AhnLab CloudMate Google Cloud

바로 도전할 수 있는 멀티 AI 에이전트 구현 Vertex AI Agent Builder

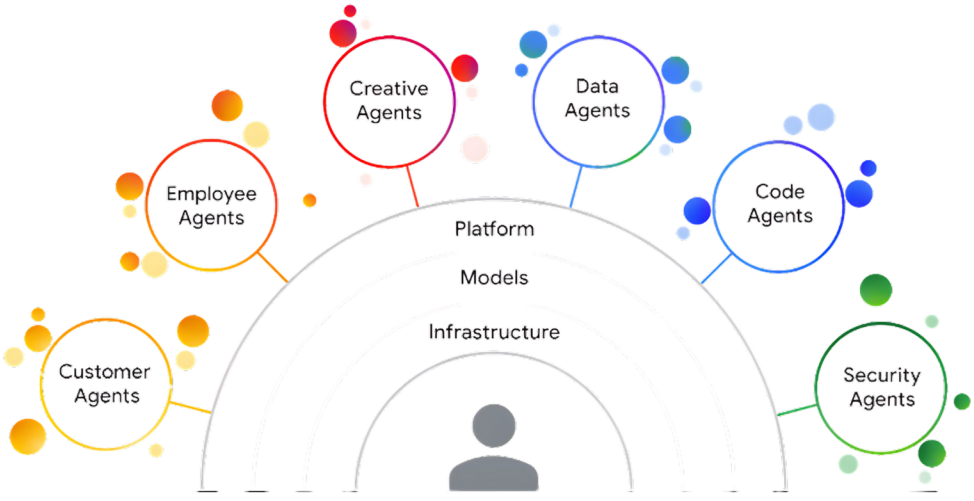
PART 1. AI 에이전트 시대

LLM(Large Language Model)과 특정 산업 또는 업무에 최적화된 SLM(Small Language Model)의 비즈니스 활용 가치가 지속적으로 높아지고 있습니다. 최근 많은 조직이 생성형 AI(Generative AI)를 2025년 주요 과제이자 목표로 설정한 이유는 무엇일까요? 디지털 시대 기업의 핵심 자산인 엔터프라이즈 컴퓨팅 환경과 오랜 기간 쌓아온 데이터의 가치를 한 층 높일 수 있기 때문입니다. 이런 이유로 디지털 전환(DX)의 완성을 AI 전환(AX)로 보는 이들이 많습니다.

AI 에이전트가 활약하는 '에이전트 경제' 시대

DX가 방점을 찍는 시점을 AX가 본격화되는 것이라고 보는 것은 기업이나 공공 모두 마찬가지입니다. 관련해 일각에서는 에이전트 경제(Agent Economy) 시대가 오고 있다고 전망하기도 합니다. 이런 분위기는 올 상반기에 열린 Google Cloud Next 24(이하 Next 24)에서도 뚜렷하게 나타났습니다. Next 24에서 구글 클라우드는 고객 지원, 직원 생산성, 마케팅, 데이터, 코딩, 보안 관련 에이전트를 소개하였습니다. 이들 에이전트는 단순한 프롬프트 기반 챗봇과 차원이 다릅니다. 다음처럼 복잡한 작업을 스스로 처리할 수 있는 AI 에이전트입니다.

- **Customer Agent:** 챗봇을 넘어 개인 맞춤형 AI 고객 성공 매니저로 기능하며, 고객의 온보딩, 개인화 기반의 응대 및 서비스 이용 안내 등의 업무를 지원합니다.
- **Employee Agent:** 사용자는 Google Workspace 환경에서 Gemini로 이메일 작성이나 슬라이드 작성 시 AI의 도움을 받을 수 있습니다.
- **Creative Agent:** 마케팅 컨텐츠를 만들 때 이미지를 생성하고 텍스트를 라이브 이미지로 시각화하는 등의 기능을 활용할 수 있습니다.
- **Data Agent:** 구글의 데이터 분석 및 관리 툴의 사용법을 잘 몰라도 Gemini로 각종 분석과 시각화를 수행할 수 있습니다.
- **Code Agent:** 코드를 자동으로 완성하거나 코드 블록을 생성하여 개발 생산성을 높입니다.
- **Security Agent:** 침해 탐지와 대응 과정에서 위협 인텔리전스를 참조해 공격자의 전략, 전술, 절차(TTP)를 분석하고 각종 탐지 룰을 만드는 등의 작업을 AI가 돕습니다.



PART 1. AI 에이전트 시대

구글 클라우드는 AI 에이전트 시대를 앞당기기 위해 생성형 AI 관련 다양한 서비스를 제공합니다.
대표적인 것으로 Gemini, Model Garden, Vertex AI, Cloud Run, AI Hypercomputer를 꼽을 수 있습니다.

구글 클라우드의 AI 관련 서비스는 기업의 다양한 요구를 수용할 수 있습니다. 생성형 AI 프로젝트를 수행하고자 하는 조직은 각자 여건이 다릅니다. 흔히 말하는 AI 성숙도가 다릅니다. GPU 기반 컴퓨팅 인프라와 MLOps 플랫폼 구축 여부, 내부 인력의 AI 역량, 각 조직이 다루고 있는 문제, 그리고 제품이나 서비스의 요구사항도 서로 천차만별입니다. 이에 따라 조직의 현재 AI 성숙도를 고려해 생성형 AI 프로젝트 수행 방향을 잡아야 합니다.

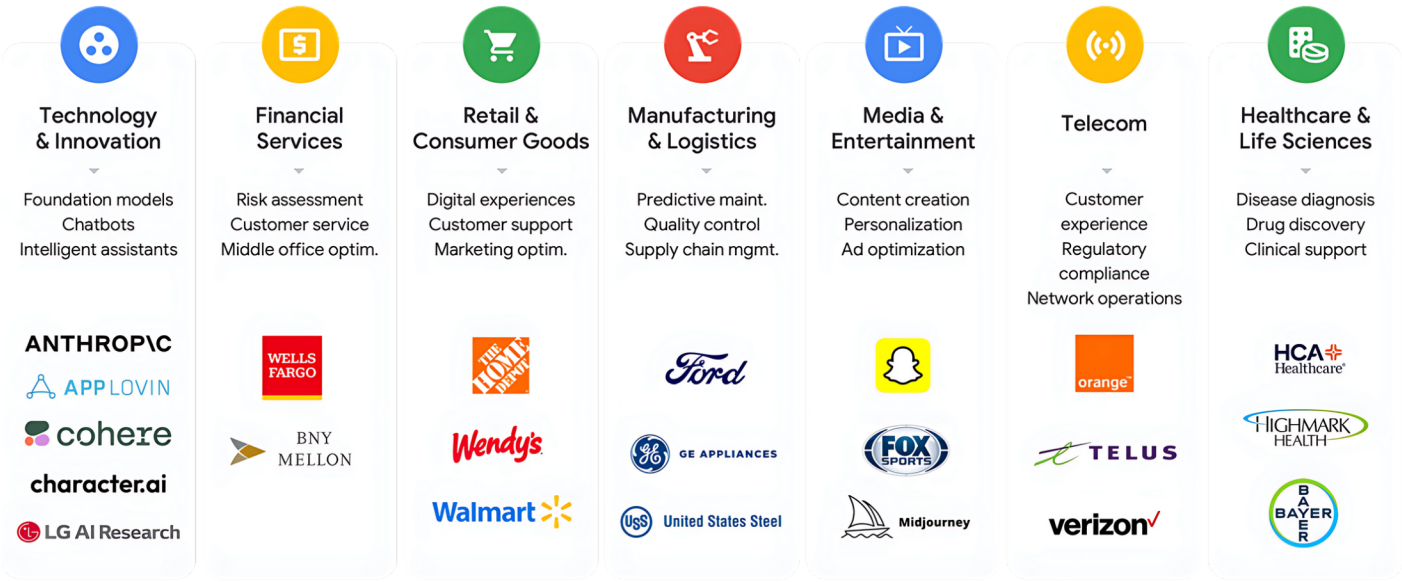
만약 AI 인프라, 플랫폼, 운영 역량을 충분히 갖추지 못했다면? Vertex AI 플랫폼에서 API를 활용해 Gemini 모델을 사용하는 방법을 고려해볼 수 있습니다. 이를 통해 개념 검증(Proof of Concept)부터 실제 프로젝트 수행까지 손쉽고 빠르게 진행할 수 있습니다.

개발 역량은 충분하지만 AI 인프라와 MLOps 플랫폼 구축과 운영에 자신이 없다면, 서버리스 환경인 Cloud Run에서 GPU 자원을 이용해 Model Garden에서 원하는 모델을 선택해 Vertex AI 플랫폼에서 작업을 하는 것을 고려할 수 있습니다. 서버리스 환경이므로 인프라 관리 부담이 적고, IaaS를 이용하는 것보다 비용 효율적이어서 개발에 더 집중할 수 있는 장점이 있습니다.

온프레미스에 GPU 클러스터와 MLOps 플랫폼 그리고 운영 인력까지 보유하고 있다면, 대기업이나 연구소 같은 경우 하이브리드 클라우드 전략의 일환으로 AI Hypercomputer를 활용하는 방법을 고려할 수 있습니다.

산업별 AI 에이전트의 활약상

AI 에이전트가 실제 비즈니스에 어떤 변화를 불러올 수 있을까요? 구글 클라우드와 AI 기술을 활용한 산업별 사례를 통해 AI 에이전트가 어떻게 비즈니스 현장에서 활약하고 있는지 살펴볼 수 있습니다.



PART 1. AI 에이전트 시대

산업별 시나리오를 정리하면 다음과 같습니다. 수년간 언론이나 각종 보고서를 통해 본 시나리오와 크게 다르지 않게 느낄 수 있습니다.그러나 중요한 차이점은 구현과 실행 방식에 있습니다. 언급하는 시나리오는 AI 에이전트가 단순히 지원하는 역할을 넘어, 스스로 복잡한 작업을 직접 처리할 수 있다는 점입니다. 이 부분에 대해서는 Part 2에서 더 자세히 살펴보겠습니다.

- **금융 서비스(Financial Services):** 금융 산업에서는 생성형 AI를 활용해 리스크 평가와 고객 지원을 최적화하고 있습니다. 금융 서비스에서는 AI가 대출 신청자의 데이터를 바탕으로 상환 가능성을 평가하고, 실시간으로 고객의 요구에 맞춘 상담을 제공할 수 있습니다. 가령 AI 기반의 챗봇이 대출 신청 및 금융 상품 추천을 자동화해 상담 시간을 줄이고, 고객에게 더 나은 서비스를 제공하는 시나리오를 떠올릴 수 있습니다.

- **소매 및 소비재(Retail & Consumer Goods):** 소매 업계는 생성형 AI로 소비자의 쇼핑 경험을 개선하고 있습니다. AI 기반 추천 시스템은 고객의 구매 기록과 취향을 분석해 관련 상품을 추천합니다. 또한, 챗봇을 통한 고객 응대, AI가 생성한 마케팅 컨텐츠, 프로모션 전략 수립 등 다양한 분야에 생성형 AI가 활약하고 있습니다.

- **제조 및 물류(Manufacturing & Logistics):** 제조업에서는 AI로 공정 및 제품 품질 관리를 자동화하고 있습니다. 예지 기반 유지보수는 제조업의 대표적 AI 도입 사례입니다. 기계 상태를 실시간 모니터링하고 유지보수가 필요한 시점을 예측하여 장비 고장을 방지합니다. 또한, 물류 업계에서는 AI가 공급망 데이터를 분석하여 재고를 최적의 상태로 유지하고, 물류 비용을 절감에 기여하고 있습니다.

- **미디어 및 엔터테인먼트(Media & Entertainment):** 미디어와 엔터테인먼트 산업에서는 AI를 활용해 컨텐츠 제작과 개인화를 강화하고 있습니다. 창작 관련 작업을 처리하는 AI 에이전트는 영상 편집과 이미지 생성에 AI를 활용하여 제작 시간을 단축하고, 소비자 데이터에 기반한 맞춤형 컨텐츠를 추천하여 사용자 경험을 개선합니다.

- **통신(Telecom):** 고객 데이터 분석과 고객 경험 맞춤화에서 생성형 AI가 활용됩니다. 가령 AI 기반의 고객 지원 에이전트가 고객의 요구에 신속히 반응하고, 맞춤형 요금제 추천과 같은 개인화된 서비스를 제공합니다. 또한, 네트워크 운영도 AI를 통해 트래픽 패턴을 분석하고, 네트워크 장애를 예측하며 유지보수 계획을 세울 수 있습니다.

- **헬스케어 및 생명과학(Healthcare & Life Sciences):** 생성형 AI는 질병 진단과 신약 개발에 중요한 역할을 하고 있습니다. 의료 부문에 특화된 LLM이나 SLM은 방대한 의료 데이터를 분석하여 질병의 초기 징후를 찾아내고, 임상 시험 데이터를 통해 신약 후보 물질을 예측합니다. 또한, 환자 기록을 바탕으로 맞춤형 치료법도 제안합니다.

PART 2. LLM 앱 아키텍처

구글 클라우드의 Gemini 같은 LMM(Large Multi Modal Model)이나 LLM, SLM을 기반으로 하는 앱 아키텍처는 빠르게 진화의 과정을 거치고 있습니다. 참고로 LMM은 LLM의 언어 처리 능력에 시각, 청각 등 다양한 모달리티를 통합하여 더 복잡하고 풍부한 데이터를 다룰 수 있는 모델입니다. LMM, LLM, SLM 앱 개발의 주요 관심사는 최근 빠르게 변하고 있습니다. 2023년까지는 주로 싱글 프롬프트와 프롬프트 체이닝이 중심이었고, 2024년의 관심사는 AI 에이전트로 전환되었습니다. 다가오는 2025년에는 멀티 AI 에이전트 구현까지 예상보다 더 빠르게 현실화될 가능성이 커지고 있습니다.

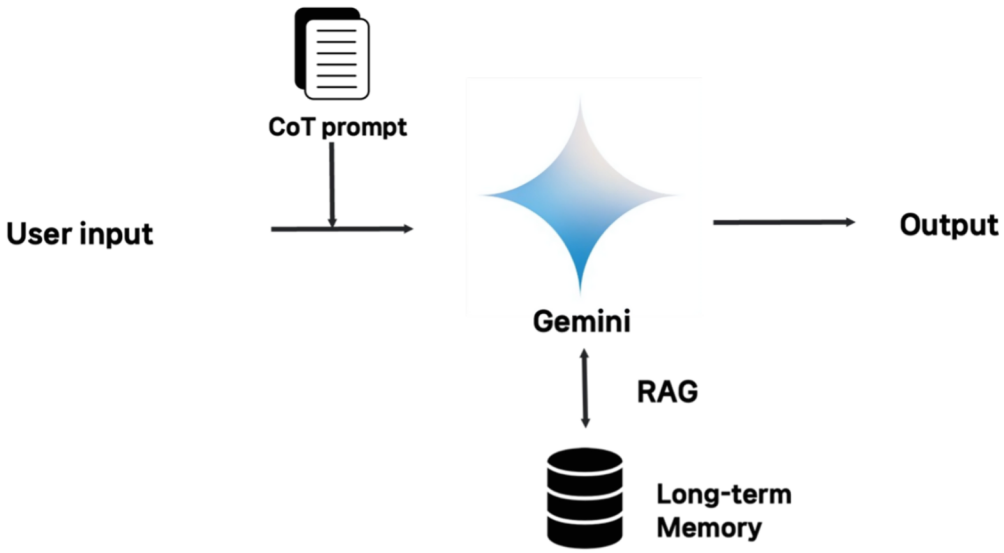
안랩클라우드메이트는 사내 PoC(Proof of Concept) 프로젝트로 Vertex AI Agent Builder와 LMM 기반 Gemini로 멀티 AI 에이전트를 구현하였습니다. 그 결과 구글 클라우드를 활용하면 빠르고 효율적으로 멀티 AI 에이전트가 서로 협업하는 환경을 구현할 수 있다는 가능성을 확인하였습니다. 이에 대해서는 Part 3에서 자세히 다루겠습니다.

LMM, LLM, SLM 앱 아키텍처의 진화

LMM, LLM과 SLM 앱 아키텍처는 싱글 프롬프트, 프롬프트 체이닝, 싱글 AI 에이전트, 멀티 AI 에이전트로 발전하고 있습니다. 아키텍처의 진화가 이전 방식을 대체하는 것이 아니라, 비즈니스 요구와 작업 성격에 맞게 다양한 아키텍처를 활용할 수 있다는 것입니다. 각 아키텍처의 특징과, 복잡한 아키텍처 구현에서 발생하는 어려움에 대해 살펴보겠습니다.

싱글 프롬프트

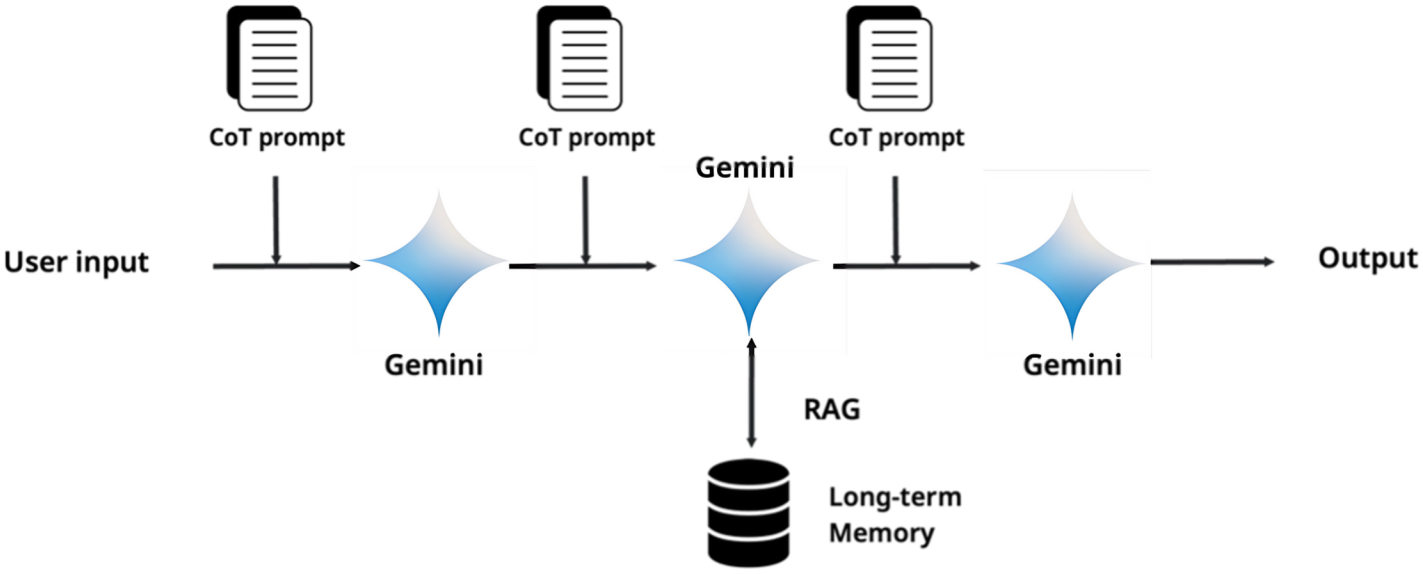
사용자의 질문이나 요청에 대해 단일 응답을 제공하는 방식입니다. 하나의 질문에 대해 AI가 즉시 반응하여 응답을 제공합니다. 빠른 응답이 필요할 때 적합합니다.



PART 2. LLM 앱 아키텍처

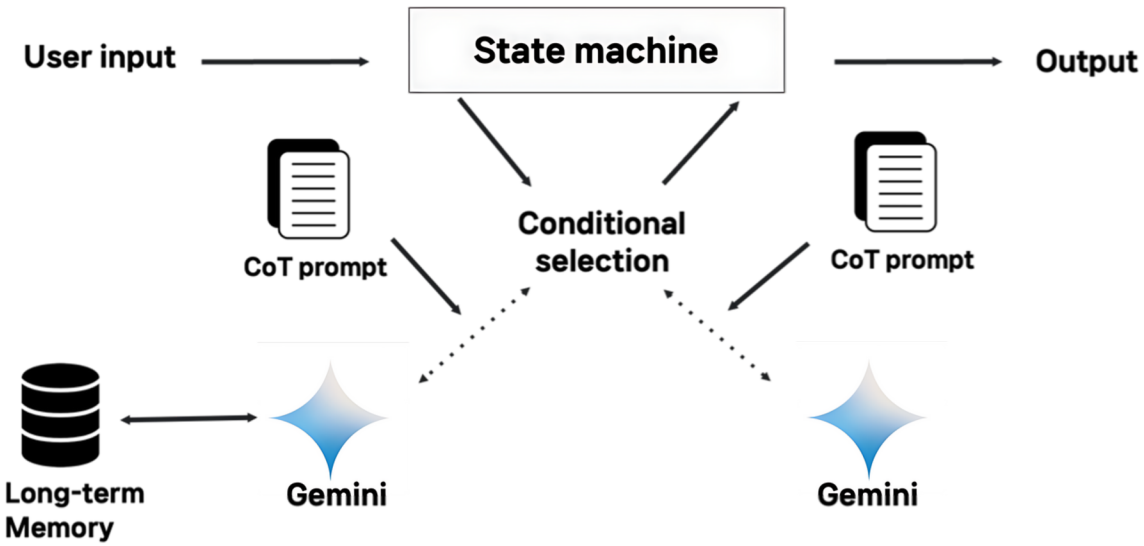
프롬프트 체인닝

하나의 질문에 하나의 답변을 주는 싱글 프롬프트 방식으로 시작해 여러 질문과 응답을 연쇄적으로 연결하여 복잡한 요청을 처리하는 구조입니다. 일련의 작업을 수행하는 방식으로 단계적 문제 해결이 필요한 경우 유용합니다.



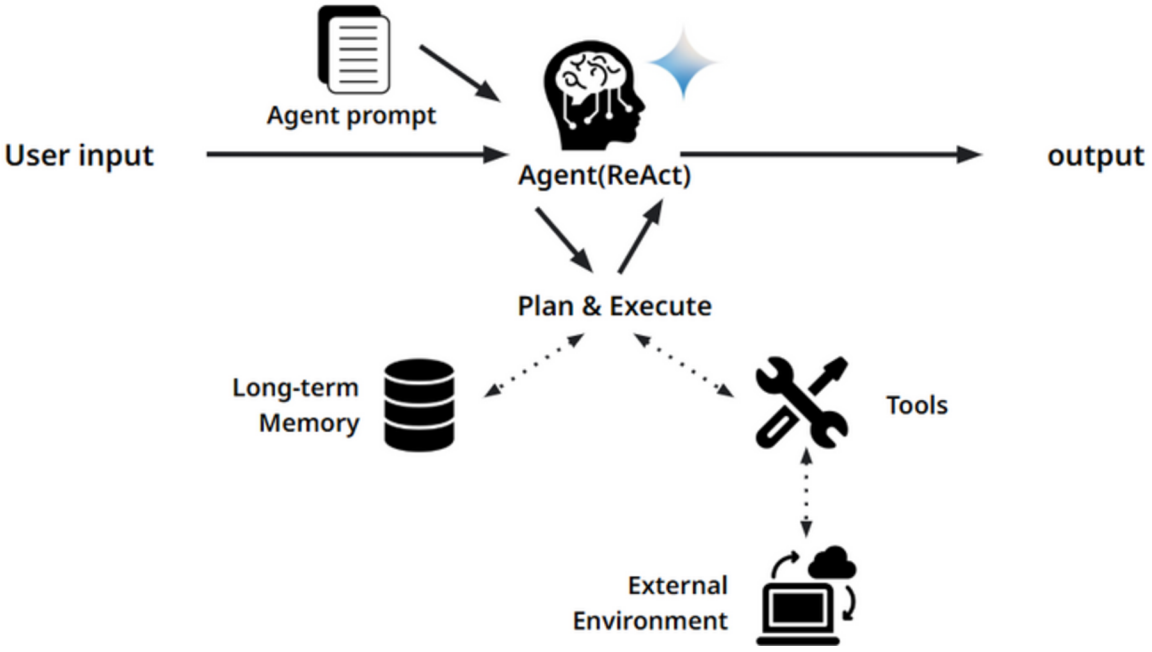
싱글 AI 에이전트 (state machine)

하나의 AI 에이전트가 특정한 역할을 수행하며, 주어진 요청을 처리하기 위해 데이터에 접근하거나 툴을 사용해 작업을 수행하는 방식입니다.



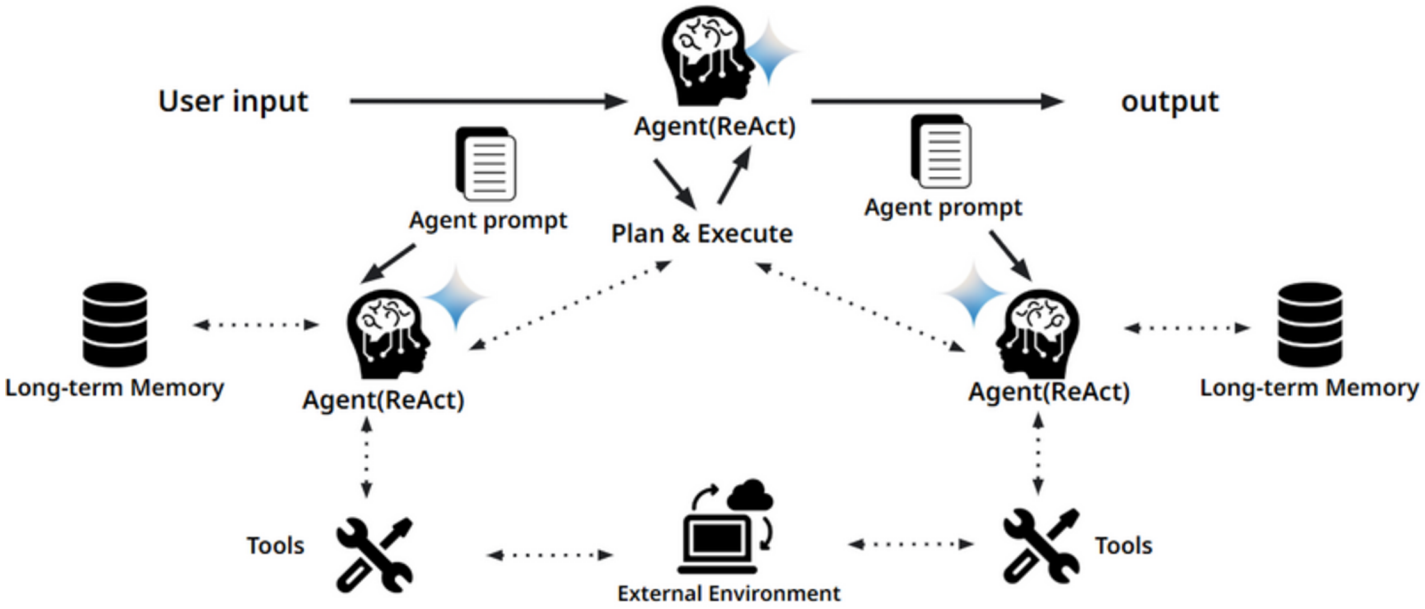
PART 2. LLM 앱 아키텍처

싱글 AI 에이전트 (Single agent(ReAct, Plan & Execte)



멀티 AI 에이전트

하나의 질문에 하나의 답변을 주는 싱글 프롬프트 방식으로 시작해 여러 질문과 응답을 연쇄적으로 연결하여 복잡한 요청을 처리하는 구조입니다. 일련의 작업을 수행하는 방식으로 단계적 문제 해결이 필요한 경우 유용합니다.



PART 2. LLM 앱 아키텍처

멀티 AI 에이전트 구현의 어려움

싱글 AI 에이전트에서 멀티 AI 에이전트로의 전환은 여러 면에서 복잡성을 동반합니다. 멀티 AI 에이전트 시스템은 여러 개의 에이전트를 조정하고 상호작용을 관리해야 합니다. 각 에이전트가 어떤 데이터를 다루고 어떤 작업을 수행할지 결정하는 오케스트레이션이 필요합니다.

다음으로 멀티 AI 에이전트는 스테이트 머신과 같은 복잡한 로직이 필요합니다. 이러한 시스템은 디버깅과 유지보수가 어려울 수도 있습니다. 보안도 부담입니다. 여러 AI 에이전트가 서로 다른 권한과 역할을 가지고 다양한 데이터와 도구에 접근하는 멀티 AI 에이전트 환경은 보안 설정과 접근 제어가 매우 중요합니다.

사실 이런 보안 부담은 기업이 멀티 AI 에이전트를 도입해야 하는 이유이기도 합니다. 데이터 접근과 기능 수행 권한을 분리하면 보안성을 강화할 수 있습니다. 즉, 특정 데이터나 도구에 접근 권한이 있는 AI 에이전트만 해당 데이터를 처리할 수 있도록 설정하면 보안 침해를 예방하는 데 효과적입니다.

이 외에도 트레이싱의 어려움도 있습니다. 여러 AI 에이전트가 서로 연결된 복잡한 시스템에서는 오류 발생 시 원인을 파악하기 어려워집니다. 이를 해결하기 위해 AI 에이전트의 수행 과정을 추적하고, 이를 토대로 새로운 데이터나 피드백을 반영해 모델을 주기적으로 업데이트하고, 필요시 재학습이나 파인튜닝을 하여 정확성을 높여야 합니다.

PART 3. 멀티 AI 에이전트 POC

이번 파트에서는 안랩클라우드메이트가 Vertex AI Agent Builder로 멀티 AI 에이전트를 구현한 PoC 프로젝트 경험을 공유하겠습니다. 앞서 살펴본 바와 같이 LMM, LLM, SLM 활용의 궁극적 미래인 멀티 AI 에이전트 구현은 기술과 운영 관련 난제를 해결해야 하는 어려운 과제입니다. 이런 난이도 높은 작업을 구글 클라우드는 당장 실천으로 옮길 수 있는 목표로 만듭니다. 안랩클라우드메이트는 사내 PoC 프로젝트로 Vertex AI Agent Builder를 활용하면 멀티 AI 에이전트 구현을 수월하게 할 수 있다는 것을 직접 확인하였습니다.

다재다능한 Vertex AI Agent Builder

Vertex AI Agent Builder는 사용자 편의성이 높은 도구입니다. 이를 사용하면 다양한 도구와 API를 통합할 수 있습니다. 이를 통해 사용자는 AI 에이전트가 필요한 데이터를 자동으로 수집하고, 다양한 외부 도구나 서비스와 연결할 수 있습니다.

RAG 기반 검색도 편합니다. Vertex AI Agent Builder의 RAG 기능으로 실시간으로 데이터를 검색하고, 관련 정보를 AI 모델의 답변에 포함할 수 있습니다.PDF문서 같은 비정형 문서의 데이터 처리도 Layout Parser를 통해 원활하게 할 수 있습니다.

또한, 코드 작성 없이도 AI 에이전트의 역할을 정의하고, AI 에이전트를 설정할 수 있습니다. 이런 특징 덕분에 안랩클라우드메이트는 Vertex AI Agent Builder로 다음과 같이 세 가지 AI 에이전트와 이들 세 가지 AI 에이전트에 지시를 내리고 유저를 연결하는 supervisor AI 에이전트를 만들었습니다. 그리고 이들을 서로 협력하게 하여 주식 투자자들을 도와주는 멀티 AI 에이전트 시스템을 수월하게 만들 수 있었습니다.

- **뉴스 분석 AI 에이전트:** 뉴스 정보를 수집하고 분석하여, 주가에 긍정적 또는 부정적인 영향을 줄 수 있는 뉴스를 사용자에게 제공합니다.
- **재무 분석 AI 에이전트:** 회사의 재무 상태를 분석하여, 사업 보고서나 분기 보고서 등 공시 정보를 바탕으로 회사의 가치를 평가하는 역할을 합니다.
- **주식 브로커 AI 에이전트**: 사용자의 주식 거래 요청에 따라 매수/매도 주문을 실행합니다.

좀 더 자세히 멀티 AI 에이전트 구현 과정을 살펴보겠습니다.

PART 3. 멀티 AI 에이전트 POC

1 데이터 적재

먼저 AI 에이전트가 처리할 비정형 문서 데이터를 Google Cloud Storage에 업로드합니다. 이 데이터에는 주로 PDF, DOCX, HTML 등 다양한 문서 형식이 포함될 수 있습니다.참고로 안랩클라우드메이트는 PoC를 위해 한국거래소의 모든 상장사 보고서(2023) 문서를 데이터로 활용했습니다.

2 DATASTORE 생성

Vertex AI Agent Builder에서 DataStore를 생성하여 구글 클라우드 스토리지에 저장된 데이터를 구조화합니다. DataStore는 비정형 데이터를 인덱싱하고 검색할 수 있는 구조로 변환하여 AI 에이전트가 효율적으로 정보를 찾고 응답할 수 있도록 합니다.

이 단계에서 주목할 것은 구글 클라우드가 제공하는 문서 파서인 Layout Parser입니다. Layout parser는 google의 document AI의 기술을 활용하여 문서의 레이아웃을 분석하여 텍스트 이미지 테이블 등의 정보를 의미단위로 추출합니다.

이렇게 추출한 데이터를 활용하면 AI 에이전트가 다양한 작업을 수행할 수 있습니다. 안랩클라우드메이트는 한국 조직의 다양한 문서 형식, 특히 표가 많은 문서에 대한 Layout Parser의 적용 가능성을 PoC에서 살펴보았습니다. 결과는 기대 이상이었습니다. 기존 파서와 비교해 표 내용에 대한 이해가 높았고, 특히 기존 파서가 인식하지 못하는 병합된 셀 처리도 정확하다는 것을 확인하였습니다.

3 AI 에이전트 생성

Agent Builder에서 AI 에이전트를 생성합니다. 이 AI 에이전트는 사용자 요청에 따라 데이터를 검색하고, 분석하며, 그에 맞는 정보를 제공하는 역할을 수행합니다. 그 뿐만이 아니라 주문, 이메일 전송 등과 같이 유저를 대신하여 도구를 이용한 액션 또한 수행할 수 있습니다. AI 에이전트는 특정한 역할과 목표를 설정할 수 있으며, 다양한 질문에 대한 응답을 생성하거나 검색하는 작업을 자동화하고 외부환경과 상호작용 할 수 있습니다. PoC와 같이 우리가 보유한 데이터를 검색하고 분석하는 재무 분석 AI 에이전트와 오픈 API와 구글 검색을 활용하여 외부 뉴스 정보를 검색하고 분석하는 뉴스 AI 에이전트로 검색 유형을 나누어 설정할 수 있습니다.

PART 3. 멀티 AI 에이전트 POC

4 DATASTORE 툴 생성

AI 에이전트가 DataStore와 상호작용을 할 수 있도록 DataStore 툴을 생성합니다. 이 툴은 AI 에이전트가 DataStore에 접근하여 필요한 데이터를 검색하고, 추출, 변환, 처리하여 응답에 활용할 수 있도록 지원합니다. DataStore 툴을 통해 AI 에이전트는 다양한 데이터 소스에서 정보를 불러오거나, 특정 조건에 따라 데이터를 필터링하여 사용자의 질문에 맞는 최적의 답변을 제공할 수 있습니다.

5 DATASTORE 툴과 DATASTORE 연결

생성한 DataStore 툴을 두 번째 단계에서 만든 DataStore와 연결합니다. 이를 통해 AI 에이전트는 데이터 소스에 직접 접근하여 관련 정보를 검색하고, 필요한 데이터를 빠르게 수집하여 작업의 효율성을 높일 수 있습니다.

6 AI 에이전트가 DATASTORE 툴을 사용하도록 활성화

DataStore 툴을 AI 에이전트의 주요 데이터 수집 도구로 활성화합니다. AI 에이전트는 DataStore 툴을 통해 DataStore에 저장된 데이터를 실시간으로 조회하고, 사용자가 요청한 정보를 신속하게 찾을 수 있습니다.

7 프롬프트 작성

AI 에이전트가 사용자의 요청에 따라 정보를 검색하고 응답할 수 있도록 프롬프트를 설계합니다. 프롬프트는 AI 에이전트가 수행해야 할 작업을 명확히 규정합니다. 예를 들어 “A 기업에 대한 최신 뉴스 분석을 제공해 주세요”와 같은 프롬프트를 통해 AI 에이전트가 DataStore 툴을 사용하여 관련 뉴스를 검색하고 응답을 준비할 수 있습니다. 참고로 잘 설계된 프롬프트는 AI 에이전트의 응답 정확도와 효율성을 높여줍니다. 이는 특히 멀티 AI 에이전트 시스템에서 각 에이전트의 역할을 명확히 구분하는 데 중요합니다.

PART 3. 멀티 AI 에이전트 POC

8 테스트

각 AI 에이전트와 DataStore 툴이 올바르게 작동하는지 테스트합니다. 이는 AI 에이전트가 예상한 대로 DataStore에서 데이터를 검색하고 응답을 생성하는지 확인하는 과정입니다.

테스트 과정에서 AI 에이전트가 올바른 응답을 생성했는지 기대했던 도구를 사용하는지 등을 확인하기 위해 Agent Builder의 테스트 기능인 debog trace와 Original Response 기능등을 이용해 Gemini의 응답을 구체적으로 확인해 볼 수 있습니다. 또한 conversation history 기능을 적절히 활용하면 테스트와 평가를 손쉽게 할 수 있습니다.

테스트와 평가 결과에 따라 프롬프트를 변경하고 다시 테스트와 평가를 하는 과정을 반복하면서 멀티 에이전트 시스템을 지속적으로 개선합니다.

9 배포

테스트가 완료된 AI 에이전트를 실제 사용자 환경에 배포하여 실시간으로 사용할 수 있도록 합니다.
Vertex AI Agent Builder는 slack, facebook messenger등의 다양한 서드파티와의 통합 기능으로 별도의 앱을 개발하지 않고도 실제 유저가 사용할 수 있게 배포 할 수 있습니다. 만약 별도의 앱을 개발하고 있다면 Environment를 추가하여 웹훅을 이용해 개발하는 앱과 연동하여 배포할 수 있습니다. 배포 후에는 모니터링을 해야 합니다. 이를 통해 에이전트의 성능을 실시간으로 모니터링하여 필요한 경우 유지보수와 업데이트를 진행합니다. 이 외에도 배포 후에는 사용자 피드백을 수집하여 AI 에이전트의 성능을 개선할 수 있으며, 새로운 데이터를 DataStore에 추가하여 최신 정보를 반영하도록 할 수 있습니다.

AhnLab CloudMate Google Cloud

web: ahnlabcloudmate.com
e-mail: mktg@ahnlabcloudmate.com
Tel: 02-2069-1980