Mixture-of-Agents Enhances Large Language Model Capabilities

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, James Zou
Together AI | Duke University | University of Chicago | Stanford University
7 Jun 2024

발표자 : 장자윤

Table of contents

Introduction

• LLM은 방대한 양의 데이터에 대해 사전 학습된 후 인간의 선호도에 맞춰 조정되어 유용하고 일관된 결과를 생성함. 그러나 여전히 모델 크기와 학습 데이터에 대한 본질적인 제약에 직면해 있음. LLM을 추가로 확장하려면 비용이 매우 많이 들고, 종종 수조 개의 토큰에 대한 광범위한 재학습이 필요하다.

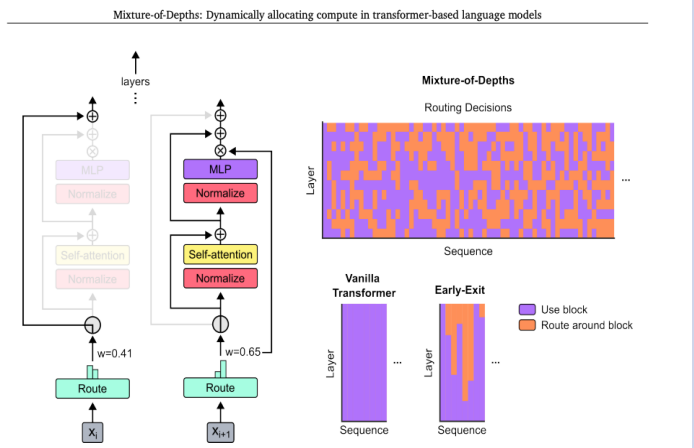
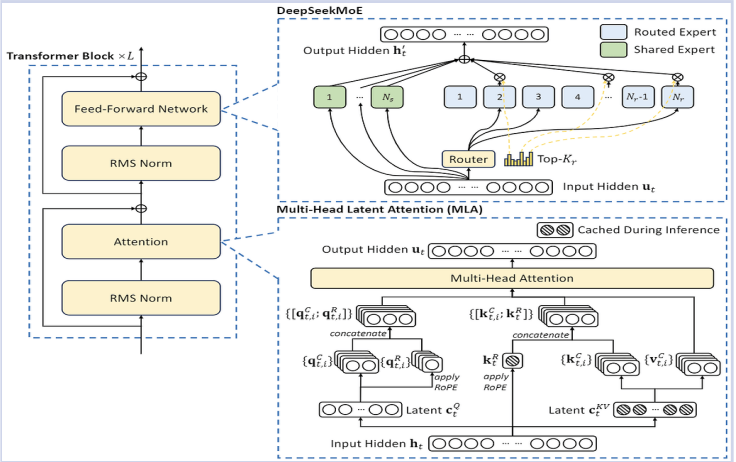
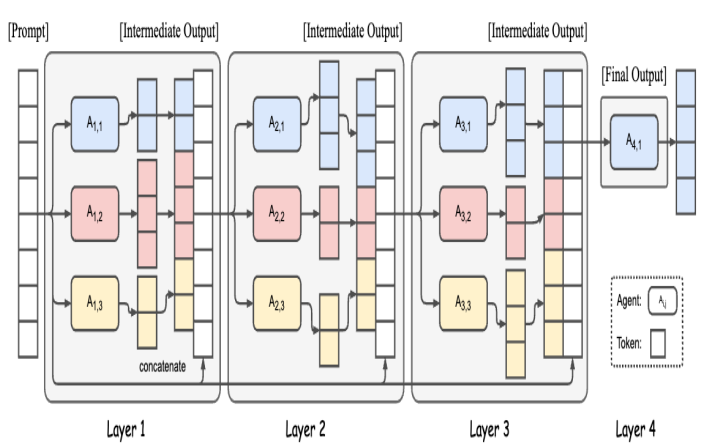
• 동시에 다양한 LLM은 고유한 강점을 보유하고 있으며 다양한 task를 전문으로 함. 예를 들어 일부 모델은 복잡한 명령을 따르는 데 탁월한 반면 다른 모델은 코드 생성에 더 적합할 수 있다. 다양한 LLM 사이의 다양성은 흥미로운 질문을 제시한다.

?Qustion

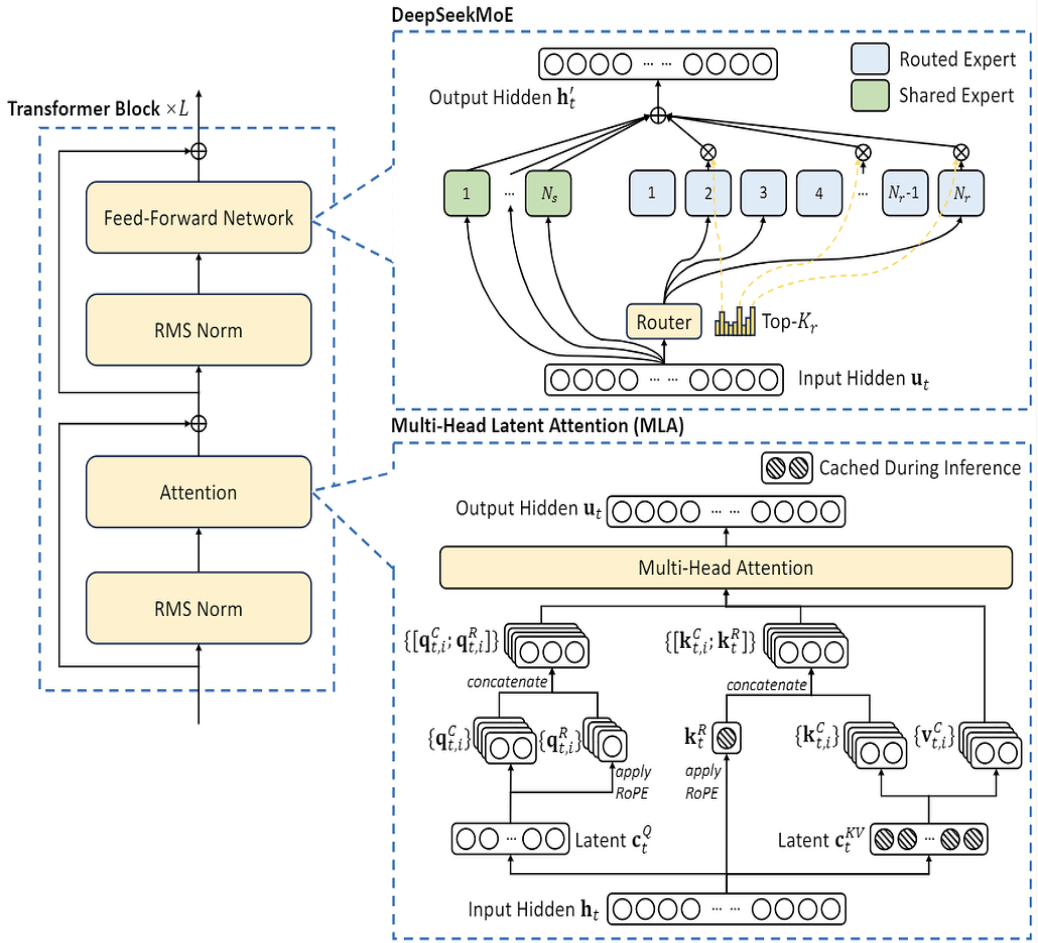
• 여러 LLM의 집단적 전문 지식을 활용하여 보다 유능하고 강력한 모델을 만들 수 있을까?

- 다수의 LLM이 계층적으로 협력하여 응답을 생성하는 새로운 프레임워크인 Mixture-of-Agents(MoA)등장.
- 기존 LLM의 구조를 변경하지 않고, 프롬프트와 샘플링만으로도 성능 향상 가능

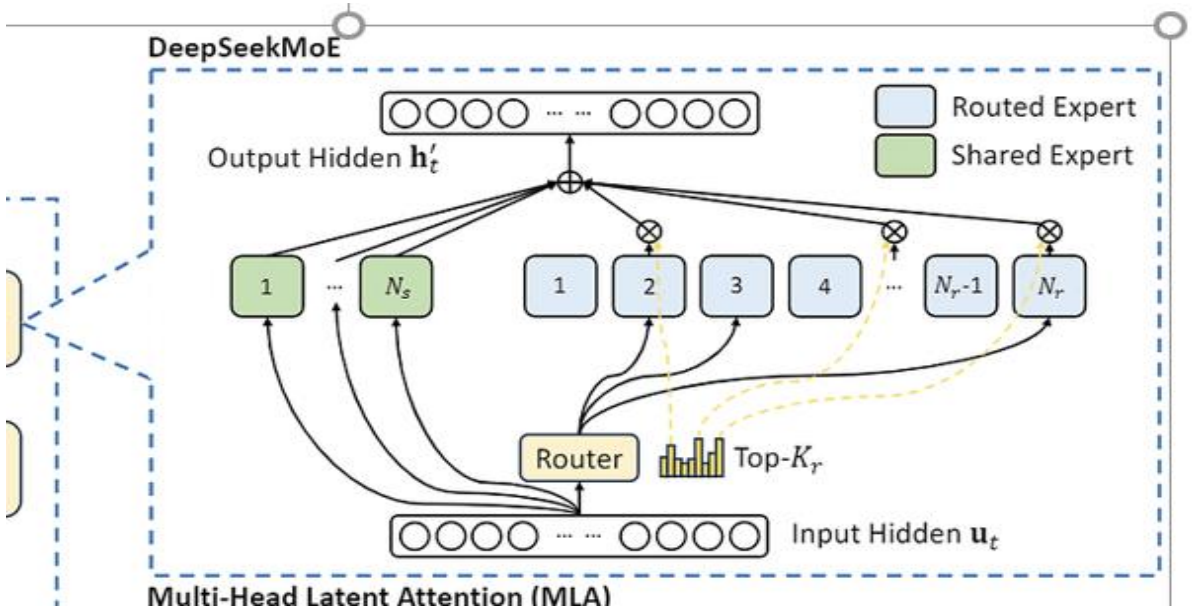
LLM 성능 향상 기법들

MoD (Mixture-of-Depths)	MoE (Mixture-of-Expert)	MoA (Mixture-of-Agent)
		
활성화된 레이어를 줄여서 깊이를 줄이는 기법	학습 시점에 여러 전문가 모델들을 포함하도록 학습하여 모델의 크기는 크지만, 실행시점에는 사용자의 질문에 대해 적절히 답변할 수 있는 일부 전문가들만 활성화하는 방식으로 모델의 성능을 향상시키는 기법	여러 개의 LLM이 계층적으로 협력하여 응답을 생성하는 프레임워크

MoE 기법



- DeepSeekMoE 아키텍처의 핵심 구성요소를 시각화한 구조도
- Transformer 기반 구조에 Mixture-of-Experts (MoE)와 Multi-Head Latent Attention (MLA) 기법을 결합한 형태



- 입력 hidden vector ut 가 들어오면, Router가 이를 받아서 Top-kr개의 전문가를 선택
- 전문가 Expert는 Routed Expert와 Shared Expert로 구성

  Routed expert = 전용 전문가
  Shared expert = 공유 전문가(모든 입력에 공통으로 적용)

- 선택된 전문가에만 입력이 전달되어 연산이 수행되고, 그 결과가 다시 출력 hidden vector ht' 로 통합
- 모든 expert를 쓰지 않고 일부만 선택하여, 연산 효율성 ↑
  일부는 공유 전문가로 설정하여 정보 일관성을 유지

MoA란?

- Mixture-of-Agents(MoA) 방법론은 여러 LLM의 집단적 전문 지식을 계층 구조를 통해 활용하는 것을 목표
- 각 계층은 여러 LLM 에이전트로 구성되며, 이전 계층의 출력물을 기반으로 응답을 생성하여 최종 출력을 점진적으로 개선
- MoA 기법의 핵심 아이디어는 대규모 언어 모델(LLM)이 다른 모델의 답변을 참고할 때 더 높은 품질의 응답을 생성할 수 있다는 점에서 시작

MoA 주요특징

- 계층 구조 : MoA 프레임워크는 여러 LLM 에이전트로 구성된 다층 구조를 사용. 각 에이전트는 이전 계층의 응답을 개선하여 최종 출력을 점진적으로 향상
- 모델 다양성 : 프레임워크는 각 계층에서 다양한 LLM을 사용하는 것을 강조합니다. 다양한 모델이 결합될 때 더 풍부하고 세밀한 응답을 생성
- 반복적 개선 : 반복적인 과정은 생성된 텍스트를 지속적으로 개선할 수 있게 하며, 여러 모델의 협력적 합성을 통해 최상의 결과를 도출

MoA의 구성 요소

- **Collaborativeness of LLMs**

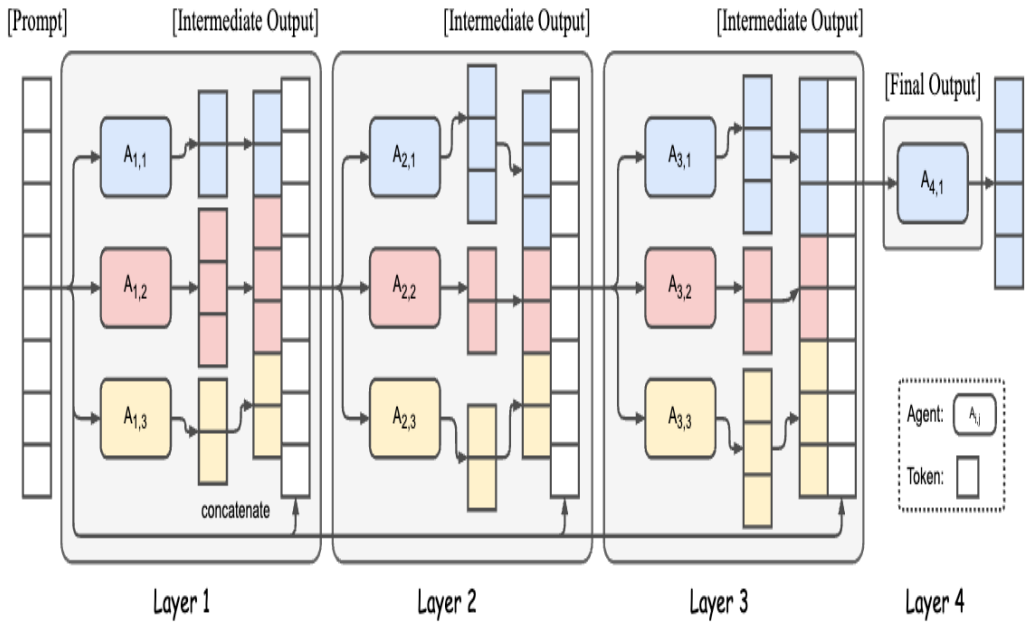
- LLM의 강점을 모아서 집단적 전문 지식을 활용하여 성능을 향상시킨다. 이러한 에이전트 혼합 기법은 모델이 다른 모델의 답변을 활용하고, 그 답변의 품질이 낮더라도 더 나은 품질의 답변을 생성하는 LLM간의 협령성에 기초하고 있다.

<LLM의 역할>

Proposer LLM : 다른 모델에서 사용할 수 있는 유용한 참조 응답을 생성하는 데 탁월한 LLM이다. 좋은 제안자는 그 자체로 반드시 높은 점수를 받는 응답을 생성하지는 않지만, 더 많은 맥락과 다양한 관점을 제공하여 궁극적으로 Aggregator와 함께 사용할 때 더 나은 최종 응답에 기여할 수 있다.

Aggregator LLM : 다른 모델의 응답을 하나의 고품질 출력으로 합성하는데 능숙한 모델이다. 효과적인 집계자는 집계자 LLM이 자체적으로 생성할 수 있는 답변의 품질보다 낮은 품질의 입력이 Propose로부터 들어왔을 때도 최종 답변의 품질을 유지하거나 향상시킬 수 있어야 한다.

MoA의 구조



- 여러 계층(Layer)으로 구성되며, 각 계층(Layer)에는 여러개의 LLM들이 포함 됨. 이 구조에서 각 계층의 에이전트는 이전 계층의 모든 출력물을 보조 정보로 사용하여 응답을 생성한다. 이 때 유의해야 할 점은 각 LLM이 동일한 계층(layer) 및 서로 다른 계층(layer)에서 재사용될 수 있다.

- 첫 번째 계층의 LLM이 독립적으로 주어진 프롬프트에 대한 응답을 생성한다. 그 다음, 이 응답은 다음 계층의 에이전트에게 전달되어 더 정교한 응답을 생성한다. 이 과정은 최종적으로 더 정밀하고 종합적인 응답이 생성될 때까지 반복한다.

- 여러 차례 반복되며, 최종적으로 더 강력하고 종합적인 응답을 얻을 수 있다. 이를 통해 개별 모델의 한계를 극복하고, 보다 다양한 정보와 관점을 통합한 고품질의 응답을 생성할 수 있으며, 특히 복잡한 문제 해결에서 매우 유용하다.

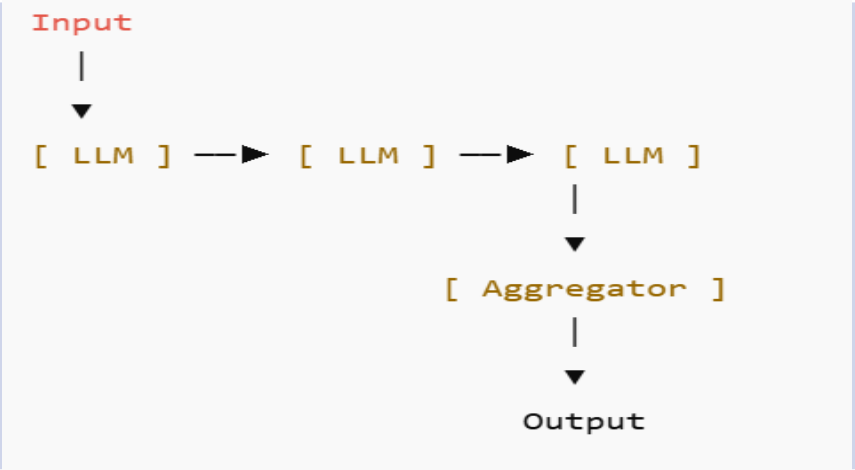
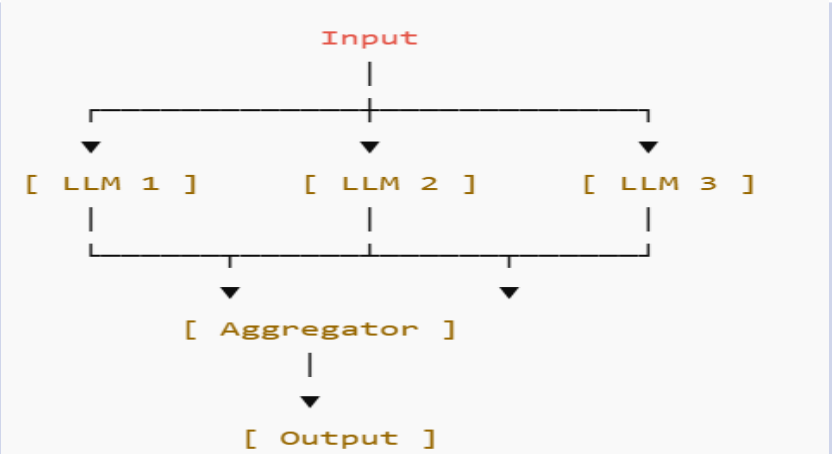
$$y_i = \oplus_{j=1}^{n} [A_{i,j}(x_i)] + x_1, \quad x_{i+1} = y_i$$

- 각 LLM Ai,j는 입력 텍스트를 처리한다.

- 입력 프롬프트 x1이 주어지면 i번째 MoA 레이어 yi의 출력은 위의 식으로 표현 된다.

- +는 텍스트의 concatenation을 의미하며, ⊕는 아래 표와 같은 Aggregate-and-Synthesize 프롬프트를 모델 출력에 적용함을 의미

- 실제로는 프롬프트와 모든 모델 응답을 concatenate할 필요가 없으므로 마지막 레이어에 하나의 LLM 만 사용하면 된다. 따라서 l번째 레이어의 LLM 출력 Al,1(xl) 을 최종 출력으로 사용

단일 제안자 구조 (Single-Proposer)	다중 제안자 구조 (Multi-Proposer)
	
- 동일한 LLM을 여러 계층(layer)에 걸쳐 반복적으로 활용 - 각 계층에서는 다른 샘플링 설정 적용하여 서로 다른 출력을 생성 - 생성된 출력들은 마지막 단계에서 집계자에 의해 통합되어 최종 응답이 만들어짐.	- 각 계층에 다양한 모델을 사용하여 각기 다른 출력을 생성 - 모델 간의 상호작용과 협력성을 최대화하여 더 포괄적이고 높은 품질의 응답을 생성 - 다중 제안자 구조는 모델의 다양성을 최대한 활용하여 문제 해결의 폭을 넓히고, 단일 모델이 가지고 있는 한계를 극복할 수 있고, 이러한 설정을 통해 다중 에이전트 구조는 보다 강력하고 종합적인 솔루션을 제공할 수 있습니다.
응답 시간과 연산 자원이 중요한 경량 챗봇, 실시간 고객 응대 시스템, 내장형 어시스턴트등의 주로 사용 됨.	응답 품질이 핵심인 의료·법률 분야의 전문 지식 응답, 연구 보조 시스템, 멀티모달 요약기등의 활용됨.

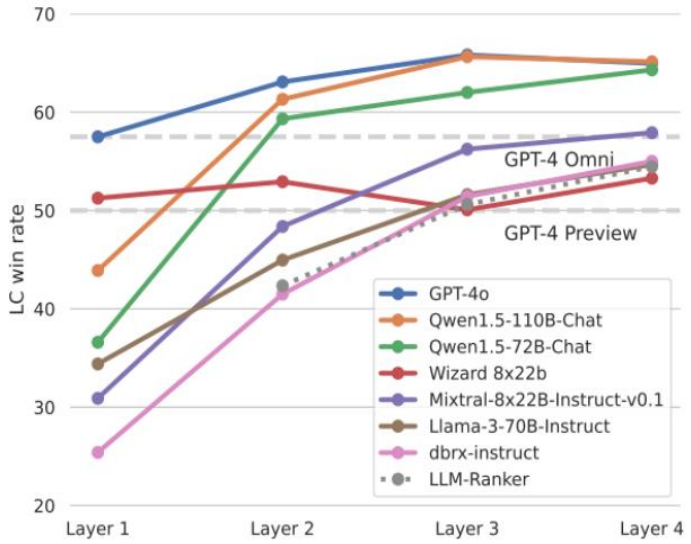
Evaluation- 성능평가

(a) AlpacaEval 2.0

Model	LC win.	win.
MoA w/ GPT-4o	$65.7_{\pm 0.7}\%$	$78.7_{\pm 0.2}\%$
MoA	$65.1_{\pm 0.6}\%$	$59.8_{\pm 0.3}\%$
MoA-Lite	$59.3_{\pm 0.2}\%$	$57.0_{\pm 0.7}\%$
GPT-4 Omni (05/13)	57.5%	51.3%
GPT-4 Turbo (04/09)	55.0%	46.1%
WizardLM 8x22B†	51.3%	62.3%
GPT-4 Preview (11/06)	50.0%	50.0%
Qwen1.5 110B Chat	43.9%	33.8%
Qwen1.5 72B Chat	36.6%	26.5%
GPT-4 (03/14)	35.3%	22.1%
Llama 3 70B Instruct	34.4%	33.2%
Mixtral 8x22B v0.1	30.9%	22.2%

(b) MT-Bench.

Model	Avg.	1st turn	2nd turn
MoA w/ GPT-4o	$9.40_{\pm 0.06}$	9.49	9.31
GPT-4 Turbo (04/09)	9.31	9.35	9.28
MoA	$9.25_{\pm 0.10}$	9.44	9.07
GPT-4 Preview (11/06)	9.20	9.38	9.03
GPT-4 Omni (05/13)	9.19	9.31	9.07
MoA-Lite	$9.18_{\pm 0.09}$	9.38	8.99
Qwen1.5 110B Chat	8.96	9.23	8.63
Llama 3 70B Instruct	8.94	9.2	8.68
Mixtral 8x22B v0.1	8.78	9.11	8.44
WizardLM 8x22B	8.78	8.96	8.61
Qwen1.5 72B Chat	8.44	8.55	8.34
GPT-4 (06/13)	8.84	9.08	8.61



AlpacaEval = Stanford에서 개발한 LLM 간 응답 비교 평가 도구

MoA w/GPT-4o = 여러 GPT-4o 모델 인스턴스를 MoA 구조로 구성하여 협업적으로 응답을 생성한 시스템

MoA-Lite = 경량화된 MoA구조를 의미하며, 원래 MoA의 구조와 원리는 유지하면서도 리소스를 줄이고, 구현 난이도를 낮춘 버전

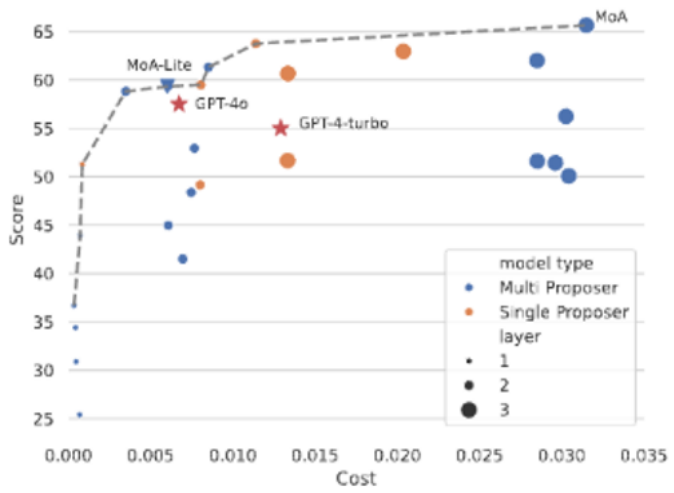
MT-Bench = 대화 흐름에 따라 모델이 얼마나 일관되고 정확한지를 평가 도구

→ MoA 구조가 일반적인 단일 LLM보다 더 뛰어난 응답 품질을 만들어냄을 명확히 보여줌

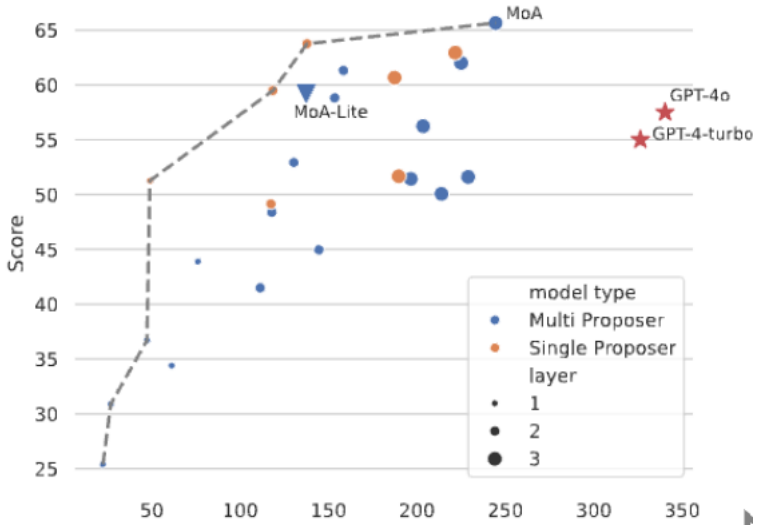
→ 특히 MoA w/ GPT-4o는 성능 지표 두 곳 모두에서 최상위 점수

→ MoA-Lite도 비용 대비 괜찮은 성능으로 실용성 있음

Evaluation-토큰 및 비용 효율성



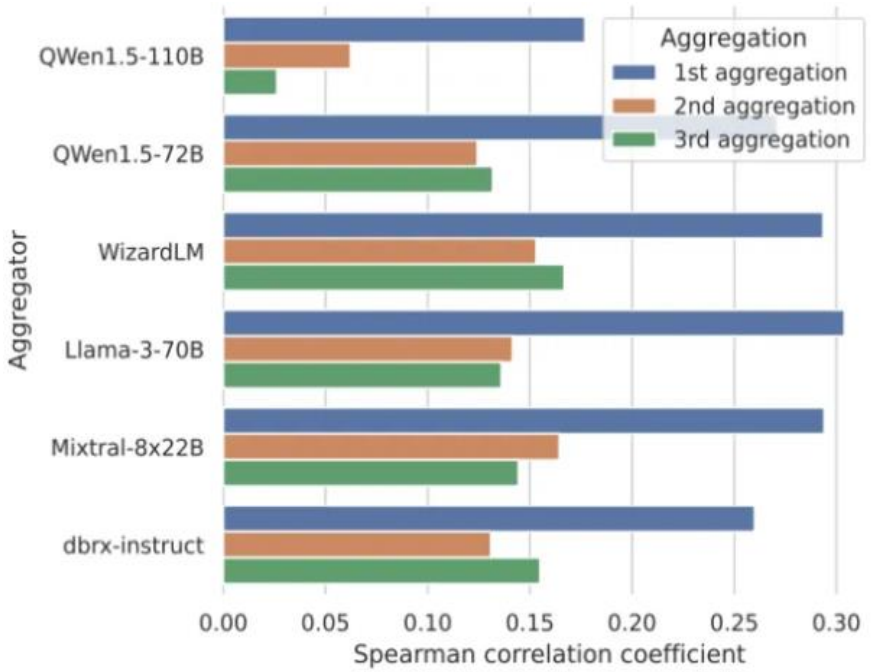
(a) LC win rate vs. cost



- 성능과 비용 사이의 trade-off를 나타낸 그래프
- 각 API 제공자의 비용을 바탕으로 계산하였으며, 이는 MoA 기법이 높은 성능을 달성하면서 과도한 비용을 발생시키지 않는 비용 효율적인 방법임을 보여줌
- 특히, MoA-Lite는 GPT-4 Turbo보다 약 4% 더 우수하면서 비용 효율성은 두 배 이상이다.

- 성능과 Tflops 사이의 trade-off를 나타낸 그래프
- Tflops수를 지연 시간(latency)를 나타낼 수 있는 값으로 대신 사용하고 있다.
- 비용 효율성 분석과 유사하게 파레토 프런티어(Pareto frontier)가 관찰된다. 즉, 연산 자원을 효율적으로 사용하면서 LC 승률을 최대화하고 있다.

Evaluation-Blue score



- BLEU 점수 (3-gram, 4-gram, 5-gram)를 통해 aggregator의 응답과 proposer의 응답을 비교한 결과
- 각 샘플 내에서 proposer가 제안한 n개의 답변을 바탕으로 n개의 BLEU 점수와 GPT-4가 평가한 n개의 선호도 점수 사이의 Spearman 상관 계수를 계산한 결과
- MoA는 제안된 가장 좋은 답변을 통합하는 경향이 있다

MoA 장점

응답 품질 향상

다양성 확보

모델구조 변경없이 적용

비용 대비 성능 우수

Conclusion

- **MoA의 한계점**

1) MoA는 모델 응답의 반복적인 집계가 필요
   모델이 마지막 MoA 레이어에 도달할 때까지 첫 번째 토큰을 결정할 수 없음을 의미

2) 잠재적으로 Time to First Token (TTFT)이 길어져 사용자 경험에 부정적인 영향을 미칠 수 있다.

해결방법: 첫 번째 응답 집계가 생성 품질을 가장 크게 향상시키므로 이 문제를 완화하기 위해 MoA 레이어 수를 제한

- **핵심내용**

1) MoA는 여러 LLM의 출력을 통합하여 정확하고 다양성 높은 응답을 생성하는 협력 구조

2) 프롬프트와 샘플링 설정만으로 성능 향상이 가능하며, 구조 변경이나 미세조정 없이도 적용

3) GPT-4 수준의 성능을 달성하며,차세대 LLM 시스템 설계에 효과적인 전략으로 주목