 모두에서 가장 우수한 성능을 보이는 것을 확인하였다. 따라서, Optimal 프롬프트는 최적의 리포트를 생성하기 위한 가장 효과적인 프롬프트로 결론지을 수 있다.

체크리스트를 활용해 정성적 평가를 진행한 결과, Optimal의 평균 값(89.8)이 다른 모형(Case 1: 82.03, Case 2: 73.83, Case 3: 82.87)보다 더 높음을 확인할 수 있다 (그림 4). 이는 Optimal 프롬프트를 통해 생성된 리포트가 내용의 정확성, 캠페인 평가에 대한 분석력, 언어 유창성, 사용성 측면에서 모두 뛰어난 성능을 발휘한다는 것을 의미한다. 반면, ROUGE Score에서는 상대적으로 낮은 성과를 보였던 Case 1과 Case 3가 체크리스트 평가에서는 상대적으로 높은 점수를 기록한 점을 고려할 때, 단어 단위의 내용 정확성과 전체적인 프롬프트의 사용성은 항상 비례하지 않음을 확인할 수 있었다.

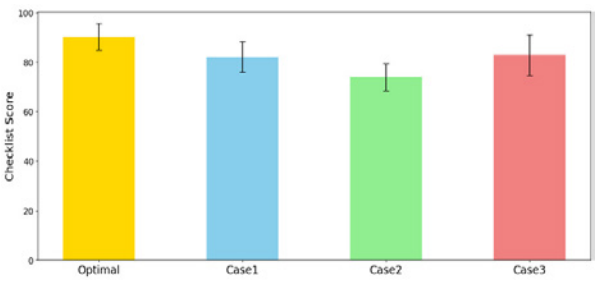

그림 4. 프롬프트 타입별 체크리스트 점수 평균
Fig. 4. Average scores for checklist by prompt type

V. 결 론

본 논문에서는 다양한 프롬프트 기법을 통합한 새로운 입력 프롬프트 구조를 설계하고 그 중 최적의 프롬프트를 선정하고자 하였다. 이를 위해 체크리스트 및 ROUGE Score를 활용하여 제안된 프롬프트 구조가 기존 프롬프트에 비해 데이터 입력의 정확성 측면에서 어떠한 차이를 보이는지 체계적으로 평가하였다. 정확성의 정량적 평가 결과, ROUGE-1, ROUGE-2, ROUGE-L을 통해 측정된 결과는 제안된 Optimal 프롬프트의 출력문이 레이블 데이터와의 유사성 면에서 기존 프롬프트보다 상당히 높은 값을 기록하였음을 확인하였다. 이는 제안된 프롬프트 구조가 명령 준수의 정확성 측면에서 우수함을 시사한다. 정성적 평가에서는 ROUGE Score로 평가하지 못하는 언어 유창성 및 사용성 부분을 평가하고자 하였으며 그 결과 다른 프롬프트 대비 Optimal 프롬프트에서 높은 언어 유창성 및 사용성을 확인할 수 있었다.

연구의 의의는 다음과 같다. 먼저, 프롬프트 구조의 개선을 통해 LLM을 활용한 데이터 분석의 정확성과 사용성을 향상시키는 방법론을 제시하였고, 특히 선행 연구에서 많이 다루지 않은 'Table-to-Text' 프롬프트 패턴을 제안하였다. 정량적 및 정성적 평가 체계를 도입하여 프롬프트 성능 평가에 대한 기준을 마련하였고, 데이터 분석에 적합한 프롬프트 구조를 제시함으로써 사무 업무 자동화와 업무 부담 감소에 기여할 것으로 기대된다. 이러한 연구 결과는 프롬프트 기반 데이터 분석 방법론의 개선뿐만 아니라, 광범위한 업무 프로세스 최적화에도 중요한 기여를 할 것으로 예상된다. 본 논문의 한계점은 다음과 같다. 다양한 업무의 실제 사례 분석에 적용 시, 해당 분야의 세부적인 정보에 대해 조정이 필요하다. 본 논문은 퍼포먼스 마케팅 데이터를 기반으로 진행되었으므로, 보다 다양한 분야와 다른 예시의 데이터들을 기반으로 조정한다면 충분히 극복할 수 있을 것으로 보인다.

Acknowledgement

해당 논문은 2024년도 한국정보기술학회 하계종합학술대회에서 발표한 논문 “생성형 AI 모델을 활용한 비즈니스 데이터 분석 리포트 생성 프롬프트 엔지니어링 연구: 퍼포먼스 마케팅 사례를 중심으로”[36]을 확장한 것임.

References

[1] K. Kook, "Artificial Intelligence Technology and Industry-Specific Application Cases", Weekly Technology Trends of the Institute for Information & Communications Technology Planning & Evaluation, No. 1888, pp. 15-27, 2019.

[2] L. Cheng, X. Li, and L. Bing, "Is GPT-4 a Good Data Analyst?", arXiv preprint arXiv:2305.15038, May 2023. https://doi.org/10.48550/arXiv.2305.15038.

[3] R. Raj, A. Singh, V. Kumar, and P. Verma, "Analyzing the Potential Benefits and Use Cases of ChatGPT as a Tool for Improving the Efficiency and Effectiveness of Business Operations", BenchCouncil Trans. Benchmarks, Standards and Evaluations, Vol. 3, No. 3, pp. 100140, Sep. 2023. https://doi.org/10.1016/j.tbench.2023.100140.

[4] S. K. Singh, S. Kumar, and P. S. Mehra, "Chat GPT & Google Bard AI: A Review", 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, pp. 1-6, Jun. 2023. https://doi.org/10.1109/ICICAT57735.2023.10263706.

[5] S. H. Lee and K. S. Song, "Prompt Engineering to Improve the Performance of Teaching and Learning Materials Recommendation of Generative Artificial Intelligence", Journal of The Korea Society of Computer and Information, Vol. 28, No. 8, pp. 195-204, Aug. 2023. https://doi.org/10.9708/jksci.2023.28.08.195.

[6] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, and D. C. Schmidt, "A Prompt Pattern Catalog to Enhance Prompt Engineering with

ChatGPT", arXiv preprint arXiv:2302.11382, Feb. 2023. https://doi.org/10.48550/arXiv.2302.11382.

[7] S. Park and J. Kang, "Analysis of Prompt Engineering Methodologies and Research Status to Improve Inference Capability of ChatGPT and Other Large Language Models", Journal of Intelligence and Information Systems, Vol. 29, No. 4, pp. 287-308, Dec. 2023.

[8] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende, "Prompt Engineering in Large Language Models", in International Conference on Data Intelligence and Cognitive Informatics, Singapore: Springer Nature Singapore, pp. 387-402, Jun. 2023. https://doi.org/10.1007/978-981-99-7962-2_30.

[9] J. Bao, D. Tang, N. Duan, Z. Yan, Y. Lv, M. Zhou, and T. Zhao, "Table-to-text: Describing table region with natural language", in Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, Apr. 2018. https://doi.org/10.1609/aaai.v32i1.11944.

[10] L. S. Lo, "The CLEAR path: A framework for enhancing information literacy through prompt engineering", The Journal of Academic Librarianship, Vol. 49, No. 4, pp. 102720, Jul. 2023. https://doi.org/10.1016/j.acalib.2023.102720.

[11] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries", in Text Summarization Branches Out, pp. 74-81, Jul. 2004.

[12] F.-H. Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration", Journal of Information Technology Case and Application Research, Vol. 25, No. 3, pp. 277-304, Jul. 2023. https://doi.org/10.1080/15228053.2023.2233814.

[13] P. W. Cardon, K. Getchell, S. Carradini, C. Fleischmann, and J. Stapp, "Generative AI in the workplace: Employee perspectives of ChatGPT benefits and organizational policies", SocArXiv Papers, Mar. 2023. https://doi.org/10.31235/osf.io/b3ezy.

[14] F. Trad and A. Chehab, "Prompt engineering or fine-tuning? A case study on phishing detection with large language models", Mach. Learn. Knowl. Extr., Vol. 6, No. 1, pp. 367-384, Jan. 2024. https://doi.org/10.3390/make6010018.

[15] E. Sabit, "Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices", TechRxiv, May 2023. https://doi.org/10.36227/techrxiv.22683919.v2.

[16] J. D. Velásquez-Henao, C. J. Franco-Cardona, and L. Cadavid-Higuita, "Prompt engineering: A methodology for optimizing interactions with AI-language models in the field of engineering", DYNA, Vol. 90, No. 230, pp. 111-700, Nov. 2023. https://doi.org/10.15446/dyna.v90n230.111700.

[17] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications", arXiv preprint arXiv:2402.07927, Feb. 2024. https://doi.org/10.48550/arXiv.2402.07927.

[18] T. B. Brown, et al., "Language models are few-shot learners", arXiv preprint arXiv:2005.14165, May 2020. https://doi.org/10./arXiv.2005.14165.

[19] J. Wei, et al., "Chain-of-thought prompting elicits reasoning in large language models", arXiv preprint arXiv:2201.11903, Jan. 2022. https://doi.org/10.48550/arXiv.2201.11903.

[20] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models", arXiv preprint arXiv:2210.03493, Oct. 2022. https://doi.org/10.48550/arXiv.2210.03493.

[21] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models", arXiv preprint arXiv:2305.10601, May 2023. https://doi.org/10.48550/arXiv.2305.10601.

[22] M. Sclar, et al., "Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting", arXiv preprint arXiv:2310.11324, Oct. 2023. https://doi.org/10.48550/arXiv.2310.11324.

[23] W. Sim, H. Jin, S. Kim, and S. Kim, "The possibility of prompt engineering for ARC problem solving", KIISE Transactions on Computing Practices, Vol. 30, No. 2, pp. 63-69, Feb. 2024. https://doi.org/10.5626/KTCP.2024.30.2.063.

[24] D. Heo, K. Kim, H. Song, and B. Suh, "Proposal of Korean CSAT customized question generator system with prompt engineering", Proceedings of the Korean HCI Conference, Gangwon, South Korea, pp. 183-189, Jan. 2024.

[25] H.-N. Lee and Y.-S. Lee, "Exploring the feasibility of automated item generation for Korean language assessment", Journal of Education and Culture, Vol. 30, No. 3, pp. 659-686, 2024.

[26] B.-C. Shin, J.-S. Lee, Y.-J. Yoo, "Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations", Mathematics Education, Vol. 63, No. 2, pp. 187-207, 2024.

[27] W. Chen, et al., "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks", arXiv preprint arXiv:2211.12588, Nov. 2022. https://doi.org/10.48550/arXiv.2211.12588.

[28] A. Shirafuji, et al., "Prompt sensitivity of language model for solving programming problems", in New Trends in Intelligent Software Methodologies, Tools and Techniques, IOS Press, pp. 346-359, Sep. 2022. https://doi.org/10.3233/FAIA220264.

[29] S. Seo, "Prompt Engineering Textbook", And&Media, 2023.

[30] J. Weston and S. Sukhibaatar, "System 2 attention (is something you might need too)", arXiv preprint arXiv:2311.11829, Nov. 2023. https://doi.org/10.48550/arXiv.2311.11829.

[31] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics", in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Canada, Vol. 1, pp. 150-157, May 2003.

[32] E. Hovy, C. Y. Lin, L. Zhou, and F. Fukumoto, "Automated summarization evaluation with basic elements", in Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, Vol. 6, pp. 604-611, May 2006.

[33] C.Y. Lin, "Rouge: A package for automatic evaluation of summaries", in Text Summarization Branches Out, 2004. Available: https://aclanthology.org/W04-1013

[34] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions", in Proceedings of the 23rd International Conference on Computational Linguistics, pp. 340-348, Association for Computational Linguistics, 2010. https://doi.org/10.1145/1873781.1873820

[35] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality", in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), 2008. https://doi.org/10.1145/1613715.1613742.

[36] Y. Park, D. Kim, S. Yoo, Y. Bae, and K. Kim, "Research on Prompt Engineering for Generating Business Data Analysis Reports Using Generative AI Models: Focused on Performance Marketing Cases", Proceedings of KIIT Conference, Jeju, Korea, May 2024.

저자소개

박 예 은 (Yeeun Park)

2024년 8월 : 한동대학교 생명과학&AI융합(학사)
2024년 8월 ~ 현재 : 분당서울대학교병원 데이터융합팀 연구원
관심분야 : 바이오빅데이터, 헬스케어, 인공지능, 데이터분석

김 동 훈 (DongHun Kim)

2024년 8월 : 한동대학교 상담심리학&데이터사이언스 (학사)
2024년 9월 ~ 현재 : 고려대학교 심리학과 석사과정
관심분야 : 상담심리학, LLM, 데이터분석

유 시 현 (Sihyeon Yoo)

2020년 3월 ~ 현재 : 한동대학교 ICT융합&데이터사이언스 학사과정
관심분야 : 데이터 분석, 인간 공학, HCI, 빅데이터

배 윤 진 (Yunjin Bae)

2019년 2월 ~ 현재 : 한동대학교 경제학&AI융합 학사과정
관심분야 : 경제학, 재무분석, 주식 데이터 분석, 데이터 분석

김 경 외 (Keungoui Kim)

2012년 2월 : 한동대학교 기계공학&전자(학사)
2016년 2월 : 서울대학교 기술경영경제정책(석사)
2019년 2월 : 서울대학교 기술경영경제정책(박사)
2021년 9월 ~ 현재 : 한동대학교 AI융합교육원 조교수
관심분야 : 전산 사회 과학, 텍스트 마이닝, 네트워크 분석