



파이토치

한국 사용자 모임

오픈소스 AI와 AI 테크맵

AI 시대, 변화하는 오픈소스 기준 및 활용 방안 제안

박정환 파이토치 한국 사용자 모임 (PyTorchKR)

9bow@pytorch.kr



파이토치

한국 사용자 모임

시작하며

Introduction

발표자 소개



박정환 (9bow)

- ✓ 파이토치 한국 사용자 모임, 운영자
- ✓ PyTorch Ambassador, 2025/09 ~
- ✓ PyTorch Community Award, 2023/10
- ✓ Microsoft MVP Awards, 2007 ~ 2011
- ✓ 오픈소스, 커뮤니티, 멀티모달 및 시계열 예측

PyTorchKR 소개

파이토치 한국어 튜토리얼 (tutorials.pytorch.kr)

파이토치(PyTorch) 한국어 튜토리얼에 오신 것을 환영합니다!

아래 튜토리얼들이 새로 추가되었습니다:

- Integrating Custom Operators with SYCL for Intel GPU
- Supporting Custom C++ Classes in torch.compile/torch.export
- Accelerating torch.save and torch.load with GPUDirect Storage
- Getting Started with Fully Sharded Data Parallel (FSDP2)

PyTorch 기본 익히기

파이토치(PyTorch) 레시피

PyTorch 시작하기

레시피 찾아보기

Filter menu: All, Attention, Audio, Ax, Backends, Best Practice, C++, CUDA, Edge, Extending PyTorch, FX, Frontend APIs, Getting Started, Image/Video, Interpretability, Memory Format, Model Optimization, NLP, ONNX, Parallel and-Distributed-Training, Production, Profiling, Recommender, Reinforcement Learning, TensorBoard, Tensorboard, TorchRec, TorchX, Transformer

파이토치 한국 사용자 모임

회원가입 로그인

카테고리 태그 카테고리 인기글 최신

공지 사항

파이토치 한국 사용자 모임의 소식, 공지사항을 공유합니다.

1 / 월 파이토치 한국 사용자 모임 제 6회 테크 세미나 10일 전

리벨리온과 스쿼드비츠와 함께 하는 오픈리인 vLLM Hands-on ... 10월 5

[종료] AMD Instinct™ MI300X GPU 무료 체험 이벤트를 진행... 9월 13

읽을거리&정보공유

다른 사람과 나누고 싶은 싶은 뉴스와 정보 등을 공유합니다.

64 / 월 Hugging Face, PyTorch를 단일 Backend로 채택한 Transfor... 19시간

Agent Network Protocol(ANP): Agentic Web을 위한 AI 에이... 1일 전

Acontext: 에이전트의 문맥을 저장하고 학습하는 데이터 플랫폼 2일 전

[2025/11/24 ~ 30] 이번 주에 살펴볼 만한 AI/ML 논문 모음 2일 전

Prompt Refiner: LLM 입력 프롬프트를 정제하고 최적화하는 초... 3일 전

tool2agent: 복잡한 비즈니스 환경을 위한 LLM 에이전트 도구 피... 4일 전

Beads: AI 코딩 에이전트의 '기억 상실'을 해결하는 Git 기반의 분... 4일 전

Step-Audio-R1: 오디오 분야에서의 추론 시 연산 시간 확장(Test... 5일 전

ERA Agent: AI가 생성한 코드의 안전한 실행을 위한 로컬 샌드박스 5일 전

[GN] AI 프로토타이핑 도구 완벽 가이드 5일 전

자유게시판

처음 방문하셨나요? 자유롭게 인사를 나눠주세요!

4 / 월 안녕하세요 반갑습니다. 6일 전

안녕하세요 파알못입니다. 6일 전

안녕하세요, 가입 인사 드립니다. 13일 전

안녕하세요 가입인사드립니다. 27일 전

'Apps in ChatGPT'를 위한 Python 기반 Framework 10월 27

묻고 답하기

파이토치 및 인공지능 관련 질문과 답변을 주고 받는 공간입니다.

1 / 월 한국어 임베딩 모델 21일 전

pytorch v2.5 설치 관련 11월 1

llm 한국어 토큰라이저 10월 31

★ 최근 공유된 클라우드 현장의 실무 질문/답변을 공유합니다 9월 19

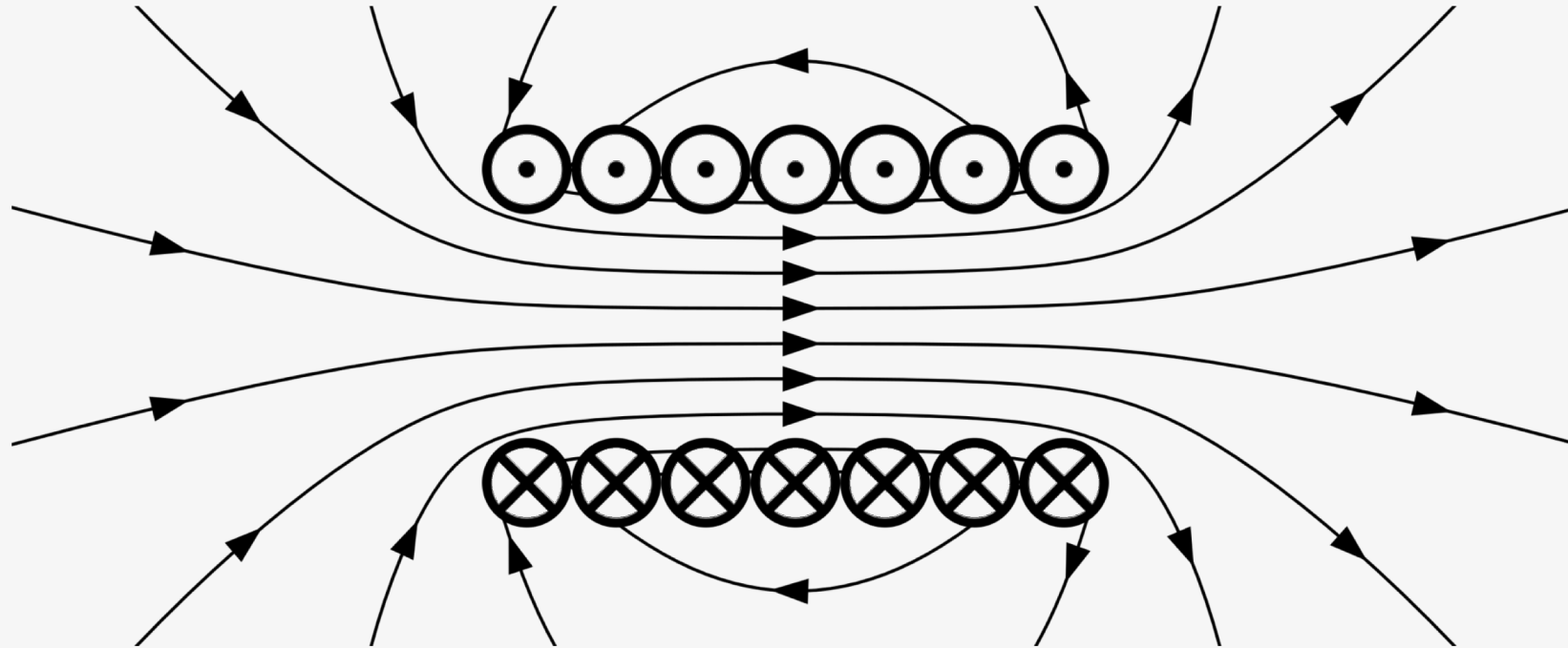
Pi Camera V3 와이드로 Yolov5 환경에서 실시간 감지하는 방법 7월 31

개인 AI 연구용 컴퓨터 세팅 7월 29

YOLO 모델 학습 관련해서 논문 작업을 하고 있습니다. 7월 28

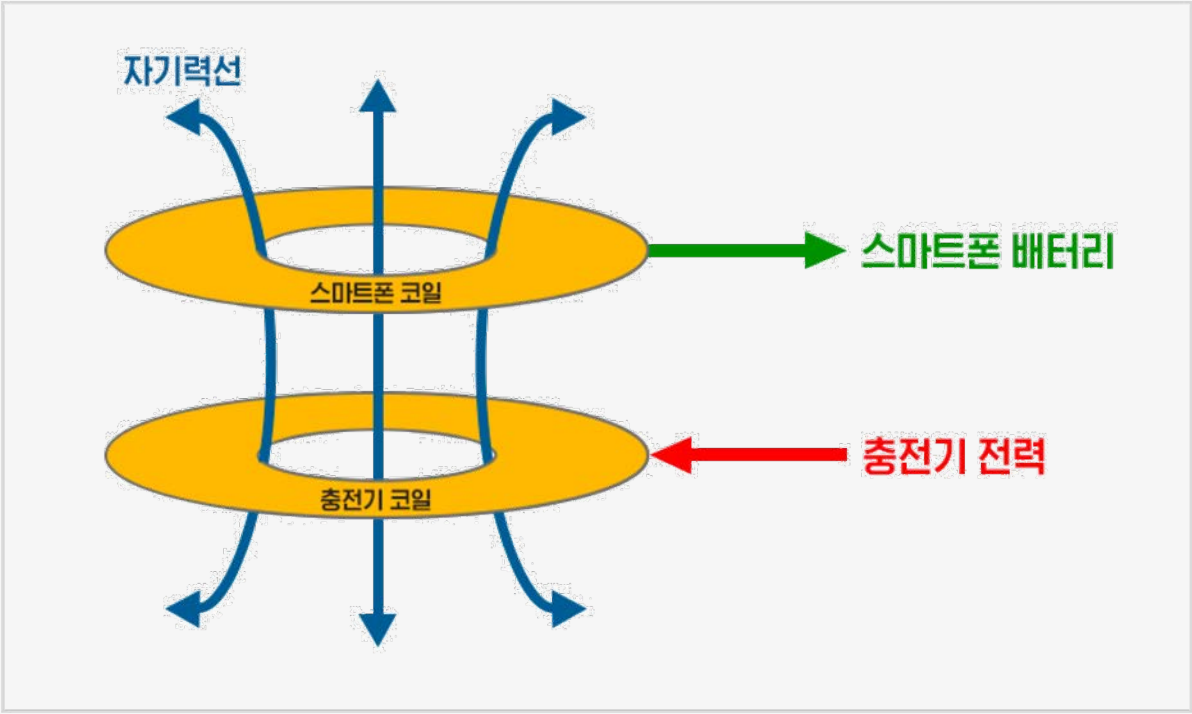
파이토치 한국어 커뮤니티 (discuss.pytorch.kr)

발표 소개



$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

발표 소개



발표 소개

Consider the time-derivative of magnetic flux through a closed boundary (loop) that can move or be deformed. The area bounded by the loop is denoted as $\Sigma(t)$, then the time-derivative can be expressed as

$$\frac{d\Phi_B}{dt} = \frac{d}{dt} \int_{\Sigma(t)} \mathbf{B}(t) \cdot d\mathbf{A}$$

The integral can change over time for two reasons: The integrand can change, or the integration region can change. These add linearly, therefore:

$$\left. \frac{d\Phi_B}{dt} \right|_{t=t_0} = \left(\int_{\Sigma(t_0)} \left. \frac{\partial \mathbf{B}}{\partial t} \right|_{t=t_0} \cdot d\mathbf{A} \right) + \left(\frac{d}{dt} \int_{\Sigma(t)} \mathbf{B}(t_0) \cdot d\mathbf{A} \right)$$

where t_0 is any given fixed time. We will show that the first term on the right-hand side corresponds to transformer emf, the second to motional emf (from the magnetic Lorentz force on charge carriers due to the motion or deformation of the conducting loop in the magnetic field). The first term on the right-hand side can be rewritten using the integral form of the Maxwell–Faraday equation:

$$\int_{\Sigma(t_0)} \left. \frac{\partial \mathbf{B}}{\partial t} \right|_{t=t_0} \cdot d\mathbf{A} = - \oint_{\partial \Sigma(t_0)} \mathbf{E}(t_0) \cdot d\mathbf{l}$$

Next, we analyze the second term on the right-hand side:

$$\frac{d}{dt} \int_{\Sigma(t)} \mathbf{B}(t_0) \cdot d\mathbf{A}$$

The proof of this is a little more difficult than the first term; more details and alternate approaches for the proof can be found in the references.^{[27][28][29]} As the loop moves and/or deforms, it sweeps out a surface (see the right figure). As a small part of the loop $d\mathbf{l}$ moves with velocity \mathbf{v}_1 over a short time dt , it sweeps out an area whose vector is $d\mathbf{A}_{\text{sweep}} = \mathbf{v}_1 dt \times d\mathbf{l}$ (note that this vector is toward out from the display in the right figure). Therefore, the change of the magnetic flux through the loop due to the deformation or movement of the loop over the time dt is

$$d\Phi_B = \int \mathbf{B} \cdot d\mathbf{A}_{\text{sweep}} = \int \mathbf{B} \cdot (\mathbf{v}_1 dt \times d\mathbf{l}) = - \int dt d\mathbf{l} \cdot (\mathbf{v}_1 \times \mathbf{B})$$

Here, [identities of triple scalar products](#) are used. Therefore,

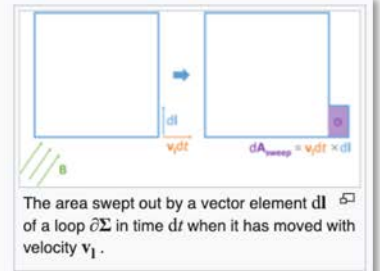
$$\frac{d}{dt} \int_{\Sigma(t)} \mathbf{B}(t_0) \cdot d\mathbf{A} = - \oint_{\partial \Sigma(t_0)} (\mathbf{v}_1(t_0) \times \mathbf{B}(t_0)) \cdot d\mathbf{l}$$

where \mathbf{v}_1 is the velocity of a part of the loop $\partial \Sigma$.

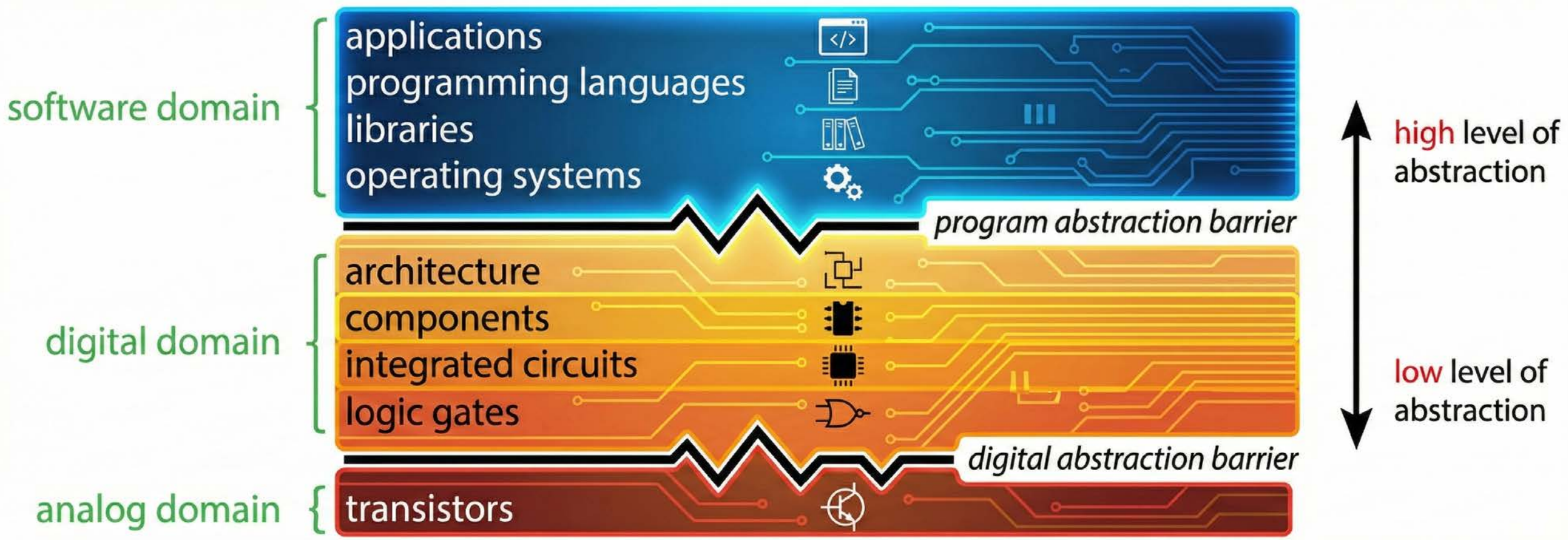
Putting these together results in,

$$\left. \frac{d\Phi_B}{dt} \right|_{t=t_0} = \left(- \oint_{\partial \Sigma(t_0)} \mathbf{E}(t_0) \cdot d\mathbf{l} \right) + \left(- \oint_{\partial \Sigma(t_0)} (\mathbf{v}_1(t_0) \times \mathbf{B}(t_0)) \cdot d\mathbf{l} \right)$$

$$\left. \frac{d\Phi_B}{dt} \right|_{t=t_0} = - \oint_{\partial \Sigma(t_0)} (\mathbf{E}(t_0) + \mathbf{v}_1(t_0) \times \mathbf{B}(t_0)) \cdot d\mathbf{l}.$$



발표 소개





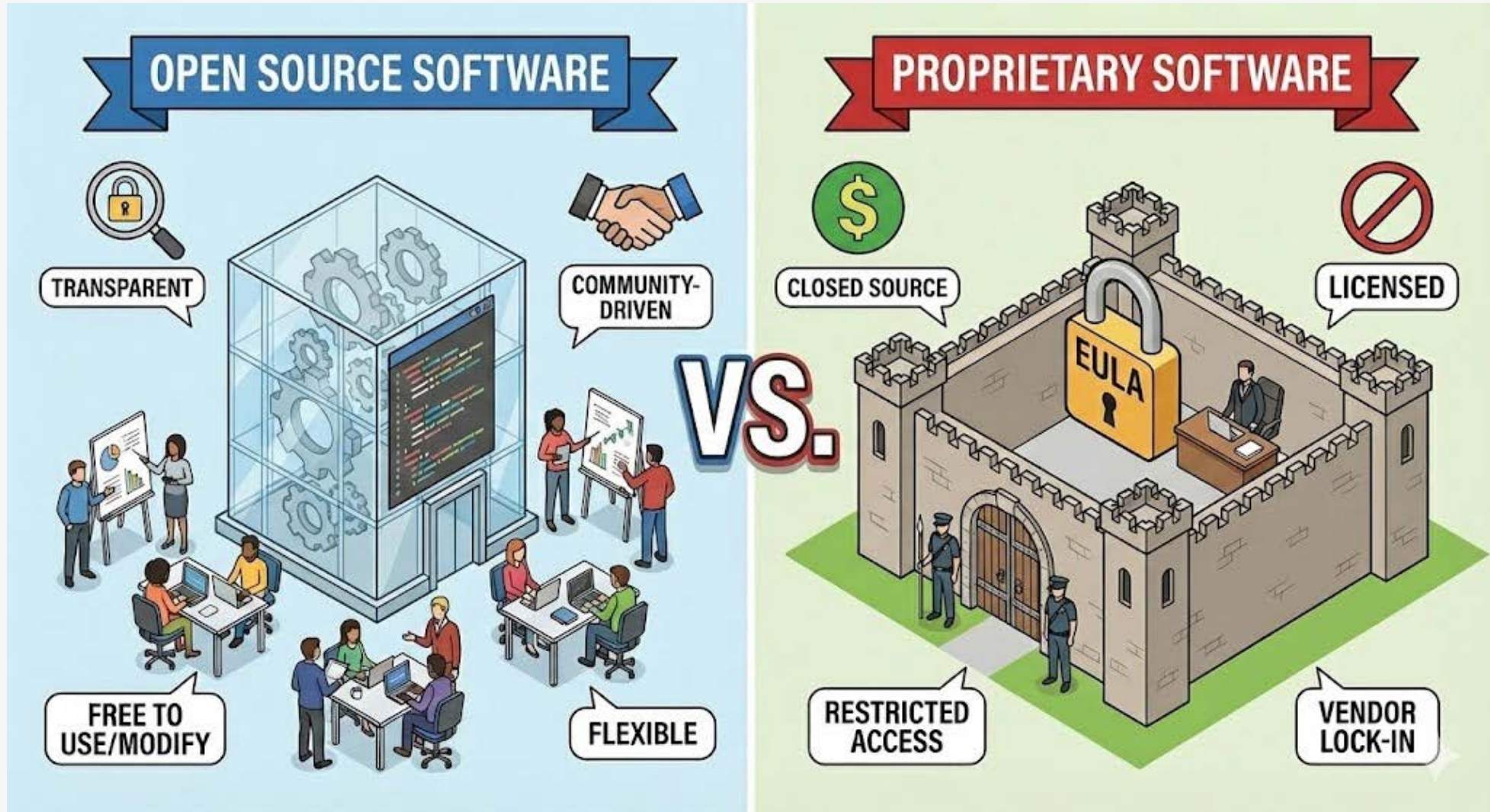
파이토치

한국 사용자 모임

오픈소스 AI

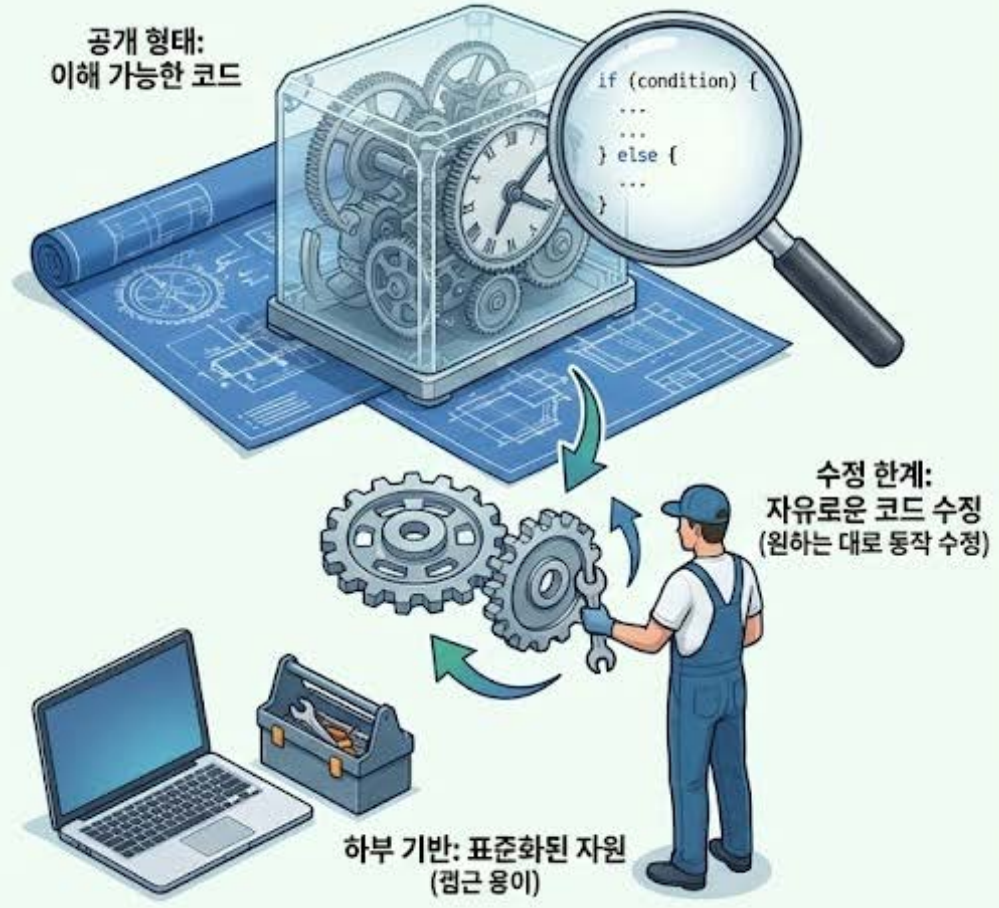
Open Source AI

오픈소스 (소프트웨어)



오픈소스... AI?

전통적 오픈소스 소프트웨어 (Traditional OSS)



오픈소스 AI (Open Source AI)



오픈소스 AI



OSI의 오픈소스 AI 정의(OSAID)

- OSI(Open Source Initiative)가 제안한 오픈소스 AI 정의 (Open Source AI Definition)
- 사용(Use), 연구(Study), 변경(Modify), 공유(Share)의 자유가 보장된 AI 시스템
- **사용**: 목적에 관계없이, 허가 없이 AI 시스템을 사용
- **연구**: AI 시스템 및 구성 요소들의 검사 및 동작 연구
- **변경**: 사용 목적에 따른 (출력 포함) 시스템의 변경
- **공유**: 목적 및 변경 여부에 관계없이 타인에게 공유

What's Open Source AI?

Following the same idea behind Open Source Software, an Open Source AI is a system made available under terms that grant users the freedoms to:

Use

Study

Modify

Share

THE OPEN SOURCE AI DEFINITION 1.0

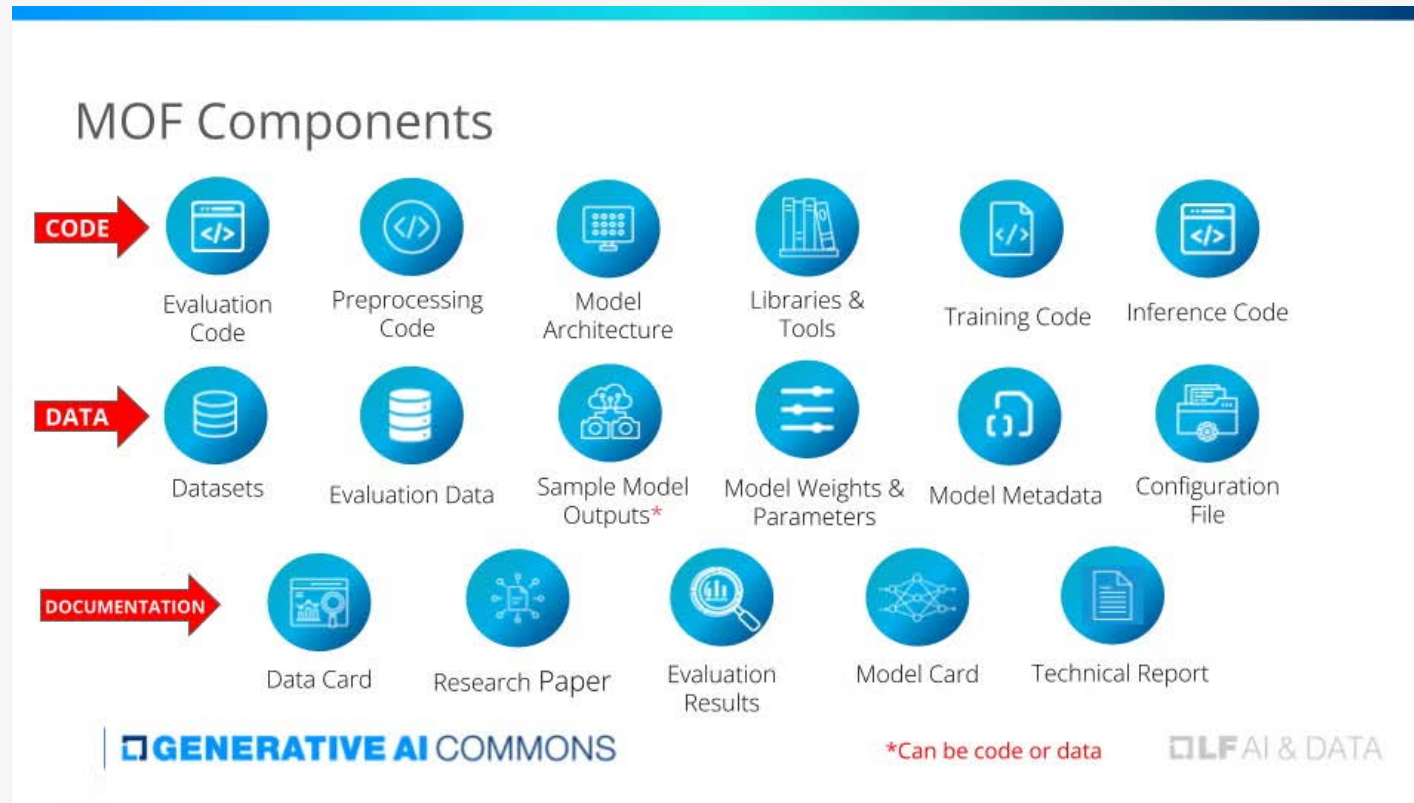
We have released the first stable version of the Definition.

[Read version 1.0](#)



Linux Foundation의 모델 개방성 평가체계(MOF)

- Linux Foundation 및 GenAI Commons 등이 함께 공개한 모델 개방성 평가체계
- 17가지 구성 요소(코드, 데이터, 문서)의 공개 정도에 따라 3단계(Class 1~3)로 구분



Linux Foundation의 모델 개방성 평가체계(MOF)

- 17가지 구성 요소(코드, 데이터, 문서)의 공개 정도에 따른 3단계 구분

- **Class 3. Open Model**

- 가장 낮은 공개 등급으로, ‘사용’이 가능한 단계
- 모델 가중치 및 기본 문서, 평가 결과만 공개

- **Class 2. Open Tooling**

- 모델 검증 및 확장, ‘변경’이 가능한 단계
- Class 3 + 학습 코드 및 주요 데이터셋 공개


- **Class 1. Open Science**

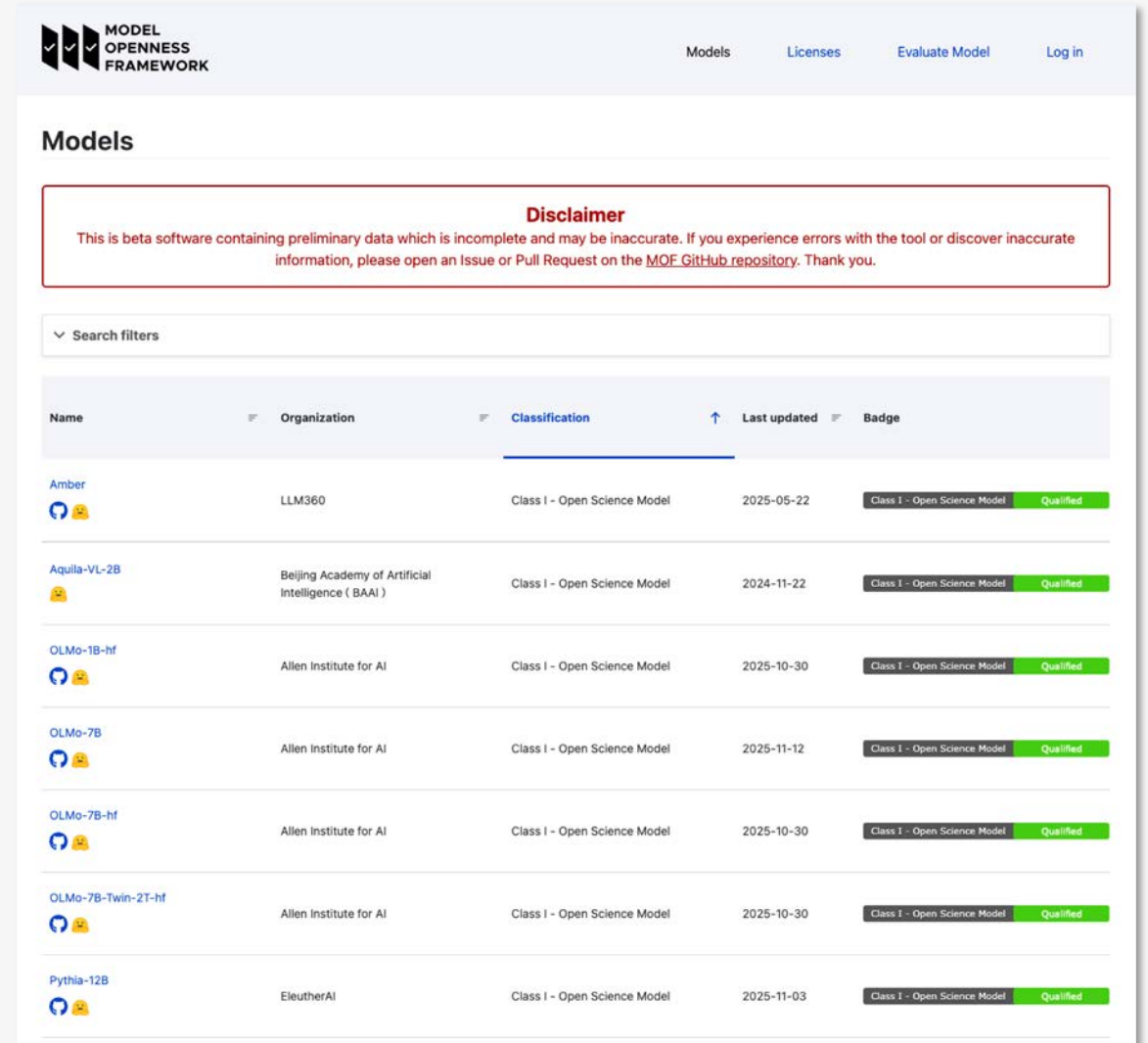
- 가장 높은 공개 등급으로, ‘재현’이 가능한 단계
- Class 2 + 학습 데이터 및 논문, 중간 결과물 등 공개

MOF Class	Components Included
Class III. Open Model	<ol style="list-style-type: none">1. Model Architecture2. Model Parameters (Final Checkpoints)3. Technical Report or Research Paper4. Evaluation Results5. Model Card6. Data Card7. Sample Model Outputs (Optional)
Class II. Open Tooling	<ol style="list-style-type: none">1. All Class III Components2. Training, Validation, and Testing Code3. Inference Code4. Evaluation Code5. Evaluation Data6. Supporting Libraries & Tools
Class I. Open Science	<ol style="list-style-type: none">1. All Class II Components2. Research Paper3. Datasets4. Data Preprocessing Code5. Model Parameters (Intermediate Checkpoints)6. Model Metadata (Optional)

Linux Foundation의 모델 개방성 평가도구(MOT)

- 모델 개방성 평가체계에 따른 평가 도구 (베타 공개 단계)
- Amber, Aquila-VL, OLMo, Pythia 등 Class 1

- Website: <https://mot.isitopen.ai/>
- GitHub:  [lfai/model_openness_tool](https://github.com/lfai/model_openness_tool)

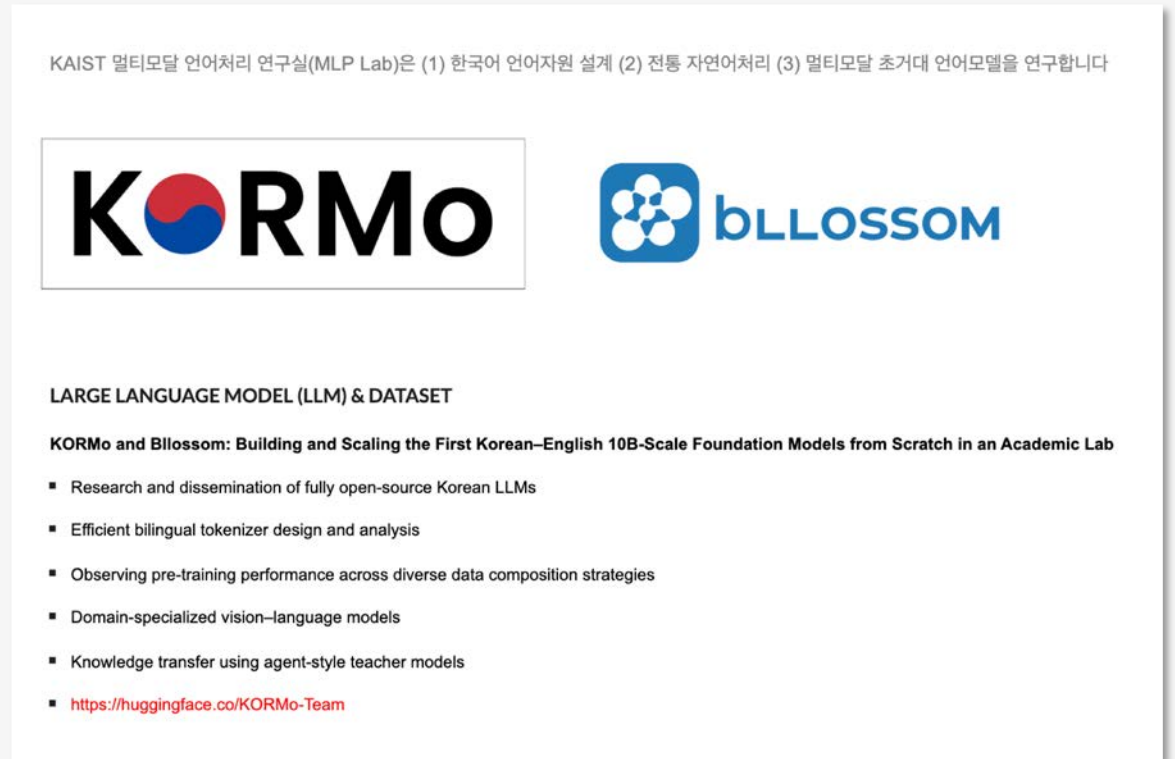
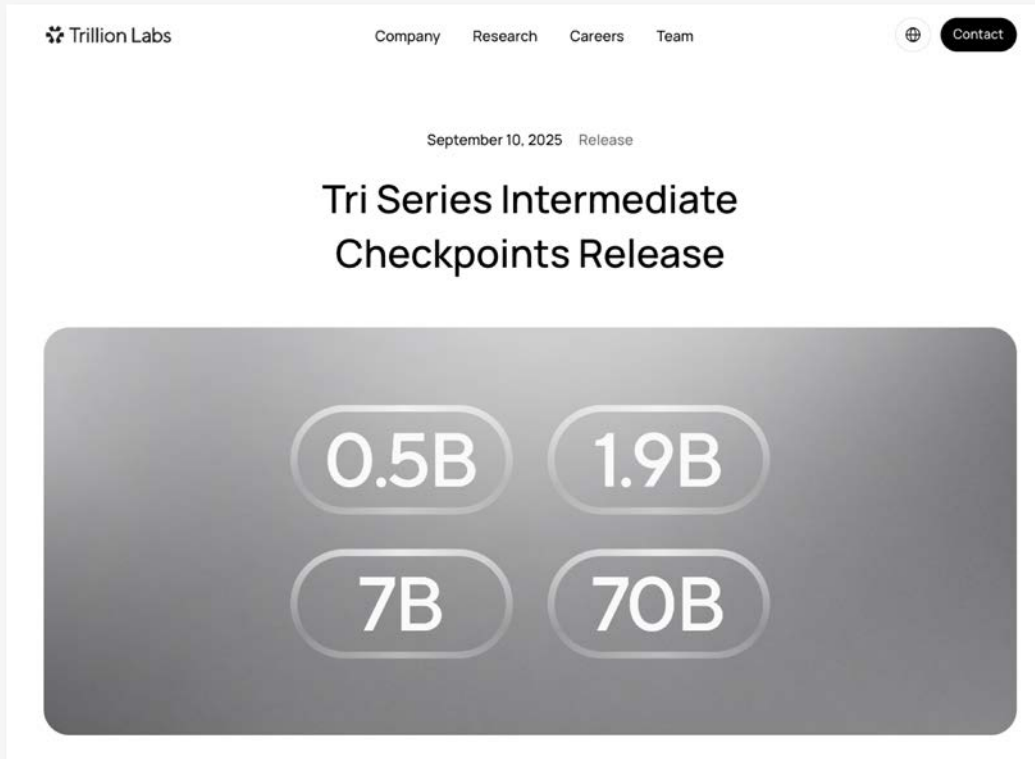


The screenshot shows the MOT website interface. At the top, there is a navigation bar with the logo 'MODEL OPENNESS FRAMEWORK' and links for 'Models', 'Licenses', 'Evaluate Model', and 'Log in'. Below the navigation bar, there is a 'Models' section with a 'Disclaimer' box stating: 'This is beta software containing preliminary data which is incomplete and may be inaccurate. If you experience errors with the tool or discover inaccurate information, please open an Issue or Pull Request on the MOT GitHub repository. Thank you.' Below the disclaimer, there is a 'Search filters' dropdown menu. The main content is a table listing models with columns for Name, Organization, Classification, Last updated, and Badge. The table contains the following data:

Name	Organization	Classification	Last updated	Badge
Amber	LLM360	Class 1 - Open Science Model	2025-05-22	Class 1 - Open Science Model Qualified
Aquila-VL-2B	Beijing Academy of Artificial Intelligence (BAAI)	Class 1 - Open Science Model	2024-11-22	Class 1 - Open Science Model Qualified
OLMo-1B-hf	Allen Institute for AI	Class 1 - Open Science Model	2025-10-30	Class 1 - Open Science Model Qualified
OLMo-7B	Allen Institute for AI	Class 1 - Open Science Model	2025-11-12	Class 1 - Open Science Model Qualified
OLMo-7B-hf	Allen Institute for AI	Class 1 - Open Science Model	2025-10-30	Class 1 - Open Science Model Qualified
OLMo-7B-Twin-2T-hf	Allen Institute for AI	Class 1 - Open Science Model	2025-10-30	Class 1 - Open Science Model Qualified
Pythia-12B	EleutherAI	Class 1 - Open Science Model	2025-11-03	Class 1 - Open Science Model Qualified

오픈소스 AI

- 단순 사용(Use)이나 변경(Modify)을 넘어, 재현(Reproduction)이 가능해야
- 한국에서는 Trillion Labs의 Tri Series, KASIT MLP Lab의 KORMo 등이 공개 중





파이토치

한국 사용자 모임

AI 테크맵

AI Techmap

제안 배경

- 지도: ‘어디’에 ‘무엇’이 있는지 확인하고, 원하는 곳을 ‘어떻게’ 찾아갈지 돕는 도구
- AI 테크맵: AI 서비스 개발 시 ‘보유하지 않은 기술’과 ‘대체 가능한 오픈소스’ 탐색을 돕는 도구

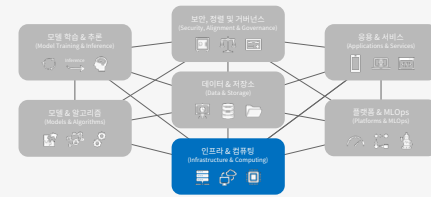


AI 테크맵 제안

- 어떠한 기술 및 도구가 필요한지 식별하기 위해 AI 기술 스택을 7가지 분류로 정리

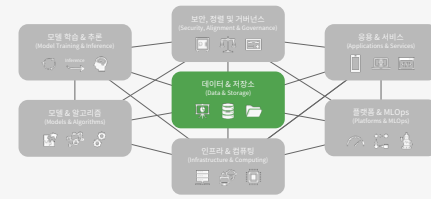


AI 테크맵 구성: 인프라 & 컴퓨팅



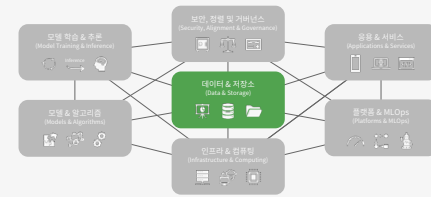
- AI 시스템 구동에 필요한 H/W, S/W 및 이에 기반한 컴퓨팅 인프라/관리 도구 일체
- 온-디바이스 AI 반도체
 - IoT 기기 등에서 AI 추론을 직접 수행하는 저전력 AI 반도체
- AI 가속기(GPU/NPU 등)
 - AI 모델 학습에 필요한 고성능 병렬 처리용 하드웨어 및 런타임 - 예. OpenCL, ROCm, Vulkan 등
- 대규모/분산 컴퓨팅 인프라
 - AI 모델 학습/배포에 필요한 인프라 자원 및 플랫폼(IaaS 등) - 예. OpenStack, CloudStack 등
- 컨테이너/오케스트레이션
 - 대규모 컴퓨팅 인프라 자원 관리 및 최적화 기술 및 도구 - 예. Docker, Kubernetes 등

AI 테크맵 구성: 데이터 & 저장소

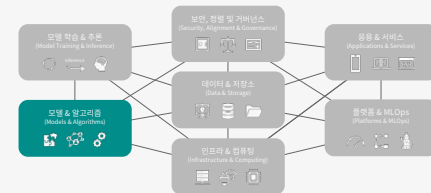


- 모델 학습, 추론 및 평가에 필요한 데이터 및 데이터 수집 - 저장 - 전처리 - 관리 도구
- **ETL & Workflow:** 다양한 데이터 원천으로부터 ETL(Extract/Transform/Load) 수행 및 관리 도구
- **카탈로그 및 버전/계보:** 메타데이터 수집/정제 및 데이터 버전/계보(Lineage) 관리 도구
- **데이터셋 및 관련 도구:** 모델 학습에 필요한 데이터셋 및 저장소/도구 등
- **합성 데이터 생성 / 관리 도구:** 프라이버시 보호 및 데이터 보안을 위해 실제와 유사한 데이터 생성/관리
- **데이터 라벨링:** 모델 학습에 필요한 데이터 라벨 생성/관리 및 협업/검수/자동화 도구
- **데이터 품질 관리/보증:** 데이터 정제 및 스키마 검증, 이상치 탐지, 정합성 확인 등 품질 관리
- **벡터 데이터베이스:** 임베딩 벡터 등을 저장/조회/검색하여 유사도 기반 검색 및 RAG 등에 활용

AI 테크맵 구성: 데이터 & 저장소



- 모델 학습, 추론 및 평가에 필요한 데이터 및 데이터 수집 - 저장 - 전처리 - 관리 도구
- **ETL & Workflow** 도구 예. Apache Airflow, Apache Oozie, Dagster, ...
- **카탈로그 및 버전/계보** 도구 예. DataHub, OpenMetadata, OpenLineage, Amundsen, ...
- **데이터셋 및 관련 도구** 예. LAION, Common Crawl, Hugging Face Datasets, ...
- **합성 데이터 생성 / 관리 도구** 예. Faker, SDV, Synthea, Syth, YData Synthetic, ...
- **데이터 라벨링** 도구 예. CVAT, Label Studio, ...
- **데이터 품질 관리/보증** 도구 예. OpenRefine, elementary, GreatExpectations, Soda Core, ...
- **벡터 데이터베이스** 예. Milvus, Qdrant, Chroma, FAISS, Weaviate, ...

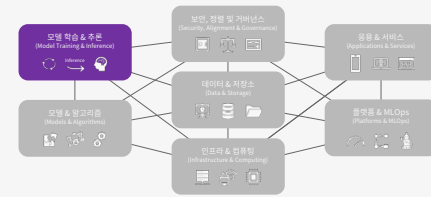


AI 테크맵 구성: 모델 & 알고리즘

- LLM 외, 전통적인 Deep Learning / Machine Learning 모델 포함
- 벤치마크 및 리더보드 등을 통해 활용 사례에 맞는 모델 검색 및 탐색 지원

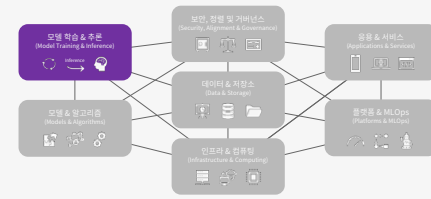


AI 테크맵 구성: 모델 학습 & 추론



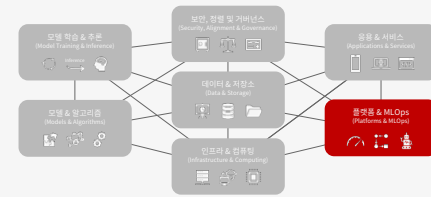
- 모델 개발 및 학습/추론을 위한 도구, 라이브러리 및 프레임워크
- **딥러닝 프레임워크**: 모델 정의 및 학습, 추론의 토대가 되는 범용 프레임워크
- **강화학습 라이브러리**: 에이전트 학습 및 문장/이미지 생성 등에 사용하는 강화학습 라이브러리
- **분산학습 프레임워크**: 여러 서버/노드를 활용하여 병렬화된 학습을 통해 학습 속도 및 확장성 강화
- **연합학습 프레임워크**: 여러 엣지 단말 등에서 데이터 이동 없이 중앙 서버를 통해 학습하는 프레임워크
- **AutoML/HPO**: 모델/파이프라인 생성, Hyper Parameter 최적화 등 모델 개발 과정 자동화 도구
- **파인튜닝/PEFT**: 학습된 모델을 추가적으로 미세조정(파인튜닝)하기 위한 도구
- **모델 경량화**: Quantization, Pruning, Knowledge Distillation 등, 추론 시 메모리 및 연산량 감소
- ...

AI 테크맵 구성: 모델 학습 & 추론



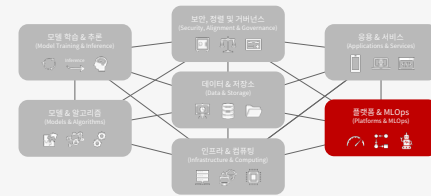
- 모델 개발 및 학습/추론을 위한 도구, 라이브러리 및 프레임워크
- 딥러닝 프레임워크 예. PyTorch, JAX, Tensorflow, Keras, ...
- 강화학습 라이브러리 예. veRL, torchRL, Hugging Face TRL, Catalyst, Ray Rllib, ...
- 분산학습 프레임워크 예. Horovod, DeepSpeed, ColossalAI, Determined, Ray Train, ...
- 연합학습 프레임워크 예. FedML, Flower, FATE, PySyft, Tensorflow Federated, ...
- AutoML/HPO 예. Auto ScikitLearn, AutoGluon, Optuna, TPOT, Hyperopt, ...
- 파인튜닝/PEFT 예. Hugging Face PEFT, LitGPT, Adapters, Ray Tune, Transformer Lab, ...
- 모델 경량화 예. Optimum, bitsandbytes, GPTQModel, Neural Compressor, ...
- ...

AI 테크맵 구성: 플랫폼 & MLOps



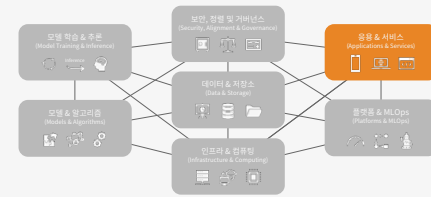
- 학습한 모델의 활용 및 운영을 위한 플랫폼 및 도구
- 모델 서빙 및 추론 엔진: 실시간 모델 서비스를 위한 고성능 서빙 및 추론 엔진
- 온디바이스 런타임: 특정 하드웨어(PC, 모바일 등)에 최적화한 모델 서빙 및 추론 런타임
- 프롬프트 관리 및 실험: 프롬프트/컨텍스트 엔지니어링 및 성능 평가, 버전 관리 등을 위한 도구
- RAG 및 GraphRAG: 모델에 외부 지식을 주입하고 최종 결과를 생성하는 RAG 파이프라인 프레임워크
- 메모리 및 지식 그래프: LLM 및 에이전트의 대화 기억, 지식 관리 등을 위한 장/단기 메모리 시스템
- 에이전트 개발 프레임워크: 에이전트 정의, 도구/상태/메모리 관리 등, 에이전트 개발 프레임워크
- 테스트 및 성능 평가: AI 시스템 전체(E2E) 또는 모듈별 기능 / 비기능 성능 테스트 및 평가 프레임워크
- ...

AI 테크맵 구성: 플랫폼 & MLOps



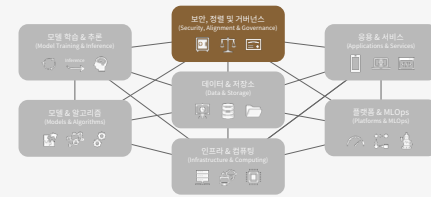
- 학습한 모델의 활용 및 운영을 위한 플랫폼 및 도구
- 모델 서빙 및 추론 엔진 예. vLLM, SGLang, Ray Serve, NVIDIA Triton Inference Server, ...
- 온디바이스 런타임 예. ollama, llama.cpp, MLC LLM, MNN, ncnn, ...
- 프롬프트 관리 및 실험 도구 예. promptfoo, DSPy, PromptTools, OpenPrompt, ...
- RAG 및 GraphRAG 프레임워크 예. GraphRAG, Haystack, RAGFlow, ...
- 메모리 및 지식 그래프 시스템 예. Mem0, cognee, Memary, memonto, txtai, ...
- 에이전트 개발 프레임워크 예. LangChain, CrewAI, AutoGen, LlamaIndex, SuperAGI, ...
- 테스트 및 성능 평가 도구 예. RAGAS, TruLens, DeepEval, ...
- ...

AI 테크맵 구성: 응용 & 서비스



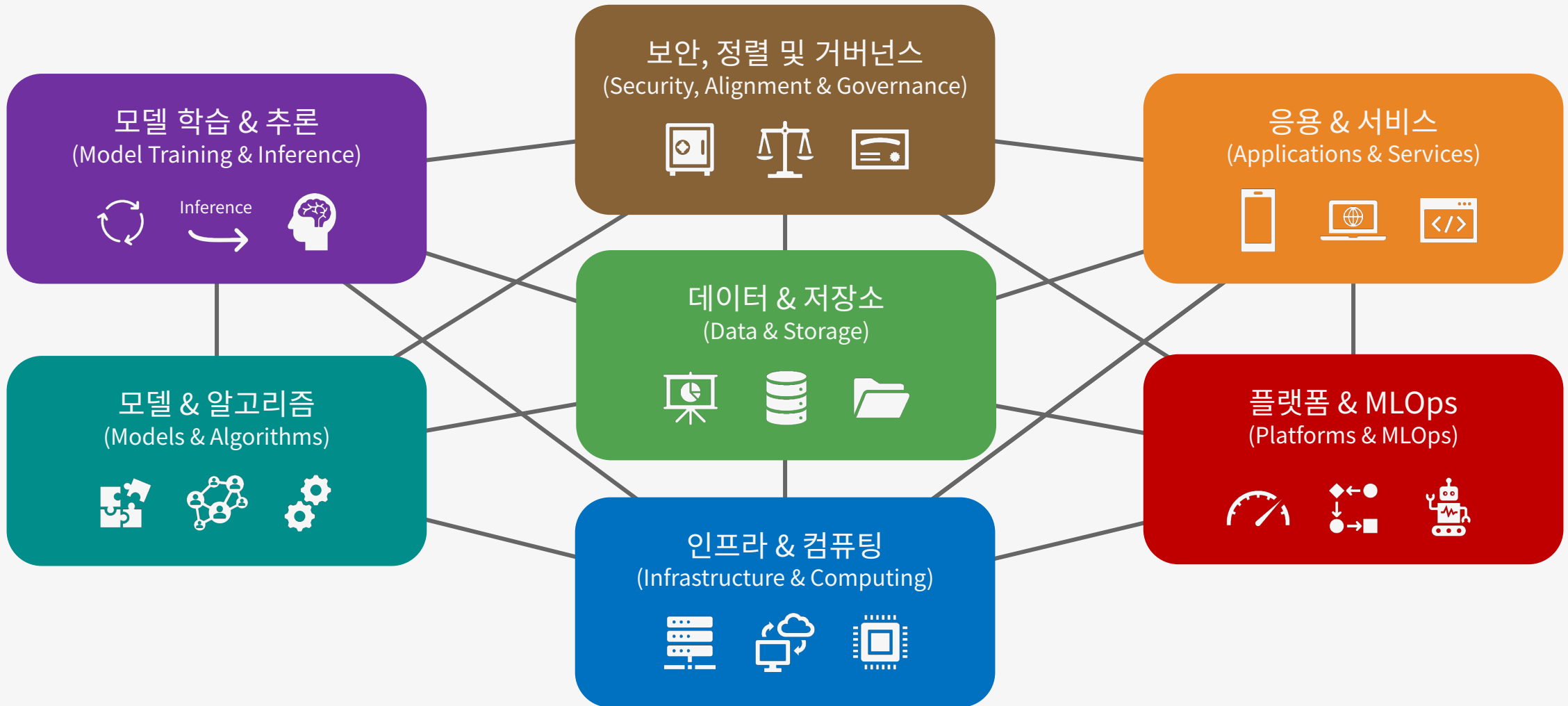
- 기업 및 소비자에게 직접적으로 제공되거나 사용하는 제품 및 솔루션
- 자율형 에이전트 서비스
 - LLM/VLM 기반, 자율적으로 복합 작업을 생성 및 수행하는 에이전트 서비스
- 산업별 AI 솔루션
 - 의료 영상 분석, 금융 사기 탐지 등, 제조 / 의료 / 금융 등 산업 도메인 특화된 AI 솔루션
- 추천 시스템 및 검색 엔진
 - 사용자 행동 또는 아이템 기반 콘텐츠, 제품 추천 및 정보 검색 서비스 솔루션
- 대화 기반 어시스턴트
 - 대화 기반으로 사용자에게 정보를 전달하거나 업무 수행을 돕는 Chatbot 등 솔루션

AI 테크맵 구성: 보안, 정렬 및 거버넌스



- AI 시스템의 유용성, 안전성 및 신뢰성 확보를 위한 정책 및 도구
- **AI 거버넌스**: AI 시스템의 윤리성, 규정/규제 준수 여부를 포함한 운영 체계 전반 검증 및 관리 체계
- **프라이버시 및 비식별화**: 개인식별정보(PII) 탐지 및 마스킹, 차등 프라이버시 적용 등 학습 데이터 보호
- **공정성 및 편향 탐지**: 특정 인종, 성별 또는 연령 등에 대한 데이터 및 모델의 편향 진단 및 공정성 확보
- **설명 가능한 AI(XAI)**: AI 모델의 의사 결정 과정을 설명하여 투명성 향상 및 의사 결정 과정 개입 등
- **가드레일 및 정책 엔진**: 답변 제한, 개인식별 정보 마스킹 등, 목적에 어긋난 대화를 하지 않도록 제한
- **보안 테스트 및 레드팀**: 탈옥, 악성 프롬프트 방어 등, AI 시스템 전반의 보안 검증 및 운영 체계 관리
- **AI 정렬(Alignment)**: 인간의 가치관/윤리관에 맞도록 AI 시스템의 유용성 / 안전성 / 신뢰성 확보
- ...

AI 테크맵 제안



AI 테크맵 활용 예시: 가상 시나리오

1. 기업 배경 및 목표 (Before)

중견/대기업용 ERP 솔루션 기업

SQL DB (정형 데이터) 규정 PDF (비정형 데이터)

현황: 온프레미스 선호, Java/Spring 숙련, AI 경험 無

목표: AI 업무 비서 & 사내 규정 Q&A

지난달 A부서 야근비 총액? → SQL 조회 결과

경조사비 지급 기준? → PDF 규정 답변

2. AI 테크맵 단계별 적용 전략 (Process - 오픈소스 조립)

Layer 1: Data (데이터 준비)

복잡한 테이블 구조 수천 페이지 PDF → 오픈소스 솔루션 도입: Unstructured.io (ETL), LangChain (SQL Chain), Milvus (Vector DB - 내부망) → 정제된 데이터 벡터 DB & 스키마 정보

Layer 2: Model (두뇌 선택)

보안 최우선 (SaaS 불가) 한국어 & SQL 능력 → 오픈소스 솔루션 도입: CodeLlama / Qwen 2.5-Coder (Text-to-SQL), Solar 10.7B (한국어 질의응답) 멀티 모델 전략 (온프레미스)

Layer 3: Serving & Adaptation (실행 및 최적화)

고가 GPU 부재 반응 속도 저하 우려 → 오픈소스 솔루션 도입: vLLM (성능 최적화), Quantization (4bit 양자화 - 경량화) → 보급형 GPU 구동 최적화

Layer 4: Orchestration (서비스 로직 연결)

잘못된 SQL 생성 위험 출쳐 명시 필요 → 오픈소스 솔루션 도입: LangGraph (사고 과정 제어 - 의도 파악 → SQL 검증 → 실행 → 오류 수정) → 안전한 AI 함수화

Layer 5: Evaluation (품질 검증)

확률적 정답 검증 어려움 → 오픈소스 솔루션 도입: Ragas (관련성 점수), Spider Benchmark (사내 데이터셋 SQL 정확도 측정) → 자동화된 정확도 측정

3. 결론 및 시사점 (After)

결과: 안전한 '온프레미스 AI ERP' 빠른 출시

기존 역량 (데이터/비즈니스 로직) 오픈소스 (모델/서빙/오케스트레이션) 전략: 바닥부터 개발 NO, 필요한 컴포넌트 조립 (Assemble)

성과: 데이터 유출 없는 보안 확보, 경쟁 우위 선점



파이토치

한국 사용자 모임

감사합니다.

발표 관련 추가 질문 또는 AI 테크맵 관련 개선 의견이 있으시면
이메일(9bow@pytorch.kr)로 연락 부탁드립니다.