Surfing on the Big Wave

Building LLM-Native Products

김주호 @ Liner

LLM Agent Application을 만들며 겪은 시행착오

김주호 @ Liner

Help People
Get More Done
With Less Time and Energy

liner

Autonomous Agent

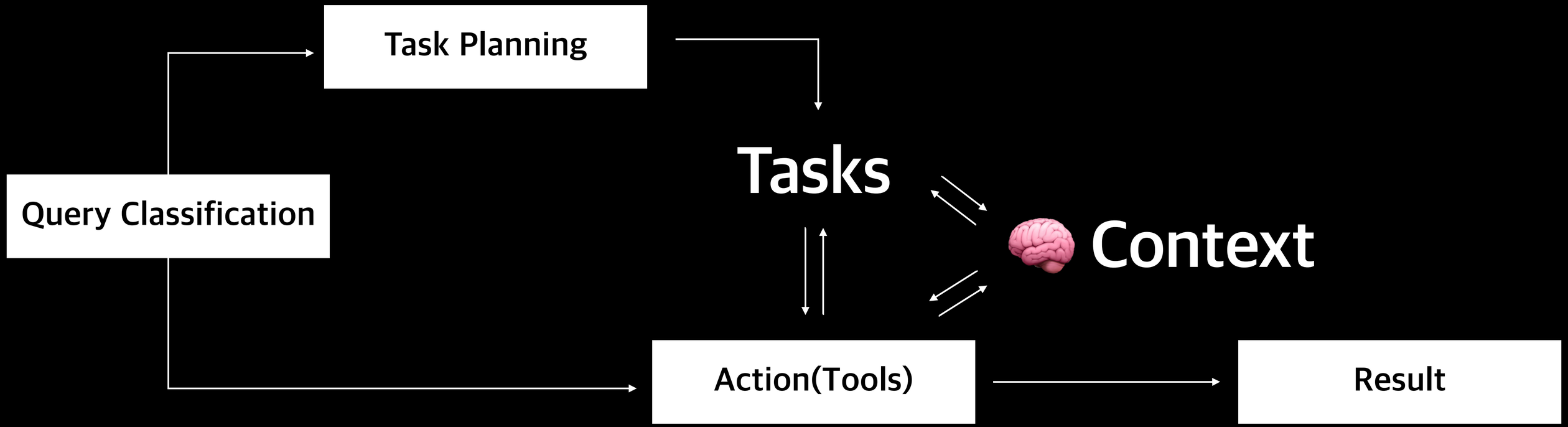
Autonomous Agent

Query Classification

Task Planning

Action(Tools)

Autonomous Agent 🤖



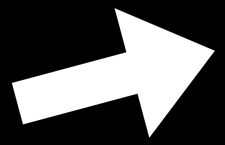
liner

“Tech 분야 최신 뉴스 브리핑해줘”

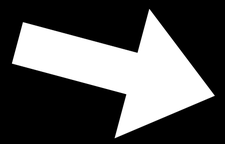
Autonomous Agent 🤖

Query Classification

“Tech 분야 최신 뉴스 브리핑해줘”



{Non-Planning}

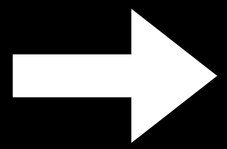


{Planning}

Autonomous Agent 🤖

Task Planning

“Tech 분야 최신 뉴스 브리핑해줘”

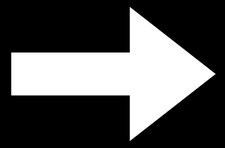


- 📄 Tech 분야 최신 뉴스 소스 탐색
- 📄 뉴스 소스로 부터 핵심 추출
- 📄 브리핑 형태로 반환

Autonomous Agent 🤖

Action(Tools)

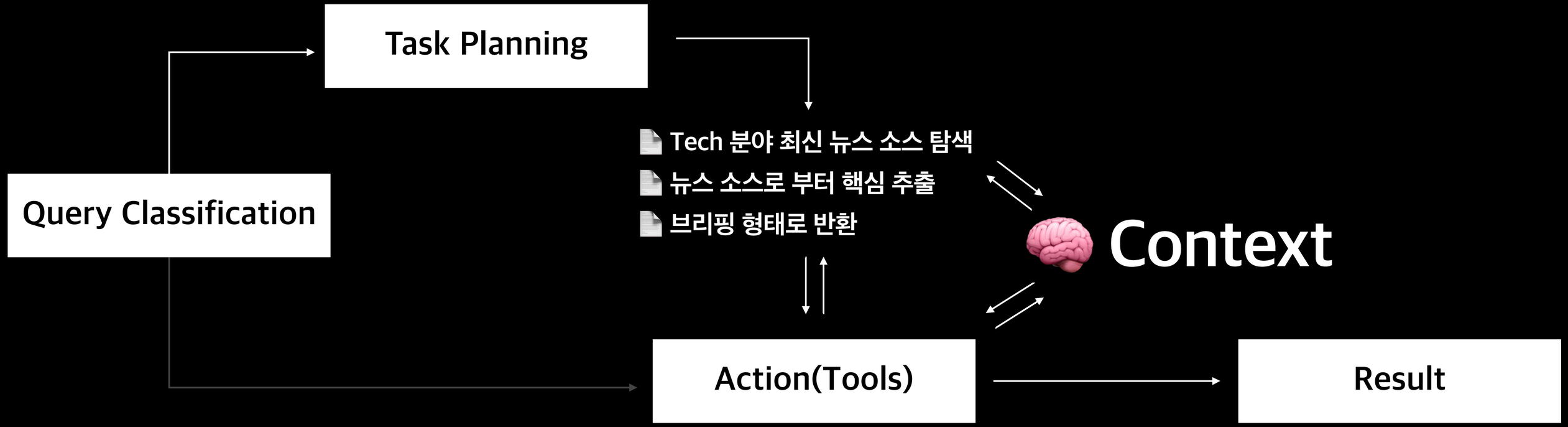
“Tech 분야 최신 뉴스 소스 탐색”



```
{
    “tool” : “search_engine”,
    “keyword” : “latest_news”,
}
```

Autonomous Agent 🤖

“Tech 분야 최신 뉴스 브리핑해줘”



Autonomous Agent 🤖

이 두 논문이 말하고자 하는 바를 비교해줘 (RAG)

이 이미지에 나온 수식 설명해줘 (Vision)


Challenges

LLM Chain System

Evaluation

Asynchronous UX

LLM Chain System

LLM Chain System

Task Planning

Tasks

Query Classification

Context

Action(Tools)

Result

LLM Chain System

Query Classification

Task Planning

Action(Tools)

Result(Aggregation)

⛓️ LLM Chain System

Query Classification

Task Planning

Action(Tools)

Result(Aggregation)

LLM Chain System

Query Classification

Task Planning

Action(Tools)

Result(Aggregation)

Optimize

LLM Chain System

Query Classification

Task Planning

Action(Tools)

Result(Aggregation)

Optimize

LLM Chain System

⛓️ LLM Chain System

🧠 추론 능력 상승

🤑 Prompt + Few-shot을 생략하여 비용, 속도 개선

⛓️ LLM Chain System

Query Classification

⏳ Throughput	Fine-Tuned GPT-3.5 (2x)	>	GPT-3.5
💸 $/Request	Fine-Tuned GPT-3.5 (1/4)	<	GPT-4
💯 Accuracy	Fine-Tuned GPT-3.5	~	GPT-4

Evaluation

Evaluation

Evaluation

Query Classification

Task Planning

Action(Tools)

Result(Aggregation)

Evaluation

컴포넌트별 측정

Query Classification

Action(Tools)

Evaluation

컴포넌트별 측정

Query Classification

Action(Tools)

Task : Classification

Metric : Accuracy, F1

Evaluation

Task Planning

- Tech 분야 최신 뉴스 소스 탐색
- 뉴스 소스로 부터 핵심 추출
- 브리핑 형태로 반환

Task : 🤔

Result : SubTaks List

Metric : ❓

Aggregation

TechCrunch 기사들을 살펴보면,
다방면의 기술과 스타트업 관련
소식이 다루어지고 있습니다

Task : QA? Summary?

Result : Text Sequence

Metric : ❓

Evaluation

Few-Shot + Detail Instruction + GPT-4

Evaluation

Query Classification

Task Planning

Action(Tools)

Aggregation

Asynchronous UX

Asynchronous UX

📮 Asynchronous UX

User

Agent

FastAPI로 ML 모델 서빙하는 예시 코드와
각 코드를 관리할 폴더구조 알려줘

📪 Asynchronous UX

User

Agent

알려드리겠습니다 ```

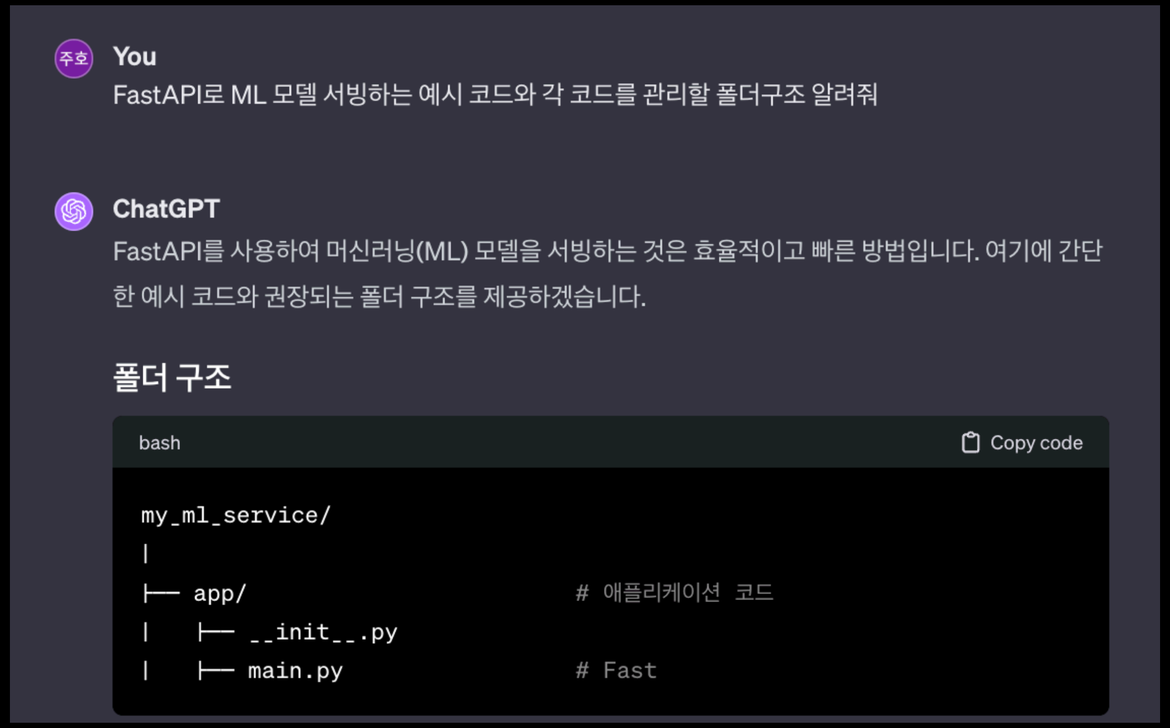
📮 Asynchronous UX

User

Agent

알려드리겠습니다 ```

📮 Asynchronous UX



You
FastAPI로 ML 모델 서빙하는 예시 코드와 각 코드를 관리할 폴더구조 알려줘

ChatGPT
FastAPI를 사용하여 머신러닝(ML) 모델을 서빙하는 것은 효율적이고 빠른 방법입니다. 여기에 간단한 예시 코드와 권장되는 폴더 구조를 제공하겠습니다.

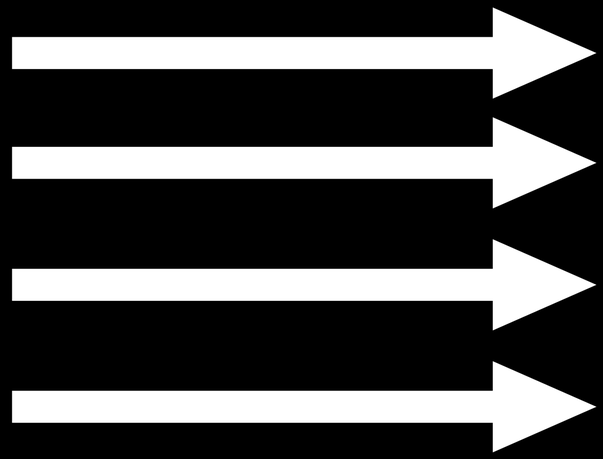
폴더 구조

bash — Copy code

```
my_ml_service/
|
├── app/                          # 애플리케이션 코드
|   ├── __init__.py
|   ├── main.py                   # Fast
```

📪 Asynchronous UX

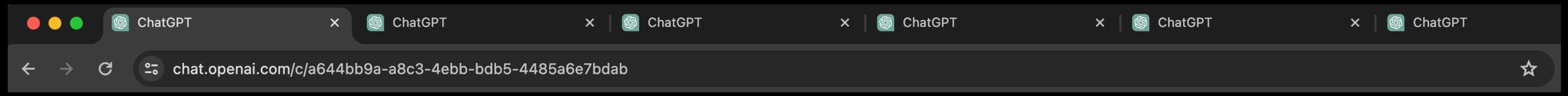
User



이것도 알려줘
저것도 알려줘
근데 이것도 수행해줘
하지만 이것도 등록해줘

Agent

📮 Asynchronous UX



📮 Asynchronous UX

liner

Autonomous Agent

📮 Asynchronous UX

Autonomous Agent

1. 동시에 여러개 요청할 수 있음
2. 이탈후 생성된 답변도 저장해줌
3. 읽지 않은 답변은 알려줌

📮 Asynchronous UX

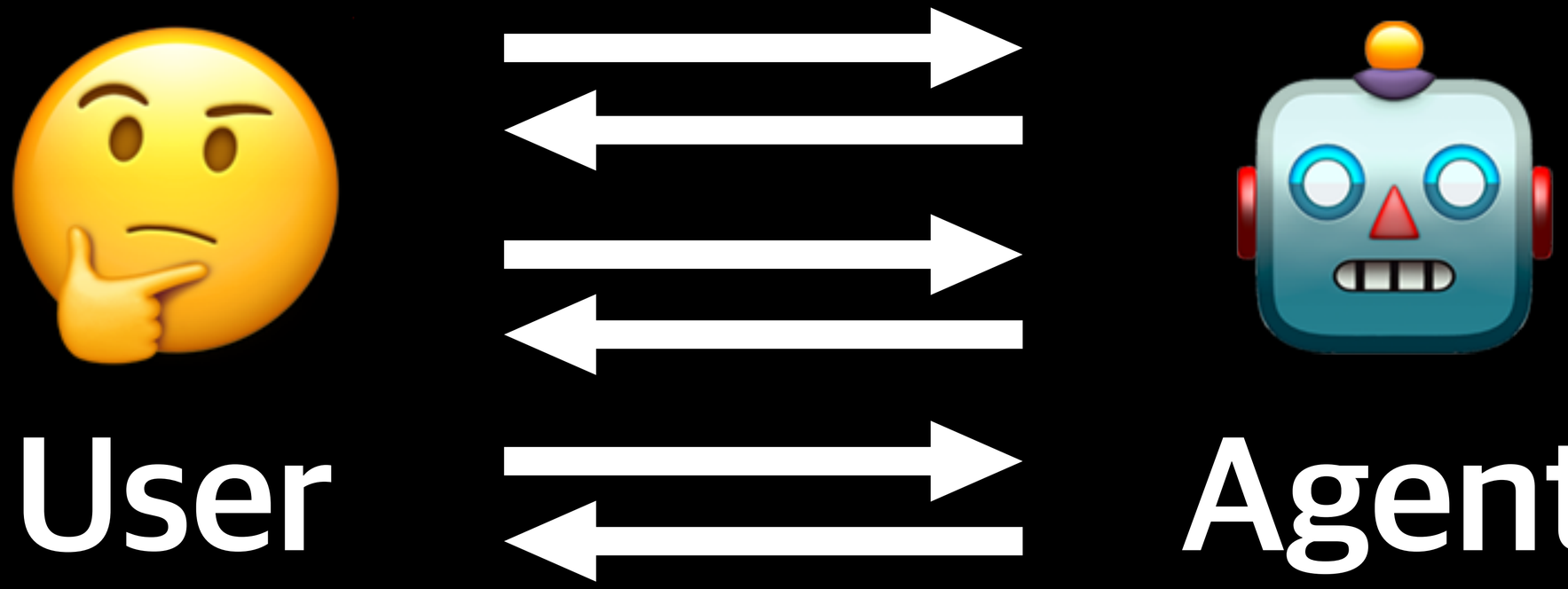
요청

답변

User

Agent

📪 Asynchronous UX



에이전트가 답변하는 동안 여기서 다른 질문을 할 수 있어요

📮 Asynchronous UX

Result

User

Agent

Wrap-Up

Wrap-Up

⛓️ LLM Chain System

결과물 좋지만 오래걸리고 비쌈

적절한 최적화 필요

Wrap-Up

 Evaluation

컴포넌트별 유닛 테스트

전통적인 Metric + GPT-4를 Evaluator로서 활용

Wrap-Up

Asynchronous UX

"Autonomus" Agent 답게
비동기 경험 가능하도록 구축


liner