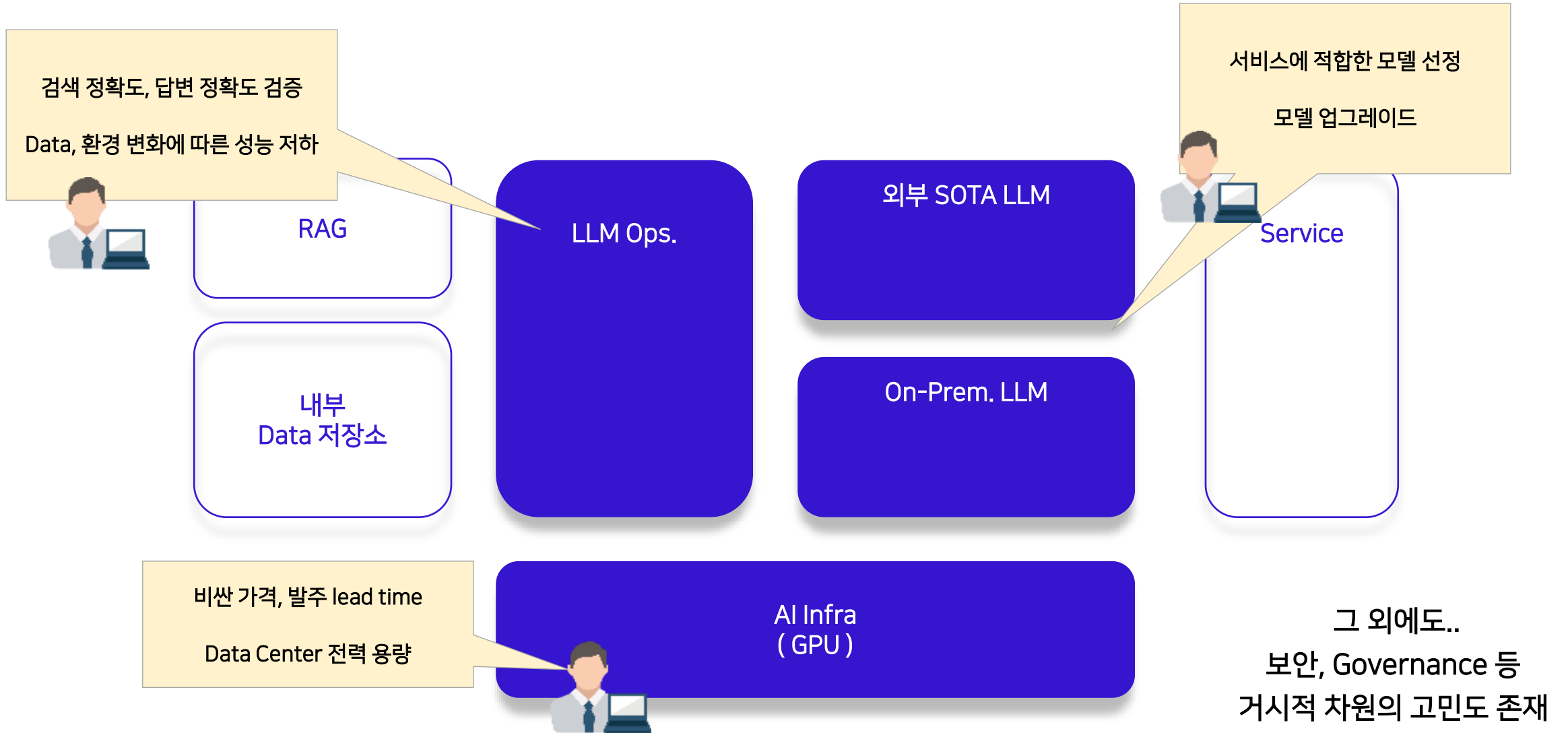


성공적인 GAI 도입을 위한 3가지 열쇠

SK텔레콤 AI Enterprise사업담당 김상목

Intro - AI 서비스 개발 시의 현실적인 고민



SK텔레콤 에이닷 서비스 개발 / 운영을 통해 직면했던 문제들과 고민 사항들에 대한 공유

다양한 서비스 활용 관점에서..

업무별 특성을 고려한
최적의 LLM을 어떻게
선정할 것인가?

Multi LLM

서비스 신뢰도 관점에서..

생성 AI 서비스가
얼마나 정확한지 검증
할 방법은 무엇인가?

Ops : Evaluation

인프라 운용 효율성 관점에서..

가성비, D/C 최적화
관점에서 GPU 인프라는
어떻게 구축할 것인가?

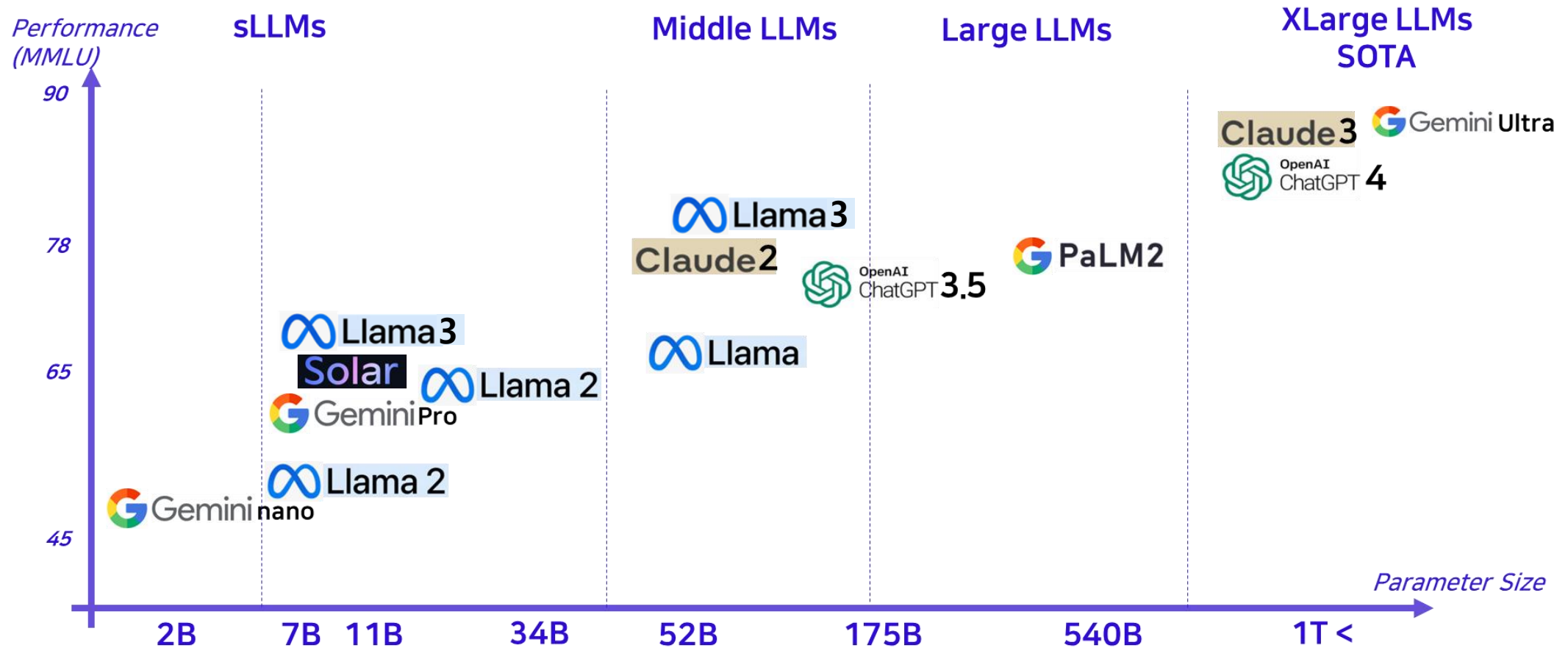
GPU

목차

1. LLM
2. Evaluation
3. GPU
4. Use Case

1. LLM - 다양한 LLM 과 성능

모델 사이즈와 범용 성능은 일반적으로 정비례하며, 빠르게 발전하고 있음 → 멀지 않은 시점에 정상 상태로 수렴 예상



출처 : <https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu>

1. LLM – XLarge LLMs

매우 만족스러운 성능을 보여주고 있음에도 아직까지는 몇가지 한계 존재

보안

해외 Region Cloud에
Data 저장
Local Dedicated 서비스 사용
시 많은 Upfront 비용 발생

많은 자원 소모

PaLM2(540B) vs. LLaMa2(34B)
비교 시,
20% 성능 차이에
15배의 GPU Infra 필요

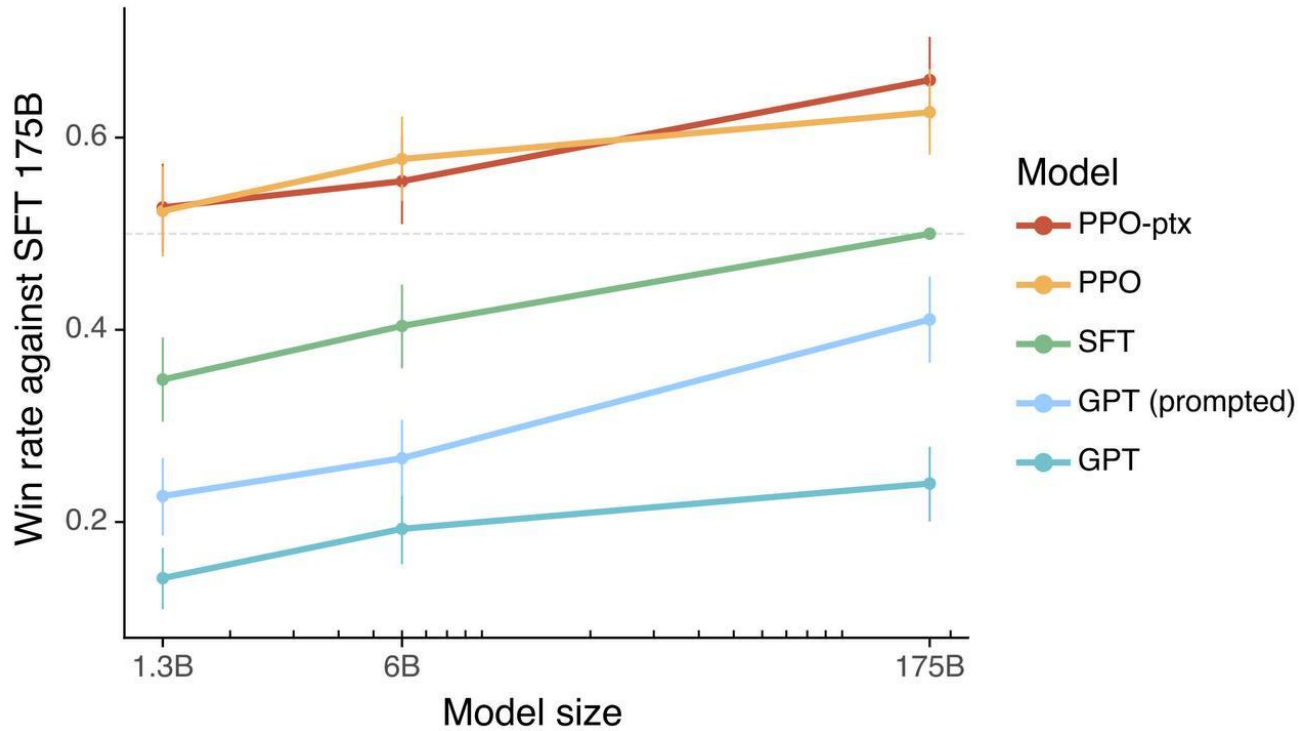
Fine-tuning 제약

성능 향상을 위해
(굉장히) 많은 Data 필요
별도 비용 발생

1. LLM - sLLMs

Fine-tuning 을 통해 특정 업무에서 LargeLLM 수준 이상의 성능을 발휘하나 기술 발전이 빠르고 숙련된 경험 필요

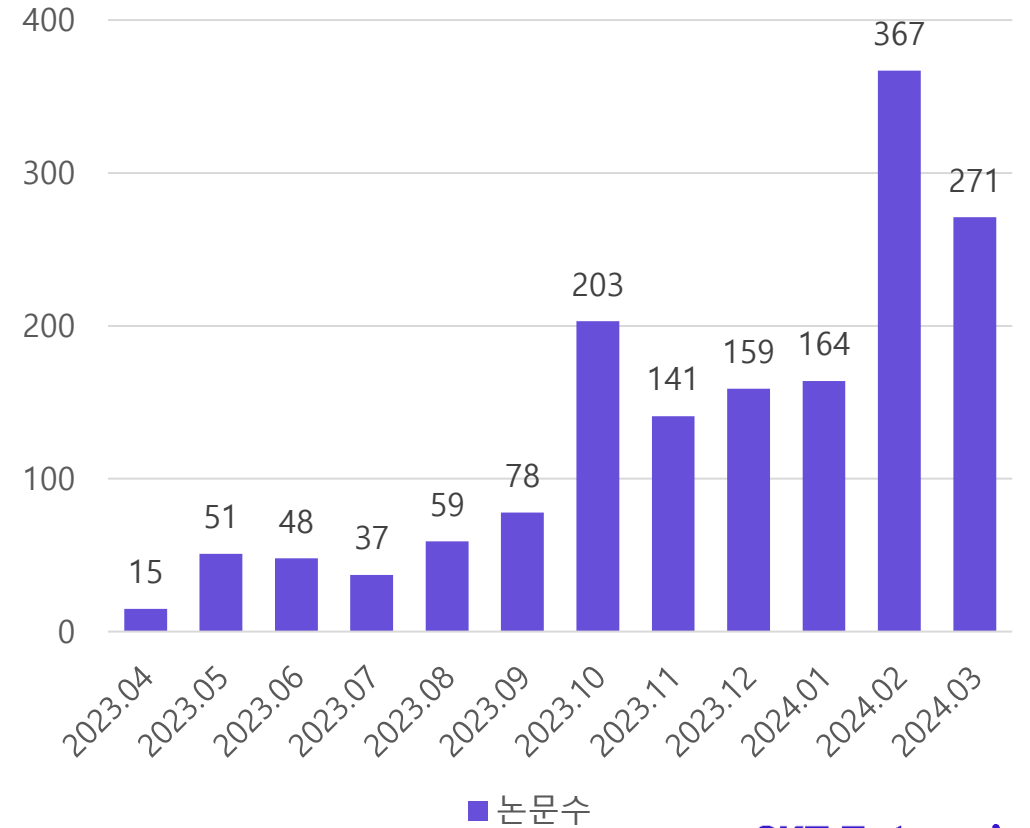
모델 사이즈별 성능 변화



출처: OpenAI 논문 Training language models to follow instructions with human feedback with PPO(Proximal Policy Optimization)



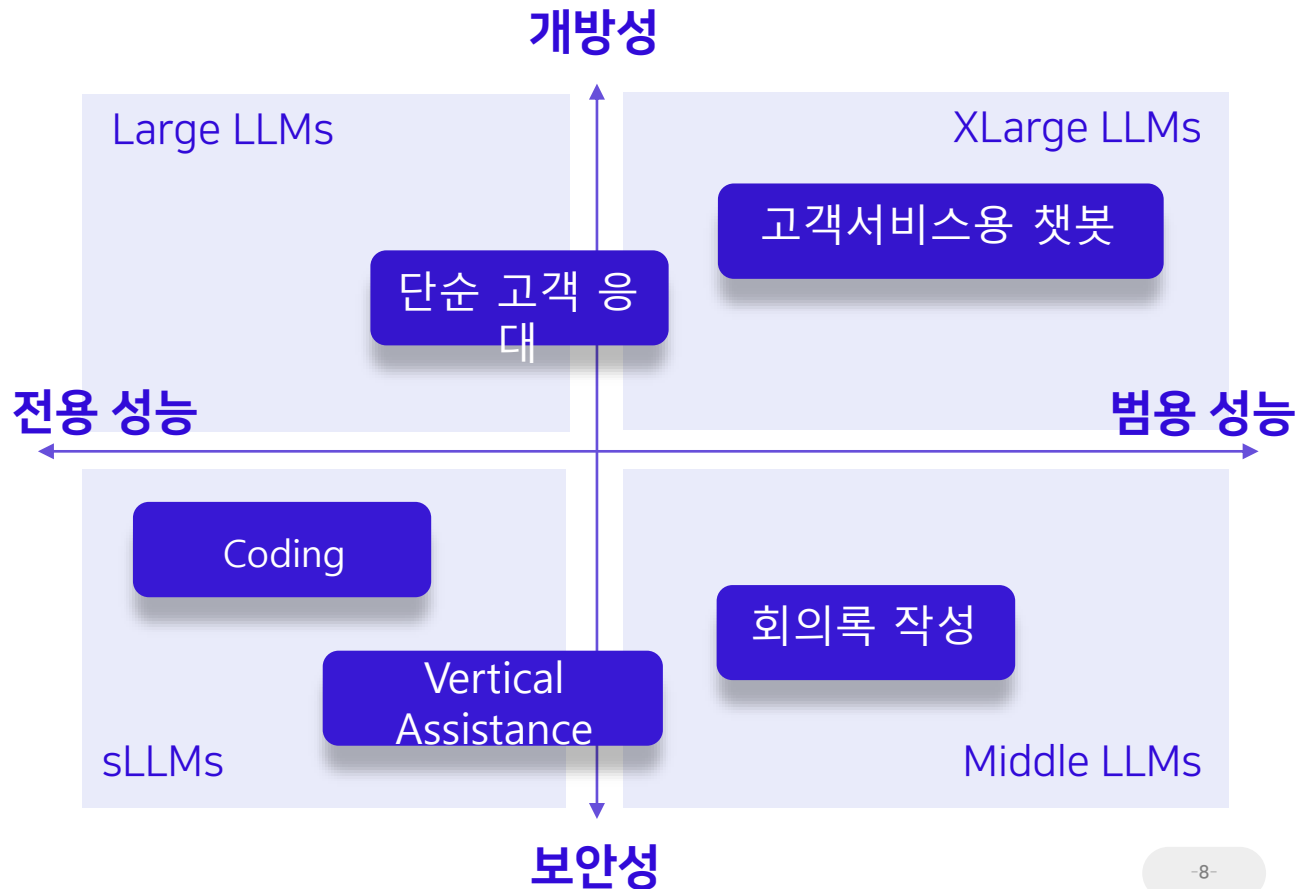
Fine-tuning 논문수



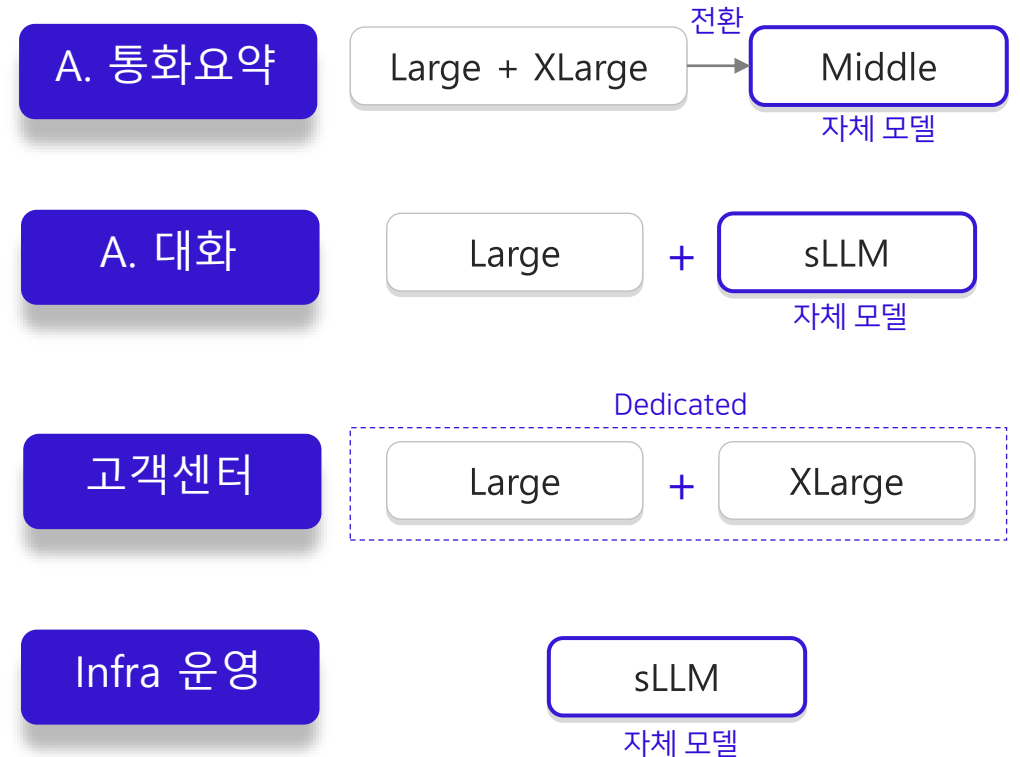
1. LLM – Multi LLM을 고려한 업무별 / 단계별 도입 전략이 중요

적용 업무 특성, 도입의 시급성, 비용효율성을 고려한 각 기업별 Multi-LLM 적용 전략 수립

Multi LLM 도입 전략 Matrix 예시

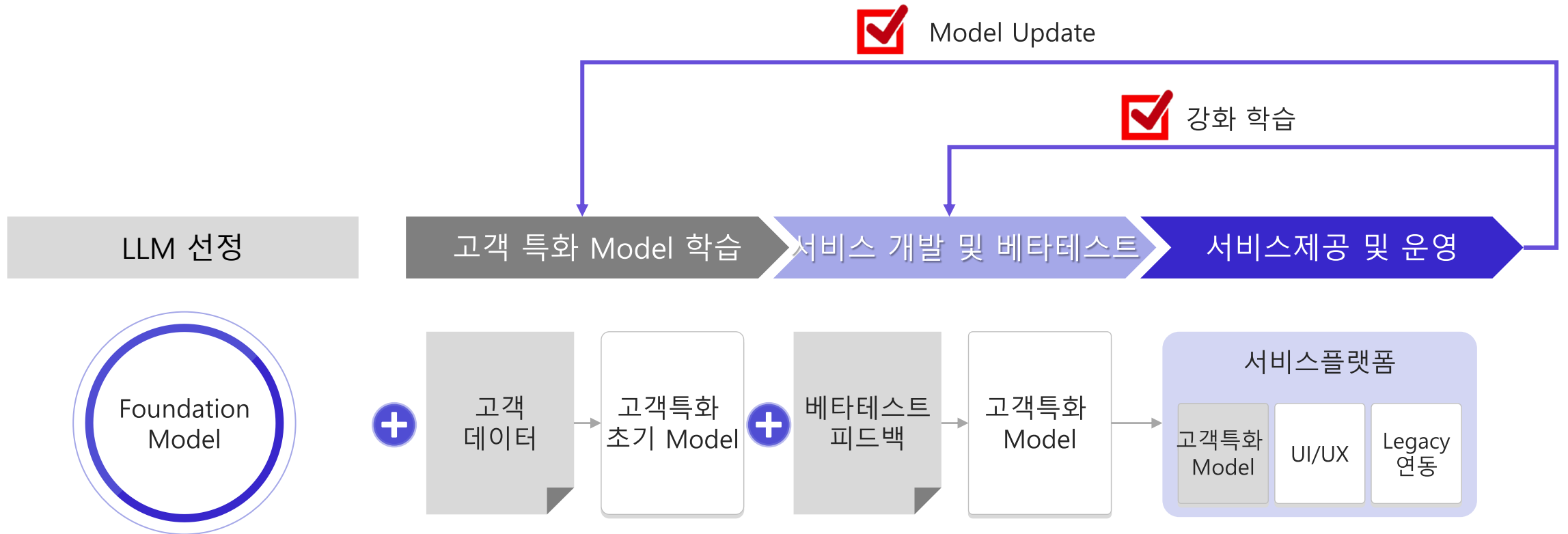


SKT Multi-LLM 적용 사례 예시



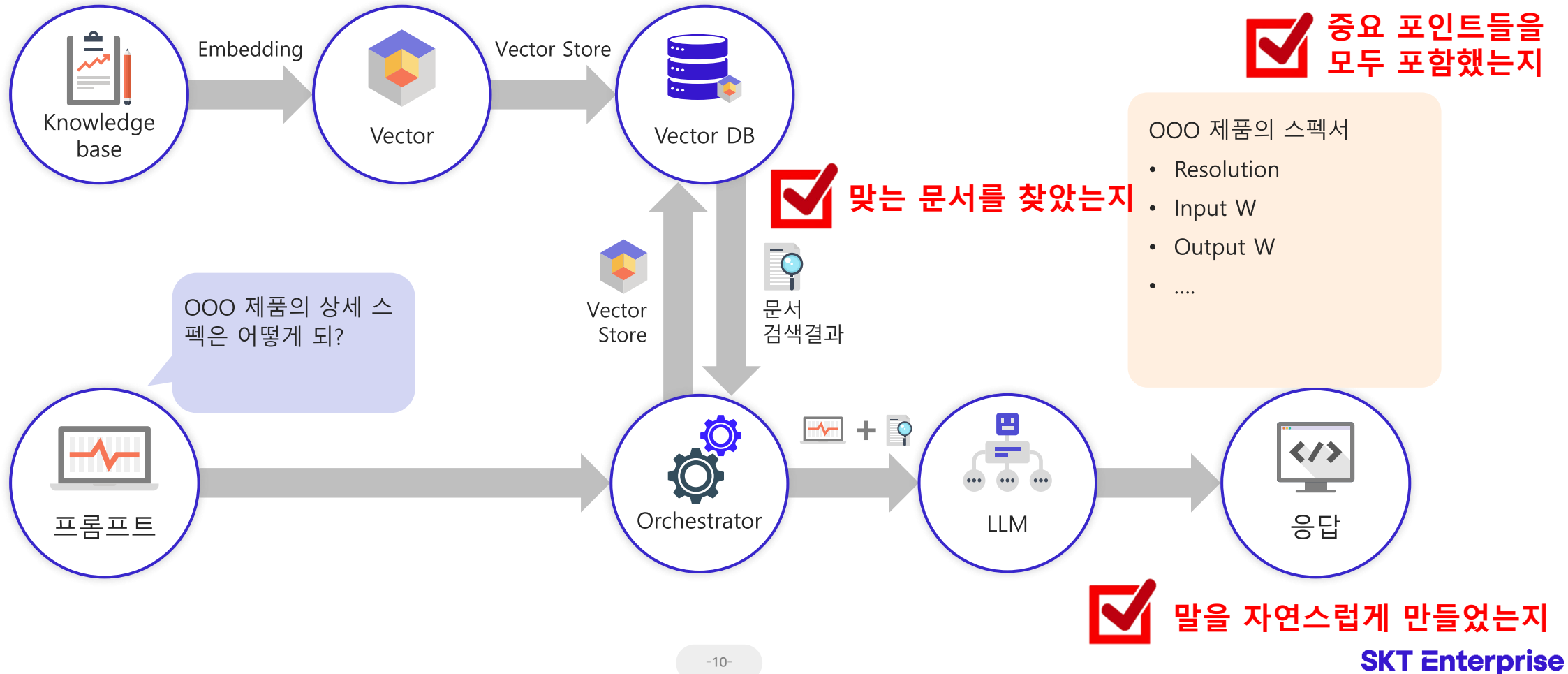
2. Evaluation - LLM Ops

AI 서비스 특성 상 지속적인 사용자 피드백을 통한 강화 학습과 Data 신규 생성 및 정책 / 환경 변화에 대응이 필수적
But, Model 변경에 따른 서비스 품질 저하 / 향상에 대한 정량적 측정 어려움



2. Evaluation – RAG의 중요성

기업 업무 적용 시 RAG를 통한 내부 문서 기반 생성이 필수적이거나 아직까지 이에 대한 체계적인 성능 검증은 미흡



2. Evaluation – Hugging Face Leaderboard

일반적인 평가 방법론은 범용 성능을 기반으로하며, 기업별/업무별 Fine-tuned Model은 평가가 어려움

🤖 Open LLM Leaderboard

The screenshot shows the Hugging Face Open LLM Leaderboard interface. It includes a search bar, filter options for columns to show (Average, ARC, HellaSwag, MMLU, TruthfulQA, Winogrande, GSM8K, Type, Architecture, Precision, Merged, Hub License, #Params (B), Hub, Model sha), and filter options for model types (pretrained, continuously pretrained, fine-tuned on domain-specific), precision (float16, bfloat16, 8bit, 4bit, GPTQ), and model sizes (in billions of parameters). A table of model performance metrics is displayed below the filters.

T	Model	Average ↑	ARC	HellaSwag	MMLU	TruthfulQA
◆	davidkim205/Rhea-72b-v0.5	81.22	79.78	91.15	77.95	74.5
💬	MISAIR/MultiVerse_70B	81	78.67	89.77	78.22	75.18
◆	MISAIR/MultiVerse_70B	80.98	78.58	89.74	78.27	75.09
◆	SF-Foundation/Ein-72B-v0.11	80.81	76.79	89.02	77.2	79.02
◆	SF-Foundation/Ein-72B-v0.13	80.79	76.19	89.44	77.07	77.82
◆	SF-Foundation/Ein-72B-v0.12	80.72	76.19	89.46	77.17	77.78
◆	abacusai/Smaug-72B-v0.1	80.48	76.02	89.27	77.15	76.67
◆	ibivibiv/alpaca-dragon-72b-v1	79.3	73.89	88.16	77.4	72.69

범용 성능에 대한 평가

- ARC (AI2 Reasoning Challenge)
초등 수준의 객관식 과학 문제 평가
- HellaSwag
가장 적합한 문장을 선택하여 추론 능력평가
- MMLU (Massive Multitask LLM Understanding)
57개 과목의 객관식 문제로 지식 평가
- TruthfulQA
금융, 법률, 건강 등 38개 범주에 대한 답변

Model 노출 우려

- 기업 내부용 Model의 외부 유출 우려
- 범용 성능과 내부 업무 목적의 성능간 괴리

2. Evaluation – Private LLM Leaderboard

자체 Fine-tuning 한 모델의 성능을 외부 공개 없이 내부 평가하여 지속적인 성능 변화 확인

엔터프라이즈 CIC LLM 리더보드

엔터프라이즈 CIC LLM 리더 보드는 다양한 Open/Private LLM 들을 평가합니다.

모델을 평가를 위해 '제출' 페이지를 참고하세요. 리더보드의 모델평가는 [Eleuther AI Language Model Evaluation Harness](#) 로 진행됩니다.

'소개' 페이지에서 자세한 내용을 확인할 수 있습니다.

LLM Benchmark | Metrics through time | 소개 | 제출

Search for your model (separate multiple queries with ';') and press ENTER...

Select columns to show

- Average ↑
- ARC
- HellaSwag
- MMLU
- TruthfulQA
- Winogrande
- GSM8K
- Type
- Architecture
- Precision
- Merged
- Hub License
- #Params (B)
- Hub
- Model sha

Show private/deleted models | Show merges | Show MoE | Show flagged models

T	Model	Average ↑	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
○	CIC/mistralopithecus-v1-dpo-7b	46.4	48.55	71.05	43.06	44.79	69.46	1.52
◆	CIC/LDCC-SOLAR-10-7B	46.4	48.55	71.05	43.06	44.79	69.46	1.52
○	0x7194633/fialka-7B-v3	46.4	48.55	71.05	43.06	44.79	69.46	1.52
○	0x7194633/fialka-13B-v3	34.58	30.97	48.83	26.36	40.58	59.43	1.29
○	0x7194633/fialka-13B-v3.1	34.11	29.95	47.28	25.41	43.03	58.48	0.53
○	0x7194633/nanoFialka-v1	28.48	22.01	28.12	25.03	45.26	50.43	0

Citation

Huggingface 와 동일한 Interface

- Huggingface 와 동일한 평가 구성
- On-Premise 구축

Model 노출 없음

- Model upload 없이 평가 가능
- 점수 평가 결과 내부 공유

지속적인 모델 성능 관리

- Fine-tuning Model별 성능 관리
- 평가 방법 추가 시 지속 update

2. Evaluation – 기업 업무 특화 LLM 성능 평가 Evaluator

실제 Task에 적용된 AI 생성 답변을 평가하는 Evaluator Tool 활용, 정량적인 평가 검증

Data Set 등록

- 질문, 답변

평가 모델 선택

- 평가 LLM

평가 Metric 선택

- Criteria

평가

- Key Criteria
- Retriever

답변 정확도 : 74%

SKT Evaluator

Dataset Evaluation 등록/수정

Model: A.X Model Version 1

Dataset: Q&A Golden Set

Metrics:

- Chatbot
- Q&A
- Summarization
- Custom

Metric 1: Description description ... (Q&A)

Metric 2: 낮을 수록 좋고, 유사도를 평가하는 ... (Summarization)

Metric 3: Description description ... (LLM, Q&A)

Metric 4: Description description ... (Summarization)

Metric 5: Description description ... (Chatbot)

RUN

3. GPU - nVIDIA 독점 시장 구조

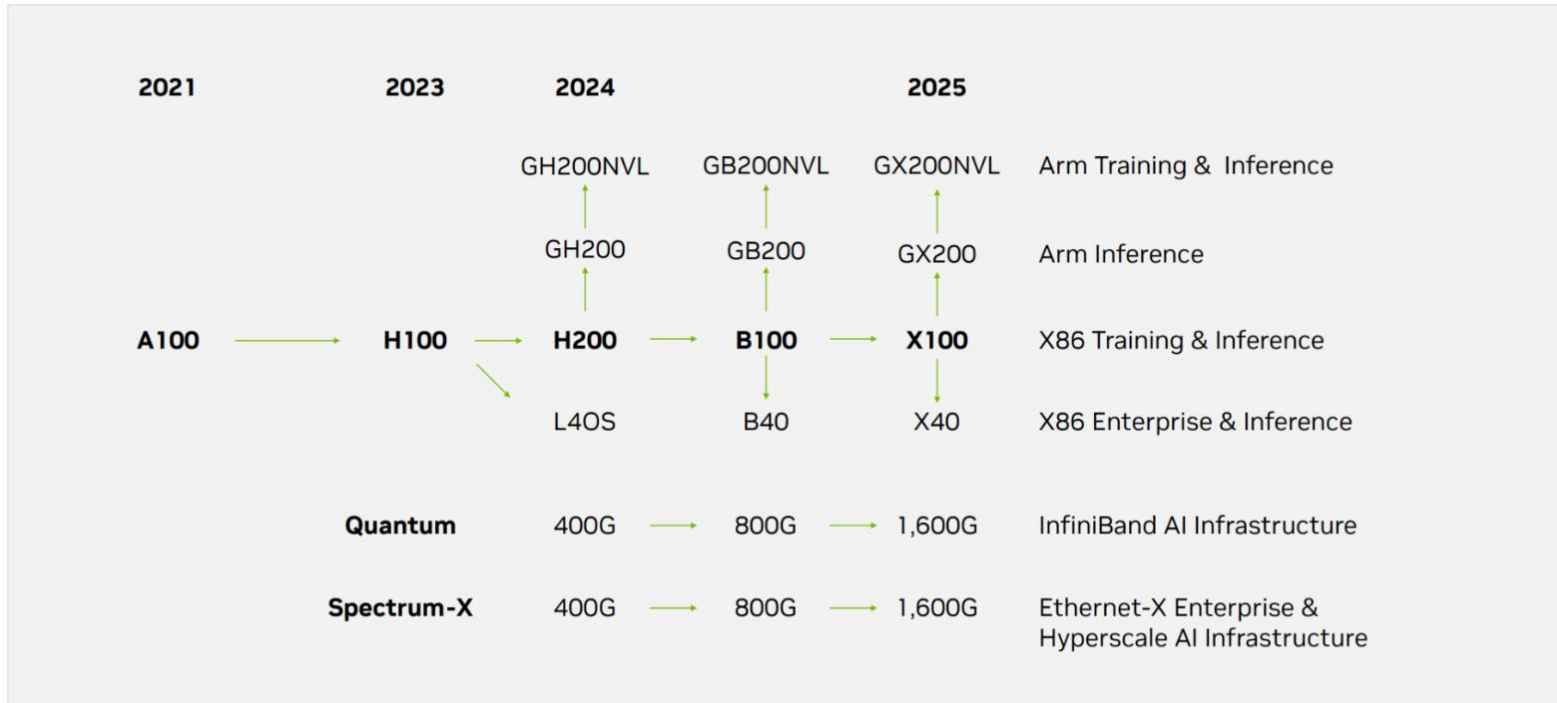
nVIDIA의 제품 Life-cycle이 짧아지고 있으며, 성능 개선과 함께 가격 상승을 수반하여 도입 의사결정이 어려움

NVIDIA AI - One Architecture | Train and Deploy Everywhere

From Two-Year Rhythm

to

One-Year Rhythm | Training & Inference | x86 & Arm | Hyperscale & Enterprise



출시주기

2 Y → 1 Y

성능수치

3 ~ 6 X

출시단가

4 ~ 50% ↑

출처: nVIDIA Datacenter/AI GPU roadmap



3. GPU - Infra 구축 전략

기업의 Data Center 환경, 서비스 개발 / 확장 계획, 예산 등을 고려한 다양한 방식의 Infra 구축 활용 전략 수립
→ 활용 용도(학습/추론 등), 학습 빈도/규모 등을 고려, 다양한 방식 혼용

On-Premise

자체 D/C 활용

자체 D/C의 전력용량,
공조설비가 충분한 경우

AI D/C Colocation

GPU 서버를 위한 별도 설비가
구축된 외부 D/C 임차
(Rack 당 40KW 이상)

On-Demand Cloud

Per Instance

필요한 시점에 필요한 만큼만
할당 받아 사용

Per Cluster

InfiniBand가 지원되는 다수
GPU를 기간 약정으로 사용

Reserved Cloud

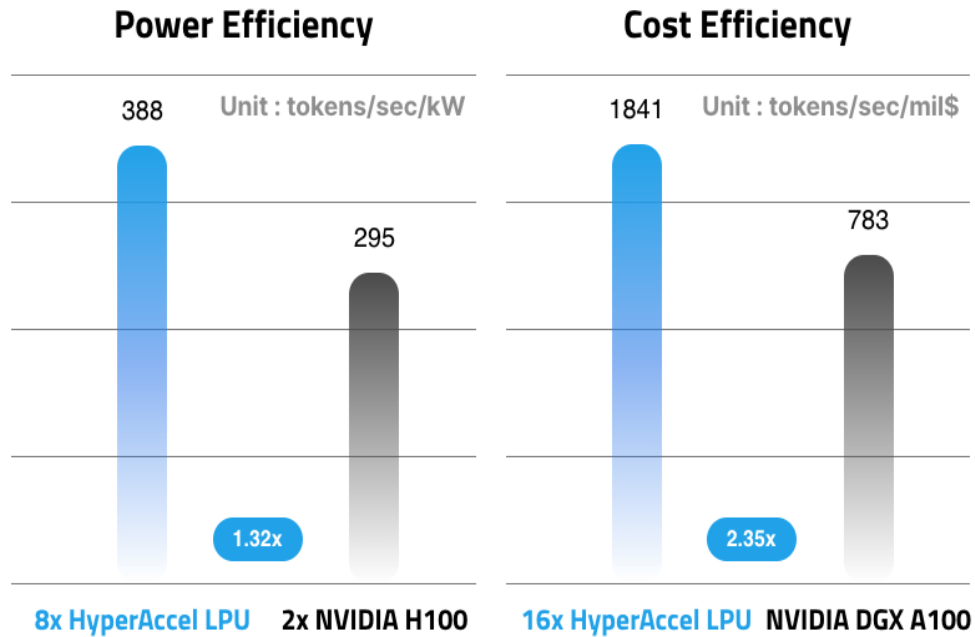
대규모의 GPU를 기간 약정,
Upfront 비율에 따라
할인된 가격으로 사용

규모에 따라 망분리가 적용된
Private Zone으로 구성

3. GPU - NPU 등 대체제에 대한 고려

중장기 대안으로 NPU를 활용한 추론 전용 AI 서버의 활용 고려

NPU 서버의 성능 예시



SAPEON, 리벨리온, 하이퍼액셀 등 NPU 기반
추론 전용 시스템이 개발 중이나,
제품 양산, 모델 Porting, 유지보수 등의
기술 지원 체계에 대한 검증 속제

출처: HyperAccel Orion Spec. Sheet

4. Use Case – SKT AI Portal

사내 Portal 에 내부 시스템과 외부 웹 검색이 가능한 AI 서비스 포털로 구성원 업무 지원

The screenshot shows the SKT AI Service Portal interface. At the top, there are navigation tabs for 'CHAT', 'IMAGE', and 'TRANSLATE'. The user is logged in as 'AI Enterprise사업 강인식 님'. On the left, there is a 'Knowledge DB' section with a dropdown menu showing 'ARRA' and '내 문서함'. The main chat area displays a conversation with the AI, including a search result for '금융 관련 최신 AI 트렌드에 대해 알려줘'. The chat interface includes a search bar, a model selector (GPT-4.0 TURBO), and a '웹 검색' toggle. Callouts highlight key features: '사내 지식관리시스템 연동하여 검색' (Search integrated with internal knowledge management system), '개인 문서 업로드하여 작업 가능' (Ability to upload personal documents for work), 'Chat, 이미지생성, 번역 지원' (Support for chat, image generation, and translation), '외부 웹 검색하여 상위 15개 활용' (Utilizing top 15 external web search results), and '멀티 LLM 선택 가능' (Multi-LLM selection capability).

Knowledge DB ①

ARRA

내 문서함 ①

문서를 업로드 해보세요

5MB 이하 파일만 업로드할 수 있습니다

문서 업로드

CHAT IMAGE TRANSLATE

AI Enterprise사업 강인식 님 FAQ 릴리즈 노트 AI Service ▾

금융 관련 최신 AI 트렌드에 대해 알려줘 me

Chat, 이미지생성, 번역 지원

- ✓ 검색 키워드를 생성합니다.
- ✓ BING을 통해 웹 문서를 검색합니다.
- ✓ 검색한 웹 문서 중 상위 15개를 활용하여 답변을 생성합니다.

금융 분야에서의 AI 트렌드는 다음과 같습니다:

1. 생성형 AI 활용: 생성형 AI는 텍스트, 코드, 대본, 음악 조각, 이메일, 편지 등과 같은 새로운 창의적인 콘텐츠를 생성할 수 있는 인공지능의 한 유형입니다. 이는 포트폴리오 최적화, 사기 탐지, 리스크 관리 등의 분야에서 활용되며, 새로운 효율성을 찾고자 하는 기업들 사이에서 빠르게 인기를 얻고 있습니다 17 19 .
2. 하이브리드 클라우드 활용: 금융 서비스 회사는 민감한 데이터를 클라우드로 마이그레이션할 수 없기에, AI 훈련과 추론에 대한 지출을 최적화하길 원합니다. 이

문서 검색 결과 (총 24건)

1. 엔비디아의 네 번째 연례 금융 서비스 AI 현...
엔비디아의 네 번째 연례 금융 서비스 AI 현황 보고서는 2024년의 현재 환경과 새로운 트렌드에 대한 인사이트...
2. 은행 및 금융 서비스 분야의 AI활용 트렌드와...
은행 및 금융 서비스 분야의 AI활용 트렌드와 사례 | UiPath. UiPath Blog. 산업별 솔루션. 은행 및 금융 서...
3. 2024년 금융 산업 it 트렌드. YS 2024. 2. 7 ...
2024년 금융 산업 it 트렌드. YS 2024. 2. 7 오후 3:12:029 min read. 2024년 금융 산업 it 트렌드. 세계는 지금 기...
4. Allganize 뉴스레터. 2024년 AI 트렌드를 ...
Allganize 뉴스레터. 2024년 AI 트렌드를 정리했습니다. 트렌드코리아 2024, IT 트렌드 2024, 이코노미스트와 ...

< 1 2 3 4 5 6 >

GPT-4.0 TURBO 웹 검색 ①

CHATGPT에게 물어보거나, "/"를 입력하여 명령어를 사용

멀티 LLM 선택 가능

새 채팅 히스토리

사내 법무 검토 관련 필요한 절차와 문서 양식을 제공

미리보기
비밀유지계약서(쌍방)

↓ 다운로드 X 닫기

비밀보호 계약서

SK텔레콤주식회사(이하 "SK텔레콤"이라 한다)와 _____(이하 "_____"이라 한다)은 다음과 같이 비밀보호계약을 체결한다.

제1조 (목적)

본 계약은 "SK텔레콤"과 "_____"가 _____과 관련하여 상호 협력함에 있어 일방당사자(이하 "제공회사"라 한다)가 상대방(이하 "수령회사"라 한다)에게 제공하는 '비밀정보'의 보호와 관련하여 필요한 제반사항 및 상대방사자의 권리·의무를 정하는 것을 목적으로 한다.

제2조 ("비밀정보"의 내용)

'비밀정보'라 함은 _____과 관련하여 구두, 시각 또는 서면 등 수단에 관계없이 양 당사자 상호간에 직접 또는 간접으로 제공되는 다음 각호의 정보를 의미하되, 이에 한하지 아니한다.

- 설계서, 계획서, 시방서, 도면, 하드웨어, 소프트웨어, 데이터, 제품계획과 예측, 시제품, 경영-기술-재무정보와 "제공회사"가 "수령회사"에게 비밀정보라고 고지한 정보 및 통상의 경우 비밀정보라고 간주되는 기타 정보 등
- "제공회사"가 제공한 제1호의 정보를 기초로 하여 "수령회사"가 가공 또는 변경한 제반 정보

T legalnet AI 계약서 Agent

항들을 계약서에 추가할 수 있습니다. 계약서는 사전에 모든 가능한 상황을 고려하고 문제 발생 시 이를 해결할 수 있는 방안을 제시해야 합니다.

본 계약 전에 무엇을 준비해야 해?

T 본 계약 체결 전에는 '양해각서' 또는 '비밀유지계약서'를 준비하는 것이 좋습니다. 이 두 계약은 계약에 앞서 당사자 간의 약속된 내용을 확인하고, 비밀 정보 유출 등을 방지하기 위해 필요합니다. 비밀유지계약서는 주로 상대방에게 제공하는 정보가 중요한 비즈니스 정보인 경우, 양해각서는 계약 체결에 앞서 당사자 간의 약속된 내용을 순서대로 확인하기 위해 필요한 사전 계약서입니다.

국문 양해각서

↓ 다운로드 Ⓞ 미리보기

비밀유지계약서(쌍방)

↓ 다운로드 Ⓞ 미리보기

미리보기로
해당 양식 조회

질문을 자연어로
입력

구체적인 설명

관련 양식 제공

4. Use Case – 마케팅 Data 분석

Analytics and Targeting Suite 3.0

김상목님 | 사용안내 | 소개 영상 | FAQ

데이터서비스 ▾
▼ 최대도달범위 / 명

All 전체필터보기

MNO ▾

앱/웹 Usage

통화 Usage

위치/이동특성

예측 모델 ▾

SK브로드밴드 ▾

SK플래닛 ▾

11번가

분석대상(개인정보활용동의)
전체회선 : 17,393,342(55.46%)
개인회선 : 16,884,636(69.18%)
데이터 기준일 2024.04.22
데이터 집계기간 Ⓞ
2024.03.23 ~ 2024.04.22

공지 사항

기능종료 2024.04.16	신규기능 2024.01.22	기능개선 2023.08.29	기능종료 2023.07.21	기능변경 2023.03.20
우주팩스 slim/mini 관련 필...	장기가입 기간 필터 추가 안내	강원도 및 군위군 위치/이동특...	거주지 건물 유형 필터 종료 안내	전사 데이터 자원 체계 개선 안내
우주팩스 slim/mini 신규 가입 중단으로 인하여 예측 모델 > 고객 > T 우주 가입 예측 / T 우주 해지 예측 필터에서 제외되었습니다.	장기가입 기간 필터가 추가되어 안내드립니다. '장기고객' 혜택 적용을 위한 선정 기간으로, 회선 유지 기간 중 정지 기간은 제외됩니다. [MNO > 회선 > 장기가입 기간] 에...	안녕하세요, 강원도 및 군위군의 metadot a 작업으로 인하여 "최근 조회 필터"에서 과거에 선택한 지역 필터를 재사용하시면 무수가 0으로 나올 수 있습니다. 해당 필터들...	안녕하세요, 데이터 제공 중단으로 인하여 '거주지 건물 유형' 필터가 2023년 7월 21일부로 종료되었으니 참고 부탁드립니다. 감사합니다.	안녕하세요, 더욱 신속하고 효율적인 운영을 위해 2023년 3월 20일부로 ATS 관련 문의 및 건의 사항은 Data Portal을 통해 접수 부탁드립니다. [Tnet > Data Porta...

전체 필터별 사용 횟수

MNO	앱/웹 Usage	통화	위치/이동특성	예측 모델	SK브로드밴드	SK플래닛	11번가
연령 2789	SKT_에이닷(ADOT) 137	부동산	세무/회계	정보광고 수신 동의 1888	넷플릭스(Netflix) 82	금융기타 91	Tmap 목적지 79
SMS 수신거부 여부 1123	SKT_에이닷TV(A.tv) 80	금융기타 91	자영업자 추정 236	단골고객 여부 816	유튜브(Youtube) 61	학습지 91	이동수단 71
부가 서비스명 758	부동산 61	농/수/축/임협 89	출퇴근 이동거리 62	단골고객 여부 816	부동산 61	Life Stage Seg. 63	5G기변예측 76

자연어 질의를 통한 Data 검색 / 분석 실행

(예. 강남 거주 아이폰 사용자 연령별 통계 뽑아줘, 이 중 넷플릭스 보는 사람 뽑아줘)

SmartBee

SmartBee

SmartBee와 대화를 통해 원하는 고객을 추출해보세요!
(ex. 아이폰을 사용하는 사람 뽑아줘)

채팅을 이어나가며 고객군을 수정하거나, 더 구체화 시킬 수 있습니다.
(ex. 이 중에서 넷플릭스 보는 사람 뽑아줘)

AI가 탑재된 ATS를 이용해보세요!

2024-04-18 14:30:31
[작업] [X]

생성 AI 서비스 도입을 위한 핵심 고려 사항

Multi LLM

성능 Evaluation

GPU

- 학습 비용과 데이터 이슈
 - 빠른 기술 발전
 - 보안
-
- 지속적인 성능 관리
 - 효과 측정을 위한 정량적 평가 지표
 - 모델 보안
-
- GPU 제품 Life-cycle 단축
 - 높은 초기 구축 비용

SKT 제공 Value

자체 개발 모델 포함
Multi LLM 플랫폼 제공

Private Leaderboard 와
Evaluator로 지속적 성능 관리

GPU as a Service('24. 4Q)
준비 중

감사합니다.