2025학년도 1학기 일반대학원생 대상 연구역량 강화 프로그램 성과

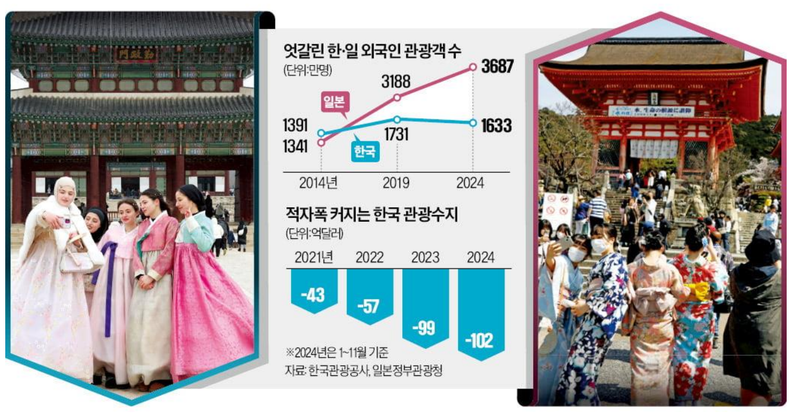
관광 활성화를 위한 거대 언어 모델 및 RAG 파이프라인 구축 연구

서울시립대학교 일반대학원 도시빅데이터융합학과 이다운

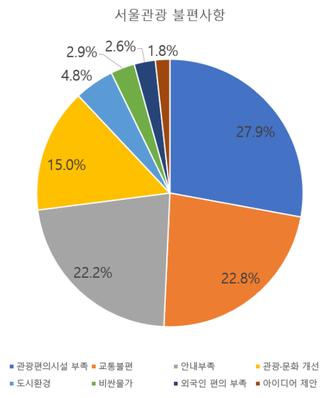
연구 요약

본 연구는 한국 방문 일본인 관광객의 언어 장벽 문제를 해결하기 위해 AI Hub의 K-Culture 관광 콘텐츠 특화 일본어 말뭉치 100K 문장쌍을 활용한 온디바이스 RAG 시스템을 개발하였다. 5종의 소형 언어모델을 MMLU 벤치마크로 평가한 결과 Qwen3-4B가 한국어 53.34%, 일본어 55.10%의 최고 성능을 보여 베이스 모델로 선정하였다. Vector RAG, Light RAG, Graph RAG를 융합한 앙상블 아키텍처를 설계하여 답변 관련성 62.2%, 응답시간 9.2초, 오류율 0%의 실용적 성능을 달성하였다. 메모리 사용량 36GB로 온디바이스 환경에서 구동 가능하며, 관광 정보 제공의 정확도와 접근성을 크게 개선한 실용적 솔루션을 제시하였다.

연구 배경 및 목적



출처 : 한국경제(이선아, 2025)



출처 : 트레블러 뉴스(심우리, 2020)

연구 배경 및 필요성

○ **외국인 관광객 급증 vs 낮은 정보 접근성**: 한국 방문 외국인 관광객이 2019년 1,750만 명에 달했으나, 언어 장벽과 디지털 서비스 한계로 정보 접근성 만족도가 여전히 낮음

○ **일본어 관광 정보 서비스 부족**: 온라인 관광 정보가 방문지 결정에 중요 영향을 미치나, 현지화된 일본어 지원 미흡으로 일본인 관광객의 정보 탐색이 제한됨

○ **AI 기반 관광 서비스 기술 필요성**: K-Culture 관광 콘텐츠 특화 일본어 말뭉치(464,932건, 177,620,461어절) 활용한 맞춤형 정보 서비스 개발 필요

연구 목표

○ **데이터 기반 구축**: AI Hub K-Culture 관광 콘텐츠 특화 일본어 말뭉치 기반 인덱싱 및 검색 데이터베이스 구축

○ **최적 모델 선정**: 온디바이스 환경에서 효율적 구동이 가능한 소형 언어모델 선정 및 일본어 관광 질의 처리 성능 최적화

○ **앙상블 RAG 개발**: Vector RAG, Light RAG, Graph RAG 융합한 앙상블 파이프라인 설계 및 구현

○ **성능 정량 평가**: 검색 정확도, 응답 속도, 일관성 측면에서 제안 시스템의 성능 정량적 검증

연구 방법

1. 데이터 준비 및 전처리

• **데이터셋**: AI Hub K-Culture 관광 콘텐츠 특화 일본어 말뭉치 활용

• **구성**: 한국 주요 관광지, 문화 콘텐츠, 관광 서비스 정보의 한국어-일본어 대역 코퍼스

• **전처리**: 온디바이스 환경 제약을 고려한 10% 샘플링 적용

2. 언어모델 선정 및 평가

•**후보 모델**: 온디바이스 구동 가능한 5종 소형 언어모델 (3-4B 파라미터)
- Qwen3-4B, Gemma 3-4b-it, Llama-3.2-3B, Korean-Bllossom-3B, HyperCLOVAX-3B

•**성능 평가**: MMLU 벤치마크를 통한 한국어/일본어 성능 측정

•**속도 평가**: Token throughput, TTFT, TPOT, ITL 지표로 추론 속도 분석

•**최종 선정**: Qwen3-4B (한국어 53.34%, 일본어 55.10% 최고 성능)

3. 앙상블 RAG 시스템 설계

•**Vector RAG**: FAISS 기반 코사인 유사도 벡터 검색

•**Light RAG**: K-means 클러스터링 기반 이중 레벨 검색

• **Graph RAG**: 문서 간 의미적 유사도 그래프 및 노드 중심성 활용

•**앙상블 융합**: 가중 평균 및 재순위화 메커니즘으로 최종 검색 결과 도출

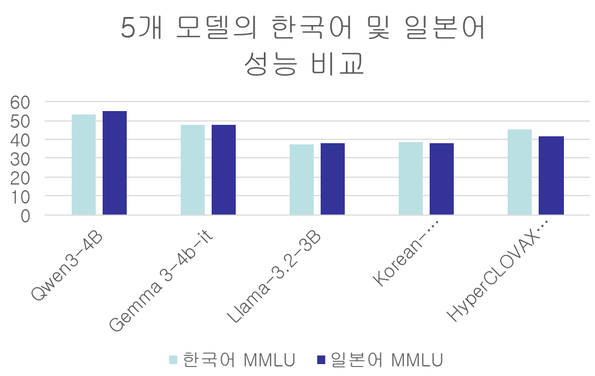
4. 성능 평가 및 분석

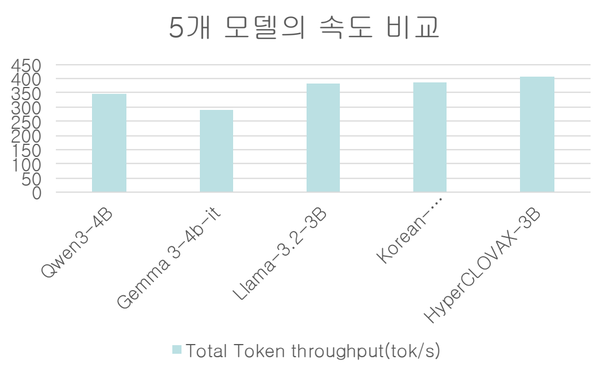
•**평가 지표**: 검색 정확도, 답변 관련성, 답변 충실도, 응답시간, BLEU/ROUGE 점수

•**비교 대상**: EnsembleRAG, LightRAG, GraphRAG, VectorRAG 4개 시스템

연구 결과

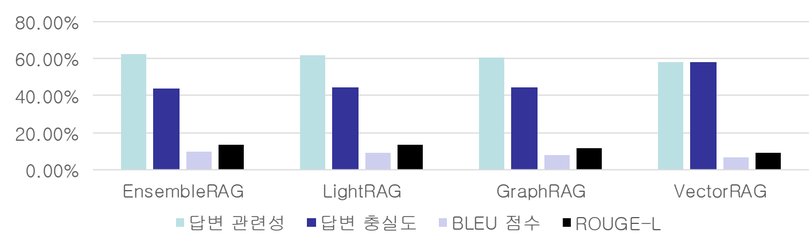
1. 언어모델 선정 및 평가

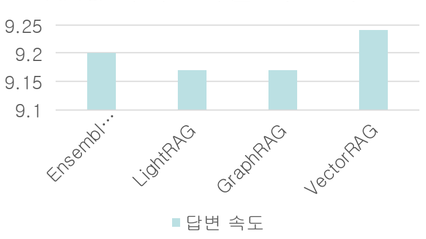




언어모델 평가에서 Qwen3-4B가 한국어 53.34%, 일본어 55.10%의 최고 성능을 달성하여 다국어 관광 정보 서비스에 가장 적합한 모델임을 확인하였다. 특히 관광 정보와 밀접한 사회과학 분야에서 한국어 58.30%, 일본어 61.55%의 뛰어난 성능을 보여 도메인 특화 효과를 입증하였다. 추론 속도 측면에서는 3B 모델들이 더 빠른 처리 속도를 보였으나, Qwen3-4B는 4B 파라미터임에도 불구하고 345.88 tok/s의 경쟁력 있는 속도를 유지하여 성능과 효율성의 균형을 달성하였다.

2. RAG 파이프라인 구성 및 평가





RAG 시스템 성능 평가에서 EnsembleRAG가 답변 관련성 62.2%로 최고 성능을 기록하며, VectorRAG(57.8%) 대비 4.4%p 향상된 결과를 보였다. 이는 여러 검색 방식의 상호 보완적 융합 효과로 해석되며, Vector RAG의 정확한 벡터 매칭, Light RAG의 의미적 클러스터링, Graph RAG의 문서 간 관계 정보가 종합적으로 작용한 결과이다. 답변 충실도에서는 LightRAG와 GraphRAG가 44.6%로 동일한 최고 성능을 보여 클러스터링과 그래프 기반 접근법이 검색된 문서와의 일치도 향상에 효과적임을 확인하였다. 모든 시스템에서 0%의 오류율을 달성하여 실용적 배포 가능성을 입증하였으며, 9.1-9.2초의 일관된 응답 시간으로 실시간 관광 정보 서비스 요구사항을 충족하였다.

언어학적 지표에서 EnsembleRAG는 BLEU 점수 9.4%, ROUGE-L 점수 13.5%를 기록하여 가장 높은 언어 품질을 보였다. 이는 앙상블 접근법이 단순히 검색 정확도뿐만 아니라 생성되는 답변의 언어적 품질까지 향상시킴을 의미한다. 검색 정확도가 모든 시스템에서 15.0%로 동일하게 나타난 것은 평가 데이터셋의 난이도와 Top-5 검색 설정의 제약으로 분석되며, 이는 향후 더 도전적인 평가 환경에서의 추가 검증 필요성을 시사한다.

결론

본 연구는 한국 방문 일본인 관광객을 위한 온디바이스 AI 관광 정보 서비스를 개발하였다. Qwen3-4B 모델이 한국어 53.34%, 일본어 55.10%로 최고 성능을 보였으며, Vector RAG, Light RAG, Graph RAG를 융합한 앙상블 시스템을 통해 답변 관련성 62.2%, 응답시간 9.2초, 오류율 0%를 달성하였다. 메모리 사용량 36GB로 실제 배포 가능한 수준의 효율성을 확보하였고, 앙상블 접근법이 단일 시스템 대비 4.4%p 성능 향상을 보여 융합 효과를 입증하였다.

본 연구는 온디바이스 환경에서 구동 가능한 다국어 RAG 시스템의 실용적 구현 방법을 제시하였으며, 관광 도메인에 특화된 성능 최적화 기법을 개발하였다. 또한 소형 언어모델의 다국어 성능 비교 분석을 통해 실제 서비스 환경에서의 모델 선정 기준을 제공하였고, 앙상블 RAG 아키텍처의 효과를 정량적으로 검증하여 향후 유사 연구의 기술적 기반을 마련하였다. 이러한 결과는 외국인 관광객 대상 정보 서비스 개선과 관련 기술 발전에 기여할 것으로 기대된다.

서울시립대학교 UNIVERSITY OF SEOUL