TANGO

GenAI 구축 전주기 지원 LLMOps 플랫폼

성 명 염익준 소 속 (주)아크릴

2025 제4회 COMMUNITY CONFERENCE

TANGO

Target Adaptive No-code neural network Generation and Operation framework

주 관 ETRI (TANGO)
주 최 과학기술정보통신부 IITP 정보통신기획평가원
문 의 parkjb@etri.re.kr / 042-860-5565

후 원 LOC lablup weda tesla system (사)한국인공지능협회 SNUH 고려대학교 KOREA UNIVERSITY 홍익대학교 HONGIK UNIVERSITY 중앙대학교 RTst KETI 한국전자기술연구원 AIVN SUREDATA ACRYL 하늘소프트 TTA 한국정보통신기술협회 Telecommunications Technology Association

content

목 차

GenAI 현황 및 도전 과제

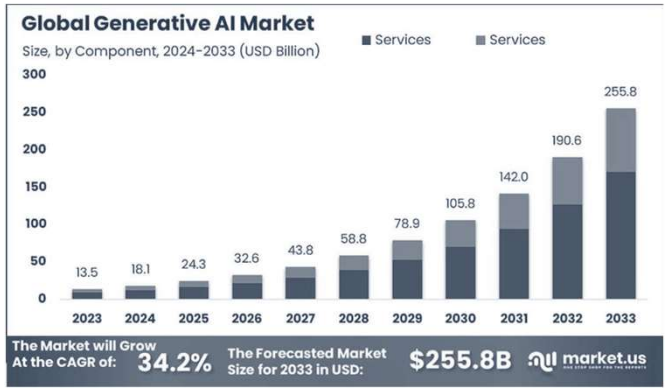
생성형 AI의 활용이 빠르게 확산되고 있으나,
실제 도입 과정에서 많은 기업과 기관이 **데이터 품질 관리, 전문 인력 확보, 운영 비용** 등 다양한 한계에 부딪히고 있습니다.

GenAI의 확산 및 파급력

[AI리포트] "생성형 AI 활용 국내 기업 2025년 55.7%에서 2026년 85%로 증가 전망"

챗GPT, 공무원 되다…미국 주정부, 생성형 AI로 예산 40억 달러 절감

생성형 AI 이용률 33%...지난해 대비 2배 증가

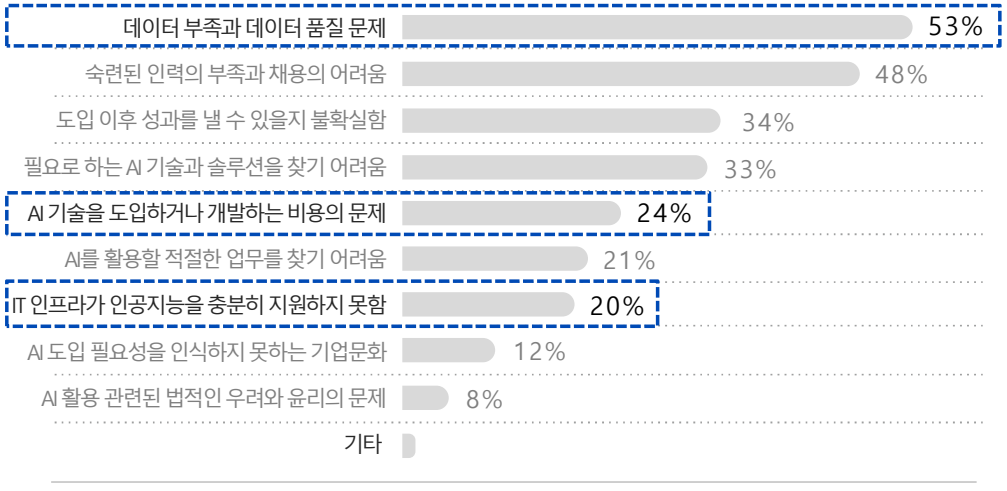


USD 13.5 Billion in 2023,

CAGR 34.2%

USD 255.8 Billion in 2023,

AI 도입 및 운영 과정에서 직면하는 문제



※ 출처: 삼성 SDS, "2023 국내 AI 도입 및 활용 현황 조사"

LLMOps란 무엇인가?

데이터 품질 관리, 인력 확보, 운영 효율성 등의 문제를 해소하기 위한 체계적 접근으로 **LLMOps**가 부상하고 있으며,
이는 **대규모 언어모델(LLM)의 개발, 배포, 운영 전 과정을 표준화·자동화**하여 모델의 품질 일관성과 생산성을 동시에 확보합니다.

LLMOps vs. MLOps 비교표

구분	MLOps (Machine Learning Ops)	LLMOps (Large Language Model Ops)
목표	머신러닝 모델의 학습·배포·운영 자동화	대규모 언어모델의 생성형 응답 품질 관리 및 최적화
주요 대상 모델	예측·분류 중심의 전통 ML 모델 (e.g, RandomForest, CNN 등)	대규모 사전학습 언어모델 (GPT, Qwen, Llama 등)
핵심 단위	데이터셋·파이프라인·모델 파라미터	프롬프트·컨텍스트·모델 버전·출력 응답
운영 초점	학습 재현성, 배포 자동화, 성능 모니터링	응답 품질, 비용·지연시간 관리, 프롬프트 버전관리
데이터 관리	Feature Store 중심 (입력데이터 관리)	Prompt Store & Vector Store 중심 (컨텍스트 관리)
평가 방식	정량적 지표 중심 (Accuracy, F1, ROC 등)	정성·정량 혼합 (Hallucination, Faithfulness, Coherence 등)
실험 관리	모델 학습 실험 추적 (MLflow 등)	프롬프트·모델 조합 실험 추적 (LangFuse, Helicone 등)
관측(Observability)	모델 성능·리소스 사용량 모니터링	응답 품질·프롬프트 로그·사용자 피드백 모니터링

LLMOps 도입의 기대효과

LLMOps를 도입함으로써 기업은 **모델 품질 향상, 운영 효율화, 거버넌스 강화, 서비스 확장성 확보** 등 생성형 AI 운영의 핵심 과제를 해결하고 지속 가능한 성과를 달성할 수 있습니다.

모델 성능 및 품질 최적화

- ✓ 프롬프트 설계, 데이터 파이프라인, 파인튜닝을 통한 모델 성능 향상
- ✓ 자동화된 평가 루프와 피드백 반영으로 응답 품질 지속 개선
- ✓ 실제 서비스 지표 기반의 품질 최적화 및 개선 주기 단축

"Vertical LLM 확보"

운영 자동화 및 워크플로우 효율화

- ✓ 데이터 수집 → 학습 → 검증 → 배포 전 과정 자동화
- ✓ 지속적 모니터링 및 품질 검증 프로세스 내재화
- ✓ 개발–운영 간 사이클 단축 및 운영 효율 극대화

"인력 투입 비용 절감"

데이터 및 모델 거버넌스 강화

- ✓ 데이터 버전·출처·품질 이력 관리로 재현성과 추적성 확보
- ✓ 모델 버전 및 학습 이력 관리로 신뢰성 강화
- ✓ 개인정보 보호, 규제 준수, 편향 데이터 방지를 위한 통제 체계 마련

"리스크 최소화"

서비스 확장성 확보

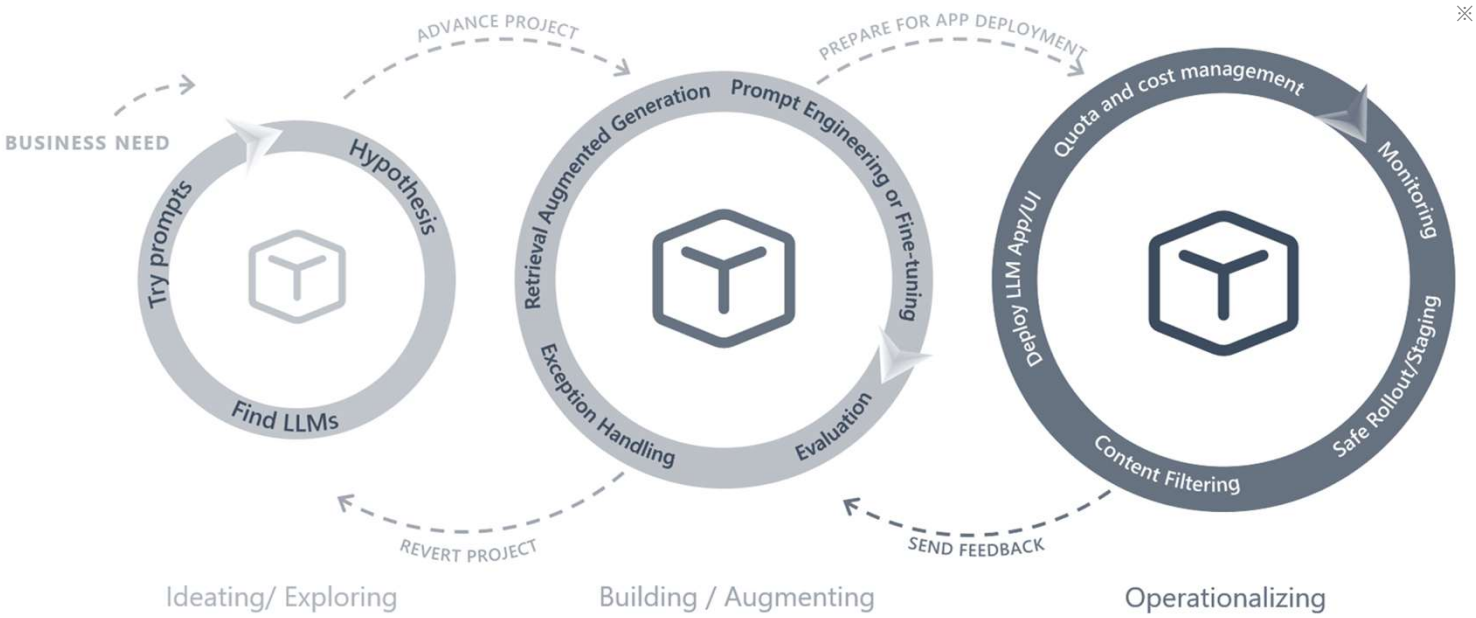
- ✓ 파이프라인 기반으로 다양한 모델·도메인·언어로 확장 용이
- ✓ 새로운 서비스나 환경으로의 빠른 적용 가능
- ✓ 확장 시에도 일관된 품질·안정성 유지

"다각적 수익 창출"

LLM 라이프사이클

LLMOps는 **데이터셋 구축부터 학습, 검증, 배포, 운영에 이르는 전 과정을 유기적으로 연결**하여,
LLM 라이프사이클 전반의 품질 관리와 재현성을 보장합니다.

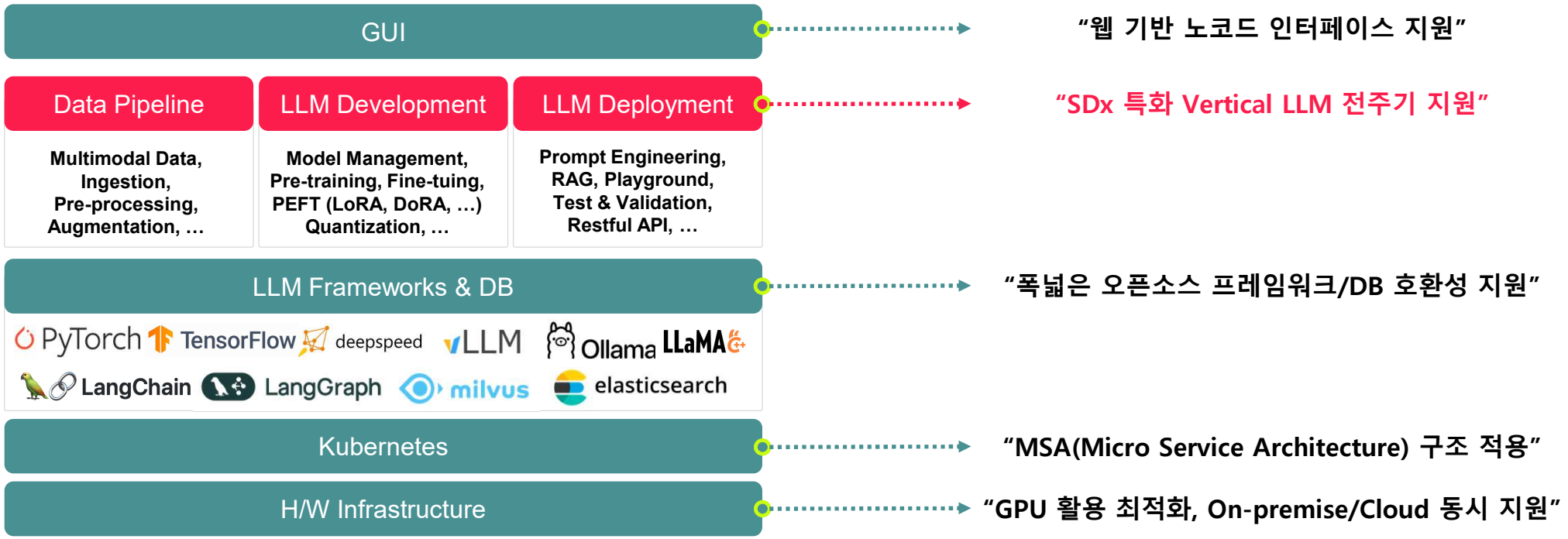
※ 출처: Microsoft



2. LLMOps 플랫폼(TANGO2) 구조 및 기능

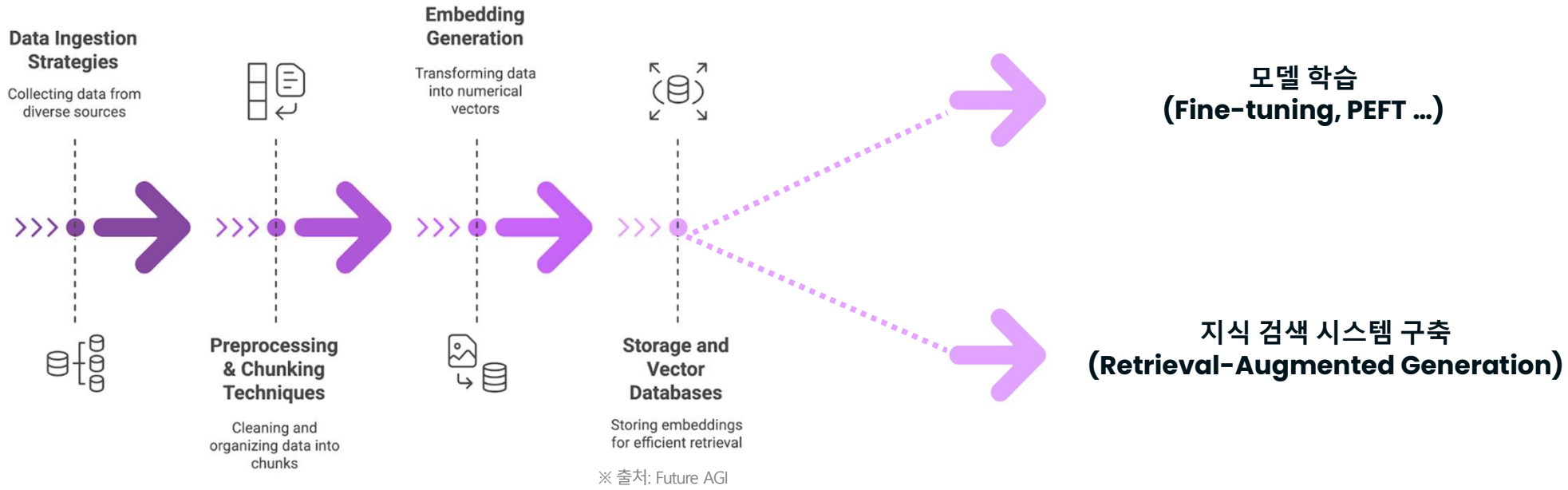
TANGO2 아키텍쳐

TANGO + JONATHAN



주요 기능 #1 – 데이터 파이프라인

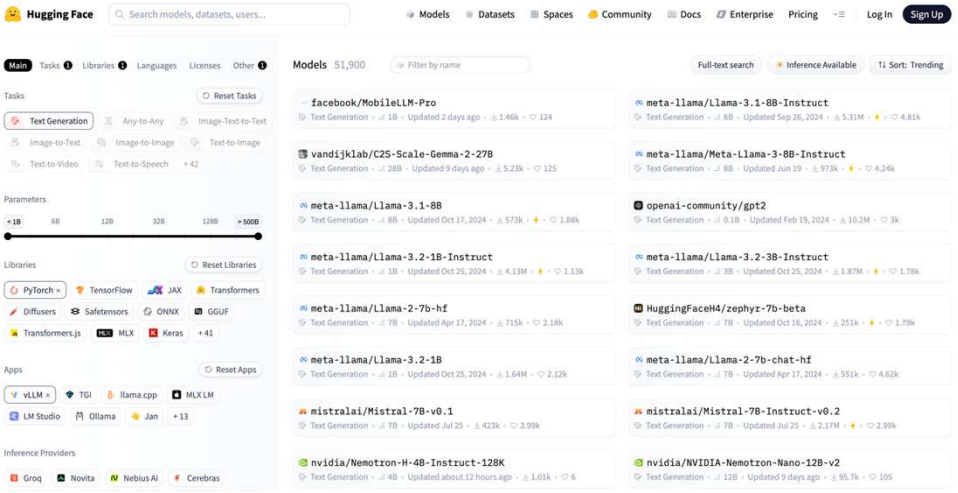
TANGO2는 데이터 수집부터 정제, 임베딩, 저장에 이르는 전 과정을 효과적으로 지원하며,
LLM 학습과 지식 검색 시스템 구축에 최적화된 고품질 데이터셋 구축을 가능하게 합니다.



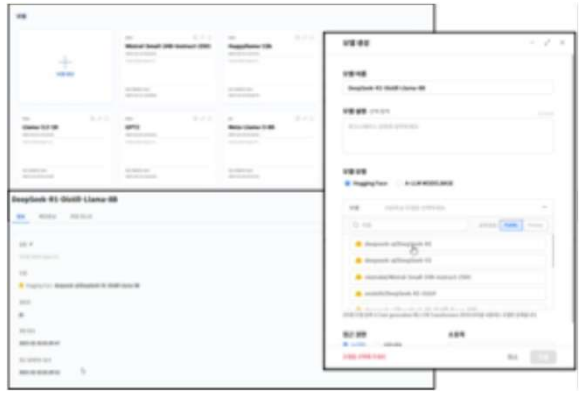
※ 출처: Future AGI

주요 기능 #2 – 모델 관리 및 파인튜닝

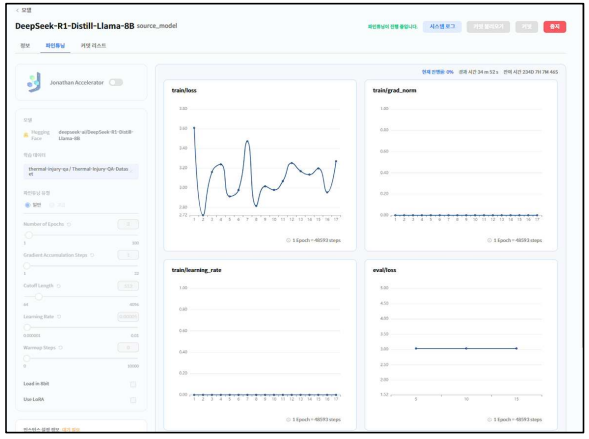
TANGO2는 다양한 오픈소스 모델을 통합 관리하고,
도메인 특화 데이터로 파인튜닝을 편리하게 수행할 수 있는 환경을 제공합니다.



※ 출처:
Huggingface



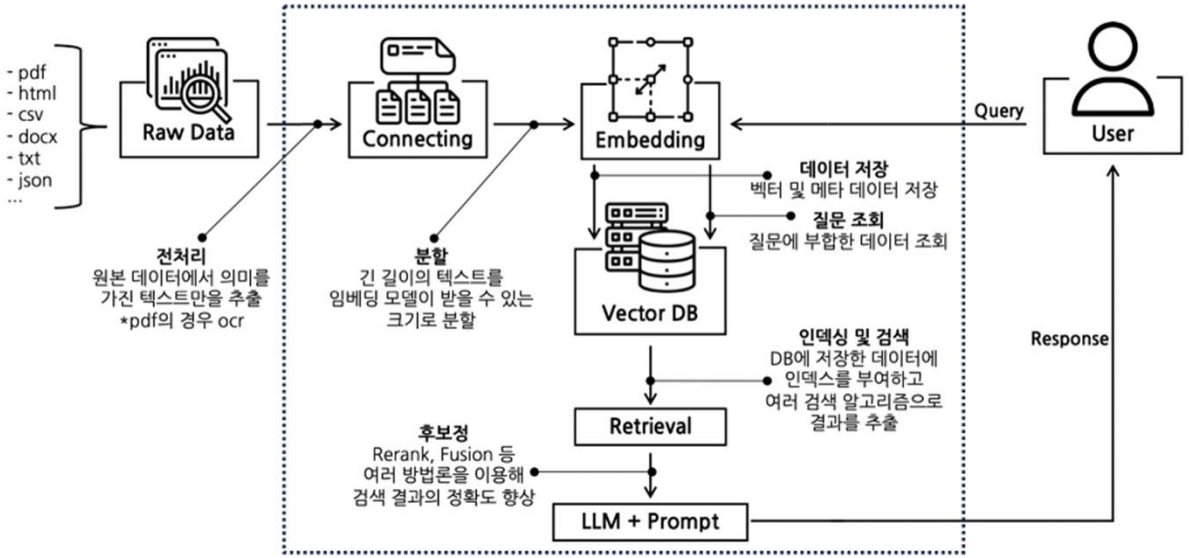
‹오픈소스 모델 로드 및 관리 기능›



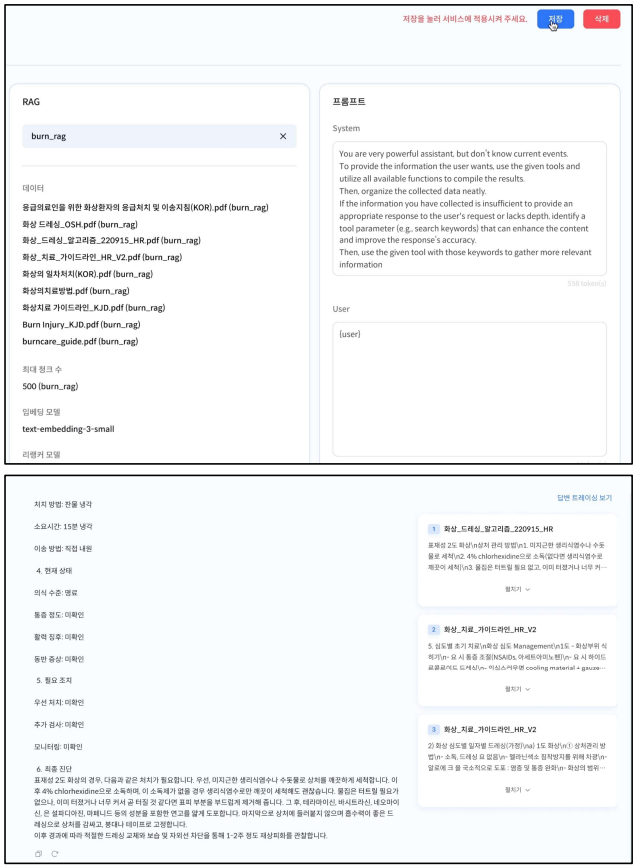
‹모델 학습 및 성능 분석 기능›

주요 기능 #3 – Prompt Engineering & RAG

TANGO2는 프롬프트 엔지니어링과 검색 기반 응답(RAG)을 통합적으로 지원하여, 정확도와 일관성이 높은 생성형 AI 서비스를 구현할 수 있는 환경을 제공합니다



※ 출처: https://velog.io/@yenoh_j/RAG-Prompt-Engineering



<Prompt-engineering/RAG 구축 및 검증 화면>

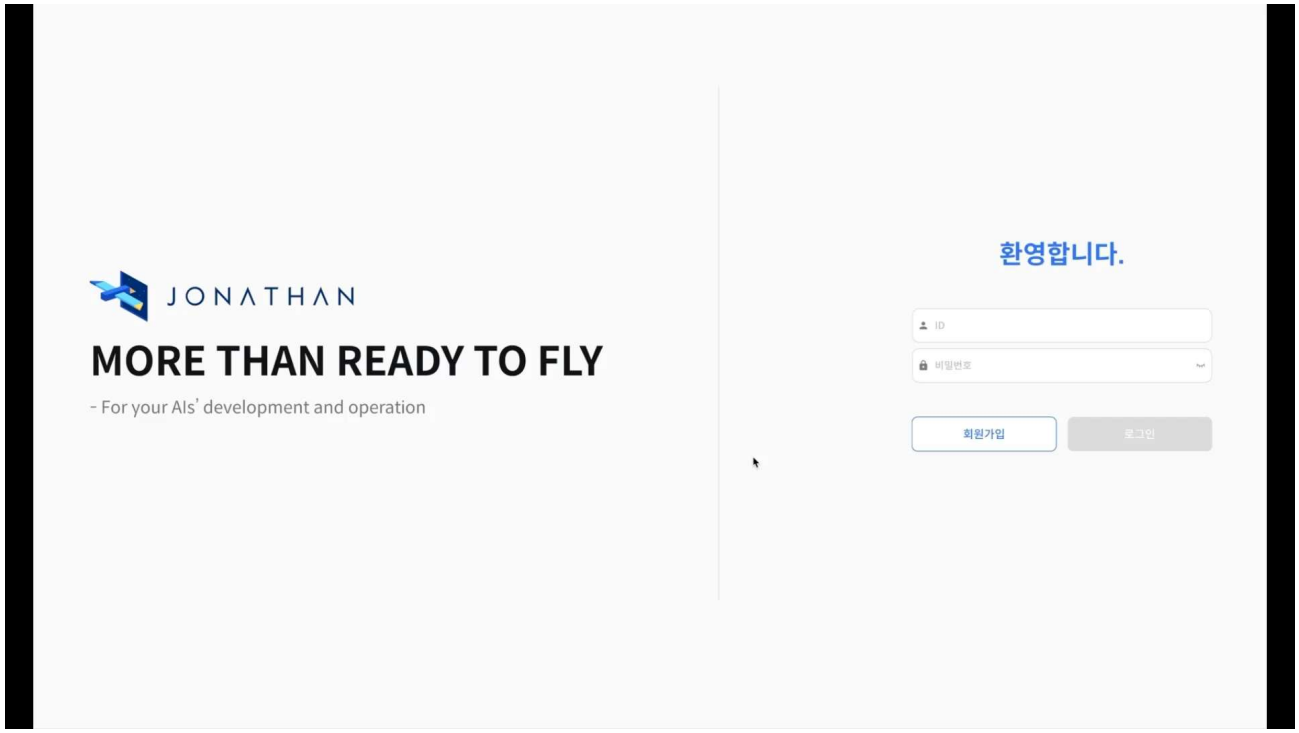
주요 기능 #4 – 배포 및 검증

TANGO2는 모델을 편리하게 배포하고 체계적으로 검증할 수 있는 환경을 제공하여,
효율적이고 안정적이며 신뢰성 높은 생성형 AI 서비스 운영을 가능하게 합니다.

일반 AI 모델과 LLM 검증의 차이점

구분	ML/DL	LLM
검증 목적	예측 정확도, 분류/회귀 성능 검증	응답 품질, 일관성, 사실성, 유용성 평가
평가 방식	정량적 지표 기반 (Accuracy, F1, RMSE 등)	정량 + 정성 평가 병행 (Human Eval, GPT Eval 등)
데이터 형태	구조화된 입력 (숫자, 이미지, 레이블 등)	비정형 텍스트 입력 (Prompt) + 자유 응답(Output)
기준 정답 존재 여부	정답(Label) 명확	절대적 정답이 없음 (주관적 품질 평가 필요)
테스트 케이스 구성	고정된 입력-출력 쌍으로 구성	다양한 Prompt 조합, Context 변화 포함
평가 주체	자동화된 Metric 기반	사람(Human Judge) + 모델 기반 자동 평가 병행
지표 예시	Accuracy, Precision, Recall, ROC-AUC	BLEU, ROUGE, BERTScore, GPT-Eval, Faithfulness Score
에러 분석 방식	오분류 샘플 분석, Feature 영향도 분석	Hallucination, 논리적 오류, 문체 불일치 분석
결과 해석	수치 중심 (객관적 판단 용이)	문맥·의도 중심 (정성적 판단 필요)

개발 현황


JONATHAN
MORE THAN READY TO FLY
- For your AIs' development and operation
환영합니다.
ID
비밀번호
회원가입
로그인

개발 계획

구분	일정															
	2025				2026				2027				2028			
	3	4	1	2	3	4	1	2	3	4	1	2	3	4		
파인튜닝 기능 개발	→	→	→													
파인튜닝 기능 고도화							→	→	→	→	→	→				
데이터 파이프라인 개발				→	→	→	→									
데이터 파이프라인 고도화							→	→	→	→	→	→				
지속학습 기능 개발							→	→	→							
지속학습 기능 고도화									→	→	→	→				
타 세부 요소기술 통합 전주기 워크스페이스 개발							→	→	→	→	→	→	→	→	→	→

생성형 AI의 전주기 운영을 통합 관리하는 LLMOps 플랫폼, TANGO2

- 데이터 품질·모델 성능·운영 효율·서비스 확장성 등 핵심 과제 해결
- 자동화된 파이프라인과 일관된 거버넌스 체계를 통한 지속 가능한 운영
- 도메인 특화 LLM 개발과 신뢰성 높은 AI 서비스 구현 지원
- 오픈소스와 연동 가능한 유연한 구조로 다양한 환경에서 적용 가능
- 기업의 AI 경쟁력과 생산성 향상을 위한 실질적 기반 제공
- **“신뢰할 수 있는 AI, 지속 가능한 AI”** 실현을 목표로 지속적 고도화 추진

TANGO

감사합니다.

주 관 ETRI (TANGO)

주 최 과학기술정보통신부 IITP 정보통신기획평가원

문 의 parkjb@etri.re.kr / 042-860-5565

후 원 LAB labiup weda teslo System (사)한국인공지능협회 Korea Artificial Intelligence Association SNUH 서울대학교병원 고려대학교 KOREA UNIVERSITY 홍익대학교 HONGIK UNIVERSITY CAU 중앙대학교 RTst Reliable & Trustworthy

KETI 한국전자기술연구원 AIVN SUREDATA ACRYL 하늘소프트 TTA 한국정보통신기술협회 Telecommunications Technology Association