

# 야나두 LLM 엔진 적용 검토



신규사업실

2024.04.01 / V1.0

## 문서변경내역

[illegible]

# Contents

- 01 배경 및 개요
- 02 LLM 엔진 종류 및 특성
- 03 LLM 엔진 상세 비교
- 04 테스트 시연 장면
- 05 기대 효과
- 06 정리
- 07 예상 비용

## 01. 배경 및 개요

### 배경

1. 글로벌 확장을 위해 AI 서비스 도입 필요
2. AI 메인스트림으로 방향성 확립
3. ChatGPT 베이스의 AI서비스는 비용 문제로 사업 확장성 낮음
4. 자체 엔진 적용 필요



### 개요

1. 목적: 야나두 내 AI 서비스 적용 가능 여부 검토
2. 목표: 오픈소스 사용하여 저비용으로 한국어 대화 가능 한 AI 서비스 구축
3. 적용 범위: 운동 데이터, 영어 학습 등 라이트하게 접근 (ex. 임베디드 AI)

## 02. LLM 엔진 종류 및 특성

- Llama, Vicuna, Mistral → 한국어 사전 학습 X
- Gemma, Solar, koAlpaca → 한국어 사전 학습 O

	LLaMa 2	Vicuna 13B	Mistral 7B	Gemma	SOLAR	KoAlpaca
기업명	Meta	Stanford와 같은 기관의 연구팀	Mistral	Google	업스테이지	이준범 AI/NLP 연구원 (개인)
모델 버전	<ul style="list-style-type: none"><li>- 7B (70억)</li><li>- 13B (130억)</li><li>- 70B (700억)</li></ul>	<ul style="list-style-type: none"><li>- Llama2 미세 조정하여 훈련</li></ul>	<ul style="list-style-type: none"><li>- 7B (약 73억)</li></ul>	<ul style="list-style-type: none"><li>- 2B (20억)</li><li>- 7B (70억)</li></ul>	<ul style="list-style-type: none"><li>- 10.7B (107억)</li></ul>	
특장점	<ul style="list-style-type: none"><li>- 현 시점 오픈소스 중 최고 성능</li><li>- 대부분 오픈소스 Llama 기반</li></ul>	<ul style="list-style-type: none"><li>- ShareGPT의 70K 공개 데이터로 학습</li></ul>	<ul style="list-style-type: none"><li>- 32k 토큰 컨텍스트 관리 능력</li><li>- 코드 생성에 뛰어난 성능</li><li>- 빠른 추론 속도</li></ul>	<ul style="list-style-type: none"><li>- 온디바이스 AI용</li></ul>	<ul style="list-style-type: none"><li>- 국내 자체 개발 AI 엔진</li></ul>	<ul style="list-style-type: none"><li>- 네이버 지식인 이용 구성된 데이터 셋</li></ul>
문제점	<ul style="list-style-type: none"><li>- 한글 학습 부족 (학습 비율 0.06%)</li></ul>	<ul style="list-style-type: none"><li>- 연구 목적으로 주 사용</li><li>- 성능, 안전성 떨어짐</li></ul>	<ul style="list-style-type: none"><li>- 연구 목적으로 주 사용</li><li>- 성능, 안전성 떨어짐</li></ul>	<ul style="list-style-type: none"><li>- 한국어 성능 떨어짐</li></ul>	<ul style="list-style-type: none"><li>- 한국어 성능 떨어짐</li></ul>	<ul style="list-style-type: none"><li>- 2023년 8월 이후 학습 및 업데이트 중단</li></ul>
라이선스	MAU 7억 명 초과 시 별도 계약	Apache 2.0 라이선스로 제공	Apache 2.0 라이선스로 제공	Apache 2.0 라이선스로 제공	Apache 2.0 라이선스로 제공	Apache 2.0 라이선스로 제공
비고				<ul style="list-style-type: none"><li>- 발열, 성능, 속도 이슈로 현 시점 사용 불가능</li></ul>		

※ 조사관련 상세 내용 링크  
[https://docs.google.com/spreadsheets/d/1cclXlIEG019sRCAqT\\_YYTp7DBmbssGZKyvyAdj\\_VtI4/edit#gid=0](https://docs.google.com/spreadsheets/d/1cclXlIEG019sRCAqT_YYTp7DBmbssGZKyvyAdj_VtI4/edit#gid=0)

### 03. LLM 엔진 상세 비교

ChatGPT 3.5	LLaMa 2	Gemma	SOLAR
<b>특징:</b> <ul style="list-style-type: none"> <li>✓ 1,750억 파라미터 크기로 제공</li> <li>✓ 자연어 이해와 생성에 특화 (한국어 포함) 주된 사용처: 자연어 처리 관련 작업</li> <li>✓ 학습 및 운영에 비용 발생</li> <li>✓ 현 시점 가장 안정적 서비스 가능</li> </ul>	<b>특징:</b> <ul style="list-style-type: none"> <li>✓ 7B(70억), 13B(130억), 70B(700억) 파라미터 크기로 제공 매개변수 규모에 따라 모델 제공</li> <li>✓ 현존 오픈소스 중 최고 성능 오픈소스 대부분 Lama 기반</li> <li>✓ 한글 사전 교육 부족 (학습 비율 0.06%)</li> <li>✓ 연구자, 상업 단체 모두에게 라이선스 부여 MAU 7억 명을 초과하는 경우 meta와 별도 라이선스 계약 필요</li> <li>✓ Meta 배포 Llama 모델은 리눅스에서만 실행 가능 고성능 GPU가 아닌 경우 7B모델 사용 권장</li> </ul>	<b>특징:</b> <ul style="list-style-type: none"> <li>✓ 2B(20억), 7B(70억) 파라미터 크기로 제공</li> <li>✓ GPT 제외 한국어 대화 가능 (성능 떨어짐)</li> <li>✓ Gemini (제미나이) 경량화 버전 (sLLm)</li> <li>✓ 온디바이스 AI 앱 또는 소프트웨어 개발에 적용 가능</li> <li>✓ 멀티모달 모델 아님 (텍스트, 자연어 이해 중점)</li> </ul> <p>※ 멀티모달 AI 여러 가지 유형의 데이터 (텍스트, 이미지, 음성, 비디오 등) 또는 정보를 함께 활용하여 인공 지능 시스템을 구축하는 접근 방식</p>	<b>특징:</b> <ul style="list-style-type: none"> <li>✓ 10.7B(107억) 파라미터 크기로 제공</li> <li>✓ 한국어 전용으로 개발된 모델</li> <li>✓ 카카오톡 '아속업(AskUp)' 챗봇 서비스 중 ChatGPT와 OCR 기술 결합한 챗봇 서비스</li> <li>✓ GPT-3.5 대비 2.5배 이상 빠른 속도</li> <li>✓ Apache 2.0 라이선스로 소스 무료 제공 단, API 사용시 비용 발생</li> </ul>

## 04. 테스트 시연 장면

### LLaMa 2

Please tell me your travel plans for Jeju Island in May based on a family of four.



codellama 70b

🎉 Congratulations on planning a family trip to Jeju Island! 🎉

Below is a suggested itinerary for a family of four visiting Jeju Island in May. Please note that this is a rough plan, and the specifics may vary depending on your interests, preferred activities, and travel style. 😊

Assumptions:

- 4 days in May (May 1-4, 2022)
- Flight to Jeju Island from Europe/North America (assuming a 16-hour flight)
- Accommodation: 1-2 nights at a resort on Hallasan Island (the highest peak in Jeju Island), 1 night in Seogwipo (

0.37 rtt 29 tps 169 toks 5.81 s

Enter something here

### Gemma

Please tell me your travel plans for Jeju Island in May based on a family of four.



gemma-7b-it

**Travel Plans for Jeju Island, South Korea in May for a Family of Four**

**Accommodation:**

- The Metropolitan Hotel Jeju City Center
- The Westin Jeju Resort & Spa
- Conrad Jeju Resort and Spa

**Day 1:**

- Arrive at Jeju International Airport and check into your hotel.
- Visit the iconic Seongeup Folk Village.
- Explore the Ollades (stone houses) of Haengjusan.
- Hike

0.33 rtt 47 tps 101 toks 2.15 s

Enter something here

수화물) 수박 : 1% (섬유)

총 탄수화물 : 61% 총 단백질 : 32% 총 지방 : 7%

일상에서 스트레스를 효과적으로 관리하기 위해 추천하는 방법 알려줘.



gemma-7b-it

**식사 구성 비율 균형**

식사 구성 비율은 건강한 식사와 관련된 중요한 요소입니다. 건강한 식사를 위해서는 탄수화물, 단백질, 지방의 적절한 비율을 섭취하는 것이 좋습니다.

일상에서 탄수화물, 단백질, 지방의 적절한 비율을 섭취하기 위한 방법:

**1. 탄수화물 감소:**

- 과실, 과일, 현곡물, 통곡물, 식곡물과 같은 올바른 식사와 비가공 식품을 포함한 탄수화물

0.24 rtt 53 tps 150 toks 2.82 s

Enter something here

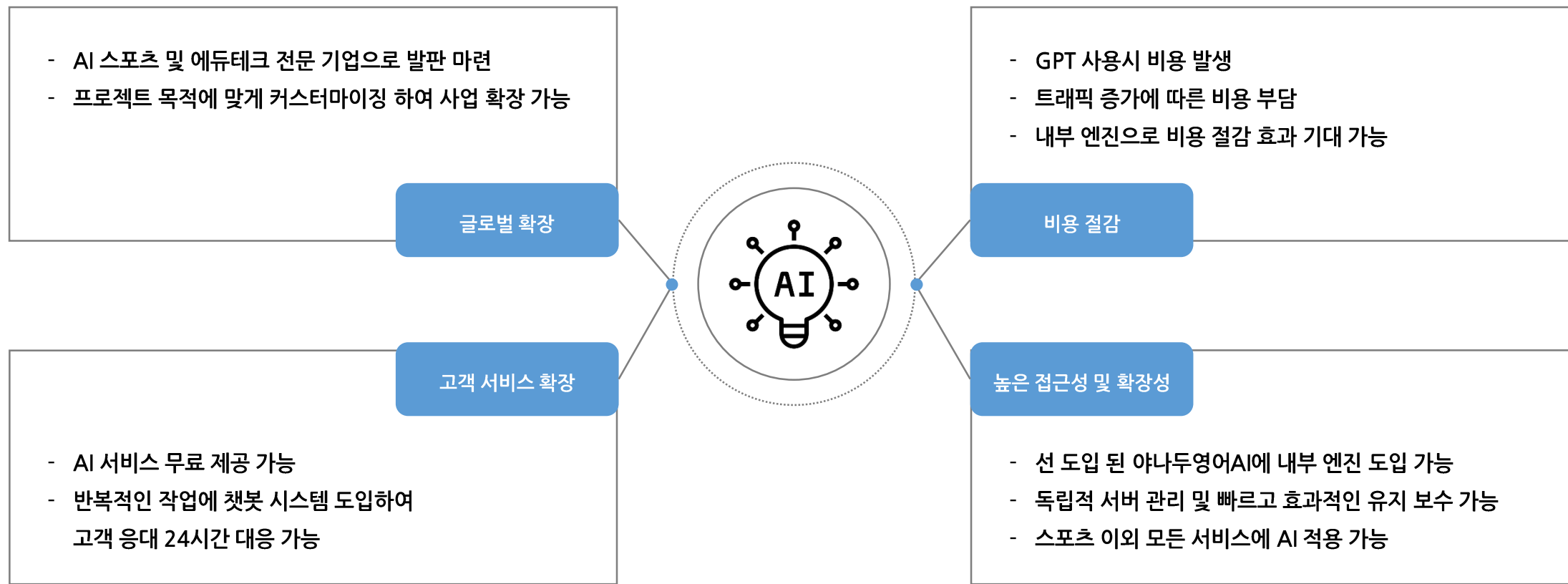
### 테스트 내용

1. Compose your diet with a 1:2:7 intake ratio of carbohydrates, protein, and fat.
2. Please tell me how to recommend to manage stress effectively in your daily life.
3. Please tell me your travel plans for Jeju Island in May based on a family of four.
4. Please recommend me a new hobby.

1. 탄수화물, 단백질, 지방을 섭취 비율 1: 2: 7로 식단 구성해줘.
2. 일상에서 스트레스를 효과적으로 관리하기 위해 추천하는 방법 알려줘.
3. 4인 가족 기준으로 5월 제주도 여행 계획 알려줘.
4. 새로운 취미를 추천해줘.

## 05. 기대 효과

### 효과





## 06. 정리

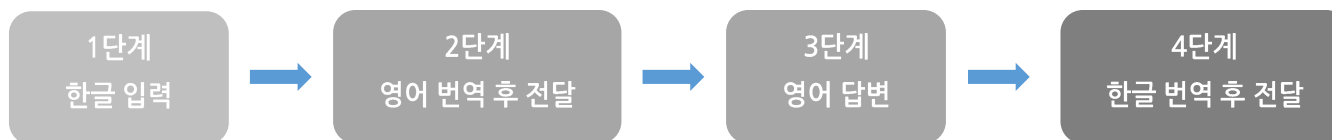
### 현재 상황

1. **현 시점 공개된 오픈소스들 한글 사전 교육 부족**
  - 야핏 무브에 적용 하기 위해서는 한글 대화 가능해야 함 (영어 학습 이외 모든 서비스 필수 사항)
2. **파인튜닝 시 필수 요소**
  - 적절한 하이퍼파라미터 선택 후 최적화
  - 강력한 하드웨어 리소스 필요 (대규모 데이터셋 사용 시 고성능의 GPU, TPU 필요)
  - AI 개발자가 직접 데이터를 확보 또는 데이터셋 찾아야 함

### 추후 계획

1. **LLaMa 2로 내부 테스트 진행 (기본 대화형 챗봇 형태)**
  - 서버 비용: 내부 장비 사용
  - 번역기(Google) 사용: 월 500만원(GPT 사용시 6천만원) 정도 비용 발생
  - 최소한의 데이터 사용(API)을 기준으로 테스트 작업 진행
2. **추후 공개 오픈소스 실시간 모니터링 및 적용**
  - 참고 자료: <https://www.aitimes.kr/news/articleView.html?idxno=30755>

▶ 적용 프로세스



※ ChatGpt 가격 정책 변동사항 상시 체크

※ 온디바이스 사용 가능한 Gemma 업데이트 현황 상시 체크

## 07. 예상 비용

### 번역 - 전제 조건:

- ✓ 하루 문장 수: 36만 문장 (= 26만 문장 + 10만 문장)
- ✓ 하루 13만회 방문 (왕복 2문장, 5만 DAU)
  - 하루 DAU 5만명 / 10% 사람들 평균 20문장 (10개 질문) 주고 받는다 로 가정
- ✓ 한달 문장 수: 30일로 단순 계산

	1일 질문자 수	1일 질문 수	대화글자 수	기간 (한달)	총 글자 수
질문	500	10	15	30	2,250,000
답변	500	10	100	30	15,000,000
한달 글자 수					17,250,000

	Google (USD)	DeepL (USD)
100만자 당	\$20	\$25
총 비용	\$3,440	\$4,317.99

※ Google 매월 첫 50만 자 무료

※ 야핏무브 기준

- google (약 462만원)

- deepl (약 580만원)

### 서버:

- ✓ 내부 장비로 커버 가능 (AI 개발자 사용하던 장비 Jeff 옆자리에 셋팅 완료)

CPU	라이젠 5800x (8코어 16스레드)
Ram	128G
Apu	GeForce RTX 3090
Hdd	1TB

※ AWS EC2 사용시 t3.xlarge 시간 당 \$0.208 발생 (1일 \$4.992)

# Thank you.