# RAG 챗봇의 기능 테스트를 위한 질문 설계 체계화 연구

최연주, 김양중*

한국공학대학교 소프트웨어융합공학과
{moonwere, *zeroplus}@tukorea.ac.kr

# A Study on the Systematic Design of Questions for Functional Testing of RAG Chatbots

Yeonjoo Choi, Yangjung Kim*
Tech University of Korea

요 약

생성형 AI 의 발전으로 RAG 기반 챗봇의 응답 정확도가 향상되고 있다. 이에 따라 RAG 챗봇의 기능을 평가할 때도, 보다 체계적인 질문 설계가 필요해졌다. 본 연구에서는 질문을 정보 요구 수준, 문장 구조, 도메인 적합성의 세 가지 기준으로 분류해, 총 60 문항을 작성하여 RAG 챗봇의 기능 테스트를 진행하고 결과를 확인하였다. 본 연구를 통해 구조화된 질문을 통해 체계적으로 RAG 챗봇의 기능을 테스트 하는 것이 중요함을 검증한다.

## Ⅰ. 서 론

ChatGPT 를 비롯한 대규모 언어 모델(LLM, Large Language Model)의 발전은 생성형 AI 기반 대화 시스템의 효율성과 대중화를 이끌며 일상에서 LLM 챗봇의 활용도가 높아졌다[1]. 이러한 기술의 발전을 통해 비전문가도 손쉽게 챗봇을 구축할 수 있는 환경이 조성되었다. 최근에는, 문서 기반 정보 검색과 자연어 생성을 결합한 Retrieval-Augmented Generation(RAG) 구조 기반이 특정 도메인의 지식을 효과적으로 활용하는 솔루션으로 주목받고 있다[2].

RAG 기반 챗봇의 성능은 단순히 모델의 응답 품질만으로 평가하기가 쉽지 않다. 질문의 구조와 표현 방식에 따라 검색 결과 및 응답 내용이 상이하게 달라질 수 있기 때문이다. 챗봇 응답 품질은 질문의 표현 방식, 의도 및 맥락 민감도 등 다양한 요소에 영향을 받으며[3], 이에 따라 단순 정답 여부만으로는 기능을 평가하는데 신뢰성을 보장할 수 없다. 한정된 테스트 단계에서 적은 질문으로 챗봇의 기능을 효율적으로 검증하기 위해서는 질문의 구조화와 체계적인 설계 기준 마련이 필요하다. 따라서, 본 연구는 RAG 챗봇의 기능 테스트를 위한 질문 설계의 체계화 방안 및 질문 유형에 따른 응답 특성을 분석해 보고, 제시한 방안을 기반으로 신뢰성 있는 챗봇 평가 기준을 마련하고자 한다.

## Ⅱ. 본론

### 1. RAG 기반 챗봇의 구조와 특징

RAG 은 외부 지식소스를 검색하여 대화 응답 생성을 보완하는 구조로, 단일 LLM 기반 챗봇보다 신뢰도 높은 응답을 제공할 수 있다[2]. 일반적으로 RAG 는 질문에 기반한 검색(Retrieval) 단계와 검색된 문서를 바탕으로 응답을 생성하는 생성(Generation) 단계로 구성된다. 이 구조는 문서 기반의 지식을 포함한 질문을 처리하기 위한 응답 시스템으로써 대표적 발전 형태이다[5]. [3][4]에서는 의도하는 바가 같다고 해도 프롬프트 표현에 따라 챗봇의 반응이나 응답 결과가 상반될 수 있다는 연구결과를 보여주고 있다. 질문 표현 방식이나 의도 해석 방식에 따라 검색 결과가 민감하게 변경될 수 있으며, 결국 생성된 응답의 품질을 신뢰할 수 없게 된다. 결국, RAG 챗봇의 기능을 평가하기 위해서는 구조화된 질문 설계가 필수적이다.

### 2. 질문 유형의 분류 체계 연구

본 논문에서는 질문 설계를 구조화하기 위해 다음 세 가지 기준을 제안한다:

- 정보 요구 수준: 사실 확인형, 추론형, 종합형
- 문장 구조: 의문형, 명령형, 비교형
- 도메인 적합성: 범용 질문, 도메인 특화 질문

이러한 유형 분류는 단순 분류가 아닌, 검색 적합도 및 응답 생성 품질의 향상과 직결된다. 이는 IBM Watson, TREC QA 평가 등 다양한 QA 시스템에서 질문 설계의 구조화가 시스템 성능뿐만 아니라 생성된 응답의 품질을 자동으로 평가하는 지표에도 영향을 미치는 것으로 나타난다 [5][6].

### 3. 실험 설계 및 결과 분석

본 논문에서는 구조적으로 설계한 질문이 RAG 기반 챗봇의 기능 테스트에서 결과에 어떠한 영향을 미치는지를 분석하였다. 이를 위해서, 앞서 제시한 질문 유형 분류 기준에 따라 총 60 문항의 테스트 질문 세트를 구성해 각 질문이 챗봇의 응답 품질에 미치는 영향을 중점으로 분석, 평가한 결과를 제시한다. 질문 세트는 동일한 정보 요구를 기반으로 하되, 표현 방식이나 구체성 등 설계 변수를 달리한 질문 쌍을 포함하였다. 예를 들어, 'OO 는 무엇인가요?'와 'OO 의 주요 기능을 예시와 함께 설명해주세요'는 동일한 정보를 요구하지만 문장 구조와 명시 수준이 상이한 질문이다.

챗봇의 응답은 다음과 같은 네 가지 항목을 중심으로 평가하였다.

- **정확도 (Correctness)**: 응답이 문서 기반 사실과 얼마나 일치하는지를 평가한다.
- **검색 적합도 (Retrieval Relevance)**: 챗봇이 활용한 문서가 질문의 주제와 논리적으로 연결되는지를 확인한다.

- **질문 민감도 (Question Sensitivity)**: 동일한 의미를 가진 질문이라도 표현 방식이 달라졌을 때, 응답 내용이 얼마나 달라지는지를 측정한다.
- **응답 일관성 (Answer Stability)**: 표현이 다른 유사 질문에 대해 챗봇이 일관된 응답을 생성하는지를 평가한다.

이러한 기준은 RAG 챗봇이 단순한 정보 검색 시스템이 아닌, 사용자의 질문 의도를 해석하고 문맥에 맞는 응답을 생성하는 복합 시스템을 고려해 설정되었다. 실험은 AI 개발자 세 명의 수작업 분석과 함께, BLEU 점수와 Sentence-BERT 기반의 Semantic Similarity 등 자동화된 평가 도구를 병행하여 수행되었다[7][8].

<표 1. 질문 유형별 응답 품질 평가표>

| 질문 유형 | 세부 분류 | BLEU 점수 | Semantic Similarity | 정확도 | 검색 적합도 | 질문 민감도 | 응답 일관성 |
|---|---|---|---|---|---|---|---|
| 정보 요구 수준 | 사실 확인형 | 0.78 | 0.89 | 4.5 | 4.6 | 4.3 | 4.4 |
| | 추론형 | 0.61 | 0.73 | 3.8 | 4.0 | 3.6 | 3.7 |
| | 종합형 | 0.67 | 0.79 | 4.2 | 4.3 | 4.1 | 4.2 |
| 문장 구조 | 의문문 | 0.75 | 0.87 | 4.6 | 4.7 | 4.4 | 4.5 |
| | 명령문 | 0.58 | 0.70 | 3.5 | 3.6 | 3.2 | 3.4 |
| | 비교문 | 0.63 | 0.76 | 4.0 | 4.2 | 3.8 | 4.0 |
| 도메인 적합성 | 범용 질문 | 0.60 | 0.72 | 3.7 | 3.9 | 3.5 | 3.6 |
| | 도메인 특화 질문 | 0.79 | 0.90 | 4.6 | 4.8 | 4.5 | 4.7 |

실험 결과, 질문 설계 방식에 따라 챗봇 응답 품질에서 확연한 차이를 보였으며, 정보 요구 수준에 따라 사실 확인형 질문은 평균 약 90%의 정확도를 기록한 반면, 추론형과 비교형 질문은 상대적으로 낮은 정확도를 보였다. 문장 구조 측면에서는 의문문 형태의 질문이 가장 높은 검색 적합도를 보였으며, 평서문이나 명령문 형태는 문맥 해석이 불명확해 응답의 정확성과 일관성이 저하되는 경향을 나타냈다. 또한 질문이 구체적일수록 응답의 정보 밀도와 구조화 수준이 향상되었으며, 유사 질문 간의 일관성도 개선되었다. 예컨대, 단순히 'OO 의 기능을 설명해주세요'라고 질의하는 것보다, 'OO 의 핵심 기능 세 가지를 예시와 함께 설명해주세요'와 같은 구체적 질문이 더 정밀한 문서 검색과 명확한 응답 생성으로 결과적 신뢰도를 제시하였다. 질문에 대한 문서 기반 도메인과 밀접하게 연관되어 있을수록 검색된 문서의 일치율과 응답의 안정성 또한 높게 도출되는 것으로, 질문 설계가 RAG 챗봇의 기능 평가에 있어 단순한 입력 요소가 아니라, 평가 결과를 좌우하는 핵심 변수임을 실증적으로 보여준다. 결국, 구조화된 질문을 기반으로 한 체계적인 기능 테스트는 RAG 기반 챗봇 검증 체계의 신뢰성과 표준화 가능성을 높이는데 기여할 수 있다.

## Ⅲ. 결론

본 연구는 RAG(Retrieval-Augmented Generation) 기반 챗봇의 기능을 정밀하게 평가하기 위해, 질문 설계의 체계화를 중심으로 기능 테스트 방법을 고찰해 검증의 방법론적 기반을 제시하였다. 기존 챗봇 평가가 단순 응답 유무나 정확성에 초점을 맞췄다면, 본 연구는 질문 자체가 응답 품질에 영향을 미치는 중요한 변수임을 실험을 통해 확인하였다. 실험 결과, 질문의 의도 유형, 문장 구조, 구체성, 도메인 적합성에 따라 응답 정확도와 일관성에 유의미한 차이가 발생하였으며, 특히 의문문 구조의 구체적이면서 체계적인 질문과 도메인 일치도가 높은 질문에서 응답 품질이 향상됨을 확인할 수 있었다.

향후 연구에서는 본 연구에서 제안한 질문 분류 기준과 평가 지표를 다양한 분야에 적용하여 일반화 가능성을 검토하고, 자동화된 질문 생성 및 평가 도구와 연계한 실용적 테스트 프레임워크로 확장할 예정이다.


## ACKNOWLEDGMENT

본 연구는 고용노동부 및 한국산업인력관리공단의 '고숙련 마이스터 사업(2025)'의 지원을 받음.


## 참 고 문 헌


[1] OpenAI. (2023). GPT-4 Technical Report. https://openai.com/research/gpt-4

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.

[3] See, A., Liu, P. J., & Manning, C. D. (2019). What Makes a Good Conversation? NAACL-HLT.

[4] Voorhees, E. M. (2001). The TREC Question Answering Track. Natural Language Engineering, 7(4), 361-378.

[5] Ferrucci, D. et al. (2010). Building Watson: An Overview of the DeepQA Project. AI Magazine, 31(3), 59-79.

[6] Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. ACL Workshop.

[7] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. ACL.

[8] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP.