LLaMA

What is LLaMA

1. LLaMA란?
   - Meta(구 Facebook)에서 개발한 LLM 시리즈
   - LLaMA 1,2,3 모델이 있으며, 최신버전은 3.3
   - 2025년 상반기 LLaMA 4 출시 예정

2. LLaMA의 특징
   - 오픈소스 모델 표방
   - 가벼운 모델(1B)부터 무거운 모델(405B)까지 다양하게 배포
   - GPT계열 LLM들과 마찬가지로 Decoder-only Transformer 구조 사용

LLaMA

Pros & Cons of LLaMA

1. LLaMA의 장점

- 접근성 : Huggingface에서 다운로드하여 온프레미스 서버 구축 가능
- 효율성 : 모델 크기에 비해 고성능
- 경제성 : 무료 또는 저렴한 가격으로 이용 가능

2. LLaMA의 한계

- 애매한 성능
- 멀티모달 기능 미지원
- 환각 이슈와 안정성 검증 부족

LLaMA

LLaMA에서 최초로 사용된 독자적인 기법은?

GQA (Grouped Query Attention)

- MHA(Multi Head Attention)의 개선된 형태
  - MHA는 query, key, value가 하나로 묶여 병렬 연산하는 방식
  - GQA는 그에 더해 query끼리 그룹화해서 더 효율적으로 연산하는 방식
  - 연산의 효율성과 속도를 최적화
- LLaMA 중에서도 큰 모델에서만 사용
  - LLaMa 2 70B 이상의 모델에서 최초로 도입
  - LLaMa 3에서는 8B 이상의 모델에서 사용
  - 그룹화하는 데 필요한 연산량 때문에 작은 모델에서는 부적합

LLaMA

LLaMA 1,2,3 비교

특성	Llama 1	Llama 2	Llama 3
파라미터 수	7B - 65B	7B - 70B	3B - 405B
학습 데이터	제한적	확장됨	더욱 풍부하고 다양함 [1]
모델 구조	기본 Transformer	개선된 Transformer	효율적인 Attention 메커니즘 [1]
메모리 관리	기본	향상됨	더 많은 입력 토큰 동시 처리 가능 [1]
미세 조정 능력	제한적	향상됨	더욱 정교해짐 [1]
한국어 능력	제한적	개선됨	더욱 자연스러운 처리 [1]
최대 모델 크기	65B	70B	405B [2]

- 학습 데이터 양을 크게 늘려나감
- 더 다양한 크기의 모델 제공
- LLaMA 3.1부터 멀티모달 기능 제공
- 다국어 지원에 집중

LLaMA

LLaMA의 과거, 현재, 미래

- 과거
  - 모델 경량화에 집중
  - 편향 및 유해성 검증, 환경 보호 등 사회적 이슈에 초점
- 현재
  - 아주 가벼운 모델부터 400B 이상의 큰 모델까지 다양한 선택지 제공
  - 잦은 업데이트와 최적화를 통해 사용성 개선
- 미래
  - AI 인프라(GPU)에 더욱 큰 투자를 통해 더 큰 모델에 도전
  - 여러 작업에 범용적으로 활용할 수 있는 AGI 개발 목표