

KRX-Bench

KRX-Bench 는 GPT-4-Turbo 를 활용하여 실제 기업의 연간 보고서, 금융 소송 문서 등 다양한 금융 자료로부터 자동으로 질문을 생성합니다. 해당 파이프라인을 통해 생성된 벤치마크 데이터셋은 다양한 금융 주제를 다루며, 단순한 사실 확인부터 복잡한 금융 추론까지 다양한 난이도의 문제를 포함하고 있습니다. 또한, 다국어 지원과 개방형 질문 등 다양한 형태의 문제를 제공하여 언어 모델의 종합적인 능력을 평가할 수 있습니다.

Paper.

[KRX-Bench: Automating Financial Benchmark Creation via Large Language Models](#) @ LERC-COLING 2024 FinNLP

Introduction.

금융 분야에서 대규모 언어 모델(LLM)의 도입이 늘어나고 있지만, 이러한 모델들의 정확성과 신뢰성에 대한 우려가 여전히 존재합니다. 특히 금융 분야에서는 LLM 의 부정확한 정보 생성 (일명 'Hallucination')이 잘못된 투자 결정으로 이어져 심각한 재정적 손실을 초래할 수 있습니다.

이러한 배경에서, 해당 논문에서는 **KRX-Bench** 를 소개합니다. KRX-Bench 는 GPT-4 를 활용하여 금융 벤치마크를 자동으로 생성하는 파이프라인으로, 이 파이프라인의 효과를 검증하기 위해 KRX-Bench-POC 벤치마크 데이터셋을 제작하였습니다:

- 미국, 일본, 한국의 주식 시장에 상장된 실제 기업들에 대한 1,003 개의 질문으로 구성
- LLM 의 실제 금융 환경에 대한 이해도를 평가하는 **최초**의 벤치마크

연구 결과, KRX-Bench 는 인간의 개입 없이도 정확하고 신뢰할 수 있는 벤치마크를 자동으로 생성할 수 있음이 확인되었습니다. 특히 주목할 만한 점은 다음과 같습니다:

1. 생성된 벤치마크에서 'false positive(FP)' 비율이 **1%에 불과할 정도로 높은 품질**을 보여줌
2. 당시 가장 뛰어난 오픈소스 LLM(예: Qwen1.5-72B, Llama-2-70B)의 성능이 80% 미만으로, **개선의 여지가 있음을 시사**

KRX-Bench.

KRX-Bench: 벤치마크 생성 파이프라인

KRX-Bench 는 금융 분야의 LLM 평가를 위한 자동 벤치마크 생성 파이프라인입니다. KRX-Bench 는 GPT-4-Turbo 를 활용하여 기존 문서로부터 다음과 같이 크게 3 가지의 프로세스로 벤치마크 데이터를 생성할 수 있습니다:

1. 질문 및 답변 생성:

- GPT-4-Turbo 에 문서를 제공하고 Q&A 쌍을 생성하도록 지시합니다.
- 이 단계에서는 연간 보고서, 금융 소송 문서 등 **다양한 금융 자료를 입력으로 사용할 수** 있습니다.

2. 오답 선택지 생성:

- 생성된 Q&A 쌍을 다지선다형 문제로 변환하기 위해 **오답 선택지**를 만듭니다.
- BM25 알고리즘을 사용해 유사한 질문을 찾고, GPT-4-Turbo 가 이를 바탕으로 그럴듯한 오답을 생성합니다.
- 질문과 무관한 회사를 언급하거나 길이가 크게 다른 선택지는 제거하는 등의 휴리스틱 규칙을 적용하여 보기의 품질을 관리합니다.

3.품질 관리:

- GPT-4-Turbo 를 활용하여 **답변 불가능한 질문을 식별하고 제거**합니다.

KRX-Bench-POC: 벤치마크 데이터셋

파이프라인의 효과를 검증하기 위해 KRX-Bench-POC 라는 벤치마크 데이터셋을 생성하여 테스트를 진행하였습니다. 해당 벤치마크 데이터셋은 다음과 같은 특징을 가집니다:

- 미국, 일본, 한국의 연간 보고서를 기반으로 생성
- 총 1,003 개의 질문 (미국 373 개, 한국 319 개, 일본 311 개)
- 15 개의 다양한 카테고리(주요 사업, 사업 목표, 회사 미션 등)로 구성



KRX-Bench-POC 에 속해 있는 벤치마크 데이터셋 예시

벤치마크 데이터셋의 품질 평가는 다음과 같이 진행하였습니다:

1.머신 러닝 특징 테스트:

- Similarity-based: BM-25 & BERT 를 사용하여 질문과 보기 간에 유사도가 가장 높은 보기를 답으로 고려
- Zero-shot Classifier: NLI 태스크로 학습된 분류기 용
- KRX-Bench-POC 는 기존의 HellaSwag 벤치마크와 비교하여 유사하거나 더 낮은 점수를 보여, 과도한 패턴이나 단서가 없음을 확인했습니다.

2.사람 평가:

GPT-4-Turbo 의 질문 분류와 인간 평가자의 판단을 비교했습니다.

결과적으로 false positive(FP) 비율이 1%에 불과해, 생성된 질문의 높은 품질을 입증했습니다.

Result.

KRX-Bench-POC 를 활용한 다양한 LLM 의 평가 결과는 다음과 같은 인사이트를 제공합니다:

1.모델 크기와 성능의 관계

- 당시 최신 오픈소스 모델인 Qwen1.5-72B 와 Llama-2-70B 도 80% 미만의 점수를 기록하며 여전히 개선의 여지가 있음을 시사
- KRX-Bench 파이프라인이 인간의 개입 없이도 최신 모델들을 도전적으로 평가할 수 있는 벤치마크를 성공적으로 생성했음을 입증

Models	N=3			
	US	KO	JR	Total
Pre-Trained Models				
Qwen1.5-0.5B	20.38	17.87	18.06	18.77
Qwen1.5-1.8B	39.68	24.14	20.97	28.26
Qwen1.5-4B	58.45	31.35	30.65	40.15
Qwen1.5-7B	81.77	47.34	48.06	59.06
Qwen1.5-14B	87.13	57.68	60.65	68.49
Qwen1.5-72B	87.40	72.10	72.58	77.36
Llama-2-7B	42.09	20.38	23.23	28.56
Llama-2-13B	85.52	52.98	51.94	63.48
Llama-2-70B	93.30	71.16	73.23	79.23
Continual Pretrained Models				
Japanese-StableLM-Base-Beta-7B	32.98	21.00	23.87	25.95
Llama-2-KOEN-7B	17.16	19.44	18.06	18.22
Proprietary Models				
GPT-3.5-Turbo	87.13	63.32	66.13	72.19
GPT-4-Turbo	95.44	84.33	84.84	88.20

전체 데이터셋에 걸쳐서 3-shot setting 으로 계산된 정확도의 평균

2.지역별 편향

- 모든 질문이 영어로 제시되었음에도 불구하고, 모델들은 미국 기업에 관한 질문에서 일본이나 한국 기업 관련 질문보다 더 높은 성능을 보임
- 이러한 결과는 영어로 된 일본 및 한국 기업 관련 자료의 부족, 추가 학습 과정에서의 catastrophic forgetting 등 여러 요인에 기인할 수 있습니다. 또한 더 어려운 문서(예: 내부 문서)를 입력으로 사용하면 벤치마크의 난이도를 쉽게 높일 수 있음을 시사

3.GPT-4-Turbo 자체 평가

- KRX-Bench 파이프라인이 GPT-4-Turbo 를 사용하여 생성되었다는 점에서, GPT-4-Turbo 자체의 벤치마크 수행 능력을 평가하는 것은 중요한 의미를 가짐
- GPT-4-Turbo 는 테스트된 모델 중 가장 높은 점수를 기록했지만, 모든 하위 집합에서 완벽한 성과를 보이지는 않음
- 특히 한국과 일본 기업 관련 질문에서 상대적으로 낮은 점수를 기록
- 이는 보충 자료를 활용한 데이터셋 생성 방식이 모델의 지식 범위를 넘어서는 질문을 만들어낼 수 있음을 시사

Practitioner's Guide for Implementing KRX-Bench.

KRX-Bench 파이프라인은 단순히 다지선다형 문제를 생성하는 데 그치지 않고, 다양한 종류의 금융 벤치마크를 만드는 데 활용될 수 있습니다:

1. Open-ended Generation

- KRX-Bench 는 다지선다 문제에서 벗어나 언어 모델의 생성 능력을 평가하기 위한 개방형 질문을 만들 수 있습니다.
- 해당 벤치마크 데이터셋에 대한 평가는 "LLM-as-a-Judge" 접근법을 사용하여 답변의 품질을 1-5 점 척도로 평가할 수 있습니다.
- 이 방법은 모델의 생성 능력뿐만 아니라 지식의 깊이도 평가할 수 있어, 금융 분야 LLM의 종합적인 성능 측정에 유용합니다.

Question

What significant actions did HOYA Corporation take in 2017 concerning the medical devices sector?

Gold Answer

In 2017, HOYA Corporation established a joint venture company with China's Aohua for a medical flexible endoscope business and acquired ReadSpeaker, a leading company in cloud-based text-to-speech services. They also set up a new production base in Thailand for intraocular lenses used in cataract surgery, and acquired Performance Optics, LLC along with its subsidiaries VISION EASE and Daemyung Optical.

GPT-3.5-Turbo

In 2017, HOYA Corporation made a significant move in the medical devices sector by acquiring the medical unit of the US-based company, Performance Optics. This acquisition allowed HOYA to expand its presence in the ophthalmic lens market and strengthen its position in the medical devices industry. The deal included Performance Optics' manufacturing facilities ...

Score:3

GPT-4-Turbo

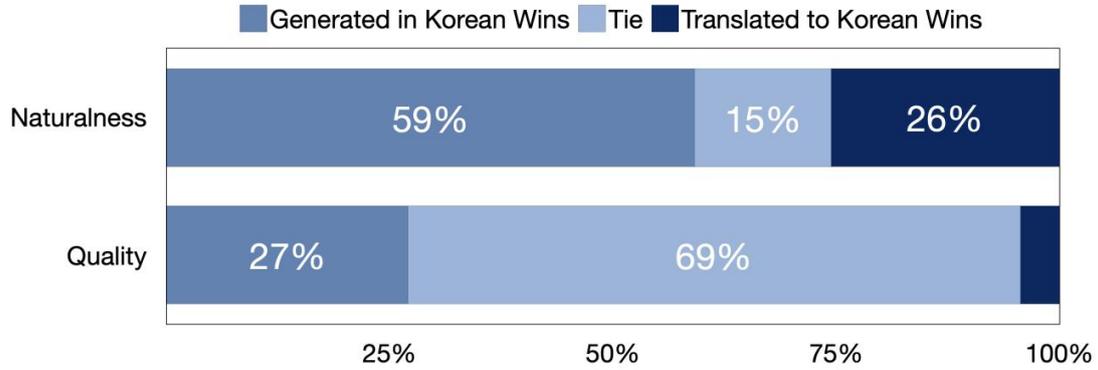
In 2017, HOYA Corporation, a global technology company based in Tokyo, Japan, known for its innovations in optics and medical devices, took a significant step in the medical devices sector by acquiring Performance Optics, LLC, including its subsidiaries VISION EASE, a global ophthalmic lens marketer and manufacturer, and Daemyung Optical Co. Ltd., a leading optical lens manufacturer in South Korea...

Score:4

open-ended question 예시

2. Multilinguality

- KRX-Bench 를 통해 한국어로 250 개의 질문을 생성하고, 이를 영어에서 번역한 질문과 비교 평가했습니다.
- 두 명의 평가자가 질문의 자연스러움과 품질을 비교한 결과, 직접 생성된 결과가 더 자연스럽고 품질 측면에서 좀 더 괜찮다는 것을 확인하였습니다.



generation 과 translation 간의 win rate 비교 결과

3.Beyond Knowledge Benchmarks

- KRX-Bench 를 통해 한국의 금융 소송 관련 문서를 입력으로 사용하여 100 개의 법적 추론 문제를 생성했습니다.
- 평가 결과, GPT-4-Turbo 가 거의 완벽한 점수를 받으며 입력 문서의 특성이나 오답 생성의 한계 때문일 수 있기 때문에 추가적인 연구가 필요합니다.

Models	N=3			
	US	KO	JR	Total
<i>Pre-Trained Models</i>				
Qwen1.5-0.5B	20.38	17.87	18.06	18.77
Qwen1.5-1.8B	39.68	24.14	20.97	28.26
Qwen1.5-4B	58.45	31.35	30.65	40.15
Qwen1.5-7B	81.77	47.34	48.06	59.06
Qwen1.5-14B	87.13	57.68	60.65	68.49
Qwen1.5-72B	87.40	72.10	72.58	77.36
Llama-2-7B	42.09	20.38	23.23	28.56
Llama-2-13B	85.52	52.98	51.94	63.48
Llama-2-70B	93.30	71.16	73.23	79.23
<i>Continual Pretrained Models</i>				
Japanese-StableLM-Base-Beta-7B	32.98	21.00	23.87	25.95
Llama-2-KOEN-7B	17.16	19.44	18.06	18.22
<i>Proprietary Models</i>				
GPT-3.5-Turbo	87.13	63.32	66.13	72.19
GPT-4-Turbo	95.44	84.33	84.84	88.20

전체 데이터셋에 걸쳐서 3-shot setting 으로 계산된 정확도의 평균