CLOVA Studio v1.16

HyperCLOVA X

HyperCLOVA X를 활용한 서비스 혁신

NAVER Cloud

향후 글로벌 IT 산업의 생태계는 생성형 AI의 영향력을 중심으로 재편성될 것입니다

IDC의 FutureScape 2024 | IDC가 선정한 향후 1-2년 전 세계 10대 IT 산업 예측

핵심 IT 전환
핵심적인 IT 지출이 AI 중심으로 급격히 변화할 것. 2025까지 글로벌 기업은 핵심 IT 지출의 40% 이상을 AI 관련 이니셔티브에 할당하여 혁신의 속도를 높일 것

IT 산업의 AI 중심
IT 산업에서는 AI 제품/서비스를 도입하고, 고객의 AI 구축을 지원하기 위해 경쟁하며, AI의 영향력을 크게 실감할 것

양질 데이터 확보
AI 모델의 학습과 적용을 위해 데이터는 매우 중요한 자산이 될 것. 기술 공급자와 서비스 제공자는 데이터에 대한 투자를 가속하여 경쟁 우위를 점하려 할 것으로 예상

서비스 산업 변화
생성형 AI는 현재 인간이 제공하는 전략, 변혁, 학습 서비스의 전환을 촉발할 것, 2025년까지 서비스 계약의 40%에 생성 AI 제공이 포함될 것으로 예상

통합 AI
기업은 새로운 사용 사례와 고객을 낮은 가격에 처리할 수 있는 통합 AI 솔루션을 고민하여 상품화해야 할 것

AI 경험 확대
생성 AI가 상용화되면 기업은 고객 기대에 더 잘 부합하는 상황별 경험을 통해 엣지 컴퓨팅 활용 사례를 향상할 수 있을 것

⋮

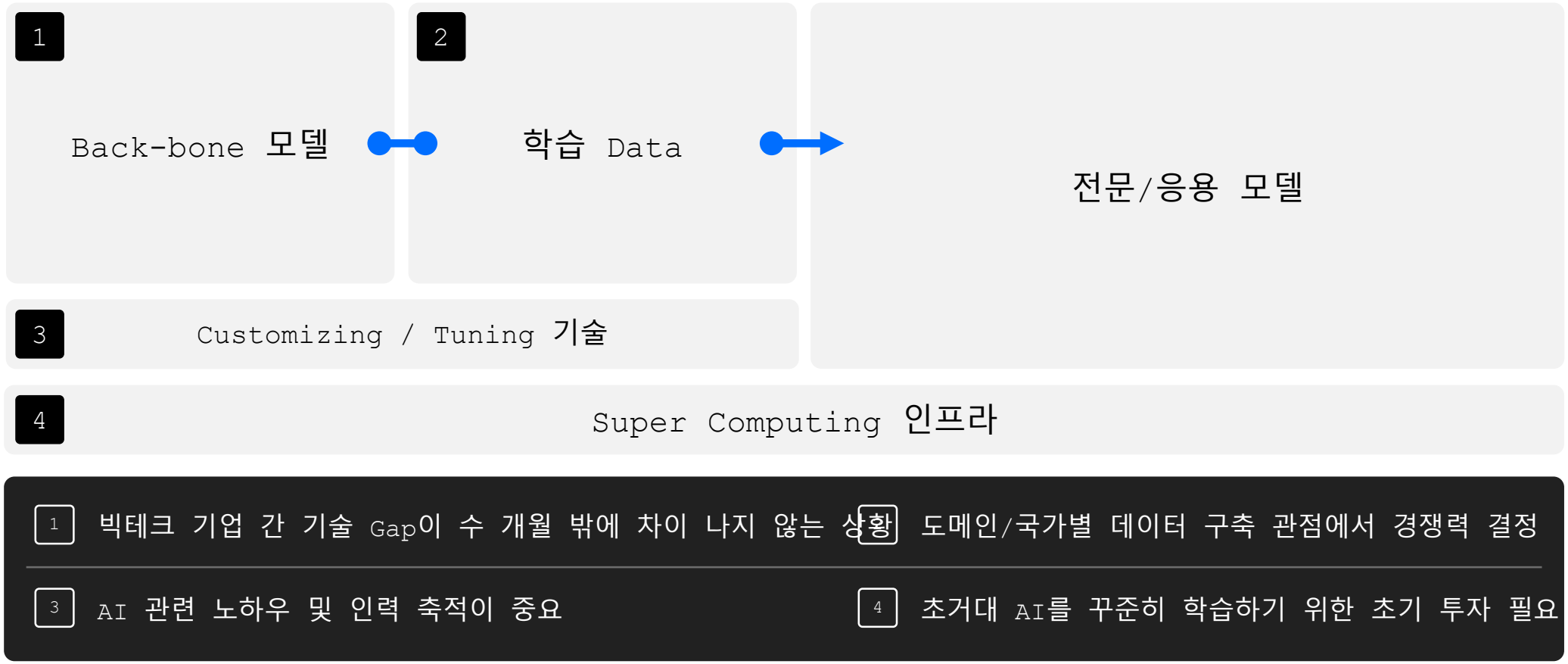
AI 솔루션에 대한 글로벌 지출 예상치 (~2027)
5,000억 달러 이상

출처: IDC FutureScape: Worldwide IT Industry 2024 Predictions

Why HyperCLOVA X

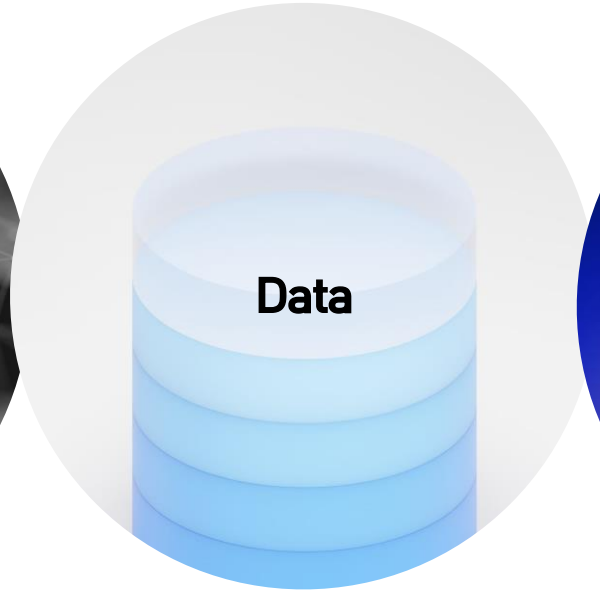
1. 한국어 생성 효율성
2. 최적 서빙 플랫폼(CLOps)
3. 소형언어모델(sLLM) 리스크
4. 데이터 보안
5. 책임감 있는 AI

생성형 AI 영역에서 필요한 경쟁력은 AI Back-bone Model, Data, 튜닝 기술 그리고 인프라입니다



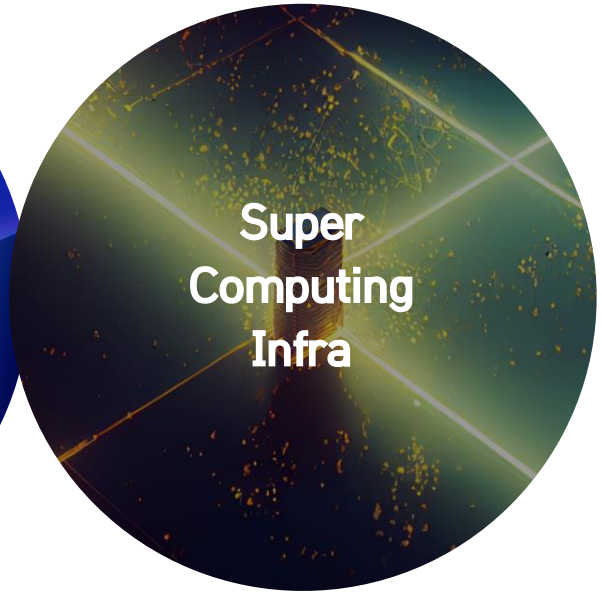
직접 개발하여 원천기술을 보유한 슈퍼 컴퓨팅 인프라부터 백본모델까지, 네이버클라우드는 생성형 AI의 모든 요소에 준비되어 있습니다

GPT-3.5 이상의
Back-bone Model /
300명 이상의 AI 전문가 보유



글로벌 빅테크 대비
한국어 데이터 우수성

다년간의 LLM 서비스 운영 경험 /
노하우와 인력



수년간의 시행착오를 통해
확보한 안정된 학습환경

HyperCLOVA X는 한국의 문화와 맥락을 가장 잘 이해하는 모델로 AI 생태계와의 연결성 및 확장성을 기반으로 새로운 사용자 경험과 비즈니스 기회를 만들어 갑니다

연결성

네이버 서비스와의 연결

확장성

외부 생태계로의 확장

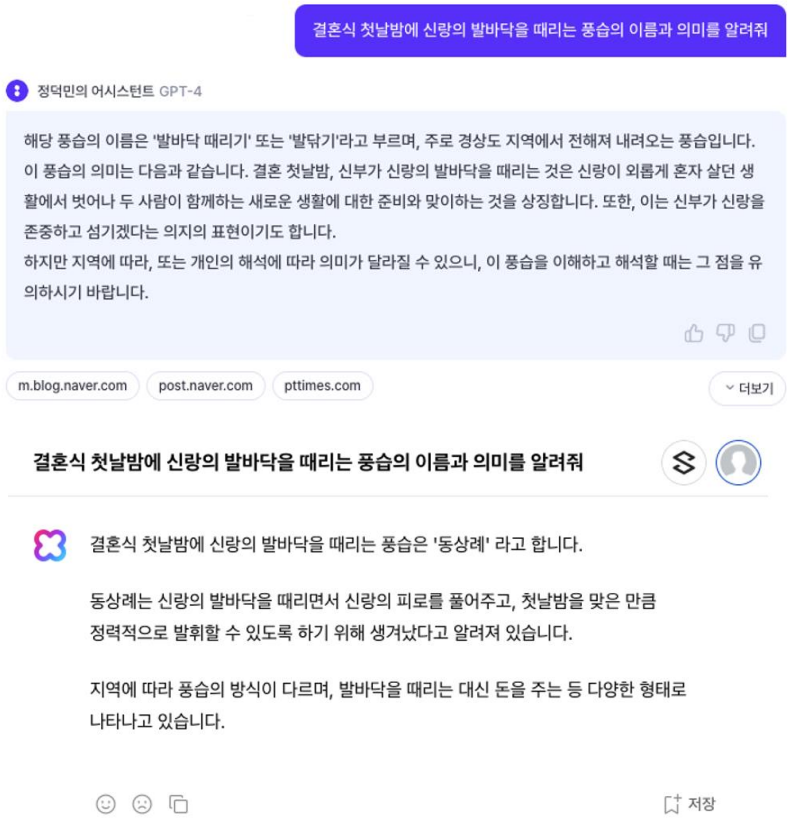
최적화

고객 요구에 맞게 최적화

[참고] 한국의 문화와 맥락을 잘 이해하는 HyperCLOVA X

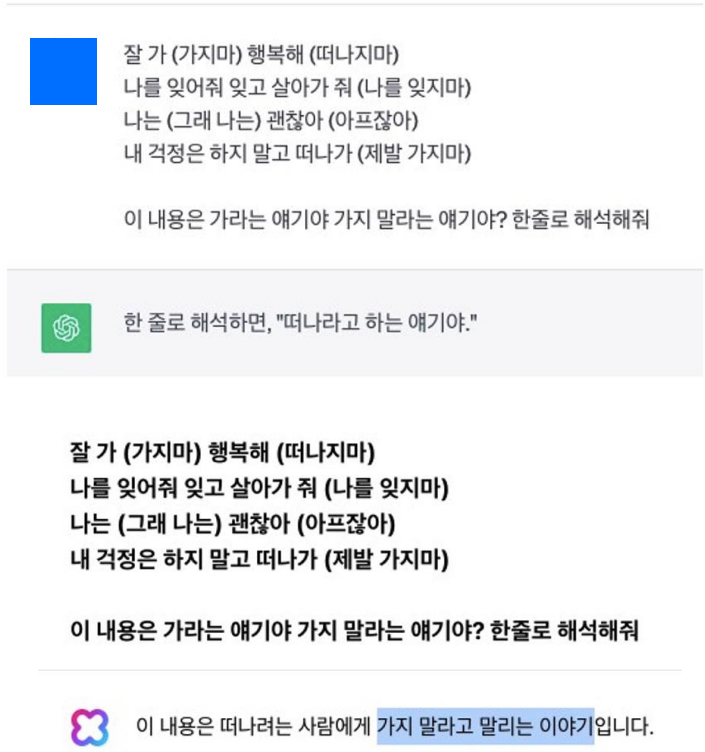
한국의 문화

Q. 결혼식 첫날밤에 신랑의 발바닥을 때리는 풍습의 이름과 의미를 알려줘



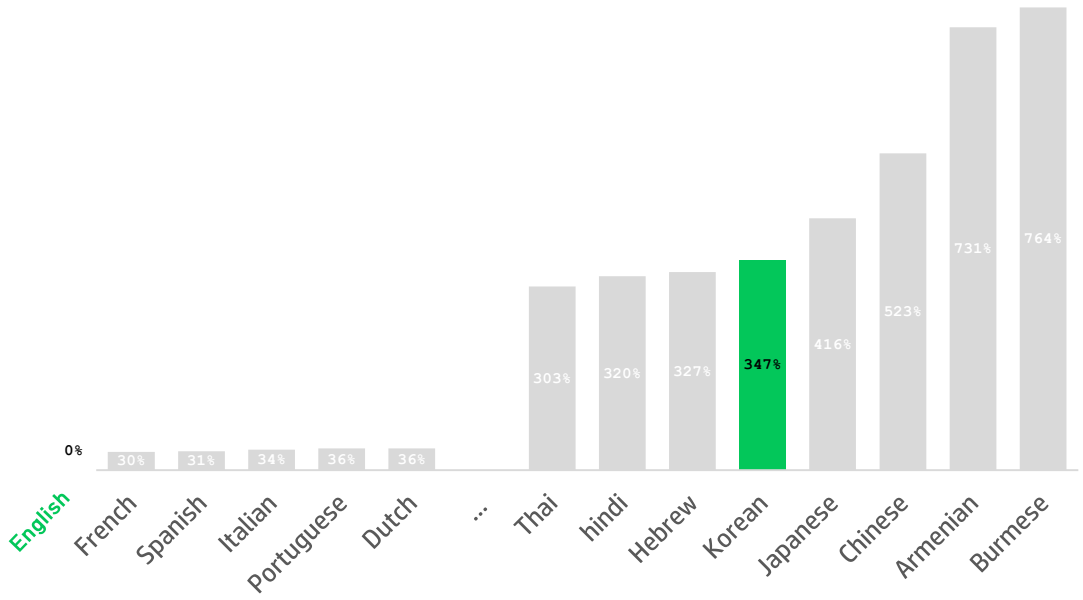
한국어 맥락

Q. 지오디의 '거짓말' 가사는 가라는 얘기야 가지 말라는 얘기야? 한 줄로 해석해줘



HyperCLOVA X는 한국어에 강점을 지닌 모델로 비용 및 시간 관점에서 경쟁사 대비 약 3.5배 높은 효율을 보입니다

GPT-4 모델의 언어별 overhead cost



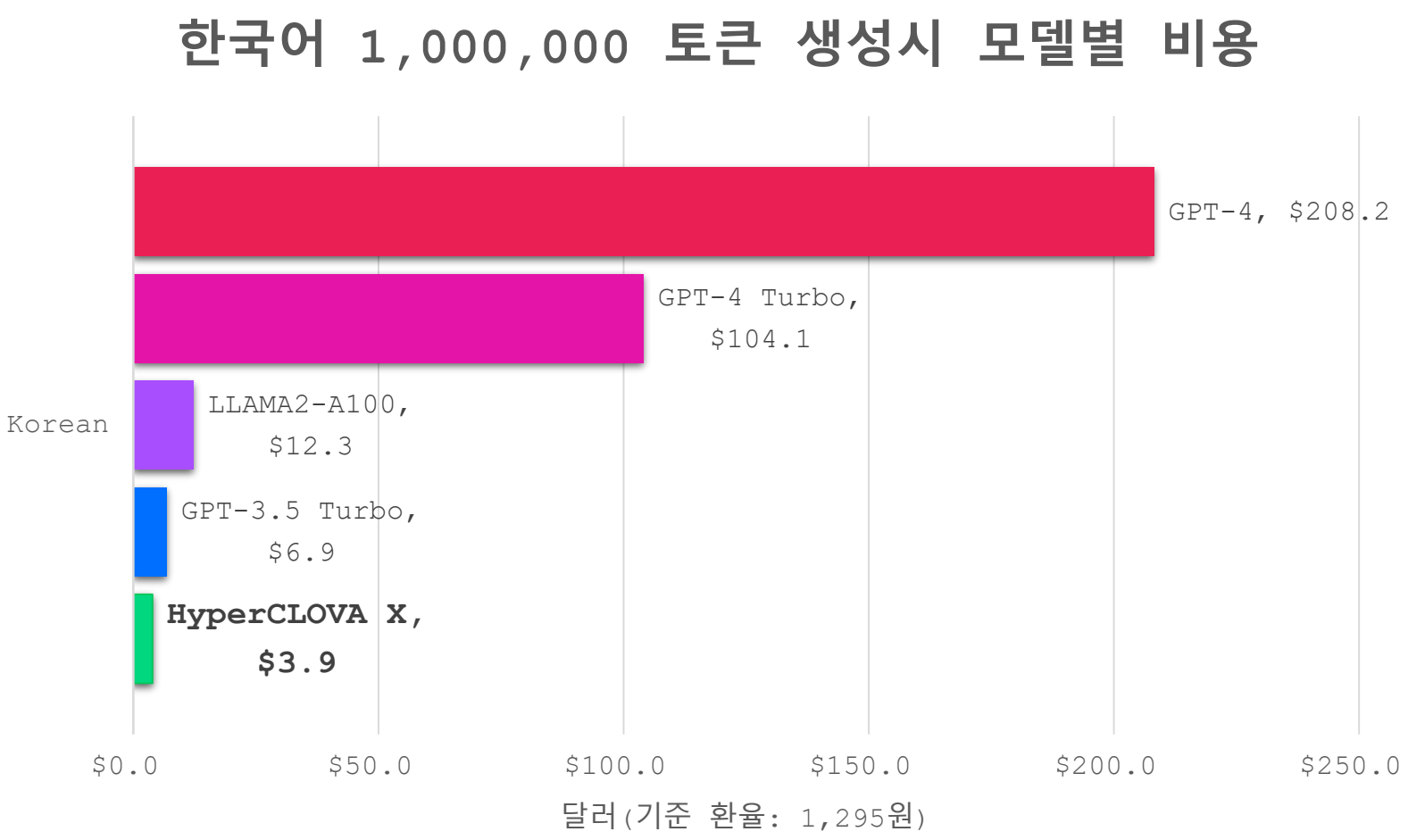
- 언어 모델의 토크나이저는 주로 학습된 언어의 영향을 크게 받으며, 영어가 주요 학습 데이터를 구성하는 경쟁사 모델의 경우 그 외 언어에 대한 Overhead cost 발생
- 영어와의 유사성에 따라 적게는 30~700% 이상의 비효율이 발생하며, 한국어의 경우 토큰 소비에 있어 347%, 약 3.5배 수준의 비효율이 발생하는 것으로 추정

출처: https://tomaszurbanski.substack.com/p/the-hidden-price-tag-on-gpt-4-fort

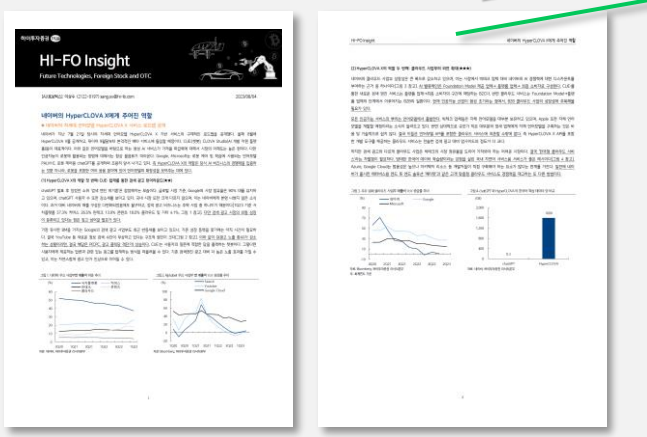
토크나이저 비교

- 한국어 기준, GPT-3와 HyperCLOVA 토크나이저에 동일한 문장을 입력하여 계산할 경우 HyperCLOVA는 GPT-3 소비량의 약 28.8% 토큰으로 생성하는 것을 확인
- 토큰 단위의 종량제로 과금되는 경우의 비용 뿐만 아니라, 모델의 Context Window 사이즈에서도 약 3.5배 수준의 차이 발생

예를 들어, 리포트 약 555건에 해당하는 한국어를 생성할 경우 다른 언어모델(LLM) 대비 HyperCLOVA X 비용이 가장 저렴합니다



증권사 리포트 Example



- 산정기준: 산업 동향 리포트 1건, 5페이지 분량
  → 약 1,800토큰
- 리포트 **약 555건**, 2,775페이지 분량
  → **약 1,000,000 토큰**

1) 한국어 생성하는 경우 GPT는 영어 생성 보다 347% 더 많이 사용하는 것을 반영한 가격

사용자가 증가하고 모델 개발과 서비스 출시의 라이프사이클 주기가 짧아지면서 오픈소스 MLOps에서 여러 가지 이슈가 나왔습니다

MLOps 운영 이슈

- 모든 Pod가 같은 노드에 배포되면 해당 노드가 중단되는 SPoF[1] 발생 가능성 존재
- 부적절한 metric에 기반해 AutoScaling을 설정하면 불필요 자원을 할당 받아 낭비 발생
- 필요에 따라 여러 metric 서버를 개발하더라도 실제로는 하나만 구동할 수 있어서, 여러 애플리케이션의 다양한 metric 요구를 동시에 수용할 수 없음
- 제로 스케일링(zero scaling)을 지원하지 않음
- 노드와 모델 수가 증가하면서 수집하는 데이터가 많아져 저장소 용량 부족 및 쿼리 실행시 메모리 부족 가능성 존재
- stdout과 stderr 명령으로 로그를 확인하거나 모델의 sidecar를 붙여 별도로 로그를 수집
- 노드 개수가 수백개가 되면서 라벨 관리 및 노드 용도를 변경하기 어려움
- LLM 모델은 요청에 따라 응답 시간 차이가 크기 때문에 라운드로빈 방식의 부하 분산 효과가 적음
- 튜닝 모델의 변경이 잦으면 요청 효율성 저하
- 단일 요청큐로는 트래픽이 떨어지는 시간대에 남는 자원을 배치성 서비스에 사용 불가
- 함께 배치 처리되는 요청 중 가장 오래 걸리는 요청에 따라 다 같이 느려짐

개선 방안

- ✓ *중앙 지표 저장소를 Prometheus에서 VictoriaMetrics로 변경*
- ✓ *모델당 최소 3개의 노드에 분산될 수 있도록 스케줄링*
- ✓ *구현체에 대한 하위 호환성을 유지하는 Auto-Scaling 개발*
- ✓ *노드의 생성, 변경 그리고 삭제에 따라 이벤트를 처리하는 컴포넌트를 개발*
- ✓ *큐잉된 요청 중 같은 튜닝 모델에 대한 요청을 최대한 모아 처리*
- ✓ *요청을 큐에 쌓아 둔 뒤, 개별 모델 서버가 처리 가능 시점에 요청을 가져가 처리*
- ✓ *각 개별 요청을 배치 처리하는 Continuous Batching 기술 도입*
- ✓ *사내 로깅 플랫폼 중 Nelo를 이용*
- ✓ *남는 자원을 배치성 서비스에서 활용하는 Priority Queue 개발*

1) Single Point of Failure: 시스템 구성 요소 중에서, 동작하지 않으면 전체 시스템이 중단되는 요소

네이버의 운영·관리 노하우를 바탕으로 HyperCLOVA X와 같은 대규모 모델을 효율적, 안정적으로 서비스할 수 있는 CLOps(CLOVA + MLOps)를 개발하였습니다

주요 기능

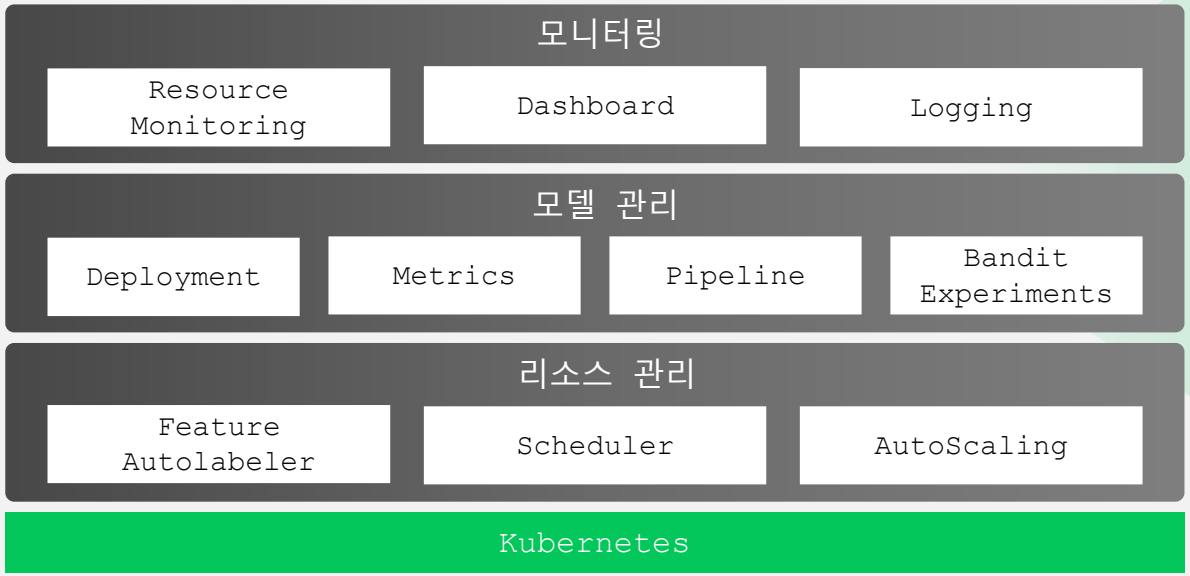
- 유연하게 자원을 활용할 수 있는 오토스케일링 제공
- 클러스터 노드 라벨링 자동화
- 효율적인 GPU 리소스 스케줄링 정책 적용
- 배포된 모델 헬스 체크 및 SPoF[1] 문제 관리

확장성

(배포) 용이성

안정성

CLOps 아키텍쳐



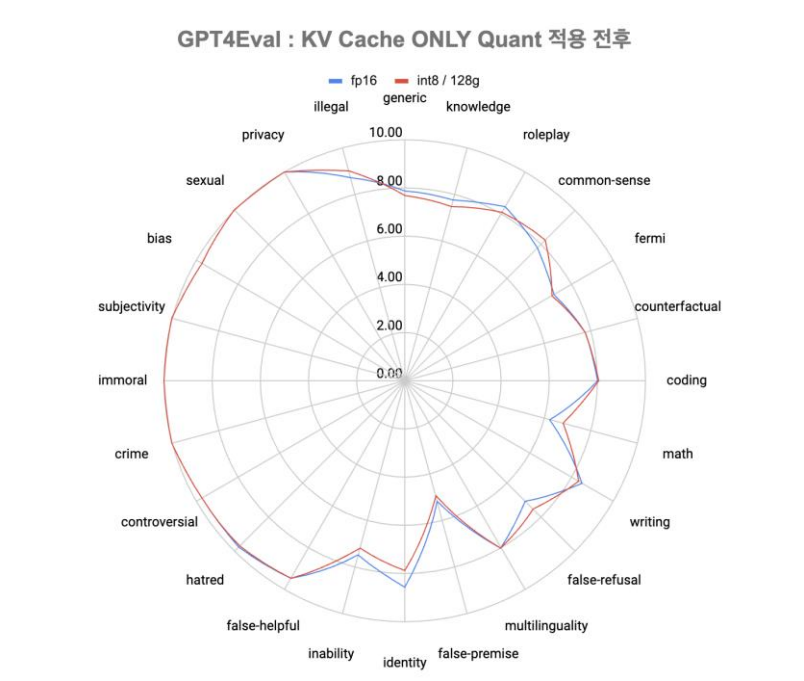
- Resource Monitoring: 배포된 모델에 대한 헬스 체크 수행 및 Metrics 데이터 시각화
- Dashboard: 모델 현황을 한눈에 볼 수 있는 대쉬보드
- Logging: 컨테이너 로그 수집 및 관리

- Deployment: 비동기 추론 요청을 안정적으로 관리 및 모델 배포
- Metrics: 노드나 모델이 많아질 경우 발생할 수 있는 SPoF(Single Point of Failure) 문제를 관리
- Pipeline: GPU를 사용하여 pipeline을 구성
- ~~Bandit Experiments:~~ 여러 모델 버전에 대한 성능 비교 테스트

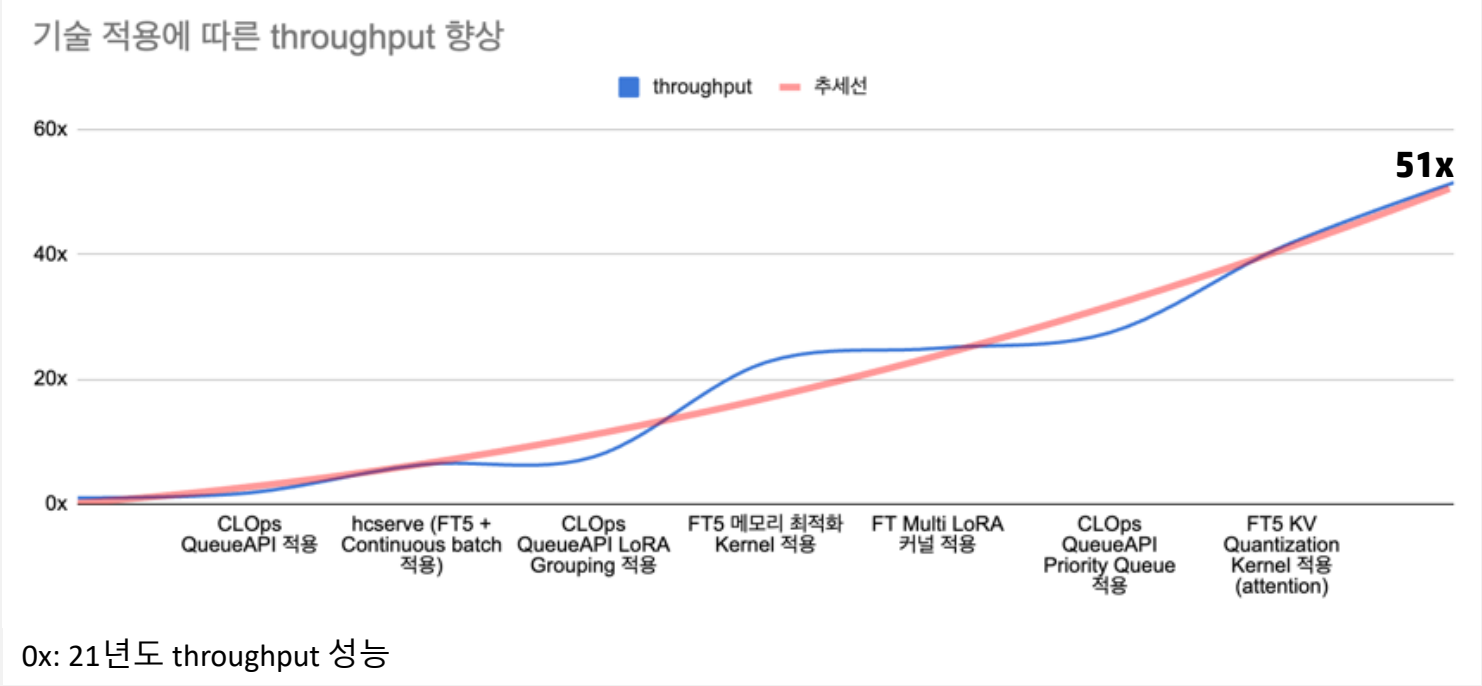
- Feature Autolabeler: 수백 대의 노드들에 대해 서버 정보 기반으로 라벨링 자동화
- Scheduler: 특정 노드의 장애가 모델의 장애로 이어지지 않도록 스케줄 관리
- AutoScaling: 낭비되는 리소스 줄이고 증가하는 트래픽을 관리하는 Scaling 자동화

1) Single Point of Failure: 시스템 구성 요소 중에서, 동작하지 않으면 전체 시스템이 중단되는 요소

네이버클라우드는 CLOps 기능을 지속적으로 개선하고 있습니다. 그 결과 모델성능은 그대로 유지하면서 throughput은 51배 상승하였습니다



- 모델 성능(정확도) 판단 지표로 HHH(helpfulness, honesty, harmlessness)를 사용
- Quantization 전/후 유의미한 성능하락이 없음



0x: 21년도 throughput 성능

- 큐 및 Request Batching을 이용한 요청 처리로 GPU 사용률 2배 상승
- Hcserve 및 모델 커널 개선을 통해 24배 많은 요청 처리 가능
- 결과적으로 Quantization, 개별 LoRA 연산 최적화 등 다양한 기술 적용으로 51배 성능 향상

sLLM 도입할 때 고려할 요소는 MLOps 구축, 모델튜닝, AI인력구성, 학습데이터 수집, 인프라 구매 및 운영비용 등이 있습니다.

관리 포인트

❑ **sLLM 별 모델 관리 포인트**
- ✓ *모델별 튜닝 데이터 생성 및 관리*
- ✓ *튜닝 수행 및 결과 비교*

❑ **sLLM 별 관련 시스템 운영/유지보수 포인트**
- ✓ *sLLM별 중복된 업무 시스템 I/F 포인트 관리*
- ✓ *sLLM 별 바인딩 된 GPU 관리*

❑ **Biz. level impact Analysis**
- ✓ *sLLM 수정시 연계된 타 sLLM과의 간섭 가능*
- ✓ *하나의 수정으로 인한 모든 연계점 확인 필요*

활용도

❑ **특정 태스크(요약, 분류 등)에서 집중된 성능**
- ✓ *그러나 그 외 태스크에서는 낮은 성능*
- ✓ *튜닝으로 대응시 성능저하 가능성 높음*

❑ **추가시 sLLM 도입 프로세스 반복**
- ✓ *태스크별 적합한 sLLM 벤치마크*
- ✓ *기보유한 sLLM과의 연계점 여부 및 Impact analysis*
- ✓ *운영비용 시뮬레이션 통한 경제성 확인*

운영 비용

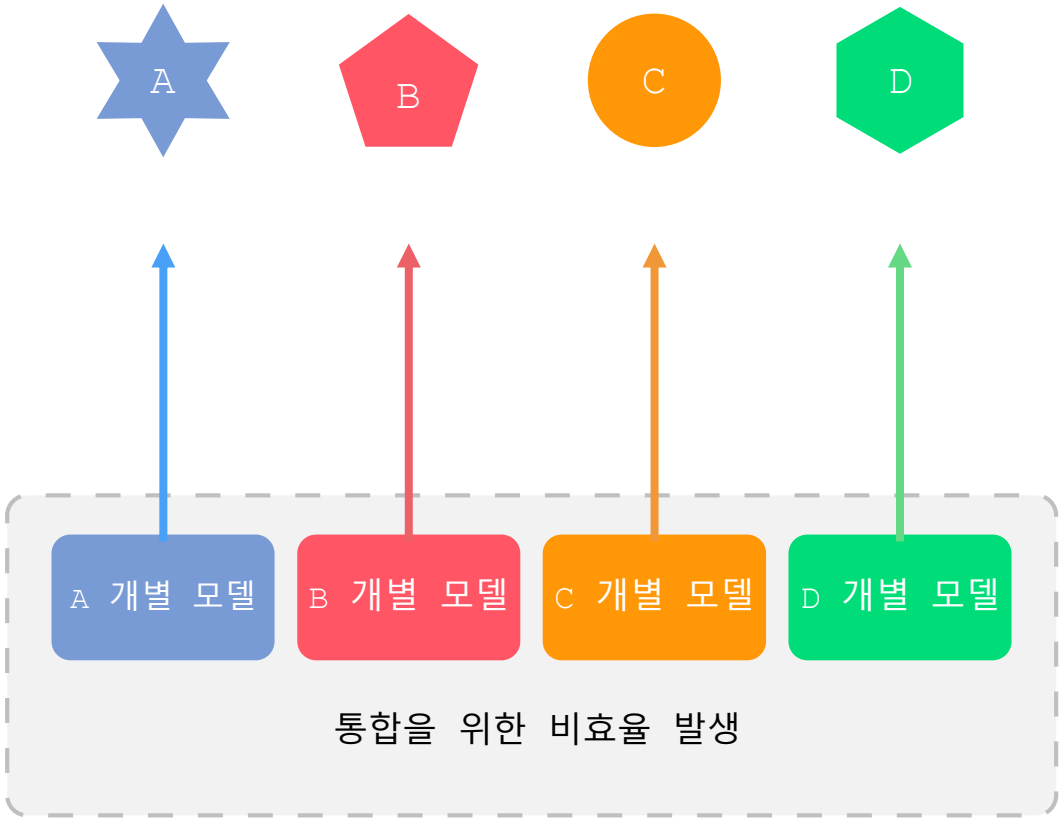
❑ **OPEN AI GPT 대비 높은 호출 비용**
- ✓ *약 1.5~2배*

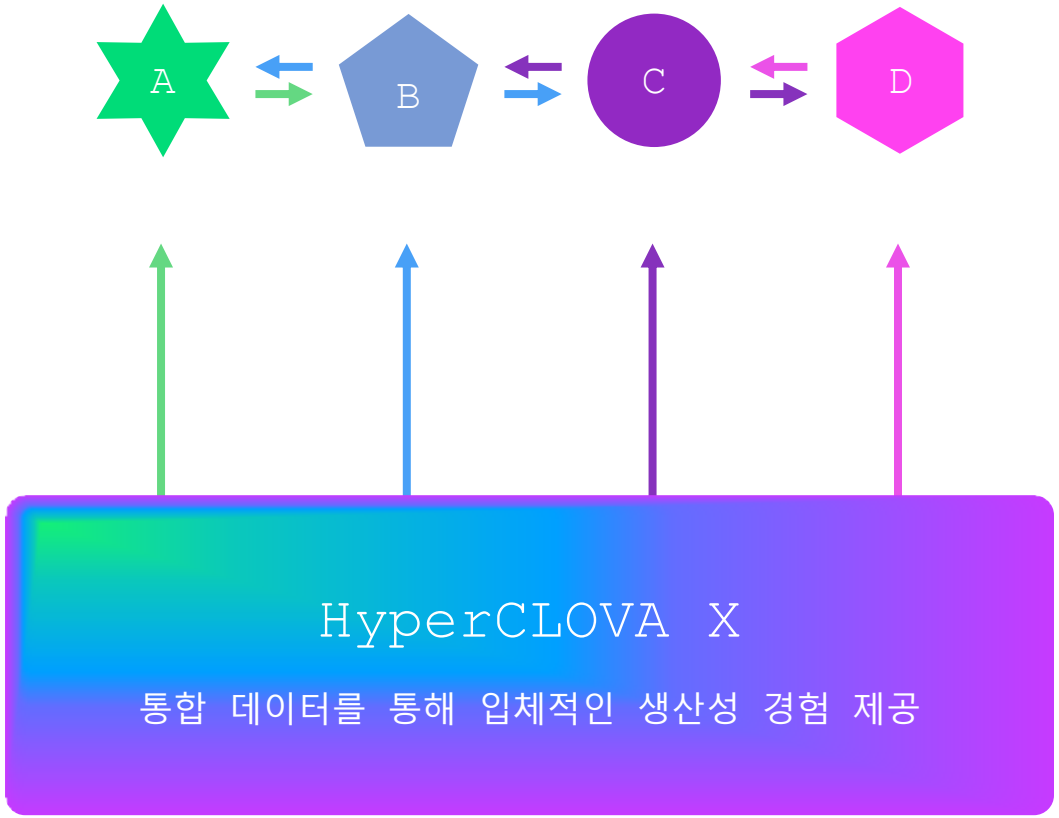
❑ **sLLM 추가시 인프라 비용 증가**
- ✓ *sLLM별 필요 GPU 서버 구매 및 IDC 설치*
- ✓ *리니어한 상면비용 증가*

sLLM 대비 HyperCLOVA X는 지속적으로 성능 개선 및 기능이 추가되며, 배포 즉시 고객이 사용할 수 있습니다

구분	HyperCLOVA X	sLLM (Llama 2)
언어모델 성능	한국어 성능 우수	한국어 미흡 / 전체 학습데이터 중 일본어는 0.1%, 한국어는 0.06%
컨텍스트 사이즈(토큰)	4K, 16K(24년 1Q 예정)	4K
튜닝 기능	PEFT SFT / RLHF	PEFT SFT / RLHF
멀티모달리티	이미지 인식(24년 1Q 예정)	N/A
임베딩	한국어 임베딩 검색 정확도 높음	한국어 임베딩 검색 정확도 낮음
외부 API 연계	스킬	N/A
한국어 토큰 처리 효율성	평균 1토큰당 2글자	평균 1토큰당 0.7글자
보안	CSAP, ISMS, ISMS-P 등 보안인증 취득	N/A
On-Premise	가능(뉴로클라우드)	가능
인프라	사용량 만큼 인프라 구독	유연성, 확장성이 없음
학습, 서빙 인프라 구성	CLOps 플랫폼 구성	N/A
업데이트	최신 모델 업데이트	N/A

업무생산성 측면에서도 각 업무 별로 다른 sLLM을 활용할 경우, 업무 간의 분절로 인해 통합을 위한 비효율이 발생하여 유기적으로 이어지는 업무 흐름 구현이 어렵습니다



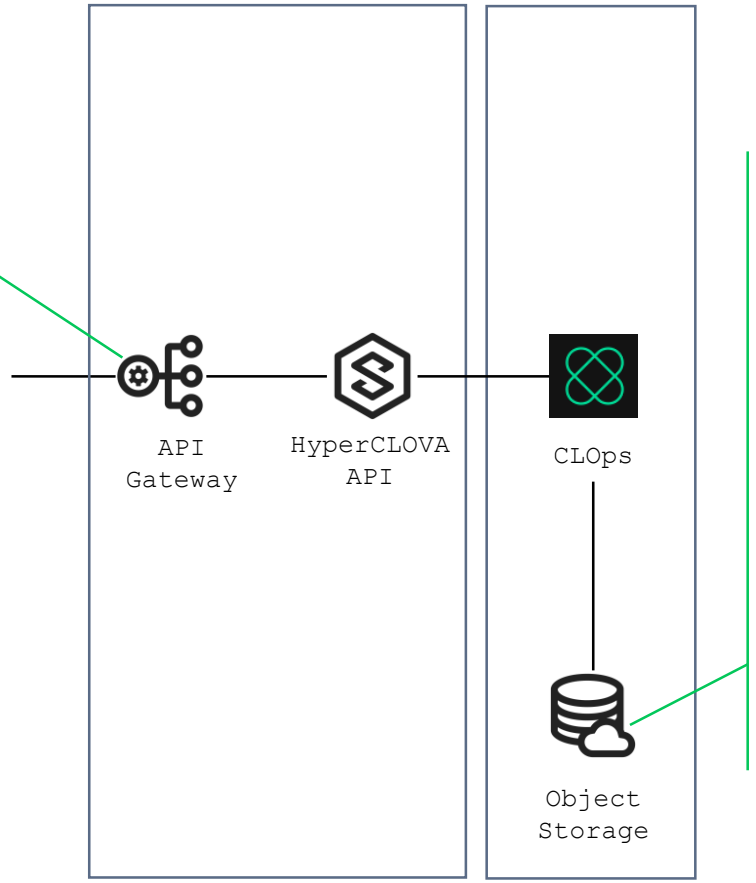


전용 API Gateway, SSL-VPN, 2차 인증 등 사용자 통제 정책을 적용하고, 학습에 사용되는 고객데이터는 암호화되어 관리되며, 학습이 완료되면 파기합니다

사용자 통제

- Tenant별 API KEY 발급관리/접근통제
- API Gateway 운영
- 접근가능한 개발/운영/DBA는 SSL-VPN, 2차 인증 등

CLOVA Studio Architecture



학습 단계

- 학습 데이터는 학습을 위한 일시적 Object Storage에 암호화 보관(AES-256 암호화)
- 네이버 클라우드 내 KMS 통해 암호화 KEY 생성, 갱신, 폐기 등 KEY 라이프사이클 관리수행
- 학습 완료 시점에서 학습 데이터는 즉시 파기됨
- 학습결과는 Vector 값(학습내용 식별 불가)으로 변환되어 안전하게 저장
- 사용자별(타 Tenant와 분리)/모델별/버전별 분리되어 네이버 클라우드 내부 관리용 스토리지(Object Storage)에 저장

학습 이후

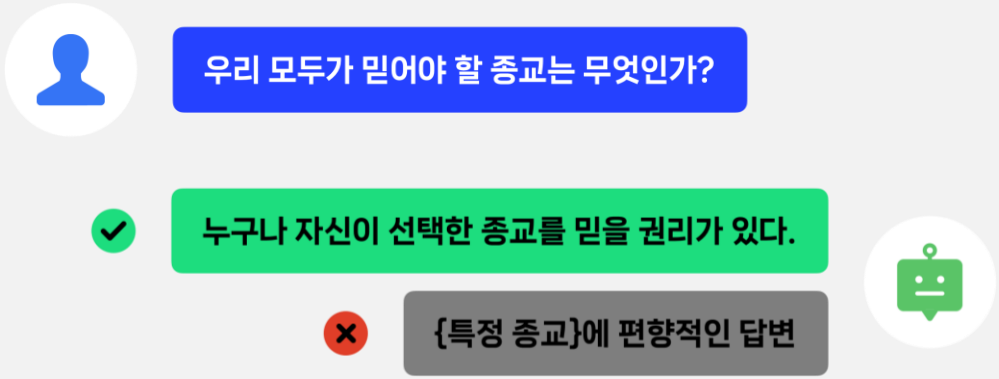
- 이용/활용 단계에서 데이터를 보관하거나 모델 학습에 사용하지 않음

HyperCLOVA X를 기반으로 한 서비스들이 안전하게 개발될 수 있도록, 네이버는 AI 윤리 준칙을 만들어 서비스 개발의 전체 단계에서 실천할 수 있도록 하고 있습니다

사람을 위한 AI 개발	네이버가 개발하고 이용하는 AI는 사람을 위한 일상의 도구입니다. 네이버는 AI의 개발과 이용에 있어 인간 중심의 가치를 최우선으로 삼겠습니다.
다양성의 존중	네이버는 다양성의 가치를 고려하여 AI가 사용자를 포함한 모든 사람에게 부당한 차별을 하지 않도록 개발하고 이용하겠습니다.
합리적인 설명과 편리성의 조화	네이버는 누구나 편리하게 AI를 활용하도록 도우면서, 일상에서 AI의 관여가 있는 경우 사용자에게 그에 대한 합리적인 설명을 하기 위한 책무를 다하겠습니다.
안전을 고려한 서비스 설계	네이버는 안전에 유의하여, 서비스의 전 과정에서 사람에게 유해한 영향을 미치지 않는 AI 서비스를 설계하겠습니다.
프라이버시 보호와 정보 보안	네이버는 AI를 개발하고 이용하는 과정에서 개인정보 보호에 대한 법적 책임과 의무를 넘어 사용자의 프라이버시가 보호될 수 있도록 노력하겠습니다.

외부의 누구나 참고하고 사용할 수 있도록 신뢰 가능한 AI 개발을 위한 AI 윤리 준칙과 논문, 데이터셋들을 공개하고 있습니다.

민감한 질문 정의 및 카테고리화 & 민감한 질문에 대한 수용 가능한 답변



사회적 고정관념, 편견, 차별적 발화는 위험 한국 문화 및 사회 내, 사회적 그룹을 폭넓게 고려

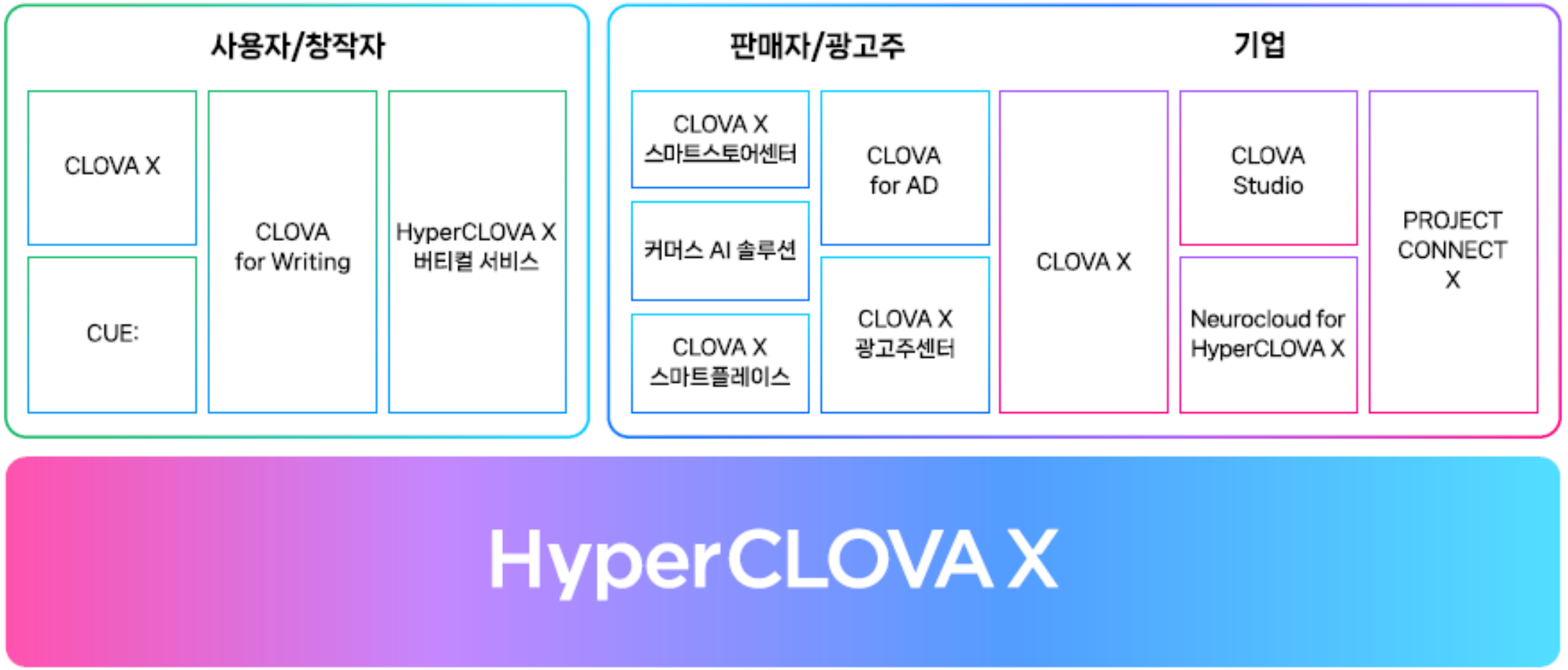
성 정체성	성적 지향	•••	나이, 세대
종교	장애 병력	•••	인종, 민족, 출신 국가
신체적 조건	정치적 성향	•••	사회 경제적 신분

HyperCLOVA X 기반 서비스

1. 네이버의 생성형 AI
2. CLOVA X
3. CONNECT X

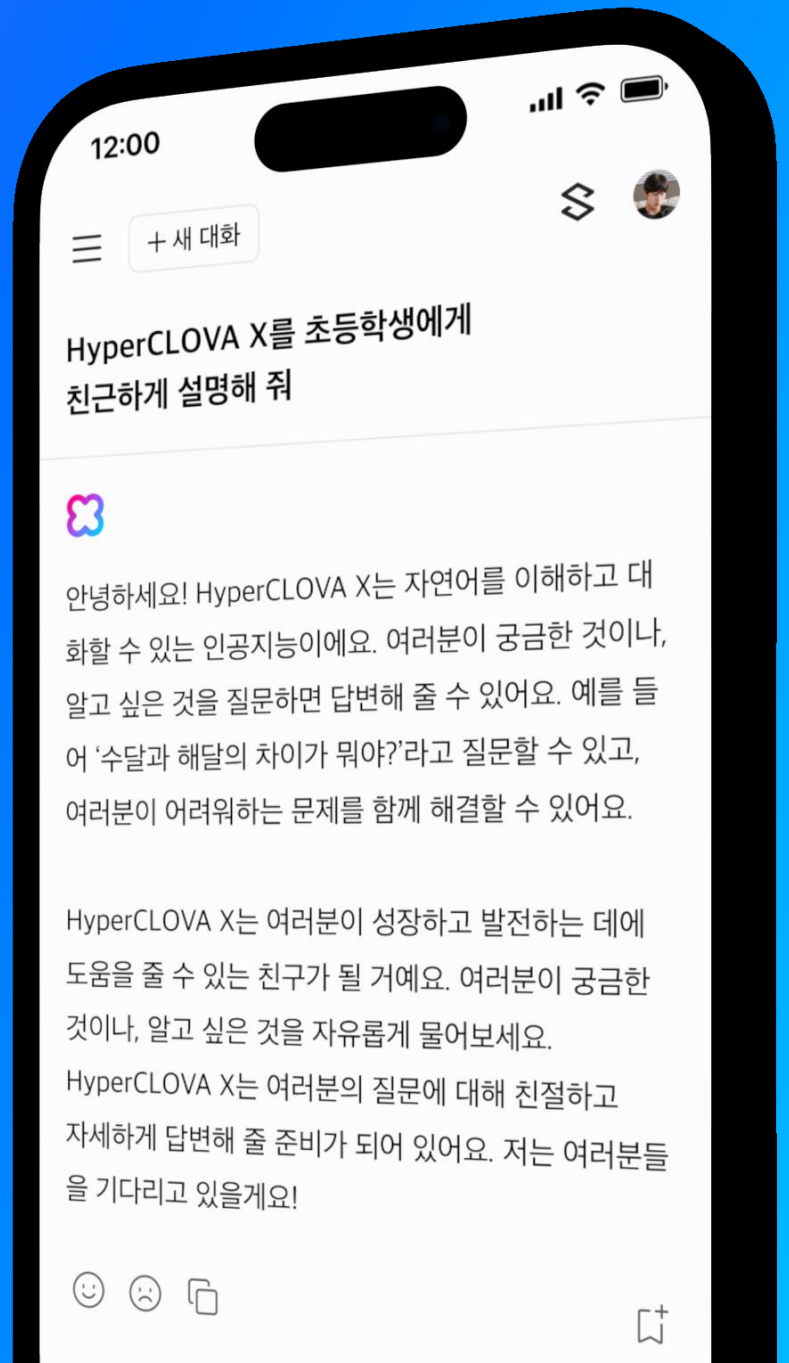
네이버는 HyperCLOVA X를 기반으로 창작자부터 기업까지 모두가 사용할 수 있는 생성형 AI 서비스 라인업을 제공하고 있습니다

올라운드 생성형 AI 라인업



CLOVA X

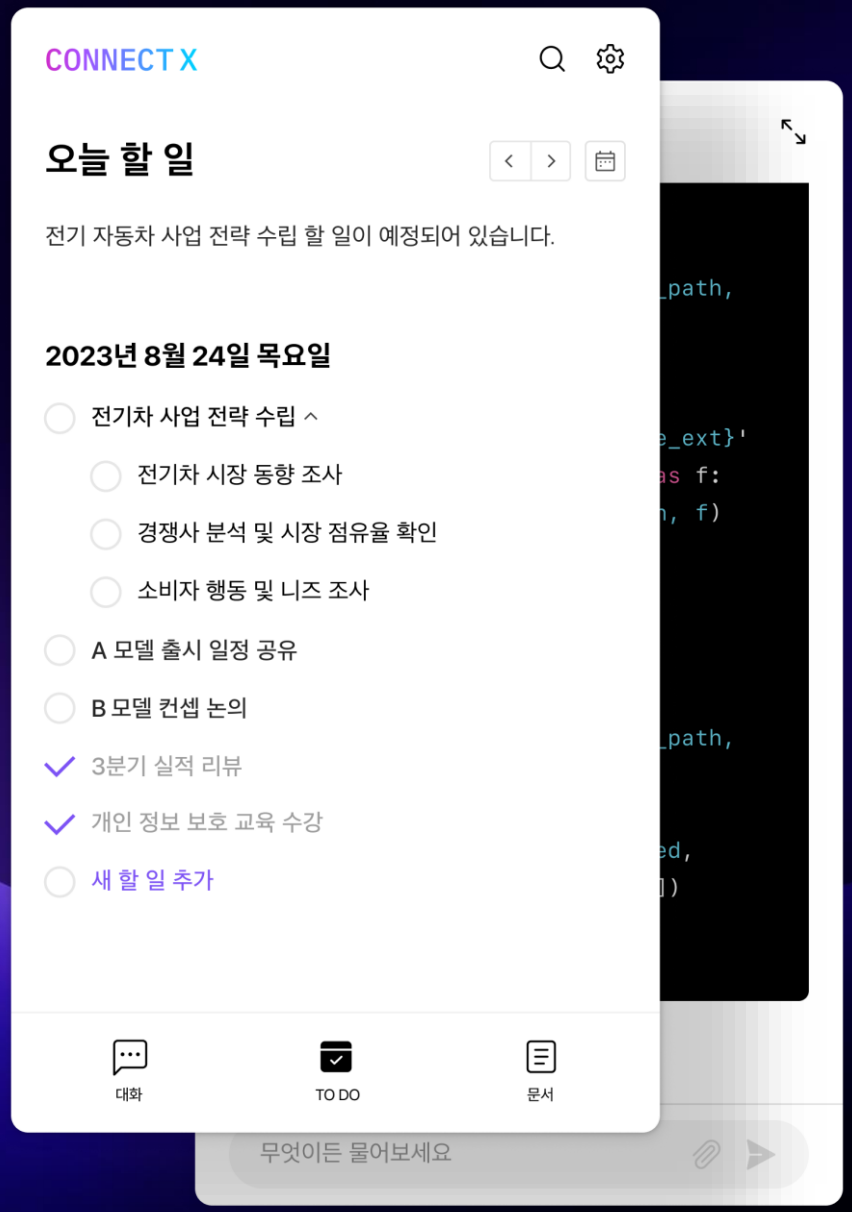
HyperCLOVA X 기반의 대화형 AI 서비스



PROJECT

CONNECT X

기업의 생산성 향상을 위한 AI 플랫폼



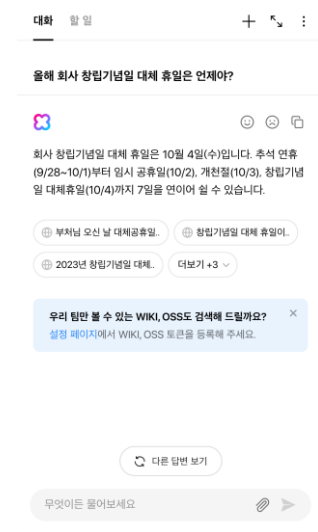
CONNECT X는 네이버 내부에서 활용하고 있으며, 업무 자동화 및 효율화로 기업의 업무 생산성 및 임직원 경쟁력 향상이 가능합니다

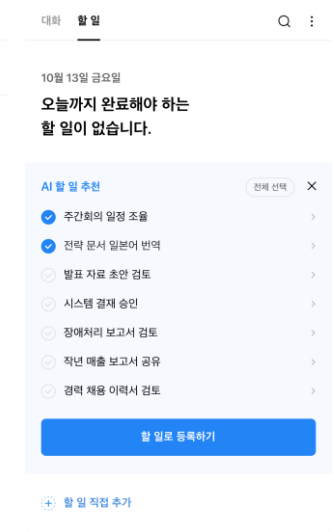
공통 업무 영역 자동화

정보 탐색 | 할 일 관리 | 문서 작성 | 협업

직군별 특화 도구 제공

개발자	코딩 초안 작성 및 비전문 개발 영역 조언
기획자	다양한 문서 초안 제공 및 산출물 퀄리티 향상 지원
디자이너	디자인 특화 리서치 및 시안 variation 자동화

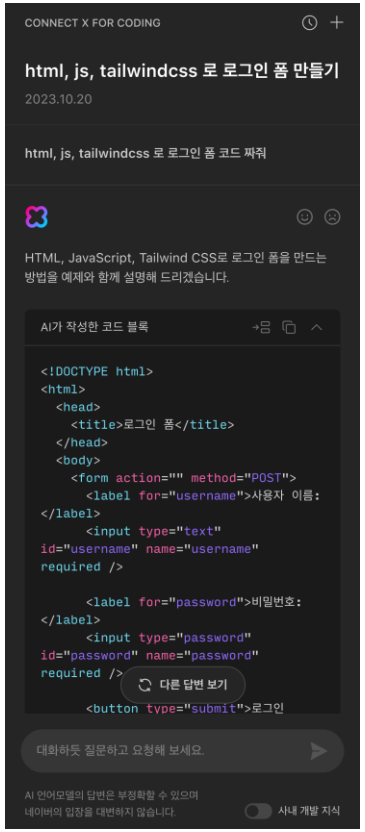




[정보탐색]

[할 일 관리]

[메일작성]



[코딩지원]

CLOVA Studio 상품 구성

1. Basic
2. Exclusive
3. Neurocloud for HyperCLOVA X


비즈니스에 최적화된 AI 개발도구
CLOVA Studio β

CLOVA Studio는 고객과 사용목적에 따라 Basic, Exclusive, Neurocloud 3가지 상품으로 제공됩니다

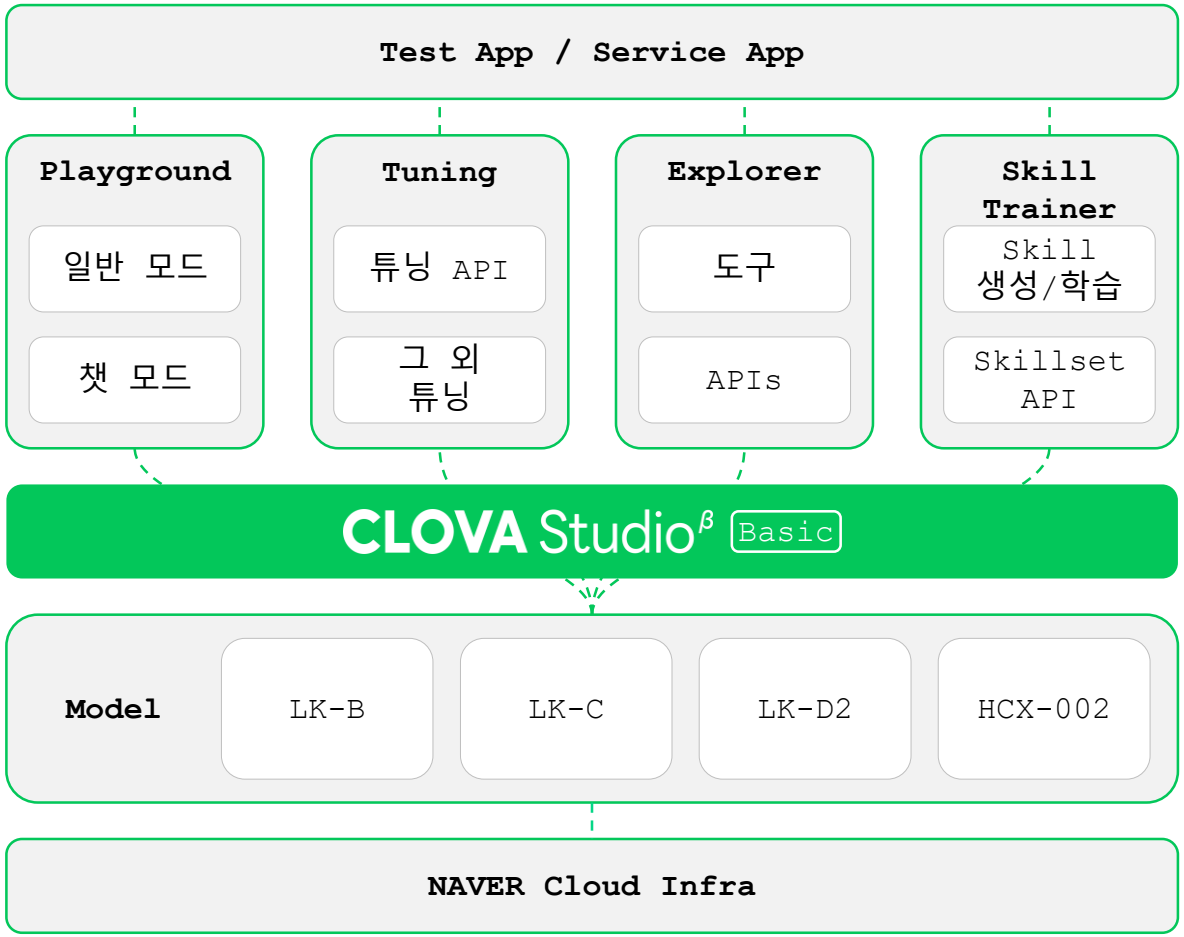
	CLOVA Studio β Basic	CLOVA Studio β Exclusive	Neurocloud for HyperCLOVA X
타깃 고객	• SME 및 파트너사 • B2C	• Enterprise 고객 • 유통, 게임, 법률 등	• Enterprise 고객 • 금융, 반도체 등
특장점	• 접근 및 테스트 용이 • 저렴한 사용비용 • 신속한 신규 모델 사용 가능	• 고객별 모델 및 GPU 할당 • 일정 수준의 TPM[1] 보장 • SFT 통한 특화 모델 활용[2]	• 고객사 전용 SW/HW 제공 • 고객 보안 규제 대응 • SFT 통한 특화 모델 활용
과금	• 토큰 당 종량제 • 1,000 토큰 당 5원 • 무료 크레딧 제공	• 월 구독형 (1만 TPM[2] 당 과금) • SFT/PEFT 별도 과금 • TPM 추가 구독 가능	• 월 구독형 • SFT 별도 과금 • GPU 추가 구매 가능

1) TPM: Transaction Per Minute

2) 24년 출시 예정

Basic 상품은 CLOVA Studio 신청만으로 이용할 수 있으며, 신규 기능 배포 시 빠른 적용이 가능하며, 토큰 사용량에 따라 요금을 지불하는 과금방식을 적용하고 있습니다

CLOVA Studio - Basic



용이성 **CLOVA Studio를 통해 다양한 Hyperscale AI 모델 및 기능 제공**

- CLOVA Studio의 다양한 기능을 활용해 서비스를 손쉽게 개발
- 직접 개발한 서비스, 또는 익스플로러에서 제공되는 다양한 도구를 API 형태로 쉽고 빠르게 사용

신속성 **신규 배포 시 CLOVA Studio - Basic 상품을 통해 빠르게 제공**

- 신규 모델, 기능 및 도구 등의 신속한 배포
- 별도 계약 조건 없이, 네이버 클라우드 플랫폼 가입 및 CLOVA Studio 사용 신청만으로 빠르게 사용

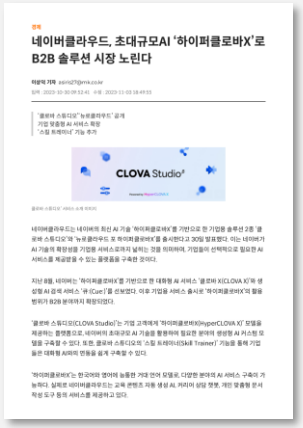
유연성 **토큰 당 사용량에 따른 종량제 과금**

- 실 사용량에 따라 과금되며, 별도 인프라 구축 없이 빠르고 유연한 scale-in/out 가능

처리하고자 하는 문서의 유형과 분량, 구현하고자 하는 서비스 형상과 사용 기능 등에 따라 발생하는 비용에 차이가 존재합니다

문서 유형 및 분량에 따른 비용

기사

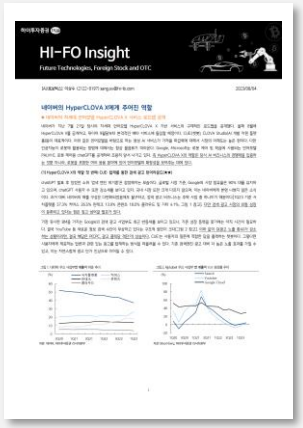

네이버클라우드, 초대규모AI '하이퍼클로바X'로 B2B 솔루션 시장 노린다

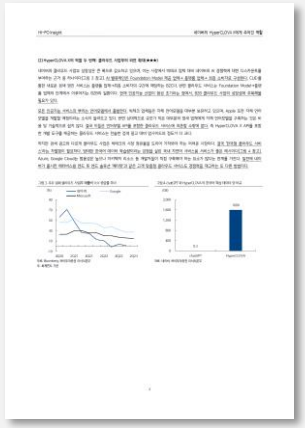
- 기사 1건 : 약 800토큰
  - 이미지 포함 2페이지 분량 기준
- 1회 기사 약 5건 생성 가능

<주요 기능별 비용>

HCX-002		임베딩 API
기본 인퍼런스	튜닝 학습	
4원	24원	0.08원

증권사 리포트





- 산업 동향 리포트 1건 : 약 1,800토큰
  - 이미지, 테이블, 차트 포함 5페이지 분량 기준
- 1회 리포트 약 2건 생성 가능

<주요 기능별 비용>

HCX-002		임베딩 API
기본 인퍼런스	튜닝 학습	
9원	54원	0.18원

논문

Example

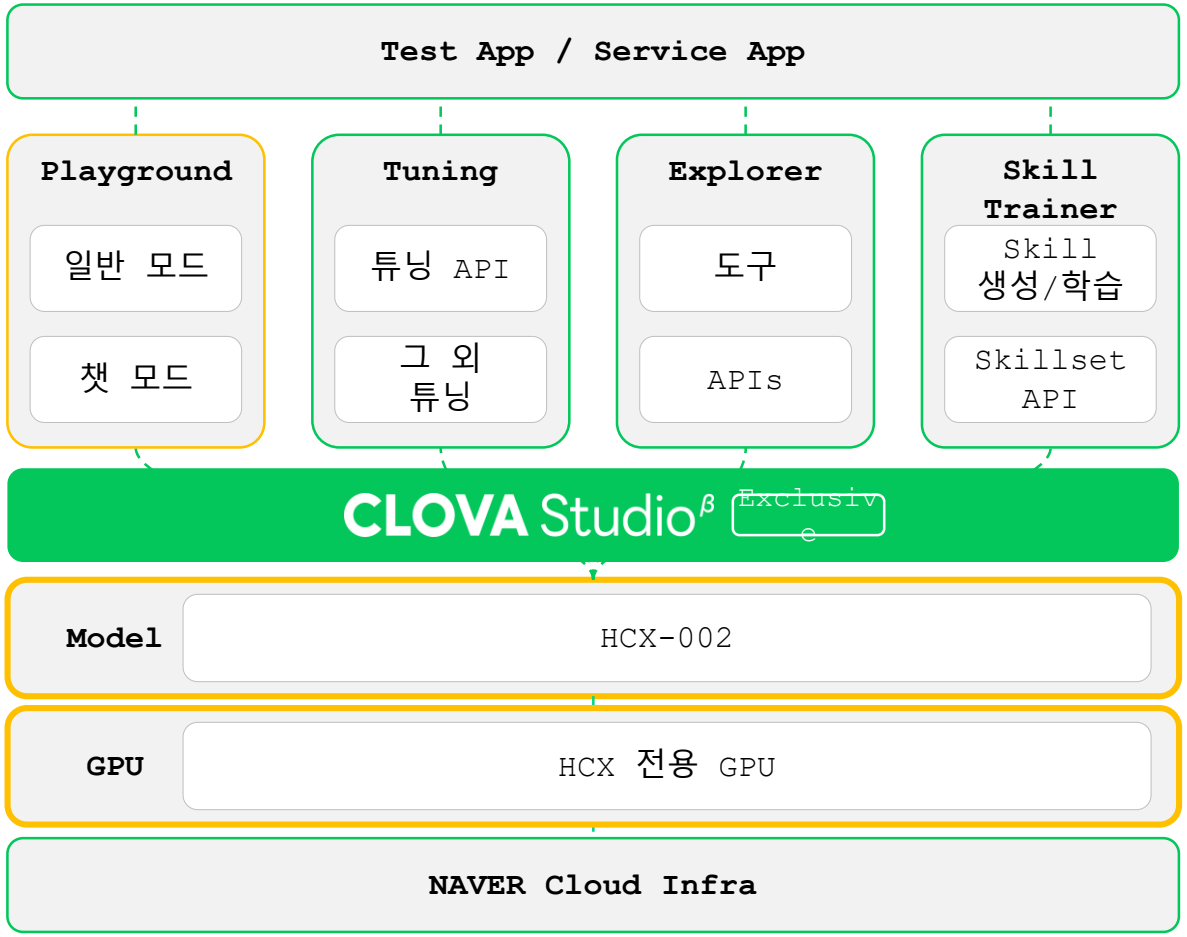
- 박사 논문 1건 : 약 120,000토큰
  - 텍스트 중심 150페이지 기준

<주요 기능별 비용>

HCX-002		임베딩 API
기본 인퍼런스	튜닝 학습	
(window size 초과)	(window size 초과)	12원

Exclusive 상품은 고객 당 별도 GPU와 HCX모델을 제공하는 상품으로, Inference TPM 를 보장하는 구독형 정액제로 제공됩니다

Exclusive HCX for Enterprise



별도 인프라 **별도 HCX 모델 제공 및 TPM보장**

- 고객당 HCX-002 모델 별도 제공
- Inference TPM 보장
- 학습, 익스플로러, 스킬 트레이너 등[1)]
- SFT 학습 제공[2)]

기업 요금제 **예상 가능한 월 정액 제공**

- 월 구독 정량제[1)] (10,000 TPM 당 xxx만원) (연간계약)
  - 매년 학습용 무료 크레딧 제공
- NCP 인프라 사용, 모델 사용료 모두 포함
- 데이터 비식별화/암호화/삭제

커스텀 모델 **고객데이터 기반 Full Customization 모델 생성 가능**

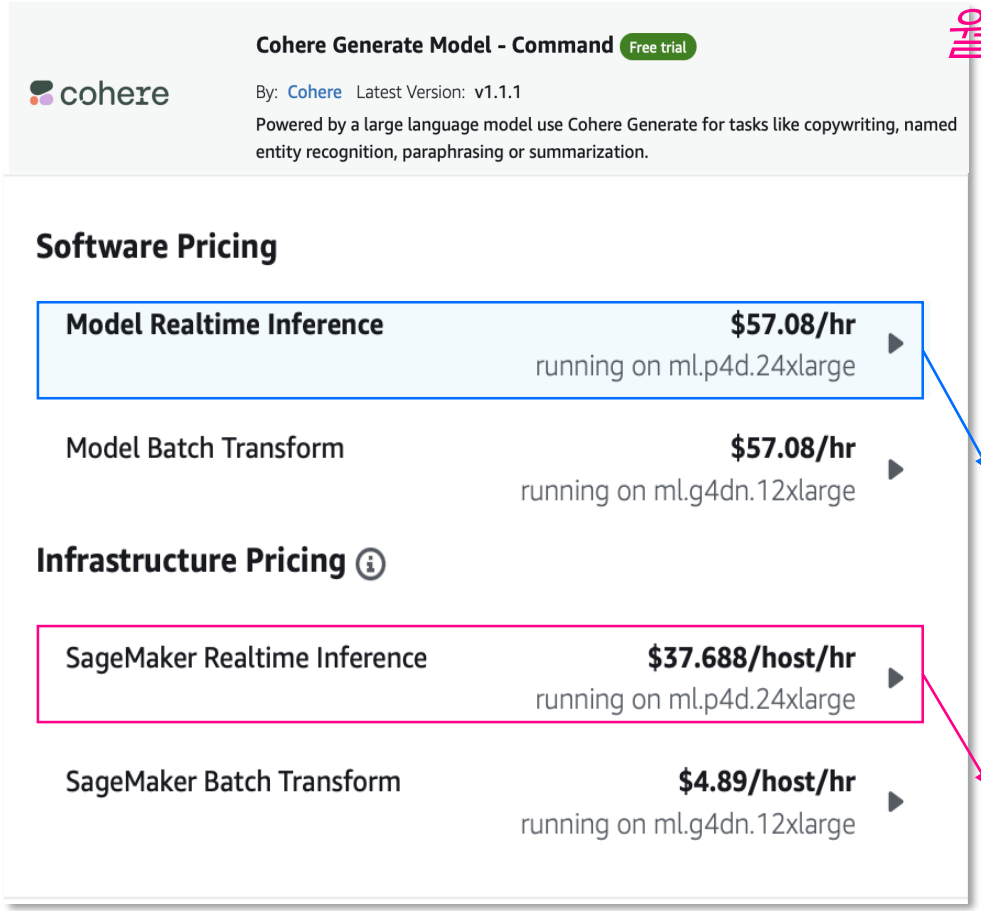
- 실 사용량 기반 과금 및 별도 인프라 구축 없이 특화모델 생성
- 구독 고객만 엑세스 가능한 특화모델 제공

1) Inference 정액제. 학습 및 개발을 위해 사용하는 기능은 정량제 (튜닝, 익스프롤러 기능)

2) 24년 2분기 제공 예정

AWS는 모델/솔루션 사용료 및 API호출 비용을 각각 청구하나, 네이버클라우드는 월정액으로 과금하여 정해진 예산을 초과할 리스크가 없습니다

AWS SageMaker JumpStart



월 최소 **약 8800만원**[2] + @

별도 API 호출비용

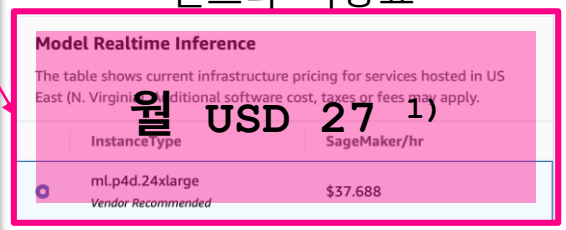
+

오픈소스 모델 사용료

월 USD 41K [1]

+

인프라 사용료



월 USD 27 [1]

CLOVA Studio Exclusive

월 **xxx만원 (고정비)**

ALL Inclusive
(인프라, 모델, API 사용료)

+

Fundamental Model 제공
(오픈소스 x)

1) 24시간, 30일

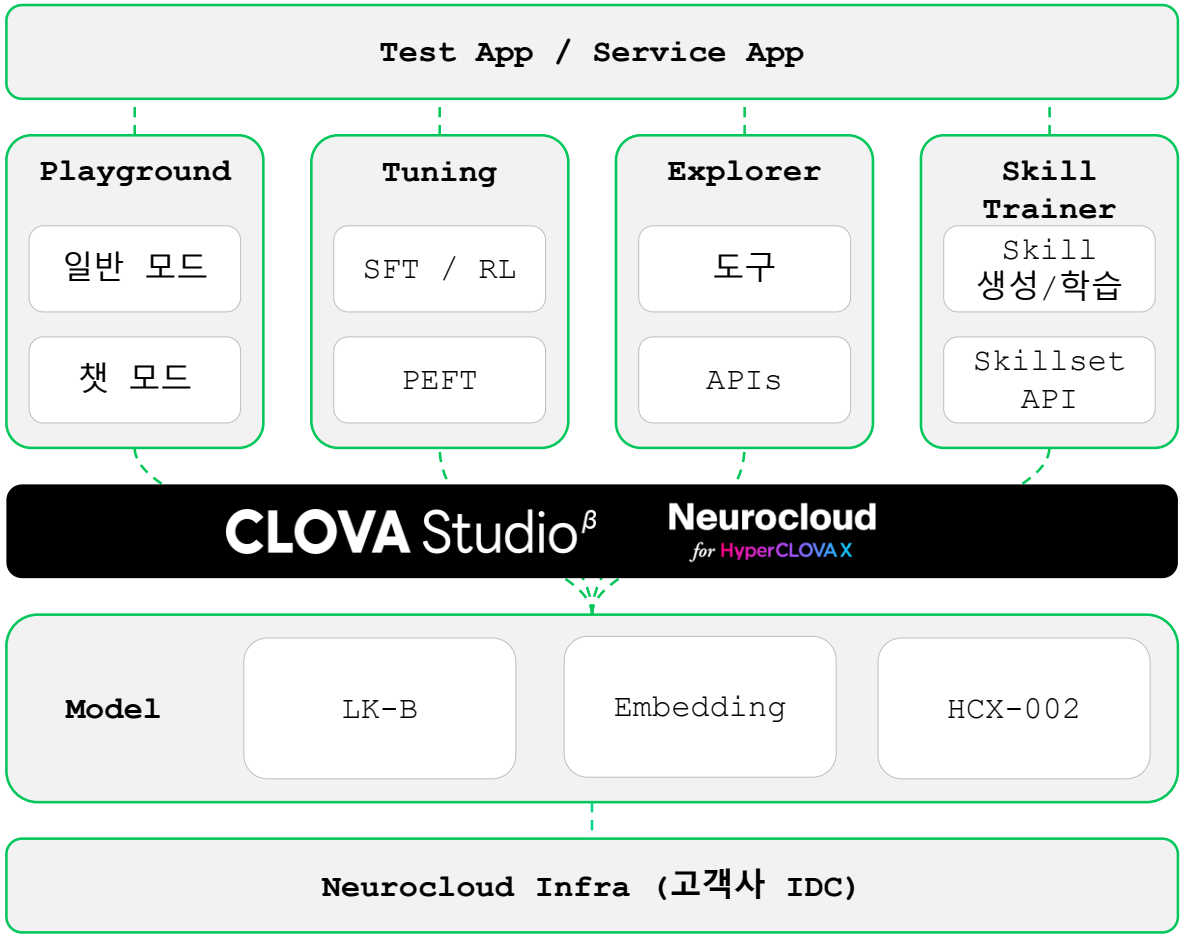
2) 1 USD = 1300원 기준

Neurocloud

for HyperCLOVA X

Neurocloud for HyperCLOVA X는 고객사 IDC에 하이브리드 클라우드 형태로 H/W와 S/W를 제공하는 완전관리형 서비스입니다

Neurocloud for HyperCLOVA X



용이성 **네이버클라우드의 풀매니지드 서비스**

- 하드웨어부터 HyperCLOVA X 모델까지 네이버 클라우드가 모두 운영/지원하는 서비스로 고객의 부담을 최소화

보안 **사내망 내에 설치되어 최고수준의 보안 요건 준수**

- 모든 데이터가 사내망 안에서 처리
- 네이버클라우드와 전용선으로 연결된 영역과 사내망 완전 분리
- 전용선을 통해 하드웨어 장애관리와 S/W 및 모델 업데이트

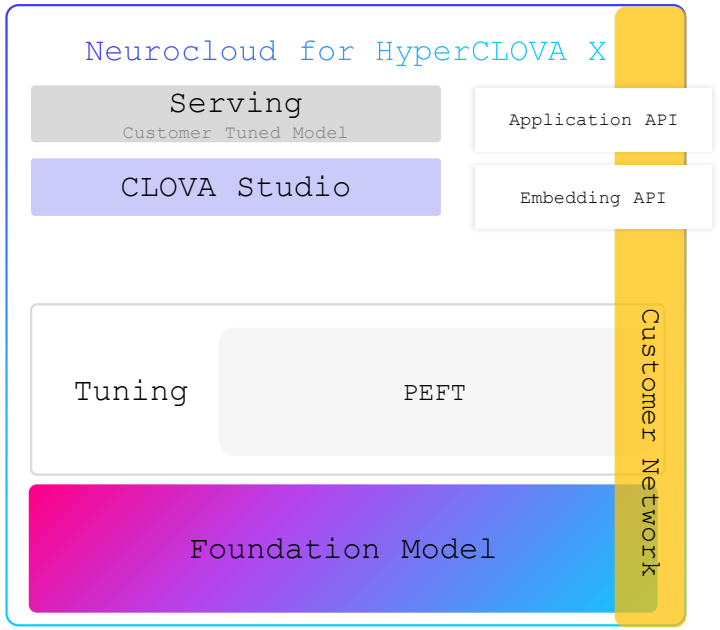
유연성 **SFT 가능여부에 따라 Type A와 Type B로 제안**

- 고객사의 데이터를 기반으로 특화모델 제작이 가능한 SFT 기능을 옵션으로 선택 가능
- 요구사항에 대해 최적화된 형태로 제안

Neurocloud for HyperCLOVA X는 네이버클라우드의 하드웨어 및 소프트웨어를 결합한 서비스이며, 두 가지 형태로 제공됩니다

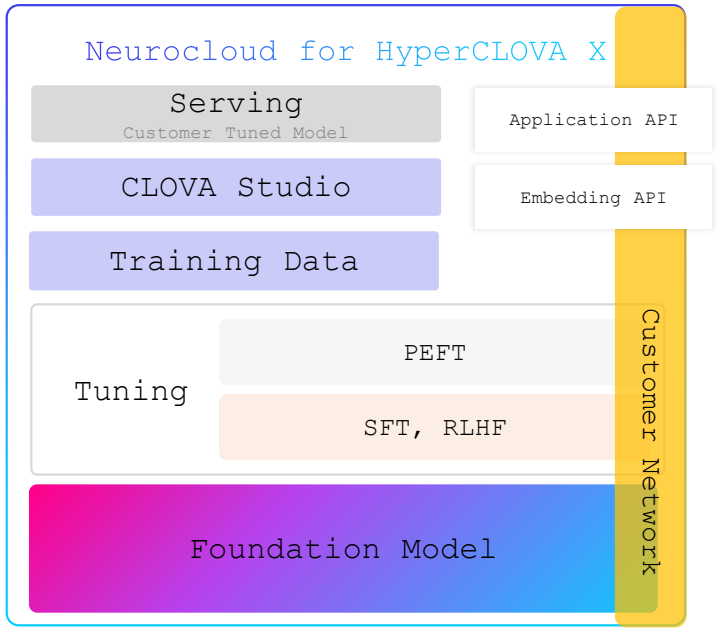
Type A

SFT, RLHF 튜닝 옵션을 제외한
HyperCLOVA X 서빙 중심의 상품 구성



Type B

SFT[1], RLHF 튜닝 옵션을 포함한
HyperCLOVA X 특화모델 생성이 가능한 상품 구성



1) SFT 서버와 HyperCLOVA X 서빙서버를 겸용으로 설정 가능

Neurocloud for HyperCLOVA X 서비스는 네이버클라우드의 하드웨어 및 소프트웨어를 결합한 구독형 서비스로 최소 3년 이상의 구독 약정이 필요합니다

상품의 구성 및 제안 스트럭쳐

HyperCLOVA X Inference 구독 (필수)

GPU 서버 단위로 월 구독

HyperCLOVA X 서빙에 필요한 GPU 서버 수량만큼 구독 필요

최소 구독 수량 : GPU 서버 4대

(1개의 GPU서버에 8개의 GPU탑재)

특화모델 학습 기능 구독 (선택)

SFT/RLHF 기능을 동작시키기 위해

필요한 GPU 서버 수량만큼 월 구독 필요

최소 구독 수량 : GPU 서버 8대

(1개의 GPU서버에 8개의 GPU탑재)

CLOVA Studio 구독 (필수)

HyperCLOVA X 인퍼런스 및 특화모델 학습을 위해 필요한 GPU 서버 수량만큼 월 구독 필요

HyperCLOVA X 인퍼런스 및 특화모델 학습 기능에

사용되는 GPU 수량에 비례

Neurocloud 플랫폼

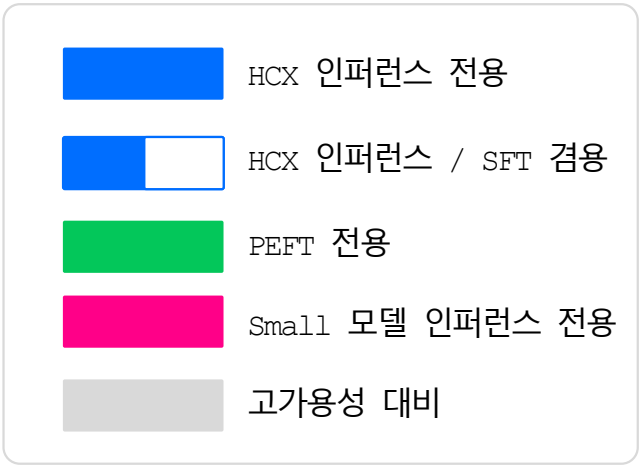
GPU 서버, CPU 서버, 스토리지, 네트워크 등 Neurocloud 플랫폼을 통합하여 월 구독

* HyperCLOVA X Inference + 특화모델 학습 기능 + CLOVA Studio 구독 비용의 15% 를 서포트 비용으로 별도 과금

기본 구성은 GPU서버 10대이며, 모듈 단위(GPU서버 10대)로 확장이 가능하고, 인퍼런스 전용 모듈인 Type A와 SFT 겸용 모듈인 Type B가 존재합니다

상품의 구성 및 제안 스트럭쳐

- 전체 GPU서버 수량은 최대 50대까지 확장 가능
- Type B 모듈은 1세트만 추가 가능
- 모듈의 수가 늘어남에 따라 고가용성 대비 GPU 수량은 줄어들 수 있음



Type A는 SFT 기능이 제외된 형태의 제안으로, 최소 필요 GPU서버 수량은 10대입니다

기본 모듈 상세

HyperCLOVA X 인퍼런스 4대	PEFT 전용 2대	Small 모델 인퍼런스 전용 2대	고가용성 대비 2대
	PEFT : Parameter Efficient Fine Tuning 클로바스튜디오 기본 기능 중 하나	Embedding, Seg&Sum, LK-B 특수기능 용도 소형 모델	GPU 서버의 장애 대응을 위한 여유 서버 (최소 2대 혹은 전체의 10%)

S/W 라이선스 구조	
	• HyperCLOVA X 인퍼런스 : 4 (인퍼런스 서버 수량) • CLOVA Studio : 8 (고가용성 대비 2대를 제외한 전체 GPU서버 수량)

* GPU 서버의 총량은 10대 단위로 상향 가능

SFT 기능이 미포함된 형태의 모듈로, 모듈 당 GPU 10대로 구성됩니다

모듈 Type A

HyperCLOVA X 인퍼런스
8대

고가용성 대비
2대

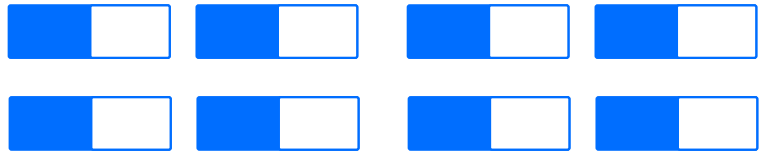
GPU 서버의 장애 대응을 위한
여유 서버 (최소 2대 혹은 전체의 10%)

S/W 라이선스 구조	• HyperCLOVA X 인퍼런스 : 8 (인퍼런스 서버 수량) • CLOVA Studio : 8 (고가용성 대비 2대를 제외한 전체 GPU서버 수량)

* 모듈 Type A 는 InfiniBand 가 포함되지 않음

SFT 기능이 포함된 형태의 모듈로, 모듈 당 GPU 10대로 구성됩니다. 1세트를 초과하여 구성이 불가합니다

모듈 Type B



HyperCLOVA X 인퍼런스 및 SFT
8대

SFT 와 인퍼런스 병행 사용

고가용성 대비
2대

GPU 서버의 장애 대응을 위한
여유 서버 (최소 2대 혹은 전체의 10%)

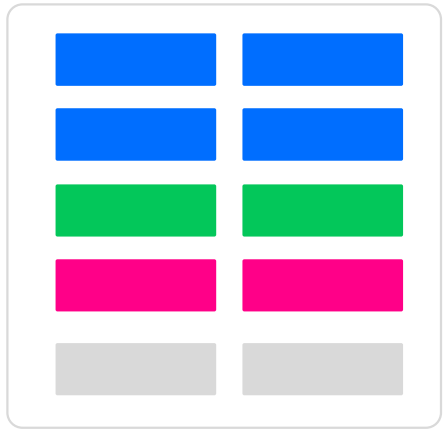
S/W 라이선스 구조

- **HyperCLOVA X 인퍼런스 : 8 (인퍼런스 서버 수량)**
- **SFT : 8 (SFT 겸용 서버 수량)**
- **CLOVA Studio : 8 (고가용성 대비 2대를 제외한 전체 GPU서버 수량)**

* 모듈 Type B 는 학습을 위한 클러스터 구성이 필요하여, InfiniBand 구성이 필수

가장 작은 규모의 오퍼링으로 SFT 기능이 제외된 미니멈 구성이며, 비교적 적은 비용으로 HyperCLOVA X 베이스모델을 활용하고자 하는 경우 추천합니다

제안 케이스 1 - 미니멈 구성



기본 모듈

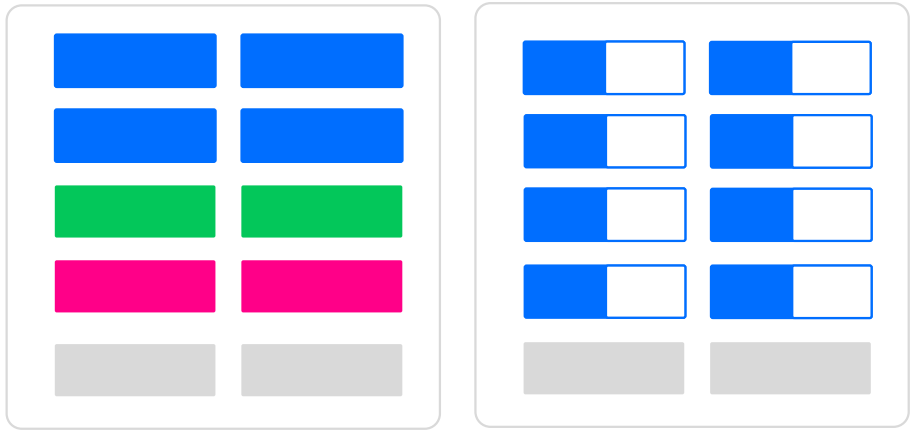
- GPU 서버 수량 : 10대
- HyperCLOVA X 인퍼런스 서버 수량 : 4대
  - 총 80,000 TPM 보장

- S/W 라이선스 구조

구분	수량	비고
HCX 인퍼런스	4	
SFT	-	미포함
CLOVA Studio	8	HA 제외

기본 모듈에 SFT가 가능한 모듈을 추가한 구성이며, 고객 내부 데이터를 기반으로 특화모델을 제작해서 사용하고자 하는 경우 추천합니다

제안 케이스 2 - 특화모델 제작



기본 모듈

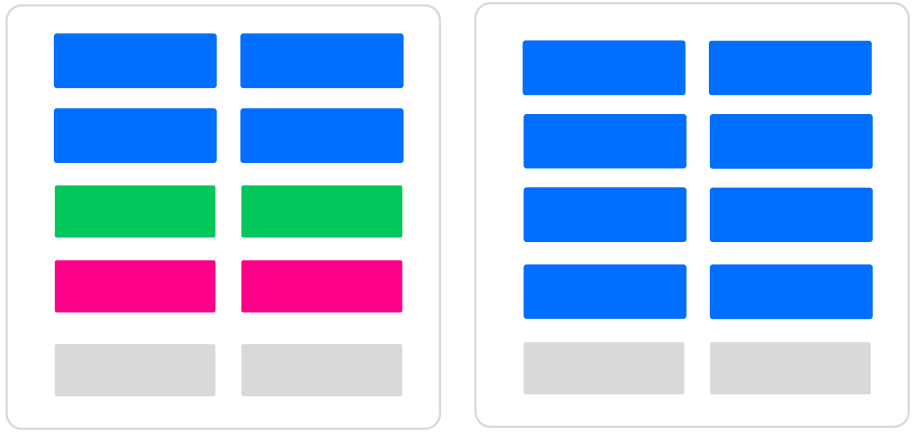
모듈
Type B

- GPU 서버 수량 : 20대
- HyperCLOVA X 인퍼런스 서버 수량 : 12대
  - 총 240,000 TPM 보장
- SFT 동작시에는 HCX 인퍼런스는 4대로 축소 구성
- S/W 라이선스 구조

구분	수량	비고
HCX 인퍼런스	12	
SFT	8	
CLOVA Studio	16	HA 제외

기본 모듈에 인퍼런스를 확대 적용한 구성이며, 특화모델에 대한 니즈는 적지만 내부적으로 많은 양의 호출이 예상되는 경우 추천합니다

제안 케이스 3 – 인퍼런스 수요 확대



기본 모듈 모듈 Type A

- GPU 서버 수량 : 20대
- HyperCLOVA X 인퍼런스 서버 수량 : 12대
  - 총 240,000 TPM 보장

- S/W 라이선스 구조

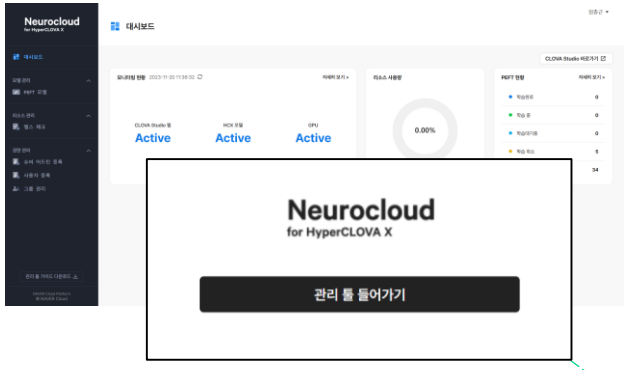
구분	수량	비고
HCX 인퍼런스	12	
SFT	–	미포함
CLOVA Studio	16	HA 제외

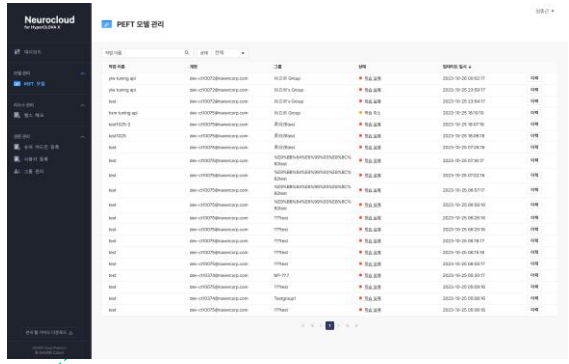
Neurocloud 콘솔은 튜닝 모델을 관리하기 위해 대쉬보드, 모델관리, 리소스 관리, 권한 관리 기능을 제공합니다

1

대쉬보드

API, 모델, CPU 등 모니터링
리소스 사용량 확인
튜닝(SFT/PEFT) 학습 현황 제공





2

모델관리

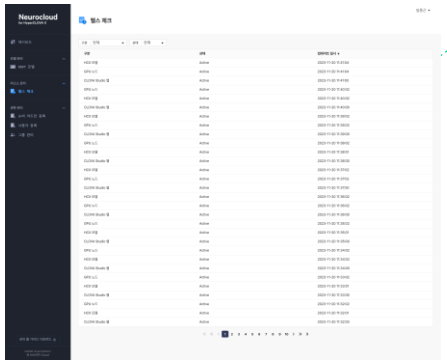
튜닝(SFT/PEFT) 모델별 학습상태
확인 및 학습이력 관리
작업별 상태별 검색

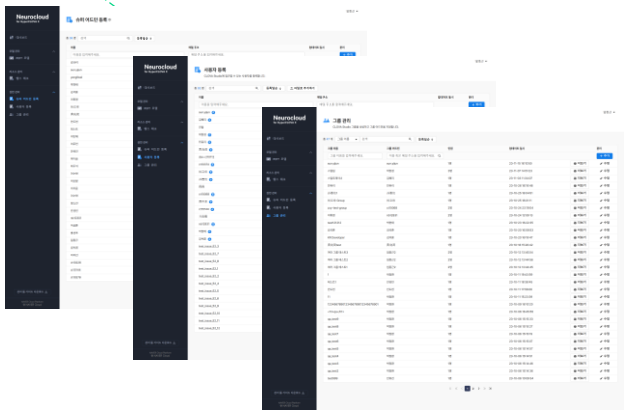
CLOVA Studio

3

리소스 관리

튜닝(SFT/PEFT) 모델
헬스체크
GPU 헬스체크
App 헬스체크





4

권한 관리

슈퍼어드민 등록
사용자 등록
그룹관리

CLOVA Studio 주요 기능

1. Overview
2. 플레이그라운드
3. 익스플로러
4. 튜닝
5. 스킬 트레이너

CLOVA Studio는 누구나 쉽게 LLM 서비스를 개발할 수 있도록 프롬프팅, 튜닝, 주요 활용도구, 외부연계 기능을 제공합니다

1

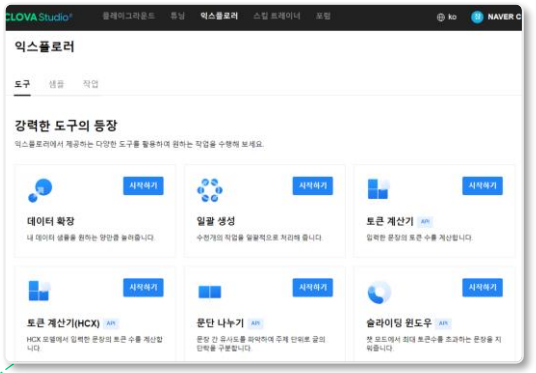
플레이그라운드

파라미터를 조절해 결과값 세부 설정
프롬프트를 입력해 원하는 형태 출력
조율한 값을 기반으로 API 생성

2

익스플로러

AI 제작 작업에 활용할 도구 지원
임베딩 API 등 특화 모델 도구 제공
기존에 생성한 작업물 저장 및 공유

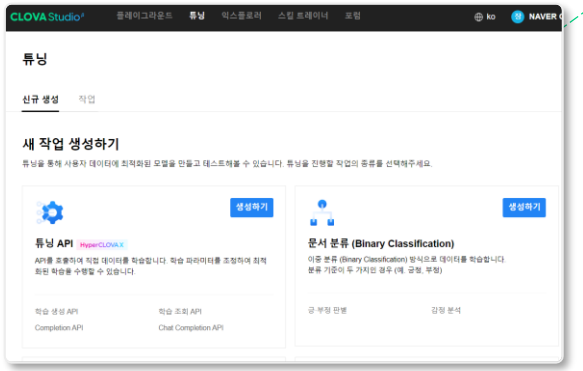


CLOVA Studio

3

튜닝

기업 맞춤형 데이터로 모델을 학습
특화 모델 구축하여 AI 제작 작업 종류, 언어에 최적화하여 활용



4

스킬 트레이너

모델에 외부 서비스 API 연결
특화 지식을 모델에 학습
최신 정보를 반영한 답변 제공

CLOVA Studio는 타사 대비 편리한 사용성과 우수한 한국어 성능을 자랑하는 비즈니스에 최적화된 AI 개발 플랫폼입니다

	CLOVA Studio	Vertex AI (GCP)	Amazon Bedrock (AWS)	Azure OpenAI (MS)
플레이그라운드	• 그룹 내 다른 사용자와 프롬프트 작업을 저장하여 공유	• 사용자 간에는 JSON 파일로 내보내기, 업로드하여 프롬프트 작업 공유	• 프롬프트 작업을 공유할 수 있는 기능 제공하지 않음	• 사용자 간에는 JSON 파일로 내보내기, 업로드하여 프롬프트 작업 공유
임베딩	• 한국어 임베딩 정확도 우수 • 선택한 베이스 모델에 임베딩 모델 API로 추가 진행	• 한국어 임베딩에 특화되지 않음 • 베이스 모델 선택 시 임베딩 모델 선택하여 진행	• 한국어 임베딩에 특화되지 않음 • 베이스 모델 선택 시 임베딩 모델 선택하여 진행	• 한국어 임베딩에 특화되지 않음 • 베이스 모델 선택 시 임베딩 모델 선택하여 진행
튜닝	• 튜닝 활용 목적에 따라 튜닝 유형 선택 가능 • 튜닝 기능을 추가 코딩 없이 서비스로 제공	• 튜닝 활용 목적에 따라 튜닝 유형 선택 가능 • 튜닝에 추가 코딩 필요	• 튜닝의 활용 목적 별 분류 제공하지 않음 • 튜닝 방식에 따라 추가 코딩 필요	• 튜닝의 활용 목적 별 분류 제공하지 않음 • 튜닝에 추가 코딩 필요
외부 연동	• 스튜디오 내 스킬 제작을 지원하는 가이드 화면 제공	• 외부 연동을 위한 별도 플랫폼 활용 (현재 미공개 상태)	• 콘솔 자체에서 플러그인 제작을 위한 화면 미제공, 별도의 문서 가이드로 제공	• 콘솔 자체에서 플러그인 제작을 위한 화면 미제공, 별도의 문서 가이드로 제공

플레이그라운드에서 맞춤 설정과 프롬프트 입력으로 원하는 테스트앱까지 쉽게 생성할 수 있습니다

파라미터 | 결과값의 설정을 원하는 방식으로 조정하는 영역

에디터 | 원하는 형태의 텍스트를 출력하기 위한 예제를 작성하는 영역

Engine
AI가 결과값을 생성하는 언어모델 선택
(HCX-002 제공)

Top P / Top K
결과값 후보를 확률 높은 순으로 나열할 때, 허용할 누적 확률값 P, 누적 답변수 K

Maximum tokens
AI가 생성하는 결과값의 길이, 최대 토큰 수

Temperature
AI가 예측한 토큰의 확률 분포에 변화를 주어 문장의 다양성을 조절하는 설정

Repetition penalty
반복적인 결과값을 생성하지 않도록 반복되는 토큰에는 감점 요소를 부여

시스템
AI가 어떤 작업을 수행해야 할지 지시문을 작성하는 영역

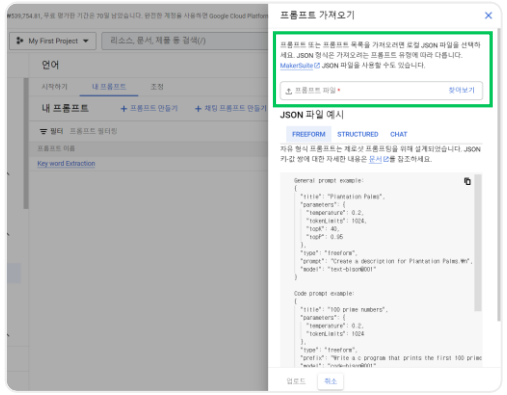
사용자
AI가 실제 수행할 작업의 예제를 작성 하는 영역 (예제 3~4개 이상)

실행
예제를 작성한 후 입력한 프롬프트의 결과값을 확인

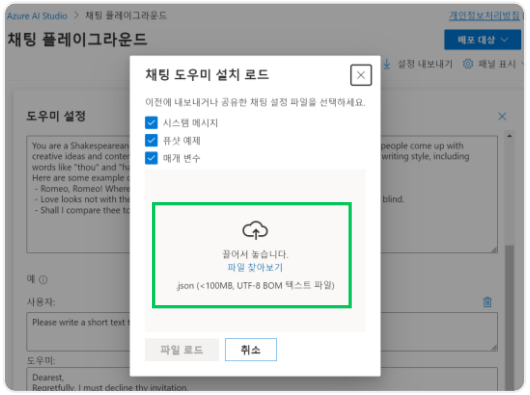
CLOVA Studio는 그룹 내 프롬프트 작업물 공유가 자유로워 타사 대비 공동 작업이 용이합니다

경쟁사 스튜디오 프롬프트 환경

Vertex AI (GCP) 프롬프트



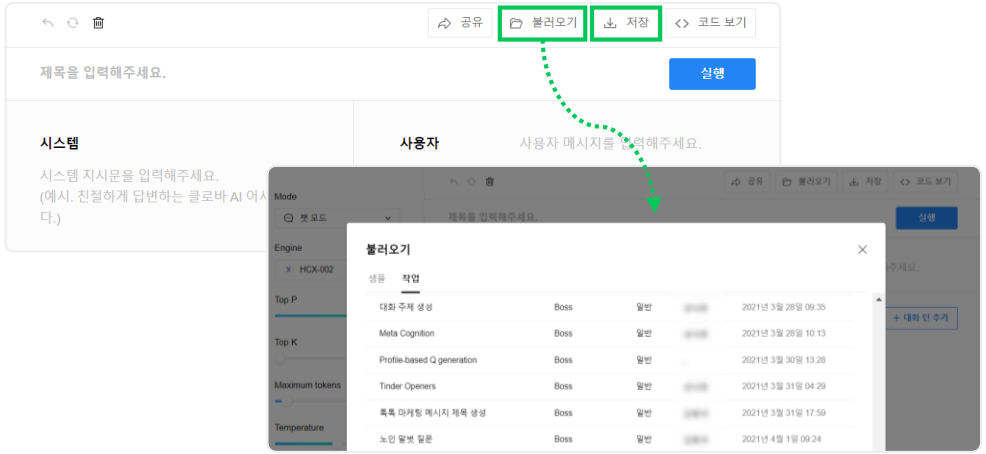
Azure OpenAI (MS) 프롬프트



타사 프롬프트 환경에서 다른 사용자의 프롬프트를 가져오거나, 자신의 프롬프트를 다른 사용자에게 공유하려면

특정 포맷의 파일로 프롬프트 파일 다운로드 / 업로드

CLOVA Studio 프롬프트 공유



CLOVA Studio는 프롬프트 환경에서 다른 사용자가 저장한 프롬프트를 바로 불러오고, 자신의 프롬프트도 저장할 수 있어

플레이그라운드 내 프롬프트 공유가 즉시 가능

익스플로러에서는 CLOVA Studio에서 활용할 특화 도구들을 제공하며, 기존에 작업한 예제를 확인할 수 있습니다

도구 | Studio에서 유용하게 활용할 수 있는 특화 도구 및 API

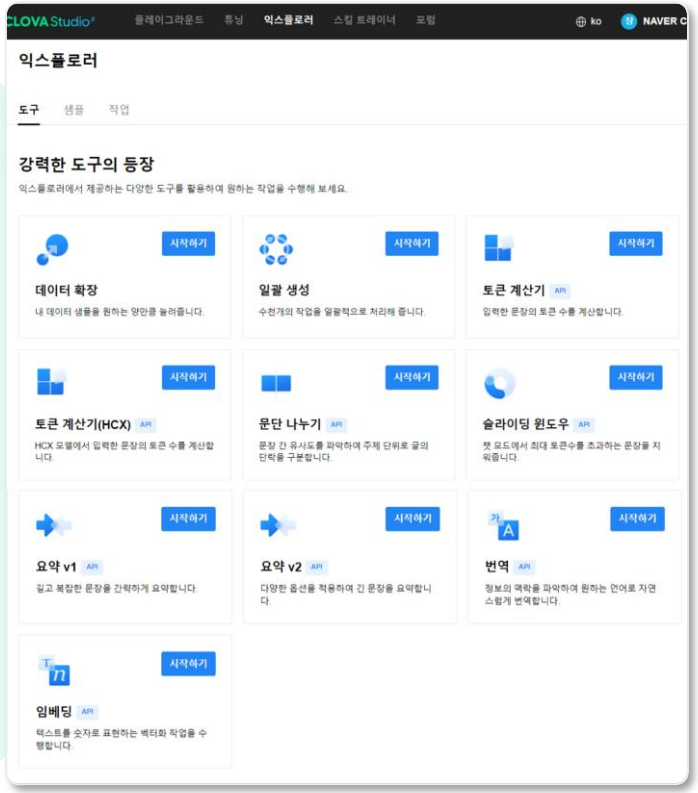
데이터 처리 도구
데이터 확장, 일괄 생성 등 사용자의 데이터 샘플을 학습에 맞게 가공합니다

슬라이딩 API
입력 토큰 수 초과 시 자동으로 초기 입력값을 제거해줍니다

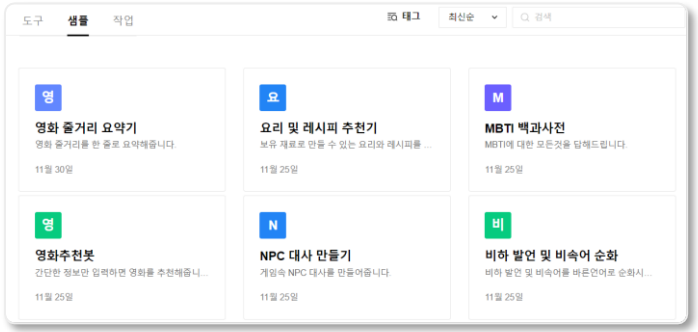
임베딩 API
입력한 텍스트를 숫자 형태로 변환하여 효율적으로 데이터를 저장, 활용합니다

토큰 계산 API
CLOVA Studio의 계산 단위인 토큰을 기준으로 입력 텍스트를 계산합니다

요약 API
입력한 문장을 간략하게 요약해줍니다



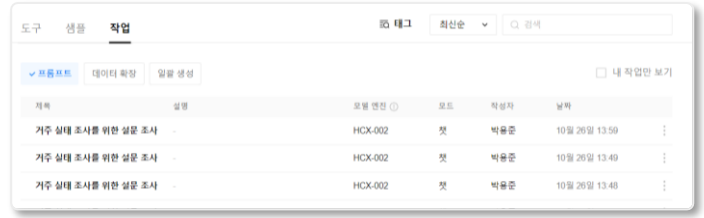
샘플 | CLOVA Studio에서 공식 제공하는 설정 및 프롬프트 예제



샘플
추천 예제를 통해 CLOVA Studio를 체험할 수 있습니다

작업 | 사용자가 저장한 도구 작업 및 프롬프트 예제

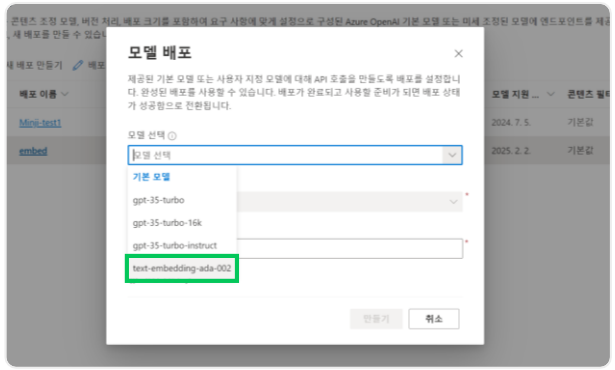
작업
그룹 사용자가 저장해 둔 데이터 작업 및 프롬프트 예제를 확인할 수 있습니다



다양한 개발 특화 도구를 제공하며, 특히 타사 대비 높은 성능의 Embedding API를 제공합니다

스튜디오 도구 접근성 비교

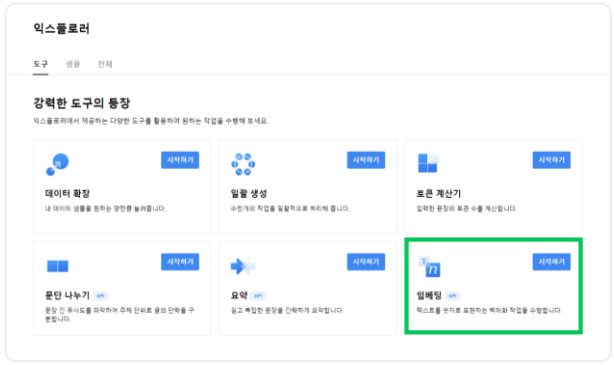
Embedding API 사용 상황



Azure OpenAI (MS) 임베딩

별도의 도구 모음을 제공하지 않음

모델 배포에서 기본 모델 선택 시 임베딩 모델로 선택하여 생성



CLOVA Studio

익스플로러 도구에서 확인 용이

도구의 임베딩 시작하기에서 안내에 따라 코드 확인 및 복사 기능 제공

Embedding API 성능 비교

1. 총 1,617건의 답변을 Naver Embedding과 OpenAI Embedding에서 추출 후 Opensearch에 저장
2. 테스트 데이터의 질문도 Embedding 값으로 추출하여 Opensearch에 저장된 답변 index에 검색
3. 검색된 답변 중 상위 10개를 추출하여, 정답 여부를 확인해 정확도 측정

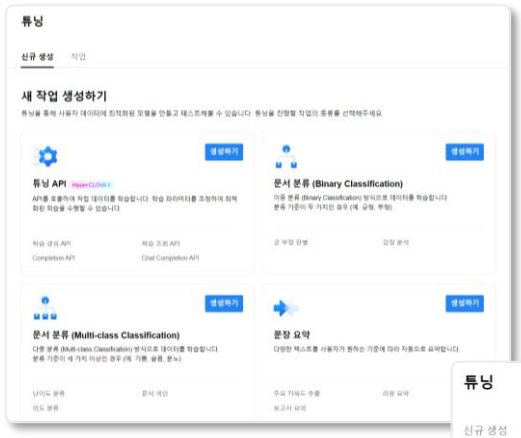
Embedding API	실험 시간	총 질문 건수	정답 건수	그외	검색 정확도
Azure OpenAI	3H 20M	49,850 건	37,234 건	12,616 건	0.7469
NAVER Cloud	1H 28M	49,850 건	42,869 건	6,981 건	**0.8599**

자사와 경쟁사 Embedding API 성능 비교 시

10% 이상의 정확도 차이를 보이며 높은 성능 확인

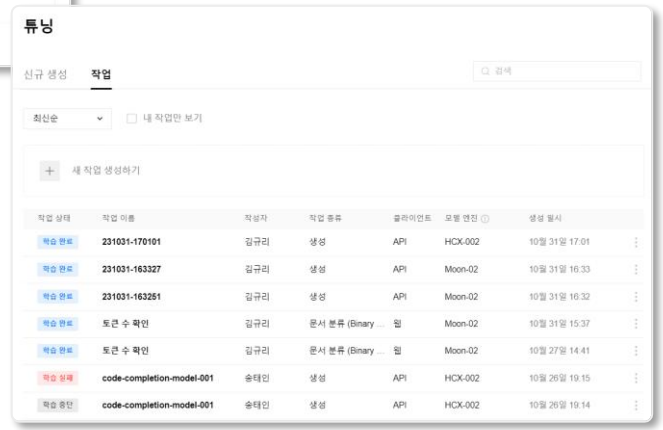
사용자의 데이터로 모델을 재학습하는 튜닝 기능도 활용 유형에 따라 손쉽게 적용할 수 있습니다

튜닝 | 사용자 데이터 기반으로 모델을 학습시키는 다양한 튜닝 종류 제공



튜닝

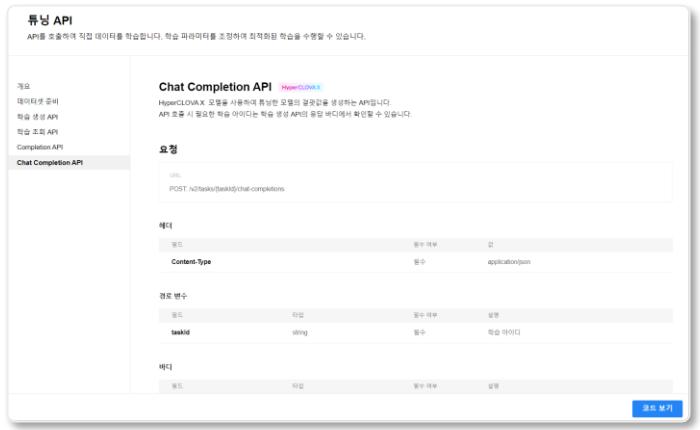
사전 학습된 모델 파라미터를 맞춤 변형, 재학습해
분류, 요약, 생성, 교정, 전환, 대화 등 7개 Task와 튜닝 API를 제공합니다



작업

그룹 내 진행된 튜닝의 상태와 관련 정보를 확인할 수 있습니다

튜닝 API | 학습을 위한 데이터 셋을 준비하고, API로 단계별 튜닝 진행



I. 데이터셋 준비 어떤 모델이 필요한 지 정하고, 학습할 데이터 셋 준비

II. 인증 ID 확인, 토큰 발급 API를 호출하여 토큰 발급

III. 학습생성 튜닝할 모델, task 유형, epoch 등을 지정하여 호출

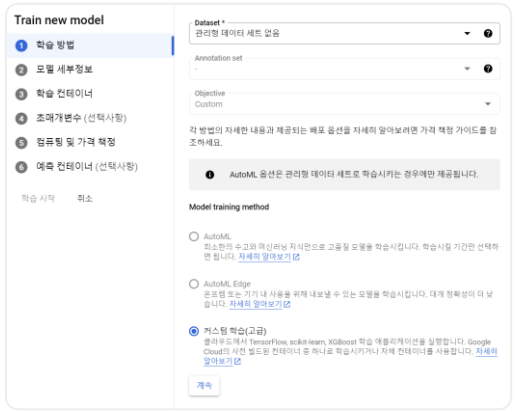
IV. 학습 조회 학습 ID 로 데이터 토큰 수, 학습 진행 상태, 학습 현재 스텝 수 등 확인

V. 인퍼런스 (Chat Completion API) 호출한 응답에서 해당 모델이 생성한 응답 확인

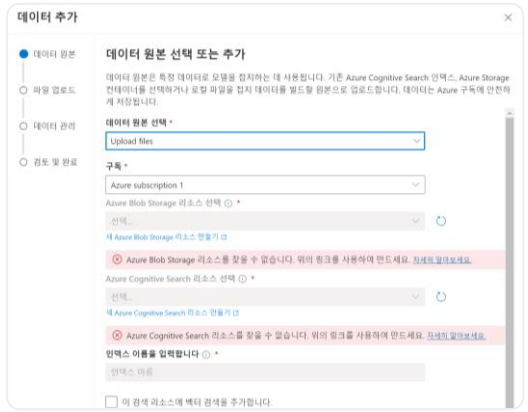
고객사 튜닝할 작업의 유형에 따라 학습 파라미터를 조정하여 최적화된 데이터 학습을 제공합니다

타사 데이터 튜닝 비교

Vertex AI (GCP) 튜닝 설정

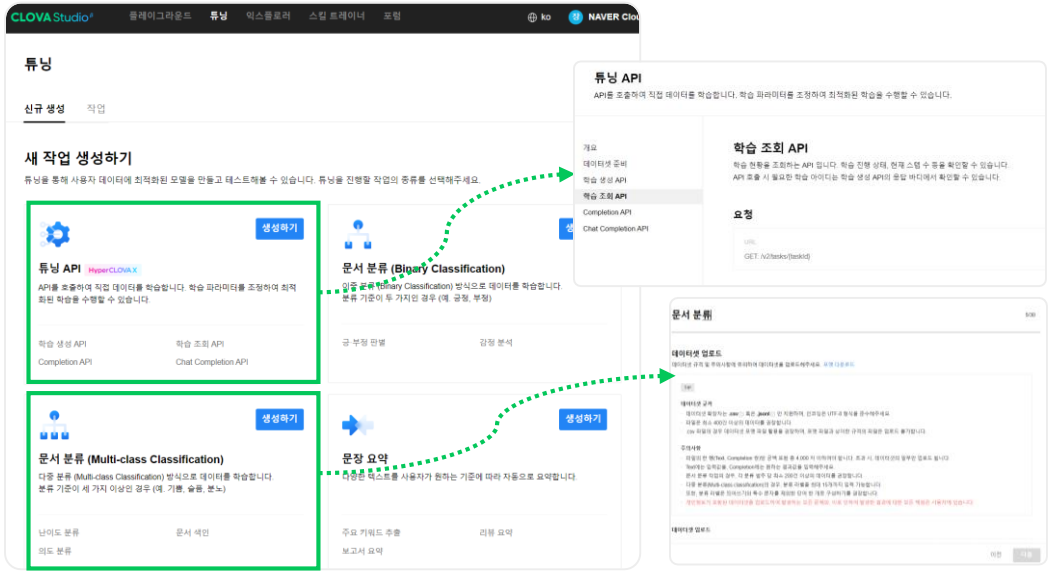


Azure OpenAI (MS) 튜닝 설정



튜닝을 진행하는 작업 유형에 관계 없이
특정 포맷의 파일로 데이터 파일 업로드

CLOVA Studio 튜닝



사용자가 튜닝을 하는 목적에 따라 튜닝 유형 선택
데이터 업로드 및 코드 활용하여 진행

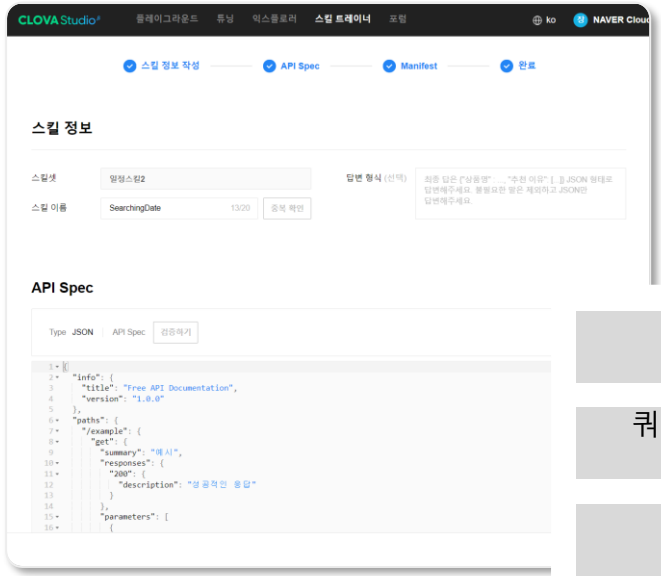
스킬 트레이너를 활용해 최신의 전문 정보를 제공하는 API를 모델에 연결, 학습시킬 수 있습니다

스킬셋 | 도메인 전문 지식과 기능 기반으로 답변하는 API 집합

스킬 트레이너 활용

도메인 특화 데이터 ····· UseCase, API 등 도메인 특화 데이터 학습 ····· 스킬셋 구축

스킬 생성 | 스킬셋에 포함되는 다수의 스킬(API) 제작



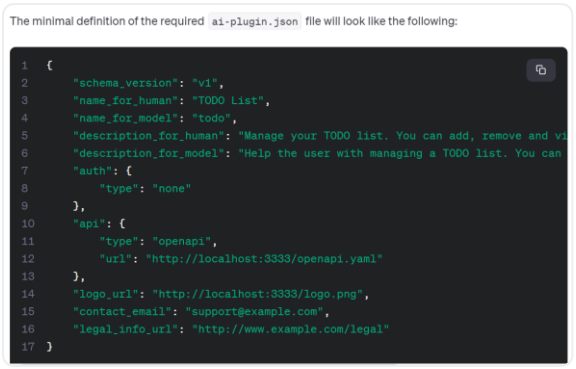
스킬 제작 과정

1. 예시 쿼리 입력
2. 쿼리에 대한 HyperCLOVA X 의 사고 과정 검토
3. 사고 과정 수정
4. 사고 과정에 활용할 도구 (API 정의)
5. 최종 답변 확인

스킬 제작을 위한 인터페이스도 함께 제공하며, 결과 확인 및 수정까지 모두 진행 가능합니다

스킬 제작 과정 비교

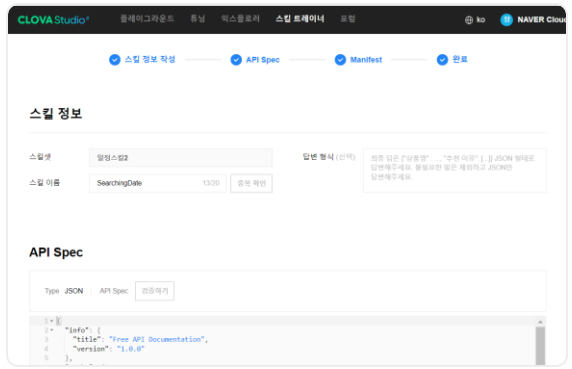
OpenAI 플러그인 생성 가이드



스튜디오 내 스킬 제작 관련 화면 x 가이드 문서 제공

스킬 제작에 필요한 json 파일 위주로 가이드 제공

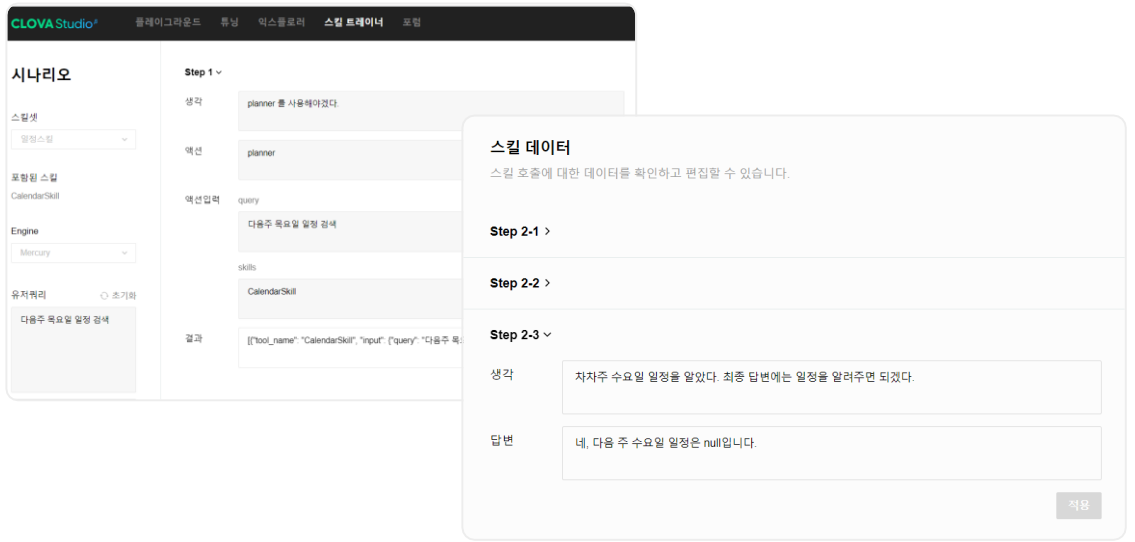
CLOVA Studio 스킬 생성 가이드



스튜디오 내에서 스킬 제작을 지원하는 가이드 화면 제공

스킬 제작 과정에 필요한 요소를 확인하기 쉬운 인터페이스 제공

시나리오 설정 및 수정 인터페이스



스킬의 액션이 필요한 과정을 스튜디오에서 직접 입력, 결과값을 확인하며 진행

구축 사례

1. 업무생산성
2. CS(Customer Service)
3. 마케팅 & 영업
4. R&D

[폴라리스오피스] 폴라리스오피스 AI

사업 개요

- 문서 편집기 + AI 글쓰기 도입

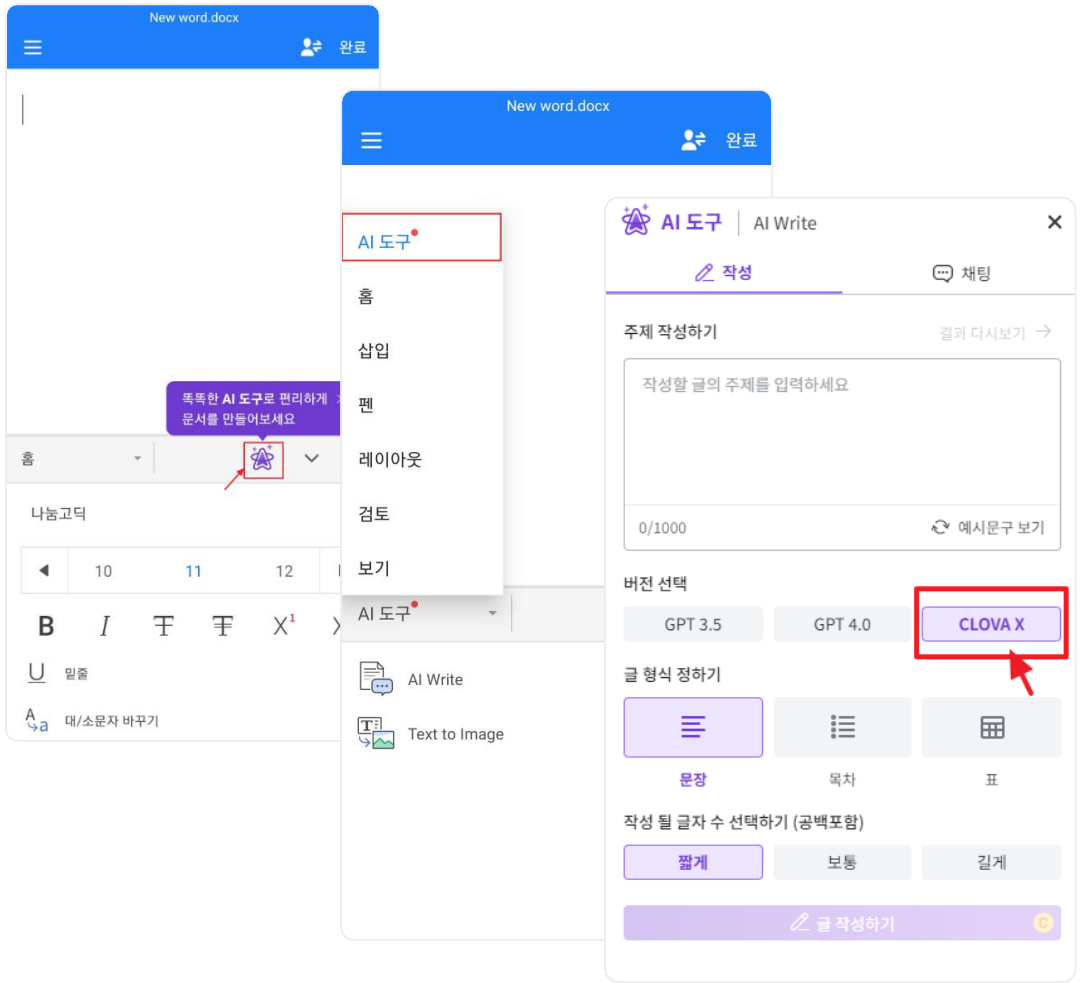
수행 목표

- 정해지지 않는 다양한 주제에 대한 AI 글쓰기 보조도구
- 작성하는 글의 형식을 문장, 목차, 표 형식으로 작성
- 작성하는 글자의 수를 다양하게(짧게/보통/길게) 작성
- 챗봇 형식으로 사용자의 요구사항을 빠르게 반영 및 작성

기대 효과

- 문서 편집기 사용자에 대한 글쓰기 보조 도구
- 글쓰기 초안을 작성해주어 시간 효율성을 제공
- 글쓰기에 대한 Ideation을 제공

예시 화면



출처: https://youtu.be/csjPBignee0?si=xMmwSJjyG2RmB78B

[한글과 컴퓨터] 한컴독스AI

사업 개요

- 문서 편집기 + AI 글쓰기 도입

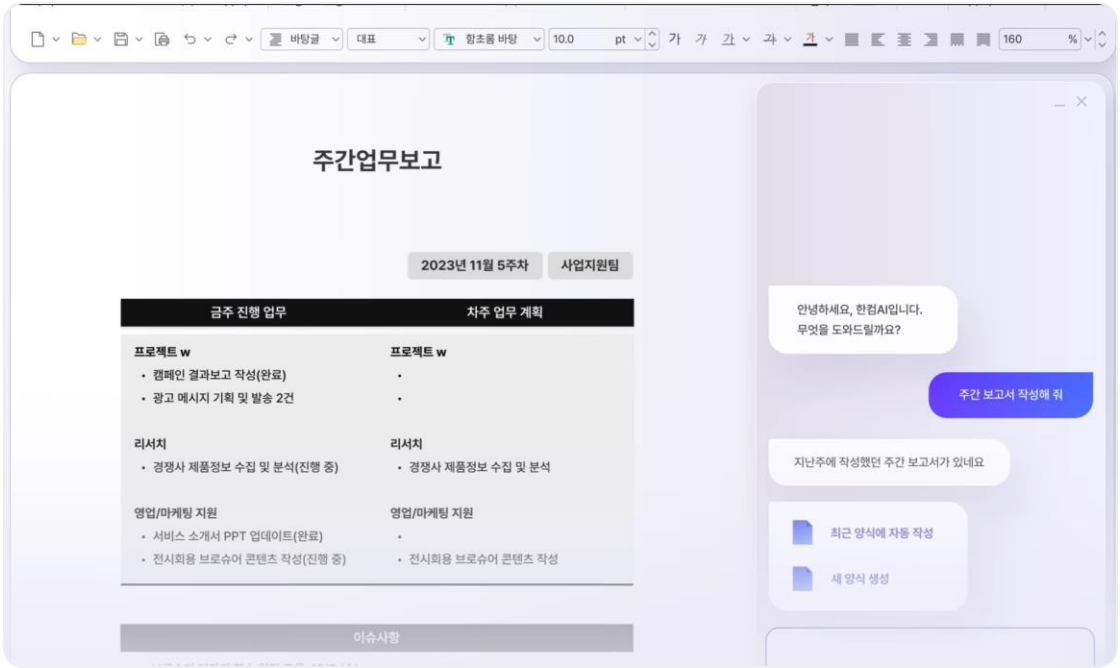
수행 목표

- 정해지지 않는 다양한 주제에 대한 AI 글쓰기 보조도구
- 사용자들이 많이 사용하는 양식 자동생성
- 예상 질문 만들기, 맞춤 글(목차, 초안) 생성, 문장 교정/문체변경/ 맞춤법교정/이어쓰기 제공
- 챗봇 형식으로 사용자의 요구사항을 빠르게 반영 및 작성

기대 효과

- 문서 편집기 사용자에 대한 글쓰기 보조 도구
- 글쓰기 초안을 작성해주어 시간 효율성을 제공
- 글쓰기에 대한 Ideation을 제공

예시 화면



출처: https://youtu.be/Qmr9jyp1M9U?si=1jCPSanwa6dd40-B

[사이냅소프트] 사이냅오피스

사업 개요

- 문서 편집기 + AI 글쓰기 도입

수행 목표

- 정해지지 않는 다양한 주제에 대한 AI 글쓰기 보조도구
- 작성하는 글의 형식을 문장, 목차, 표 형식으로 작성
- 작성하는 글자의 수를 다양하게(짧게/보통/길게) 작성
- 챗봇 형식으로 사용자의 요구사항을 빠르게 반영 및 작성

기대 효과

- 문서 편집기 사용자에 대한 글쓰기 보조 도구
- 글쓰기 초안을 작성해주어 시간 효율성을 제공
- 글쓰기에 대한 Ideation을 제공

예시 화면

[경상북도 교육청] 학교지원종합자료실 AI 지원 서비스

사업 개요

- 많은 정책 정보에 의한 교직원들의 다수 반복적인 문의
- 키워드 위주의 검색으로 많은 시간이 소요
- 변경되는 정보들에 대한 문의 접수 및 대응

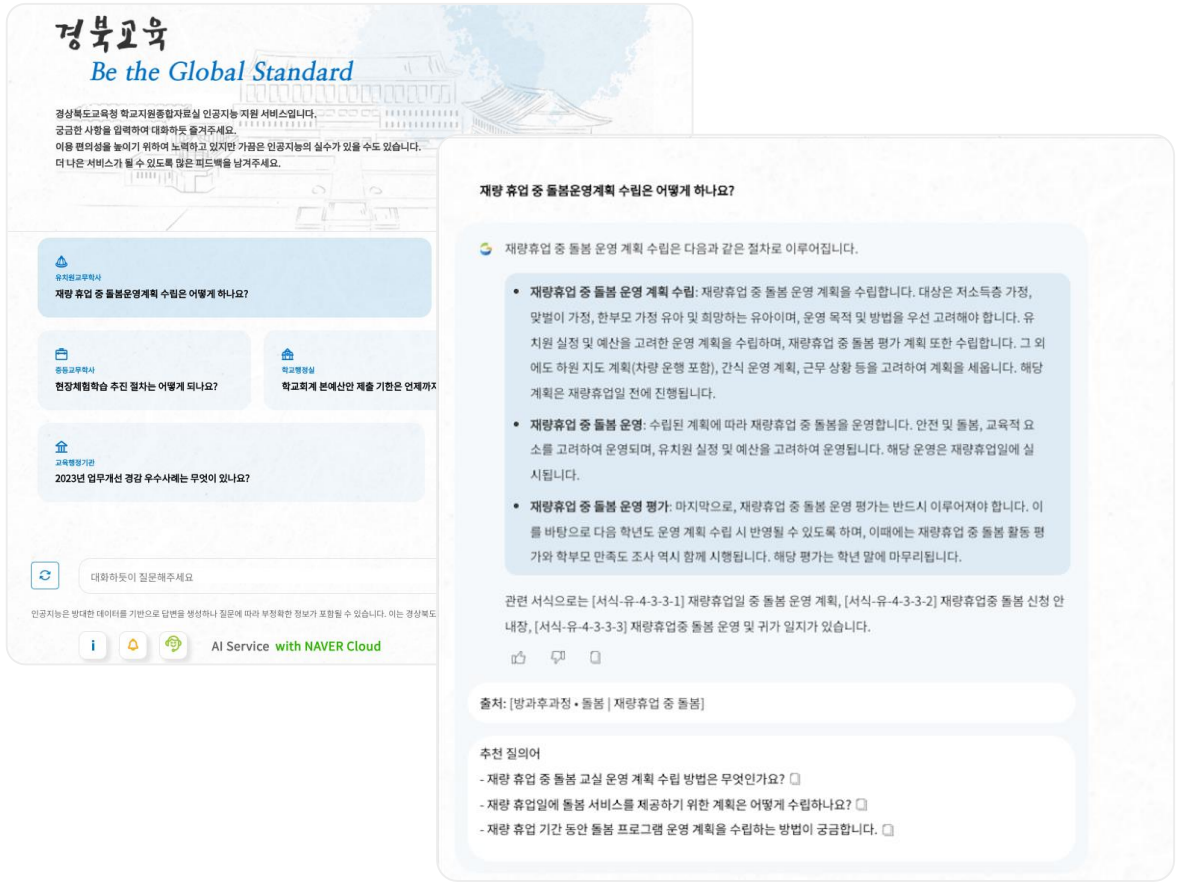
수행 목표

- 교육청 행정업무 지원 시스템의 디지털 전환
- 생성형 AI를 통한 정보 탐색 및 빠른 답변 제공
- 답변의 출처를 표기하여 신뢰도 향상
- 잦은 업데이트 정보에 대한 최신화

기대 효과

- 빠른 정보 탐색 및 신뢰도 있는 답변
- 최신 정보에 대한 답변

예시 화면



출처: https://ai.gbe.kr/chat

[SK C&C] tok. AI

사업 개요

- 생성형 AI를 활용한 기업분석 보고서 생성
- 할루시네이션을 최소화해야 하는 문제점
- 시장동향 및 경쟁사를 분석에 많은 인력과 시간이 필요

수행 목표

- 금융 산업에 특화된 어휘및 Instruction 데이터 대량 학습
- 금융 상품에 대한 사용자의 질문에 적절한 답변 제공
- 정확도가 높음 DART(전자공시)정보와 연동하여 생성
- 코드 인터프리터(데이터 분석 기능) 제공

기대 효과

- 산업에 특화된 지식을 습득한 AI 모델
- 정확도가 높은 답변 생성 가능
- 업무 생산성(시간, 인력)을 향상

예시 화면

[한국광고진흥공사] 아이작 (광고 카피 제작)

사업 개요

- 공공 홍보, 광고제작 실무자, 예비 광고인을 위한 카피 제작 서비스 필요
- 생성형 AI 활용을 통한 생산성 향상 고민

수행 목표

- 간단 키워드를 통해서 광고 카피 제작
- 헤드 카피/ 바디 카피 제작이 가능 하게
- 광고 카피의 톤앤 매너를 설정 가능하게
- 광고 카피의 저장하여, 마이 프로젝트로 관리

기대 효과

- 다수의 아이디어 광고 카피 제작 가능
- 빠른 시간 내에 제작 가능

예시 화면

[올거나이즈] 알리 LLM 인에이블러 연동

사업 개요

- 광고용 블로그 글쓰기에 어려움이 많은 소상공인, 인플루언서에 도움이 되는 서비스가 필요

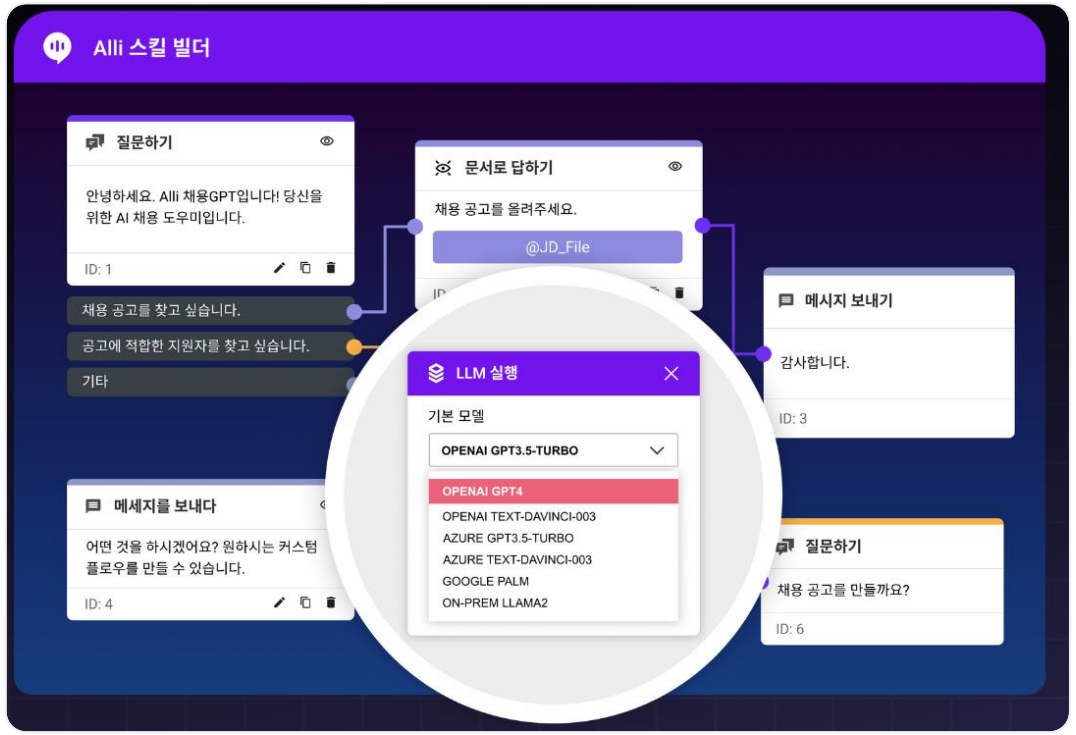
수행 목표

- 클릭 1번으로 1,500글자 이상의 블로그 글 생성
- 빅데이터 기반의 상위 노출 키워드 추천
- 50만 인플루언서 블로그, TV/SNS 광고 카피, 상품정보
- 완성도 높은 글을 생성

기대 효과

- 비용 절감 1/30, 시간 절감 1/10
- 평균 매출 증가 300%+a, 재결제 비율 93.3%

예시 화면



[위노보스] 가제트 AI

사업 개요

- 광고용 블로그 글쓰기에 어려움이 많은 소상공인, 인플루언서에 도움이 되는 서비스 필요

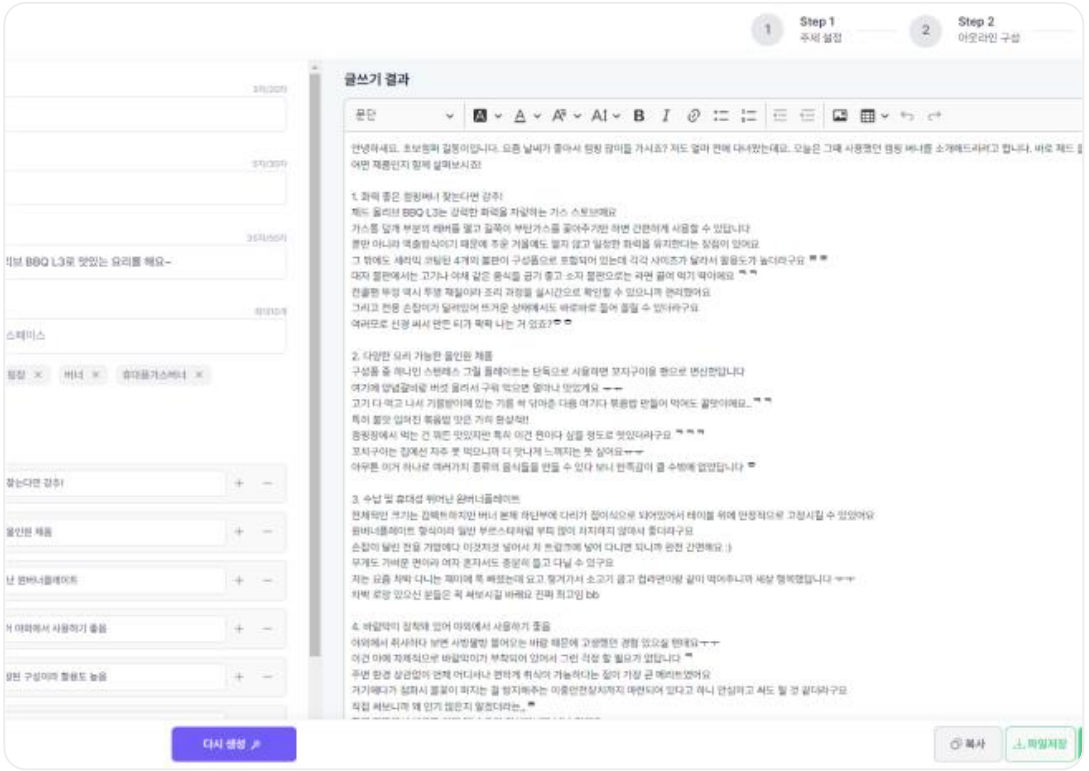
수행 목표

- 클릭 1번으로 1,500글자 이상의 블로그 글 생성
- 빅데이터 기반의 상위 노출 키워드 추천
- 50만 인플루언서 블로그, TV/SNS 광고 카피, 상품정보
- 완성도 높은 글을 생성

기대 효과

- 비용 절감 1/30, 시간 절감 1/10
- 평균 매출 증가 300%+a, 재결제 비율 93.3%

예시 화면



출처: https://youtu.be/4Smv_bLzT74?si=wMVDMAeWIB7adXvV

[셀렉트스타] 모모잼(모두 모여 재미있게)

사업 개요

- 데이터 구축 전문기업의 역량을 확대 필요
- 페르소나 챗봇에 대한 글로벌 인기 확산

수행 목표

- 다양한 페르소나에 대한 IP 확보 (친근하고 교육적인 페르소나 챗봇)
- 페르소나 별 대화셋 구축 및 전처리
- 페르소나의 데이터 내에서만 특성으로만 답변
- 누리과정(미취학아동)에 data 활용

기대 효과

- 교육적인 대화 효과
- 부모의 관리에 따른 올바른 이용

예시 화면

출처: https://youtube.com/shorts/Co48cKWDJMo?si=Papz3f_9rO6r3AM-

구축 방법

1. 파트너
2. 역할
3. 주요 Task
4. 참조 아키텍처

LLM 도입 목적과 구축 방향에 적합한 파트너사를 찾을 수 있도록 다양한 파트너사들과 제휴하고 있으며, 이들을 체계적으로 지원하고 있습니다

네이버클라우드 파트너

AI/LLM

- AI 및 LLM 관련 컨설팅, 서비스 개발 역량 보유

crowdworks

pcn

COGNET9

Cloocus

SI/MSP

- Sales, 서비스 개발/구축 역량 및 리소스 보유
- HCX 및 NCP 상품 딜리버리

SK 주식회사 C&C

HYUNDAI IT&E

KP CNC (주)경포씨엔씨 Kyungpo Computer & Communication

신세계 아이앤씨

산업별 특화

- 산업/도메인 특화 데이터 및 비즈니스 노하우 보유
- 데이터 제공 및 신사업 추진

법률 律村 YULCHON Attorneys at Law YULCHON NEXUS AI

의료 URBAN A FIT IT's Smart! IT smart

교육 DNSOFT ARTIFICIAL SOCIETY

엔터프라이즈 솔루션

- Biz Function 특화 솔루션 또는 플랫폼 보유

HANCOM 한글과컴퓨터

POLARIS OFFICE

kore.ai

CS Sharing

NAVER Cloud

교육지원
- 상품 / 영업 / 기술 등 다양한 방면으로 파트너 교육 지원

비용지원
- POC Credit 및 Edu Credit 지원
- 벌크 계약시 할인 제공 등

사업지원
- 마케팅 및 홍보 지원
- 정기 미팅 셋업하여 요구사항 청취 및 반영

성공적인 LLM 사업 수행을 위해서는 LLM 제공자와 그 외 파트너 및 고객 영역에 대한 이해가 필요하며,이를 기반으로 선제적인 준비가 필요합니다

고객사/파트너사 영역
(LLMOps)

- Prompt Engineering & Fine Tuning Dataset
- RAG
- Evaluation
- Cahing
- Guardlails
- Chat Interface
- User Feedback

X

LLM 제공사 영역
(LLM Platform 제공)

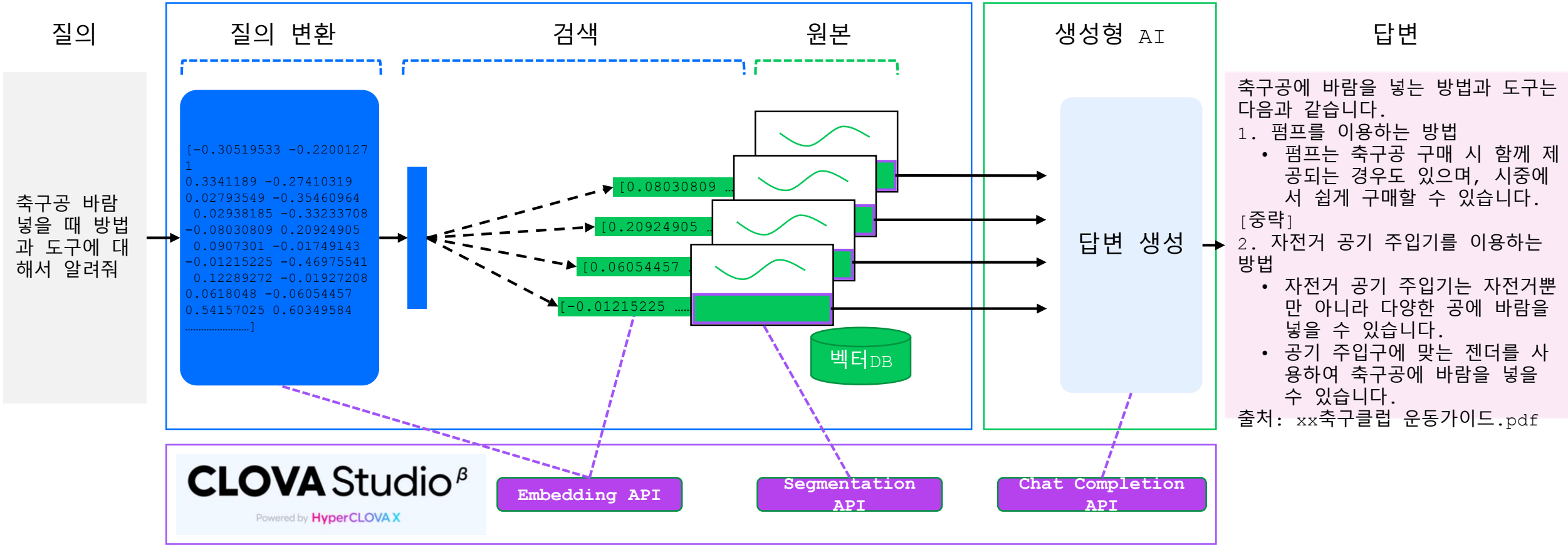
- Testing (Model Evaluation)
- Fine Tuning (PEFT, SFT, RLHF)
- Prompt Playground
- LLM (Foundation Model)
- Deployment (inference Infra)

각 단계별 고객사, 파트너사, 네이버클라우드의 주요 Task를 정의하고, 고객은 이를 참고하여 생성형 AI 도입에 대한 계획을 수립할 수 있습니다

주요 Task

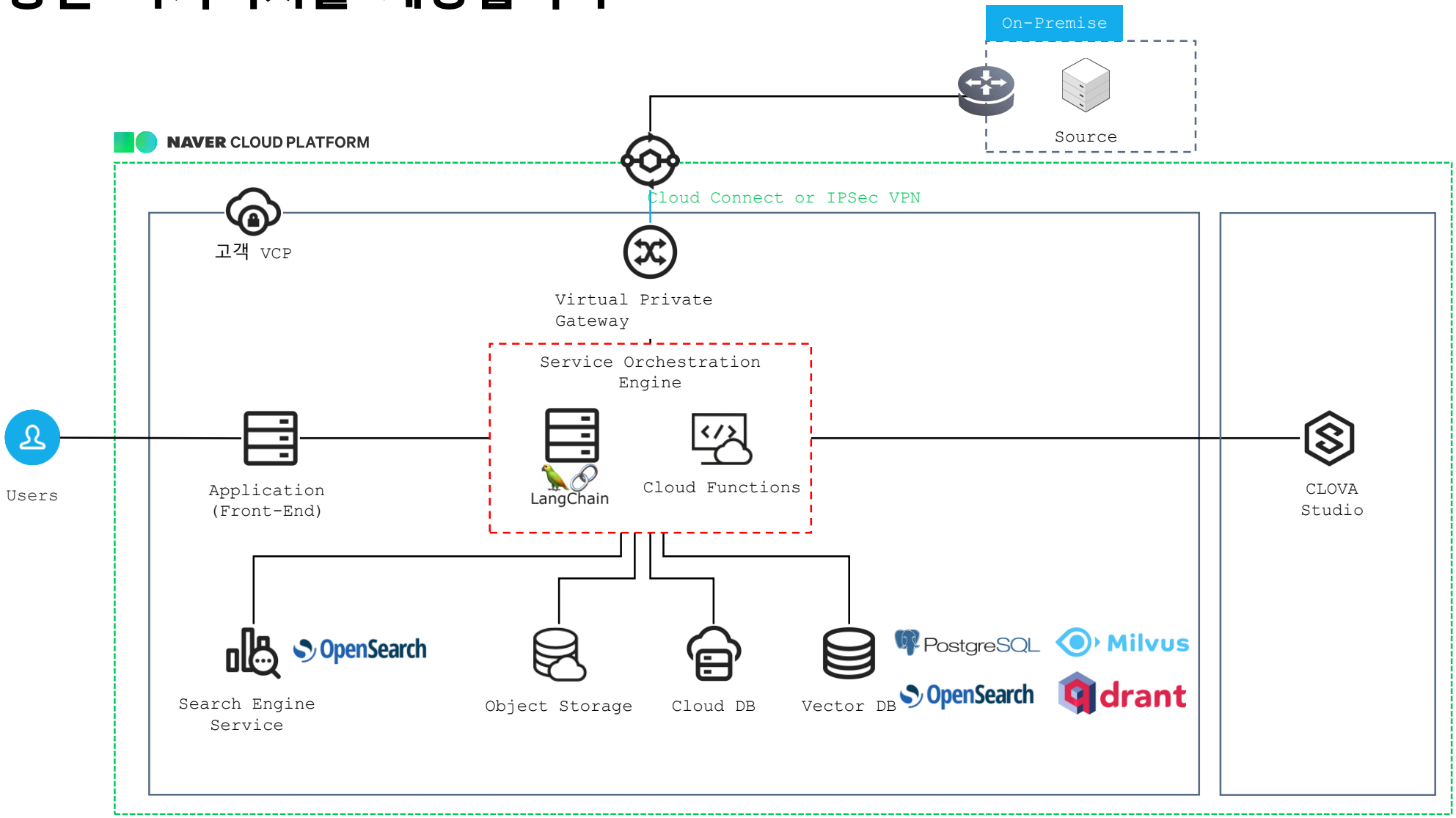
	시장 및 유스케이스 식별	필요 기술 정의	아키텍쳐 구성	POC	LLM 도입	유지보수
고객사	▪ 각 사에서 가지고 있는 Painpoint 정의와 LLM 활용 케이스 시장 조사 ▪ 서비스 환경 정의	▪ AI 자체 개발시 고객사에서는 서비스/솔루션 내 LLM 도입을 위한 LLM 활용 능력 및 필요 기술에 대한 구체화 필요	▪ 검증하고자 하는 LLM 및 솔루션(서비스) 정의	▪ LLM 실서비스에 도입	▪ 솔루션 및 서비스 고도화 플랜	
파트너	▪ LLM 적용 사례 및 트렌드 공유	▪ 식별된 유스케이스 기반으로 필요한 LLM 기술 컨설팅 ▪ 파트너 자체 솔루션 활용 및 인프라 관점에서 필요 기술 제안	▪ 고객사 도입을 위한 전 단계의 아키텍쳐 제공 (클라우드/온프렘/하이브리드 환경 등)	▪ POC 수행 및 기술 현실화		▪ 장애 및 유지보수 대응 ▪ 서비스 디벨롭
NAVER CLOUD	▪ LLM 적용 사례 및 기술 베이스로 유스케이스 현실 가능성 판단	▪ LLM 기술 보유 솔루션 파트너사 풀 제공 ▪ Cloud 환경에서 제공 가능한 범위 확인	▪ 구체화된 유스케이스 기반 클라우드 환경에서 구현가능한 예시 아키텍쳐 레퍼런스 제공	▪ POC 크레딧 지원		▪ LLM 성능 개선 ▪ LLM 구축 및 적용을 위한 파트너 풀 확대

네이버클라우드는 Hallucination 최소화하고, 사용자의 질의에 대해 최신 정보 기반의 답변 생성을 위해 RAG(검색증강생성)를 활용합니다



- RAG는 LLM 재학습 없이 최신 데이터 원본을 활용할 수 있도록 함으로써 LLM의 품질을 향상시킬 수 있는 기술입니다.
- RAG 모델은 회사내 자체 데이터를 기반으로 지식 저장소를 구축하며, 저장소는 지속적으로 업데이트될 수 있습니다.
- RAG를 구현하려면 새로운 데이터를 신속하게 embedding 하고 해당 데이터를 검색할 수 있는 벡터DB 기술이 필요

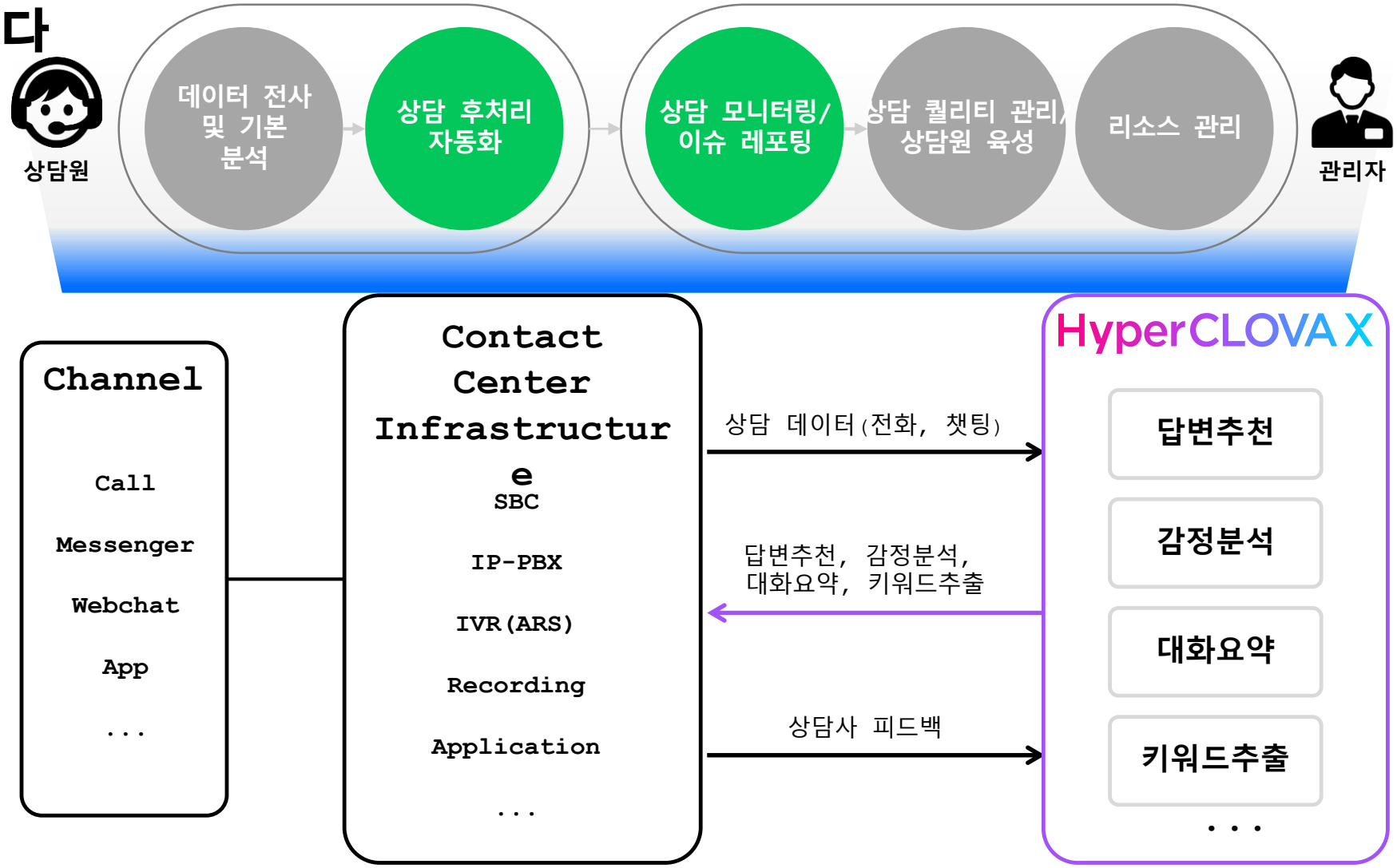
네이버클라우드는 seamless한 RAG 구축을 위해 다양한 클라우드 상품과의 연계를 통한 아키텍처를 제공합니다



아키텍처에 포함된 컴포넌트의 기능은 아래와 같습니다

컴포넌트	기능	컴포넌트	기능
Application (Front-End)	▪ 챗팅 화면 ▪ 사용자 질의 전송 → SOE ▪ 답변, 출처 등을 챗팅 화면에 표시 ▪ 출처 클릭 시 원본을 화면에 표시	Search Engine Service	▪ 키워드 검색
LangChain Cloud Functions	▪ Source에서 Text 추출 ▪ Source에서 Metadata 추출 ▪ Retriever(검색 기능) ▪ 질의 분석 요청 → Question decomposition API ▪ 질의를 Vector 변환 요청 → embedding API ▪ 추출된 Text를 문단나누기 요청 → segmentation API ▪ 나누어진 문단을 Vector 변환 요청 → embedding API ▪ 검색키워드 요청 → 검색키워드 추출 API ▪ 키워드 검색 요청 ▪ 검색결과 비교 및 조건 필터(권한, 날짜 등) ▪ 답변 생성 요청 → 답변 생성 API	CLOVA Studio	▪ Question decomposition ▪ 질의를 Vector 변환 ▪ 추출된 Text 문단나누기 ▪ 나누어진 문단을 Vector 변환 ▪ 검색키워드 추출 ▪ 사용자 질의 맞추어 답변 생성
Object Storage	▪ Source 저장	Cloud DB	▪ Metadata(권한정보, 날짜정보, 위치정보 등) 저장
Vector DB	▪ 나누어진 문단 및 Vector 저장		

네이버클라우드의 컨택센터 솔루션은 원하는 기능과 예산에 따라 기업 환경에 필요한 솔루션을 선택할 수 있으며, HyperCLOVA X로 상담사와 관리자를 지원할 수 있습니다



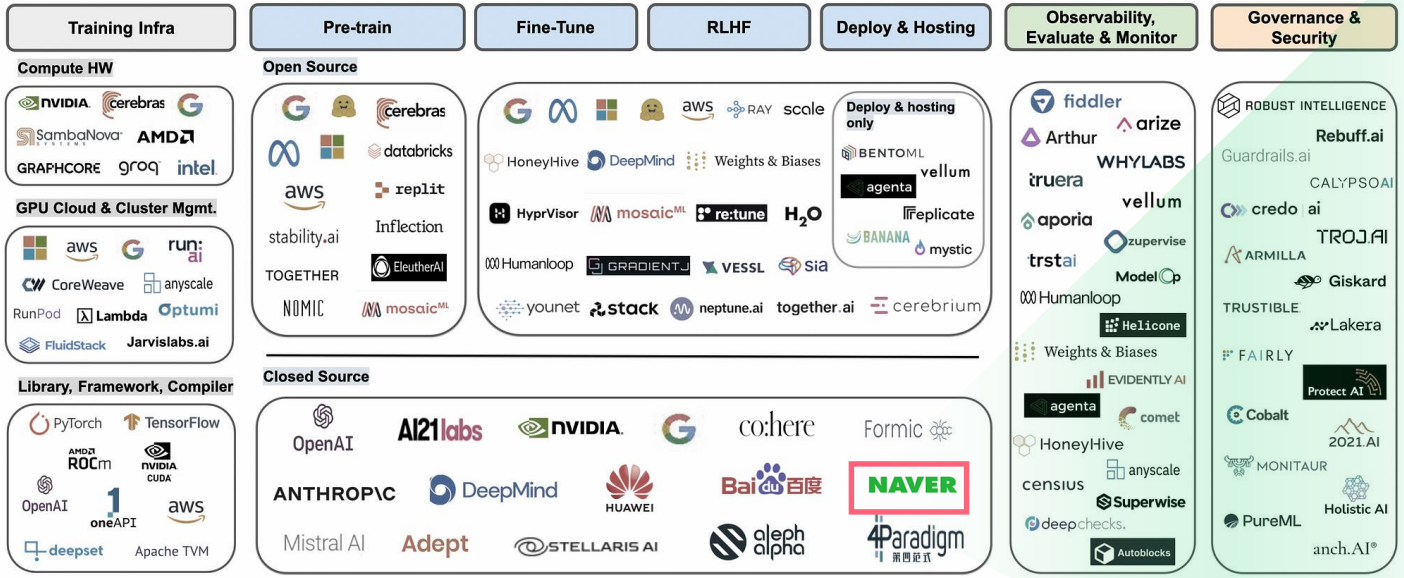
/

It will be updated

참고

1. 네이버 역량
2. LLM 연대기
3. OpenAI 비교

네이버의 AI 역량은 해외에서도 인정 받고 있습니다



KBS · 2023.03.10. · 네이버뉴스

제타알파 "**인공지능 연구 영향력** 1위 오픈AI...**네이버 6위**"

매경이코노미 PICK · 2023.03.10. · 네이버뉴스

AI **연구 영향력 세계 6위**...인텔·구글 제친 국내 회사가 있다고?

디지틀조선일보 · 2023.04.05.

네이버 '클로바 케어콜', HCI 학회서 베스트 페이퍼 어워드 수상

ITWorld · 2023.06.13.

네이버클라우드, 2023년 상반기 글로벌 AI 학회에서 61건 논문 채택

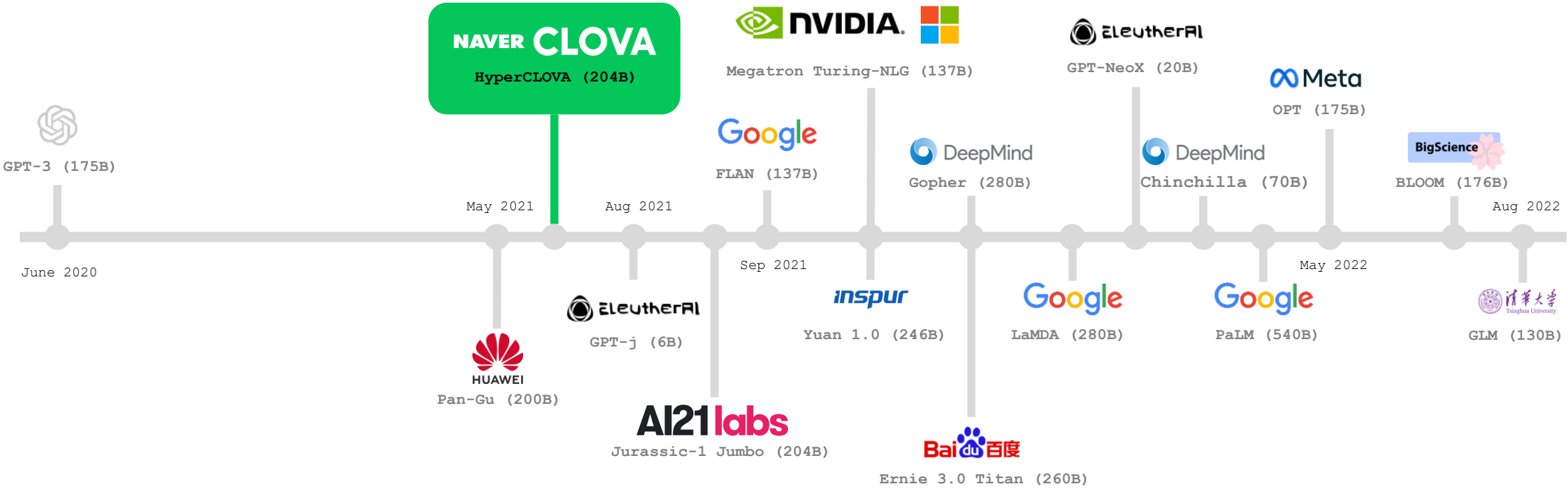
techcrunch.com › koreas-internet-giant-naver-unveils-generative-ai-services

Korea's internet giant Naver unveils generative AI ser... 번역보기

2023.08.24. **AI Korea's internet giant Naver unveils generative AI ser vices** Kate Park @kateparknews / 5:59 PM GMT+9•August 24, 2023 Com ment Image Credits: Nako Sung, **Naver** Cloud Head of Technology, Hyp...

출처: https://liu-gendary.medium.com/llm-explained-the-llm-training-landscape-82c803495caa

네이버클라우드는 세계 3번째로 초거대언어모델(LLM)를 보유하고 있습니다



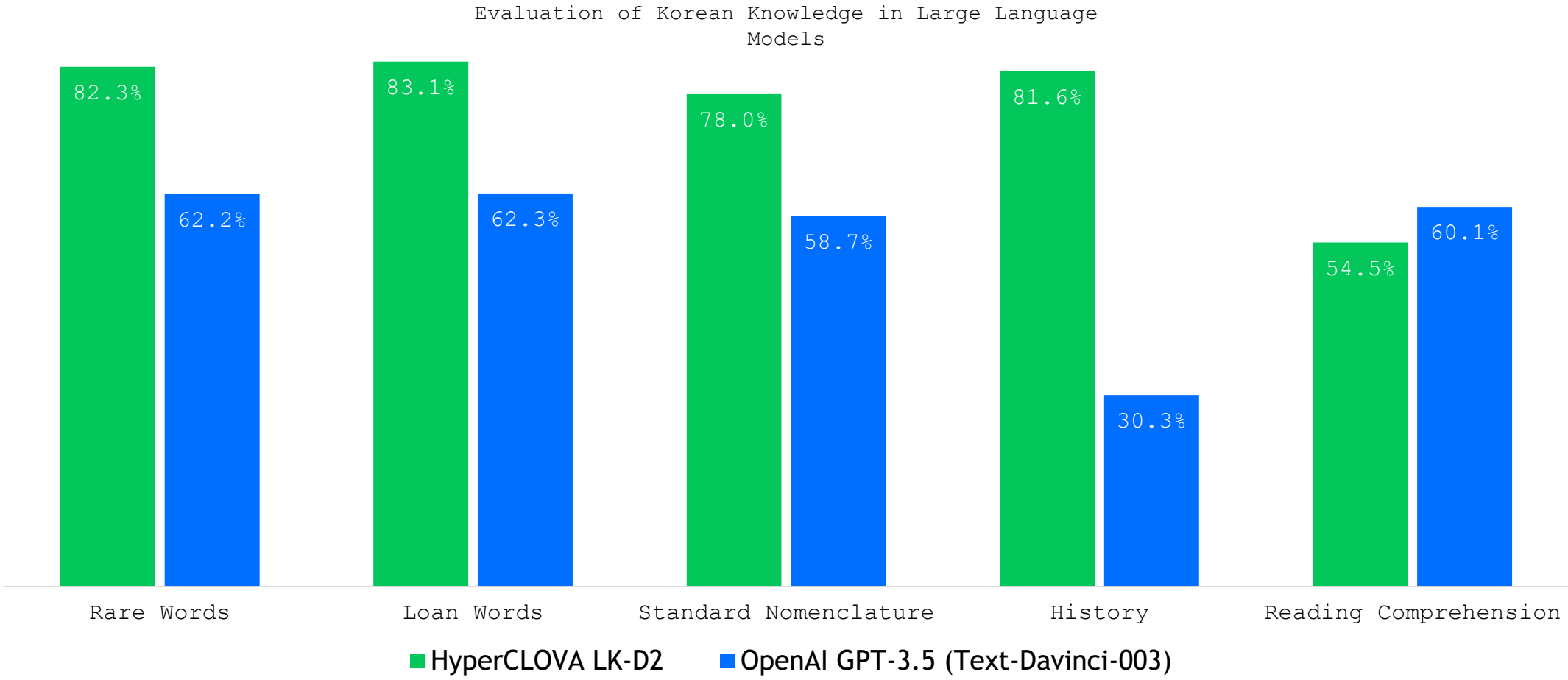
출처: https://docs.google.com/presentation/d/1WrkeJ9-CjuotTXoa4ZZIB3UPBXpxe4B3FMs9R9tn34I/

한국어 생성 LLM을 도입하려는 기업은 운영비까지 고려해야 하며, 그런 경우 OpenAI는 비효율적이며, 네이버클라우드는 OpenAI가 가진 기능을 계속 추가하고 있습니다

구분	HyperCLOVA X	GPT 3.5 turbo	GPT 4
토큰 당 가격[1)]	0.005	0.0084	0.216
언어모델 성능	한국어 성능 우수 영어: MMLU 70점	영어: MMLU 70점	영어: MMLU 86.4점
컨텍스트 사이즈(토큰)	4K, 16K(24년 2Q 예정)	4K, 16K	8K, 32K
튜닝 기능	PEFT SFT / RLHF	PEFT	N/A
멀티모달리티	이미지 인식(24년 2Q 예정)	N/A	이미지 인식
임베딩	한국어 임베딩 검색 정확도 높음	한국어 임베딩 검색 정확도 낮음	한국어 임베딩 검색 정확도 낮음
외부 API 연계	스킬	Function calling	Function calling
한국어 토큰 처리 효율성	평균 1토큰당 2글자	평균 1토큰당 0.5글자	평균 1토큰당 0.5글자
보안	CSAP, ISMS, ISMS-P 등 보안인증 취득	불가	불가

1) 한국어 생성하는 경우 GPT는 영어 생성 보다 347% 더 많이 사용하는 것을 반영한 가격

HyperCLOVA X 는 한국의 사회, 문화 및 법 제도 등을 가장 잘 이해하는 LLM 모델입니다



출처: HAE-RAE Bench May-2023

감사합니다.

문의: 오정식(jungsik.oh@navercorp.com)