LLM 기반의 네이버 검색 서비스 Cue:

NAVER ▌유홍연·유승학·김용범

1. 서 론

최근 몇 년 동안 GPT-4, Llama, PaLM과 같은 Large Language Models (LLMs)이 자연어 처리(NLP) 분야에서 주목할 만한 성과를 이루며 다양한 분야에서 적용을 위한 연구가 진행되고 있다[1-4]. 이러한 모델들은 기존의 대량의 레이블링된 학습 데이터가 필요한 한계점에서 벗어나 zero-shot 이나 few-shot을 통해서도 상당한 유연성과 함께 뛰어난 성능을 보여주는 장점이 있다[5]. 특히 고품질의 사용자 지시(instruction)-응답(response) 데이터 기반으로 사용자 요청에 대한 정교하고 상세한 응답을 생성할 수 있도록 alignment fine-tuning하는 경우 언어 이해, 자연어 생성, 추론 등 다양한 분야에서 탁월한 능력을 입증했다[6].

LLM의 성공은 정보 검색 분야(IR)에서도 많은 관심을 끌고 있으며, LLM의 능력을 정보 검색 분야에 어떻게 활용할 것인지 많은 연구가 진행되어지고 있다. LLM이 생성한 결과물의 정확도와 다양성이 기존의 IR 시스템을 대체하거나 새로운 검색 엔진 알고리즘을 개발하는 데 이용될 수 있음을 발견하였고, 현재 진행 중인 다수의 연구에서는 이러한 LLM의 텍스트 생성 능력을 어떻게 효과적으로 IR 시스템의 구조와 알고리즘에 통합할 수 있을지를 연구하고 있으며, 이 과정에서 새로운 접근 방법이 개발되고 있다. 일반적으로 질의 재작성, 문서 확장 등과 같이 IR에 많이 사용되는 각 Task 품질을 향상 시키는 것도 가능하지만, 랭킹, 인덱싱, 문서 신뢰도 평가 등 전반적인 정보 검색 프로세스 내의 언어적인 추론 능력이 필요한 다양한 부분에 적용될 수 있다.

그러나 LLM의 강력한 생성 및 추론 능력에도 불구하고, 정보 검색 분야에 적용하기에는 몇 가지 중요한 한계점들이 존재한다. 가장 중요한 문제 중 하나는 환각(hallucination) 현상이다. LLMs는 사실과 일치하지 않는 정보를 생성하기도 하며, 이러한 오류는 종종 사실과도 구분하기 쉽지 않은 수준으로 그럴듯하게 표현되어 출처가 없는 응답에 대해서는 사용자의 입장에서 제공된 정보를 비판적으로 평가하고 다시 별도의 정보 검색 엔진을 통해 사실을 확인할 필요성을 증가시키는 등의 모델의 신뢰성 문제를 야기할 수 있다[7,8].

환각 문제를 해결하고자, 여러 연구자들은 검색 문서 기반으로 응답을 생성하는 RAG (Retrieval Augmented Generation), 그리고 사용된 문서를 참고 문헌으로 표현하는 기술 등을 포함한 다양한 새로운 패러다임을 연구하고 있다. 이러한 접근법들은 LLM이 외부 지식 소스에서 정보를 검색하여 환각 현상을 줄이고 사용자의 신뢰를 향상시킬 수 있는 방법을 제시한다. 특히, RAG기반 모델은 필요한 정보를 외부 데이터베이스 또는 정보 검색 엔진을 통해 검색하고, 이를 바탕으로 응답을 생성함으로써, 모델 응답의 신뢰성을 향상시킬 수 있다. 더불어, 모델이 제공하는 응답에 출처를 인용하는 것은 사용자가 해당 정보의 신뢰성을 직접 검증할 수 있는 방안을 제공한다[9-11].

이러한 기술은 이미 몇몇 서비스에서 실용화되기 시작했다. 예를 들어, 최근에 뉴빙(New Bing)과 Perplexity.ai와 같은 플랫폼에 적용되어 사용자의 질문에 기반한 검색 결과를 이용하여 응답을 생성하고 참고 문서글 명시하도록 설계되어 있다[12,13]. 이러한 생성형 검색 서비스는 사용자 입장에서 기존의 키워드 매칭 기반 검색 방식에서 벗어나 복잡하고 다양한 질문을 통해 원하는 정보를 검색할 수 있게 되었다. 더불어 대화형 형태로 검색 단계를 점진적으로 진행할 수 있어, 사용자가 최종 검색 검색 목적을 더 효과적이고 정확하게 달성할 수 있게 돕는다. 생성형 검색은 사용자의 질문을 이해하고, 연관된 정보를 실시간으로 생성하여 제시함으로써, 사용자에게 더 나은 검색 경험과 향상된 결과의 질을 제공할 수 있다.

특히, 본 문서를 통해 다루는 네이버의 검색 서비스 Cue:는 한 단계 더 나아간 서비스를 론칭하였다. Cue:는 단순한 질문에 대한 검색을 넘어서 사용자의

질문 뒤에 숨겨진 실제 목적을 파악하고, 사용자가 목적을 달성할 수 있도록 도움을 준다. 또한, 네이버 생태계 내의 다양한 서비스와의 연동을 통해 사용자에게 더 넓은 범위의 정보와 서비스를 제공한다. 예를 들어, 네이버의 Cue: 서비스에서는 지식 베이스, 통합 검색 기반의 신뢰성있는 정보를 획득할 수 있을 뿐만 아니라, 쇼핑, 플레이스와 연결되어 실제 제품을 구매하고, 장소를 예약하는 등 잠재적 목적에 달성 할 수 있다. 본 문서에서는 네이버 Cue:를 소개하기 위하여 2장에서는 사람 처럼 검색하는 Human-like searching에 관련하여 소개하고, 3장에서는 네이버 생태계를 연결하는 Connected service에 대해 소개한다.

2. Human-like searching

현대의 정보 검색은 단순한 키워드 일치를 넘어 사용자의 의도와 맥락을 파악하는 지능적인 방식을 필요로 한다. 실제 사람들은 검색 엔진을 활용하여 질문에 대한 답변 해야하는 상황에서 일반적으로 여러 단계의 추론 과정을 수행한다. Cue:는 이러한 사람의 사고 및 추론 과정을 모델링한 멀티 스텝 추론(Multi-Step Reasoning) 기술을 이용하여 사람과 비슷한 사고 과정을 통해 생각하고 검색하여 사용자의 질문 요청에 최적화된 결과를 제공한다.

Cue:는 멀티 스텝 추론(Multi-Step Reasoning)을 통해 네이버의 서비스들을 어떻게 사용하여 사용자의 검색 목적을 달성할 수 있는지 계획(Planning)한다. 이러한 추론 과정을 사용자에게 보여줌으로써 사용자는 Cue:가 어떤 이유로 해당 답변을 제공하는지 논리의 흐름을 명확히 알 수 있다. 그리고 네이버의 서비스들을 툴로 사용하면서 수립된 검색 계획을 수행하고(Tool Usage), 검색된 결과를 바탕으로 답변을 생성한다(Retrieval-Augmented Generation). 답변 생성 과정에서 할루시네이션(Hallucination)을 줄이기 위해 보다 신뢰성 있는 결과를 선택하고(Evidence Selection Process), 검색 결과와 답변의 사실성이 일치되도록 답변을 생성하며(Entailment-Based Factually Consistent Generation), 사실적 일관성의 확인을 위해 모델이 자신의 답을 점검하는 자기 성찰(Self-Reflection)기법을 사용하여 사용자에게 사실적 일관성이 비약적으로 향상된 검색 결과에 기반한(Grounding) 답변을 제공한다.

다시 말해서, Cue:에게 기존 검색 엔진에서 한 번의 검색으로 찾기 어려웠던 복잡한 질문을 하면 질문 의도를 단계적으로 파악하고, 여러 단계의 검색을 수행

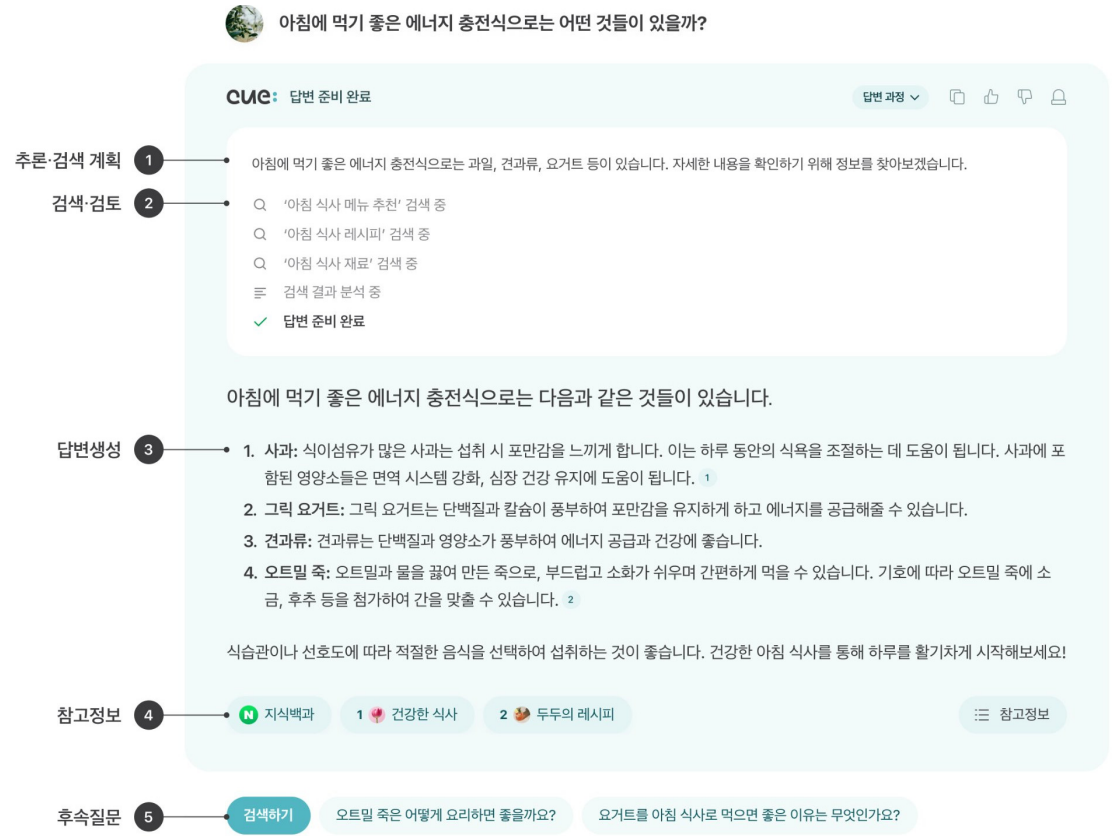

그림 1 Cue: 서비스의 단계별 구성 요소

하고, 검색된 문서중 최적의 문서들만을 이용하여 원하는 답변을 만들어 준다. 예를 들어, 아래의 예시와 같이 '아침에 먹기 좋은 에너지 충전식으로는 어떤 것들이 있을까?' 와 같은 질문은 Cue:에서 내부적으로 여러 단계를 거쳐 한 번에 답을 찾아줄 수 있다.

2.1. Multi-step reasoning

Multi-step reasoning (MSR) 과정은 Cue: 서비스에서 사용자 질문에 응답을 구성하기 위한 핵심적인 사고 프로세스를 모델링한 것을 말한다. 이 과정은 사람의 질문을 받고 수행하는 사고 과정을 유사하게 모델링하여 반영한 것이며, 사용자의 질문 의도를 명확히 이해하고, 적절한 응답을 생성하기 위한 계획을 수립하며, 필요한 정보를 검색하고 검증하는 단계를 포함한다. MSR 과정은 아래와 같이 구성된다:

- **인지 단계(Understanding)**
  - 질문 이해: 사용자의 질문 의도를 정확하게 파악하며, 관련 키워드와 문맥을 분석한다.
- **다단계 추론(Multi-step Reasoning)**
  - 계획 수립: 사용자의 질문에 응답하기 위해 필요한 정보를 어떻게 얻을 것인지, 어떤 검색 질의를 수행하거나 어떤 서비스를 활용할 것인지 결정한다.
  - 정보 검색: 계획한 검색 질의 및 파라미터를 이용하여 필요한 정보를 각 서비스를 통해 획득한다. 이 단계에서는 최신화된 정보를 얻기 위해 다양한 데이터 소스를 활용할 수 있고, 목적 달성을 위해 특정 서비스의 기능을 활용할 수 있다.
  - 정보 검증: 획득한 정보가 사용자의 질문에 대한 충분한 응답을 제공하는지 평가한다. 정보가 불충분할 경우, 계획 수립 단계로 되돌아가 추가적인 검색 또는 서비스 활용을 계획한다.
- **응답 단계(Responding)**
  - 응답 생성: 획득 및 검증된 정보를 기반으로, 사용자의 질문에 대한 명확하고 정확한 응답을 구성하여 참고 문서와 함께 제공한다.

MSR 과정은 사용자의 질문에 더욱 정확하고 효과적인 응답을 제공하기 위해, Cue: 서비스에 내재된 핵심 프로세스의 일부이다. 이 과정을 통해, Cue:는 사용자의 질문과 관련된 명확하고 신뢰할 수 있는 정보를 획득하고, 이를 기반으로 사용자에게 명확하고 유용한 응답을 제공할 수 있다. 이렇게 사람의 사고 과정을 모델링한 기술을 통해 얻을 수 있는 가장 핵심이 되는 Cue:의 기능은 추론 과정을 통해 복잡한 의도를 단계별로 해결하는 Multi-hop search를 수행하는 것이다. 이는 한 번의 검색이나 서비스의 응답이 사용

그림 2 질문 '부부의세계 작가 다른 작품 알려줘'에 대한 Cue:의 응답 과정

자의 목적을 충족시키지 못할 때, 새로운 계획을 수립하여 추가적인 검색을 통해 적절한 응답을 제공하는 것을 의미한다. 예를 들어, '삼성전자와 네이버 주가 알려줘'와 같은 복합적인 질문에 대해, Cue:는 '삼성전자 주가'와 '네이버 주가'를 각각의 독립적인 검색으로 분할하여, 사용자의 다중 의도를 이해하고 그에 따른 정보를 제공할 수 있다.

특히, '부부의세계 작가 다른 작품 알려줘'와 같은 질문에 대해서는, 초기 검색을 통해 '부부의 세계'의 작가인 '주현'의 정보를 획득한 후, 이를 기반으로 '주현 작가 작품'에 대해 추가 검색을 수행함으로써, 최종적으로 사용자에게 작가 '주현'의 다른 작품 정보인 '변혁의 사랑 (2017)', '욱씨남정기 (2016)'을 제공할 수 있다. 그림 2에서 보여지는 것처럼, 이러한 핵심 기능은 Cue: 서비스가 복합적인 사용자 질문에 대해 단계별로 문제를 분해하고, 각 단계의 결과를 재활용하여 최종 응답을 구성함으로써, 사용자의 질문에 대한 더욱 정확하고 완벽한 응답을 제공할 수 있게 만든다. 이는 실제 사용자가 검색을 수행하면서 사고하는 과정을 유사하게 모델링 했기때문에 가능해지는 결과이다.

2.2. Factually Consistent Generation

Cue:에서는 신뢰성이 부족하고 정확하지 않은 답변을 하는 환각(hallucination)을 최소화하기 위해 RAG (Retrieval-Augmented Generation) 기반의 응답을 생성한다. 일반적인 LLM은 검색 결과나 외부 지식에 의존하지 않고 기존 학습된 정보를 이용하여 연속적인 단어 예측을 통해 문장을 구성한다. 이 방식은 대량의 학습 데이터에 기반하여 고품질의 답변을 생성하지만, 이 과정에서 정보의 왜곡이나 할루시네이션이 발생할 수 있다. Cue:는 검색 결과를 기본 데이터로 활용하여 높은 신뢰성을 보장하기위해 RAG기반으로 응답을 생성하며, 세부적으로는 'Reasoning', 'Evidence Selector', 'Factually Consistent Generation'등의 기술을 통해 오류를 줄이고 정확한 정보를 제공하고 있다.

- **Reasoning**: 복잡하거나 다양한 관점의 질문에 대해 명확하고 구체적인 답변을 제공하기 위해, Reasoning 기술을 통해 질문의 본질을 분석하고 해석한다. 질문의 논리적 타당성을 검토하고, 검색 결과를 통해 질문의 의도를 재차 검증을 수행하여 질문의 의도를 명확하게 파악하고 전체적인 검색 결과의 타당성을 검증한다.
- **Evidence Selector**: 다양하고 많은 검색 결과 중에서 가장 관련성이 높은 문서나 정보를 선별하기 위해 Evidence Selector 기술을 활용한다. 문서 내의 정보 중에서 사용자 의도와 가장 일치하는 정보를 추출하며, 필요한 정보가 누락되거나 부족한 경우, 이러한 시그널을 함께 전달하여 답변의 정확성과 신뢰성을 향상시킨다.
- **Factually Consistent Generation**: 답변 생성 과정에서 정보의 출처와 신뢰성을 보장하기 위해, Reasoning 단계와 Evidence Selector 단계의 결과를 활용하여 최종 검증을 수행하며 응답을 생성한다. 응답 생성 시에는 evidence가 존재했던 문서를 참고 정보로 인용하여, 사용자들이 실제 원본 문서를 확인하며 응답의 신뢰성을 다시 평가할 수 있게 도와준다.

이와 같이, Cue:는 RAG 기반의 메커니즘과 함께 여러 단계별 검증 기술을 도입함으로써, 사용자에게 높은 신뢰성과 정확성을 보장하는 Factually consistent generation을 통해 최적의 응답을 제공하고 있다.

3. Connected service

Cue:는 사용자의 검색 흐름 및 패턴을 철저히 분석하여 통합적이며 입체적인 답변을 제공하는 서비스이다. 그 중심에는 네이버의 다양한 서비스와의 연결성에 있다. 무엇보다도 Cue:의 핵심적인 경쟁력은 네이버의 포괄적인 생태계와의 원활한 연계에 있다. 이러한 연결성은 Cue:를 단순한 검색 엔진으로서의 기능을 넘어서게 만든다.

사용자가 정보를 검색하는 것은 종종 그 정보를 기반으로 특정 행동을 취하길 원하는 것을 의미한다. 예를 들어, 사용자가 제품을 검색한다면, 그것은 구매의도를 내포하고 있을 수 있다. Cue:는 이러한 사용자의 의도를 파악하고, 네이버 쇼핑 같은 관련 서비스로 바로 연결하여 구매를 쉽게 할 수 있도록 돕는다. 또한 여행 숙소에 대한 검색의 경우, 사용자를 네이버의 예약 서비스로 안내하여 숙박 예약까지의 프로세스를 원활하게 진행할 수 있다. 이러한 통합적인 서비스 제공은 네이버가 쇼핑, 로컬 및 다른 다양한 분야에서 구축한 탄탄한 생태계 덕분이다. 그 결과, 사용자는 단순한 정보 검색을 넘어, 실제 필요한 서비스까지의 연계를 체감하며 더욱 만족스러운 경험을 얻을 수 있다.

LLM에 검색 엔진과 더불어 네이버 서비스 생태계를 연결하는 서비스가 가능해지기 위해 Cue:는 하나의 대형 언어 모델을 사용하는 것이 아닌, 크기가 다르고 기능들이 각자 다른 다수 언어 모델들을 사용한다. 이런 언어 모델들을 합쳐서 모듈화된 LLM 플랫

폼(Modularized LLM Platform)으로 설계하여 모든 사용자의 요청에 대해 전체 모델이 지속적으로 동작하는 것이 아니라, 사용자의 검색 의도를 만족시키는데 필수적인 부분만 동작 시킴으로써 효율성(Efficiency)을 향상시키고 응답 속도를 끌어올렸다. Connected service를 위해 네이버의 각 서비스들은 하나의 툴(Tool)로 모듈화되며 모듈화된 Tool들은 Tool orchestrator의 관리를 받아 네이버 생태계를 확장성 있게 연결하는데 핵심적인 역할을 하고 있다.

3.1. Tool Orchestrator

Tool orchestrator는 네이버의 각 서비스 및 기능을 Modularized Tool의 단위로 관리하고 총괄하는 핵심 구성 요소이다. Tool orchestrator의 주된 역할은 사용자의 요청을 분석하고 이를 적절한 Tool에 할당하여 처리하는 것이다. 이 과정에서, Tool orchestrator는 각 Tool의 역할과 책임을 명확하게 정의하고 있으며, 동시에 각 도구의 작업 진행 상황을 모니터링하고 관리한다. 이는 높은 병렬성과 효율성을 담보로 하여 시스템의 전반적인 성능을 향상시키는 데 기여한다. 더불어, 복잡한 의도나 중의성이 있는 경우에는 다양한 Tool을 동시에 이용하여 병렬 호출하고 응답을 검토하여 최종 응답에 활용할 Tool의 결과를 선택하거나 조합 및 재구성하여 활용하며, 사용자의 잠재적 목적을 달성하기위해 요청하지 않았더라도 사용자가 필요하다고 판단되는 경우 관련 기능을 제공하는 특징이 있다.

예를 들어, 사용자가 특정 제품에 대한 검색을 요청할 경우, 이 요청을 네이버 쇼핑 서비스와 연결하여 사용자에게 적절한 제품 정보를 제공할 뿐아니라 구매 가능한 링크를 동시에 제공한다. 또한, 특정 음식의 레시피를 찾는 요청에 해당 레시피에 포함되는 재료들을 선택하여 구매할 수 있는 기능을 제공한다. 이외에도, 여행 숙소나 레스토랑 예약과 같은 다양한 검색 요청에 대해, 관련된 네이버의 예약 서비스와 연계하여 사용자에게 원활한 예약 경험을 제공한다. 이와 같이 단일 검색이나 특정 서비스에 종속되지 않고 다양한 서비스를 적절히 연결하는 방식으로 사용자의 검색 요청과 네이버의 다양한 서비스 간의 효율적인 연계를 보장하며, 이는 사용자에게 높은 만족도와 유용성을 제공한다.

3.2. Modularized Tools

Cue:는 네이버 생태계에 존재하는 컬랙션, 서비스, 기능들을 각각 하나의 툴(Tool)로 구성하며, 툴 단위로 신뢰성있는 정보를 검색해오거나, 사용자 목적 달성에 도움이 되는 적절한 기능 제공을 수행한다. 이렇게 각각의 툴이 특정 역할을 수행함으로써, Cue:는 사용자에게 맞춤형 결과 및 서비스 연결을 제공할 수 있다. 툴은 검색 엔진, API, LLM 모델 등 다양한 형태로 구성될 수 있으며, 초기 서비스 론칭 단계에서의 대표적인 툴은 다음과 같다:

- **통합 검색**: 네이버 검색 서비스는 이용자들이 원하는 정확한 정보를 이용하기 편리하고 신속하게 제공하는 것을 가장 중요한 목표로 삼고 있다. 이러한 목표를 달성하기 위해 네이버 내·외부를 떠나 이용자에게 필요한 다양한 정보를 확보하고 있으며, 이용자의 이용 환경과 패턴을 분석하여 이용자의 환경에 최적화된 검색 결과를 제공을 하기 위해 노력하고 있다. Cue:는 기본적으로 네이버의 통합 검색을 신뢰하고 검색 결과를 사용자 질문에 대한 응답 생성에 적극적으로 활용한다.
- **지식베이스**: 네이버의 지식베이스는 네이버 컨탠츠 검색의 기반이되는 데이터베이스이며 인물, 영화, 프로그램, 증권, 날씨, 환율 등 수십가지가 넘는 다양한 카테고리의 정확도 높은 정보가 저장되어있으며 실시간 최신성을 유지하고 있다. Cue:에서는 네이버 지식베이스 기반의 컨탠츠 검색을 정답형 응답을 위한 하나의 연결된 서비스로 활용하여 답변의 신뢰도를 향상시킨다.
- **쇼핑**: 네이버 쇼핑은 일반적인 제품 검색 뿐만아니라, 제품의 정보와 특정 카테고리별 랭킹, 특정 연령대의 선호 제품 등 온라인 쇼핑몰에서 사용자들이 일반적으로 사용할 수 있는 다양한 기능을 제공한다. 사용자는 Cue:의 통합 검색 Tool을 통해 패션 트랜드를 검색하다가도 자연스럽게 쇼핑 Tool로 넘어와 검색하던 패션에 관련된 제품을 구매할 수 있다. 이 모든 것은 내부적으로 동작하고, 사용자는 Cue:에게 원하는 내용을 질문만 하면된다. 추가적으로 쇼핑 툴에서는 사용자의 잠재적 목적을 파악하고 구매 옵션까지 이어지는 기능을 제공한다. 예를 들어, 질문 '내일 김치찌개 만들어야하는데 맛있게 만드는 레시피 알려줘'에 대해 잠재적의도인 김치째개 레시피의 재료 구매에 대한 목적을 파악하고, 각 재료를 손쉽게 구매할 수 있는 장보기 기능까지 연결한다.
- **플레이스**: 네이버 플레이스는 네이버 지도 연계하여 가게, 업체, 호텔 등 다양한 장소에 관련된 상세정보를 검색하고 확인할 수 있는 서비스이

다. 위치, 영업시간 및 휴무일, 주차장 유무, 업체 사진, 간단한 업체 소개글, 업체 공식 홈페이지나 블로그 등의 웹주소, 메뉴, 서비스 등 가게와 업체를 이용하는데 필요한 다양한 정보를 확인할 수 있다. Cue:에서는 특정 위치의 맛집 요청이나, 데이트 할때 갈만한 곳 등 다양한 의도의 요청을 처리 할 수 있으며 실제 맛집 예약 등 사용자의 목적까지 연결해주는 기능을 갖추고 있다.

- **이미지**: Cue:에서는 텍스트 뿐만 아니라 이미지와 같은 형태의 결과도 응답으로 제공 가능하다. 기본적으로 네이버 이미지 검색 서비스를 툴로써 활용하며, '1990년대 서울 거리 사진 보여줘' 등과 같은 요청에 대응 가능하다.
- **동영상**: 동영상 검색은 네이버의 서비스뿐 아니라 외부 서비스의 동영상 또한 검색 결과로 제공한다. 썸네일이나 제목을 클릭하면 영상을 바로 재생하거나 원본 페이지로 이동할 수 있다. 특히, '드라마 연인 관련 최근 영상 보여줘'와 같이 직접적인 동영상을 보여달라는 의도에 대한 응답도 가능하지만, '골프 비거리 늘리는 방법 설명해줘'와 같이 동영상을 특정한 의도가 아니더라도, 일반 검색 결과 기반의 응답과 함께 조합되어 사용자에게 제공 가능하다.

이러한 서비스들은 각각의 툴로 구성되어 Cue:의 기능성과 효율성을 높이며, 사용자에게 더욱 풍부하고 정확한 정보를 제공하는데 기여한다. Cue:는 초기 론칭 단계 이후에 서비스 로그 분석을 통해 사용자의 의도와 선호를 더욱 깊이 파악하여, 이를 기반으로 서비스 및 기능 기반의 툴 확장과 개선을 계속 진행할 예정이다.

4. 결론 및 향후 계획

본 문서에서는 LLM 기반의 새로운 대화형 검색 서비스 Cue:의 개발과 구현에 관한 사례를 소개하였다. 기존의 검색 엔진은 키워드 기반 매칭을 통해 단편화된 정보 검색에 그쳤으나, 실제로 정보 검색은 종종 어떠한 잠재 목적을 달성하기 위한 과정으로 이루어진다. 예를 들어, 사용자는 제품을 구매하기 위해 제품 관련 정보 및 트랜드를 검색하거나, 식당 예약을 위해 맛집을 찾는 경우가 있을 수 있다. Cue:는 이러한 잠재적 목적 달성을 쉽게 도와주는 데 많은 기여를 하고 있다. Cue:는 Human-like searching과 Connected service의 기술적 통합을 통해 사용자에게 키워드 기반 검색을 넘어서 복잡한 의도를 가지고 있더라도 효율적이며 자연스러운 정보 검색 경험을 제공하며, 네이버 서비스 생태계와 연결된 경험을 제공함으로써 사용자의 잠재적 목적 달성에 기여하였다. Cue:를 사용하는 사용자들은 본 문서에서 소개한 바와 같이 제품 구매, 정보 검색, 장소 예약 등 다양한 목적을 달성하기 위해 자연스럽게 각 서비스의 경계를 넘나드는 경험을 할 수 있는 특징이 있다.

또한 Cue:는 핵심 기술 중 하나인 Multi-step reasoning을 이용하여 복잡한 사용자 의도에 대한 단계별 해결을 가능케 함으로써, 정보 검색과 처리의 정확성과 효율성을 극대화하였다. 더불어, Factually Consistent Generation 기술은 할루시네이션 현상을 최소화하고, 정보의 출처를 명시함으로써 신뢰성을 확보하였고, Modularized LLM 기술은 네이버의 다양한 서비스와 기능을 End-to-End Platform-Level로 통합함으로써, 네이버 생태계와 연결성과 확장성있는 서비스 경험을 제공할 수 있게 되었다.

네이버의 Cue:는 사용자의 검색 경험을 혁신적으로 바꾸는 데 중요한 첫 걸음이며, 앞으로 더 다양한 서비스와 기능을 확장하고 개선할 예정이다. 이를 통해 사용자의 잠재적인 목적을 더욱 효과적으로 달성할 수 있도록 도움을 주는 서비스를 제공할 계획이다. 예를 들어, 네이버 부동산, 네이버 항공권과 같은 다양한 서비스를 Cue:에 통합하여 사용자에게 더욱 풍부하고 유용한 정보를 제공하고, 잠재적 목적 달성을 도울 예정이다. 또한, Cue:는 텍스트 뿐만 아니라 이미지, 음성, 동영상과 같은 다양한 형태의 데이터와 크로스 도메인간의 정보를 복합적으로 처리할 수 있는 기능을 개발 중에 있어, 사용자가 원하는 정보를 더욱 풍부하고 다양한 방식으로 제공받을 수 있도록 서비스를 확장할 예정이다. 이러한 확장성은 Cue:를 통해 사용자들이 더욱 다양하고 실용적인 정보를 얻을 수 있게 하며, 네이버의 Cue:는 이러한 노력을 통해 사용자들에게 더욱 향상된 검색 경험을 제공할 것이다.

추가적으로, Cue:는 서비스 품질의 지속적 향상을 목표로 사용자의 질문, 행동 및 피드백을 기반으로 리워드 모델링을 도입할 계획이다. 이러한 접근은 사용자의 질문을 더욱 정확하게 이해하고, 그에 따라 사용자의 의도와 목적에 부합하는 답변 및 서비스를 제공하는 데 중점을 둔다. 이는 사용자에게 더 나은 서비스 경험을 제공하기 위한 노력의 일부로, 사용자의 실제 서비스 내 행동을 분석하여 사용자 의도와 목적을 더 정확하게 파악하는 데 도움을 줄 것이다. 특히, 사용자 행동 기반의 리워드 모델링을 통해 Cue:는 더

정확하고 유용한 답변을 제공할 뿐아니라, 사용자의 새로운 잠재적 목적을 분석하여 추가 서비스 및 기능들을 확장성 있게 제공할 수 있을 것이라 기대하고 있다.

이러한 노력과 기술적 혁신을 통해 네이버의 Cue:는 사용자의 검색 경험을 혁신적으로 변화시키며, 더욱 진보된 대화형 검색 서비스의 가능성을 제시하였다. Cue:의 지속적인 발전을 통해 사용자는 더욱 향상된 정보 접근성과 편의성을 경험할 수 있을 것이며, 이는 결국 검색 기술의 패러다임을 전환하는 기술의 시작이 될 것이다.

참고 문헌

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

[2] OpenAI. 2023. Gpt-4 technical report. ArXiv, abs/2303.08774.

[3] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, 2023. LLaMA: Open and Efficient Foundation Language Models, ArXiv, abs/2302.13971.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, Noah Fiedel, 2022. PaLM: Scaling Language Modeling with Pathways, ArXiv, abs/2204.02311.

[5] Weiwei Sun, Lingyoun Yan, Xinyu Ma, Pengjie ren, Dawei Yin, Zhaochun Ren, 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent, ArXiv, abs/2304.09542.

[6] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

[7] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. CoRR, abs/2301.12652.

[8] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation, ArXiv, abs/2305.06983.

[9] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback, ArXiv, abs/2112.09332.

[10] Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, Ji-Rong Wen, RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit, 2023. ArXiv, abs/2306.05212.

[11] Tianyu Gao, Howard Yen, Jiatong Yu, Danqi Chen, Enabling Large Language Models to Generate Text with Citations, 2023. ArXiv, abs/2305.14627.

[12] Bing Chat, https://www.bing.com/new

[13] Perplxity.ai, https://www.perplexity.ai

약 력

유홍연

2017 동아대학교 컴퓨터공학과 졸업(학사)
2019 동아대학교 컴퓨터공학과 졸업(석사)
2020~현재 Naver Search
관심분야 : 자연어처리, 정보검색, 머신러닝 등
Email : hongyeon.yu@navercorp.com

유승학

2015 고려대학교 졸업(박사)
2015~2018 Samsung Research Staff Engineer
2019~2020 MIT Postdoctoral Associate
2021~2022 Amazon Applied Scientist
2022~현재 Naver Search US Staff Scientist
관심분야 : 자연어처리, 정보검색, 머신러닝 등
Email : seunghak.yu@navercorp.com

김용범

전) Senior Applied Scientist, Microsoft AI and Research
전) Head of Applied Science, Amazon Alexa
2022~현재 Naver Search US Chief Scientist
관심분야 : 자연어처리, 정보검색, 머신러닝 등
Email : youngbum.kim@navercorp.com