How to Customize ChatGPT/LLM

나의 지식을 불어 넣는 방법 4가지

한대희 @ 뉴럴웍스랩

daehee@neuralworks.io

010-2101-0255

http://neuralworks.io/

NeuralWorks

상식 vs 특화 지식

- ChatGPT/LLM :
  - 인터넷에 공개된 방대한 자료를 학습함
  - 언어 표현력, 어휘력, 기억력, 방대한 상식 자료를 보유함
  - **→ 모든 분야에 상식이 많고 어휘력이 뛰어난 사람**
  - **→ 상식(Common Sense) 임베딩 되어 있음**

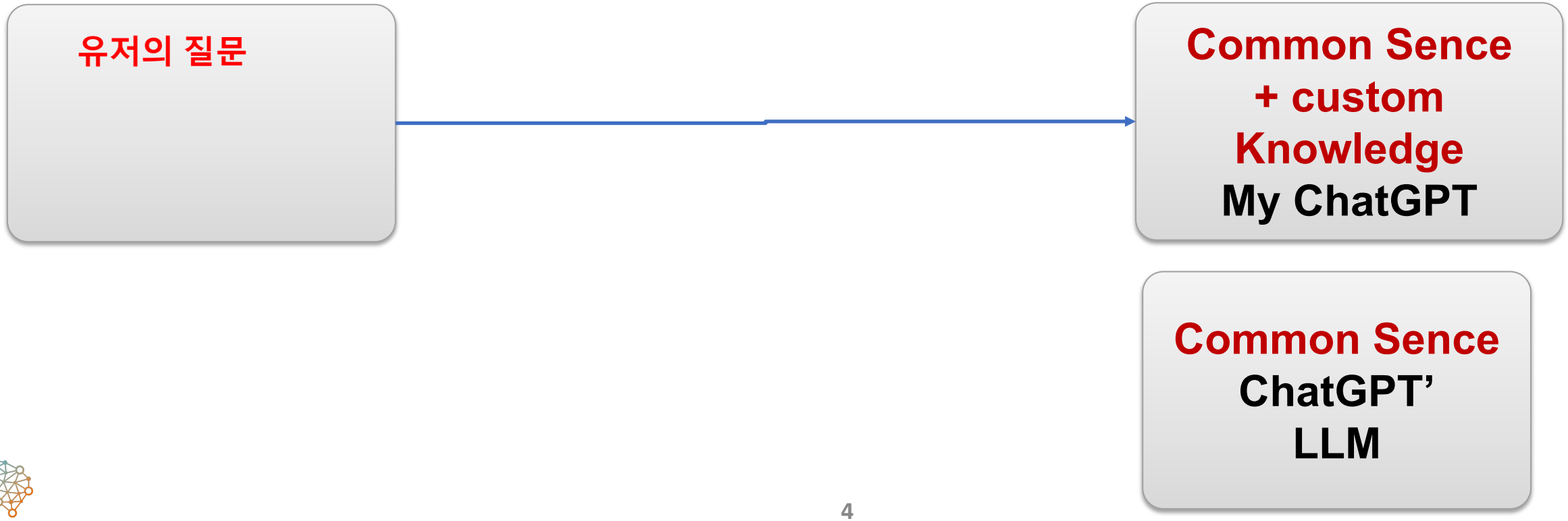
- **인터넷에 공개된 데이터가 아닌, 특정 기업/분야에 특화된 전문지식을 학습 및 구축할 수 있는가?**
- **예) 사내 매뉴얼, 보험 약관, …**
- **→ 지식 (knowledge) 구현(임베딩?)**

1) Prompt 방식

- 유저의 질문 + 프람프트 전달
- ChatGPT 모델/API 그대로 사용
- 단점: prompt의 길이 제한

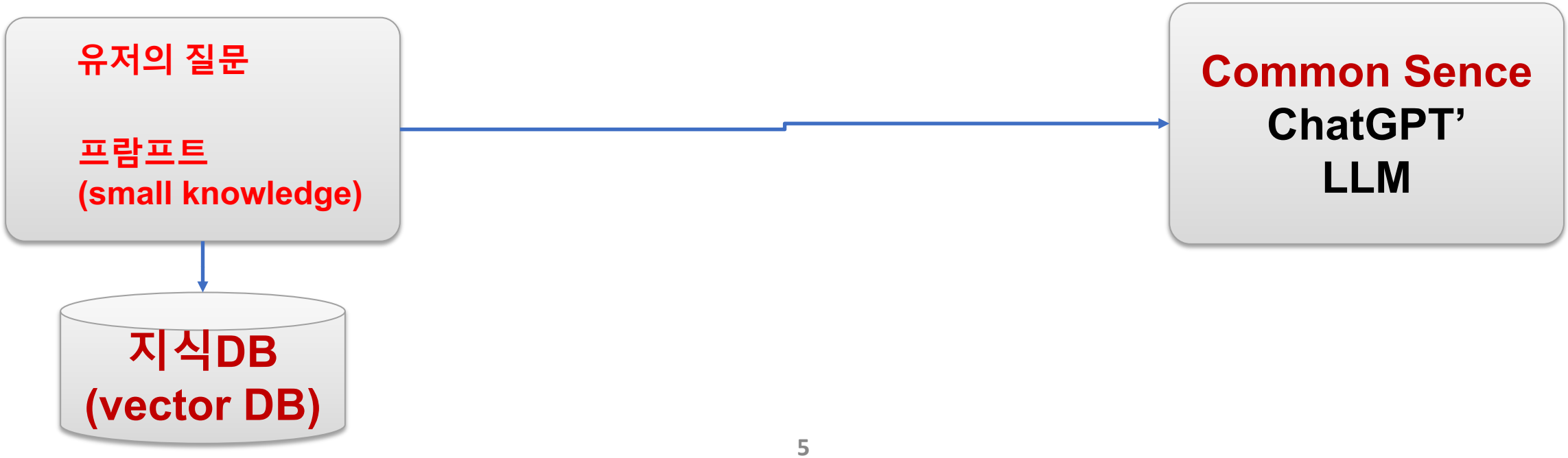
2) Fine-tuning 방식

- 유저의 질문만 보냄
- 내가 튜닝한 ChatGPT 모델 사용. API 그대로 사용
- 단점: 조회비용이 높음



3) 지식DB 조회 + ChatGPT 방식

- 유저의 질문 + 프람프트 전달
- 프람프트 내용이 고정되지 않고, 유저의 질문에 따라 달라짐(지식DB에서 조회) → RAG (Retrieval Augmented Generation)
- ChatGPT 모델/API 그대로 사용
- 단점: 지식DB 조회결과에 의존. 길이제약

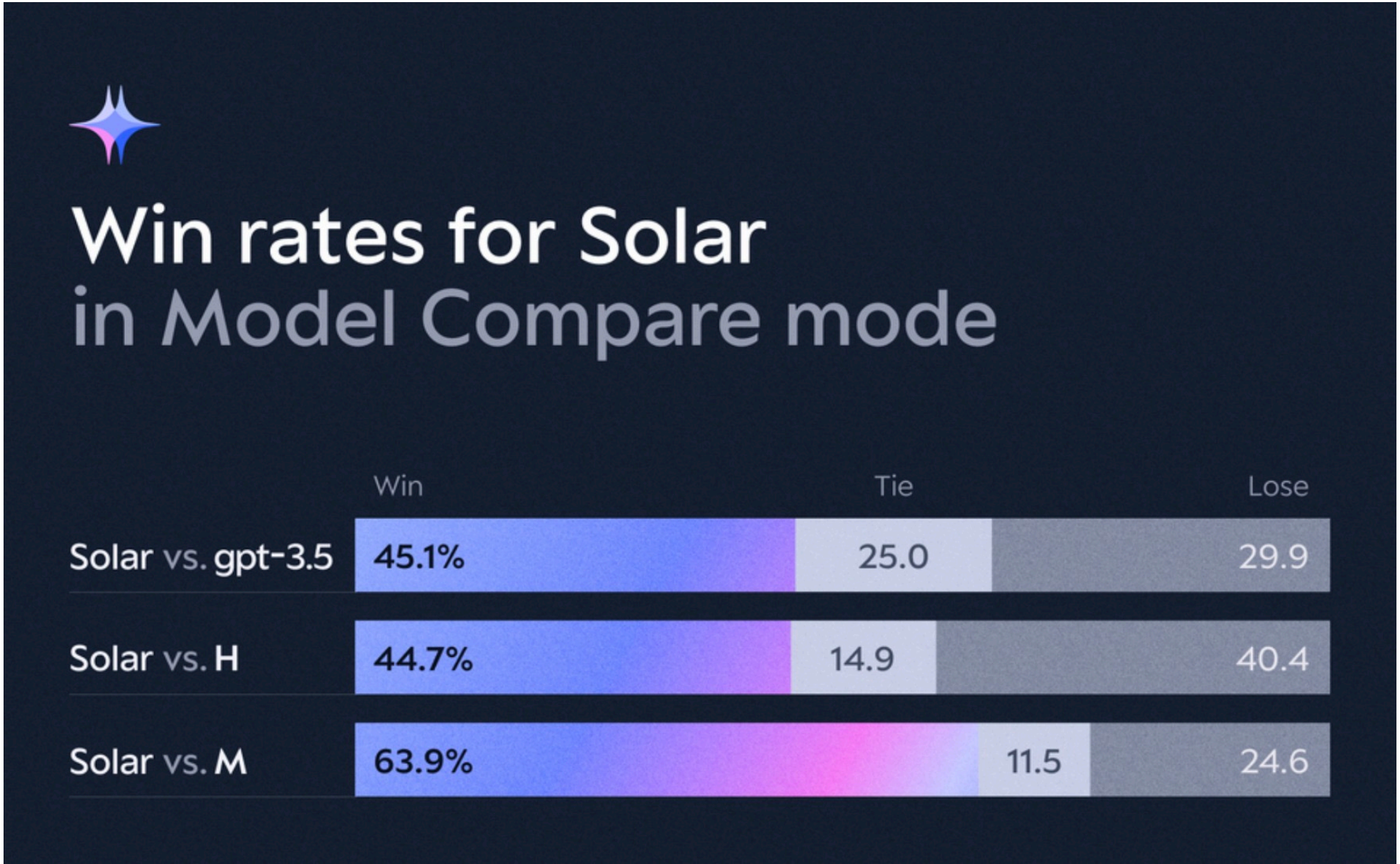


4) 나만의 LLM 구축

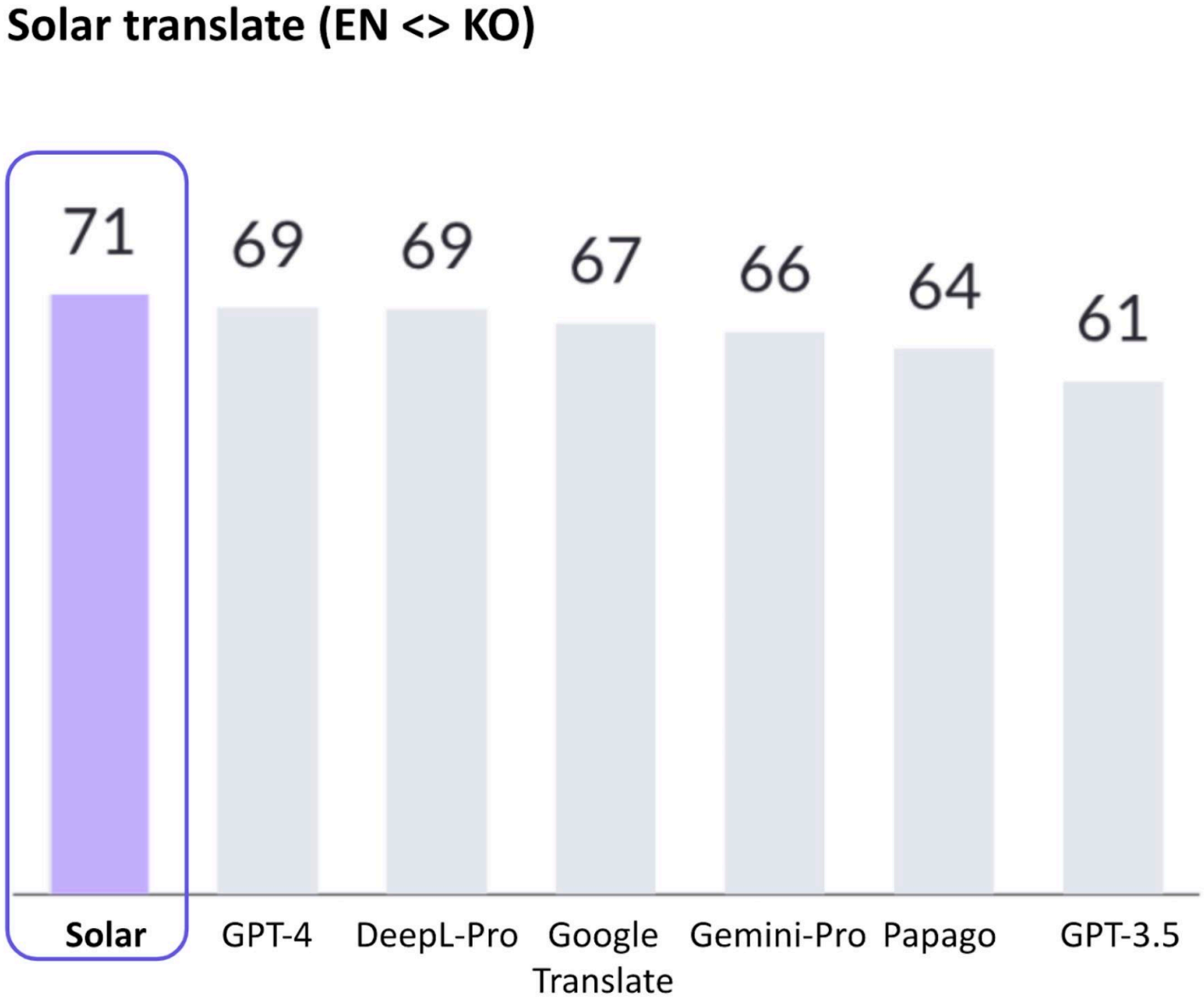
- 지식을 LLM에 직접 학습. 나의 소스, 인프라 사용
- 단점: 인프라 비용 (학습, 운영)

Huggingface, Leaderboard

Upstage의 LLM , Solar



Upstage의 LLM , Solar



Benchmark for machine translation from Meta https://ai.meta.com/tools/flores/

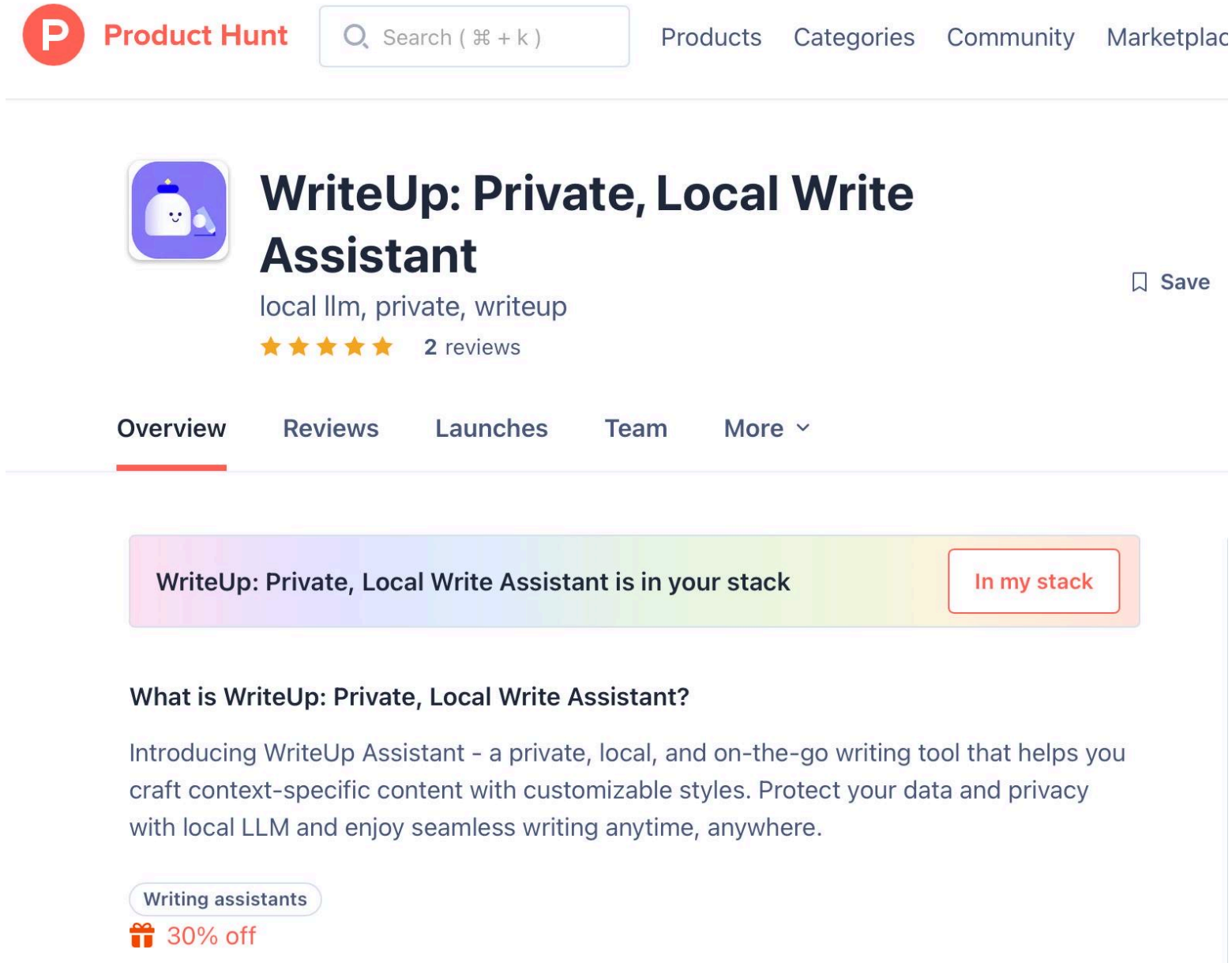
Upstage의 LLM , Solar

- Conversational Multi-Doc QA

Table 2: Performance of different LLMs on the eval dataset.

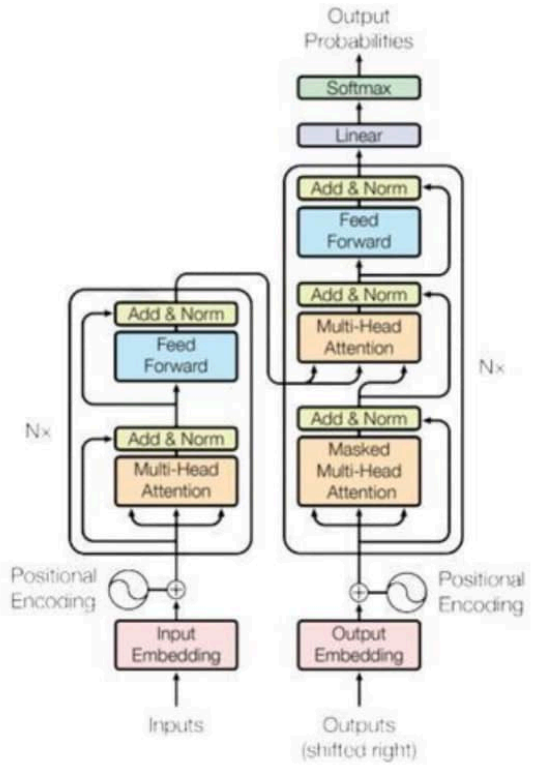
LLM	C-ROUGE-L	KR
ChatGPT (zero-shot) [12]	0.4815	0.5387
GPT-4 (zero-shot) [13]	0.5069	0.5537
Yi-6B [14]	0.5286	0.6464
ChatGLM3-6B [16]	0.5740	0.6068
ChatGLM3-6B-Base [16]	0.5782	0.6166
DeepSeek 7B Base [1]	0.5783	0.6184
Llama 2 Chat 13B [10]	0.5821	0.6266
Yi-6B-Chat [14]	0.5833	0.6365
Llama 2 13B [10]	0.5845	0.6359
Mistral 7B [3]	0.6031	0.6489
Mistral 7B-Instruct [3]	0.6048	0.6558
SOLAR 10.7B [4]	0.6099	0.6627
SOLAR 10.7B-Instruct [4]	**0.6104**	**0.6691**

sLLM의 활용 예시 1)


Product Hunt
Search (⌘ + k)
Products
Categories
Community
Marketplac
WriteUp: Private, Local Write Assistant
local llm, private, writeup
2 reviews
Save
Overview
Reviews
Launches
Team
More
WriteUp: Private, Local Write Assistant is in your stack
In my stack
What is WriteUp: Private, Local Write Assistant?
Introducing WriteUp Assistant - a private, local, and on-the-go writing tool that helps you craft context-specific content with customizable styles. Protect your data and privacy with local LLM and enjoy seamless writing anytime, anywhere.
Writing assistants
30% off

허깅페이스

AI Engineers

The Interview	The Job
	`import transformers`

LLM, AI 반도체

마크 저커버그 28일 일정

12:23

- 서울 여의도 LG트윈타워 도착
- 조주완 LG전자 사장 등 경영진과 오찬 · 회의 진행
- 메타 퀘스트3 등 XR 사업전략 · 차세대 기기 개발 논의

15:00

- 메타코리아 도착, 업스테이지 등 국내 XR 및 인공지능(AI) 업체 5곳과 간담회 진행
- 김성훈 업스테이지 대표 및 박은정 업스테이지 최고과학책임자(CSO) 등 참석
- 메타 LLM **'라마2'** 국내 성과 및 최신기술 논의

18:15

- 서울 이태원동 삼성 승지원 도착 이재용 삼성전자 회장과 만찬 · 회의 진행
- 메타 LLM **'라마3'** 구동 자체 AI 반도체 개발 · 생산 논의

온디바이스(on-device) AI, small LLM

- 업스테이지(대표 김성훈)가 LG전자(대표 조주완)와 인공지능(AI) 사업 MOU를 체결, '온디바이스 AI' 구축에 나선다고 6일 밝혔다.
- 온디바이스 AI는 스마트폰, 노트북, 태블릿 등 전자 단말기 내부에서 정보를 처리하기 때문에 클라우드 기반 AI보다 빠른 작업 속도, 낮은 전력 소모의 특징을 보인다. 아울러 개인정보 유출 위험 없이 보안 문제를 해결할 수 있으며, 인터넷 연결이 불안정하거나 끊어져도 구동이 가능하다.
- 업스테이지는 자체 개발한 고성능 소형언어모델(sLM) '솔라(SOLAR)'를 활용할 예정이다. 현재 10.7B(107억개) 매개변수보다 작은 온디바이스 AI용 sLM을 개발, LG 노트북 '그램'에 탑재할 계획이다. 또 가전용 온디바이스 AI 개발을 검토 중이다.

출처 : AI타임스(https://www.aitimes.com)