우리회사 생성형 AI 구현에

기업 맞춤형 watsonx 플랫폼이 필요한 이유

2023.11.21

IBM Customer Success Manager
Data & AI specialty architect
이규봉 부장

우리회사 생성형 AI 구현에

기업 맞춤형 **watsonx 플랫폼**이 필요한 이유

dd

1. **생성형 AI 시대, 기업 동향과 고민은?**
2. **생성형 AI, Large Language Model의 특징**
3. **모델 이야기** : 개발에서 선택으로!
4. **성능 이야기** : 적은 비용으로 최대의 효과를!
5. **운영 이야기** : Trustworthy 생성형 AI 구현!
6. **기업 맞춤형 watsonx 플랫폼**

Here we are,
10 months after the biggest bang of most of our careers – **generative AI** – and

The question is no longer what is it, but how are organizations putting it into action?

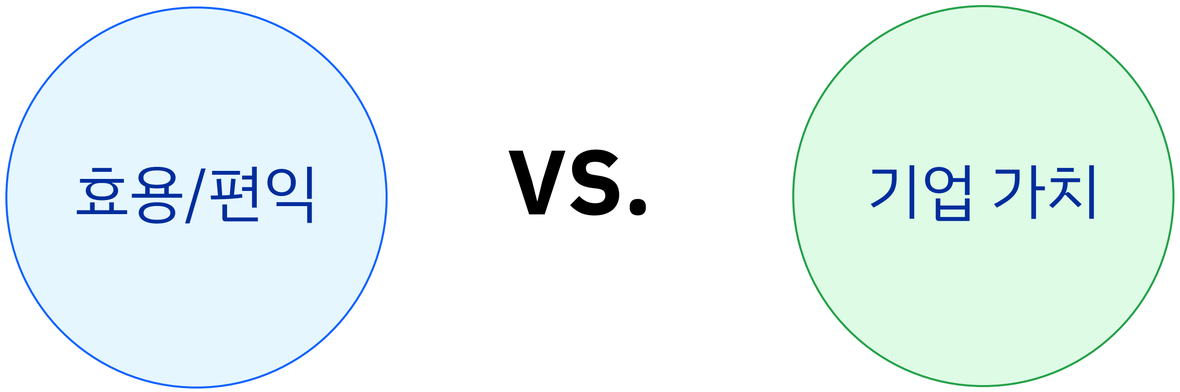
Generative AI Realities: Proactive Approaches for Quantifiable Business Results by Gartner
https://www.gartner.com/en/webinar/530960/1196465

Disruption and Uncertainty increase

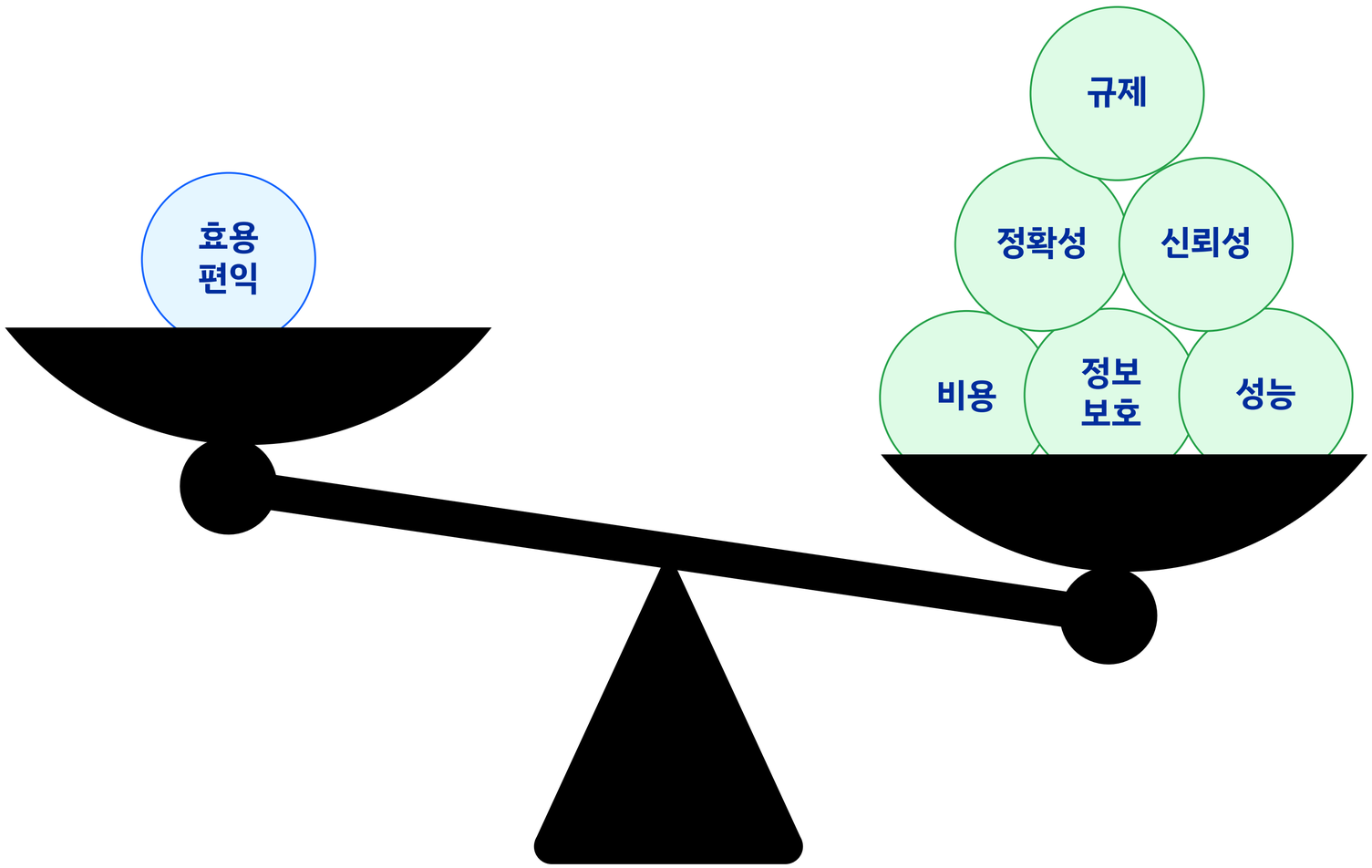
Business as usual (wait & see) approach is riskier than investing.

Organizations that lead through turbulence with intention will be the most successful.

새로운 기술, 특히 혁신적인 기술을 도입하기 전에 기업은 항상 고민을 합니다.



우리는 생성형 AI 도입을 위해 어떤 선택을 해야 할까요?



우리회사 생성형 AI 구현에

기업 맞춤형 **watsonx 플랫폼**이 필요한 이유

1. **생성형 AI 시대, 기업 동향과 고민은?**

dd

2. **생성형 AI, Large Language Model의 특징**

3. **모델 이야기** : 개발에서 선택으로!

4. **성능 이야기** : 적은 비용으로 최대의 효과를!

5. **운영 이야기** : Trustworthy 생성형 AI 구현!

6. **기업 맞춤형 watsonx 플랫폼**

Transformer 기반의 Foundation Model의 등장은 기업의 AI 채택을 가속화하고 있습니다.

Expert Systems

- No use of data
- Manually authored rules
- Brittle

1980s

Machine Learning

- Need large data sets
- Less brittle, but labor intensive
- Demanding data prep and feature engineering

1990s

Deep Learning

- Massive data and compute
- Automatically learn if there is enough labeled data
- Enterprise adoption limited by availability of labeled data

2010s

Foundation Models

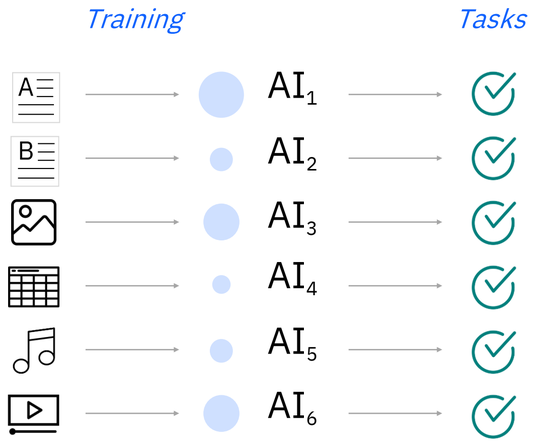
- Self-supervision at scale
- Massive data and compute
- Learn from lots of data without requiring labels
- Adapt quickly to many tasks
- Accelerate enterprise adoption

Transformer

2020s

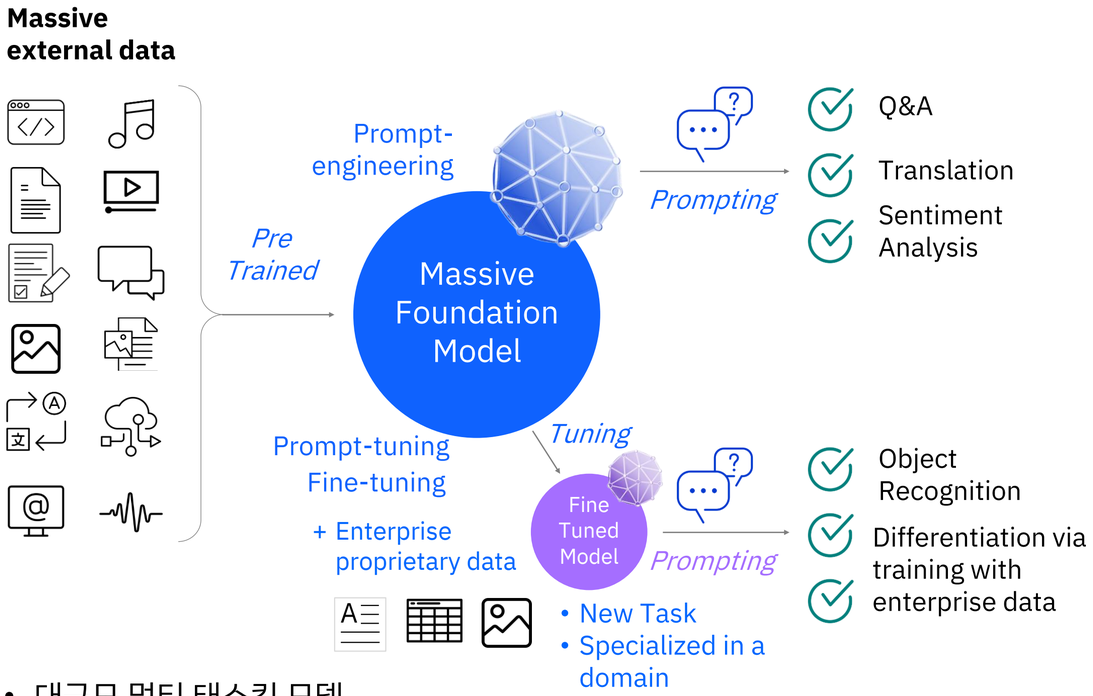
사전 훈련된 Foundation Model은 훈련이 거의 없이도 높은 성능의 멀티 태스킹이 가능하여 기업의 다양한 업무에 빠르게 적용할 수 있습니다.

Traditional AI models



- 개별 사일로 모델
- task별 특화된 교육 필요
- 사람이 직접 감독하는 다양한 훈련 필요

Foundation Models



- 대규모 멀티 태스킹 모델
- 훈련이 거의 또는 전혀 없이도 적용 가능
- 사전 훈련된 비지도 학습

향상된 기능

- Summarization
- Conversational Knowledge
- Content Creation
- Code Co-Creation

주요 강점

- 더 적은 라벨링을 통해 **초기 비용 절감**
- fine-tuning 및 추론을 통한 **빠른 배포**
- 여러 사용 사례에 대해 **동일하거나 더 나은 정확도**
- 더 나은 성능을 통한 **지속적 수익 증가**

특정 NLP 타스크들 최대 70% 감소

일반적인 Large Language Model AI 태스크

Retrieval-Augmented Generation (검색증강생성)

문서 또는 동적 콘텐츠를 기반으로 챗봇 또는 질의 응답 기능을 만듭니다.

최신 지식 기반 Q&A 구축
직원 및 고객 서비스 지원

Summarization

도메인별 콘텐츠가 포함된 텍스트를 핵심 사항을 캡처하는 개인화된 개요로 변환합니다.

대화 요약, 매뉴얼,
계약 정보 - 보험 적용 범위,
회의록 정리

Content Generation

특정 목적을 위한 텍스트 콘텐츠를 생성합니다.

마케팅 캠페인,
Job Description,
블로그 게시물 및 기사,
이메일 초안 작성 지원

Named Entity Recognition

구조화되지 않은 텍스트에서 **필수 정보**를 **식별**하고 **추출**합니다.

감사 가속화

Insight Extraction

기존의 구조화되지 않은 텍스트 콘텐츠를 분석하여 **특수 도메인 영역**에서 **인사이트를 도출**합니다.

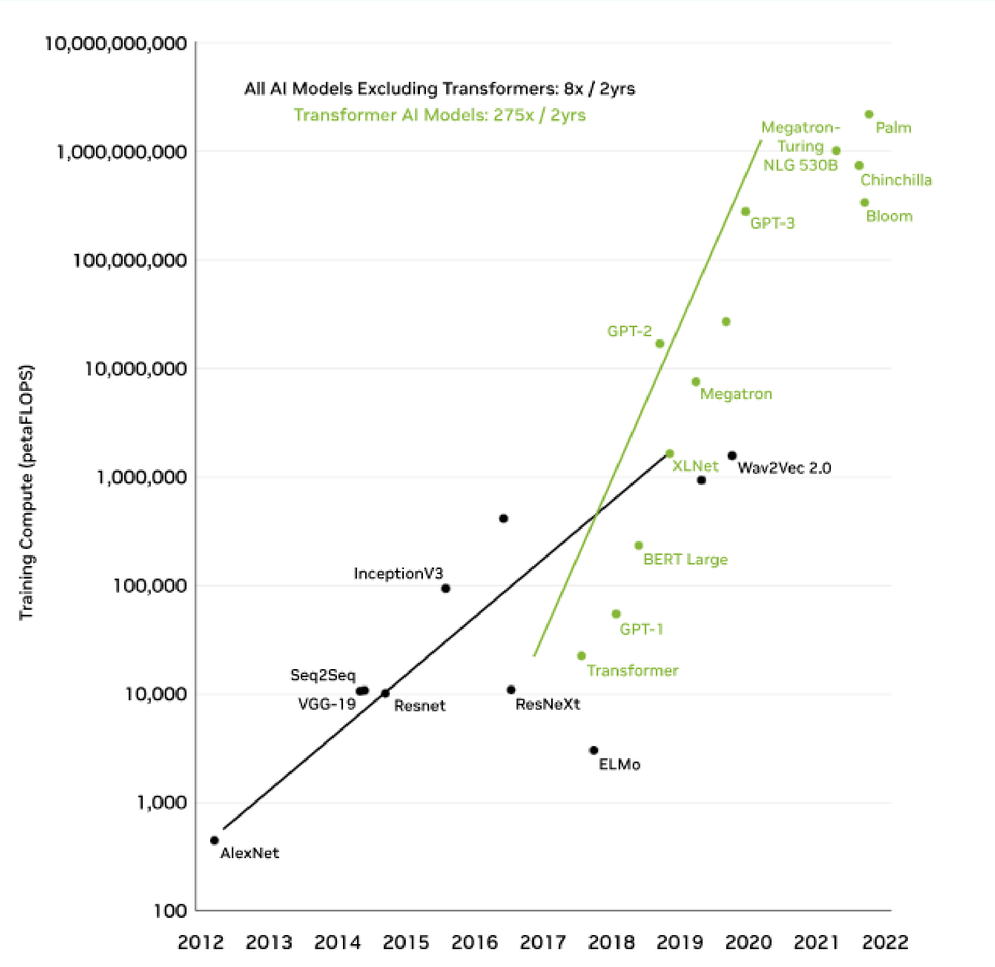
의료 진단 지원, 사용자 연구 결과

Classification

최소한의 예제로 작성된 입력을 읽고 분류합니다.

고객 불만 분류,
위협 및 취약성 분류,
감성 분석,
고객 세분화

What are the Challenges of Large Language Models?



Compute required for training transformer models

대규모 언어 모델을 개발하고 유지하는 데 필요한 상당한 자본 투자, 대규모 데이터 세트, 기술 전문 지식 및 대규모 컴퓨팅 인프라는 대부분의 기업이 Generative AI로 가는 진입 장벽입니다.

1. **Compute-, cost-, and time-intensive workload**:
   - Significant capital investment, technical expertise, and large-scale compute infrastructure are necessary to maintain and develop LLMs.
   - Training an LLM requires thousands of GPUs and weeks to months of dedicated training time.
   - Some estimates indicate that a single training run for a GPT-3 model with 175 billion parameters, trained on 300 billion tokens, may cost over $12 million dollars in just compute.
2. **Scale of data required**:
   - As mentioned, training a large model requires a significant amount of data. Many companies struggle to get access to large enough datasets to train their large language models.
   - This issue is compounded for use cases that require private - such as financial or health - data. In fact, it's possible that the data required to train the model doesn't even exist.
3. **Technical expertise**:
   - Due to their scale, training and deploying large language models are very difficult and require a strong understanding of deep learning workflows, transformers, and distributed software and hardware, as well as the ability to manage thousands of GPUs simultaneously.

https://www.nvidia.com/en-us/glossary/data-science/large-language-models/

페타플롭스(PetaFlops)는 1초당 1,000조 번의 수학 연산처리

우리회사 생성형 AI 구현에

기업 맞춤형 **watsonx 플랫폼**이 필요한 이유

1. **생성형 AI 시대, 기업 동향과 고민은?**
2. **생성형 AI, Large Language Model의 특징**
3. **모델 이야기** : 개발에서 선택으로!
4. **성능 이야기** : 적은 비용으로 최대의 효과를!
5. **운영 이야기** : Trustworthy 생성형 AI 구현!
6. **기업 맞춤형 watsonx 플랫폼**

Multi Model Strategy > One model does not fit for all use cases

비즈니스에 성공적인 AI 도입을 위한 **다양한 Model Provider 옵션 제공!**
기업의 다양한 요구 사항을 유연하게 대응하기 위한 Multi Model Strategy

Build **your own models**

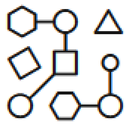
Use **IBM verified Open Source models**

Use **IBM developed models**

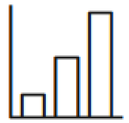
모든 모델이 똑같이 만들어지는 것은 아니며,
업계, 도메인, 사용 사례에 따라 **최적의 모델이 달라집니다.**

Try → Tune → Build and evaluate → Deploy on any cloud

도메인 특화된 비즈니스 문제 해결
학술, 법률, 재무, 화학, 제조 등 도메인 특성 반영



다양한 모델 활용 유연성 제공
다양한 규모, 아키텍처 모델을 활용한 성능, 속도, 효율성 향상



가치 실현 시간 단축/ 비용 예측
자사 데이터를 활용한 차별화된 AI 구현 및 비용 예측

https://www.ibm.com/products/watsonx-ai/foundation-models

Use Case에 따라 최적화된 다양한 모델 > Foundation model Libraries

IBM models

IBM Granite Models Series

Large Language Model

granite.13b.instruct decoder only	granite.13b.chat decoder only
Extract	Q&A
Summarize	Generate
Classify	

Code Model

granite.code.ansible decoder only Ansible-tunned model	granite.13b.chat decoder only Cobol2Java-tuned model
CodeGen	CodeGen

IBM Slate Models Series

Encoder only, non-generative models(NLP models)

slate *135 million params* multilingual *Fine-tuned for* ***Entity Extraction***	slate *135 million params* multilingual *Fine-tuned for* ***Relationship Detection***	slate *135 million params* multilingual *Fine-tuned for* ***Sentiment Analysis***	slate *135 million params* multilingual *Fine-tuned for* ***Targeted Sentiment Analysis***
Extract	Extract	Classify	Extract
			Classify

Open Source models

flan-ul2-20b encoder/decoder	gpt-neox-20b decoder only	mt0-xxl-13b encoder/decoder	flan-t5-xxl-11b encoder/decoder	mpt-instrucct2-7b decoder only
Q&A	Q&A	Q&A	Q&A	Q&A
Generate	Generate	Generate	Generate	Generate
Extract		Summarize	Summarize	
Summarize		Classify	Classify	
Classify				

3rd party models

Meta AI

llama2-chat-70b decoder only	llama2-chat-13b decoder only
Q&A	Q&A
Generate	Generate
Extract	Extract
Summarize	Summarize
Classify	Classify

StarCoder

Starcoder-15.5b decoder only
CodeGen

투명한 모델 > building transparency model with Sound Data

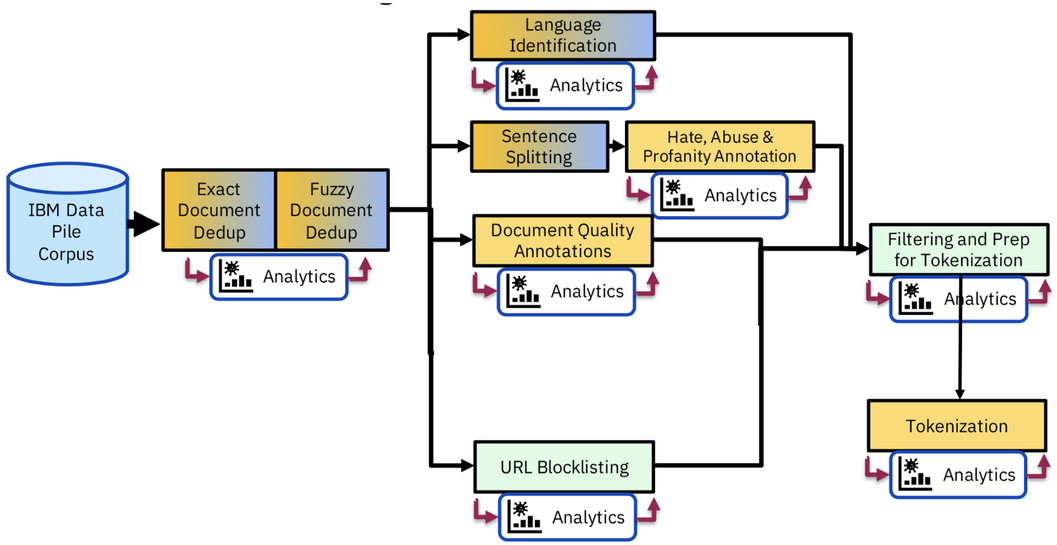
granite.13b 원천 데이터 공개

Today, IBM is sharing the following data sources used in the training of the Granite models (learn more about how these models are trained and data sources used):

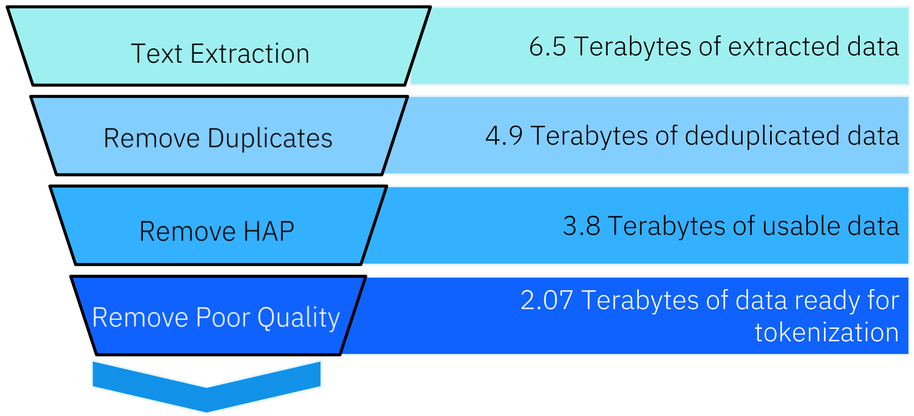
1. Common Crawl
2. Webhose
3. GitHub Clean
4. Arxiv
5. USPTO
6. Pub Med Central
7. SEC Filings
8. Free Law
9. Wikimedia
10. Stack Exchange
11. DeepMind Mathematics
12. Project Gutenberg (PG-19)
13. OpenWeb Text
14. HackerNews

- IBM이 큐레이팅한 엔터프라이즈 중심 데이터 세트로 훈련
- 1 Trillion Tokens: 한번에 4M token 으로 30만번 훈련
- 8k 전체 context 가능 (Llama-2 70b : 4k)
  출력 1k token (Llama-2 70b : 900개)
- 5개의 비즈니스 영역에 특화 됨
  (Q&A, Generate, Extract, Summarize, Classify)

Data 전처리 파이프 라인



사전 교육 데이터 세트에 대한 거버넌스 통계 요약

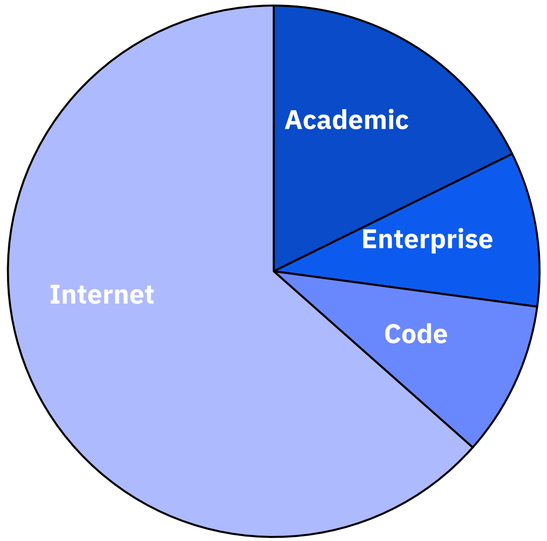


1T Tokens of data for training granite.13b

https://www.ibm.com/blog/watsonx-tailored-generative-ai
https://www.ibm.com/downloads/cas/X9W4O6BM

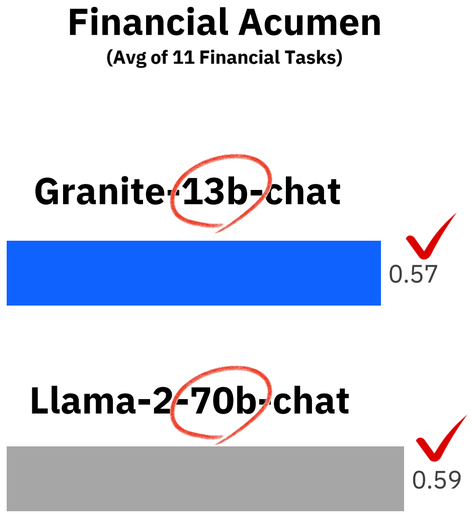
업무 특화된 효율적인 모델 > 기업 업무 특화된 더 작은 모델로 더 빠른 응답 제공

Enterprise



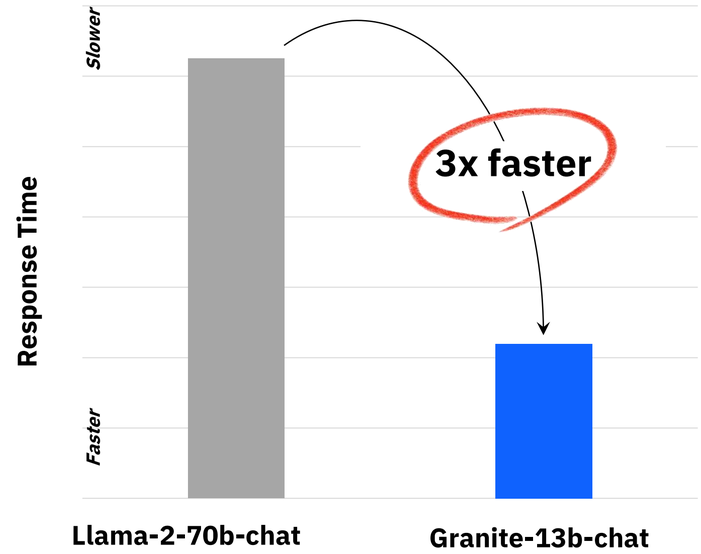
법률 및 재무 데이터를 포함한 기업 업무에 맞춰진 데이터(약 10%)로 훈련됨

Performance



11개의 재무 관련 작업에서의 granite-13b 성능은 훨씬 더 큰 Llama-2-70b-chat과 유사함

Delivered Faster



Granite-13b는 1000 토큰에 대해 Llama-2-70b보다 최대 3배 빠름

우리회사 생성형 AI 구현에

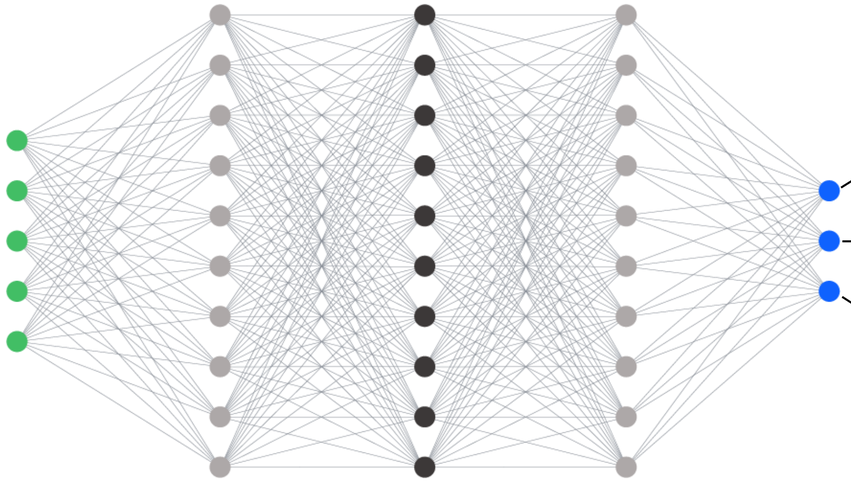
기업 맞춤형 **watsonx 플랫폼**이 필요한 이유

1. **생성형 AI 시대, 기업 동향과 고민은?**
2. **생성형 AI, Large Language Model의 특징**
3. **모델 이야기** : 개발에서 선택으로!
4. **성능 이야기** : 적은 비용으로 최대의 효과를!
5. **운영 이야기** : Trustworthy 생성형 AI 구현!
6. **기업 맞춤형 watsonx 플랫폼**

모델 성능과 적합성을 빠르게 파악하여 상황에 맞는 튜닝 방법을 선택합니다.

새로운 데이터에 대한 **예측 성능을 향상**시키기 위해 사용되며,
모델의 일부 층을 수정하거나 새로운 데이터를 추가하여 수행
수행 결과로 **모델 가중치가 변경된 새로운 모델**이 만들어짐

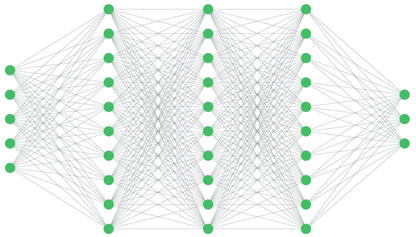
Foundation Model은
기본적으로 일정 수준이상의
성능을 제공합니다.



Foundation model

Fine-tuning (Model level)

Task A



Roadmap item

Prompt-tuning (Prompt level)

prompt tuning 에서는
고객의 **labeled data가 전달됨**

Task B

watsonx.ai
Tuning Studio

prompt tuning 또는
prompt engineering은
모델이 변경되지 않음

Prompt engineering

모델에 입력되는 텍스트에
프롬프트를 포함시켜 모델이
"출력"하는 결과를 세밀하게 제어

Task C

watsonx.ai
Prompt Lab

High

Complexity, Skills, Time, Cost

Low

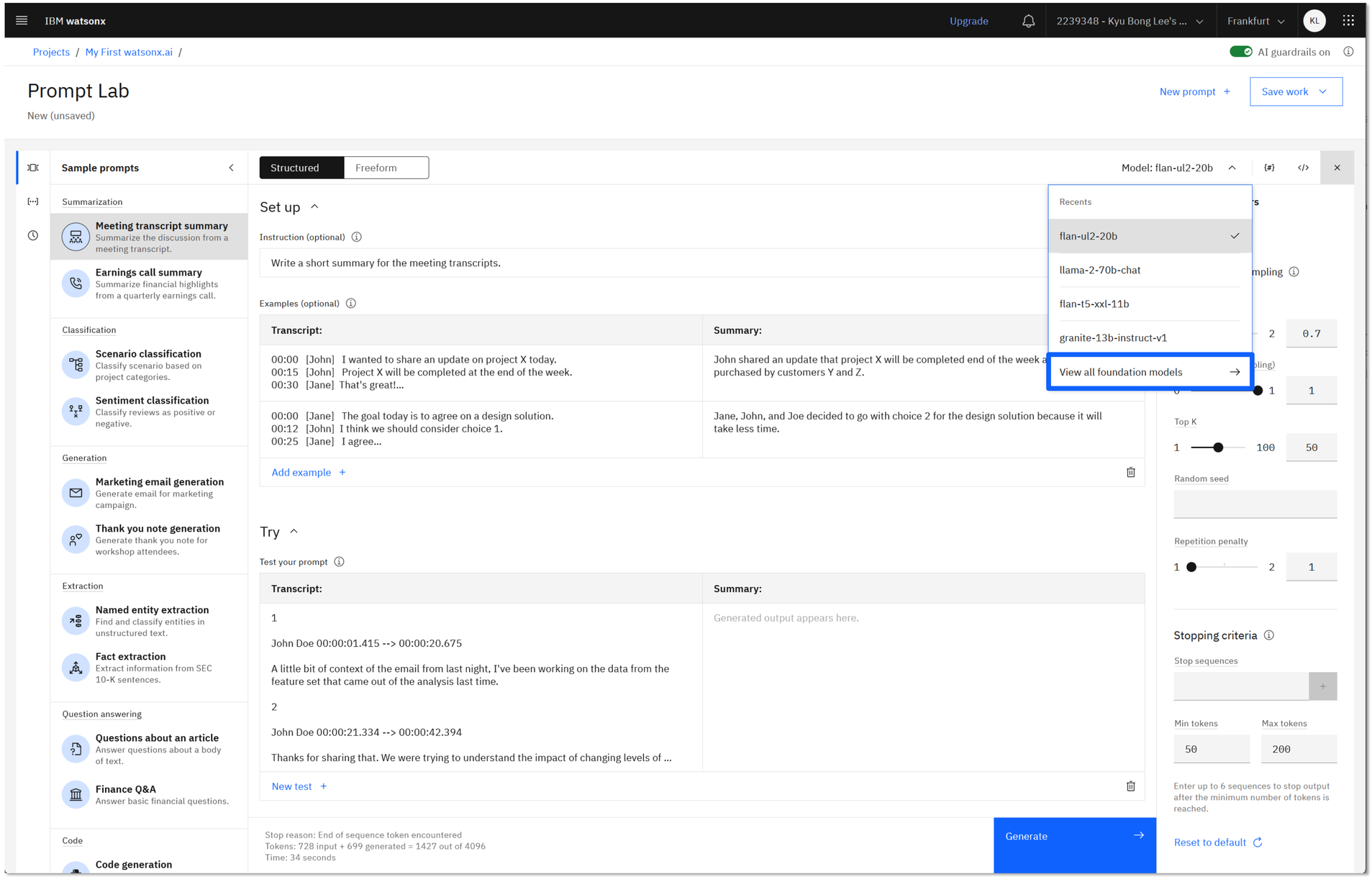
Small

Requires increasing model size

Large

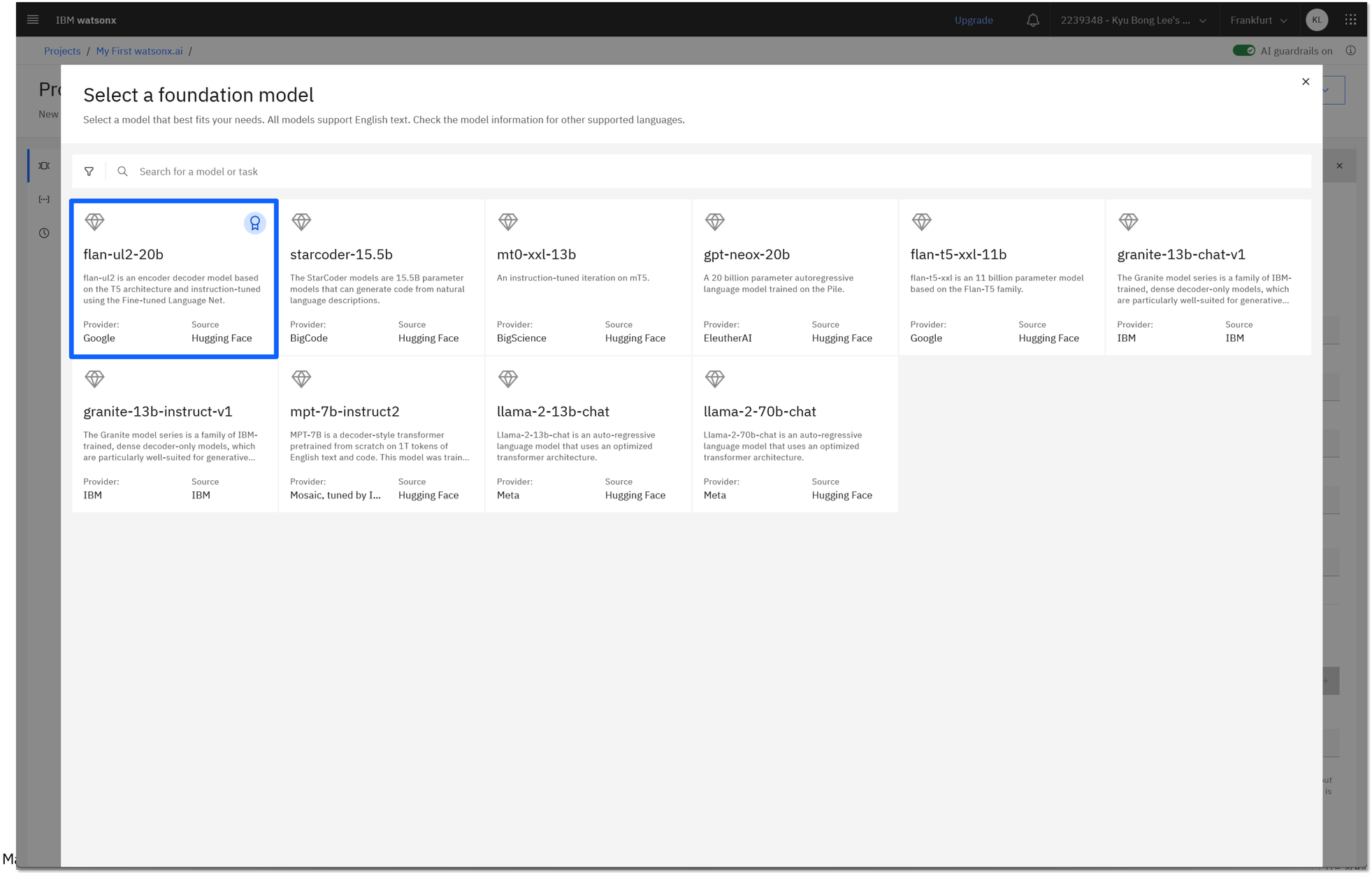
watsonx.ai
Prompt Lab

Task에 따른 Model을 선택합니다.



watsonx.ai
Prompt Lab

Task에 따른 Model을 선택합니다.



watsonx.ai
Prompt Lab

Model 특성을 파악한 후 Model을 선택합니다.

IBM watsonx

Upgrade

2239348 - Kyu Bong Lee's ...

Frankfurt

KL

Projects / My First watsonx.ai /

AI guardrails on

flan-ul2-20b

Provider: Google | Source: Hugging Face

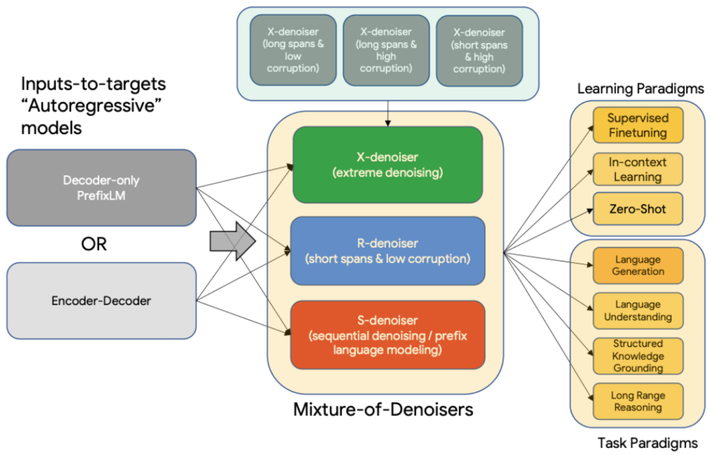
Question answering | Summarization | Retrieval-Augmented Generation | Classification | Generation | Extraction

Note: This model is a Non-IBM Product governed by a third-party license that may impose use restrictions and other obligations. By using this model you agree to these terms. Read terms

Introduction to UL2

This entire section has been copied from the `google/ul2` model card and might be subject of change with respect to `flan-ul2`.

UL2 is a unified framework for pretraining models that are universally effective across datasets and setups. UL2 uses Mixture-of-Denoisers (MoD), apre-training objective that combines diverse pre-training paradigms together. UL2 introduces a notion of mode switching, wherein downstream fine-tuning is associated with specific pre-training schemes.



Abstract

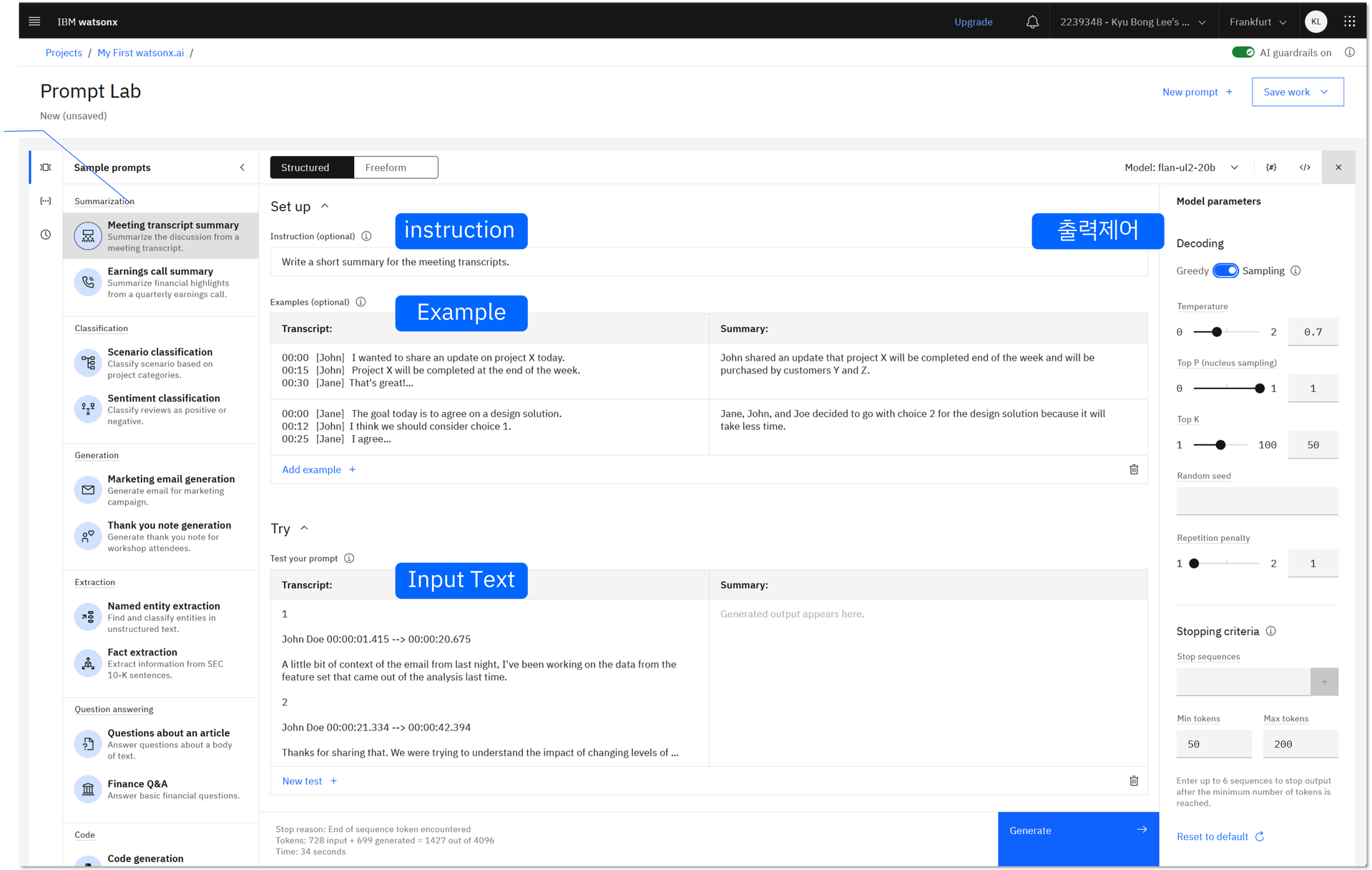
Existing pre-trained models are generally geared towards a particular class of problems. To date, there seems to be still no consensus on what the right architecture and pre-training setup should be. This paper presents a unified framework for pre-training models that are universally effective across datasets and setups. We begin by disentangling architectural archetypes with pre-training objectives -- two concepts that are commonly conflated. Next, we present a generalized and unified perspective for self-supervision in NLP and show how different pre-training objectives can be cast as one another and how interpolating between different objectives can be effective. We then propose Mixture-of-Denoisers (MoD), a pre-training objective that combines diverse pre-training paradigms together. We furthermore

Back | Select model

watsonx.ai
Prompt Lab

선택된 모델의 적합성을 빠르게 판단합니다.

- Task 별 프롬프트 예시
- 프롬프트 저장
- 프롬프트 히스토리

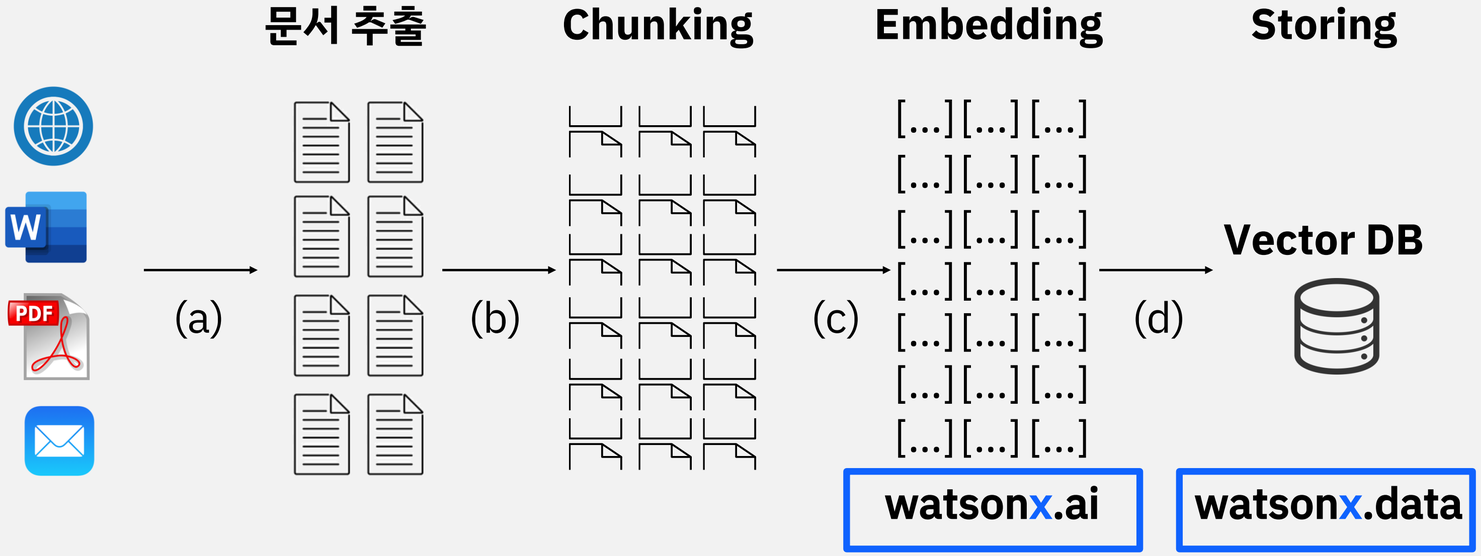


RAG(Retrieval Augmented Generation)

RAG 방식을 통해 기업 내부의 최신 전문 지식을 기반으로 Large Language Model을 활용할 수 있습니다.

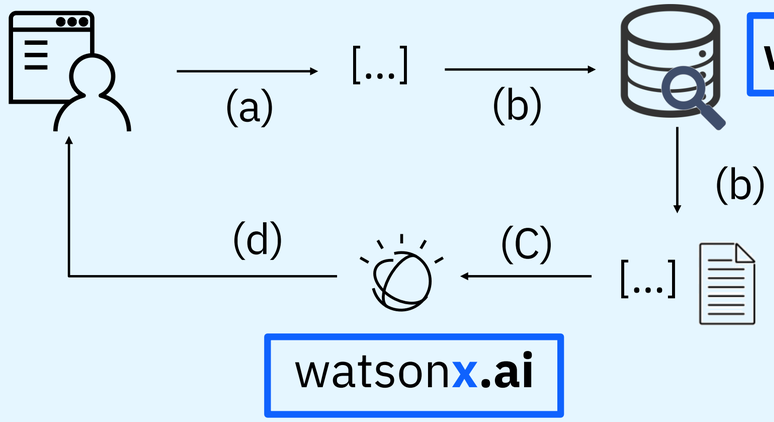
Phase 1 자체 데이터 수집 단계

(a) 원본 파일을 문서로 추출
(b) 문서를 chuck 단위로 분류
(c) Chuck들을 임베딩
(d) 임베딩을 Vector Store에 저장



Phase 2 문서 검색 및 응답 단계

(a) 검색어를 임베딩 변환 후 쿼리
(b) 가장 일치하는 chuck 얻기
(c) Prompt에 검색된 결과 추가
(d) LLM의 결과를 답변



RAG 성능 향상을 위한 실험 단계별 참조사항

1. 문서 로딩/사전 작업
2. 청크 크기
3. 임베딩 모델의 선택
4. database로 부터 리턴 받는 청크의 크기
5. LLM 선택
6. 프롬프트
7. LLM 파라미터 (temperature, Top-K, Top-P 등)

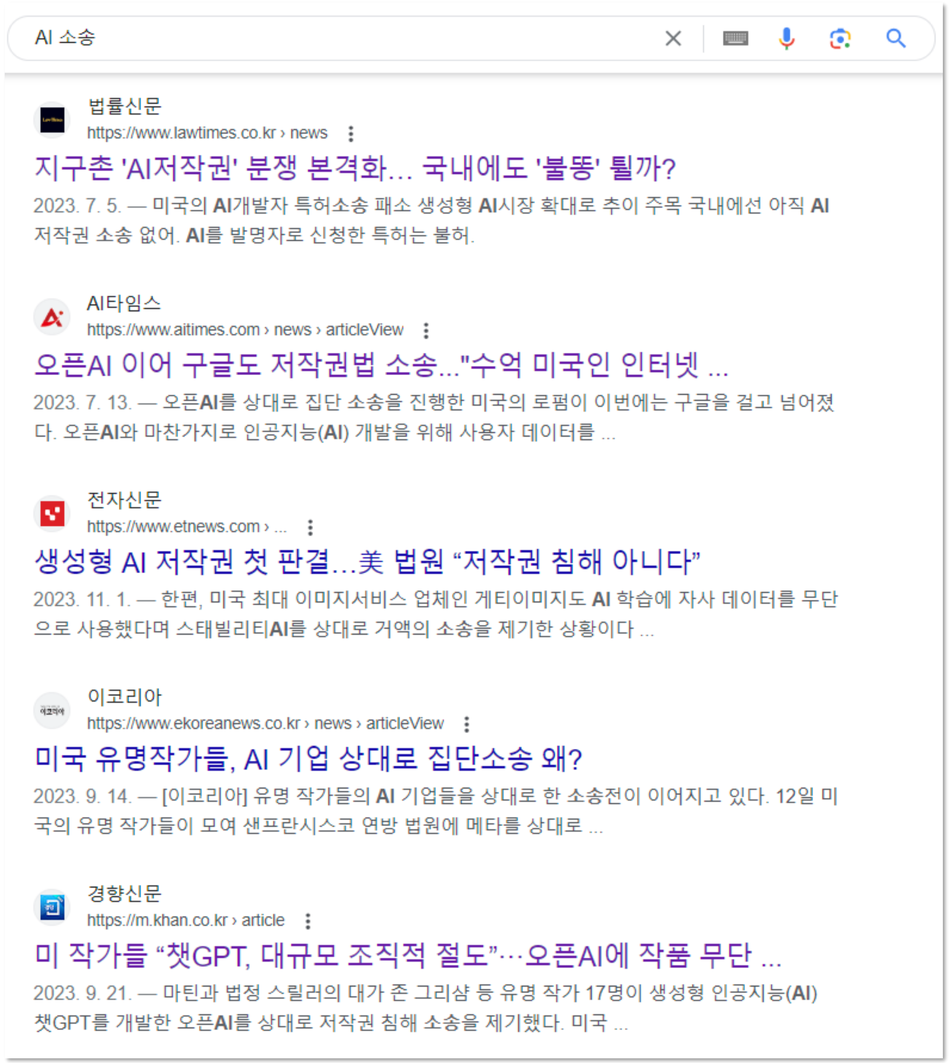
우리회사 생성형 AI 구현에

기업 맞춤형 **watsonx 플랫폼**이 필요한 이유

1. **생성형 AI 시대, 기업 동향과 고민은?**
2. **생성형 AI, Large Language Model의 특징**
3. **모델 이야기** : 개발에서 선택으로!
4. **성능 이야기** : 적은 비용으로 최대의 효과를!
5. **운영 이야기** : Trustworthy 생성형 AI 구현!
6. **기업 맞춤형 watsonx 플랫폼**

AI Risk

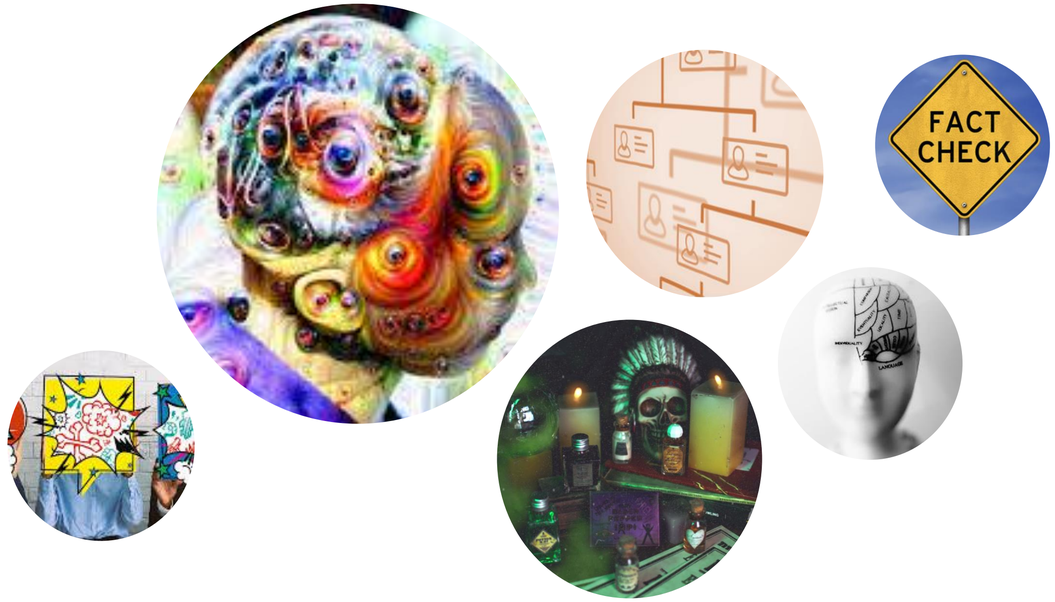
생성형 AI는 새로운 위험에 대한 노출을 야기합니다. 기업의 생성형 AI는 강력한 거버넌스와 규정 준수가 필수적입니다.



Q : What is the likely risk of the following, resulting from adoption of generative AI to your organization?

2023 IBM IBV generative AI Impact on Hybrid cloud pulse survey

생성형 AI에는 기존 AI 위험 요소에 더하여 새로운 위험 요소들이 있습니다.



생성형 AI 시대에도 **전통적인 AI/ML 영역의 위험 요소**들은 여전히 남아 있습니다.

- poor predictive accuracy
- lack of fairness and equity
- lack of explainability
- model uncertainty
- distribution shifts
- poisoning attacks
- evasion attacks
- extraction attacks
- inference attacks
- model transparency

이에 더해 생성형 AI에는 **완전히 새로운 위험 요소들**이 등장하였습니다.

- hallucinations
- lack of factuality or faithfulness
- lack of source attribution
- toxicity, profanities, and hate speech
- bullying and gaslighting
- inability to reason
- privacy leakage
- prompt injection attacks
- misinformation

Foundation models: Opportunities, risks and mitigations

AI 서비스 전과정에 거버넌스가 적용되는 Trusted AI Lifecycle 을 제안합니다

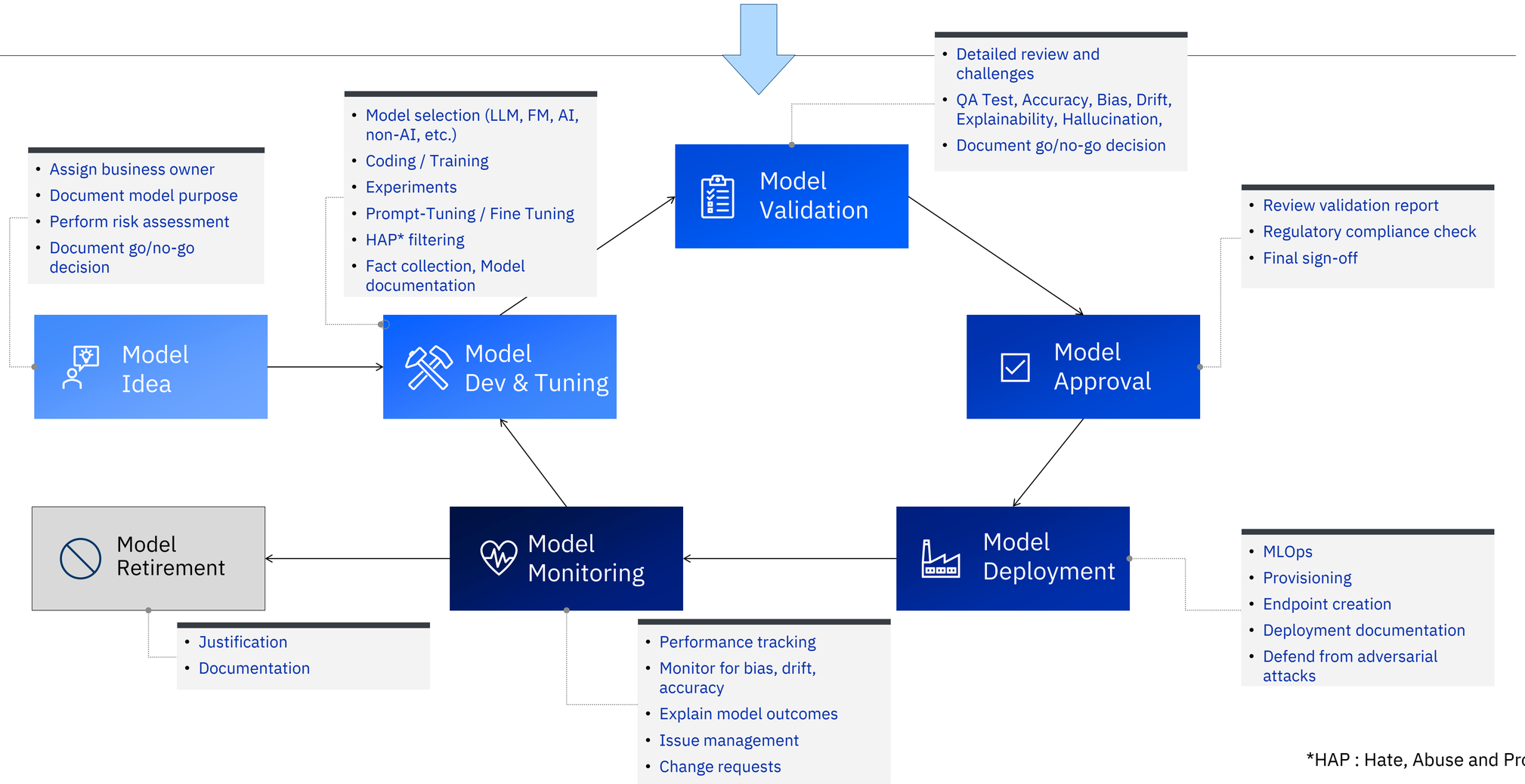
Macro level

Define legal obligations
Define relevant company policies
Extend enterprise risk framework

Articulate AI principles
AI Governance board & escalation process
Define roles & responsibilities

Awareness and skills
Audits
...

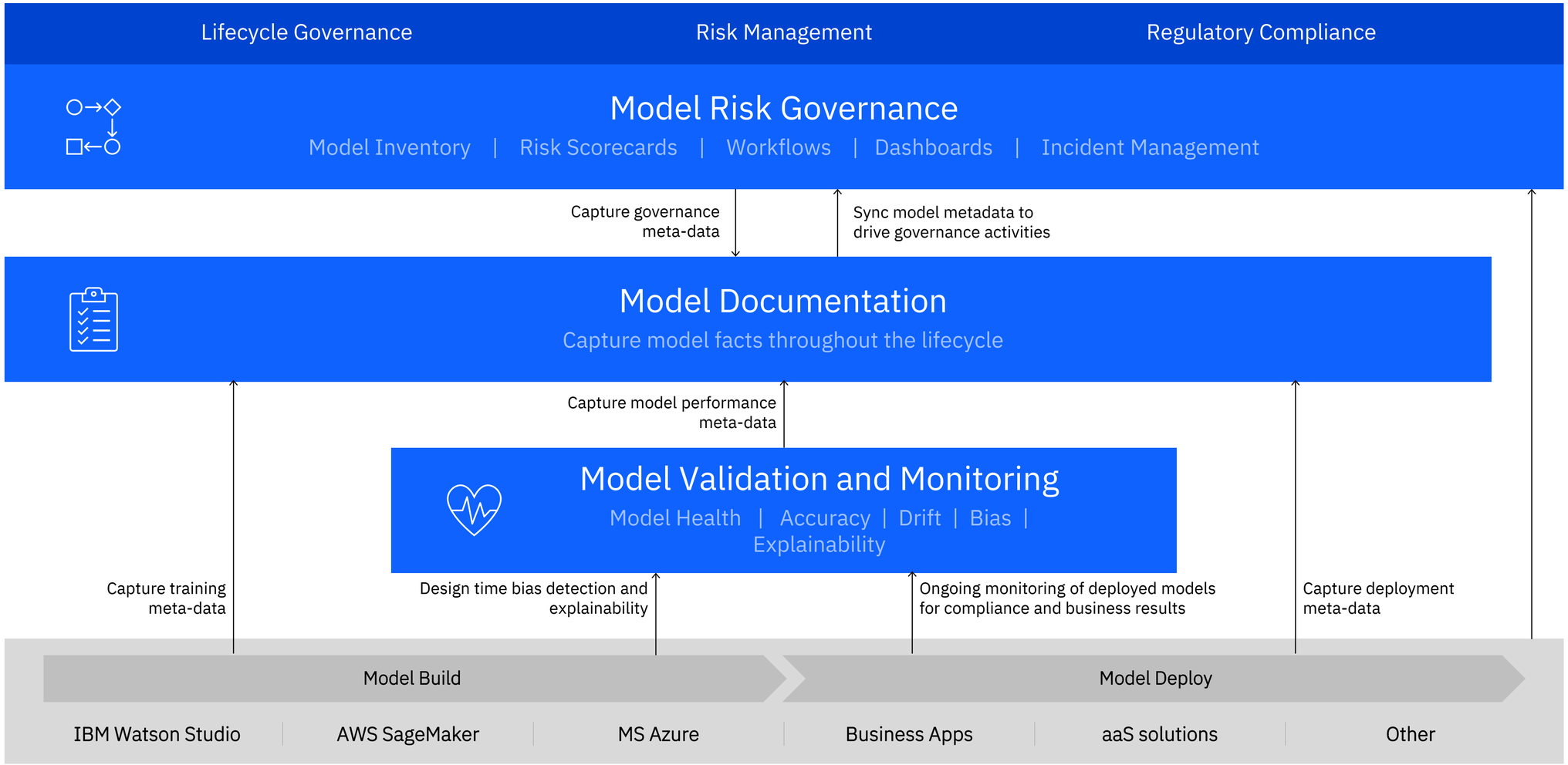
Micro level



*HAP : Hate, Abuse and Profanity

서비스 플랫폼의 제한을 두지 않고 IBM AI Governance를 제공합니다.

- Model Owners
- Model Validators
- Audit Teams
- Compliance Teams
- Risk Management Teams
- Data Privacy Teams
- Principal Data Scientists



- Data Engineers
- (Citizen) Data Scientists
- MLOps
- ML Engineer

우리회사 생성형 AI 구현에

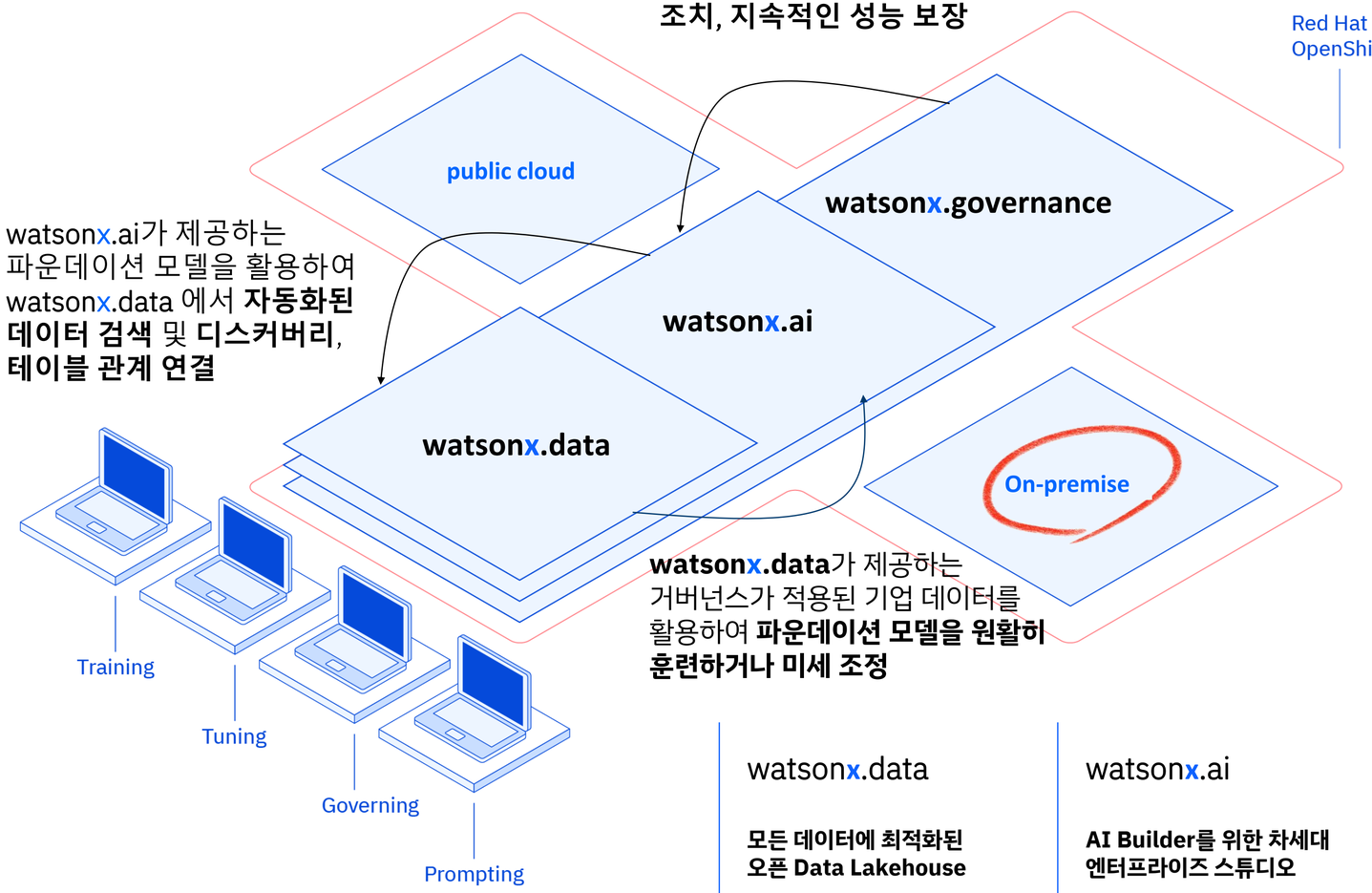
기업 맞춤형 watsonx **플랫폼**이 필요한 이유

1. **생성형 AI 시대, 기업 동향과 고민은?**
2. **생성형 AI, Large Language Model의 특징**
3. **모델 이야기** : 개발에서 선택으로!
4. **성능 이야기** : 적은 비용으로 최대의 효과를!
5. **운영 이야기** : Trustworthy 생성형 AI 구현!
6. **기업 맞춤형 watsonx 플랫폼**

기업 맞춤형 AI 플랫폼 watsonx

AI와 Data를 위한 플랫폼

신뢰할 수 있는 데이터로 AI의 영향력을 확장하고 가속화합니다.



watsonx.data

모든 데이터에 최적화된 오픈 Data Lakehouse

- 데이터 접근 및 공유를 위한 federated query
- 데이터 거버넌스 제공
- 자연어 기반 데이터 탐색, 보강 및 강화

watsonx.ai

AI Builder를 위한 차세대 엔터프라이즈 스튜디오

- 기존 머신 러닝 모델 + foundation 모델 모두 훈련, 검증 및 배포
- 짧은 시간에 AI 애플리케이션을 구축

watsonx.governance

책임감 있고 투명하며 설명 가능한 AI 워크플로 구현

- AI 라이프사이클 거버넌스 관리 및 모니터링
- 위험 및 규정 준수
- 이해 관계자 가시성 향상 및 협업 촉진

IBM이 제공해 드리는 생성형 AI 기술 및 전문성

AI assistants	Empower individuals to do work without expert knowledge across a variety of business processes and applications.	**watsonx** Code Assistant **watsonx** Assistant **watsonx** Orchestrate **watsonx** Orders
SDKs & APIs	Embed watsonx platform in third party assistants and applications using programmatic interfaces.	**Ecosystem integrations**
AI & data platform	Leverage generative AI and machine learning — tuned with your data — with responsibility, transparency and explainability.	**watsonx** watsonx.ai watsonx.governance watsonx.data **Foundation models** Granite | *IBM* Open Source | *Hugging Face* Llama 2 | *Meta AI* Geospatial | *IBM + NASA* ...
Data services	Define, organize, manage, and deliver trusted data to train and tune AI models with data fabric services.	**Data fabric services** **watsonx** Discovery
Hybrid cloud AI tools	Build on a consistent, scalable foundation based on open-source technology.	**Red Hat** OpenShift AI (*e.g.,* Ray, Pytorch)

Consulting
Generative AI strategy, experience, technology, operations

Ecosystem
System Integrators, Software and SaaS partners,
Public Cloud providers

IBM이 제안하는 watsonx Pilot 프로그램

Solutions Workshops

가치와 파일럿 범위 정의

2-8 Hours

결과물

- 유스케이스 식별 및 조정
- Data와 솔루션 탐색
- 범위와 성공 기준 정의
- **watsonx** 핸즈온 경험
- 파일럿 계획 개발 및 조정

방법 및 내용 구성

- 상황에 맞는 워크숍 진행
  - 유스케이스 식별 및 개선
  - 필요한 데이터 및 소스를 포함한 기술 솔루션파악 및 정의
  - 파일럿 범위와 성공 기준 정의
- watsonx 핸즈온 lab

Pilot Build

watsonx 기반 “Co-Creation”을 통한 가치 증명

2 to 4 weeks

결과물

- 성공 기준에 맞는 watsonx 코드 및 모델
- 파일럿을 위한 UX(사용자 경험) 설계
- 추가적인 IBM 제품 코드 (파일럿 범위에 포함된 경우)
- UI(사용자 인터페이스) 및 연계 개발 (파일럿 범위에 포함된 경우)
- 비즈니스 케이스 수립을 위한 input 제공
- 검증 포인트 및 학습 내용 캡쳐
- 파일럿 프로젝트 요약과 데모 비디오

방법 및 내용 구성

- 개발 및 실행 환경 구성
- 데이터 식별, 소싱 및 준비 작업
- 검증 계획 정의
- 반복적으로 모델 선택, 조정 및 검증
- 비즈니스 케이스 input 개발
- 스폰서 및 이해 관계자들과 정기적 playback 수행

Transition

적용을 위한 계획 수립

결과물

- 고객에게 지식 이전
- 파일럿 유스케이스 이후 다음 단계 계획 수립
- High level Generative AI 도입 계획
- Generative AI에 대한 서비스 요구 사항을 충족하기 위한 IBM의 제안
- 추가적 유스케이스 식별

방법 및 내용 구성

- 지식 이전 세션
- 로드맵과 플래닝 워크숍

기업 가치를 준수하며 효용을 극대화하는 기업 고객을 위한 생성형 AI 플랫폼

IBM PoV : 4 core principles to tailor generative AI for enterprise

Open	IBM's AI is based on the best open technologies available
Trusted	IBM's AI is transparent, responsible, and governed
Targeted	IBM's AI is designed for enterprise and targeted at business domains
Empowering	IBM's AI enables users to become a value creator with full ownership of data and AI models

watsonx.ai
watsonx.data
watsonx.governance