

컴퓨터교육학회 논문지 2026년 제29권 제7호
<https://doi.org/10.32431/kace.2026.29.7.005>



SLM을 위한 CoT 기반 한국어 관광 도메인 데이터셋 구축에 관한 연구*

A Study on Constructing a CoT-Based Korean Tourism Domain Dataset for Small Language Models

오민수[†] · 윤여찬^{††}
 Minsu Oh[†] · Yeochan Yoon^{††}

요약

본 연구는 여행·관광 도메인에서 low parameter LLM의 한국어 응답 품질을 향상시키기 위해 여행·관광 도메인 데이터를 수집하고, 이를 입력-추론-출력 구조의 데이터셋으로 구축하였다. 구축한 데이터를 이용하여 low parameter LLM인 Llama3.2 1B/3B 모델과 Qwen3 1.7B/4B 모델을 LoRA 학습을 진행하였다. 학습한 결과, Llama 3.2 1B 모델의 BERTScore는 0.8295에서 0.8687로 향상되었으며 다른 모델의 성능도 향상되는 경향을 보였다. 이는 구조화된 도메인 데이터가 low parameter LLM의 한국어 여행·관광 질의응답 성능 개선에 효과적임을 보여준다.

주제어 LLM, 여행·관광 질의응답, 도메인 적응, 입력-추론-출력 구조, LoRA, 한국어 데이터셋

ABSTRACT

This study aims to improve the quality of Korean responses generated by low-parameter large language models (LLMs) in the travel and tourism domain. To this end, travel and tourism domain data were collected and organized into a structured dataset consisting of input, reasoning, and output components. Using the constructed dataset, LoRA fine-tuning was performed on Llama 3.2 1B/3B and Qwen3 1.7B/4B models. The experimental results showed that the BERTScore of Llama 3.2 1B improved from 0.8295 to 0.8687, while the other models also demonstrated performance improvements. These findings indicate that structured domain-specific data can effectively enhance the Korean travel and tourism question-answering performance of low-parameter LLMs.

Keywords Large language models, Travel and tourism question answering, Domain adaptation, Input-Reasoning-Output structure, LoRA, Korean dataset

* Large language models, Travel and tourism question answering, Domain adaptation, Input-Reasoning-Output structure, LoRA, Korean dataset

†종신회원 제주대학교 소프트웨어학부 인공지능
전공 학사과정
 ††정회원 제주대학교 인공지능학과 교수(교신저자)
 논문투고 2026년 03월 23일
 심사완료 2026년 05월 08일
 게재확정 2026년 06월 08일
 발행일자 2026년 07월 22일

* 본 연구는 제주대학교의 2026년도 교육·연구 및 학생
지도비 지원사업의 지원을 받아 수행되었습니다.

1. 서론

대규모 언어 모델(Large Language Models, LLMs)은 자연어 이해 능력의 빠른 향상과 이를 뒷받침하는 인프라의 빠른 확충을 바탕으로 질의응답, 추천, 요약, 대화형 서비스 등 다양한 응용 분야로 활용 범위를 넓혀가고 있다. 그러나 실제 서비스 환경에서는 연산 비용, 배포 효율성, 응답 지연, 메모리 제약과 같은 현실적 조건을 함께 고려해야 하므로, 상대적으로 작은 규모의 언어 모델과 parameter-efficient fine-tuning 기법의 활용이 중요해지고 있다. LoRA[1]와 QLoRA[2]는 이러한 흐름을 대표하는 기법이며, Phi-3 기술 보고서[3]는 소형 모델 또한 적절한 학습 전략과 데이터 설계를 결합할 경우 높은 실용성을 가질 수 있음을 보여주었다.

여행 및 관광 도메인의 질의응답은 이러한 소형 모델의 한계가 특히 두드러지게 나타나는 영역이다. 여행 추천은 사용자의 선호와 방문 목적뿐 아니라 시간 예산, 이동 가능성, 지역 맥락, 관광지 간 관계와 같은 다양한 요인을 종합적으로 고려해야 하므로 단순한 정보 회수보다 높은 수준의 문맥 이해를 요구한다[4-6]. 최근에는 관광 도메인에 특화된 LLM, RAG, 멀티모달 모델을 활용하여 여행 계획이나 관광 질의를 처리하려는 연구도 활발히 진행되고 있다[7-11]. 그러나 기존 연구는 주로 중국어 중심 데이터셋, 외부 지식 결합형 프레임워크, 실제 여행 계획 벤치마크, 혹은 멀티모달 관광 지원에 초점을 두고 있어, 한국어 중심의 구조화된 학습 데이터와 소형 오픈모델 적응을 동시에 다룬 연구는 충분하지 않다.

또한 복잡한 조건을 고려해야 하는 과업에서는 단순한 질문-응답 쌍만으로는 충분한 감독 신호를 제공하기 어렵다. Chain-of-thought prompting[12], instruction tuning [13], Self-Instruct [14]와 같은 연구는 중간 추론 과정을 반영한 데이터 구성이 모델의 일반화와 응답 품질 향상에 기여할 수 있음을 보여주었다. 관광 도메인에서도 LETToT[15]와 같이 추론 구조를 활용한 평가 및 분석 시도가 등장하고 있으나, 한국어 여행/관광 질의를 대상으로 입력-추론-출력 구조를 갖춘 학습 데이터셋을 구축하고 이를 소형 언어모델에 적용한 실증적 연구는 여전히 제한적이다.

이에 본 연구에서는 여행/관광 질의를 대상으로 한국어 데이터셋을 구축하고, 이를 입력(Input), 추론(CoT: Chain of Thought), 출력(Output)의 3필드 구조로 정제하였다. 여기서 추론(CoT)은 모델의 사고 과정을 Chain of Thought 방식으로 모델의 추론 방향을 서술한 데이터 필드를 지칭한다. 또한 해당 데이터를 활용하여 Llama 3.2 1B/3B와 Qwen3 1.7B/4B 모델에 대해 LoRA 기반 파인튜닝을 수행하고, 원본 모델과의 비교를 통해 구조화된 도메인 데이터가 소형 언어모델의 한국어 응답 품질에 미치는 영향을 분석하였다. 본 연구를 통해 여행/관광 도메인의 데이터 셋을 구축하였으며 구축한 데이터셋을 이용하여 소형 언어모델의 성능을 향상시킬 수 있음을 보였다.

2. 관련 연구

2.1 도메인 추천 및 질의응답 연구

여행 및 관광 추천 연구는 오랫동안 사용자의 선호, 시간 예산, 이동 가능성, 지역 정보와 같은 맥락을 함께 반영하는 문제로 다루어져 왔다[4-6]. Bahramian et al.[4]은 온톨로지 and spreading activation을 활용하여 사용자 선호와 맥락 정보를 반영하는 관광 추천 시스템을 제안하였고, He et al.[5]은 POI의 인기도, 동시 방문 관계, 사용자 선호를 함께 학습하는 context-aware POI embedding을 통해 여행 일정 추천 성능을 개선하였다. 또한 Tsai et al.[6]은 POI의 암묵적·동적 정보를 고려하는 투어 추천 시스템을 제안함으로써, 관광 추천이 단순한 장소 나열이 아니라 사용자의 상황과 제약을 반영해야 하는 과업임을 보여주었다.

최근에는 관광 도메인에 특화된 대규모 언어모델 및 검색 결합형 접근도 활발히 보고되고 있다. Wei et al.[7]은 관광 지식을 반영한 지도 미세조정 모델을 제안하여 관광지 설명과 여행 계획 생성 품질을 개선하고자 하였다. Song et al.[8]은 다층 지식그래프 기반의 TravelRAG 프레임워크를 제안하여 관광지 검색과 생성의 결합 가능성을 보여주었다. Karlović et al.[9]은 retrieval-augmented large language models와 semantic re-ranking을 결합한 맥락 인지형 관광 추천 방식을 제시하였고, Deng et al.[10]은 실제 환경 제약과 암묵적 요구를 반영하는 여행 계획 벤치마크를 통해 현실 세계 여행 계획 과업의 난이도를 분석하였다. 또한 Chu et al.[11]은 도시 장면 이해와 여행 지원을 위한 멀티모달 모델을 제안하여 여행 도메인에서 텍스트와 시각 정보를 함께 다루는 가능성을 확장하였다.

이러한 연구는 관광 분야에서 LLM의 활용 가능성을 보여주지만, 한국어 여행/관광 질의를 위한 구조화된 지도 데이터와 소형 오픈모델 적응이라는 관점에서는 여전히 공백이 존재한다. 따라서 본 연구는 기존 관광 추천 및 관광 특화 LLM 연구를 확장하여, 한국어 중심 데이터 구조와 소형 모델 적응 전략의 효과를 실험적으로 검증하고자 한다.

2.2 추론 구조를 반영한 데이터 구성 연구

언어모델 학습에서 단순한 질문-응답 쌍을 넘어 중간 추론 과정을 명시하는 접근은 복합 과업의 성능 향상과 직결되는 중요한 연구 흐름으로 발전해 왔다[12-14]. Wei et al.[12]은 chain-of-thought prompting이 산술, 상식, 기호 추론 과제에서 유의미한 성능 향상을 가져올 수 있음을 보였고, Chung et al.[13]은 instruction tuning과 CoT 데이터를 함께 활용할 때 다양한 미시 과업에 대한 일반화 성능이 개선됨을 보고하였다. 또한 Wang et al.[14]의 Self-Instruct는 모델이 스스로 생성한 instruction 데이터와 필터링 절차를 이용해 instruction-following 능력을 향상시킬 수 있음을 제시하였다.

관광 도메인에서도 reasoning 구조를 반영하려는 시

도가 점차 등장하고 있다. LETToT[15]는 관광 QA 평가에 expert tree-of-thought를 적용하여 도메인 지식과 추론 구조를 결합한 평가 프레임워크를 제안하였고, TravelLaMA[11]는 여행 관련 텍스트와 시각 질의응답에서 구조화된 추론의 필요성을 보여주었다. 그러나 기존 연구는 주로 평가 프레임워크, 멀티모달 데이터셋, 혹은 특정 언어권 중심의 관광 데이터에 초점을 두고 있어, 한국어 중심의 입력-추론-출력 구조를 갖춘 학습 데이터 설계와 소형 오픈모델 적응을 동시에 다룬 연구와는 차이가 있다.

본 연구는 이러한 선행연구의 흐름을 바탕으로, 여행/관광 질의응답에서 사용자의 의도 해석, 제약 조건 파악, 지역 정보 평가, 후보 필터링과 같은 중간 판단 과정을 추론 필드로 구조화하여 지도 학습에 활용하고자 한다. 이는 최종 응답만을 학습하는 방식보다 풍부한 감독 신호를 제공함으로써, 응답의 일관성과 적합성을 높이기 위한 전략으로 이해할 수 있다.

2.3 소형 언어모델의 도메인 적응과 효율적 파인튜닝

자원 제약 환경에서 언어모델을 활용하기 위한 연구는 소형 언어모델과 parameter-efficient fine-tuning 기법에 주목하고 있다[1-3]. Hu et al.[1]이 제안한 LoRA는 사전학습된 가중치를 고정해 채 저랭크 행렬만 학습함으로써 학습 가능한 파라미터 수와 GPU 메모리 요구량을 크게 줄이는 방법이며, Dettmers et al.[2]의 QLoRA는 4-bit 양자화 모델 위에 LoRA를 적용하여 더 적은 자원으로 미세조정을 가능하게 하였다. 또한 Abidin et al.[3]은 Phi-3 기술 보고서를 통해 소형 모델이 모바일 환경까지 고려한 배포 대상이 될 수 있을 정도로 실용성이 높아지고 있음을 보여주었다.

이러한 흐름은 도메인 특화 문제에서도 반드시 초대 모델만을 사용할 필요는 없음을 시사한다. 적절한 데이터셋과 효율적 파인튜닝 기법을 결합할 경우, 소형 모델 또한 특정 도메인에서 실용적인 대안이 될 수 있다. 본 연구가 사용하는 Llama 3.2 1B/3B와 Qwen3 1.7B/4B 역시 이러한 소형 모델 활용 흐름 위에 놓여 있으며, 본 연구는 한국어 여행/관광 도메인 데이터 설계와 소형 오픈모델의 효율적 적응을 함께 검증하고자 한다.

3. 방법론

3.1 연구 설계

본 연구는 여행/관광 도메인에서 소형 언어모델의 한국어 응답 성능을 향상시키기 위한 데이터 중심 접근을 제안

한다. 이를 위해 연구는 크게 네 단계로 구성하였다. 첫째, 여행/관광 상황에서 활용 가능한 도메인 특화 질의응답 데이터를 구축하였다. 둘째, 구축한 데이터를 다양한 소형 언어모델에 적용할 수 있도록 입력(Input), 추론(CoT), 출력(Output)의 구조를 갖는 학습 데이터로 정제하였다. 셋째, 정제된 데이터를 이용하여 Llama 3.2 및 Qwen3 계열의 소형 모델에 대해 LoRA 기반 파인튜닝을 수행하였다. 넷째, 파인튜닝 이후의 모델을 원본 모델과 비교하여 도메인 특화 데이터 학습의 효과를 분석하였다.

본 연구의 초점은 특정 오류 현상을 독립적으로 탐지하거나 분류하는 데 있지 않다. 오히려 여행/관광과 같이 지명, 장소 특성, 사용자 제약, 추천 근거가 함께 고려되어야 하는 도메인에서, 구조화된 데이터 학습이 소형 언어모델의 응답 품질에 어떠한 영향을 미치는지를 검증하는 데 목적이 있다. 따라서 본 연구의 방법론은 데이터셋 자체의 설계와 학습 구조를 중심으로 구성되며, 실험 또한 원본 모델과 파인튜닝 모델 간의 성능 비교를 통해 제안 방식의 효과를 확인하는 방향으로 진행하였다.

3.2 여행/관광 도메인 데이터셋 구축

본 연구에서는 여행/관광 도메인에 특화된 질의응답 데이터셋을 구축하였다. 데이터셋은 원천 자료 수집, 추론 가능 구조 반영, 한국어 변환 및 정제, 품질 검토의 순서로 구축하였다. 데이터셋이 여행/관광 특성을 반영할 수 있도록 직접 여행 블로그, 여행 웹사이트, 지역 관광 블로그, TripAdvisor, Redit 등 공개 커뮤니티의 정보를 수집하였다.

수집된 자료는 단순 참고 문서로 사용하지 않고, 실제 사용자가 여행 추천 시스템에 입력할 법한 자연어 질의 형태의 프롬프트로 재구성하였다. 이 과정에서 각 사례는 핵심 여행 과업을 중심으로 정리되었으며, 지역별 어휘와 날짜·단위 표기, 구어적 표현, 팁 문화, 사원 복장 규정, 계절별 기후와 혼잡도, 행사 시기, 예산·접근성·식이 제한 등 지역·문화적 맥락이 함께 반영되도록 정제하였다. 또한 민감한 지정학적 표현은 완화하고, 이중언어 검토와 모의 사용자 테스트를 통해 문장의 자연스러움과 사실성을 점검하였다.

이후 각 프롬프트에는 추론 가능한 구조를 부여하였다. 먼저 날짜, 위치, 예산 상한, 이동성, 식이 제한 등의 명시적 제약을 추출한 뒤, 계절 및 공휴일 확인, 활동 가능성 검토, 교통 동선 설계, 문화적 적합성 판단, 개인화 조건 반영, 일정 순서화의 단계를 거치는 추론(CoT) 경로를 구성하였다. 각 단계에는 간단한 근거를 함께 기록하여 판단 과정의 투명성을 확보하였고, 최종 응답은 이러한 중간 추론 결과와 순

```
{
  "Input": "나는 한국의 주요 도시, 역사적 유적지, 자연 경관을 탐험하며 한 달을 보내고 싶어요. 숙박, 교통, 그리고 통신 상태를 유지하는...",
  "CoT": "1단계: 사용자가 도시, 문화, 자연, 그리고 음식 탐험을 원한다는 목표를 파악한다.
    2단계: 핵심 도시와 경험을 제안한다.
    3단계: 숙박과 현지 교통편에 대해 조언한다.
    4단계: SIM 카드 또는 포켓 와이파이와 같은 통신 팁을 제공한다.
    5단계: 한 달간 여행 계획을 요약한다.",
  "Output": "서울에서 도시의 명소를 시작으로 여행을 시작한 후, 부산의 해변과 해산물을 방문하세요. 경주의 역사 유적지를 탐험하고, 자연의..."
}
```

Figure 1. Example of the Input – CoT – Output structure used in the dataset

서를 반영한 사용자 친화적 추천 문장 또는 일정 형태로 작성하였다.

번역과 정제 과정에서는 단순한 직역이 아니라 한국어 사용자 관점에서 자연스럽게 일관된 질의응답 형식이 되도록 표현을 재구성하였다. 특히 장소명, 지역 맥락, 추천 조건, 이동 제약, 일정 관련 표현이 응답 과정에서 자연스럽게 드러나도록 문장 구조와 어휘를 조정하였다. 또한 수집된 원천 자료는 실제 사용자가 질문하는 방식에 가깝도록 질의 형태로 재작성하고, 이에 대응하는 응답 역시 여행·관광 도메인에서 요구되는 정보 배치와 설명 방식에 맞추어 정리하였다. 이를 통해 여행·관광 도메인에서 빈번하게 등장하는 조건부 추천 상황과 지역 맥락을 학습 가능한 형태로 반영하고자 하였다. 이러한 절차를 통해 총 96,672개의 데이터를 구축하였으며, 최종 데이터의 형태는 Figure.1의 json 형식으로 구성하였다.

3.3 입력-추론-출력 구조의 학습 데이터 설계

본 연구에서 구축한 데이터셋은 총 96,672개의 데이터로 구성되었으며 입력(Input), 추론(CoT), 출력(Output)의 3필드 구조로 설계하였다. 입력은 사용자가 실제로 제시하는 질의에 해당하며, 출력은 해당 질의에 대해 모델이 최종적으로 생성해야 하는 응답을 의미한다. 그리고 두 요소 사이에 위치하는 추론 필드는, 사용자의 요구를 해석하고 적절한 응답을 도출하는 과정에서 필요한 판단 절차를 명시적으로 기술하도록 구성하였다. 각 필드의 데이터는 공개 커뮤니티를 통해 앞서 수집한 정보를 토대로 지역별, 관광지별로 데이터를 일차적으로 정리하고, 정리한 자료를 토대로 각 필드에 맞추어서 직접 작성하였다.

추론 필드는 크게 네 가지 흐름을 반영한다. 첫째, 사용자의 의도를 해석하고 제약 조건을 파악하는 단계이다. 여기서는 사용자가 무엇을 원하는지, 응답 생성 시 반드시 고려해야 할 조건이 무엇인지 정리한다. 둘째, 상황적 및 지역적 정보를 평가하는 단계이다. 여행/관광 질의는 지역명, 장소 특성, 이동 가능성, 관광 맥락과 밀접하게 연결되므로, 이러한 정보를 판단 과정에 포함할 필요가 있다. 셋째, 사용자 요청에 따라 후보를 필터링하는 단계이다. 이는 여러 가능한 응답 가운데 사용자의 목적에 부합하는 선택지를 남기는 과정으로 볼 수 있다. 넷째, 최종적으로 하나의 추천이나 설명을 선택하여 응답으로 연결하는 단계이다.

이와 같은 구조는 단순한 질문-답변 쌍보다 풍부한 감독 신호를 제공한다는 점에서 의의를 가진다. 일반적인 입력-출력 형식은 최종 결과만 학습하게 하지만, 입력-추론-출력 구조는 모델이 응답을 생성하기까지의 중간 판단 과정을 함께 학습하도록 돕는다. 이는 chain-of-thought 기반 접근이 복합 과제에서 효과를 보인다는 선행연구[12]의 관점과도 연결된다. 특히 여행/관광 도메인처럼 조건과 맥락을 단계적으로 고려해야 하는 과업에서는, 최종 응답의 표면적 문장뿐 아니라 판단의 순서와 근거를 학습 데이터에 포함하는 것이 응답의 일관성과 적합성을 높이는 데 유리하다.

또한 본 연구의 3필드 구조는 다양한 소형 언어모델에 공통적으로 적용할 수 있도록 설계되었다는 점에서도 의미가 있다. 이는 특정 모델에만 맞춘 일회성 데이터가 아니라, 여러 모델 계열에 동일한 의미 구조를 제공하는 데이터 설계라는 점에서 소형 오픈모델 전반의 도메인 적응 가능성을 검토하기 위한 기반으로 기능한다.

3.4 실험 모델

본 연구에서는 소형 언어모델을 대상으로 데이터셋 기반 파인튜닝의 효과를 검증하기 위해 Llama 3.2 1B/3B와 Qwen3 1.7B/4B를 실험 모델로 선정하였다. 이들 네 모델은 서로 다른 모델 계열에 속하면서도 비교적 작은 규모의 오픈모델이라는 공통점을 가지므로, 본 연구의 목적과 부합한다.

두 개의 서로 다른 모델 계열을 함께 선택한 이유는 제한된 데이터셋과 학습 구조의 효과를 보다 폭넓게 확인하기 위함이다. 하나의 모델군만 사용할 경우 특정 아키텍처나 사전학습 특성에 따라 결과가 제한적으로 해석될 가능성이 있다. 반면 Llama 계열과 Qwen 계열을 함께 사용하면, 모델의 계열이 달라지더라도 여행/관광 도메인 데이터 학습이 일관된 성능 향상으로 이어지는지를 비교할 수 있다. 또한 1B급과 3B-4B급 모델을 함께 포함함으로써, 모델 규모 차이에 따라 도메인 적응 효과가 어떻게 나타나는지도 함께 관찰할 수 있다.

아울러 본 연구가 소형 모델에 주목한 이유는 실제 활용 가능성과도 관련이 있다. 상대적으로 작은 모델은 연산 자원과 메모리 측면에서 효율적이기 때문에, 향후 교육·서비스 환경이나 경량 배포 환경에서 적용 가능성이 높다. 그러나 이러한 장점에도 불구하고, 소형 모델은 범용 성능만으로 도메인 특화 질의를 충분히 처리하지 못하는 경우가 있다. 따라서 여행/관광 분야와 같이 맥락 의존성이 강한 과업에서 소형 모델을 대상으로 도메인 특화 학습을 수행하는 것은 연구적·실용적 의미를 동시에 가진다.

3.5 학습 설정

앞선 과정에서 구축한 데이터를 이용하여 파인튜닝을 진행하였다. 파인튜닝은 LoRA 기반의 supervised fine-tuning 방식으로 수행하였다. 손실 함수는 Cross Entropy Loss를 사용하였고, 학습은 5개의 GPU를 활용한 분산 학습 환경에서 진행하였다. 배치 구성은 8행 단위로 설정하였고, 4개 배치마다 gradient accumulation을 수행하도록 구성하였다. 이러한 설정은 제한된 자원 환경에서도 여러 소형 모델에 대해 동일한 도메인 적응 실험을 수행할 수 있도록 하기 위한 것이다.

LoRA는 사전학습된 모델의 전체 파라미터를 모두 갱신하는 대신, 저랭크 적응 행렬만을 추가로 학습하는 방식으로 제한된 파라미터 효율적 미세조정 기법이다[1]. 이 방식은 전체 모델을 전면적으로 업데이트하는 방법에 비해 학습

가능한 파라미터 수를 줄이고, GPU 메모리 부담을 완화하면서도 도메인 적응 효과를 확보할 수 있다는 장점이 있다. 본 연구처럼 복수의 소형 오픈모델을 동일한 데이터셋으로 비교해야 하는 상황에서는, LoRA를 활용한 파인튜닝이 실험 효율성과 재현 가능성 측면에서 적절한 선택이 된다.

또한 입력과 출력 사이에 추론 필드를 포함하는 본 연구의 데이터 구조는 일반적인 짧은 instruction 데이터보다 시퀀스 길이가 길어질 가능성이 있다. 따라서 배치 크기를 작게 유지하면서 gradient accumulation을 통해 학습 안정성을 확보하는 구성은 자원 효율성과 학습 품질을 함께 고려한 실용적 설정으로 볼 수 있다.

다만 Llama 3.2 1B, 3B 모델은 출력을 생성하는 과정에서 추론 기능을 제공하지 않기에 학습에는 추론 필드의 데이터를 제외한 입력 - 출력 필드의 데이터만 사용하여 파인튜닝을 진행하였다.

4. 실험

4.1 실험 설정

본 연구에서는 여행/관광 도메인 데이터셋 기반 학습이 소형 언어모델의 한국어 응답 품질에 미치는 영향을 검증하기 위해, 원본 모델(Original)과 제안 방법으로 파인튜닝한 모델(Ours)을 비교하였다. 비교 대상 모델은 Llama 3.2 1B, Llama 3.2 3B, Qwen3 1.7B, Qwen3 4B의 네 가지로 구성하였다. Llama 3.2 1B, 3B 모델은 reasoning 기능을 제공하지 않지만, Qwen3 1.7B, 3B 모델은 reasoning 기능을 제공하기에 Qwen3 모델에서는 reasoning 기능을 사용하도록 설정하였다. 모든 모델에 동일한 입력 필드의 데이터를 사용하여 출력을 생성하도록 하였다. 출력을 생성함에 있어서 입력 필드의 데이터를 제외한 시스템 프롬프트는 제공하지 않았다. 또한 본 연구의 초점이 데이터셋 기반 학습의 효과 검증에 있는 만큼, 결과 분석은 원본 모델과 파인튜닝 모델의 차이를 중심으로 수행하였다.

4.2 평가 지표

모델의 생성 품질 평가는 BERTScore, BLEU, ROUGE 계열 지표를 중심으로 수행하였다. BERTScore는 생성 문장과 정답 문장 간의 의미적 유사도를 contextual embedding 기반으로 측정하는 지표이며, BLEU와 ROUGE는 참조 문장과 n-gram 중첩 정도를 기반으로 생성 결과의 정합성을 평가하는 대표적인 자동 평가 지표이다[16-18]. 본 연구에서는 BERTScore의 F1, Precision, Recall을 산출하였고, BLEU는 1-gram부터 3-gram까지, ROUGE는 ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L을 사용하였다.

본 논문의 핵심 목적은 여행/관광 도메인 데이터셋 학습이 응답 품질을 얼마나 개선하는지 확인하는 데 있으므로, 결과 분석에서는 BERTScore(F1), BLEU-1, ROUGE-L을

대표 지표로 우선 해석하고, 나머지 세부 지표는 이를 보조하는 근거로 활용하였다.

4.3 실험 결과 및 분석

Table 1과 Table 2는 Llama 3.2 계열 모델에 대한 실험 결과를 보여준다. 먼저 1B 모델의 경우, 제안 방법 적용 후 BERTScore(F1)는 0.8295에서 0.8687로 향상되었고, BLEU-1은 17.5227에서 30.9466로, ROUGE-L은 12.7561에서 28.4025로 증가하였다. 또한 BLEU-2와 BLEU-3은 각각 6.6777에서 22.1059, 2.8598에서 16.8712로 개선되었다. 이는 제안한 데이터셋 기반 학습이 단순한 의미 보존 수준을 넘어, 참조 응답과의 어휘적·구조적 정합성을 함께 향상시켰음을 시사한다.

Llama 3.2 3B 모델에서도 동일한 경향이 확인되었다. BERTScore(F1)는 0.8456에서 0.8763로 향상되었고, BLEU-1은 22.0922에서 32.5194로, ROUGE-L은 14.7433에서 30.6090로 증가하였다. 세부적으로 보면 BLEU-2는 9.2038에서 24.0436로, BLEU-3은 4.2895에서 18.9116으로 향상되었으며, ROUGE-2는 4.4435에서 22.5083, ROUGE-3은 1.0792에서 14.7594로 상승하였다. 특히 ROUGE 계열 지표의 향상폭이 크게 나타난 것은 파인튜닝 이후 생성 응답이 기준 응답의 정보 구성과 문장 전개를 보다 안정적으로 반영하게 되었음을 보여준다. 1B와 3B 모두에서 원본 모델 대비 일관된 성능 향상이 나타났으며, 상대적으로 큰 모델인 3B가 파인튜닝 후에도 전반적으로 더 높은 성능을 기록하였다. 이는 도메인 특화 데이터셋의 효과가 모델 규모와 무관하게 유효하면서도, 일정 수준 이상의 모델 용량이 확보될 경우 그 효과가 더욱 안정적으로 발현될 수 있음을 보여준다.

Table 1. Llama3.2 1B result

Metric	Original	Ours
BERT-Score(F1)	0.8295	0.8687
BERT-Score(P)	0.8321	0.8885
BERT-Score(R)	0.8274	0.8782
BLEU-1	17.5227	30.9466
BLEU-2	6.6777	22.1059
BLEU-3	2.8598	16.8712
ROUGE-1	20.7484	38.9386
ROUGE-2	3.0401	19.5333
ROUGE-3	0.6353	12.0549
ROUGE-L	12.7561	28.4025

Table 2. Llama3.2 3B result

Metric	Original	Ours
BERT-Score(F1)	0.8456	0.8763
BERT-Score(P)	0.8500	0.8948
BERT-Score(R)	0.8415	0.8852
BLEU-1	22.0922	32.5194

Metric	Original	Ours
BLEU-2	9.2038	24.0436
BLEU-3	4.2895	18.9116
ROUGE-1	25.9821	41.5815
ROUGE-2	4.4435	22.5083
ROUGE-3	1.0792	14.7594
ROUGE-L	14.7433	30.6090

Table 3과 Table 4는 Qwen3 계열 모델의 결과를 나타낸다. 1.7B 모델의 경우, BERTScore(F1)는 0.8174에서 0.8717로 향상되었고, BLEU-1은 6.6678에서 25.0055으로, ROUGE-L은 5.9550에서 30.0669로 증가하였다. BLEU-2와 BLEU-3 역시 각각 0.6843에서 18.5202, 0.1893에서 14.5825로 크게 향상되었다. 이러한 결과는 원본 모델이 여행/관광 도메인의 한국어 응답 구조를 충분히 반영하지 못했던 반면, 제안한 데이터셋 기반 학습 이후에는 참조 응답과의 정합성이 현저히 개선되었음을 의미한다.

Table 3. Qwen3 1.7B result

Metric	Original	Ours
BERT-Score(F1)	0.8174	0.8717
BERT-Score(P)	0.8181	0.8969
BERT-Score(R)	0.8169	0.8840
BLEU-1	6.6678	25.0055
BLEU-2	0.6843	18.5202
BLEU-3	0.1893	14.5825
ROUGE-1	6.7485	40.2239
ROUGE-2	0.0782	21.6985
ROUGE-3	0.0164	14.1951
ROUGE-L	5.9550	30.0669

Table 4. Qwen3 4B result

Metric	Original	Ours
BERT-Score(F1)	0.8167	0.8774
BERT-Score(P)	0.8169	0.9011
BERT-Score(R)	0.8167	0.8890
BLEU-1	6.1656	26.2102
BLEU-2	0.6121	19.9142
BLEU-3	0.1656	16.0237
ROUGE-1	6.3002	42.5222
ROUGE-2	0.0673	24.2225
ROUGE-3	0.0134	16.5146
ROUGE-L	5.6259	32.2617

Qwen3 4B 모델은 본 실험에 포함된 모델 가운데 가장 높은 성능을 기록하였다. BERTScore(F1)는 0.8167에서 0.8774로 향상되었고, BLEU-1은 6.1656에서 26.2102로, ROUGE-L은 5.6259에서 32.2617로 증가하였다. 추가적으로 BLEU-2는 0.6121에서 19.9142로, BLEU-3은 0.1656에서 16.0237로 상승하였으며, ROUGE-2는 0.0673에서

24.2225, ROUGE-3은 0.0134에서 16.5146으로 향상되었다. 특히 Qwen3 4B는 본 실험에 포함된 모델 가운데 가장 높은 BERTScore(F1)와 ROUGE-L을 기록하여, 제안한 데이터셋이 의미적 유사도와 문장 수준 포괄성 측면에서 매우 효과적으로 작동했음을 보여준다.

종합하면, 제안 방법은 네 가지 모든 실험 모델에서 원본 모델 대비 일관된 성능 향상을 보였다. BERTScore(F1)는 최소 0.0307에서 최대 0.0607까지 향상되었고, BLEU-1은 10.4272에서 20.0446 범위의 증가를 보였으며, ROUGE-L은 15.6464에서 26.6358까지 향상되었다. 이는 본 연구에서 구축한 여행/관광 도메인 데이터셋과 입력-추론-출력 구조의 학습 방식이 소형 언어모델 전반에 걸쳐 안정적인 성능 개선 효과를 제공하였음을 의미한다. 특히 BERTScore의 상승폭은 비교적 완만한 반면 BLEU와 ROUGE의 향상은 훨씬 크게 나타났는데, 이는 원본 모델도 일정 수준의 의미적 관련성은 유지하고 있었으나, 제안 방법 적용 후에는 응답의 표현 방식과 정보 구성, 문장 구조가 참조 응답에 더 가깝게 정렬되었음을 시사한다. 다시 말해, 본 연구의 데이터셋은 단순히 정답 의미를 맞추는 데 그치지 않고, 여행/관광 도메인에서 요구되는 응답 형식과 서술 구조까지 학습시키는 데 기여한 것으로 해석할 수 있다.

5. 결론

본 연구는 여행/관광 도메인에서 소형 언어모델의 한국어 응답 품질을 향상시키기 위한 데이터 중심 접근을 제안하였다. 이를 위해 여행/관광 질의를 대상으로 한국어 데이터셋을 구축하고, 이를 입력(Input), 추론(CoT), 출력(Output)의 3필드 구조로 정제하였다. 또한 해당 데이터를 활용하여 Llama 3.2 1B/3B와 Qwen3 1.7B/4B 모델에 대해 LoRA 기반 파인튜닝을 수행하고, 원본 모델과의 비교를 통해 구조화된 도메인 데이터가 소형 모델의 응답 생성에 미치는 영향을 분석하였다.

실험 결과, 제안한 방법은 모든 실험 모델에서 원본 모델 대비 일관된 성능 향상을 보였다. 예를 들어 Llama 3.2 3B 모델은 BERTScore(F1)가 0.8456에서 0.8763로, ROUGE-L이 14.7433에서 30.6090로 향상되었으며, Qwen3 4B 모델은 BERTScore(F1)가 0.8167에서 0.8774로, ROUGE-L이 5.6259에서 32.2617로 증가하였다. 이러한 결과는 여행/관광 도메인에 맞추어 설계된 학습 데이터와 추론 구조가 소형 언어모델의 한국어 응답 품질을 실질적으로 개선하는 데 효과적임을 보여준다. 특히 본 연구는 초거대 모델이 아닌 소형 오픈모델에서도 도메인 특화 데이터셋을 통한 성능 향상이 가능함을 확인하였다는 점에서 의의를 가진다.

다만 본 연구는 자동 평가 지표 중심으로 성능을 분석하였으며, 실제 사용자 만족도나 추천 적절성을 반영하는 사람 평가를 포함하지 못했다는 한계가 있다. 또한 데이터셋

이 여행/관광 도메인에 한정되어 있어 다른 도메인으로의 일반화 가능성은 추가 검토가 필요하다. 향후 연구에서는 사람 평가를 포함한 다면적 성능 분석을 수행하고, 데이터 규모 및 유형을 확장하며, 다양한 소형 모델 및 외부 지식 활용 방식과의 비교를 통해 도메인 적응 전략의 효과를 보다 정교하게 검증할 필요가 있다.

참고문헌

- [1] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2106.09685>
- [2] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 10088-10115. <https://doi.org/10.48550/arXiv.2305.14314>
- [3] Abidin, M. I., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H. H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Mendes, C. C. T., Chen, W., Chaudhary, V., Chopra, P., Del Giorno, A., de Rosa, G., Dixon, M., Eldan, R., Iter, D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Huynh, J., Javaheripi, M., Jin, X., Kauffmann, P., Karampatziakis, N., Kim, D., Khademi, M., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Liang, C., Liu, W., Lin, X., Lin, Z., Madan, P., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Rosset, C., Roy, S., Saarikivi, O., Saied, A., Salim, A., Santacrose, M., Shah, S., Shang, N., Sharma, H., Song, X., Ruwase, O., Wang, X., Ward, R., Wang, G., Witte, P., Wyatt, M., Xu, C., Xu, J., Xu, W., Yadav, S., Yang, F., Yang, Z., Yu, D., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone (MSR-TR-2024-12). Microsoft. <https://doi.org/10.48550/arXiv.2404.14219>
- [4] Bahramian, Z., Ali Abbaspour, R., & Claramunt, C. (2017). A context-aware tourism recommender system based on a spreading activation method. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W4, 333-339. <https://doi.org/10.5194/isprs-archives-XLII-4-W4-333-2017>
- [5] He, J., Qi, J., & Ramamohanarao, K. (2018). A jointly learned context-aware place of interest embedding for trip recommendations. *arXiv preprint arXiv:1808.08023*. <https://doi.org/10.48550/arXiv.1808.08023>
- [6] Tsai, C.-Y., Chuang, K.-W., Jen, H.-Y., & Huang, H. (2024). A tour recommendation system considering implicit and dynamic information of POI. *Applied Sciences*, 14(20), 9271. <https://doi.org/10.3390/app14209271>
- [7] Wei, Q., Yang, M., Wang, J., Mao, W., Xu, J., & Ning, H. (2025). CTourLLM: Enhancing LLMs with Chinese tourism knowledge. *Cluster Computing*, 28(6), 351. <https://doi.org/10.1007/s10586-024-05009-x>
- [8] Song, S., Yang, C., Xu, L., Shang, H., Li, Z., & Chang, Y. (2024). TravelRAG: A tourist attraction retrieval framework based on multi-layer knowledge graph. *ISPRS International Journal of Geo-Information*, 13(11), 414. <https://doi.org/10.3390/ijgi13110414>
- [9] Karlović, R., Rovis, M., Smajić, A., Sever, L., & Lorencin, I. (2025). Context-aware tourism recommendations using retrieval-augmented large language models and semantic re-ranking. *Electronics*, 14(22), 4448. <https://doi.org/10.3390/electronics14224448>
- [10] Deng, B., Feng, Y., Liu, Z., Wei, Q., Zhu, X., Chen, S., Guo, Y., & Wang, Y. (2025). RETAIL: Towards real-world travel planning for large language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China*, 14870-14902. <https://doi.org/10.18653/v1/2025.emnlp-main.752>
- [11] Chu, M., Chen, Y., Gui, H., Yu, S., Wang, Y., & Jia, J. (2026). TravelLaMA: A multimodal travel assistant with large-scale dataset and structured reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5), 3390-3398. <https://doi.org/10.1609/aaai.v40i5.37335>
- [12] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837. <https://doi.org/10.48550/arXiv.2201.11903>
- [13] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., & Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70), 1-53. <https://www.jmlr.org/papers/v25/23-0870.html>
- [14] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-Instruct: Aligning language models with self-generated instructions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, 13484-13508. <https://doi.org/10.18653/v1/2023.acl-long.754>
- [15] Qi, R., Wen, C., Zhou, W., Liang, S., & Li, L. (2025). LETToT: Label-free evaluation of large language models on tourism using expert tree-of-thought. *arXiv preprint arXiv:2508.11280*. <https://doi.org/10.48550/arXiv.2508.11280>
- [16] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q.,

& Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1904.09675>

- [17] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA*, 311-318. <https://doi.org/10.3115/1073083.1073135>
- [18] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, Barcelona, Spain, 74-81.



오민수

· 2022년~현재 제주대학교 소프트웨어학부 인공지능학과 학사과정

✚ 관심분야 : 자연어처리, 딥러닝

✉ dhals0517@gmail.com



윤여찬

· 2004년 고려대학교 컴퓨터학과 졸업(학사)

· 2007년 고려대학교 컴퓨터학과 졸업(석사)

· 2020년 고려대학교 컴퓨터학과 졸업(박사)

· 2007~2022년 한국전자통신연구원 책임연구원

· 2022년~현재 제주대학교 인공지능전공 조교수

✚ 관심분야 : 자연어처리, 멀티모달 딥러닝

✉ ycyoon@jejunu.ac.kr