AI Service INTRO

인공지능 서비스 개발 교육 과정

AI 서비스란?

AI 기술을 활용해 사용자·기업의 문제를 해결하고 새로운 가치를 창출하는 서비스

- **자동화된 처리: 반복적인 업무를 자동으로 수행**
- **데이터 기반 의사결정: 대량의 데이터를 분석하여 패턴을 찾고 예측**
- **개인화된 경험**: 사용자별 맞춤 서비스 제공
- **실시간 분석**: 실시간으로 데이터를 처리하고 응답

🤔 이전에도 되던 것 아닌가? 왜 AI가 강조되는 걸까?

AI 서비스란?

1. 정확도와 효율성의 비약적 향상

- 기존 자동화: 사람이 작성한 규칙 기반 → 예외 상황에 취약
- AI 자동화: 머신러닝/딥러닝이 데이터에서 **패턴을 스스로 학습** → 예외나 복잡한 상황에도 더 잘 대응

2. 데이터 규모와 복잡도 처리 능력

- 전통적 통계나 BI 툴은 “샘플링 + 단순 회귀/분류” 중심
- AI는 **수십억 건 규모 데이터**를 학습해 복잡한 상관관계를 찾아냄
- 예: 제조업 불량률 → 과거엔 몇 가지 주요 원인만 분석 가능, 지금은 센서 수천 개 신호까지 학습해 미세한 이상 감지 가능

즉, 예전에도 “가능”은 했지만, AI는 **정확도·스케일·개인화 정밀도·실시간 대응력**에서 기존 기술과는 차원이 다른 성과를 내면서, 같은 카테고리의 서비스도 **혁신 수준**으로 끌어올렸기 때문에 강조되는 것.

AI 서비스 트렌드

1. 업무 자동화 및 운영비용 절감

- **의미**: 단순 반복 업무를 AI가 대체하면서 인건비·시간 절감 효과를 가져옴.
- **사례**: 고객센터 → 콜봇·챗봇 도입으로 상담 인력 30~50% 절감

2. 비결정성(Non-determinism)과 평가 난이도

- **의미**: 같은 입력과 설정에서도 결과가 다를 수 있는 AI 특성 → 일관성·재현성 관리가 어렵다.
  - 벤치마크 테스트만으로는 서비스 품질 보장 불가
  - 다회 샘플링, LLM-as-judge, 사용자 피드백 루프 필요
- **사례**:
  - 번역 서비스 → 같은 문장을 번역해도 미묘하게 다른 결과 → 품질 검증 프로세스 강화 필요

AI 서비스 트렌드

3. 외부 지식·도구 결합 (RAG & 함수 호출)

- **의미**: 모델이 학습한 파라미터에만 의존하지 않고, **실시간 외부 데이터와 시스템을 호출해 보완**하는 아키텍처
  - API / DB 검색 / 사내 지식베이스 접속
  - 스프레드시트·업무툴 연동
- **사례**:
  - 법률 서비스: 최신 판례·법령 DB와 결합한 LLM 검색 → 법률 자문 정확도 강화

4. 확장 트렌드

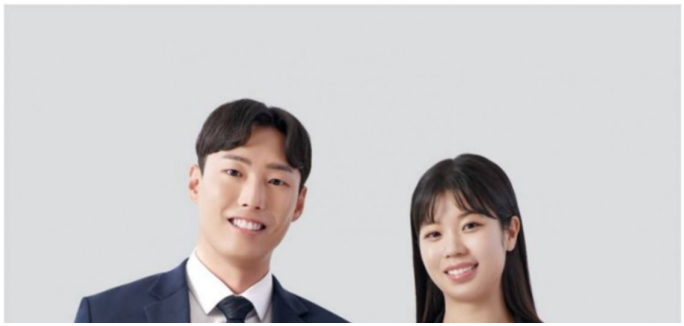
- **멀티모달 AI**: 텍스트뿐 아니라 이미지·음성·영상 동시 처리 (예: Copilot + Vision)
- **규제와 윤리**: EU AI Act, 개인정보보호법, 저작권 이슈 대응 필요
- **MLOps·AIOps**: 실제 운영환경에서 AI 성능 유지·모니터링·업데이트 자동화

AI 서비스 예시

업계 최초, ‘AI OCR’ 활용 新 위험률 개발… 한화생명, 주요 암(위·간·폐) 특약 출시

입력 2024.04.15 09:55 수정 2024.04.15 10:07 가가

☐ 비정형 진료비세부내역서 등 5백만건, ‘AI OCR’로 읽어 ‘경험통계’로 집적
☐ 주요 3대 암(위·간·폐) 중심으로 분석, ‘시그니처 암보험 3.0’에 특약 탑재
☐ ‘AI OCR’ 기술 활용 新위험률 확대… “새로운 보장으로 시장 선도해 갈 것”



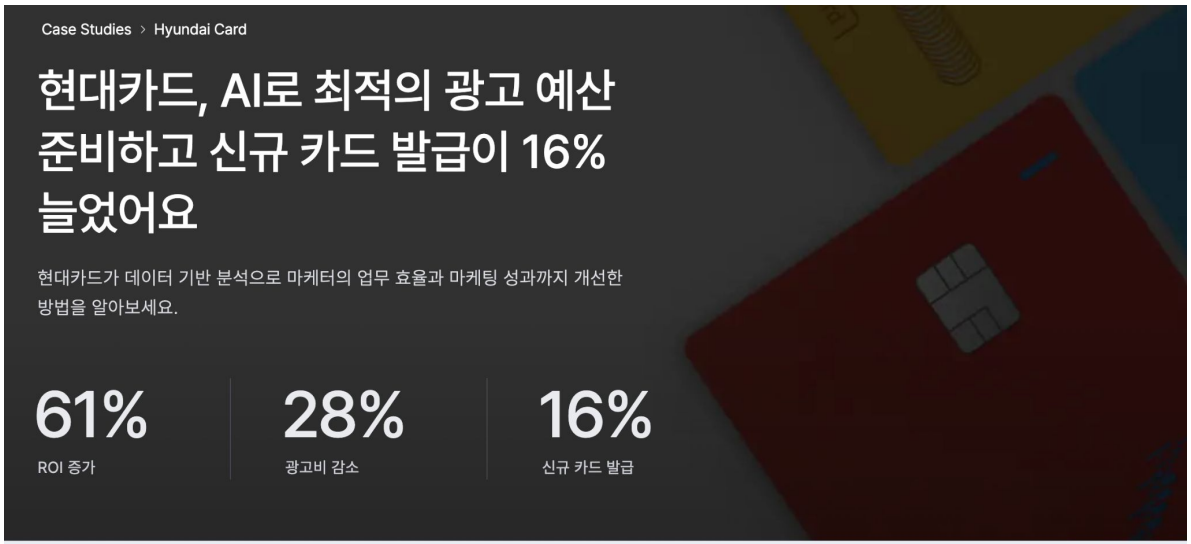
많이 본 뉴스

1 "10년 만에 연락"...김나영, 환자복 입고 눈물 보인 이...
2 '신혼여행 성지'의 추락...13년 만에 항공 운항 중단 선언
3 "다이소 미쳤나"...커피 전문가 경악한 '5000원 아이템'
4 "1억이 두 달 만에 5000만원 됐다"...개미들 '패닉'
5 '신현준 아내', 첼리스트 접고 육아 전념하더니...깜짝 변신

ADVERTISEMENT

https://www.hankyung.com/article/202404150830P?utm_source=chatgpt.com

AI 서비스 예시



"에어브릿지가 광고비 절감과 퍼포먼스 향상 뿐만 아니라 업무 부담도 크게 줄여주었습니다. 에어브릿지 AI가 제안한 예산안 덕분에 광고비는 28% 감소하고 신규 카드 발급은 16% 증가해서 ROI가 61% 증가하는 놀라운 결과를 달성했습니다. 현대카드 온라인 마케팅 팀은 에어브릿지를 활용하여 고객 여정을 자세히 분석하고, 광고 채널별로 신규 회원 모집 기여도를 확인합니다. 그리고 정확한 데이터에 기반하여 보다 전략적이고 직접적인 전환에 도움을 주는 의사 결정을 내리고 있습니다.

https://www.airbridge.io/ko/case-studies/hyundai-card-ad-spend-optimization?utm_source=chatgpt.com

AI 서비스 예시

- **제조업**
  - 스마트팩토리로 센서 데이터 기반 자동 제어를 구현 -> 중단 없는 연속 작업과 불순물 관리 정밀화를 달성
  - 반도체에선 공정 미세화로 인한 간섭·누설을 AI로 최적화하고 웨이퍼 손실 원인을 분석해 불량률↓·수율↑
- **금융**
  - 생성형 AI 챗봇·음성봇으로 창구 상담을 대체하고 고객 데이터 기반 맞춤 응대를 제공
  - 로보어드바이저가 광범위한 금융지표를 학습해 개인별 포트폴리오를 제안하며 운용 규모가 빠르게 확대
- **유통**
  - 과거 판매·고객·날씨·프로모션 데이터를 학습해 지역·시점별 수요를 정밀 예측, 재고·폐기 최소화
  - RMN(소매 미디어 네트워크)로 구매 시점 타게팅 광고를 집행하고 멤버십으로 데이터 확보·추천 고도화
- **게임**
  - LLM으로 NPC 대화를 자연어 기반으로 구현해 맥락 이해·응답 다양화로 몰입도 향상
  - 플레이어 실력·상황에 맞춘 적응형 AI로 몬스터/퀘스트 패턴을 동적으로 변화시켜 난이도·전략성을 강화

https://www.samsungsds.com/kr/insights/ai_use_cases.html

AI 서비스 예시

- **Character.ai & Janitor.ai**
  - 가상 캐릭터와 대화할 수 있는 AI 플랫폼들로, Character.ai는 구글 출신 인물들이 설립한 서비스로 다양한 게임, 애니메이션, 실제 인물 등과 소통할 수 있음
  - anitor.ai는 60,000개 이상의 사전 제작된 AI 캐릭터를 제공하며 더 자유로운 대화가 가능하며, 애니메이션 팬, 작가, 게이머들 사이에서 특히 인기가 높음

- **ElevenLabs & LALAL.AI & Speechify**
  - 30초 정도의 음성 샘플만으로도 누구의 목소리든 복제할 수 있는 AI 서비스들입니다. 사랑하는 사람의 목소리를 보존하거나, 팟캐스트 제작, 오디오북 생성 등에 활용할 수 있음
  - 죽은 분의 목소리를 재현하는 것도 인기를 끌고 있음

https://www.samsungsds.com/kr/insights/ai_use_cases.html

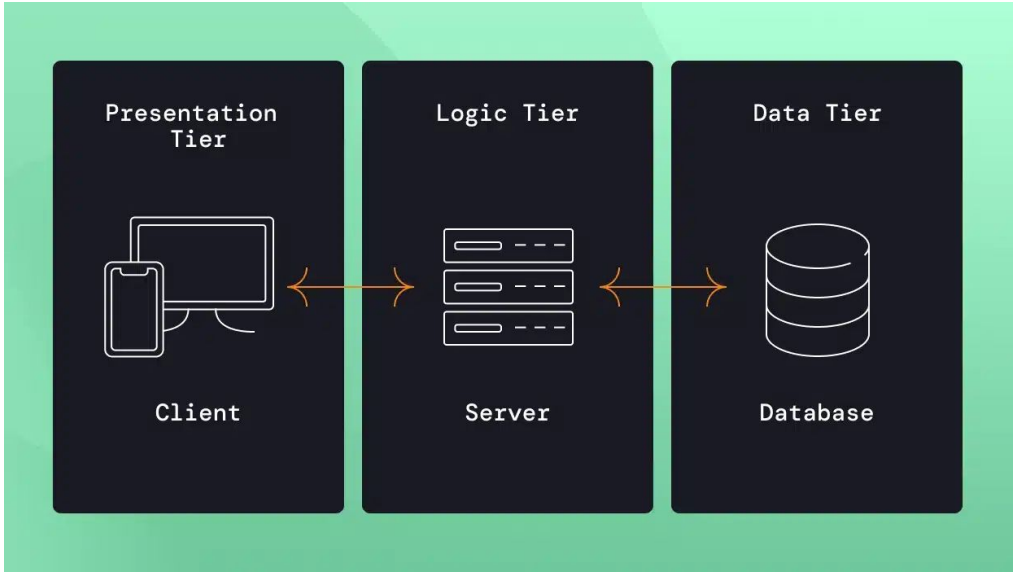
AI 서비스 예시

- **카리즈AI**
  - 카카오톡 대화 분석을 통해 연애 상대방의 호감도, 밀당 상황을 분석하고 전략적 대응책을 제안하는 한국형 AI 연애 코치
  - 심리학 기반의 연애 분석과 맞춤형 컨설팅을 제공함

- **Gamma & presentations.ai**
  - 몇 초 만에 멋진 파워포인트 프레젠테이션을 만들 수 있는 'PPT용 ChatGPT'라고 불리는 서비스로, 아이디어를 입력하면 전문 PPT를 자동 생성

https://www.samsungsds.com/kr/insights/ai_use_cases.html

AI 서비스 아키텍쳐



전통적인 3 tier 아키텍처

- **프레젠테이션 계층(Presentation Tier)**

  최종 사용자가 시스템과 상호작용하는 사용자 인터페이스(예: 웹 브라우저, 모바일 앱)

- **로직 계층(Logic Tier)**

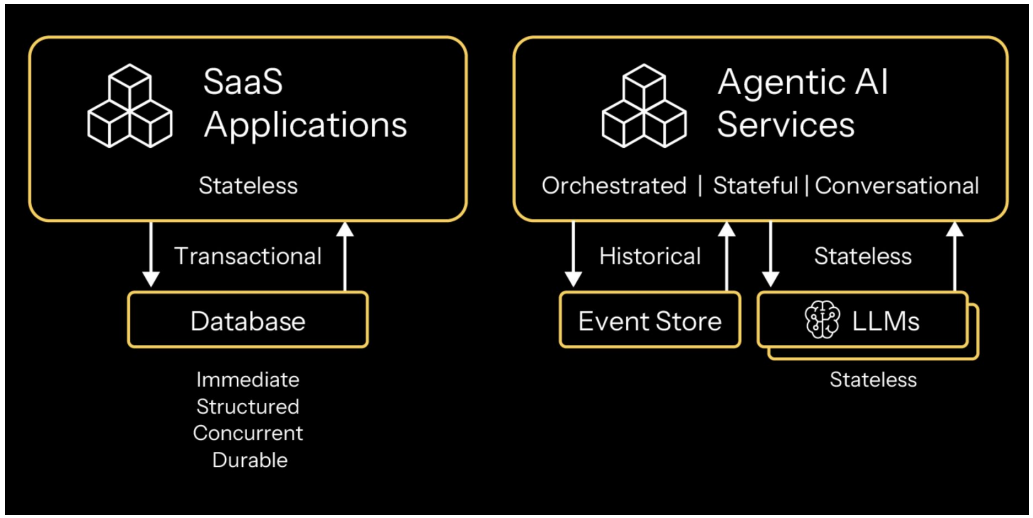
  아키텍처의 중간 계층으로, 애플리케이션의 핵심 처리, 비즈니스 규칙, 계산을 담당

- **데이터 계층(Data Tier)**

  애플리케이션 데이터의 저장, 조회, 조작을 관리하며, 일반적으로 데이터베이스를 사용

https://vfunction.com/blog/3-tier-application/

AI 서비스 아키텍쳐



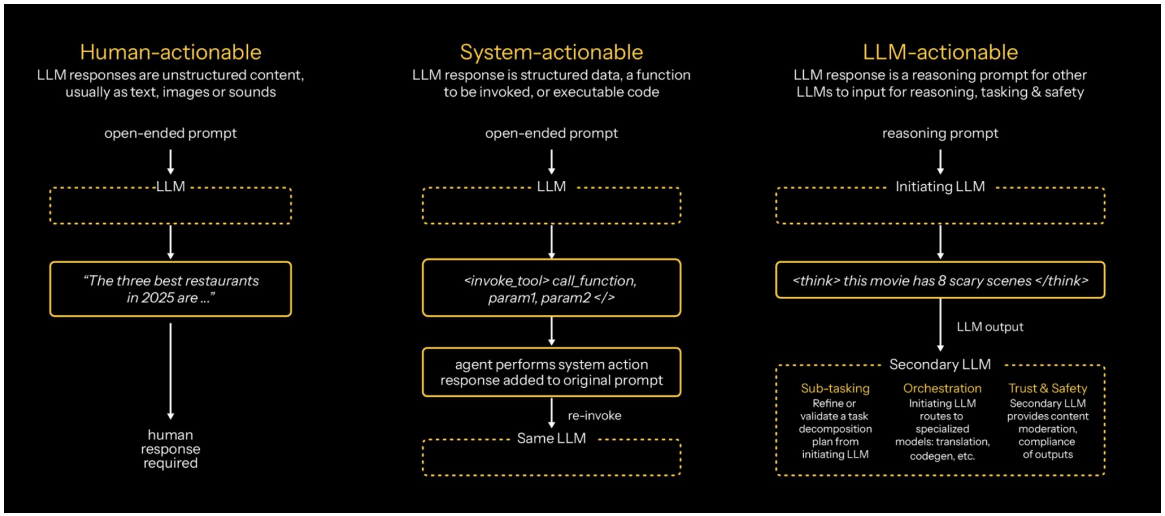
Agentic AI 서비스

- 에이전트형(Agentic) AI 서비스는 전통적인 SaaS 애플리케이션의 트랜잭션 중심(요청–응답 패러다임)에서, 대규모 언어 모델(LLM)을 중심으로 한 대화 중심의 반복적 패러다임으로의 근본적 전환을 의미

- AI 에이전트의 핵심에는 **대규모 언어 모델(LLM)**이 있음.
    - LLM은 방대한 텍스트로 학습된 강력한 신경망으로, 인간과 유사한 응답을 생성하며, 이들은 추론, 요약, 의사결정에 뛰어나므로 에이전트형 AI의 핵심 구성 요소로 작용함

https://akka.io/what-is-agentic-ai

AI 서비스 아키텍쳐



Agentic AI 서비스

- LLM은 **실행 가능한 출력물을 생성**함
- LLM의 출력물은 사람, 시스템, 또는 다른 LLM 가운데 하나가 처리
  - Human-actionable 출력:

    특정 프롬프트에 답하는 비정형 텍스트나 이미지처럼, 사람이 바로 이해하고 행동할 수 있는 결과
  - System-actionable 출력:

    함수 호출, 실행 가능한 코드, 구조화된 데이터처럼, 시스템이 바로 처리·실행할 수 있는 결과
  - LLM-actionable 출력:

    다른 LLM이 추가 추론이나 작업 수행을 위해 입력으로 사용할 수 있는 프롬프트와 컨텍스트

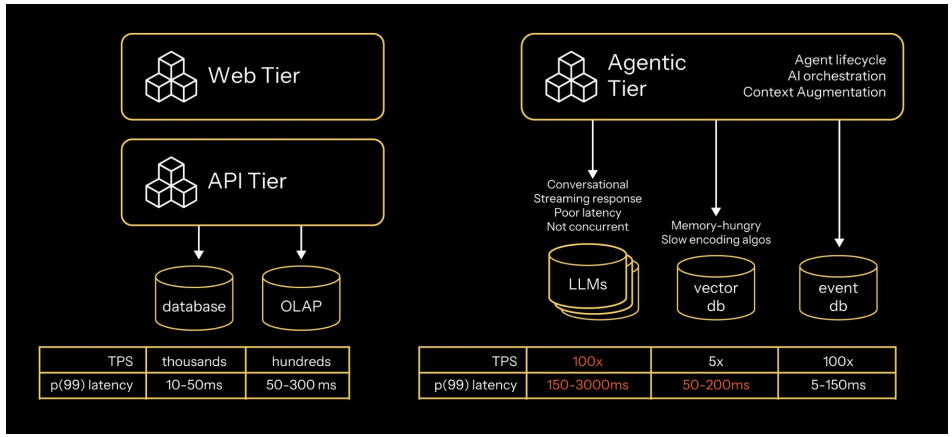
https://akka.io/what-is-agentic-ai

AI 서비스 아키텍쳐

SaaS Applications	Agentic AI
Transaction-centered: short, stateless interactions	**Conversation-centered**: Long-lived, stateful interactions
Immediate response: system processes request and returns a result	**Context-aware responses**: system remembers past interactions to stay relevant
Minimal compute and memory per transaction	**High compute and memory usage per conversation**
Predictable scaling: More users = more transactions = predictable load	**Unpredictable scaling**: More users lead to exponentially higher workload due to longer interactions and agent-driven actions

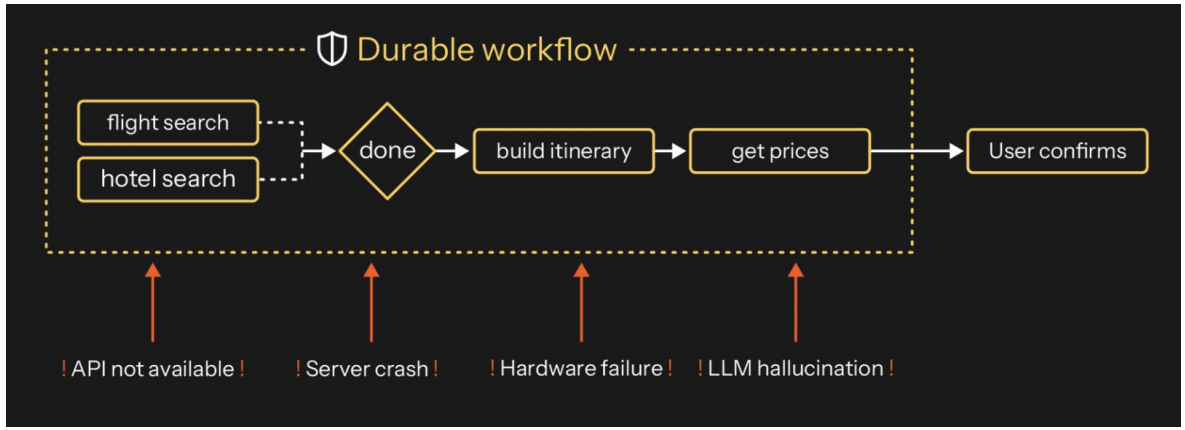
Agentic AI 서비스

- 시스템에 LLM을 추가하면 전통적 소프트웨어로는 구현하기 어려운 추론, 요약, 종합, 의사결정과 같은 강력한 능력이 생김
- 그러나 이러한 역량을 제대로 활용하려면 시스템 설계 방식의 전환이 필요함



https://akka.io/what-is-agentic-ai

AI 서비스 아키텍쳐

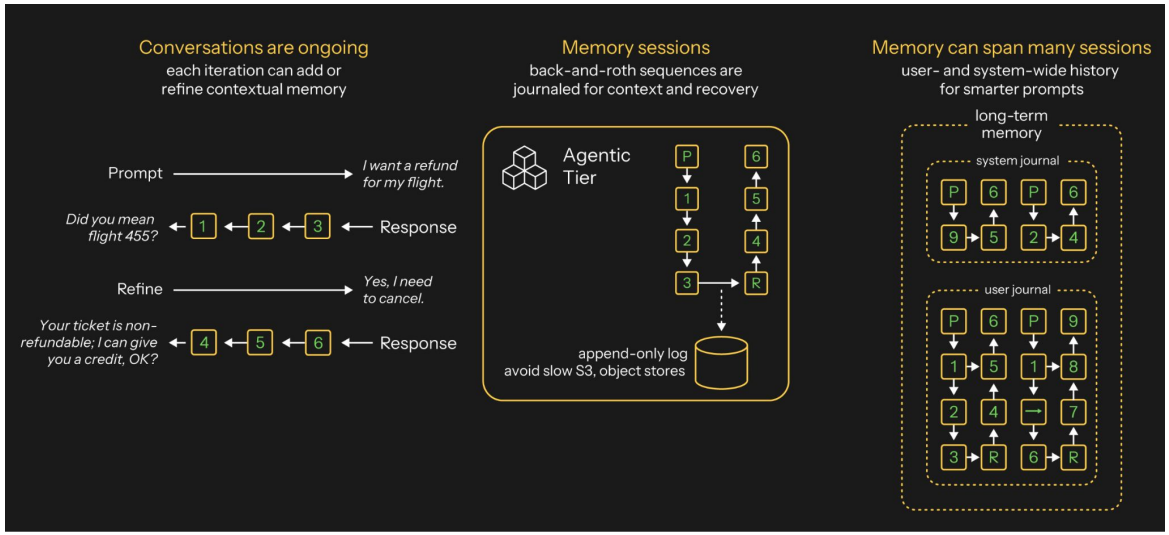


Agentic AI 서비스

- (1) Agent Orchestration
  - 하드웨어 장애, 타임아웃, 환각(hallucination), 재시작이 발생하더라도 에이전트의 동작과 LLM 호출이 신뢰성 있게 실행되도록, 장시간 실행되는 다단계 프로세스를 관리하는 견고한 워크플로우 필요

https://akka.io/what-is-agentic-ai

AI 서비스 아키텍쳐

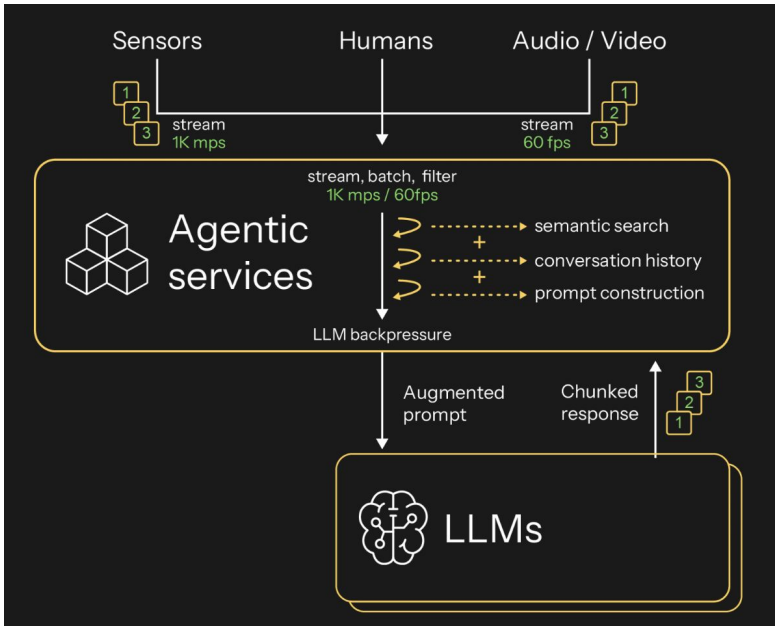


Agentic AI 서비스

- (2) Memory
  - 대화형 상호작용 전반에서 문맥을 유지하기 위한 지속적인 단기 및 장기 메모리 제공

https://akka.io/what-is-agentic-ai

AI 서비스 아키텍쳐

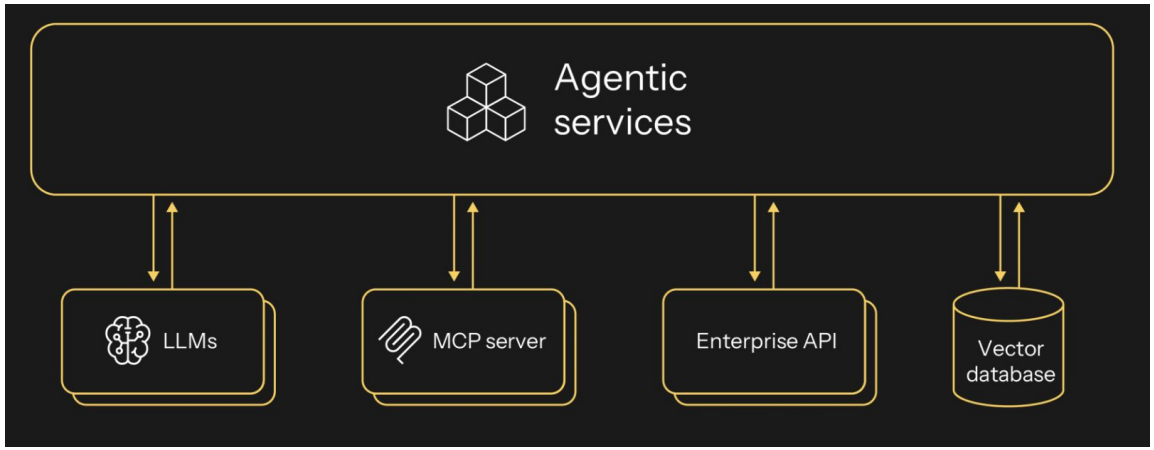


Agentic AI 서비스

- (3) Streaming
  - 비디오·오디오·IoT 데이터·메트릭·이벤트 스트림 등 대용량 데이터를 에이전트가 신속히 처리하고 응답하도록 하며, LLM 지연(latency)을 무리 없이 다뤄 응답성을 보장해야함
  - 스트리밍은 앰비언트 에이전트(ambient agents, 이벤트 스트림을 지속적으로 모니터링하다가 필요할 때만 응답하는 AI 에이전트)를 지원함

https://akka.io/what-is-agentic-ai

AI 서비스 아키텍쳐



Agentic AI 서비스

- (4) Integrations and Tool Support
  - 엔터프라이즈 API, 데이터베이스, 외부 도구와의 네이티브 연동을 제공하며, OpenAPI 같은 확립된 표준과 Model Context Protocol(MCP) 같은 표준을 활용함

https://akka.io/what-is-agentic-ai

다양한 AI 모델들 - OpenAI

Frontier models OpenAI's most advanced models, recommended for most tasks.

GPT-5
The best model for coding and agentic tasks across domains

GPT-5 mini
A faster, cost-efficient version of GPT-5 for well-defined tasks

GPT-5 nano
Fastest, most cost-efficient version of GPT-5

GPT-4.1
Smartest non-reasoning model

다양한 AI 모델들 - OpenAI

Specialized models Purpose-built for specific tasks.

o3-deep-research
Our most powerful deep research model

o4-mini-deep-research
Faster, more affordable deep research model

GPT Image 1
State-of-the-art image generation model

DALL·E 3
Previous generation image generation model

GPT-4o mini TTS
Text-to-speech model powered by GPT-4o mini

GPT-4o Transcribe
Speech-to-text model powered by GPT-4o

GPT-4o mini Transcribe
Speech-to-text model powered by GPT-4o mini

다양한 AI 모델들 - OpenAI

Realtime and audio models Models for audio use cases and realtime inputs and outputs.

gpt-realtime
Model capable of realtime text and audio inputs and outputs

gpt-audio
For audio inputs and outputs with Chat Completions API

ChatGPT models Models used in ChatGPT, not recommended for API use.

GPT-5 Chat
GPT-5 model used in ChatGPT

ChatGPT-4o
GPT-4o model used in ChatGPT

다양한 AI 모델들 - Anthropic

Feature	Claude Opus 4.1	Claude Opus 4	Claude Sonnet 4	Claude Sonnet 3.7	Claude Haiku 3.5	Claude Haiku 3
Description	Our most capable model	Our previous flagship model	High-performance model	High-performance model with early extended thinking	Our fastest model	Fast and compact model for near-instant responsiveness
Strengths	Highest level of intelligence and capability	Very high intelligence and capability	High intelligence and balanced performance	High intelligence with toggleable extended thinking	Intelligence at blazing speeds	Quick and accurate targeted performance

다양한 AI 모델들 - Google

정식 버전 Gemini 모델

Gemini 2.5 Pro
Google의 가장 진보된 추론 모델

Gemini 2.5 Flash
가격 대비 성능이 우수하고 기능이 다양한 최적의 모델

Gemini 2.0 Flash
차세대 기능과 향상된 기능을 갖춘 최신 멀티모달 모델

Gemini 2.0 Flash-Lite
비용 효율성과 지연 시간 단축에 최적화된 Gemini 2.0 Flash 모델

다양한 AI 모델들 - Google

Gemma 모델

Gemma 3
텍스트 및 이미지 입력으로 다양한 작업을 해결하고 140개 이상의 언어를 지원하며 긴 128K 컨텍스트 창을 갖춘 최신 Gemma 오픈 모델

Gemma
텍스트 생성, 요약, 추출을 지원하는 소형 경량 오픈 모델

PaliGemma
SigLIP와 Gemma를 결합한 Google의 오픈 비전 언어 모델

TxGemma
치료 관련 데이터를 기반으로 예측, 분류 또는 텍스트를 생성하며, 적은 데이터와 컴퓨팅으로 치료 관련 작업을 위한 AI 모델을 효율적으로 빌드하는 데 사용할 수 있습니다.

Gemma 2
텍스트 생성, 요약, 추출을 지원하는 두 번째 세대의 오픈 모델

ShieldGemma 2
정의된 일련의 안전 정책에 따라 텍스트와 이미지의 안전성을 평가하기 위한 명령 조정 모델

CodeGemma
중간 코드 완성, 코드 생성, 자연어 이해, 수학적 추론, 명령어 따르기와 같은 다양한 코딩 작업을 실행할 수 있는 강력하고 가벼운 오픈 모델

다양한 AI 모델들 - Google

프리뷰 Imagen 모델

Imagen 4 for Generation
텍스트 프롬프트를 사용하여 이전 이미지 생성 모델보다 품질이 우수한 새로운 이미지를 생성합니다.

Imagen 4 for Fast Generation
텍스트 프롬프트를 사용하여 이전 이미지 생성 모델보다 품질이 우수하고 지연 시간이 짧은 새로운 이미지를 생성합니다.

Imagen 4 for Ultra Generation
텍스트 프롬프트를 사용하여 이전 이미지 생성 모델보다 품질이 우수하고 프롬프트 준수가 더 우수한 새로운 이미지를 생성합니다.

Veo 모델

Veo 2 for Generation
텍스트 프롬프트와 이미지를 사용하여 새로운 동영상 생성

Veo 모델 허용 목록 미리보기

Veo 3 for Generation
텍스트 프롬프트와 이미지를 사용하여 새로운 동영상 생성

다양한 AI 모델들 - Google

http://www.youtube.com/watch?v=FKBL-bmLugg

python의 필요성과 AI에서의 역할

1. Python 생태계에서의 우위

- **이유**:
  - Python은 단순한 문법과 방대한 라이브러리 생태계를 기반으로 AI·데이터 과학 연구의 기본 언어로 자리 잡음.
  - 에이전트 오케스트레이션, RAG, 관측성, 스트리밍 등 **최신 AI 아키텍처 요소**들이 대부분 Python 환경에서 가장 먼저 구현되고 검증됨.

- **핵심 생태계**:
  - **PyTorch / Transformers** → 모델 학습 및 추론
  - **LlamaIndex / Haystack** → RAG 구현
  - **Weaviate, Milvus, Pinecone** → 벡터 DB 클라이언트 Python SDK 제공

👉 따라서 “새로운 기능은 먼저 Python에 나온다”는 말이 있을 정도로 **표준 플랫폼** 역할 수행.

python의 필요성과 AI에서의 역할

2. 스트리밍 / 실시간 처리

- **강점**: Python의 비동기 처리(asyncio), 이벤트 기반 프레임워크(FastAPI, Starlette)로 손쉽게 SSE(Server-Sent Events)·WebSocket 스트리밍 구현 가능.

- **활용 예시**:
  - ChatGPT 스타일의 실시간 응답 스트리밍
  - 실시간 데이터 모니터링 대시보드 (금융 거래 이상 탐지, IoT 센서 데이터 분석)

  👉 Python 기반으로 **프로토타입 → 프로덕션 서비스**로 빠르게 전환할 수 있음.

python의 필요성과 AI에서의 역할

3. Agent 모니터링 및 평가

- **관점**: AI 서비스는 단순 모델 호출이 아니라, 여러 도구·API를 오케스트레이션하는 **에이전트 환경**으로 확장됨.
- **필요성**: 이런 복잡한 플로우에서 **모니터링·평가 체계** 없이는 서비스 품질 보장 불가.
- **도구**:
  - OpenTelemetry GenAI → LLM 호출 및 툴 사용 기록
  - LangSmith, OpenAI Evals → 품질 지표 자동 수집·분석

  👉 Python 환경에서 바로 연동되므로, **테스트·관측·피드백 루프**가 쉬움.

python의 필요성과 AI에서의 역할

4. RAG 스택 구성 용이

- **이유**: RAG(Retrieval-Augmented Generation) = LLM + 검색/DB/지식베이스 조합 > Python SDK로 대부분 제공됨.

- **예시**:
  - Weaviate / Pinecone / Milvus → Python 클라이언트로 벡터 검색
  - LlamaIndex / Haystack → RAG 파이프라인 구축 라이브러리

  👉 따라서 Python은 **RAG 구현의 디폴트 언어**라 할 수 있음.

python의 필요성과 AI에서의 역할

Python – AI/데이터 표준

- 방대한 AI 생태계 (PyTorch, Transformers, LlamaIndex 등)
- 빠른 프로토타이핑 & 연구–프로덕션 전환 용이
- RAG·에이전트·관측성 프레임워크 최우선 지원 언어

JavaScript – 웹·실시간 UX

- WebSocket·실시간 처리 최적화
- 프론트엔드·브라우저 연동 관련이라면 ok
- LangChain.js 등 일부 AI 지원 있으나 제한적

Java – 안정성과 확장성

- 금융·통신 등 엔터프라이즈 중심
- 안정적 운영과 대규모 시스템에 강점
- AI 전용 라이브러리 부족 → Python 보완 필요

다른 언어들도 조금씩 올라오고 있으나,

AI 기반의 서비스를 한다면 기술 지원 / 편의성 / 인력 수급 등에서 Python이 우위

쉬는시간

Anaconda 설치하기

- **윈도우**
  - https://wikidocs.net/254434
- **맥**
  - https://soundprovider.tistory.com/entry/Miniconda-Ubuntu%EC%97%90-Miniconda-%EC%84%A4%EC%B9%98%ED%95%98%EA%B8%B0

실습환경

Jupyter Notebook

- 대화형 프로그래밍을 지원하는 웹 어플리케이션으로 중간 결과 확인 및 검색을 쉽게 진행할 수 있음
- jupyter를 통해 웹에서 실시간 프로그래밍이 가능함

Google Colaboratory (Colab)

- Google에서 제공해주는 무료 클라우드 기반 Jupyter notebook 환경
- 무료 GPU 및 TPU를 제공하여 딥러닝을 효율적으로 수행할 수 있음
- Google Driver와 통합 되어 데이터를 쉽게 저장하고 공유할 수 있음