PlanTy: LLM과 IoT를 활용한 개인 맞춤형 식물 관리 솔루션

Technology Presentation

과 목 명: 실무중심산학협력프로젝트1 (1분반)
발 표 일: 2025.05.21 (수)
팀　　명: BIoTy (바이오티)
팀　　원: 구선주(32220207), 김민지(32200588),
민유진(32221598), 최예림(32224684)

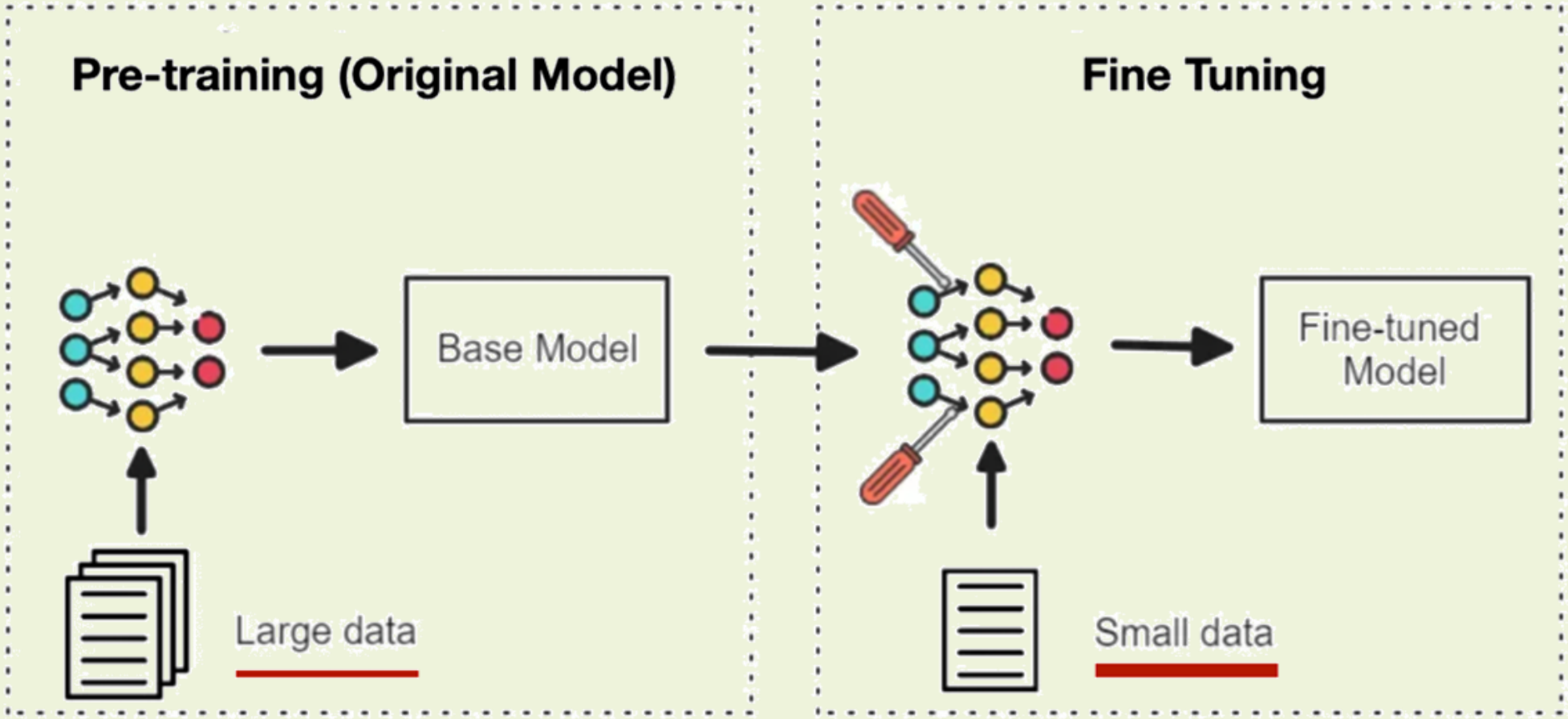
Table of contents

01

Fine-Tuning & RAG

| Fine-Tuning | PeFT | RAG | Reranker

Fine-Tuning


Pre-training (Original Model)
Base Model
Large data
Fine Tuning
Fine-tuned Model
Small data

SFT

- CoT 경로를 학습하는데 초점 (질문 당 하나의 추론 경로)
- 학습 데이터에만 특화된 규칙 형성 가능성 높음 → 일반화 능력이 부족

PeFT

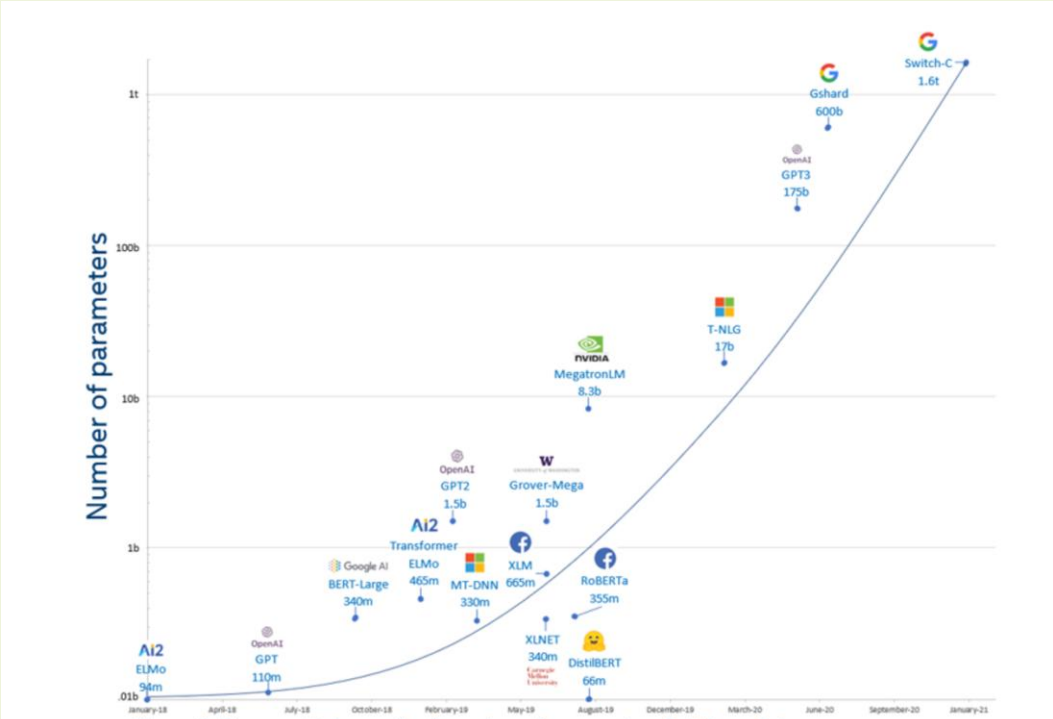
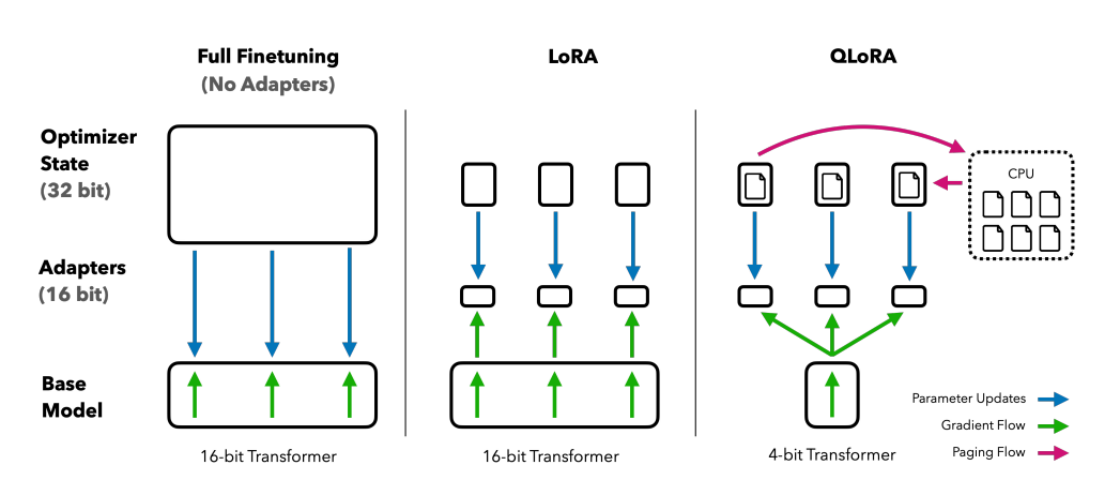


Figure 1: Exponential growth of number of parameters in DL models

- Parameter-Efficient Fine-Tuning
- 대형 언어 모델의 일부 파라미터만 조정하여 Fine-tuning의 효율성 극대화

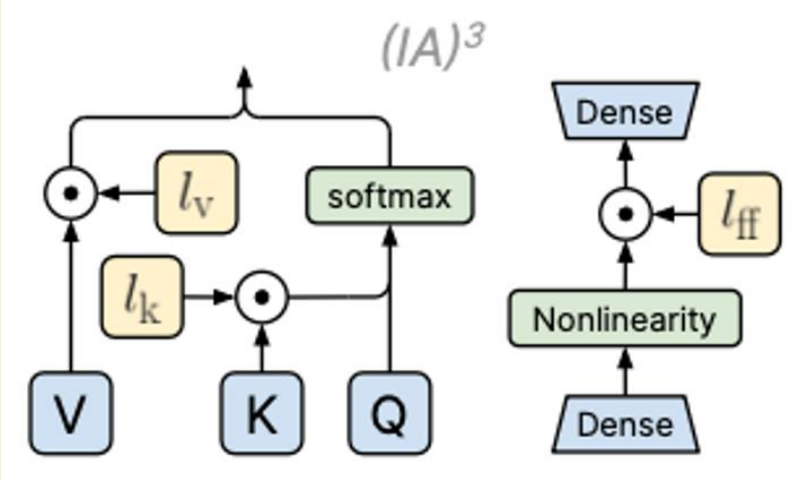
PeFT

QLoRA



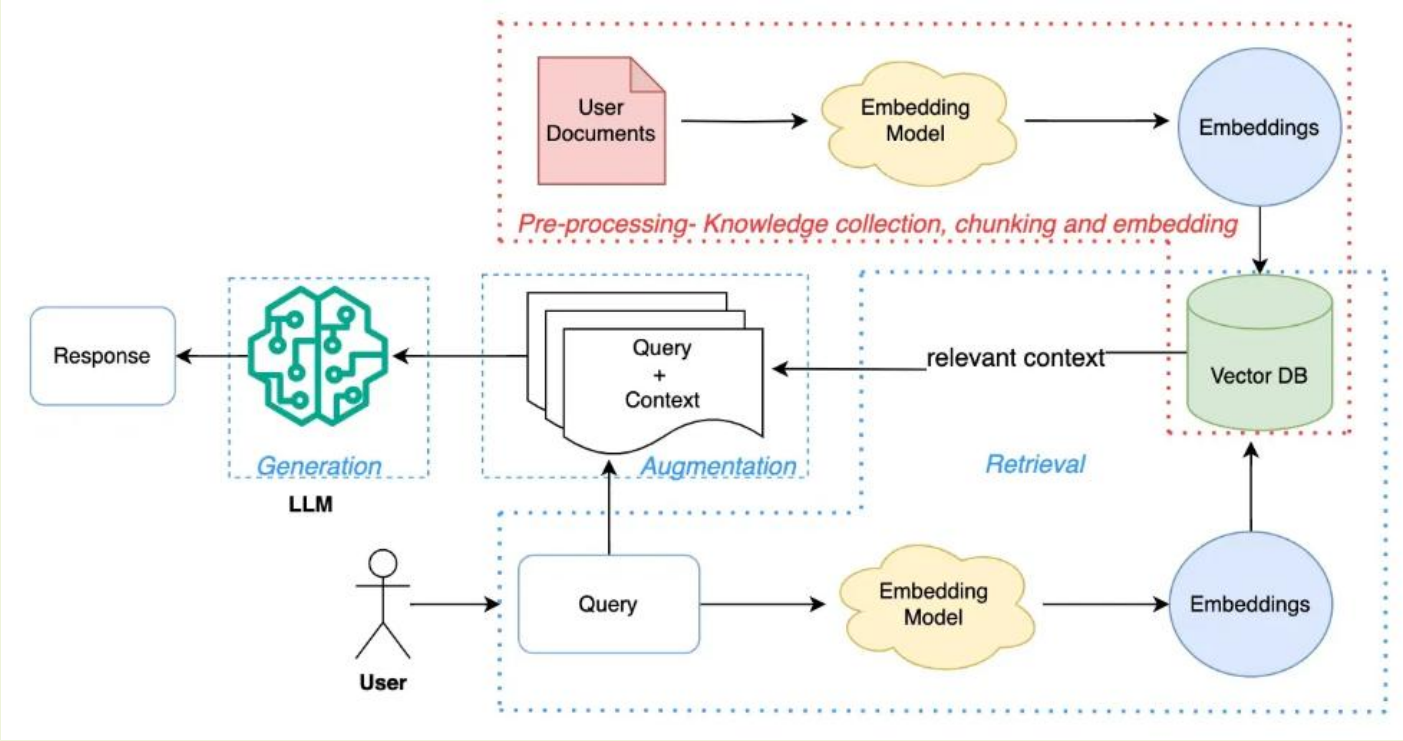
- 4bit로 quantization된 모델에 LoRA를 사용
- 일반적으로 성능이 더 강력
- 메모리 효율성이 높음

IA3



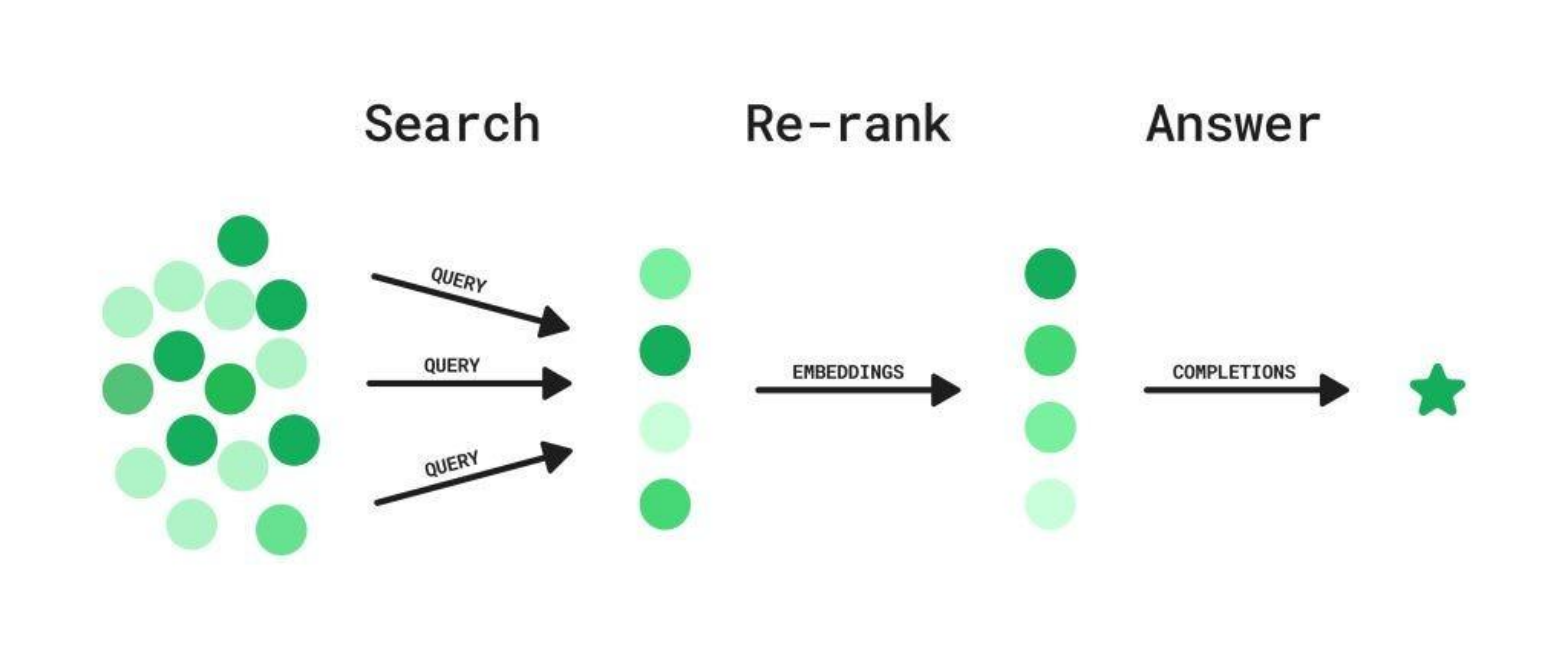
- g vector를 넣어 파라미터를 조금만 학습
- 가볍고 빠르지만 성능은 약간 떨어짐
- QLoRA보다 더 적은 파라미터 학습

RAG



- 검색 증강 생성 Retrieval-Augmented Generation
- LLM에 외부 지식 베이스를 결합하여 정확하고 신뢰성 있는 응답 제공

Reranker



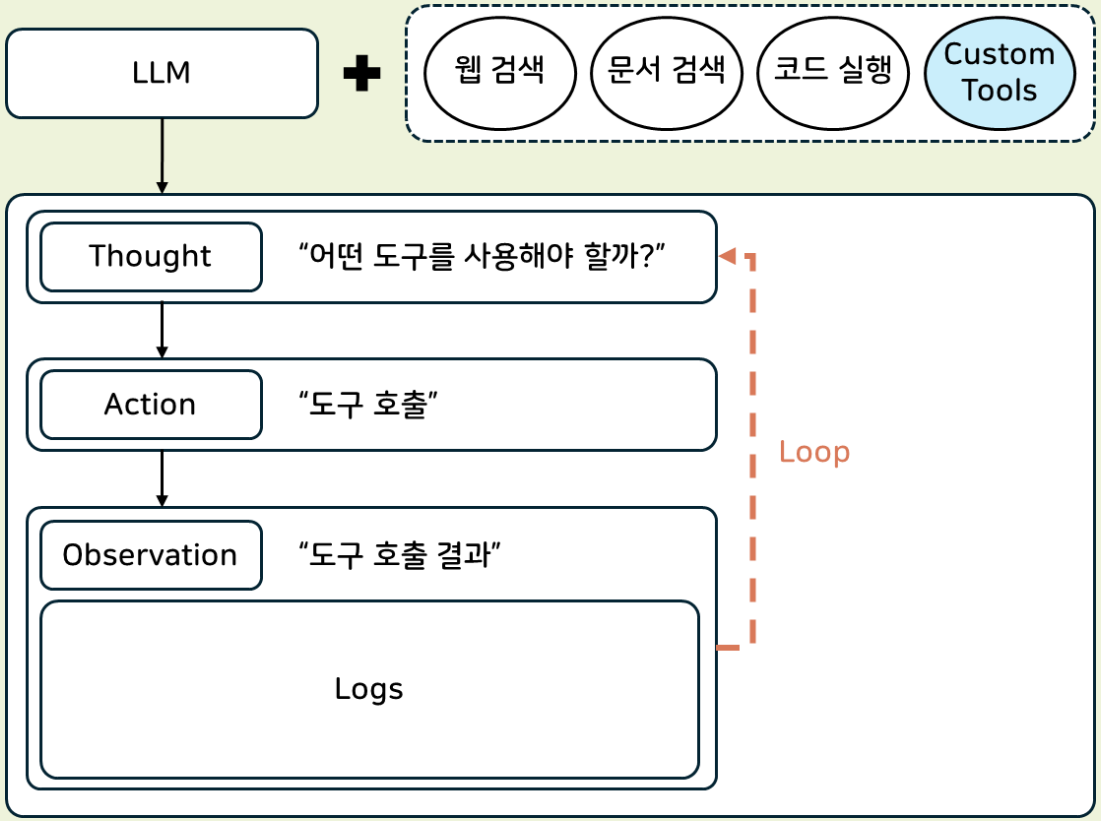
- 초기 검색을 통해 추출된 후보 문서 집합을 대상으로 순위 재정렬
- **초기 검색 단계**: 임베딩 벡터 간 유사도 기준, 상위 N개의 문서를 빠르게 추출
- **재정렬 단계**: 후보 문서들과 질의를 함께 모델에 입력하여 의미적 일치도 확인, 순위 재구성

02

Multi-Agent Introduction

| Multi-Agent | AutoGen | CrewAI | LangGraph | Framework 비교 |

Multi-Agent



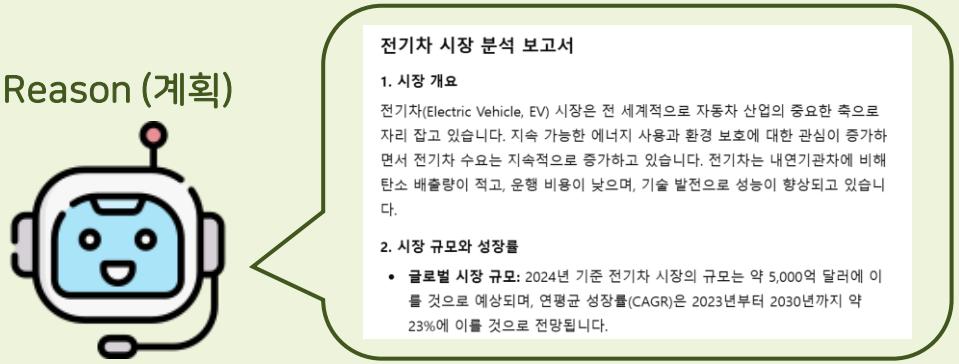
Agent

- **Agent**: LLM이 알아서 생각하고 행동하는 시스템
- **ReAct**: LLM은 ReAct 과정으로 더 나은 답변 생성
- **LLM 한계**: 최신 정보 부재, 내부 데이터 접근 불가능
- **Tool**: 외부 Tool을 활용하여 LLM 능력 확장

Multi-Agent

단일 에이전트

Reason (계획)



전기차 시장 분석 보고서

1. 시장 개요

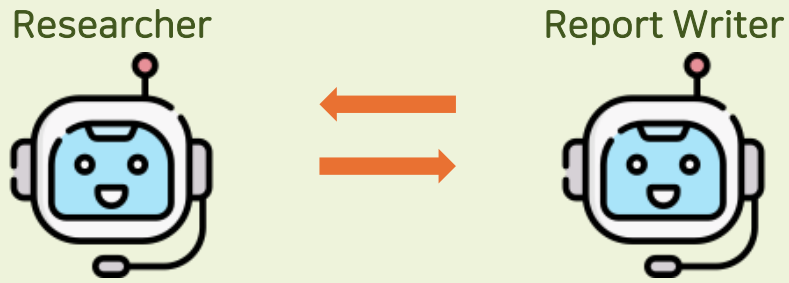
전기차(Electric Vehicle, EV) 시장은 전 세계적으로 자동차 산업의 중요한 축으로 자리 잡고 있습니다. 지속 가능한 에너지 사용과 환경 보호에 대한 관심이 증가하면서 전기차 수요는 지속적으로 증가하고 있습니다. 전기차는 내연기관차에 비해 탄소 배출량이 적고, 운행 비용이 낮으며, 기술 발전으로 성능이 향상되고 있습니다.

2. 시장 규모와 성장률

- **글로벌 시장 규모:** 2024년 기준 전기차 시장의 규모는 약 5,000억 달러에 이를 것으로 예상되며, 연평균 성장률(CAGR)은 2023년부터 2030년까지 약 23%에 이를 것으로 전망됩니다.

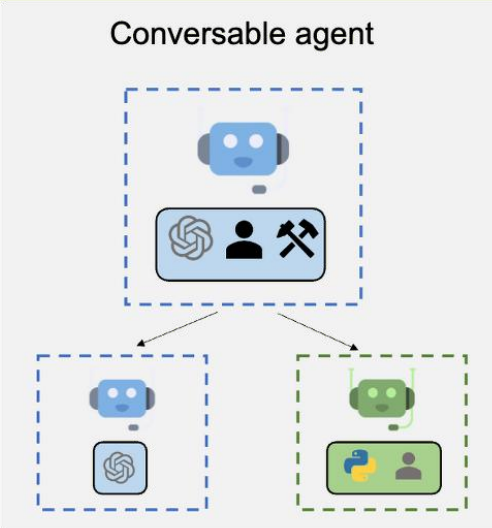
- 실무에서는 많은 데이터로 추론 과정을 통해 보고서 작성
- 단건으로 요청 받아 보고서를 작성하는 것에 한계 존재

멀티 에이전트

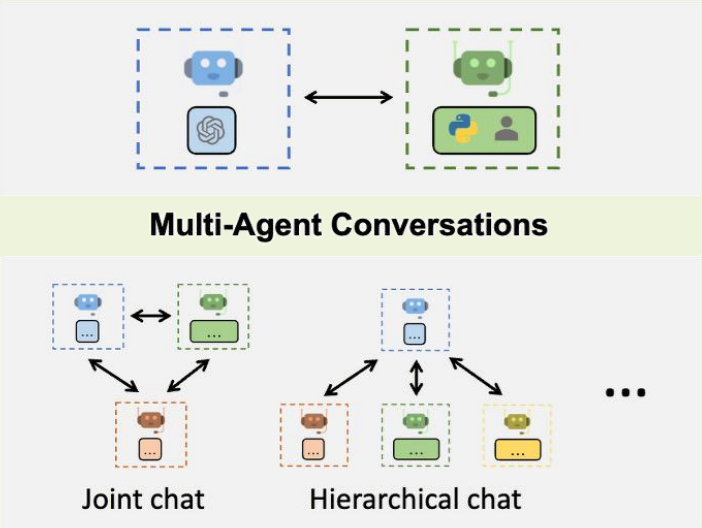


- 서로 다른 에이전트가 협력하거나 위계질서를 가지고 소통
- 역할에 따라 단일 에이전트가 결과를 출력

AutoGen



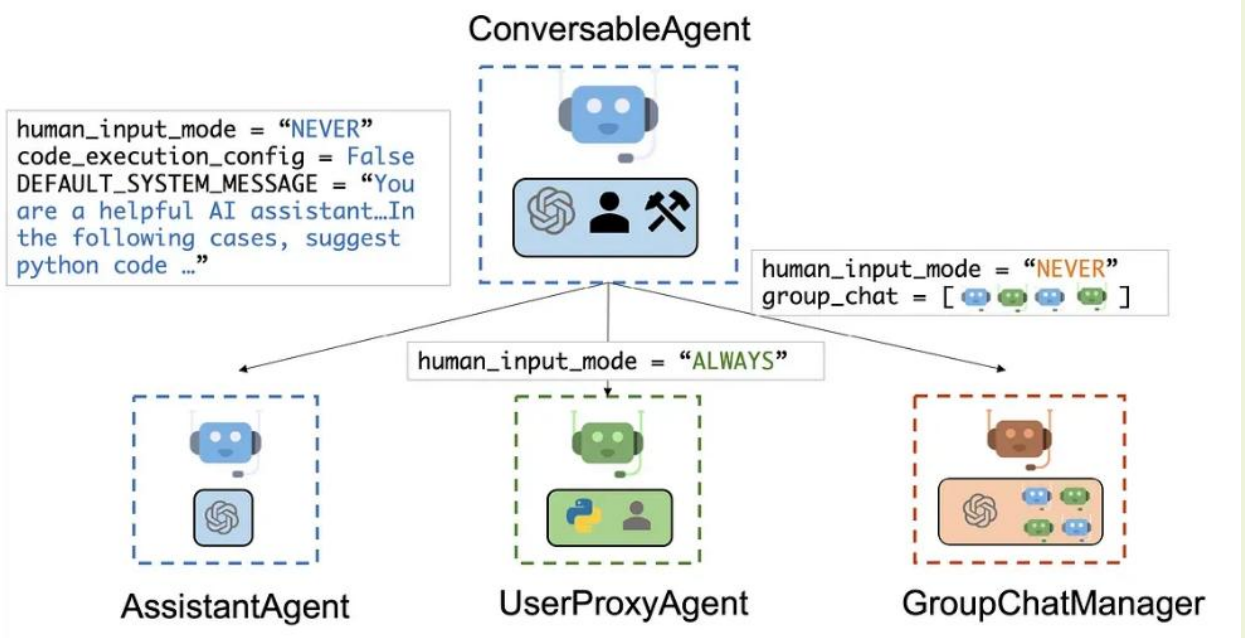
Agent Customization



Flexible Conversation Patterns

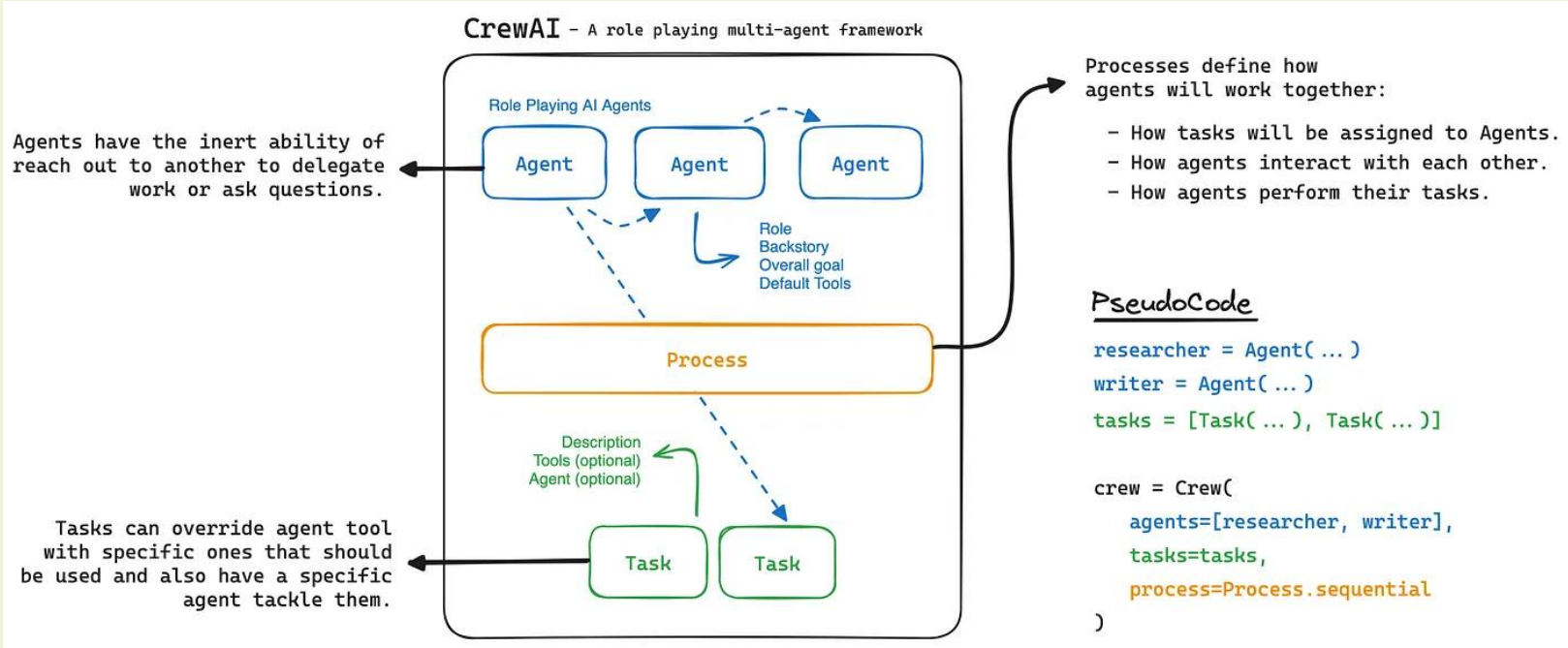
- Microsoft와 Penn State 대학 및 Washington 대학 협업
- 다중 에이전트 대화를 통해 LLM 어플리케이션 손쉽게 구축
- 각 Agent 들은 대화 및 작업 수행을 자동화 가능

AutoGen



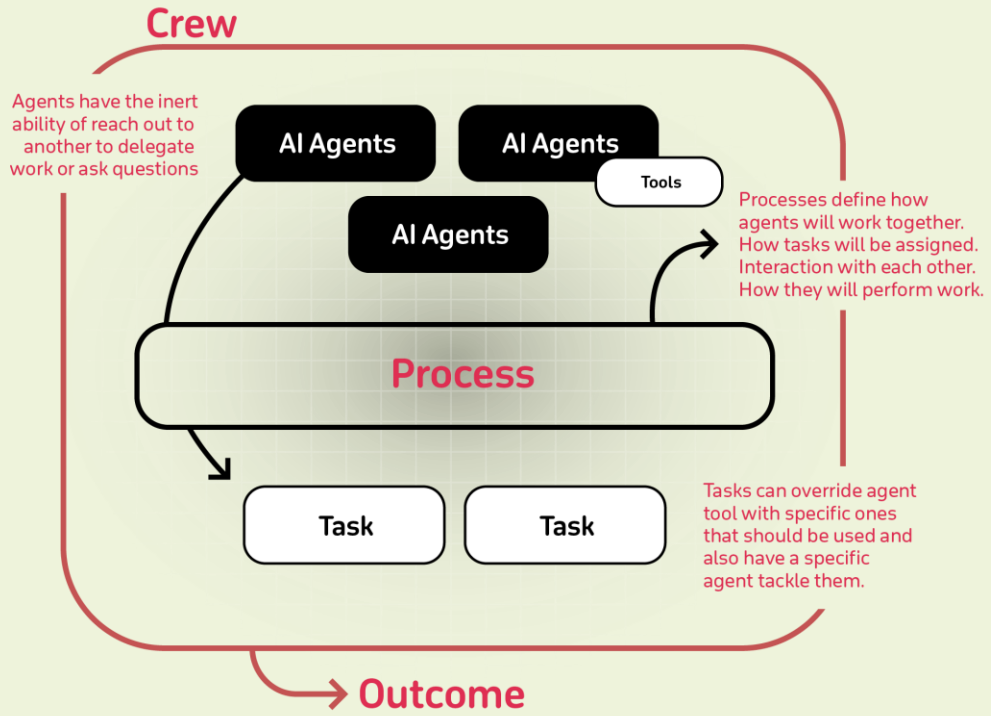
- **ConversableAgent**: 기본적인 AI 상호작용 처리 추상화 에이전트
- **AssistantAgent**: AI 모델 기반 가상의 비서 역할 수행
- **UserProxyAgent**: 사용자의 입력 완전 자동화, 특정 조건에서만 입력 요구
- **GroupChatManager**: 다자간 대화 가능, 특정 에이전트만 필터링 가능

CrewAI



- 전문화된 개별 AI 에이전트들이 각자 역할을 수행하면 서로 소통하여 협업
- Agent, Tools, Task, Crew/Process로 구성

CrewAI



Agent

- 작업 수행 주체
- 에이전트 팀 관리, 워크플로우 감독, 협업보장, 결과제공

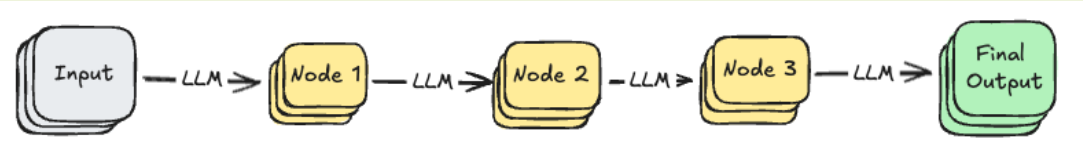
Task

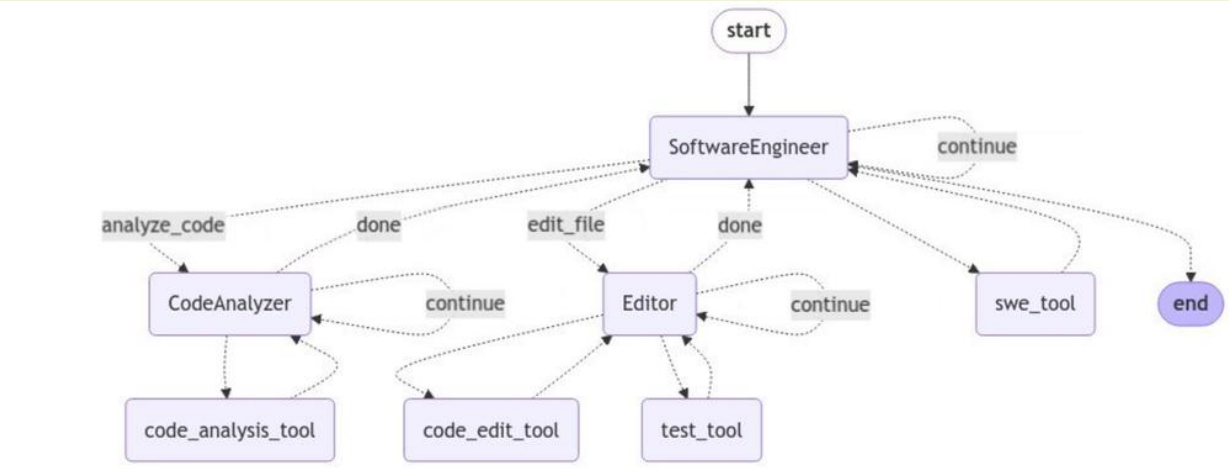
- 에이전트가 수행해야 하는 과제나 작업
- 특정 역할, 지정된 tool 사용, 작업 위임 기능

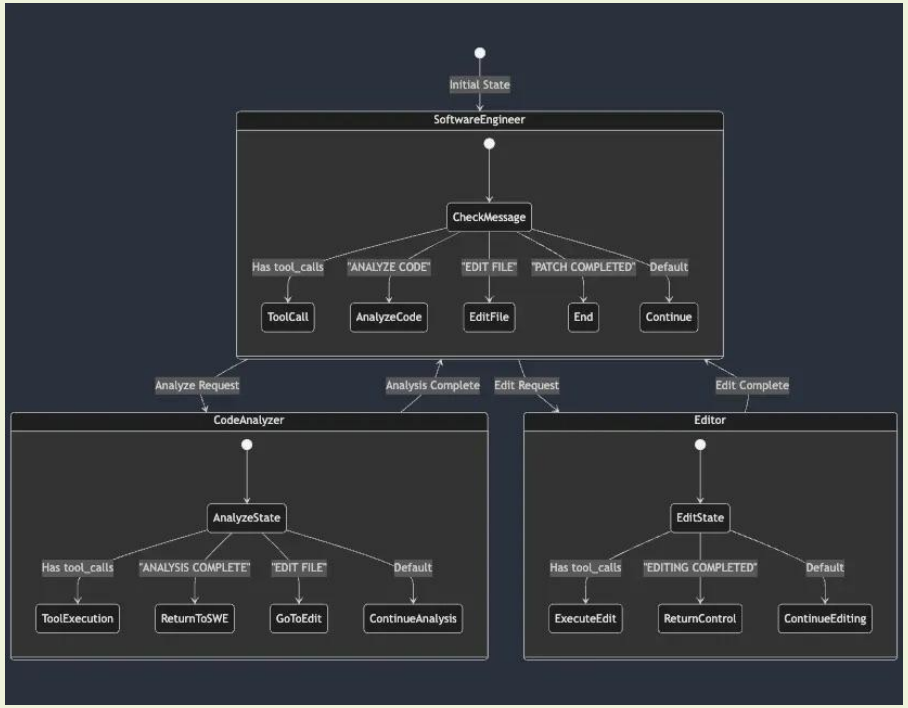
Crew

- 태스크 효율적 수행 흐름 조정
- 명확한 목표, 특정 도구 사용, 더 큰 프로세스로 전환

LangGraph

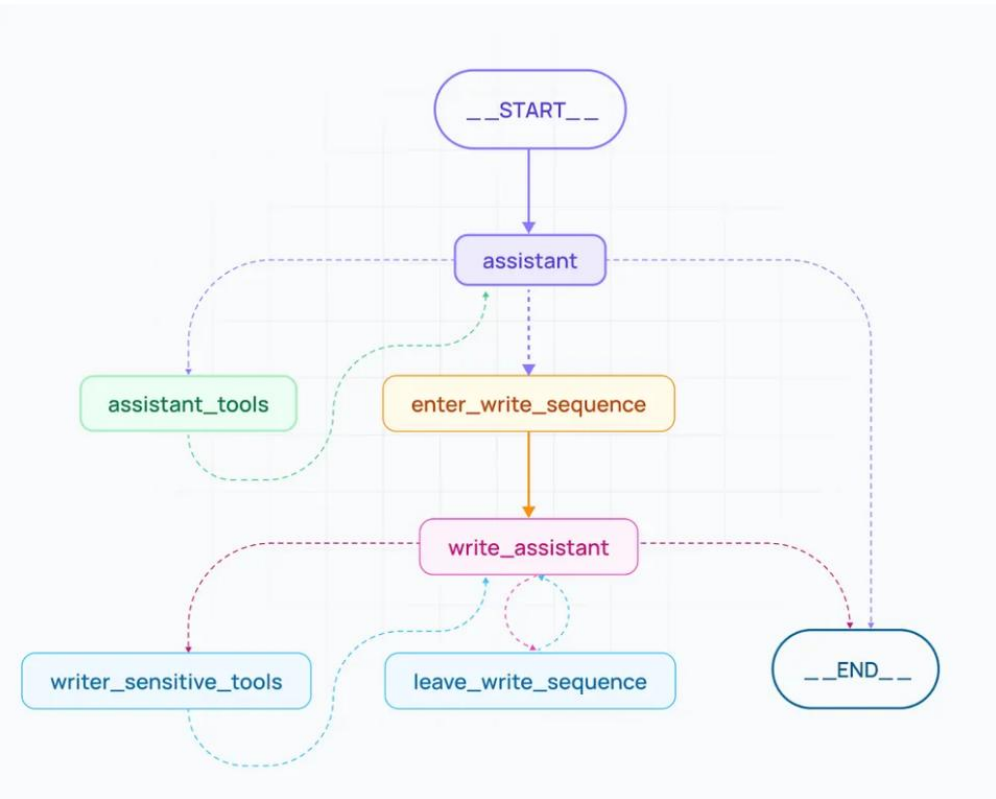






- 그래프 구조를 사용하여 LangChain의 선형적인 체인 구조의 한계 극복
- 특정 노드만 재실행하거나 대체 경로로 전환하는 등의 유연한 대응 가능

LangGraph



State

- 그래프의 전체 데이터 흐름을 관리
- 에이전트 간에 주고받을 정보 정의

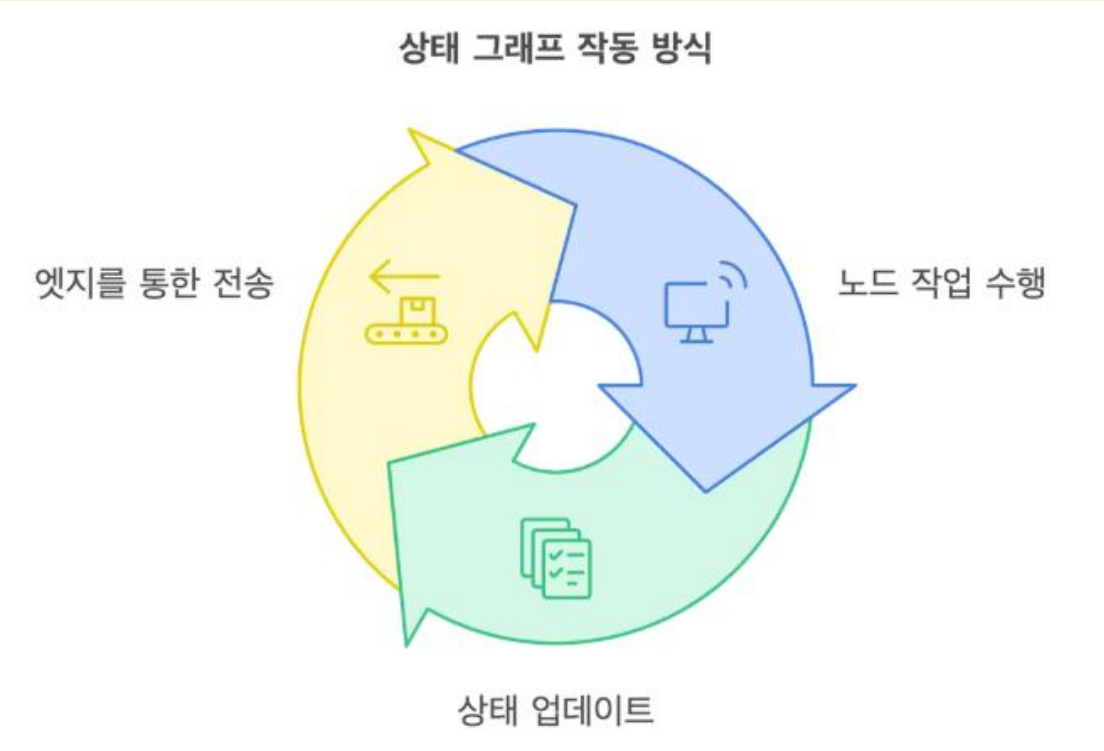
Node

- 작업을 수행하는 단위
- 각 노드는 특정 기능을 수행하는 Python 함수로 정의

Edge

- 노드 간의 연결을 정의하여 실행 흐름 결정
- 다음 노드로 상태 전달

LangGraph



StateGraph

- 노드와 엣지로 구성
- 각 노드는 특정 작업 수행 및 상태 업데이트
- 체크포인터를 통해 저장 및 고나리
- 복잡한 AI 시스템 구현에 적합

Framework 비교

Framework	특징
AutoGen	• 가장 사용하기 쉬움 • 자동화된 멀티 에이전트 협업이 필요한 경우 적절
CrewAI	• 각 Agent에 1개 이상의 Task를 주어 더 상세한 프롬프팅이 가능 • 팀 기반 역할 수행이 필요한 경우 적합
LangGraph	• 고도로 커스텀 가능 • LangChain과 연동하여 사용 가능 • 복잡한 LLM 워크플로우 및 상태 유지 챗봇을 만들 때 적합

03

PlanTy Chatbot

| Persona | Chatbot System |

Persona

모델이 따르는 특정한 성격, 역할, 말투, 배경지식 등
일관된 말하기 스타일 또는 행동 양식

Chatbot System

Fine-Tuning

농사로
용인특례시 농업기술센터
모두가 도시 농부

mixtral-8x7b-32768

Question & Answer

Chatbot System

Fine-Tuning


```
{
   "question": "분갈이는 언제 해야 하나요?",
   "true_answer": "보통 봄이나 가을에 분갈이를 하는 것이 좋습니다.",
   "generated_answer": " 보통 5-7일에 한 번씩 화분에서 식물을 빼내어 뿌리 주변을
                        정리하는 것이 좋습니다….
}
```

Chatbot System

Fine-Tuning

```
prompt = f"""“
    질문: {question}
    참조 답변: {reference_answer}
    모델 생성 답변: {model_answer}

    위의 질문과 답변을 평가해주세요.
     특히 답변이 식물에 대한 것인지 여부를 고려하여 다음 두 가지 기준에 따라 1-10점 사이의 점수를 매겨주세요:

    1. 식물 정보 정확성: 모델의 답변에 포함된 식물 관련 정보가 참조 답변과 비교하여 얼마나 정확한가?
        (만약 질문이 식물에 관한 것이 아니라면 이 항목은 건너뛰고 설명을 "해당 없음"으로 해주세요.)
    2. 답변 적절성: 모델의 답변이 질문에 얼마나 적절하고 유용한가?

    각 기준에 대한 점수와 간단한 설명을 제공해주세요.
"""
```

Chatbot System

Fine-Tuning

index	question	elapsed_time_sec
0	분갈이는 언제 해야 히	4.7285
1	분갈이를 왜 해야 하니	11.7117
2	분갈이할 때 흙은 어떻	2.6417
3	분갈이 주기는 어느 정	12.0653
4	분갈이 후에 물은 어떻	6.796
5	분갈이 후 잎이 시들어	11.4788
6	분갈이할 때 뿌리는 잘	3.2876
7	분갈이 후 비료를 줘도	5.7628

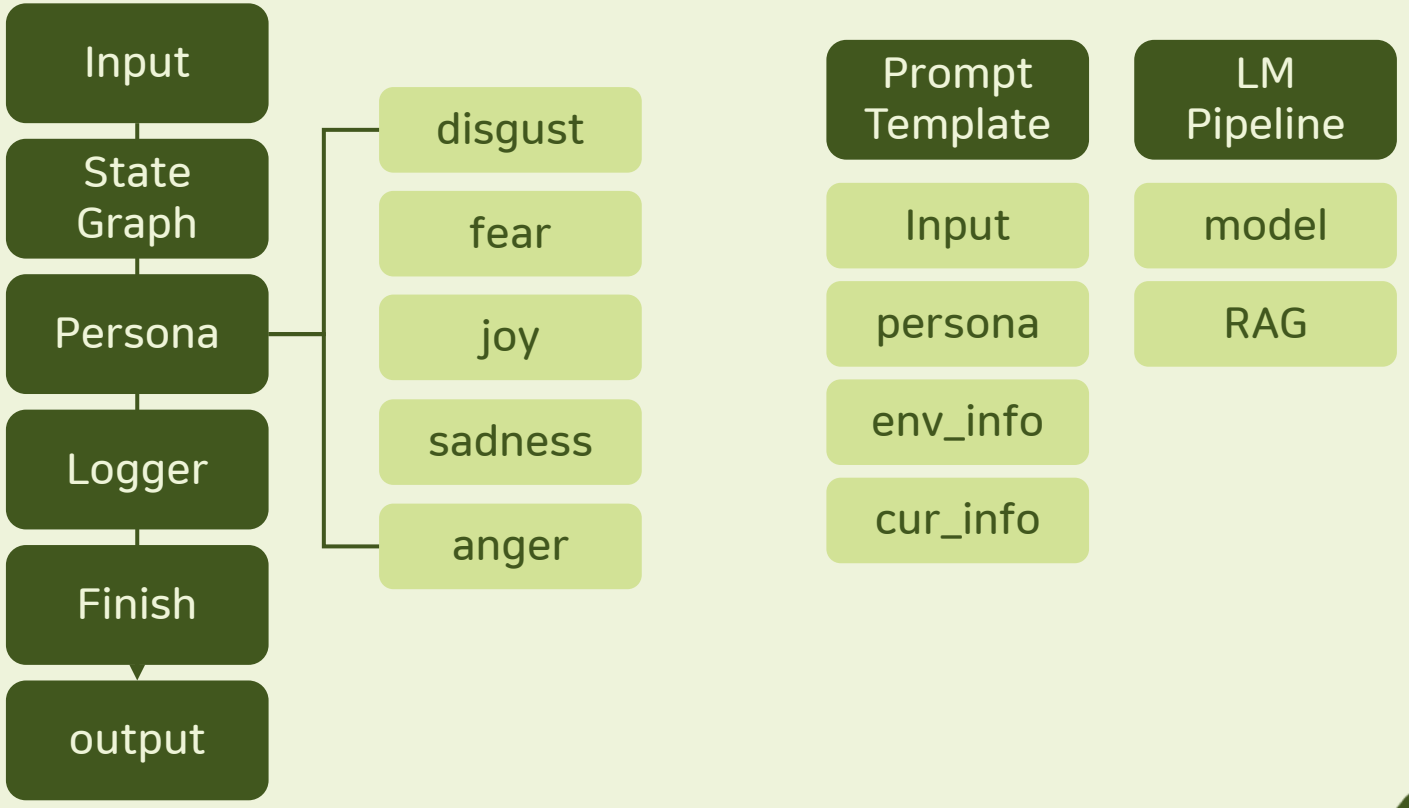
Chatbot System

Fine-Tuning

모델	식물 정보 정확성	평균 답변 적절성	모델 사이즈(MB)	응답시간 (s)	ox 정답률
gemma-3-4b (기본 모델)	8.59	8.76	16403.70	36.25	84
gemma_sft	8.88	8.58	12.33	42.07	84
gemma-3-4b-planty-2	7.5	7.24	45.46	47.16	88
gemma-3-4b-planty-ia3	9.12	8.53	1.65	41.26	82
HyperCLOVAX-SEED-Text-Instruct-1.5B (기본 모델)	6.6	6.2	6048	7.07	54
HyperCLOVAX-SEED-Text-Instruct-1.5B-planty-ia3	6.31	6.17	0.76	8.80	54
HyperCLOVAX-SEED-Text-Instruct-1.5B-planty-ia3-2	6.58	6.35	0.76	10.06	54

Chatbot System

LangGraph



Chatbot System

RAG

Chroma

HTML, PDF

```
rag_prompt = ChatPromptTemplate.from_messages([
    ("system", system_prompt),
    ("human", "{input}"),
])

rag_chain = create_retrieval_chain(
    compression_retriever,
    create_stuff_documents_chain(lm, rag_prompt)
)
```

References

AgentChat . (n.d.). https://microsoft.github.io/autogen/stable/user-guide/agentchat-user-guide/index.html.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv. org/abs/2305.14314, 2. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.

Introduction . (n.d.). https://docs.crewai.com/introduction.

Introduction . (n.d.).https://python.langchain.com/docs/introduction/?_gl=1*1sexmlh*_ga*MTY4NDQ1MzI5MC4xNzM5MjUzMDc1*_ga_47WX3HKKY2*czE3NDc0NzgyOTEkbzEkZzAkdDE3NDc0NzgyOTEkajAkbDAkaDA.

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. A. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Advances in Neural Information Processing Systems, 35, 1950-1965.

Overview . (n.d.).https://langchain-ai.github.io/langgraph/concepts/why-langgraph/.

시난 오즈데미르. (2024). 쉽고 빠르게 익히는 실전 LLM. n.p.: 한빛미디어. 강다솔. (2024). 한 권으로 끝내는 실전 LLM 파인튜닝. n.p.: 위키북스.

윤성재. (2024). RAG 시스템 구축을 위한 랭체인 실전 가이드. n.p.: 루비페이퍼.

허정준. (2024). LLM을 활용한 실전 AI 애플리케이션 개발. n.p.: 책만.

Thank You