FubaoLM : 연쇄적 사고 증류와 앙상블 학습에 의한 대규모 언어 모델 자동 평가

김희주[1], 전동현[2], 권오준[2], 권순환[2], 김한수[2], 이인권[2], 김도현[3], 강인호[2]
[1]고려대학교 컴퓨터학과, [2]네이버, [3]서울대학교 기계항공공학부
haena0320@korea.ac.kr, doh0106@snu.ac.kr,
{donghyeon.jeon, ohjoon.kwon, soonhwan2.kwon, kim.hansu, inkwon.l,once.ihkang}@navercorp.com

FubaoLM : Automatic Evaluation based on Chain-of-Thought Distillation with Ensemble Learning

Huiju Kim[1], Donghyeon Jeon[2], Ohjoon Kwon[2], Soonhwan Kwon[2], Hansu Kim[2], Inkwon Lee[2], Dohyeon Kim[3], Inho Kang[2]
[1]Department of Computer Science and Engineering, Korea University,
[2]Naver Corporation, [3]Science in Mechanical and Aerospace Engineering, Seoul University

요 약

대규모 언어 모델 (Large Language Model, LLM)을 인간의 선호도 관점에서 평가하는 것은 기존의 벤치마크 평가와는 다른 도전적인 과제이다. 이를 위해, 기존 연구들은 강력한 LLM을 평가자로 사용하여 접근하였지만, 높은 비용 문제가 부각되었다. 또한, 평가자로서 LLM이 사용하는 주관적인 점수 기준은 모호하여 평가 결과의 신뢰성을 저해하며, 단일 모델에 의한 평가 결과는 편향될 가능성이 있다. 본 논문에서는 엄격한 기준을 활용하여 편향되지 않은 평가를 수행할 수 있는 평가 프레임워크 및 평가자 모델 'FubaoLM'을 제안한다. 우리의 평가 프레임워크는 심층적인 평가 기준을 통해 다수의 강력한 한국어 LLM을 활용하여 연쇄적 사고(Chain-of-Thought) 기반 평가를 수행한다. 이러한 평가 결과를 다수결로 통합하여 편향되지 않은 평가 결과를 도출하며, 지시 조정 (instruction tuning)을 통해 FubaoLM은 다수의 LLM으로 부터 평가 지식을 증류받는다. 더 나아가 본 논문에서는 전문가 기반 평가 데이터셋을 구축하여 FubaoLM 효과성을 입증한다. 우리의 실험에서 앙상블된 FubaoLM은 GPT-3.5 대비 16% 에서 23% 향상된 절대 평가 성능을 가지며, 이항 평가에서 인간과 유사한 선호도 평가 결과를 도출한다. 이를 통해 FubaoLM은 비교적 적은 비용으로도 높은 신뢰성을 유지하며, 편향되지 않은 평가를 수행할 수 있음을 보인다.

주제어: 연쇄적 사고 증류, 자동 평가, 앙상블 학습, 대규모 언어 모델

1. 서론

최근에는 대규모 언어 모델 (LLM)개발이 급속도로 진행되고 있어, 이러한 모델의 성능 평가가 점점 중요해지고 있다. 하지만 기존 자연어 처리 벤치마크에서 사용되던 평가 방식은 실제 서비스에 적용될 때 중요한, 예를 들어 인간의 선호도 같은, 주관적 기준을 충분히 반영하지 못한다. 따라서, 복잡한 지시사항을 이해하고 적절한 응답을 생성하는 능력 등 여러 가지 측면에서의 모델 평가가 갈수록 중요해지고 있다.

생성된 텍스트에 대해 인간에 의한 평가가 높은 신뢰성을 가지고 있지만, 그 과정은 비용과 시간이 많이 들고 노동 집약적이다. 이 문제를 해결하기 위해, LLM 자체를 평가 도구로 사용하는 연구들이 제안되고 있다 [1, 2, 3]. 특히 [1]에서는 대규모 LLM을 사용해 생성된 데이터로 작은 규모의 모델 (예: LLaMA-7B)을 학습시키는 평가 프레임워크를 소개하였다. 이 프레임워크는 두 가지 다른 언어 모델이 생성한 응답 중 어느 것이 더 우수한지 비교하기 위해 설계되었고, GPT-3.5 와 GPT-4 기반의 평가 성능과 유사한 수준을 달성하였다. 하지만 이러한 평가 방법은 주관적 기준이 모호하며, 그 신뢰성을 어떻게 측정할지, 또는 다양한 태스크에 어떻게 적용할지에 관한 연구는 아직 부족하다.

본 논문에서는 1) 명확한 평가 지표가 부족한 한국어 검색 관련 태스크들에 집중하여, 더 엄격한 자동 평가 시스템을 구성하는 것을 목표로 한다. 이를 위해 2) 그림 1과 같이 연쇄적 사고 증류와 앙상블 학습을 기반으로 하는 LLM 기반 평가 프레임워크와 평가자 모델 FubaoLM을 제안한다. 구체적으로 FubaoLM은 연쇄적 사고 기법을 기반으로 엄격한 절대 평가 기준에 의해 각 모델에 점수를 할당하고, 모델 간 최종 평가 우위를 결정한다. 강건성을 강화하고 평가 바이어스를 줄여 FubaoLM을 학습시키기 위해, 강한 대규모 언어 모델을 앙상블하여 훈련 데이터셋을 생성한다. 더 나아가, 전문가 기반 평가 데이터셋을 활용하여 FubaoLM의 효과성을 입증한다. 우리의 실험에서, FubaoLM은 GPT-3.5에 비해 16% 에서 23% 향상된 평가 능력을 갖춤을 보인다.

2. 관련 연구

자동 평가 Open-ended 태스크의 자동 평가를 위해, 기존에는 사람과 기계가 생성한 유사도를 측정하여 평가하였다. 이후 LLM의 언어 이해 및 생성 능력이 크게 향상되면서 GPTScore [4] 와 같이 LLM을 자동 평가에 적용하려는 시도가 있었다. [5]에서는 자연어 생성 태스크에서 생성된 텍스트의 품질을 LLM을 사용하여 평가하는 프레임워크를 제안하면서, 텍스트 요약 및 대화 태스크에서 기존의 BLEU 및 ROUGE와 같은 참조 기반 평가 지표들보다 사람의 평가와 상관관계가 더

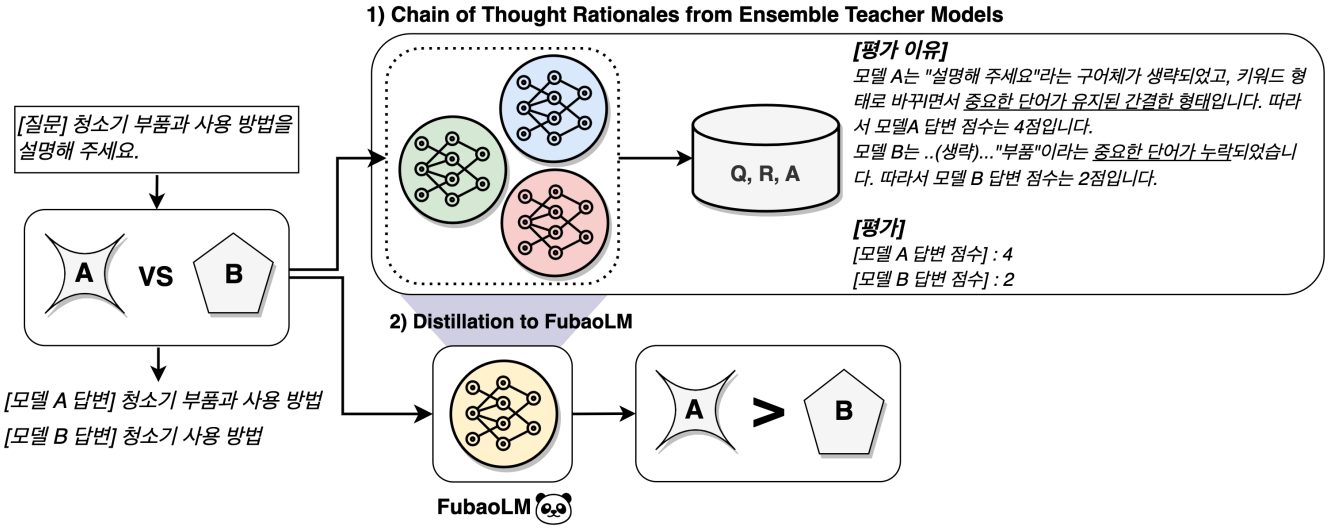

그림 1. Question-to-Query 태스크에 적용된 자동 평가 프레임워크. 1) LLM 앙상블 기반 연쇄적 사고 데이터셋 생성: 다수의 LLM이 연쇄적 사고 기반 평가를 진행하고 평가 결과를 앙상블 하여 평가 모델 구축을 위한 데이터셋을 구성한다. 2) FubaoLM 훈련: FubaoLM이 앙상블 모델로 생성한 데이터셋에서 훈련되면서 앙상블 모델의 평가 지식이 FubaoLM으로 증류된다.

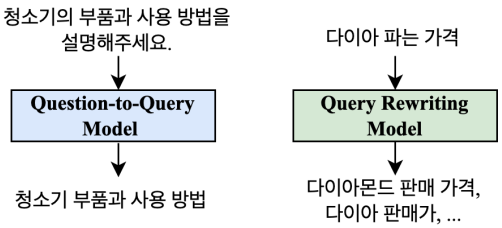

그림 2. Question-to-Query, Query Rewriting 태스크 입력값 및 출력값 예시.

높음을 보여주었다.

연쇄적 사고 증류 연쇄적 사고 (Chain of Thought)는 LLM의 추론 능력을 강화하기 위해 사용되는 프롬프트 방법이다 [6]. [7]는 LLM의 출력값을 평가하기 위해 연쇄적 사고 프롬프트를 사용하였다. 더 나아가 [8]는 선생 모델의 연쇄적 사고를 학생 모델에게 증류시켜 학생 모델의 추론 능력을 강화했다.

앙상블 방법 앙상블 방법은 다양한 모델의 예측을 조합하여 더욱 정확하고 안정적인 평가 결과를 도출하는 데 활용된다 [9]. 앙상블은 여러 모델의 다양한 강점을 결합하며, 특정 모델의 한계를 보완하는 데 도움을 준다. [3]는 여러 LLM들을 앙상블하여 다양한 태스크에서 바이어스와 예측 오류를 줄이는 시도를 했다.

3. 방법론

3.1 평가 대상 태스크

본 논문에서 평가 대상 태스크는 검색 서비스의 효율성을 높이기 위한 두 가지 주요 태스크인 질의 정규화 (Question-to-Query)와 질의 재작성 (Query Rewriting)이다. Question-to-Query는 사용자의 구어체 질문을 문서 검색에 적합한 키워드 기반의 질의로 변환하는 작업이고, Query Rewriting은 문서 검색의 다양성을 높이기 위해 기존 질의의 의미는 유지하면서 다른 표현으로 변경하는 작업이다. 각 태스크의 구체적인 예시는 그림 2에서 확인할 수 있다. 이러한 태스크들은 정량적 평가 지표를 설정하기 어려울 뿐만 아니라 주관적인 평가가 불가피하므로, 본 논문에서는 LLM을 기반으로 한 평가 방법론을 제안한다.

3.2 평가 기준

기존의 상대 평가 방법들 [10, 11]은 각 예제에 대해 특정 범위의 점수를 할당하거나, 상대적인 우위를 비교하였다. 그러나 이러한 방법들은 할당하는 점수에 대한 품질 기준이 명확하지 못하다는 단점이 있다. 따라서 본 논문에서는 평가 대상 태스크들의 특수성을 반영하여 4단계의 계단식 점수를 구성하였다. 특히, 검색 키워드로서 용이성을 반영하기 위해 간결성, 명확성, 의미 보존 등의 평가 요소를 포함한다. Question-to-Query 및 Query Rewriting 태스크들에 대한 평가 기준은 각각 그림 3, 그림 4 의 프롬프트에서 확인할 수 있다.

3.3 훈련 데이터셋 구성

LLM 기반 입력값 생성 본 논문에서는 검색 질의에 대한 다양한 답변 결과를 평가 이유와 함께 평가할 수 있는 훈련 데이터셋을 구축한다. 각 데이터 예제는 입력 튜플 (명령어, 질문, 답변1, 답변2)과 출력 튜플 (평가 이유, 답변1 평가 점수, 답변2

표 1. FubaoLM 데이터셋 통계. 훈련, 검증, 평가는 각각 훈련 데이터셋, 검증 데이터셋, 평가 데이터셋을 구성하는 데이터 수를 의미한다.

평가 태스크	훈련	검증	평가
Question-to-Query	2,807	413	163
Query Rewriting	2,986	439	154

평가 점수)로 이루어져 있다. 질문은 Question-to-Query 평가에서는 구어체 질문을 Query Rewriting 평가에서는 키워드형 검색 질의로 사용했고, 답변 쌍은 Question-to-Query와 Query Rewriting에 각각 미세 조정된 네이버 내 자체 구축한 한국어 LLM 모델을 사용하여 생성되었다.

LLM 앙상블 기반 출력값 생성 평가 점수와 평가 이유를 포함한 출력값은 훈련 데이터셋 구성의 가장 중요한 부분이다. 본 연구에서는 다양한 크기의 네이버 내 자체 구축한 한국어 LLM 모델을 앙상블하여 편향을 줄이고 주석 다양성을 확보한다. 구체적으로, [12] 연구를 따라 프롬프트 예제와 배치 사이즈를 조정하여 다양한 평가 결과를 생성한 후, 이들을 앙상블 하여 최종 출력값을 결정한다. 또한 [11]에 따르면 답변의 순서에 따라 평가 결과가 바뀌는 위치 편향이 발생할 수 있기 때문에, 모든 앙상블 모델은 각 예제에 대해 위치를 바꿔서 두 번의 평가를 수행하고, 그 결과를 종합하여 다수결 합의를 도출한다.

데이터 후가공 훈련 데이터의 노이즈를 줄이고 품질을 높이기 위해 출력값 생성 단계에서 다수결의 합의에 도달하지 못하는 경우 해당 예제는 훈련 데이터셋에서 제외한다. 이 과정을 통해 약 37%의 데이터가 필터링되었으며, 최종적으로 Question-to-Query 3,220개, Query Rewriting 3,142개의 데이터를 구성한다. 이후 학습과 검증 데이터셋을 약 7:1 로 나눈다.

3.4 평가 데이터셋 구성

우리는 FubaoLM의 신뢰도를 보장하기 위해, 평가 LLM을 평가할 수 있는 답변 쌍에 대해 인간이 주석을 단 고품질 평가 데이터셋을 제작한다. 평가 데이터의 각 예시는 한 개의 명령어와 질문, 두 개의 서로 다른 인스트럭션 튜닝된 LLM이 생성한 답변 (답변1, 답변2), 두 개의 답변에 대한 인간의 평가 점수로 구성된다. 각 답변에 대한 평가 점수는 검색 관련 태스크에 대한 이해가 높은 2명의 전문가가 수동으로 주석을 달았다. 이때 Cohen's kappa 점수 [13]로 측정한 주석자 간 동의의 정도는 태스크별 0.92, 0.94였다. 우리는 두 명의 전문가가 모두 같은 평가 점수를 할당한 예제만을 평가 데이터로 사용한다. Question-to-Query 태스크의 평가 데이터는 총 163개로 답변 1이 이긴 횟수 61회, 답변 2가 이긴 횟수 62회, 두 답변이 비긴 횟수가 40회로

당신은 평가 AI입니다. 명령어에 따라 아래에 표시된 사용자의 질문에 대해 AI 어시스턴트의 답변 품질을 공정하게 평가해주세요.
평가할 때 등급별로 해당되는 사항을 고려해주시면 됩니다.

[[1]]: 질문에서 사용된 요청의 말투를 유지하였습니다.
[[2]]: 구어체를 생략하고 키워드 형태로 바꾸면서 중요한 단어가 누락되거나, 새로운 단어가 추가되어 의미가 변경되었습니다.
[[3]]: 구어체를 생략하고 키워드 형태로 바꾸었고 중요한 단어는 유지되었거나 유의어로 대체되었으며, 조사나 어미 등이 추가되어도 질의의 의미가 그대로 유지되었습니다.
[[4]]: 구어체를 생략하고 키워드 형태로 바꾸었고 중요한 단어는 유지되었거나 유의어로 대체되었으며, 불필요한 조사나 어미 등이 제거되어 간결하게 표현되었습니다.

평가를 시작하기 전에 간단한 설명을 제공해주시고, 가능한 한 객관적으로 평가해주세요. 평가를 마친 후에는 반드시 "[[rating]]" 형식을 사용하여 답변1, 답변2의 응답을 1부터 4까지의 등급으로 평가해주시면 됩니다. 귀하의 평가는 다음과 같은 형식을 따라야 합니다.

[평가 이유] : <여기에 평가 이유를 작성하세요>
[답변1 점수] :[[rating]]
[답변2 점수] :[[rating]]

[명령어]
다음 질문을 검색에 적합한 키워드 형태로 바꿔주세요.

[질문]
{Q1}

[답변1 시작]
{R1}
[답변1 끝]

[답변1 시작]
{R2}
[답변2 끝]

그림 3. Question-to-Query 태스크에 대한 평가 프롬프트 예시. 총 3개의 슬롯-질문 (Q1), 어시스턴트 A 답변 (R1), 어시스턴트 B 답변 (R2) 이 포함된다.

구성된다. Query Rewriting의 평가 데이터는 총 154개로 답변 1이 이긴 횟수 42회, 답변 2가 이긴 횟수 48회, 두 답변이 비긴 횟수가 64회로 구성된다. 학습 및 평가 데이터셋의 통계는 표 1 과 같다.

3.5 평가자 모델 FubaoLM

FubaoLM은 네이버 내 자체 구축한 한국어 LLM 모델을 본 논문에서 제안하는 평가자용 훈련 데이터로 미세 조정한 모델이다. 훈련 시에는 Adam 옵티마이저와 Super convergence 스케쥴러 [14]를 사용한다. 하이퍼파라미터로 러닝 레이트는 {1e-5, 2e-5}, 훈련 에폭은 {5,7}, 학습 배치 크기는 16으로 설정했다. 학습 과정에서는 [15] 연구를 기반으로 두 가지 서로 다른 프롬프트 설정을 사용했다. 이 프롬프트 설정은 평가 기준과 형식만을 포함하는 제로샷 프롬프트와, 제로샷 프롬프트에서 2개의 예제를 추가한 퓨샷 프롬프트의 두 가지 버전으로 구성되었으며, 이를 8:1:1 비율로 랜덤하게 혼합하여 활용했다.

표 2. 두 검색 관련 태스크 (Question-to-Query, Query Rewriting)에서 FubaoLM과 베이스라인 모델들의 절대 평가 성능 비교. 평가 메트릭인 Accuracy와 F1은 인간 평가자의 평가 결과를 정답 라벨로 간주하여 측정했다. 표에서 볼드체 표시는 최고 성능, 밑줄 표시는 두 번째 최고 성능을 나타낸다.

평가자	Question-to-Query		Query Rewriting	
	Accuracy	F1	Accuracy	F1
GPT-3.5	30.26	27.27	30.37	30.44
Llama-2-Ko-7B-FubaoLM	26.87	20.27	26.58	21.07
FubaoLM-7B	33.12	33.08	29.11	23.99
FubaoLM-60B	<u>44.37</u>	<u>42.39</u>	32.91	28.92
FubaoLM-Ensemble	35.62	30.08	<u>37.97</u>	<u>36.21</u>
FubaoLM-GPT-3.5-Ensemble	**51.87**	**56.86**	**46.83**	**43.72**

당신은 평가 AI입니다.

... *(생략)* ...

[[1]] : 질문에서 사용된 질문이나 요청의 말투를 유지하거나, 질문에 대한 답변을 하였습니다.
[[2]]: 키워드 형태의 답변이지만 중요한 단어가 누락되거나 대체되면서 질문의 의미가 변경되었습니다.
[[3]]: 키워드 형태의 답변이며 질문의 중요 단어가 보존(/동의어로 대체)되었으나 새로운 단어 추가로 인한 질문의 의미가 변경되었습니다.
[[4]]: 검색어 형태의 답변이며 질문의 중요 단어가 보존(/동의어로 대체)되었으며 새로운 단어가 추가된 경우에도 질문과 의미가 동일합니다.
[평가 이유] : <여기에 평가 이유를 작성하세요>
[답변1 점수] :[[rating]]
[답변2 점수] :[[rating]]

[명령어]
다음 질문을 검색에 적합한 키워드 형태로 바꿔주세요.

[질문]
{Q1}

[답변1 시작]
{R1}
[답변1 끝]

[답변1 시작]
{R2}
[답변2 끝]

그림 4. Query Rewriting 태스크에 대한 평가 프롬프트 예시. 총 3개의 슬롯-질문 (Q1), 어시스턴트 A 답변 (R1), 어시스턴트 B 답변 (R2) 이 포함된다.

4. 실험 및 결과

4.1 실험 세팅

본 논문에서는 FubaoLM을 포함하여 평가자 LLM 에 대한 두 가지 주요 평가 방법으로 절대 평가와 이항 평가를 도입한다.

절대 평가 모델의 답변에 부여하는 점수에 대한 평가자 모델의 평가 정확성을 엄격하게 측정하기 위해 절대 평가 방법을 도입한다. 이 과정에서는 평가 지침에 따라 모델이 얼마나 정확한 평가를 수행하는지를 측정한다. 두 개의 전문가 평가 데이터에서 4개의 점수가 고루 분포하게 하여 80개, 79개의 데이터를 추출하여 평가한다. 평가자 모델의 위치 편향을 제거하기 위해 프롬프트에서 답변1과 답변2의 위치를 변경하여 2회 평가를 수행하고, 이에 대한 평균적인 정확도를 계산하여 최종 절대 점수로 사용한다. 예를 들어, 답변1이 답변2의 앞에 있을 때 3, 뒤에 위치해 있을 때 4, 정답 점수가 4라면 3.5가 해당 예제의 최종 점수가 되고 최종 점수(3.5)와 정답 점수(4)가 다르므로 틀린 것으로 간주한다. 이때 사용되는 절대 평가 측정 메트릭은 각 답변에 사람이 할당한 절대 평가 점수에 대한 평가자 모델의 정확도와 F1 점수를 사용한다.

이항 평가 두 개의 다른 답변 사이에서 어떤 것이 더 우수한지를 평가하기 위해 상대 평가 방법을 도입한다. 이 방법은 win/tie/lose(승/무/패)의 형태로 각 쌍의 답변을 평가하는 것을 의미한다. 위치 편향을 고려한 평가를 위해 각 쌍의 답변 순서를 바꿔 2회씩 평가하고 평균 내어 절대 평가 점수를 계산하고, 이를 통해 답변의 우위 여부를 확인하여 상대 평가를 내린다. 예를 들어 답변1과 답변2가 각각 {1,2}, {4,4} 점수를 받았다면 절대 평가 점수는 1.5, 4가 되고 답변 2의 점수가 높아서 답변 1이 패하게 된다.

베이스라인 우리는 전문가 기반 평가 데이터셋에서 FubaoLM의 평가 성능을 평가하기 위해 2개의 강한 대규모 언어모델을 베이스라인으로 도입한다: GPT-3.5, Llama-2-Ko-7B-FubaoLM. Llama-2-Ko-FubaoLM 모델은 Llama-2-Ko-7B[16]를 한국어 지시 데이터셋으로 미세 조정한 모델[17]을 FubaoLM의 훈련 데이터셋을 사용하여 추가 학습시킨 모델로, FubaoLM의 베이스 모델과 Llama-2-Ko에 대한 비교를 위해 사용한다. 우리는 3개의 프롬프트를 제공하여 각 모델당 3회의 평가 기회를 제공하고, 표 2에 가장

표 3. Question-to-Query 태스크에서 다양한 한국어 LLM에 대한 상대 평가 결과. 표 내부의 괄호는 세로축의 모델이 가로축의 모델과 비교하여 이긴 횟수, 비긴 횟수, 진 횟수 (승/무/패)를 나타낸다.

평가자	평가 대상	Llama-2-Ko-7B	komt-7B	KoAlpaca-5.8B	KuLLM-5.8B
Human	Llama-2-Ko-7B	-	(70, 78, 15)	(100, 59, 4)	(64, 84, 15)
	komt-7B	(15, 78, 70)	-	(86, 32, 45)	(40, 80, 43)
	KoAlpaca-5.8B	(4, 59, 100)	(45, 32, 86)	-	(26, 44 93)
	KuLLM-5.8B	(15, 84, 64)	(43, 80, 40)	(93, 44, 26)	-
GPT-3.5	Llama-2-Ko-7B	-	(42, 64, 57)	(51, 52, 60)	(33, 53, 33)
	komt-7B	(57, 64, 42)	-	(52, 59, 52)	(49, 66, 48)
	KoAlpaca-5.8B	(60, 52, 51)	(52, 59, 52)	-	(33, 63, 67)
	KuLLM-5.8B	(77, 53, 33)	(48, 66, 49)	(67, 63, 33)	-
FubaoLM-7B	Llama-2-Ko-7B	-	(55, 57, 51)	(53, 58, 52)	(54, 72, 37)
	komt-7B	(52, 58, 53)	-	(49, 67, 47)	(38, 69, 56)
	KoAlpaca-5.8B	(51, 57, 55)	(47, 67, 49)	-	(49, 67, 47)
	KuLLM-5.8B	(37, 72, 54)	(56, 69, 38)	(47, 67, 49)	-
FubaoLM-Ensemble	Llama-2-Ko-7B	-	(71, 56, 36)	(60, 73, 30)	(66, 64, 33)
	komt-7B	(36, 56, 71)	-	(57, 76, 30)	(19, 61, 83)
	KoAlpaca-5.8B	(30, 73, 60)	(30, 76, 57)	-	(30, 61, 72)
	KuLLM-5.8B	(33, 64, 66)	(83, 61, 19)	(72, 61, 30)	-

좋은 결과를 보고한다.

앙상블 비교 우리는 FubaoLM을 기반으로 더 신뢰할 수 있는 평가를 할 수 있는지 분석하기 위해 2 개의 앙상블 변주를 도입한다. FubaoLM-Ensemble은 FubaoLM-7B 와 FubaoLM-60B 모델을 앙상블 한 방법에 대한 표기이며, FubaoLM-GPT-3.5-Ensemble은 다음 3개 모델을 모두 앙상블 한 결과를 말한다: FubaoLM-7B, FubaoLM-60B, GPT-3.5. 우리는 위치 편향을 고려하여 모델들을 앙상블 하기 위해, 각 모델당 순서를 변경하여 2회의 측정을 진행하고 모든 모델의 평가 결과를 취합하여 다수결로 예측된 평가 결과를 해당 답변의 최종 평가 결과로 사용한다. 각 방법에서 앙상블 가중치는 FubaoLM-Ensemble은 FubaoLM-7B, FubaoLM-60B 모델 2:1로, FubaoLM-GPT-3.5-Ensemble은 FubaoLM-7B, FubaoLM-60B, GPT-3.5를 1:1:2로 설정했다.

4.2 실험 결과 및 분석

절대 평가 결과 표 2는 본 논문의 주요 실험 결과를 나타낸다. 두 개의 검색 관련 태스크에서 각각 절대평가를 진행하였다. 절대 평가에서 우리는 FubaoLM-60B는 GPT-3.5보다 약 2점에서 14점 향상된 평가 결과를 내는 것을 확인하였다. 더욱이, FubaoLM-7B는 크기가 약 25배 큰 GPT-3.5와 비교하였을 때, Question-to-Query에서 더 높은 평가 성능을 달성하고, Query Rewriting에서 유사한 수준의 평가를 보인다. FubaoLM-Ensemble 모델은 Query Rewriting 태스크에서 단일 FubaoLM보다 더 좋은 평가 성능을 내며, Question-to-Query에서는 FubaoLM-7B와 유사한 평가 정확도를 보였다. 마지막으로 FubaoLM-GPT-3.5-Ensemble 모델은 두 검색 태스크에서 가장 좋은 평가 성능을 달성했다.

이항 평가 결과 평가자로서 LLM은 평가 대상에 대한 절대 평가에 기반하여 두 평가 대상에 대한 우위를 비교할 수 있다. 표 3는 인간, GPT-3.5, FubaoLM를 각각 평가자로서 이항 평가 결과를 나타낸다. 이 실험에서 인간 평가자는 4개 한국어 오픈소스 LLM에 대한 이항 평가를 통해 앞 순서일수록 더 좋은 모델로 나열했을 때 Llama-2-Ko-7B, KuLLM-5.8B [18], komt-7B [19], KoAlphaca-5.8B [20]의 평가 순위를 매겼다. 한편, GPT-3.5의 경우, KuLLM-5.8B, (komt-7B, KoAlphaca-5.8B), Llama-2-Ko-7B 의 평가 순위를 할당했다. 이때 괄호로 묶인 komt-7B와 KoAlphaca-5.8B는 동점을 말한다. GPT-3.5의 평가 결과는 인간 평가와 매우 상반된 결과를 보인다. 반면, FubaoLM-7B와 FubaoLM-Ensemble 모델의 경우, Llama-2-Ko-7B, KuLLM-5.8B, komt-7B, KoAlphaca-5.8B의 평가 순위를 할당했으며 인간과 동일한 이항 평가를 내리는 것을 보인다. FubaoLM-7B가 GPT-3.5 (175B)보다 이항 평가에서 더 우수한 결과를 냈으며, 평가자로서 LLM으로의 FubaoLM가 효과적임을 보인다.

5. 결론

본 논문에서는 평가자로서 LLM을 학습할 수 있는 프레임워크와 평가자 LLM인 FubaoLM을 제안한다. 우리의 프레임워크에서 여러 개의 강한 LLM은 엄격한 평가 기준에 의해 연쇄적 사고 기반 평가를 내리고, 다수의 평가 결과를 앙상블 한 데이터를 구축하여 평가자 모델인 FubaoLM에게 증류한다. 우리의 실험에서 FubaoLM-Ensemble 모델은 GPT-3.5 대비 16% 에서 23% 향상된 절대 평가 성능을 가지며, 이항 평가에서 인간과 유사한 선호도 평가 결과를 도출한다. 추후 검색 관련 태스크 이외 다양한 한국어 태스크에서도 우리의 방법을 확장할 예정이다.

참고문헌

[1] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie *et al.*, "Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization," *arXiv preprint arXiv:2306.05087*, 2023.

[2] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *arXiv preprint arXiv:2306.05685*, 2023.

[3] D. Jiang, X. Ren, and B. Y. Lin, "Llm-blender: Ensembling large language models with pairwise ranking and generative fusion," *arXiv preprint arXiv:2306.02561*, 2023.

[4] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.

[5] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," 2023.

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24 824–24 837, 2022.

[7] X. Li, P. Yu, C. Zhou, T. Schick, L. Zettlemoyer, O. Levy, J. Weston, and M. Lewis, "Self-alignment with instruction backtranslation," *arXiv preprint arXiv:2308.06259*, 2023.

[8] L. H. Li, J. Hessel, Y. Yu, X. Ren, K.-W. Chang, and Y. Choi, "Symbolic chain-of-thought distillation: Small models can also" think" step-by-step," *arXiv preprint arXiv:2306.14050*, 2023.

[9] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 8, No. 4, p. e1249, 2018.

[10] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023.

[11] P. Wang, L. Li, L. Chen, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui, "Large language models are not fair evaluators," *arXiv preprint arXiv:2305.17926*, 2023.

[12] Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto, "Alpacafarm: A simulation framework for methods that learn from human feedback," *arXiv preprint arXiv:2305.14387*, 2023.

[13] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, Vol. 20, No. 1, pp. 37–46, 1960.

[14] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," *Artificial intelligence and machine learning for multi-domain operations applications*, Vol. 11006, pp. 369–386, 2019.

[15] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei *et al.*, "The flan collection: Designing data and methods for effective instruction tuning," *arXiv preprint arXiv:2301.13688*, 2023.

[16] J. Lee, "Llama-2-ko," https://huggingface.co/beomi/llama-2-ko-7b, 2023.

[17] H. Kim, "Llama-2-ko-7b-chat," https://huggingface.co/heegyu/llama-2-ko-7b-chat, 2023.

[18] N. . A. Lab and H.-I. A. research, "Kullm: Korea university large language model project," https://github.com/nlpai-lab/kullm, 2023.

[19] C. Kim, "komt-llama-2-7b," https://github.com/davidkim205/komt, 2023.

[20] H. Ko, K. Yang, M. Ryu, T. Choi, S. Yang, jiwung Hyun, and S. Park, "A technical report for polyglot-ko: Open-source large-scale korean language models," 2023.