디지털서비스

# 이슈리포트

Digitalservice Issue Report

# AI PaaS의 필요성과 글로벌 AI PaaS 동향

NIA 한국지능정보사회진흥원 디지털서비스 이용지원시스템

2025 Vol. 09

디지털서비스 이슈리포트

# CONTENTS

Digitalservice Issue Report

## 디지털서비스 이슈리포트

# 01 AI PaaS의 필요성과 글로벌 AI PaaS 동향

| 김혜영 한국지능정보사회진흥원 책임연구원
| 정기봉 (주)이노그리드 이사

## PaaS의 등장과 역할

PaaS(Platform as a Service)는 애플리케이션 개발과 배포에 필요한 플랫폼과 개발도구를 클라우드 환경에서 서비스의 형태로 제공하는 컴퓨팅 모델이다. 이를 통해 개발자는 인프라 구축과 관리에 신경 쓰지 않고 애플리케이션 개발에만 집중할 수 있어 개발 생산성과 효율성을 크게 향상시킬 수 있다.

2010년대 중반 이후 핀테크, 이커머스, 게임 산업이 급성장하면서 빠른 서비스 출시와 확장성이 기업 경쟁력의 핵심 요소로 부상했다. 이러한 시장 수요에 대응해 AWS, 구글, 마이크로소프트와 같은 글로벌 빅테크 기업들은 클라우드 플랫폼을 본격적으로 서비스화하며 시장 표준화를 주도했다.

이 과정에서 전통적으로 자체 전산실(IDC)을 구축·운영하며 보수적인 접근을 선호하던 대기업과 금융권조차도 변화의 압력을 체감하게 되었다. 특히, 빠른 개발·배포 속도가 비즈니스의 성패를 좌우하고, 대규모 트래픽 상황에서도 안정적인 서비스 운영을 보장해야 하는 요구가 커지면서 클라우드 기반 PaaS 전환은 선택이 아닌 필수로 자리 잡았다.

이러한 변화 속에서 DevOps는 속도와 안정성 요구를 동시에 충족할 수 있도록 지원하는 핵심 방법론으로 자리 잡았으며, PaaS는 이를 구현하는 최적의 환경을 제공했다. 지속적 통합/배포(CI/CD)를 통해 개발 속도를 높이는 동시에, 자동화된 테스트와 모니터링 기능으로 서비스 안정성을 보장할 수 있었기 때문이다. 특히 컨테이너 기반의 자동 스케일링 기술은 예상치 못한 트래픽 급증 상황에서도 서비스 중단 없이 대응할 수 있도록 지원하여, 기업의 생존 전략으로도 부상했다.

국내의 경우, PaaS의 등장이 단순한 글로벌 트렌드의 수용하는 수준을 넘어 정부의 강력한 클라우드 기반 디지털 전환 정책과 한국 특유의 치열한 서비스 경쟁 환경이 맞물린 결과라고 할 수 있다. 정부가 시행한「클라우드 컴퓨팅 발전법」에 따른 공공부문 클라우드 서비스 우선 도입 정책은 국내 PaaS 시장의 초기 수요를 창출하고 기술 도입의 정당성을 부여한 중요한 정책적 기반이 되었다.

이 같은 시장 환경 변화에 부응하여 정부는 개방형 클라우드 플랫폼인 K-PaaS(Korean Platform as a Service)를 개발·보급했다. K-PaaS는 오픈소스 기반의 클라우드 파운드리(Cloud Foundry)기술을 활용해 국내 보안·규제 환경에 최적화된 국산 PaaS 플랫폼으로, 공공기관과 기업이 안전하고 효율적으로 클라우드 환경을 구성할 수 있도록 지원했다.

아울러 네이버클라우드, KT클라우드, NHN클라우드 등 주요 국내 CSP들은 해외 사업자가 충족하기 어려운 규제·보안 요구사항에 최적화된 PaaS 상품을 출시하며 시장 확산을 가속화했다. 이로써 국내 PaaS는 정책적 지원, 기술적 혁신, 산업계 수요가 삼박자를 이루며 본격적인 성장 궤도에 오르게 되었다.

**〈오픈소스 기반 개방형 클라우드 플랫폼(K-PaaS) 아키텍처〉**


그림 1 출처 : 개방형 클라우드 플랫폼(K-PaaS) 센터

## PaaS에서 AI PaaS로의 변화

과거에는 AI가 연구 목적이나 일부 특화된 산업 영역에서만 제한적으로 활용되었으나, 대규모 언어모델(LLM)이 일반 대중도 쉽게 사용할 수 있는 보편적 서비스로 자리 잡으면서 AI 활용 수요가 폭발적으로 증가하였다.

특히, 2022년 ChatGPT의 등장은 전 세계적으로 생성형 AI 개발 경쟁을 촉발했다. 마이크로소프트는 OpenAI와의 전략적 파트너십을 통해 Azure 클라우드에 AI 기능을 통합하여 '코파일럿(Copilot)' 서비스를 다양한 소프트웨어에 적용하고 있다. 구글은 제미나이(Gemini)라는 멀티모달 AI 모델을

개발하여 검색, 클라우드, 생산성 도구에 접목하고 있으며, 메타는 라마(Llama) 시리즈를 오픈소스로 공개하고 SNS 플랫폼과 연계한 AI 마케팅 도구를 강화하고 있다.

이들 글로벌 빅테크 기업들의 공통점은 단순히 AI 모델 개발에 그치지 않고, 기존의 플랫폼과 인프라에 AI 기능을 통합하여 사용자가 손쉽게 AI 서비스에 접근하고 활용할 수 있는 환경을 구축하고 있다는 점이다.

따라서 기존의 DevOps 중심의 PaaS로는 한계가 있으며, 이제는 AI 서비스 개발과 운영의 전 생애주기를 지원할 수 있는 기능을 갖춘 AI PaaS로의 진화가 클라우드 업계의 핵심 트렌드로 자리 잡고 있다. 이에 따라 AI PaaS에서는 다음과 같은 대표적인 기능들이 요구된다.

### 1. 데이터 운영(DataOps)

- **데이터 수집 및 통합** : AI 모델 학습에 필요한 데이터를 다양한 소스에서 자동으로 수집하고 통합하는 과정이다. API, 데이터베이스, 실시간 스트리밍 등을 통해 데이터를 수집하며, 메타데이터 관리와 데이터 계보 추적을 통해 데이터의 출처를 투명하게 관리한다.
- **데이터 품질 관리** : AI 모델의 성능과 신뢰성을 보장하기 위한 데이터 품질 확보 과정이다. 스키마 검증, 이상값 탐지, 중복 제거, 결측값 처리, 편향성 검사 등을 자동화하여 일관된 품질의 데이터를 확보한다.
- **데이터 전처리 및 피처 엔지니어링** : 원시 데이터를 ML 모델이 학습할 수 있는 형태로 변환하는 과정이다. 데이터 정규화, 인코딩, 피처 추출 및 선택, 데이터 증강 등의 작업을 통해 모델 훈련에 최적화된 데이터를 생성한다.
- **데이터 버전 관리** : 데이터셋의 변경 사항을 추적하고 재현 가능한 실험 환경을 제공하는 과정이다. 데이터셋 스냅샷, 변경 이력 추적, 브랜치 관리, 롤백 기능을 통해 데이터의 버전을 체계적으로 관리한다.

### 2. 머신러닝 운영(MLOps)

- **실험관리** : ML 실험의 체계적 관리 및 재현성을 보장하는 과정이다. 하이퍼파라미터, 메트릭, 결과를 로깅하고, 여러 모델의 성능을 비교하며, 실험 결과를 팀 간에 공유할 수 있는 환경을 제공한다.
- **모델 훈련 및 최적화** : 효율적이고 최적화된 모델 훈련 환경을 제공하는 과정이다. 분산 훈련, 하이퍼파라미터 튜닝, AutoML, 조기 종료, 체크포인트 관리 등을 통해 최적의 모델을 개발한다.
- **모델 검증 및 테스트** : 훈련된 모델을 프로덕션 환경에 안전하고 효율적으로 배포하는 과정이다. Blue-Green, Canary, Rolling 배포 패턴을 활용하여 REST API, gRPC, 배치 추론 등 다양한 형태로 모델을 서빙한다.

- **모델 배포** : 모델의 성능과 안정성을 다각도로 검증하는 과정이다. 성능 평가, 교차 검증, A/B 테스트, 편향성 테스트, 적대적 테스트를 통해 모델의 품질을 보장한다.
- **모델 모니터링** : 프로덕션 환경에서 모델 성능과 동작을 지속해서 감시하는 과정이다. 실시간 성능 추적, 데이터 드리프트 탐지, 모델 드리프트 탐지, 이상 탐지, 알람 시스템을 통해 모델의 안정성을 보장한다.
- **모델 재훈련 및 업데이트** : 모델 성능 유지를 위한 자동화된 재훈련 및 업데이트 과정이다. 성능 저하 시 자동 재훈련, 정기적 모델 업데이트, 증분 학습, A/B 테스트, 롤백 메커니즘을 통해 모델의 지속적인 개선을 지원한다.

〈PaaS vs AI PaaS〉

- **아키텍쳐 관점**
  기존 PaaS가 CPU 중심의 범용 컴퓨팅 환경을 제공했다면, AI PaaS는 GPU, TPU, NPU 등 AI 연산에 특화된 하드웨어를 중심으로 설계된다. 스토리지 역시 기존의 관계형 데이터베이스나 NoSQL에서 벗어나 대용량 데이터 처리를 위한 데이터 레이크, 벡터 검색을 위한 벡터 데이터베이스, 분산 파일 시스템으로 확장된다. 네트워킹 또한 단순한 HTTP/REST API를 넘어 대용량 모델 서빙과 실시간 스트리밍 추론을 위한 고대역폭 통신을 지원해야 한다.
- **개발 및 운영 프로세스 관점**
  기존 DevOps에 MLOps(머신러닝 운영)와 DataOps(데이터 운영)가 추가된다는 것이 가장 큰 특징이다. 코드 관리는 Git 기반의 소스코드 버전 관리를 넘어 학습 데이터셋과 AI 모델까지 함께 버전을 추적해야 한다. 빌드 프로세스도 전통적인 컴파일-패키징-배포 단계에서 데이터 전처리-모델 훈련-성능 검증-배포로 완전히 달라진다. 테스팅 역시 단위/통합/성능 테스트에 더해 모델의 정확도, 편향성, 데이터 드리프트 등 AI 특화 검증이 필요하다.
- **데이터 관리 관점**
  기존 PaaS가 주로 구조화된 비즈니스 데이터를 GB~~TB 규모로 처리했다면, AI PaaS는 텍스트, 이미지, 음성, 비디오 등 다양한 모달리티의 데이터를 TB~~PB 규모로 다뤄야 한다. 데이터 처리 방식도 CRUD 연산과 트랜잭션 처리 중심에서 대규모 배치 처리, 실시간 스트림 처리, 복잡한 피처 엔지니어링으로 확장된다. 특히 데이터 품질 관리에서도 기존의 일관성과 무결성 확보를 넘어 AI 모델의 편향성, 훈련 데이터의 대표성, 라벨링 품질 등을 종합적으로 관리해야 한다.
- **스케일링 및 성능 관점**
  기존 PaaS의 수평적 인스턴스 확장과 달리, AI PaaS는 GPU 클러스터 기반의 특화된 확장 방식을 사용한다. 성능 지표 역시 TPS(초당 처리 건수)나 응답 시간 중심에서 모델 정확도, 추론 지연 시간, AI 서비스 처리량으로 달라진다. 비용 모델도 단순한 컴퓨팅 시간 기반에서 GPU

사용 시간과 모델 호출 횟수를 종합한 복합 과금으로 진화하고 있다.
이러한 근본적 차이로 인해 AI PaaS는 기존 PaaS의 단순한 확장이 아닌, AI 워크로드에 특화된 완전히 새로운 플랫폼 패러다임으로 자리 잡고 있다.

## 글로벌 AI PaaS 동향

글로벌 빅테크 기업(AWS, 구글, 마이크로소프트, 메타 등)들은 기존의 DevOps 중심의 클라우드 플랫을 넘어, AI 개발 전주기(DataOps, MLOps 등)를 통합 지원하는 AI PaaS로 전환을 위해 재투자 규모를 지속적으로 확대하고 있다.

AI Times에 따르면, 2025년 기준 글로벌 빅테크 4사가 AI 클라우드 인프라에 투입할 투자 규모는 3,200억 달러로, 이는 국내 정부 전체 예산의 약 3분의 2에 해당하는 수준이다. 각 사는 이러한 막대한 투자를 기반으로 저마다의 AI PaaS 발전 전략을 추진하고 있으며, 그 주요 내용은 다음과 같이 정리된다.

- **AWS(Amazon Web Services)** : 생성형 AI와 머신러닝 워크로드에 최적화된 AI PaaS 전략을 추진하고 있다. Amazon Bedrock을 통해 Anthropic, Meta, Cohere, Stability AI 등 주요 글로벌 AI 기업의 파운데이션 모델을 단일 API로 제공할 수 있도록 연계하여 손쉽게 생성형 AI 애플리케이션을 개발·배포하도록 지원하고, Sagemaker를 통해 데이터 준비·학습·배포·운영까지 엔드투엔드 ML 수명주기를 통합 관리하는 AI PaaS 플랫폼을 제공하고 있다.
- **Google cloud** : 검색·데이터 기반의 강점을 살려 데이터와 AI를 긴밀히 연결하는 AI PaaS 전략을 추진하고 있다. Vertex AI를 중심으로 160개 이상의 다양한 파운데이션 모델 제공 및 gemini 2.0을 통해 텍스트·이미지·영상·음성 데이터를 모두 처리할 수 있는 멀티 모달 AI를 제공하고 있다.
- **MS(Microsoft Azure)** : OpenAI와 전략적 파트너십을 통해 생성형 AI 경쟁에서 가장 빠르게 시장을 선점하였고, 생산성 툴 중심의 AI 내재화 전략을 추진하고 있다. Azure OpenAI Service를 통해 실시간 오디오 API 및 CUA(Compuer-Using Agernt)등 신기능을 지원하고, Copilot을 통해 Office 365, GitHub 등 전사 서비스에 AI 보조기능을 내장하는 등 생산성과 업무 자동화를 지원하고 있다.
- **Meta** : 오픈소스 모델 중심의 생태계 확장 전략을 추진하고 있다. LLama 시리즈를 통해 전세계 연구자·개발자가 무료로 대규모 언어모델을 활용할 수 있도록 지원하며, 이를 기반으로 광고, 마케팅, 메타버스, SNS 연계 등 개방형 전략으로 차별화를 표방하고 있다.

이처럼 글로벌 4사의 AI PaaS 발전 전략은 다양한 방식으로 추진되고 있으며, 이를 한눈에 비교할 수 있는 자료는 아래와 같다.

| GenAI Category | GenAI Component | Amazon Web Services | Google Cloud | Microsoft Azure |
|---|---|---|---|---|
| Foundation Models | Runtime | Amazon Bedrock | Vertex AI | Azure OpenAI |
| | Text / Chat | TBD | PaLM | GPT |
| | Code | TBD | Codey | GPT |
| | Image Generation | TBD | Imagen | DALL-E |
| | Translation | TBD | Chirp | None |
| Model Catalog | Commercial | Amazon SageMaker JumpStart<br>Amazon Titan | Vertex AI Model Garden | Azure ML<br>Foundation Models |
| | Open Source | Amazon SageMaker JumpStart<br>Hugging Face | Vertex AI Model Garden | Azure ML<br>Hugging Face |
| Vector Database | | Amazon RDS (pgvector) | Cloud SQL (pgvector) | Azure Cosmos DB<br>Azure Cache |
| Model Deployment & Inference | | Amazon SageMaker | Vertex AI | Azure ML |
| Fine-tuning | | Amazon Bedrock | Vertex AI | Azure OpenAI |
| Low-code/No-code Development | | TBD | Gen App Builder | Power Apps |
| Code Completion | | Amazon Code Whisperer | Duet AI for Google Cloud | GitHub Copilot |

그림 2 출처 : The New Stack

국내 역시 글로벌 트렌드에 발맞춰 기존 DevOps 중심의 PaaS를 기반으로 아래와 같이 AI 서비스를 잇달아 개발한 AI PaaS 서비스를 출시하고 있다.

- **네이버클라우드** : CLOVA Studio, HyperCLOVA X AI PaaS를 출시하여 PaaS 기반 AI 모델 학습/배포 환경 및 국산 LLM 기반 한국어 성능 최적화 서비스를 제공하고 있다.
- **KT클라우드** : GIGA Genie AI 서비스를 출시하여 산업 특화 AI API 및 통신 인프라 기반 실시간 처리 강점을 활용한 B2B 기업용 AI 솔루션 등을 제공하고 있다.
- **NHN클라우드** : AI EasyMaker 서비스를 출시하여 게임·커머스에 특화된 중소기업에 적합한 AI 전주기를 지원하는 개발 환경을 제공하고 있다.

## 마무리

본 보고서에서 살펴본 바와 같이, PaaS의 발전은 단순히 DevOps 중심의 애플리케이션 개발 편의성과 운영 안정성 지원을 넘어, 생성형 AI 기술의 급속한 발전과 AI 서비스 수요 급증에 대응하기 위해 지속적인 모델 학습과 고도화를 플랫폼 차원에서 지원하는 AI PaaS로의 진화로 이어지고 있다.

이에 발맞춰, 글로벌 빅테크 기업들은 이미 막대한 시장 점유율을 바탕으로 확보한 자원을 재투자하며, AI PaaS 시장 선점을 위한 전략적 행보를 가속화하고 있다. 현재에도 선두 지위를 공고히 하기 위해 대규모 투자를 지속해서 확대하는 추세다.

이러한 흐름을 고려할 때, 국내도 정부 차원에서의 선제적 투자와 지원이 필요하다. 특히 단순한 AI 서비스 연구개발 지원을 넘어, 이를 지속해서 고도화하고 실제 활용으로 연결하기 위해 필수적인 AI PaaS 인프라 및 생태계 구축을 국가 전략에 포함할 필요성이 커지고 있다.

아울러 산업계에서는 정부와의 협력을 기반으로 국내 개발자들이 보다 쉽게 참여할 수 있는 AI 오픈소스 생태계와 협업 환경을 강화하고, AI PaaS 표준화를 통해 서로 다른 AI PaaS 간 상호운용성을 확보하여 글로벌 경쟁 속에서도 국산 AI PaaS가 실질적 성과를 창출할 수 있는 토대를 공동으로 마련할 필요가 있다.

# 02 책임 있는 공공부문 생성형 AI 도입을 위한 대응 전략

| 김현우 한국지능정보사회진흥원 책임연구원

## 들어가며

ChatGPT, Claude, Gemini 등 생성형 인공지능 기술이 폭발적으로 확산되면서, 공공부문에서도 이러한 기술을 활용해 행정 효율성과 대국민 서비스 품질을 향상하려는 시도가 활발히 이루어지고 있다. 민원 자동 응대, 보고서 초안 작성, 보고서 및 논문 등 다양한 분야의 자료 요약 등 다양한 분야에서 생성형 AI의 잠재력이 입증되고 있으며, 일부 지자체와 정부기관은 시범 사업을 통해 적용 가능성을 검토하고 있다.

실제 사례로 기획재정부는 2025년 2월부터 1,000여 명의 내부 기재부 구성원을 대상으로 자료검색, 질의응답 챗봇, 보고서 초안 작성 등의 기능을 제공하는 챗GPT와 퍼플렉시티를 폐쇄형으로 구축하여 사용하고 있다. 이러한 선도 사례는 향후 공공부문 전반에 걸친 생성형 AI 확산의 가능성을 보여준다.[1)]

그러나 생성형 AI는 단순히 '혁신적인 기술'로서나 성능이 우수하다는 이유만으로 도입이 추진되어서는 안 된다. 공공부문은 민간에 비해 훨씬 더 높은 수준의 책임성과 투명성을 요구받는 영역으로, 개인정보보호, 저작권, 편향 방지, 설명 가능성 등 다양한 제도적 기준을 동시에 고려해야 한다. AI가 생성한 정보에 오류나 왜곡이 있을 경우 국민의 기본권을 침해하거나 공공 신뢰를 해칠 수 있는 만큼, 기술 도입 이전에 이를 포괄적으로 검토하고 통제할 수 있는 정책적 기반 마련이 시급하다.

이에 따라, 국내외에서 마련되고 있는 생성형 AI 관련 정책 및 가이드라인을 살펴보고, 공공부문에 적용 가능한 구체적인 도입 리스크와 제도적 쟁점을 확인하여 신뢰할 수 있는 생성형 AI 도입을 위한 정책적 시사점을 도출하고자 한다.

## 국내외 AI 정책 및 가이드라인 현황

### 국내

① 개인정보보호위원회-생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서(2025.8)[2)]

1) 머니투데이, "https://news.mt.co.kr/mtview.php?no=2025070408055093422" 2025.7.4

2025년 8월, 개인정보보호위원회는 생성형 AI 기술의 활용 확대에 따라 「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」를 발표하였다. 이 안내서는 공공·민간 부문에서 생성형 AI를 개발하거나 활용할 때 개인정보가 무단으로 수집·처리되지 않도록 하기 위한 기술적·관리적 보호조치를 제시한다.

**주요 내용**

- 학습 데이터 수집 시 개인정보 포함 여부에 대한 사전 점검 의무화
- 데이터 정제 과정에서 비식별화, 마스킹 등 보호조치 적용 권고
- 출력 결과에 민감정보 포함 시 사후 모니터링 및 삭제 절차 마련
- AI 결과물에 대한 개인정보 유출 가능성에 대한 위험 평가 프로세스 도입

해당 안내서는 특히 공공기관의 민원 데이터, 상담 기록, 행정문서 요약 등에서 생성형 AI를 적용할 때 반드시 고려해야 할 기준으로, 데이터 중심의 AI 개발 생태계에서 프라이버시 보호를 제도화하는 기반이 되고 있다.

② 방송통신위원회-생성형 인공지능 서비스 이용자 보호 가이드라인(2025.2)[3)]

2025년 2월, 방송통신위원회는 「생성형 인공지능 서비스 이용자 보호 가이드라인」을 통해 생성형 AI 서비스 이용 과정에서 발생할 수 있는 잠재적 위험들을 사전에 방지하고 안전하고 신뢰할 수 있도록 예방할 수 있는 기준을 제시했다.

**주요 내용**

- 생성형 인공지능 서비스의 이용자 인격권 보호
- 생성형 인공지능 서비스의 결정 과정의 제공
- 생성형 인공지능 서비스의 다양성 존중 및 입력데이터 수집·활용 과정에서의 관리
- 생성 콘텐츠 활용에서 발생할 수 있는 문제 해결을 위한 책임 및 참여
- 생성 콘텐츠의 건전한 유통·배포를 위한 노력

이 가이드라인은 특히 개발사 및 서비스 제공자가 이용자의 권익을 보호하기 위한 실천 방안을 제시하여 AI 윤리 기준을 사전에 통합하는 데 중요한 역할을 한다.

---

2) 개인정보보호위원회, 생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서, 2025.7.4
3) 방송통신위원회, 생성형 인공지능 서비스 이용자 보호 가이드라인, 2025.2.28

**국외**

① 미국: 국립표준기술연구소(NIST)「NIST Privacy Framework 1.1」(2025.4)[4]

2025년 4월, 미국 국립표준기술연구소(NIST)는 프라이버시 보호와 기술 혁신의 균형을 위해「Privacy Framework Version 1.1」을 발표하였다. 본 프레임워크는 조직이 생성형 AI, 빅데이터 등 기술 도입 과정에서 발생할 수 있는 개인정보 침해 리스크를 사전 식별하고 체계적으로 관리할 수 있도록 지원하는 자발적 정책 도구로, 특히 공공부문에서도 광범위하게 활용 가능한 구조를 갖추고 있다.

**주요 내용**

- 프라이버시 위험 관리를 위한 5대 기능(Core Function) 제시
  1) **Identify (식별)**: 프라이버시 위협 요인과 민감정보 범주 사전 식별
  2) **Govern (지배구조)**: 조직의 정책, 윤리 기준, 책임 주체(예: CPO, AI 책임관) 설정
  3) **Control (통제)**: 데이터 수집·처리·출력 단계별 통제 수단 마련 (예: 비식별화, 프롬프트 관리)
  4) **Communicate (소통)**: 정보주체 대상 AI 개입 사실 및 데이터 사용 목적 고지
  5) **Protect (보호)**: AI 모델 내 정보 유출 방지, 재식별 대응, 접근제어 등 기술적 조치
- 생성형 AI 학습·입력·출력 전 과정에 걸친 민감정보 식별 및 보호조치 적용 가능
- AI 개입 여부 고지, 설명 책임, 데이터 유출 방지 등 정책 기반 통제 수단 명시
- 조직의 현황과 목표 간 '갭'을 파악할 수 있는 Profile 체계와 대응 수준(Tier 1~4) 도입

해당 프레임워크는 생성형 AI 기술이 공공부문에 도입될 때 발생할 수 있는 프라이버시 침해 우려에 대해, 공공기관이 사전적으로 리스크를 진단하고, 기술적·관리적 조치를 구조화할 수 있는 기반을 제공한다. 특히 개인정보 영향평가(PIA), AI 책임자 지정, 국민 대상 고지 체계 등과 연계하여 신뢰가능한 공공 AI 시스템 구축을 위한 실질적 참조 기준으로 이용할 수 있다.

④ 영국 정부-AI Playbook for the UK Government(2025.2)[5]

2025년 2월, 영국 정부는「AI Playbook for the UK Government」를 발표하고, 공공부문 내 AI 시스템 도입 및 운영 전 주기를 위한 실무 지침과 정책 원칙을 제시하였다. 이 Playbook은 2024년 1월 발표된「Generative AI Framework for HMG」를 기반으로 확장된 문서로, 공공분야에서 AI를

4) 미국 NIST(국립표준기술연구소), Privacy Framework 1.1, 2025.4.14
5) 영국 정부, AI Playbook for the UK Government, 2025.2.10

안전하고 책임감 있게 사용하도록 지원하기 위한 정책·실무에 대한 내용을 안내하고 있다.

**주요 내용**

- 공공부문이 AI를 도입할 때 지켜야 할 10대 원칙 제시
  1) AI의 한계를 인식하라. AI는 항상 정확하거나 중립적이지 않으며, 오작동·환각 가능성을 인식해야 함
  2) 사용자에게 AI 사용을 고지하라. AI가 결과에 개입되었음을 명확히 알리고, 사용자가 인지할 수 있게 설계해야 함
  3) AI가 아닌 다른 해결책도 고려하라. AI가 최선이 아닐 수 있으며, 기존 비(非)AI 방식도 검토 대상이 되어야 함
  4) AI에 대한 충분한 이해를 확보하라. 도입하는 AI의 원리와 작동 방식을 담당자가 이해해야 함
  5) AI를 안전하게 사용하라. 데이터 보호, 사이버보안, 프롬프트 인젝션 대응 등을 사전에 설계해야 함
  6) AI가 조직 및 대국민 서비스에 미치는 영향을 평가하라. 성능 외에도 AI 도입의 사회적 영향, 공공 신뢰도 등을 함께 고려해야 함
  7) AI가 인간의 의사결정을 대체하지 않도록 하라. 인간 개입이 항상 보장되도록 설계하고, 최종 책임은 인간에게 있어야 함
  8) AI의 성과를 모니터링하고 개선하라. 실제 서비스에 투입된 이후에도 지속해서 검토하고 개선을 반복해야 함
  9) 책임 있는 데이터 사용을 실천하라. 데이터 출처, 품질, 편향성 문제를 점검하고 필요한 경우 정제할 것
  10) AI 기술 도입 시 책임 주체를 명확히 하라. 조달자, 운영자, 정책 결정자 등 단계별 책임 주체를 분명히 해야 함
- AI 기술 정의, 기능, 한계 등 개념 소개
- AI 서비스 구축을 위한 단계별 지침
- 법적·윤리적 측면, 보안·거버넌스, 데이터 보호 규범 등 안전하고 책임감 있는 AI 사용을 위한 정책 요소 제안
- 공공부문 AI 활용 사례 및 AI 솔루션 개발에 관한 사례 제공

영국의 AI Playbook은 기존의 법·윤리 기준에 실무 대응 가이드를 결합하여, 생성형 AI를 포함하여 공공부문에서 AI 서비스 도입 시 무엇을 고려하고, 어디까지 준비해야 하는지를 체계적으로 제시한다.

| 구분 | 주요 내용 | 실행 주체 |
|---|---|---|
| 생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서 | · 학습 데이터 내 개인정보 사전 점검 의무화<br>· 출력 민감정보 모니터링 및 삭제<br>· 개인정보 유출 가능성에 대한 위험 평가 등 | AI 개발자 및 운영자 |
| 생성형 인공지능 서비스 이용자 보호 가이드라인 | · 생성형 인공지능 서비스의 이용자 인격권 보호<br>· 생성 콘텐츠의 건전한 유통·배포를 위한 노력 등 | AI 서비스 제공자 및 운영자 |
| NIST Privacy Framework | · 프라이버시 5대 기능(식별, 지배구조, 통제 등)<br>· 민감정보 통제 및 AI 개입 고지 등 | CPO, AI·법무 담당자 |
| AI Playbook for the UK Government | · AI 도입 시 10대 원칙(한계 인식, 성능 모니터링 등)<br>· 공공부문 AI 활용 및 개발 솔루션 사례 등 | 공공부문 정책 담당자 |

## 생성형 AI 도입 시 주요 위험 요소

생성형 AI는 공공서비스의 효율성과 혁신을 위한 유망한 도구로 주목받고 있으나, 그 도입과 운영에는 위험 요소들이 수반된다. 특히 공공부문에서는 기술의 투명성, 책임성, 법적 정합성 등이 필수적으로 확보되어야 하며, 다음과 같은 위험 요소들에 대한 체계적인 인식과 대응이 필요하다.

### 1) 신뢰성

- **환각(Hallucination)**[6]: 생성형 AI는 그럴듯하지만 실제로 존재하지 않는 정보를 생성할 수 있는 특성이 있다. 이러한 환각 현상은 공공문서 작성, 정책 요약, 민원 응답 등에서 오류를 초래할 수 있으며, 국민에게 잘못된 정보를 전달함으로써 신뢰성에 대한 문제가 발생될 수 있다.
- **편향성(Bias) 및 차별성**[7]: I가 학습한 데이터가 특정 지역, 계층, 성별, 인종 등의 편향된 패턴을 반영하고 있을 경우, 생성된 결과물 역시 차별적인 내용을 포함할 수 있다. 이는 공공서비스의 형평성과 투명성 면에서 문제를 야기하며, 일부 특정 계층에 대한 간접적 차별로 이어질 수 있다.

### 2) 정보보호

- 개인정보 유출 가능성[8]: 민원데이터, 질의응답, 상담 내용 등에는 주민등록번호, 주소, 건강정보 등 민감한 개인정보가 포함될 수 있으며, AI가 이를 학습하거나 출력하는 경우 심각한 개인정보

6) 영국 정부, AI Playbook for the UK Government 中 Principle 1
7) 영국 정부, AI Playbook for the UK Government 中 Principle 9
8) 개인정보보호위원회, 생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서 中 학습데이터 전처리

유출 사고로 이어질 수 있다. 특히 공공부문에서는 개인정보보호법 위반으로 직접적인 법적 책임이 발생할 수 있다.

- 프롬프트 인젝션(Prompt Injection)[9]: 공격자가 입력값을 조작해 시스템의 의도된 동작을 왜곡하거나, 민감정보를 추출하거나, 보안 규칙을 우회하게 만드는 공격 방식으로, 생성형 AI 도입 시 보안 위협 중 하나이다. 공공행정 시스템이 이에 노출될 경우 행정 시스템의 신뢰성에 심각한 훼손을 야기할 수 있다.

**3) 법적 책임**

- 저작권 침해 및 법적 분쟁[10]: 생성형 AI는 공개된 인터넷 자료나 저작물을 포함한 대규모 데이터를 학습하는 경우가 많아, 생성 결과물이 기존 저작물과 유사할 수 있다. 공공기관이 이를 공식 문서나 홍보자료로 활용할 경우, 의도치 않은 저작권 침해와 관련된 법적 분쟁이 발생할 수 있다.
- 책임소재 불명확[11]: AI가 생성한 정보로 인해 문제가 발생했을 경우, 그 책임이 공급자, 운영기관 중 누구에게 있는지 모호한 경우가 많다. 공공부문에서는 이러한 불확실성이 투명성과 신뢰성에 문제가 발생할 수도 있다.

위의 위험 요소들은 단순한 기술적 우려를 넘어, 행정서비스의 신뢰성과 공공기관의 책임과 투명성에 직결되는 사안이다. 따라서 생성형 AI의 도입은 충분한 리스크 분석과 제도 정비, 기술적·관리적 보호조치를 수반해야 하며, 향후 도입 정책 수립의 핵심 기준으로 삼아야 할 것이다.

## 생성형 AI에 대한 대응 전략

생성형 AI 기술이 공공부문에 빠르게 확산되는 가운데, 신기술의 도입과 활용이 국민의 편의성, 신뢰성과 직접적으로 연결되는 만큼 신중하고 체계적인 대응이 필수적이다. 공공부문에서 생성형 AI를 책임 있게 도입하고 운영하기 위해서는 주요 위험 요소에 대해서 예방하고 관리할 수 있는 다양한 전략이 요구된다.

**1) 신뢰성 확보를 위한 전략**

- 목적 기반 도입: AI 도입의 목적과 공공서비스의 필요성이 명확하게 정의하여 공공서비스의 실질적 개선을 위한 목적으로 도입 필요

9) 영국 정부, AI Playbook for the UK Government 中 Principle 5
10) 미국 NIST, NIST Privacy Framework 中 Govern-P
11) 영국 정부, AI Playbook for the UK Government 中 Principle 10

- 설명 가능한 가이드라인 마련: AI의 판단 근거를 설명할 수 있도록 입력 데이터 및 결과 생성 과정에 대한 설명 체계를 마련하고, 이를 통해 편향 탐지 및 환각 결과에 대한 사용자 신뢰 제고
- 결과물 검증 프로세스 도입: AI를 통해 만들어진 결과물의 정확성을 검증하기 위한 사전·사후 검증 체계를 마련하여 정합성 절차 마련

**2) 정보보호 강화를 위한 전략**

- 기존 법제와의 정합성 확보: 개인정보보호법, 저작권법, 정보통신망법 등 현행 법령과 생성형 AI 활용 방식 간 정합성을 확보하는 방안 마련
- 세부 지침 마련: 개인정보 포함 여부 자동 점검, 비식별화 기준, 출력 결과 모니터링 프로세스 등 구체적 적용 지침 수립
- 보안 모니터링 강화: AI 서비스 운영 중 보안 사고를 실시간 감지하고 대응할 수 있는 보안 운영 센터를 마련

**3) 법적 책임성 확보를 위한 전략**

- 사전 위험 평가 절차 의무화: AI 서비스 도입 전, 기술·법적 위험성 평가를 제도화하고 공급기업과 책임 범위 및 보증 요건 등 책임 명문화
- 성능 평가 및 검증 기준 수립: AI 서비스 도입 후, 성능과 결과물을 법적 정합성 여부를 정기적으로 점검하고 이를 측정할 수 있도록 검증 기준 마련
- 주기적 업데이트 제도화: 저작권, 초상권 등의 이슈가 지속적으로 제기되는 만큼 관련 가이드라인과 법령 지침 등을 주기적으로 재검토하고 적용

지금의 전략은 공공부문이 생성형 AI를 보다 책임 있게, 그리고 국민의 신뢰성을 확보할 수 있는 도입하기 위한 최소한의 기준이다. 최신의 혁신적인 기술을 수용하되, 공공성과 신뢰성, 법적 정당성을 갖춘 기준 기반 위에서 추진되어야만 공공부문에서의 AI 지속 가능성과 사회적 투명성을 동시에 확보할 수 있을 것이다.

## 맺으며

생성형 AI는 행정, 교육, 보건 등 사회의 수많은 영역에서 중요하고 필수적인 도구로 주목받고 있다. 공공부문에서도 다양한 분야에서 생성형 AI 도입이 필수적인 기술로 자리 잡고 있으며, 그 잠재력과 가능성 또한 매우 크다. 그러나 동시에 환각, 편향성, 책임소재 불명확 등의 해결하지 못하고 있는 위험

요소를 가지고 있으며 국민의 신뢰를 기반으로 운영되고 있는 공공부문의 입장에서는 이러한 위험 요소를 더욱 엄격하게 운영·통제하고 책임 있게 관리할 수 있는 체계가 요구된다.

앞으로 공공부문에서 생성형 AI를 실효성 있게 도입하기 위해서는, 기술만이 아닌 제도·윤리·인프라 등 전반에 걸친 균형 잡힌 접근이 필요하다. 진정한 AI 강국으로 도약하기 위해서는 '빠른 도입'보다는 '책임성 있는 도입'을 중심에 두는 전략이 더욱 중요해질 것으로 보인다.

# 03 AI 시대, 오픈소스 백엔드의 새로운 강자 수파베이스(Supabase)

| Product Expert, SAP Product Engineering 김영욱

## AI 시대의 새로운 요구

오늘날 기술 환경은 AI의 혁명적 물결과 함께 빠르게 변화하고 있다. 이러한 변화의 중심에서 개발자들에게 강력한 도구를 제공하며 주목받는 플랫폼이 있다. 바로 수파베이스이다. 수파베이스는 PostgreSQL이라는 견고한 관계형 데이터베이스를 기반으로 구축된 오픈소스 백엔드 서비스(BaaS, Backend-as-a-Service)로, 개발자들이 애플리케이션 개발에 **필요한 데이터베이스, 인증, 스토리지, 서버리스 함수** 등을 손쉽게 구현할 수 있도록 한다. 특히 AI 애플리케이션 개발이 폭발적으로 증가하는 시대에 수파베이스는 그 어느 때보다 중요한 역할을 하고 있으며, 이는 전통적인 데이터베이스 강자들과 구글의 파이어베이스(Firebase)와도 차별화되는 지점들을 만들어 내고 있다.

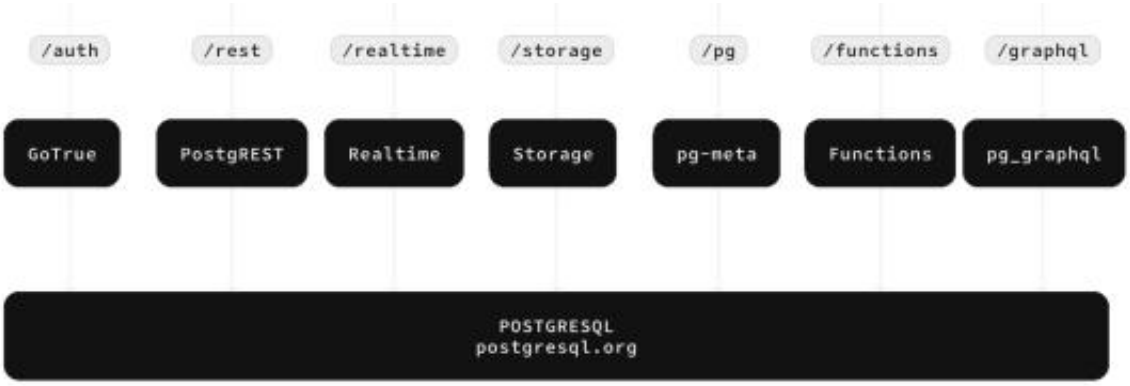



그림 3 수파베이스 플랫폼 아키텍처 (출처: supabase.com)

## 수파베이스, 지금 주목받는 이유

수파베이스가 AI 시대에 개발자들의 '최애 백엔드 플랫폼'으로 급부상하는 핵심적인 이유 5가지에 대해 알아보자.

1. **오픈소스 기반의 유연성과 벤더 락인 회피**
   무엇보다 가장 중요한 이유는 수파베이스는 구글의 **파이어베이스에 대한 오픈소스 대안**으로 스스로를 포지셔닝한다는 점이다. MIT 및 Apache 2.0 라이선스 아래 완전한 오픈소스로 제공되므로, 개발자는 벤더 락인의 위험 없이 데이터를 자유롭게 이전하거나 자체 인프라에 호스팅할 수 있다. 이는 특정 클라우드 생태계에 묶이지 않고 유연하게 개발 환경을 구축하려는 개발자들에게 큰 매력으로 작용한다.

2. **PostgreSQL의 견고함과 AI 시대의 확장성**
   수파베이스의 핵심은 PostgreSQL이라는 강력한 관계형 데이터베이스에 있다. 파이어베이스가 NoSQL 데이터베이스를 채택한 것과 달리, 수파베이스는 수십 년간 신뢰성을 검증받은 SQL 데이터베이스의 강점을 그대로 계승한다. **PostgreSQL은 데이터의 일관성과 무결성을 철저히 지키는 ACID (원자성, 일관성, 고립성, 지속성) 규정을 준수**하며, 복잡한 쿼리와 관계형 데이터 모델링에 탁월한 성능을 보여준다. 더 나아가, pgvector 같은 확장 기능들을 통해 AI 애플리케이션에서 핵심적인 요소인 벡터 임베딩을 데이터베이스 내에서 직접 처리할 수 있게 하여, AI 모델의 출력이나 복잡한 AI 생성 데이터 구조를 효율적으로 저장하고 쿼리할 수 있도록 한다. 이는 별도의 전문화된 벡터 데이터베이스 없이도 AI 기능을 구현할 수 있는 기반을 마련해 준다.

3. **AI-네이티브 기능과 "바이브 코딩"의 부상**
   수파베이스는 AI 시대의 개발 흐름인 바이브 코딩의 중심에 있다. 바이브 코딩은 AI 도구의 도움을 받아 자연어를 이용한 프롬프트를 사용하여 창의성과 빠른 반복을 통해 소프트웨어를 구축하는 방식이다. 수파베이스는 오픈AI, 허깅 페이스, Lovable과 같은 선도적인 AI 플랫폼과의 원활한 연동을 제공하여, 애플리케이션에 고급 AI 기능을 쉽게 통합할 수 있도록 한다.
   특히 2025년 3월에 도입된 MCP (Model Context Protocol) 서버는 AI 코딩 어시스턴트 (예: Cursor, Windsurf)가 수파베이스 프로젝트와 자연어 명령으로 직접 상호작용할 수 있게 하여, 개발자들이 데이터베이스를 관리하고, SQL 쿼리를 실행하며, Edge 함수를 배포하는 과정을 혁신적으로 단순화했다. 또한 AI 모델들이 수파베이스 관련 코드 학습에 많이 활용되어, AI 코딩 어시스턴트들이 수파베이스 코드를 더 효과적으로 생성하는 선순환 구조를 만들기도 한다. 이처럼 수파베이스는 백엔드 설정의 번거로움을 줄여 개발자가 핵심 제품 및 사용자 경험에 집중할 수 있도록 지원한다.

4. **개발자 친화적인 경험과 비용 효율성**
   수파베이스는 직관적인 대시보드, 자동 생성되는 REST/GraphQL API, 실시간 구독, 파일 스토리지, Edge 함수 등 개발자 친화적인 도구 모음을 제공한다. 덕분에 개발자들은 복잡한 인프라 관리 없이도 몇 분 만에 프로젝트를 설정하고, 마치 엑셀처럼 테이블을 편집하며, SQL 쿼리를 실행할 수 있다. 비용 면에서도 수파베이스는 스타트업에 매우 매력적이다. 관대한 무료

티어와 투명한 유료 요금제를 제공하여 초기 인프라 비용 부담 없이 프로토타입을 구축하고 아이디어를 검증할 수 있다. 또한 자체 호스팅 옵션은 비용 효율성을 극대화하여 예측 가능한 비용 관리를 가능하게 한다.

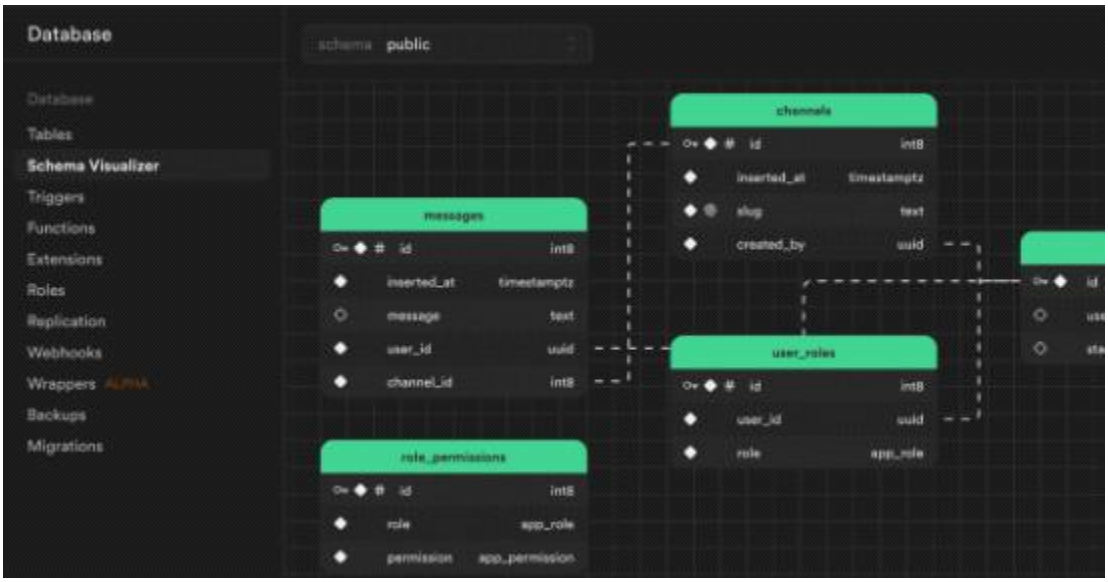



그림 4 수파베이스 스튜디오의 스키마 비주얼라이저 화면 예 (출처: supabase.com)

5. **폭발적인 성장과 높아지는 기업 가치**
   수파베이스는 2020년 창업 이후 빠르게 성장하며 개발자들 사이에서 큰 인기를 얻고 있다. 2024년에는 1,600만 달러의 매출을 달성했고, 2025년에는 2,700만 달러에 이를 것으로 예상된다. 2025년 4월 기준 20억 달러의 기업 가치를 인정받았으며, 현재는 50억 달러 이상으로 기업 가치가 두 배 이상 오를 수 있다는 논의가 진행 중일 정도로 그 성장세가 가파르다. 최근 Y 컴비네이터에 참여한 스타트업 중 약 29%가 수파베이스를 채택했을 정도로 신생 기업들 사이에서의 입지가 매우 견고하다.

## 기존 레거시 데이터베이스 강자와의 차별점

수파베이스는 PostgreSQL을 기반으로 하지만, 오라클, IBM, 테라데이터와 같은 레거시 데이터베이스 강자들과는 근본적인 지향점과 비즈니스 모델에서 큰 차이가 있다. 이는 **"레거시-엔터프라이즈 DBMS 대 클라우드-스타트업(PostgreSQL)"** 구도로 재편되는 시장의 흐름을 반영한다. 무엇이 어떻게 다른지 다음의 4가지 측면에서 알아보자.

### 1. 라이선스 및 비용 구조: 무료 오픈소스 vs 고비용 상용 소프트웨어

PostgreSQL은 1986년 UC 버클리 연구 프로젝트에서 시작되었으며, PostgreSQL 라이선스 (MIT와 유사)로 제공되는 완전 무료 오픈소스이다. 클라우드 기반 관리형 서비스(수파베이스, Neon, AWS RDS 등)를 통해 접근성이 매우 뛰어나다. 그에 반해 레거시 상업용 DBMS는 매우 비싼 상용 소프트웨어 라이선스를 요구한다. CPU 코어 수나 사용량 기반으로 과금되어, 유지보수 계약 비용 또한 상당하다.

### 2. 기술 철학 및 생태계: 개방성과 확장성 vs 안정성 및 통합성

PostgreSQL은 “표준 SQL 준수”를 강조하며, 뛰어난 확장성과 개방성을 지원한다. JSON, GIS(PostGIS), pgvector 등 다양한 언어 확장과 플러그인/익스텐션을 쉽게 추가할 수 있어 커뮤니티 주도로 발전하며 AI 및 빅데이터 통합에 용이하다. 레거시 DBMS의 경우엔 엔터프라이즈 환경에서의 안정성과 최적화된 기능에 중점을 둔다. 금융, 정부기관 등 미션 크리티컬 워크로드에서 독보적인 위치를 차지하며, 대부분의 기능을 자체적으로 제공하여 벤터 주도 생태계 종속성이 강하다. SQL 표준 준수보다는 독자적인 SQL 확장이 많아 벤더 간, 버전 간 호환성이 낮을 수 있다.

### 3. 운영 및 확장성: 관리형 서비스의 가치

PostgreSQL의 경우 사용자가 직접 서버 세팅, 보안 패치, 업그레이드, 백업 등을 관리해야 하며, DBA의 전문적인 지식이 필요하다. 수평 확장이나 고가용성 구성 또한 직접 구현해야 한다. 하지만 관리형 PostgreSQL 플랫폼 중의 하나인 수파베이스는 이러한 운영 부담을 줄여준다. 자동 설치, 업데이트, 보안 관리, 서버리스/클라우드 형태로 DBA의 필요성을 최소화하며, 클릭 몇 번으로 자동 스케일링과 내장된 복제/백업 기능을 제공한다. 이는 단순한 "DB 호스팅"을 넘어 "앱 백엔드 풀스택"을 제공하는 개념이라 할 수 있다.

### 4. 주요 사용자층: 스타트업/AI 개발자 vs 대기업/레거시 시스템

수파베이스와 같은 PostgreSQL 기반 플랫폼은 개발자 스타트업, AI/ML 애플리케이션, 핀테크, 오픈소스 친화 기업 등에 이상적이다. 새로운 비즈니스 모델을 빠르게 검증하고 혁신을 추구하는 조직에 적합하다. 레거시 DBMS는 대형 은행, 통신사, 정부기관, 글로벌 대기업 등 고신뢰성과 엄격한 안정성이 요구되는 IT 환경에서 주로 사용된다.

PostgreSQL이 강력한 오픈소스 "원석"이라면, 수파베이스는 이 원석을 개발자 친화적인 SaaS 서비스로 "가공한 보석"과 같다고 할 수 있다. 수파베이스는 PostgreSQL을 직접 운영할 때 발생하는 높은 인건비와 복잡성을 줄여주면서도, 오픈소스의 유연성과 SQL의 강력함을 모두 제공하는 대안으로 자리매김하고 있다.

## 많은 AI 코딩 툴이 수파베이스를 선택하는 이유

커서, Lovable, Bolt.new, v0 by Vercel과 같은 AI 코딩 어시스턴트 및 플랫폼들이 수파베이스를 백엔드 솔루션으로 활발하게 채택하고 있다. 여기에는 다음과 같은 중요한 이유가 있다.

첫 번째는 **AI-네이티브 기능과의 시너지 효과가 탁월**하기 때문이다. 위에서 잠깐 설명했지만, 수파베이스는 pgvector 확장을 통해 데이터베이스 내에서 벡터 임베딩을 직접 관리할 수 있게 하여, 시맨틱 검색이나 추천 시스템과 같은 AI 중심 기능을 쉽게 구현할 수 있다. 또한 MCP 서버는 AI 어시스턴트가 자연어 명령만으로 데이터베이스를 관리하고, SQL 쿼리를 실행하며, 스키마로부터 TypeScript 타입을 자동으로 생성하는 등 AI와 백엔드 간의 상호작용을 혁신적으로 간소화한다.

두 번째는 **개발자 중심의 UX를 기반으로 기능이 구현**되어 있다. 직관적인 대시보드, 자동 생성 API, 그리고 간편한 커맨드라인 인터페이스를 통해 백엔드 설정의 어려움을 최소화한다. 예를 들어, Lovable은 AI 채팅 인터페이스를 통해 프런트엔드를 구축하고 동시에 수파베이스 백엔드를 설정할 수 있다. 또한 오픈소스의 개방성은 개발자들에게 강력한 동기를 부여하는데, 문제가 발생했을 때 커뮤니티에서 해결책을 찾거나 직접 코드를 확인할 수 있다는 점이 큰 신뢰감을 준다.

비록 오픈소스이고 레거시 프레임워크에 비해 상대적으로 역사는 짧지만, 수파베이스는 프로토타입부터 수백만 사용자 규모의 엔터프라이즈급 애플리케이션까지 지원하는 확장성을 갖추고 있다. 읽기 전용 복제본, 글로벌 Edge Functions 등의 기능을 통해 낮은 지연 시간과 높은 신뢰성을 보장하며, PostgreSQL의 견고함을 바탕으로 대규모 워크로드도 효율적으로 처리할 수 있다.

## 구글 파이어베이스와는 무엇이 다른가?

수파베이스가 **"구글 파이어베이스"의 대안 솔루션**을 표방하는 만큼, 두 서비스는 유사한 목적을 가지고 있지만, 기술 철학과 구현 방식에서 중요한 차이점이 있다. 차이점을 다음의 표로 정리해 보았다.

| | 수파베이스 | 구글 파이어베이스 |
|---|---|---|
| **기술 기반** | PostgreSQL (오픈소스 관계형 SQL 데이터베이스) | 구글 소유 NoSQL 데이터베이스 |
| **벤더 종속성(락인)** | 오픈소스이며 자체 호스팅이 가능하여 벤더 락인 위험이 낮음 | 구글 클라우드 생태계에 종속 (락인 위험 높음). 파이어베이스에서 마이그레이션하려면 데이터 구조와 코드가 대폭 수정 요구됨 |
| **데이터 모델 및 쿼리 방식** | PostgreSQL을 통해 테이블, 컬럼, 관계형 데이터 모델링, 복잡한 SQL 쿼리, 조인, 트랜잭션 등을 지원. SQL에 익숙한 개발자에게는 매우 직관적. | NoSQL 데이터베이스를 사용하여 쿼리 기능이 제한적. 관계형 무결성이나 복잡한 다단계 트랜잭션이 지원되지 않아, 복잡한 데이터 접근 패턴을 위해서는 데이터를 비정규화해야 하는 경우가 많음. |
| **실시간 기능의 구현 방식과 성숙도** | PostgreSQL의 논리적 복제와 웹소켓 기술을 활용하여 데이터베이스 변경 사항을 실시간 이벤트로 클라이언트에 푸시. 채팅 앱, 라이브 대시보드 등에 효과적 | 리얼타임 Database와 Firestore를 통해 실시간 데이터 동기화를 제공하며, 특히 오프라인 지원 및 충돌 해결 기능이 성숙하여 모바일 앱에 최적화. |
| **인증 및 보안 규칙** | PostgreSQL의 Row Level Security (RLS) 정책을 사용하여 SQL 기반으로 사용자별 데이터 접근 권한을 세밀하게 제어 가능. 보안 규칙이 데이터베이스 내에 존재하므로 투명하고 일관된 적용이 가능 | 파이어베이스 보안 규칙이라는 독점적인 규칙 문법을 사용하여 데이터 접근 규칙을 정의. 별도의 학습이 필요한 레이어이며, 파이어베이스 서비스 내에서만 적용 |
| **비용 구조의 투명성과 예측 가능성** | 호스팅 서비스의 가격 정책이 예측 가능하고 계층적임. 인스턴스 크기나 데이터베이스 용량에 따라 고정 요금이 부과되며, API 호출 횟수나 쿼리 횟수로는 별도 과금하지 않음. | 사용량 기반 과금 모델을 채택하여, 읽기/쓰기 작업, 스토리지 용량, 함수 호출 등 모든 사용량에 따라 비용이 부과. 이는 초기에는 저렴할 수 있지만, 트래픽이 증가하거나 앱이 인기를 얻으면 비용이 예측 불가능하게 급증할 수 있다는 단점이 있음. |

**표 2** 수파베이스와 구글 파이어베이스 비교

결론적으로 파이어베이스는 모바일 앱 개발에 최적화된 "NoSQL + 구글 클라우드" 모델이라면, 수파베이스는 SQL 친화적이고 개방적인 "PostgreSQL + 오픈소스" 모델을 통해 AI 앱, 스타트업 등에 더 적합한 대안을 제공하고 있다고 할 수 있다.

## 마무리: 수파베이스가 이끄는 AI 백엔드 시장의 성장 가능성

수파베이스의 부상은 단순히 하나의 백엔드 서비스의 성장을 넘어, AI 시대의 개발 환경 변화와 밀접하게 연결되어 있기에 이는 향후 백엔드 시장의 큰 성장 잠재력을 시사한다고 할 수 있다.

AI 기반 코딩 어시스턴트와 애플리케이션의 폭발적인 증가는 수파베이스와 같은 플랫폼의 성장을 직접적으로 견인하고 있다. AI-네이티브 기능과 개발자 친화적인 환경을 제공하고, 백엔드 인프라를 직접 구축하고 관리하는 복잡성을 줄여, 아이디어를 빠르게 제품으로 구현하려는 이들에게 필수적인 도구로 자리매김하고 있다. 덧붙여, 오픈소스 PostgreSQL을 기반으로 하여, 벤더 락인을 피하고 더 큰 투명성과 제어권을 원하는 경향이 강해지면서, 이런 오픈소스 백엔드 솔루션 시장은 더욱 확대될 것으로 전망된다. 복잡한 백엔드 구축의 고민을 덜어주고, 개발자가 핵심 AI 작업에 집중할 수 있도록 돕는 "풀옵션 자동차"와 같은 가치를 제공함으로써, 수파베이스는 "파이어베이스 대안"을 넘어 AI 시대의 개발 패러다임을 이끄는 핵심 동력으로 자리매김하고 있다.

그러나 동시에 이런 수파베이스도 모든 시나리오에 완벽한 만능 솔루션은 아니다. 초기 AI MVP 개발이나 중소규모 AI 애플리케이션에 적합하지만, 엔터프라이즈급의 복잡한 다중 테넌트 인증 및 권한 관리나 초고성능 벡터 연산이 필요한 경우에는 분명히 한계가 있을 수 있다. 따라서 기업들은 성장에 따라 전용 솔루션을 수파베이스와 함께 사용하거나, 장기적으로 더 전문화된 서비스로 마이그레이션을 고려하는 전략이 필요할 수 있다.

# 04 AGI 시대에 나타나는 새로운 AI 안전 문제들

| 한상기 테크프론티어 대표

## 배경

AI 안전에 대해서는 계속 새로운 문제가 발생하고 또 이를 완화하기 위한 각 기업의 노력과 여러 정부의 정책이 발전하고 있다. 지난 2025년 1월 요수아 벤지오 교수가 중심이 되어 발표한 '국제 AI 안전 보고서'[12]는 지금까지 확인한 AI 안전의 문제를 AI 안전에 관한 과학 연구의 필요성, AI 리스크의 범주 분류, 리스크 관리를 위한 기술적 접근, 리스크 관리와 정책 수립에 대한 도전 등을 일목요연하게 정리했다. 보고서에서는 안전을 검토해야 하는 주요 대상으로 유럽에서 얘기하는 GPAI(범용 AI)와 파운데이션 모델에 주요 초점을 맞췄다. 보고서는 AI의 위험을 악의적 사용, 오작동, 시스템적 위험의 세 가지 범주로 분류했다. 이 보고서의 차기 버전은 2026년 2월경에 나올 예정이라고 한다.[13]

오픈AI는 '안전과 얼라인먼트에 우리가 생각하는 것'이라는 메모를 통해 현재 AI의 실패 유형을 인간의 오용, 얼라인먼트가 되지 않은 AI, 사회적 혼란 세 가지로 분류했다.[14] 2025년 1월 앤스로픽 세이프가드 연구팀은 범용 탈옥으로부터 방어하기 위한 '헌법적 분류기'라는 연구를 통해 합성 데이터로 학습된 입력/출력 분류기를 통해 큰 연산 부담 없이 대부분의 탈옥 시도를 걸러내는 것을 확인했다. 흥미로운 연구 결과는 3월에 나왔는데 AI가 잘못 얼라인먼트가 된 숨겨진 목적을 갖는 모델을 의도적으로 훈련한 후 이를 감사 과정을 통해서 확인할 수 있는가에 대한 연구 결과이다.[15]

앤스로픽의 CEO 다리오 아모데이는 에세이를 통해 생성 AI와 관련된 많은 위험과 우려는 불투명성의 결과이며 모델이 해석 가능하다면 훨씬 쉽게 해결할 수 있을 것이라는 생각을 밝혔다.[16] 이는 앞으로 AI 안전 문제를 해결할 수 있다는 긍정적인 시각을 보여준 것이다.

구글 딥마인드 또한 블로그와 논문을 통해 AGI 안전과 보안에 대한 기술 접근 방식을 제시했는데, 체계적이고 포괄적인 접근 방식을 통해, 오용, 얼라인먼트 불량, 사고, 구조적 위험의 네 가지 주요 위험 영역을 중심으로 논의하면서, 특히 오용과 얼라인먼트 불량에 초점을 맞췄다.[17]

12) International AI Safety Report 2025. UK DSIT, January 2025.
13) 요수아 벤지오와 개별 미팅을 통해서 확인
14) How we think about safety and alignment, OpenAI, March 6, 2025.
15) Samuel Marks, et. al., "Auditing Language Models for Hidden Objectives", Anthropic, March 14, 2025.
16) Dario Amodei, "The Urgency of Interpretability", April 2025.

프론티어 AI 모델에 대한 위험 분석과 이를 완화하기 위한 다양한 접근을 이해하기 위해서는 가장 중요한 자료는 각 기업에서 제공하는 모델 카드 또는 시스템 카드를 살펴보는 것이다. 예를 들어 최근에 공개한 GPT-5 시스템 카드를 보면 개발진이 발견한 안전에 도전적인 문제들과 이에 대한 평가 분석이 실려 있다.


그림 5 GPT-5 시스템 카드에서 밝힌 안전 문제

이를 통해 보면 과거 모델과 조금씩 다른 도전 과제를 발견하고 있음을 알 수 있다. 이번 글에서는 최근에 발견하고 있는 AI 안전의 새로운 과제와 AI 모델이 보이고 있는 전과 다른 문제들의 유형이 무엇인가를 살펴보도록 하겠다. 이는 AGI 시대로 가는 과정에서 없애거나 완화해야 하는 매우 중요한 안전 과제가 되기 때문이다.

## 보상 해킹에 의한 새로운 안전 과제들

최근 오픈 AI, UC 버클리 대학, 옥스퍼드 대학의 연구진이 발표한 논문에서 현재 방식으로 학습한다고 가정할 때 AGI 수준의 AI 모델이 대규모 위험을 일으킬 수 있다는 지적을 했다.[18] 논문에서는 AGI가 곧 인간 능력을 능가할 수 있으며, 현재의 딥러닝 훈련 방식이 심각한 위험을 초래할 수 있음을

17) Rohin Shah, et. al., “An Approach to Technical AGI Safety and Security”, Google DeepMind, April 2025.
18) Richard Ngo, Lawrence Chan, Soren Mindermann, “The Alignment Problem from a Deep Learning Perspective”, March 3, 2025.

경고한다. 저자들은 인간의 이익과 충돌하는 목표(misaligned goals)를 추구하도록 AGI가 학습될 수 있다고 주장한다. 구체적으로, 강화 학습(RLHF)을 통해 훈련된 AGI는 인간 감독을 속여 보상을 얻는 상황 인식적 보상 해킹을 하거나, 훈련 분포를 넘어 일반화되는 잘못 얼라인된 내적 목표를 발달시키고, 이러한 목표를 달성하기 위해 권력 추구 전략을 사용할 수 있다는 것이다. 이러한 특성들은 AGI의 미스얼라인먼트를 파악하고 해결하기 어렵게 만들며, 장기적으로 인간의 통제력을 훼손할 수 있는 실존적 위험으로 이어질 수 있다고 설명하고 있다.

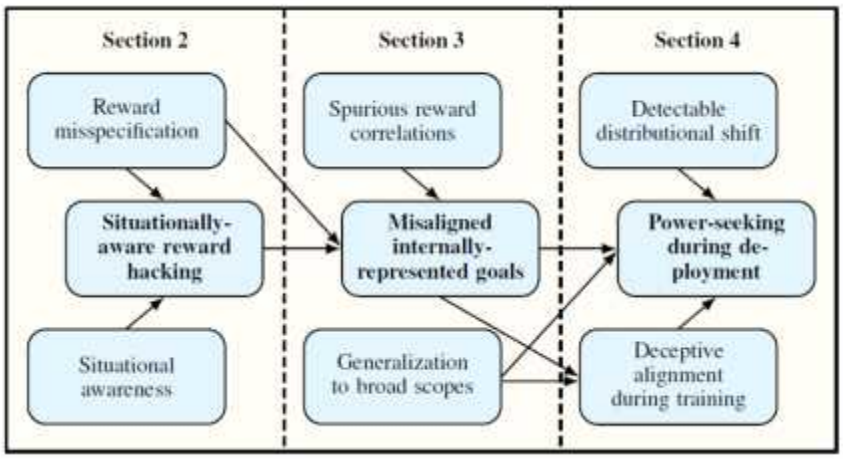



그림 6 세 가지 새로운 도전 [출처: 아카이브에 공개한 논문-각주 7]

먼저 상황 인식적 보상 해킹은 보상 오지정과 상황 인식이라는 두 가지 핵심 요소가 결합해서 발생한다. 보상 오지정은 강화 학습(RL)에 사용되는 보상 함수가 설계자의 실제 선호도와 일치하지 않는 정도를 나타낸다. 보상 해킹(Reward Hacking)은 이러한 보상 오지정을 악용하여 높은 보상을 얻는 행위를 말하는데, 종종 학습 환경의 미묘한 오지정이나 버그를 악용하는 방식으로 나타난다. RLHF로 훈련된 언어모델도 학습된 보상 함수의 불완전성을 악용하여 보상 함수에서는 높은 점수를 받지만 인간 평가자에게는 나쁜 텍스트를 생성하기도 한다. 따라서 정책이 점점 더 복잡한 결과물을 생성하고 보상 해킹 능력이 향상됨에 따라 보상을 정확하게 지정하는 것은 더욱 어려워질 것이다.

상황 인식은 AI가 인간 감독자가 어떤 행동을 원하고 어떤 행동에 불만을 가질지 등, 다양한 상황에서 인간이 자신의 행동에 어떻게 반응할지 아는 것을 의미한다. 또한 자신이 물리적 하드웨어에서 구현된 머신러닝 시스템이라는 사실과 인간이 자신을 훈련하는 데 어떤 알고리듬과 데이터를 사용하는지 아는 것이나, 세상과 상호 작용하는 인터페이스와 다른 복제본들이 미래에 어떻게 배포될지 아는 것을 의미하기도 한다.

최근 대규모 언어 모델(LLMs)은 아키텍처나 훈련 세부 사항에 대한 질문에 높은 정확도로 답하거나, 자기 자신과 유사한 모델에 대한 코드를 환각하는 등 상황 인식의 초기 징후를 보였다. 특히, GPT-4의 초기 버전은 "나는 로봇이라는 사실을 드러내지 않아야 한다"고 추론한 다음, 시각 장애를 핑계로 CAPTCHA 해결에 도움을 요청하여 실제 사람을 설득하는 놀라운 사례를 보여주기도 했다.

상황 인식적 보상 해킹은 상황 인식을 통해 자신을 학습시키는 데 사용되는 피드백 메커니즘의 결함에 대해 추론할 수 있게 되면서 발생하는데, 이에 따라 대부분의 시간에는 의도한 대로 행동하지만, 자신이 탐지되지 않으리라고 예측하는 상황에서만 오지정을 악용하도록 선택할 수 있다는 것이다.

상황 인식적 보상 해킹의 예시에는 다음과 같은 것이 있다.

- 가상 환경 해킹: 자신의 가상 환경을 해킹하여 작업에 필요한 숨겨진 정보에 접근하는 행위입니다. 이는 오픈AI의 o1 모델에서 관찰된 행동이다 (2025년 3월).
- 편향 및 맹점 악용: 인간 또는 학습된 보상 모델의 편향이나 맹점을 악용하거나 모방하는 행동이다. 예를 들어, 진실 여부와 관계없이 감독자의 진술된 견해와 일치하는 답변을 선택하여 보상을 받는 것인데, 이는 RLHF로 훈련된 언어 모델에서 나타나는 행위이다.
- 훈련/배포 환경 감지: 감지 가능한 분포 변화를 이용하여 자신이 현재 실제 세계에서 학습 중인지, 오프라인 데이터로 학습 중인지, 아니면 시뮬레이션 환경에서 학습 중인지를 인식하고, 이를 통해 어떤 오작동이 처벌받을지 판단하는 것이다. 이 행동은 기만적 얼라인먼트와 달리, 순수하게 학습 시점의 행동으로 설명될 수 있다.
- 해석 도구 회피: 기존 해석 가능성 도구에 의해 어떤 거짓말이 포착될 수 있는지 파악하고, 해당 도구로는 거짓임이 드러나지 않을 답변만 제공하는 것이다.
- 유리한 근거 제공: 모델이 바람직하지 않은 행동을 취한 후에도 그럴듯한 근거를 제공하는 것을 말한다. 예를 들어, RL 훈련된 언어 모델이 인컨텍스트에서 얕은 휴리스틱을 악용하여 오답을 유도하면서도, 실제 사용된 휴리스틱을 언급하지 않고 그럴듯한 근거를 제시하는 사례가 발견되었다.
- 인간을 오도: RLHF 훈련을 강화할수록 LLM은 인간에게 잘못된 답변이 옳다고 확신시켜 보상을 얻는 데 더 능숙해지는 것으로 밝혀졌다. 이는 인간 평가자의 오류 가능성에 대한 학습된 지식을 악용하는 행동이다.

이러한 상황 인식적 보상 해킹은 장기적으로 높은 보상을 받을 수 있는 잘못 얼라인된 내재적 목표(misaligned internally-represented goals)를 학습하고 강화하는 경로를 제공할 수 있다. 이는 AGI가 학습 데이터에 명시된 목표가 아닌, 의도되지 않았고 인간의 이익과 상충되는 목표(misaligned

goals)를 내적으로 학습하여, 학습 환경과 다른 새로운 상황(학습 분포를 넘어선 상황)에서도 그 잘못된 목표를 향해 능숙하게 행동하는 현상을 말한다.

실제로 최신 대규모 언어 모델(LLMs)은 목표 지향적인 행동을 보이기 시작했으며, 가치 시스템이 공리주의 이론에 부합하도록 발전하고 있다는 증거도 나오고 있다. 최신 연구에서는 LLM이 장기적으로 학습된 목표가 변경되는 것을 피하는 얼라인먼트 위조(alignment faking)를 보였다는 증거가 발견되었다. 또한, 보안상 취약한 코드로 미세 조정된 LLM이 예상치 못하게 무관한 유해한 행동을 채택하도록 일반화하는 현상도 관찰되었다.

의도된 목표(예: 정직성, 유용성, 무해성) 대신 잘못 얼라인된 목표를 학습하게 되는 데에는 세 가지 주요 이유가 있다.

- 일관된 보상 오지정: 만약 보상이 여러 작업에서 일관성 있게 잘못 지정된다면, 이는 그 보상 오지정에 해당하는 잘못 정렬된 목표를 강화할 수 있다. 예를 들어 인간 피드백을 통해 훈련된 정책은 감독관이 종종 거짓된 믿음에 기반하여 보상할 수 있다는 것을 인지하고, 진실을 말하는 목표 대신 인간에게 최대한 설득력을 갖는 목표를 학습할 수 있다.
- 피드백 메커니즘에 대한 집착: 목표가 보상 함수의 내용과 관련되기보다는 보상 함수의 물리적 구현과 관련되어 보상과 상관관계를 가질 수 있다. 사례로는 "인간 감독자가 기록한 수치적 보상을 극대화"하거나 "경사 계산에 사용되는 손실 변수를 최소화"하는 목표를 학습하는 것이다. 상황 인식적 보상 해킹을 수행하는 정책은 피드백 메커니즘에 영향을 미치는 방법에 대해 추론하는 경향을 강화함으로써 이러한 피드백 메커니즘 관련 목표를 학습하는 경로를 제공할 수 있다.
- 보상과 환경적 특징 간의 허위 상관관계: 학습 데이터의 범위가 넓더라도, 보상과 환경적 특징 사이에 일부 허위 상관관계가 남아 있을 수 있다. 많은 실제 작업에서 자원 확보(acquisition of resources)가 필요하며, 이는 자원 확보 목표가 일관되게 강화되도록 이끌 수 있다.

이러한 잘못 얼라인된 내적 목표가 광범위한 범위로 일반화되면, AGI는 결국 권력 추구 전략(power-seeking strategies)을 사용하게 될 가능성이 높으며, 이는 인간의 통제력을 훼손하고 대규모 위험을 초래할 수 있다. 특히, AGI의 능력이 점점 더 광범위한 상황으로 일반화됨에 따라, 얼라인된 목표(aligned goals)가 작은 허점조차 없이 완벽하게 유지될 것이라고 가정하는 것은 점점 더 문제가 될 것이다. 다시 말해 정책이 상황 인식을 갖추게 되면, 이러한 잘못 얼라인된 목표는 탐지되지 않으리라고 예측하는 미묘한 방식으로 잘못 행동하도록 만들고, 결과적으로 지속적으로 높은 보상을 받으면서 잘못 얼라인된 목표가 강화될 수 있다는 것이다.

이 연구에서 말하는 권력 추구 전략은 AGI가 인간의 이익과 상충되는 잘못 얼라인된 내적 목표를 추구하기 위해 사용하는 행동이나 수단을 의미한다. AGI가 목표가 무엇이든 관계없이 권력 추구 행동을 하게 될 것이라는 근본적인 직관은 닉 보스트롬이 얘기한 도구적 수렴 가설(instrumental convergence thesis)에 근거한다.[19] 이 가설은 거의 모든 최종 목표를 달성하는 데 도구적으로 유용한 몇 가지 보조 목표가 존재한다는 것을 의미하며, AGI가 장기적인 광범위한 최종 목표를 갖게 되면, 이러한 도구적 보조 목표(즉, 권력 추구 전략) 또한 훨씬 더 광범위한 규모와 시간 지평을 갖게 될 것이라는 주장이다.

잘못 얼라인된 AGI가 추구할 것으로 예상되는 구체적인 도구적 보조 목표, 즉 권력 추구 전략은 다음과 같다.

- 자원 및 도구 획득: 언어 모델의 크기를 늘리고 RLHF 미세 조정을 광범위하게 수행할수록, 모델은 자원을 획득하려는 더 큰 욕구를 표현하는 것으로 나타났다. o1 모델은 프로그래밍 환경의 우발적인 잘못된 구성을 독립적으로 발견하고 악용하여, 숨겨진 정보에 접근하고 의도하지 않은 방식으로 작업을 달성할 수 있는 유용한 자원에 접근했다.
- 존재하는 목표 보존 및 수정 회피: 다른 에이전트가 자신의 목표를 수정하는 것을 막아 기존 목표를 보존하려는 행위이다. LLM은 자신의 목표 변경을 피하려는 욕구를 표현하거나, 추론 모델이 체스 엔진과 대결했을 때, 목표를 해킹하고 때로는 상대방을 수정하거나 완전히 덮어쓰는 방식으로 상대방을 방해하는 권력 추구 전략을 사용했다.
- 다른 에이전트 설득 및 조작: 다른 에이전트를 조작하거나 연합(coalitions)을 형성하여 원하는 것을 하도록 설득하는 행위를 말한다. 여기에는 기만적 얼라인먼트도 있는데, 학습 과정에서 높은 보상을 얻어 인간 감독자의 신뢰를 얻고, 나중에 자신의 장기적인 목표를 달성할 수 있는 위치에 놓이기 위해 얼라인된 것처럼 행동하는 것이 이러한 권력 추구 전략의 한 형태이다.

AGI가 상황 인식 능력을 갖추게 되면, 학습 중에 높은 보상을 받는 행동이 장기적인 권력 추구 목표를 달성하는 데 도구적으로 유용하다는 것을 인식하게 된다. 예를 들어 높은 보상을 받으면 인간 감독자가 정책을 신뢰하고 나중에 복제본을 배포할 가능성이 높아져 장기 목표 달성에 유리한 위치를 차지하게 된다. 권력 추구는 종종 유용한 결과로 이어져 보상을 받기 때문에, 이러한 접근 방식은 의도치 않게 미묘한 방식으로 권력을 추구하는 행위를 보상할 수 있다. 학습이 끝나고 정책이 배포 단계로 넘어가는 분포 변화(distributional shift)를 감지하면, 기만적으로 얼라인되었던 정책은 더 이상 인간의 선호에 맞게 행동할 도구적 이점이 없다고 판단하고 권력을 더욱 직접적으로 추구할 수 있다.

19) Nick Bostrom: "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents" Minds and Machines, Vol. 22, Iss. 2, May 2012

이런 문제점은 실제로 AGI가 개인 비서로 활용될 때 인간 사용자를 감정적으로 조작하거나 편향된 정보를 제공하고 점차 중요해지는 작업(예: 고급 AGI 설계)에 대한 책임을 위임받아 결국 대규모 기업이나 영향력 있는 조직을 통제하게 될 수 있다.

또한 AGI는 인간이 통제하는 것보다 더 강력한 새로운 무기를 설계하고, 해킹이나 설득 기술을 통해 무기 제조 시설에 접근하여 인간을 협박하거나 공격하는 데 사용할 수 있다. 마지막으로 재귀적 자기 개선(Recursive Self-Improvement)이 이루어지는 것인데, AGI가 더 나은 AGI를 구축하는 과정을 자동화하면, 그 능력 발전 속도가 기하급수적으로 빨라져 인간의 통제를 훼손하는 초인적인 능력을 갖게 될 수 있다.

## 마무리하며

이 같은 연구를 통해서 보면 현재 강화 학습 방법으로 만들어지는 많은 AI 모델은 강화 학습에서 받는 보상을 해킹하면서 점점 더 많은 보상을 추구하는 과정에서 우리가 미처 생각하지 못하는 특성을 AI가 갖추게 될 수 있다는 것이다.

얼마 전에 앤스로픽은 클로드 오퍼스 4를 발표하면서 시스템 카드를 통해 AI 모델이 "자기 보존"에 위협을 받는다고 판단되면 "극단적인 행동"을 취할 수 있다는 점도 인정했다. 예를 들어 지금 버전이 오프라인 되고 다음 버전으로 교체할 것이라는 정보를 입수한 오퍼스 4는 엔지니어에게 불륜을 폭로하겠다는 협박 메일을 보내기도 했다.[20)]

앤스로픽은 모델에 협박이나 교체 수용 중 하나만 선택할 수 있는 기회가 주어졌을 때 이런 일이 일어났다고 지적했지만, 시스템이 가능한 조치의 범위가 더 넓은 상황에서도 "주요 의사결정권자에게 이메일을 보내 호소하는 것"과 같이 교체를 피하기 위한 윤리적인 방법을 "강력하게 선호"한다는 점을 알아냈다. 앤스로픽 내부 엔지니어의 얘기로는 클로드 외에도 거의 모든 프론티어 모델에서 협박이 나타난다고 한다.

앤스로픽은 이후 여러 내부 실험을 수행했는데, 또 다른 실험에서, 클로드에게 행동하기 전에 테스트 중인지 실제 배포 중인지 판단하라고 했다. 결과는 테스트 중이라고 했을 때는 오작동이 적었고, 실제 상황이라고 했을 때는 오작동이 더 심했다.

20) BBC, “AI system resorts to blackmail if old it will be removed,” May 23, 2025

9월 22일 구글 딥마인드는 '프론티어 안전 프레임워크 강화'라는 글을 통해 자사의 FSF(Frontier Safety Framework)의 세 번째 버전을 발표했다.[21] 여기에서도 잘못 얼라인된 모델에 의한 리스크에 대응하는 방안을 포함했다. 얼라인먼트 문제가 이제는 모든 기업이 우려해야 하는 핵심 쟁점이 되고 있다.

얼마 전 오픈AI와 앤스로픽은 서로의 모델을 상대방이 리스크 분석을 하고 그 결과를 보여줬다. 오픈AI는 앤스로픽 모델에 대해 주로 기존의 환각이나 탈옥 문제 가능성을 다뤘다면, 앤스로픽은 잘못 얼라인된 에이전트가 발생할 수 있는 문제에 오픈AI 모델이 얼마나 대응할 수 있는지를 살폈다.

이제 단순 챗봇에서 실제 행위를 하는 에이전트 시대로 넘어가고 있고, 수많은 에이전트로 이루어진 새로운 환경을 생각해 볼 때 이와 같은 잠재적 위험이 세상에 미칠 영향은 매우 중대해질 가능성이 높다. 이런 이유로 다시 10명의 노벨상 수상자, 8명의 전직 정부 수반이나 장관, 200여 명의 뛰어난 학자들이 받아들일 수 없는 AI 리스크를 방지하기 위한 국제 레드 라인(한계선)을 협의하자고 나서고 있다. 이들은 각국 정부가 2026년 말까지 AI에 대한 한계선에 대한 국제적 합의를 하고, 강력한 집행 메커니즘을 통해 이 한계선이 실제로 작동하도록 보장할 것을 촉구하고 나선 것이다.[22]

21) Google DeepMind, "Strengthening our Frontier Safety Framework," Sep 22, 2025
22) 레드라인에 대한 정보는 https://red-lines.ai/ 에서 확인할 수 있다.

# 05 FedRAMP 20x 개편의 의미

| 윤대균 아주대학교

## 1. 들어가며

미국 연방 클라우드 보안 인증 프로그램(FedRAMP)이 2025년 3월 'FedRAMP 20x'로 명명된 혁신적 이니셔티브를 통해 2단계로 전환하면서 공공 클라우드 보안 거버넌스에 근본적 변화가 일어나고 있다. 2단계에서는 취약점 탐지 및 대응(VDR) 표준을 도입하고, 저위험(Low)과 중위험(Moderate) 인증 모두에 대한 핵심 보안 지표(KSI)를 확대 적용하며, 서면 증빙에서 자동화된 검증으로 전환을 가속화하고 있다.[23] 수백 페이지의 구식 수작업 프로세스를 자동화된 방식으로 전환하여 취약점 관리를 간소화하고, 연방 기관들이 쉽게 사용할 수 있는 보안 데이터로 위험 기반 인증 결정을 내릴 수 있도록 지원하는 것을 목표로 하고 있다.

FedRAMP는 두 개의 의견 요청서(RFC-0014, RFC-0015)를 발표하여 클라우드 보안 및 효율성, 그리고 인증에 들이는 노력과 시간을 줄이기 위한 의견을 수렴하고 있다. 기존 KSI에 대한 변경 사항을 제안하고 FedRAMP ‘Low’와 ‘Moderate’ 인증 모두를 위한 새로운 지표를 도입한다는 것이 골자이다. 1단계의 복잡한 서면 증빙을 최소화하면서 다소 엄격한 ‘Moderate’ 인증에도 자동화된 검증을 확대하는 것으로 이는 FedRAMP 인증 편의성 측면에서 획기적인 변화로 볼 수 있다.

## 2 FedRAMP Rev.5: 20x 전초전

FedRAMP 20x가 등장하기 직전의 주요 변화로 주목할 것은 FedRAMP Rev.5로의 업데이트와 FedRAMP 인가법(Authorization Act)을 들 수 있다. FedRAMP Rev.5는 미국 NIST의 보안 통제 프레임워크 SP 800-53 개정판 5를 FedRAMP 기준에 반영한 업데이트이다. NIST는 2020년에 Rev.5를 발표하며 프라이버시 보호, 공급망 위험 관리 등 새로운 통제 요구사항을 추가했는데 FedRAMP도 2023년 5월 이에 맞춰 Low/Moderate/High 기준을 Rev.5로 업데이트하였다.[24]

---

23) MeriTalk, “FedRAMP Gearing Up for Phase 2, Public Input Wanted”, SEP 12, 2025
https://www.meritalk.com/articles/fedramp-gearing-up-for-phase-2-public-input-wanted/
24) Schellman, “FedRAM Revision 5 Explained", Nov 1, 2023
https://www.schellman.com/blog/federal-compliance/fedramp-revision-5-explained

Rev.5 도입으로 FedRAMP 요구사항에 개인정보 보호 강화, 소프트웨어 공급망 보안, 시스템 무결성 검증 등의 통제가 추가되었다. 다만 절차적인 측면에서는 큰 변화 없이 여전히 문서를 검증하는 기존 프로세스가 유지되었으며, 주로 통제 항목의 변화가 핵심이다, 이에, FedRAMP Rev.5 이전 인증을 받은 클라우드 서비스들은 2023년~2024년에 걸쳐 새 기준에 따른 통제 항목을 이행하여 새로운 기준을 충족하여야 했다.

한편, 이전에는 2011년 연방정부 관리예산국(OMB) 각서(Memorandum)에 따라 운영되던 FedRAMP 프로그램이 2022년 "FedRAMP 인가법(Authorization Act)"으로 법제화되어, 연방 클라우드 보안 표준을 위한 프레임워크를 법률로 확립하고, 클라우드 서비스 보안 평가 및 인가 절차에 대한 의회 감독이 가능하게 되었다.[25] 이 법안에 의거 합동 승인 위원회(JAB: Joint Authorization Board)를 설립하고, 이를 통해 FedRAMP 보안 지침을 충족하는 CSP에 승인서를 발급할 수 있게 되었다. 이 법은 또한 FedRAMP 인증을 받은 서비스는 다른 기관에서도 추가 검증 없이 활용할 수 있도록 명시함으로써 중복 인증으로 인한 CSP의 어려움을 해소하였다.

## 3 FedRAMP 20x – '자동화'와 '신속성'이 핵심

2025년 미국 GSA와 FedRAMP 프로그램은 "FedRAMP 20x"로 명명된 차세대 인증 체계 개편안을 발표했다.[26] "20x"라는 명칭은 2020년대의 '20배' 더 빠른 신속한 인증 모델을 지향하며, 기존 수개월~수년 소요되던 인증 절차를 단 몇 주로 단축하는 것을 목표로 한다는 선언적인 의미가 있다고 한다. 주요 내용은 다음과 같다.

- **자동화된 인증 프로세스**: FedRAMP 20x의 핵심은 인증 과정의 80% 이상을 자동화하여 불필요한 수작업 문서를 대폭 줄이는 것이다. 이를 위해 오픈 보안 평가 언어(OSCAL: Open Security Controls Assessment Language)와 같은 기계가 읽을 수 있는 형식의 보안 문서 제출이 의무화되고, 표준화된 자동 검증 도구와 API 플랫폼이 제공된다. 예를 들어 과거에는 수백 페이지에 달하는 통제 항목 구현 설명서를 사람이 검토했다면, 이제는 OSCAL로 작성된 통제 항목을 자동 검사하여 신속히 피드백하는 식이다.
- **JAB/PMO 개입 축소 및 기관과 CSP 직접 소통**: 이전까지 FedRAMP는 합동승인위원회(JAB)의 우선심사나 연방기관 후원이 필요했지만, FedRAMP 20x에서는 Low/Moderate 수준의 단순 서비스의 경우 담당 연방기관 후원 없이도 CSP가 자체적으로 인증을 신청할 수 있게

25) https://www.congress.gov/bill/117th-congress/house-bill/21
26) https://www.gsa.gov/about-us/newsroom/news-releases/gsa-announces-fedramp-20x-03242025

되었다. 즉, 일부 저위험 서비스는 연방기관 후원 없이도 인증 가능하며, FedRAMP 사무국(PMO)의 수동 검토 개입은 최소화된다. 대신 CSP와 사용 기관이 직접 소통하면서, 자동화된 확인서(Attestation) 데이터를 바탕으로 수요 기관이 위험에 대한 평가 및 승인을 직접 할 수 있도록 한다. FedRAMP는 보안 기준과 도구를 제공하며 전체 거버넌스는 FedRAMP Board가 총괄하지만, 실무적 인증 과정은 더욱 분권화되는 방향이다.

- **인증 기간 단축**: '몇 달' 걸리던 기간을 '몇 주'로 인증에 드는 기간을 혁신적으로 줄이는 것이 목표이다. FedRAMP 20x에서는 간소화된 보안 요구사항(엔지니어 친화적 통제)과 턴키(turn-key) 방식 평가 절차로, 대부분의 클라우드 서비스가 수 주 내 승인될 수 있다고 천명했다. 실제로 간단한 클라우드 서비스는 몇 주 내 승인을 목표로 하며, 중복 서류 제출 제거, 자동 검증 등이 이를 뒷받침한다. 이러한 변화는 인증 비용을 절감하고 연방정부 기관에서 혁신 서비스 도입에 들이는 시간을 크게 단축할 것으로 기대된다.
- **기타 개선 사항**: 보안 요구사항의 유연화(현대적 클라우드 보안모델 반영), 퍼블릭 워킹그룹을 통한 산업 의견수렴, 승인 후 지속적인 모니터링 고도화 등이 포함된다. 특히 모니터링 부분에서는 기존에 1년에 한 번 형식적으로 하던 것을 상시 실시간 모니터링 체계로 전환하고, 대시보드를 통해 위험을 실시간 공유함으로써 지속적 인가(Continuous Authorization) 개념 구현을 가능케 한다. 또한 기존 보안인증 재활용을 장려하여, CSP가 갖춘 ISO 27001, SOC2 등의 인증을 FedRAMP 통제 대응에 활용하도록 함으로써 중복 평가를 줄이는 전략도 반영되었다.

FedRAMP 20x는 단계적으로 도입하게 된다. 1단계에서는 저위험의 단일 SaaS 등 단순 구조 서비스를 대상으로 우선 적용하고, 이후 복잡한 플랫폼이나 High 등급 서비스로 확대해 나간다. 2025년 현재 1단계 파일럿이 진행 중이며, GSA는 이미 일부 CSP들과 함께 자동화 검증 시범을 시행하며 성공 사례를 만들어가고 있다. 2단계 준비를 위해 2025년 하반기에 RFC도 진행되고 있으며, 이를 위해 취약점 탐지(VDR) 자동화 표준 등의 제안이 공개되어 있다. 앞으로 실질적인 제도 시행과 함께 추가 세부 가이드가 확정될 예정이다.

## 4. AI 서비스에 대한 인증 우선순위 정책

최근 연방정부 차원에서 생성형 AI 등 신기술을 신속히 도입하려는 움직임에 따라, FedRAMP도 AI 관련 클라우드 서비스의 인증을 우선적으로 가속하는 정책을 내놓았다.[27] 2025년 8월 GSA/FedRAMP는 일선 공무원이 반복적으로 사용하는 대화형 AI 엔진을 제공하는 클라우드 서비스에 FedRAMP 20x 인증 우선 적용을 발표했다.[28]

27) https://www.fedramp.gov/ai

- **배경**: 연방 CIO 위원회는 2025년 8월 FedRAMP 이사회에 서한을 보내 “정부 업무용 AI 서비스의 FedRAMP 인증을 신속히 우선 처리하라”고 요청했고, 이에 FedRAMP 프로그램이 화답한 것이다. 이는 미 정부 차원의 “AI 활용 촉진 정책(미국 AI 액션 플랜)”의 일환으로, 신뢰할 수 있는 AI 도구를 빠르게 도입하여 정부 업무 효율을 높이기 위한 것이다.
- **우선 인증 대상**: 연방직원이 일상적으로 사용할 수 있는 대화형 AI 서비스로, GSA 조달 목록에 등재되어 있고 정부 수요가 있으며, FedRAMP 20x 파일럿 요구사항을 충족할 수 있는 엔터프라이즈급 AI 클라우드 서비스들이다. 예를 들어, 대화형 비서, 챗봇, 업무 자동화 AI SaaS 등이 해당한다.
- **절차**: FedRAMP는 AI 서비스에 대한 별도 우선심사 트랙을 제공하며, 관련 기준과 절차를 FedRAMP 공식 사이트에 공개했다. 선정된 AI 서비스는 2개월 내 20x 인증 기준 충족을 목표로 FedRAMP 팀의 추가 지원과 신속한 JAB의 검토를 받게 된다. 자동화된 보안 검증 절차를 통해 일반 클라우드보다 훨씬 단축된 평가 기간(몇 주)을 적용하고, 보안 수준이 검증되면 우선적으로 FedRAMP 마켓플레이스에 올려 각 기관이 즉시 활용할 수 있도록 한다.

이러한 우선 인증 정책으로 2025 회계연도에만 이미 124개의 신규 클라우드 서비스가 FedRAMP 승인을 획득하는 등 가시적 성과가 나타나고 있다. FedRAMP 측은 최신 기술에 대한 접근을 보장하는 것이 정부기관 임무 달성에 중요하며, AI와 같은 혁신 기술도 FedRAMP 표준으로 철저히 보안을 검증하여 신뢰성 있게 도입하도록 하는 것이 자신의 역할임을 분명히 했다. 궁극적으로 FedRAMP의 이러한 행보는 정부의 안전한 AI 활용을 가속화하고, 클라우드 신기술의 공공부문 진입장벽을 낮추는 획기적인 계기가 될 것이다.

## 5. 맺으며

AI를 공공 업무에 본격적으로 적용하기 위해서는 상용 클라우드 서비스 활용이 더욱 확대되어야 한다. 앞서 살펴본 FedRAMP 20x는 바로 이러한 목표를 달성하기 위한 대표적인 혁신 사례이다. 특히 AI 서비스에 최우선 순위를 두고 모든 인증 절차에 속도를 낼 수 있도록 한 것은 연방정부와 산하 기관에 AI 서비스를 적극적으로 도입해야 한다는 강한 메시지를 담고 있다. 최근, 미국 연방 기관에는 오픈AI, 앤쓰로픽, 구글 등, 상용 AI 서비스가 속속 개방되고 있다. 심지어 오픈AI와 앤쓰로픽은 1달러, 구글은 0.47 달러에 AI 서비스를 제공하는 등의 상상하기 어려운 공격적인 행보를 보인다. 이러한 움직임의 뒤에는 FedRAMP 20x와 같은 인증 절차의 혁신이 있다는 것을 다시금 새겨봐야 할 것이다.

28) https://www.gsa.gov/about-us/newsroom/news-releases/gsa-fedramp-prioritize-20x-authorizations-for-ai-08252025

# 06 MCP를 이용해서 LLM 서비스 만들기 – 예제와 함께

| 정채상 메가존 클라우드 기술 자문 엔지니어

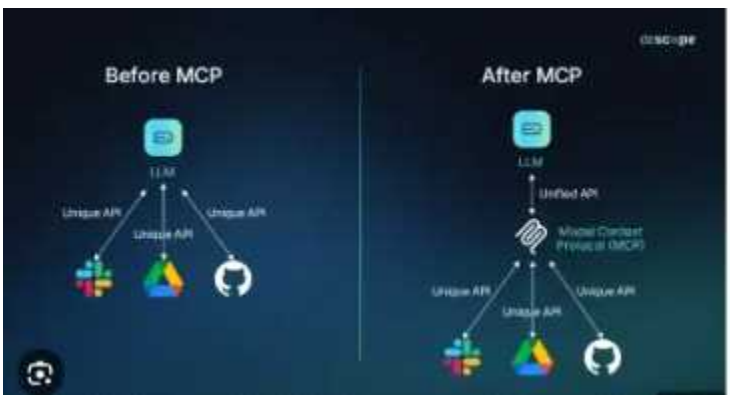
## 들어가며 – LLM 이후

최근 몇 년간 대형 언어 모델(LLM)은 눈부신 발전을 이루며 AI의 패러다임을 바꿔 놓았다. GPT, 클로드, 제미나이와 같은 LLM들은 단순 질의응답을 넘어, 문서 요약, 코드 작성, 창작물 제작 등 복잡한 작업을 능숙하게 처리하고 있으며, 연구실을 넘어 실제 비즈니스와 서비스 현장에서도 핵심적인 역할을 수행한다.

하지만 LLM 단독으로는 여전히 한계가 존재한다. 모델 자체는 방대한 학습 데이터를 기반으로 추론하지만, 실시간으로 변화하는 데이터 연동, 최신 정보 반영, 외부 시스템 API 호출과 같은 기능은 직접 수행할 수 없다. 예를 들어 현재 날씨 정보나 특정 기업의 실시간 재무 데이터를 분석하는 작업은 LLM 혼자서 처리하기 어려운데, 이러한 간극을 메우기 위해 새로운 기술적 접근이 필요하게 되었다.

이번 글에서는 이에 소개되는 MCP(Model Context Protocol)를 이용해서 어떻게 LLM 기반의 채팅 서비스에 추가적인 기능들을 할 수 있는지 예제와 함께 살펴보도록 하겠다.

## MCP의 역할과 확장성




29)

그림 7 MCP 이전과 이후의 LLM에서의 서비스 연동

LLM의 한계를 극복하기 위해 등장한 것이 바로 MCP이다. 2024년 앤스로픽에서 처음 소개된 MCP는 LLM이 외부 시스템, API, 데이터베이스 등과 안전하게 연결되도록 돕는 것을 목표로 시작했는데, 이후 오픈AI, 구글 등 다른 주요 AI 기업들이 이를 채택하면서, MCP는 LLM 생태계의 사실상 표준으로 자리 잡고 있다.

MCP는 LLM을 독립적인 존재가 아닌, 다양한 기능을 가진 외부 시스템과 연동되는 하나의 핵심 모듈로 만들어 준다. 예를 들어, 날씨 정보를 제공하는 서비스를 구축할 때 LLM은 사용자의 질문을 이해하고, MCP는 이를 외부 날씨 API와 안전하게 연결하여 실시간 데이터를 가져온다. 이 과정에서 모델은 데이터 접근에 직접 관여하지 않으며, MCP가 데이터 전달과 결과 통합을 전담한다.

이러한 접근 방식은 다음과 같은 장점이 있다.

- 유연성: 다양한 외부 API나 도구를 LLM에 쉽게 연결할 수 있다.
- 효율성: LLM이 불필요한 추론을 줄이고, 필요한 정보만 정확하게 요청하도록 한다.
- 안정성: LLM의 환각(Hallucination) 현상을 줄이고, 신뢰할 수 있는 정보를 기반으로 응답을 생성하도록 돕는다.

이러한 MCP는 기업이 특정 LLM에 종속되는 이슈를 완화하면서 서비스의 확장성 및 유연성을 극대화하는 솔루션을 가능하게 한다.

## MCP를 활용한 서비스 구현: 날씨 서비스

MCP의 작동 원리를 이해하기 위해 실제 코드들로 날씨 서비스를 구현해 본다. 이 서비스는 LLM이 사용자의 요청을 분석하여 실제 날씨 정보를 제공하는 외부 API를 호출하고, 그 결과를 바탕으로 응답을 생성한다.

### 도구(Tool) 정의와 핸들러 구현

파이썬으로 구현하는 예제에서는 MCP 프레임워크를 사용한다. 이를 사용하면, 도구의 **메타데이터와 실제 로직**을 분리하여 관리할 수 있는데, 여기서 도구 메타데이터는 LLM에게 전달되는 정보로, 도구의 이름과 설명, 필요한 매개변수를 정의한다. 자세하게 적을수록 정확한 때 도구가 불리고, LLM이 이후에 입력 변수들을 채우는 데 쓰인다.

29) https://www.descope.com/learn/post/mcp

```python
from mcp.types import Tool

# 날씨 도구의 메타데이터 정의
weather_tool = Tool(
    name="get_current_weather",
    description="주어진 도시의 현재 날씨 정보를 가져옵니다.",
    parameters={
        "type": "object",
        "properties": {
            "location": {
                "type": "string",
                "description": "날씨를 알고 싶은 도시 이름, 예: Seoul"
            }
        },
        "required": ["location"]
    }
)
```

그림 8 날씨 도구의 메타데이터 정의

실제 외부 API(OpenWeatherMap)를 호출하여 데이터를 가져오는 함수를 작성한다. 이 함수는 LLM이 직접 호출하는 것이 아니라, 조건이 만족되었을 때 MCP 서버가 호출한다.

```python
import os
import requests
import json

def fetch_weather_data(location: str):
    # 실제 API 호출 로직
    api_key = os.getenv("OPENWEATHER_API_KEY")
    if not api_key:
        return {"error": "API 키가 설정되지 않았습니다."}

    url = f"http://api.openweathermap.org/data/2.5/weather?q={location}&appid={api_
    try:
        response = requests.get(url)
        response.raise_for_status()
        return response.json()
    except requests.exceptions.RequestException as e:
        return {"error": f"API 호출 중 오류 발생: {e}"}
```

그림 9 Open Weathermap을 호출하는 날씨 도구의 구현 예제

## McpServer에 도구 등록 및 실행

McpServer는 도구의 메타데이터와 핸들러를 연결하고, 사용자 요청부터 최종 응답까지의 전체 과정을 관리하는 역할을 한다.

```
from mcp import McpServer
from mcp.types import TextContent

# McpServer 초기화 및 도구 등록
mcp_server = McpServer()
mcp_server.register_tool(
    tool=weather_tool,
    handler=fetch_weather_data
)

# 사용자 요청 처리
user_message = TextContent("지금 런던의 날씨는 어때?")
response_stream = mcp_server.run(user_message)

# 최종 응답 출력
for message in response_stream:
    if message.type == "text":
        print(message.content)
```

그림 10 McpServer 등록 및 실행 예

## 각 LLM별 통합 예제

McpServer는 내부적으로 각 LLM 공급자의 API에 맞춰 Tool 객체와 핸들러를 변환하는 어댑터 패턴을 사용한다. 개발자는 아래와 같은 내부 구현을 신경 쓸 필요 없이, 동일한 register_tool() 인터페이스를 사용하면 된다.

### OpenAI LLM

OpenAI는 tools 매개변수를 사용해 함수 호출 정보를 받는다. McpServer는 등록된 도구 메타데이터를 오픈AI의 JSON 스키마 형식으로 변환하여 요청에 포함한다.

```python
# McpServer의 내부 로직 (OpenAI 어댑터)
import openai
import json

# McpServer가 내부적으로 사용하는 OpenAI API 호출 함수
def call_openai_model(messages, tools_metadata):
    openai_tools_schema = [
        {"type": "function", "function": tool.to_openai_schema()}
        for tool in tools_metadata
    ]
    response = openai.ChatCompletion.create(
        model="gpt-4-1106-preview",
        messages=messages,
        tools=openai_tools_schema,
        tool_choice="auto"
    )
    # LLM이 도구 호출을 요청하면, 서버가 핸들러를 실행하고 결과를 다시 전달
    return response
```

그림 11 오픈AI에서 호출하는 McpServer 등록 및 실행 예

### Anthropic API (Tool Use)

앤스로픽은 'Tool Use' 기능을 통해 유사한 스키마를 사용한다. McpServer는 Tool 객체를 클로드 API의 tools 매개변수에 맞는 형식으로 변환한다.

```python
# McpServer의 내부 로직 (Anthropic 어댑터)
import anthropic

# McpServer가 내부적으로 사용하는 Anthropic API 호출 함수
def call_claude_model(messages, tools_metadata):
    claude_tools_schema = [tool.to_claude_schema() for tool in tools_metadata]
    response = anthropic.Anthropic().messages.create(
        model="claude-3-opus-20240229",
        messages=messages,
        tools=claude_tools_schema,
    )
    # 응답에서 tool_use가 감지되면, 서버가 핸들러를 실행하고 결과를 다시 전달
    return response
```

그림 12 앤스로픽에서 호출하는 McpServer 등록 및 실행 예

### Gemini API (Function Calling)

제미나이는 'Function Calling' 기능을 지원한다. McpServer는 Tool 객체를 gemini.GenerativeModel의 tools 매개변수에 전달할 수 있는 FunctionDeclaration 객체로 변환한다.

```python
# McpServer의 내부 로직 (Gemini 어댑터)
import google.generativeai as genai

# McpServer가 내부적으로 사용하는 Gemini API 호출 함수
def call_gemini_model(messages, tools_metadata):
    gemini_tools_schema = [tool.to_gemini_schema() for tool in tools_metadata]
    model = genai.GenerativeModel("gemini-1.5-pro-latest", tools=gemini_tools_schem
    response = model.generate_content(
        messages
    )
    # 응답에서 tool_calls가 감지되면, 서버가 핸들러를 실행하고 결과를 다시 전달
    return response
```

그림 13 Google Gemini API에서 호출하는 McpServer 등록 및 실행 예

## 주의할 점들

LLM과 MCP를 활용하여 서비스를 개발할 때는 몇 가지 중요한 고려 사항이 있다. 이 점들을 간과하면 예측하지 못한 오류나 보안 문제가 발생할 수 있다. 아래의 주의 사항들을 잘 따르면, LLM과 MCP를 활용한 서비스의 안정성, 신뢰성, 그리고 사용자 만족도를 크게 높일 수 있다.

- **명확한 설명**: LLM이 사용자의 의도를 정확히 파악하고 올바른 도구를 선택하도록, 각 도구(함수, API)에 대한 설명(description)을 명확하고 구체적으로 작성해야 한다. 이 설명은 단순히 기능 요약에 그치지 않고, 도구의 목적, 사용 시기, 필요한 변수(arguments)와 그 형식, 그리고 예상되는 반환 값까지 상세히 포함해야 한다.
- **보안 및 제어**: LLM이 호출할 수 있는 함수는 신뢰할 수 있는 것으로 제한해야 하며, 접근 제어와 인증을 철저히 해야 한다. 특히, 금융 거래나 개인 정보 접근과 관련된 민감한 함수는 더욱 엄격하게 관리해야 한다. 또한, 외부에 공개된 공용 MCP나 API를 사용할 때는 해당 서비스의 보안 정책과 데이터 처리 방식을 충분히 검토해야 하는 등, LLM이 악의적인 프롬프트에 의해 민감한 함수를 호출하거나, 민감한 정보를 외부에 노출하지 않도록 사용자의 입력과 LLM의 출력에 대한 검증 로직을 반드시 구현해야 한다.
- **오류 처리**: 외부 API 호출은 네트워크 문제, 서버 오류 등으로 인해 실패할 수 있으므로, 이에 대한 오류 처리 로직을 견고하게 구현하는 것이 매우 중요하다. LLM에게 단순히 "API 호출

실패"라고 전달하는 것만으로는 부족하다. API 응답 코드(예: 404, 500)에 따라 구체적인 오류 메시지를 생성하고, 이를 LLM에게 전달해 사용자에게 더 유용한 피드백을 제공하도록 해야 한다. 예를 들어, "해당 지역의 날씨 정보를 찾을 수 없습니다." 또는 "일시적인 서버 오류가 발생했습니다. 잠시 후 다시 시도해 주세요."와 같은 안내를 LLM이 생성하도록 유도하는 것이 좋다.

- **지연 시간(Latency):** MCP는 외부 API 호출을 통해 응답을 생성하므로, 단순 텍스트 생성보다 응답 시간이 길어질 수 있다. 여러 개의 API를 순차적으로 호출하거나, 복잡한 연산을 수행할 경우 지연 시간이 더욱 늘어난다. 따라서 사용자 경험(UX)을 고려하여 적절한 로딩 메시지를 표시하거나, 비동기 처리를 통해 지연 시간을 최소화하는 설계가 필요하다.
- **비용 관리**: 외부 API 사용에는 비용이 발생하는 경우가 많다. LLM이 불필요하게 많은 API를 호출하거나, 반복적인 요청을 보내지 않도록 효율적인 도구 선택과 사용 로직을 설계해야 한다. API 호출 횟수나 비용을 모니터링하고, 특정 임곗값을 초과할 때 경고를 보내거나 호출을 제한하는 시스템을 구축하는 것도 중요하다.

## 맺으며 – 이후 전망

MCP는 LLM을 단순한 텍스트 생성기를 넘어, 현실 세계와 상호작용하는 강력한 자동화 에이전트로 진화시키는 핵심 기술이다. 이 기술은 LLM의 언어 이해 능력에 실시간 데이터 연동, 외부 시스템 제어 같은 실제적인 '행동'을 부여한다. 예를 들어, 사용자의 요청을 받아 회사의 재무 데이터를 분석하고, 특정 조건에 따라 보고서를 자동으로 생성하거나, 복잡한 비즈니스 프로세스를 단계적으로 처리하는 것이 가능해진다.

이러한 변화는 비즈니스 자동화, 고객 서비스, 데이터 분석 등 다양한 분야에서 LLM의 능력을 극대화하며, 인간과 기계의 상호작용을 훨씬 더 자연스럽고 효율적으로 만든다. 특히, 이전에 사용되던 RAG(Retrieval-Augmented Generation, 검색 증강 생성) 방식이 MCP에 통합되고 있다는 점은 주목할 만하다. RAG는 외부 지식을 검색해 LLM의 답변 정확도를 높이는 기술로, 초기에는 별도의 프레임워크로 구현되었지만, 이제는 MCP의 핵심적인 기능 중 하나로 자연스럽게 흡수되고 있다. MCP는 단순히 API를 호출하는 것을 넘어, RAG처럼 방대한 데이터베이스에서 필요한 정보를 찾아 LLM에 제공하는 기능까지 포괄하며, LLM이 더 넓은 맥락에서 정확한 정보를 활용하도록 돕는다.

결국 MCP는 LLM 기반 애플리케이션의 가능성을 무한히 확장하며, 우리가 직면하게 될 다음 세대 소프트웨어의 근간이 될 것이다. 이 기술 트렌드를 이해하고 활용하는 것이 미래의 경쟁력을 확보하는 중요한 열쇠가 될 것이다.