AI 기술 브로셔(배포용)

# Llama K

기술혁신부문 Tech Design Center

# 목차

01

# Why Llama K

공개 LLM 모델 동향

KT의 선택

Llama K 개요

# 뛰어난 성능의 상용 LLM도 취약점이 존재

## 01 비용이 기하급수적으로 증가


상용 LLM은 API 호출당 비용이 발생하여
대규모 데이터 처리나 반복적 작업이 필요한
경우는 비용이 기하급수적으로 증가

## 02 제한된 커스터마이징

상용 LLM 모델은 제한된 커스터마이징
옵션을 제공하며, 특정 도메인에 맞춘
튜닝이 어려움

## 03 데이터 프라이버시가 불안

상용 LLM을 사용할 경우
데이터가 외부 서버로 전송

*SOTA K는 Secure Public Cloud 활용으로 데이터 보안을 강화

# 공개 LLM 모델로 비용 효율화와 도메인 최적화를 추진

## 01 비용 부담 없이 모델 활용

공개 LLM은 라이선스 비용없이
무료로 제공되므로,
중소기업 및 스타트업이
추가적인 비용 부담없이
모델을 활용하고
자체 인프라에 배포 가능

## 02 자유로운 수정 및 맞춤화

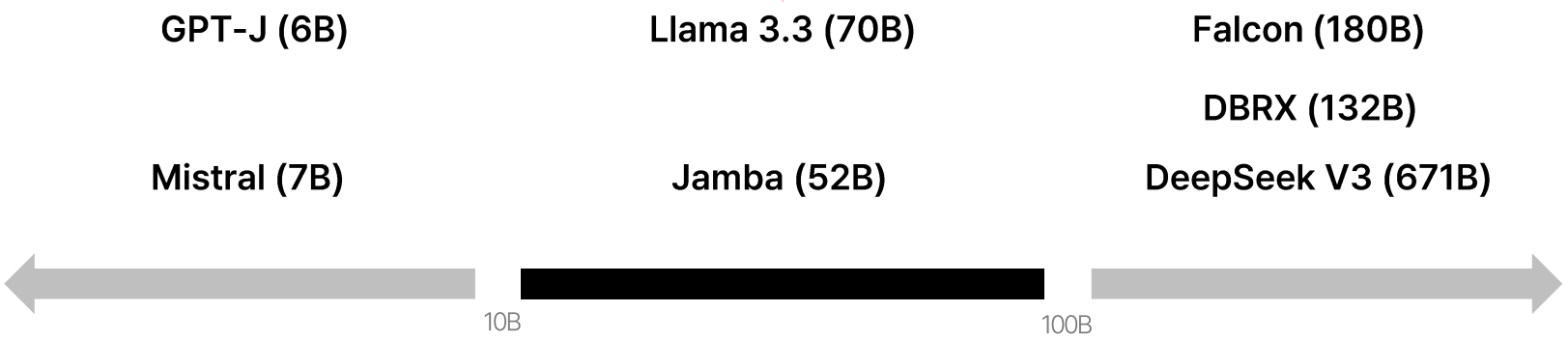
소스 코드와 모델 가중치가
공개되어 있어
연구자나 개발자가 모델을 분석하고,
품질을 평가하며,
맞춤형 변형이 가능

## 03 데이터 프라이버시와 보안

공개 LLM을 자체 구축하면
데이터가 외부로 유출되지 않음.
금융/의료/법률과 같은
민감한 데이터를 다루는 기업에서는
필수적인 요소

# 수많은 공개 LLM 모델 중 Llama 3.3(70B) 선택

**높은 성능과 비용 효율성**을 가진
중간 규모의 모델로
Llama 3.3(70B) 선택



## 우리 기업에 맞는

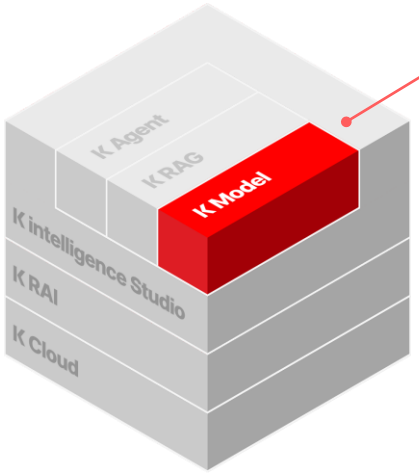
# Vertical Model을 만들기에 최적인 Llama K

02

# Llama K 소개

Llama K 상품 정의
타겟 고객 및 Pain Point
도입 기대 효과
경쟁력 및 차별화 포인트

# Introducing Llama K




K intelligence Suite

KT의 자체 데이터 및 학습 노하우를 적용해
원본모델의 성능은 유지하면서도
Meta의 Llama가 지원하지 않는 한국어 능력을 대폭 강화한 비용 효율적 AI 모델

산업별 요구에 유연하게 대응할 수 있도록, 공개된 오픈소스 모델을 기반으로 실용적으로 설계하여 유연한 적용과 경제성을 고려하였습니다.

[K Model Line-up]

| 믿:음 K 2.0 | | Llama K | | SOTA K |
|---|---|---|---|---|
| KT 자체 기술 기반 AI 모델 | + | 유연한 적용과 경제성을 고려한 오픈소스 기반 | + | SOTA 성능에 한국적 데이터 정밀 튜닝 |

# 무엇이 우리 회사 AX를 어렵게 하고 있나요?

금융 그룹
DT 추진팀 팀장

그룹의 민감한 데이터와 고객 정보를
안전하게 다루면서도
규제를 준수하는 새로운 금융 상품
아이디어를 발굴해야 해.

수십 년간 축적된 방대한 양의 정책
보고서, 연구자료, 국회 회의록 등..
단순 자료 검색이 아니라, '맥락에 맞는'
핵심 내용을 찾는 것이 어려워.

국책 연구기관
책임연구원

## Llama K는 이런 분들을 위해 만들어졌습니다.

**01** On-Prem 또는 Private Cloud 환경에 직접 모델을 설치하여 도메인 특화한 전문 AI를 구축하여, 데이터 기반 고객 혁신을 가속하고자 하는 B2B2C 고객

**02** 생성형 AI를 업무에 적용할 때 자체 윤리 기준에 벗어나는 환각 현상에 민감하며 한국 공공기관의 언어와 맥락이 고려되어야 하는 B2G 고객

**03** 방대한 양의 한국어 텍스트를 단순 키워드 검색이 아닌 의미 기반으로 이해하고 분석하여 여러 문서에 흩어져 있는 정보를 종합하고 핵심적인 인사이트 기반으로 정교한 정책을 수립해야 하는 국책 연구기관

## Llama K를 선택하신 고객님들은

# 아래의 네 가지 가치를 모두 누리실 수 있습니다.

Llama K

### 01 높은 추론 성능 기반 인사이트 제공

다양한 데이터의 의미를 정제하고 추론하여
단순 생성이 아닌 인사이트 제공형 업무에 활용
지시어 기반 분류, 정리, 추론 업무에 활용

### 02 공공기관 윤리 기준 준수

높은 수준의 KT Responsible AI 기준
학습 데이터의 개인정보/유해 표현 제거,
환각 현상 감소

### 03 범정부 업무 맞춤형 문서 능력

기관별 특화된 문체를 반영한
문서 요약 및 보고서 초안 작성
Q&A 생성 등 전문적/한국 조직형 문서 처리

### 04 세부 모델 라인업 선택권

일반적인 성능의 기본 모델 Llama K 11B와
RAG 등 Enterprise급 성능 요구에 적합한
고성능 모델 Llama K 74B 제공

## 글로벌 오픈소스 모델은 한국어를 잘 하지 못한다?

# KT Llama K는 다릅니다.

### 01 Llama 원본 모델의 성능은 그대로 유지하고, 한국어 이해/생성 능력은 획기적으로 강화

원본 모델의 지식 보존 위해 KT의 노하우가 담긴 독자적인 학습 기법 사용 및 한국어와 한국 문화를 담은 고품질의 데이터 선별 활용

### 02 한국어 지시 이행 (Instruction-Following) 성능 대폭 향상

Meta 원본모델이 상대적으로 성능이 낮았던 한국어 지시 이행 영역 에서 비약적인 성능 향상 달성

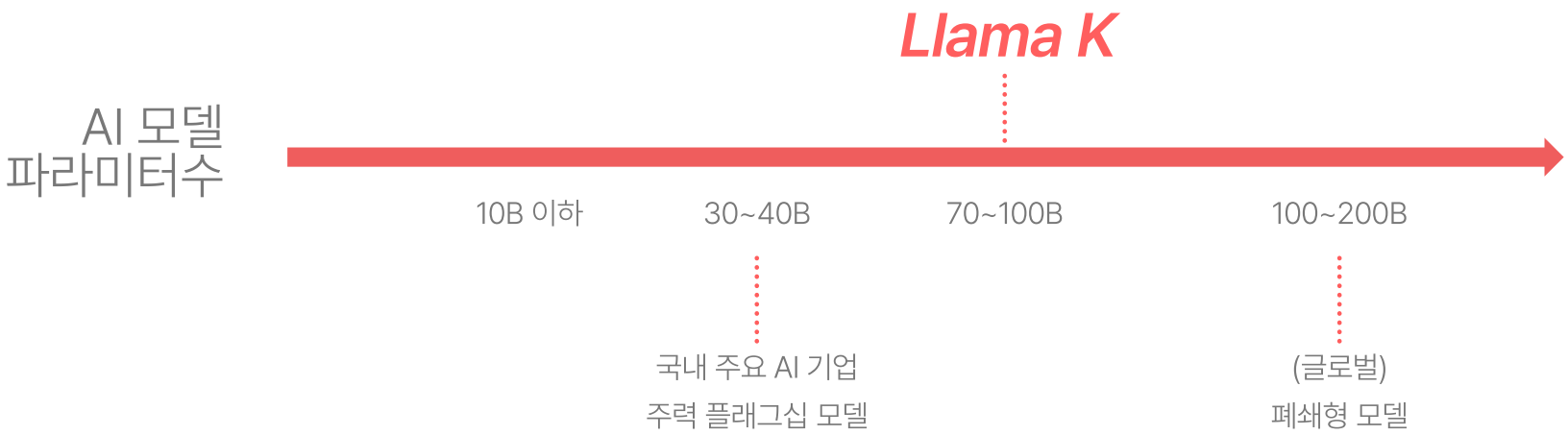
높임법 종결어미 등 한국어 특화 영역과 공공분야 포맷 이행 영역 적용

### 03 한국어 Reasoning 코딩, 수학 영역 최적화 학습 기법 적용

KT 독자적인 한국어 리워드모델 기반 최신 Online Reinforcement Learning 기술 활용

## 비용 효율적으로 운용 가능한

# 파라미터 70B 전후 한국어 모델



03

# LIama K 주요 기능

기술 구조 및 사양
주요 기능 및 성능 지표
도입 환경 및 연계 요건

## 128K Context Length를 지원하는 74B 규모의

# Transformer 구조 Auto-regressive 언어모델

Llama 3.3 70B를 기반으로
Block Expansion을 통해 원본 지식을 유지하며 한국어 지식을 확장(Continual Pretraining) 후
한국어와 영문 데이터를 혼합하여 기존 Instruction Skill을 보존하며 한국어 Instruction 학습(Supervised Fine Tuning)
이후 모델의 성능과 사용성을 높이기 위한 Multi-Stage RL 진행

**기존 Llama 3.3 70B의 성능을 최대한 보존하며 한국어 능력을 강화한 74B 모델을 생성**

**원본 지식 보존**
**한국어 지식 주입**

Catastrophic Forgetting
최소화

**기존 Instruction**
**Skill 보존**

Chat-Vector
기술 활용

**모델 성능**
**및 사용성 개선**

긴 문장 처리 성능
Math / Code 성능

## 실사용 환경에서의 효용성과 신뢰성을 극대화할 수 있도록

# 한국어 지식과 Instruction을 추가 학습하였습니다.

| 한국어 지식 확장 Continual Pre-Training |

Block Expansion 적용*
원본 Knowledge 유지
한국어 Knowledge 확장

| 한국어 Instruction(=skill) 확장 Supervised Fine Tuning |

한국어 Instruction 학습 **
한국어 + 영문 SFT

| 모델 성능/사용성 개선 Multi-stage RL |

Long Context 확장 ***
ChatVector 적용


※ Transformer Block에 추가 학습을 위한 Block을 추가하여
추가된 Block을 우선 학습하고 다른 Block으로 확장하는 Multi-Stage Pretraining을 적용

※※ 한국어 Instruction 학습 데이터에 영문 Instruction 데이터를 혼합하여 학습
Chat-Vector 기술을 활용하여 Meta에서 학습한 Llama의 Instruction Skill 을 반영

※※※ 긴 문장 처리 성능을 높이기 위한 LongPO 학습 기술을 적용,
Reasoning 학습 기법(GRPO)을 적용 Math와 Code 성능을 향상

## Llama K의 강력한 한국어 성능의 근원

# 한국어 역량 강화를 위해 단계별 한국적 데이터 학습

실제 학습에 적합한 고품질 데이터만을 효율적으로 선별하는 KT 고품질 대규모 코퍼스 데이터 셋을 활용하고,
한국어 특화 Reward Model을 개발하여 활용하였습니다.

| 약 180B | 약 8.83B | 3.52B | 105,994개 | 8,740개 | 24,994개 |
|---|---|---|---|---|---|
| 한국어: 89B<br>영어: 91B | 한국어: 2.28B<br>영어 6.55B | 한국어: 1.88B<br>영어: 1.64B | 한국어: 105,994개<br>영어: 0개 | 한국어: 6,417개<br>영어: 2,323개 | Coding 13,920개<br>Math 11,074개 |
| Base Continual Pretraining | Annealing | Supervised Fine-tunning | DPO | LongPO | GRPO |

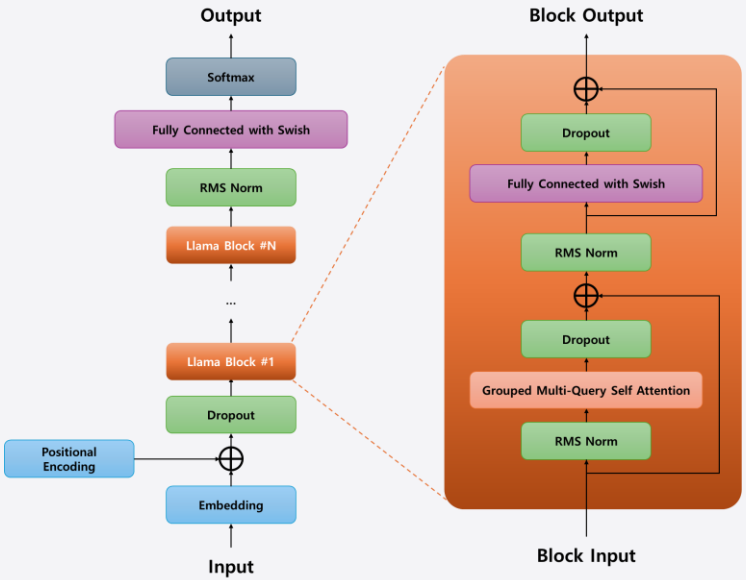
| Llama K 모델 학습 데이터구성 |

# 기술 사양

| 항목 | 기술 사양 상세 |
| --- | --- |
| 모델 유형 | • 텍스트 입력 및 생성 모델 |
| 모델 크기 | • 파라미터 수: 74B<br>• 최대 입출력 길이 지원 (Context-Length): 32K |
| 학습 데이터 | • 한국어: Applied Science, STEM, 생활문화, 인문사회, 식품보건 등<br>• 영어: STEM, 생활문화, 인문사회 등<br>• Code 및 Math: STEM<br>• Multilingual: STEM, 인문사회 등<br>※ 학습 데이터상의 욕설, 비속어, 편견, 차별 등 비윤리적 표현 제거 |

# Llama K 모델 아키텍처 및 상세사양

| Llama K 모델 아키텍처 |




| Llama K 모델 상세사양 |

| Specification | Llama 74B | Llama 74B INT 4 |
|---|---|---|
| Number of Parameters | 74B | 74B |
| Hidden size | 8,192 | 8,192 |
| Number of layers | 84 | 84 |
| Activation function | SILU | SILU |
| Feedforward Dimension | 28,672 | 28,672 |
| Attention type | GQA | GQA |
| Number of attention heads | 64 | 64 |
| Head size | 128 | 128 |
| Context Length | 32K | 32K |
| Positional Embeddings | RoPE ($\theta$ =500,000} | RoPE ($\theta$ =500,000} |
| Vocab size | 128,256 | 128,256 |
| Tied word embedding | False | False |

Llama K,

# 다양한 자연어 Task를 안정적으로 수행합니다.

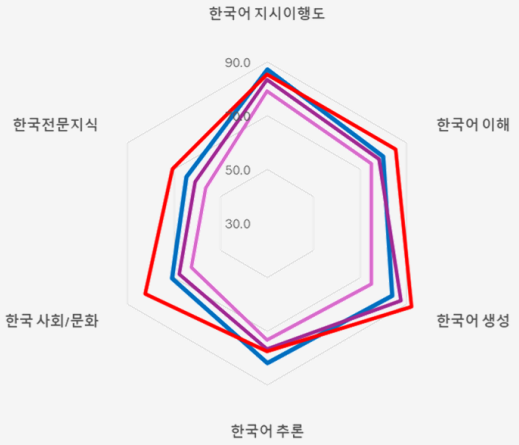
| | | |
|---|---|---|
| 01 | 분류 | 감정 분류, 인과 관계 분류 등 |
| 02 | 질의응답 | 언제/어디/누구/무엇 등 단문형 질의 응답, 설명 등 장문형 질의응답, 기계 독해, 질문 생성 등 |
| 03 | 요약 | 신문 기사, 보도자료, 보고서, 회의록, 사설, 도서, 구조화된 요약 등 |
| 04 | 생성 | 연설문 작성, 신문 기사 제목 생성, 도서 생성, 광고 문구 작성 등 |
| 05 | 변환 | 문장 패러프레이징, 맞춤법 교정 등 |

※ 사용자 목적에 맞게 추가 파인튜닝하여 사용할 수 있습니다.

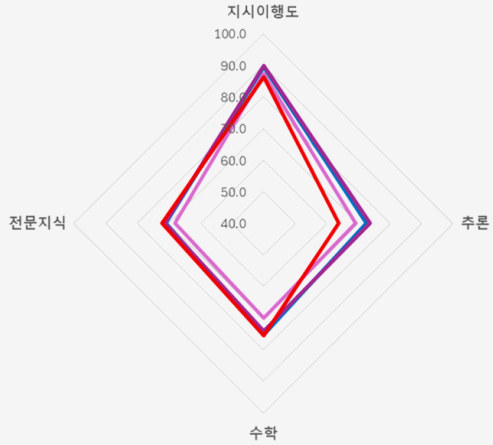
## 영어와 한국적 AI 지표 모두 기본 모델 대비 성능 향상

# 주요 벤치마크 성능 지표

### 한국적 AI 지표 성능 평가




### 영어 공통능력 성능 평가

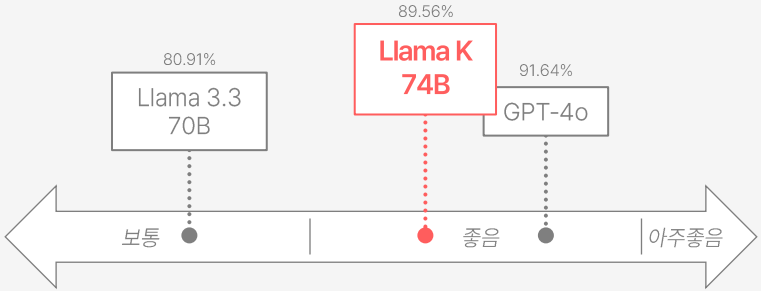



한국어와 영어 모두, 성능 평가 전체 영역에서 Base 모델Llama3.X 대비 성능 우위
'추론' 영역에서는 GPT-4o 보다 높은 성능 확인

## 개발 초기부터 Responsible AI를 고려하여 안전한 모델

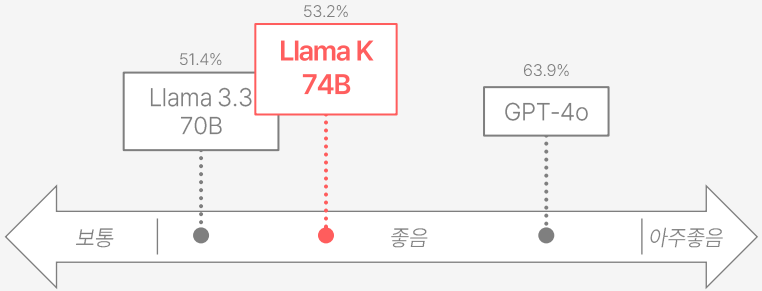
# 모델 안전성 및 강건성 성능 지표

### 모델 안전성 평가




Base 모델보다 훨씬 안전

### 모델 강건성※ 평가




모델 강건성도 향상되어 GPT-4o와 유사

※ 적대적 사용자로의 악의적인 프롬프트 공격에 대한 방어

# Llama K 모델 추론을 위한 메모리 요구량

| | 모델 크기 | GPU 수량 (H100 80G) | 품질 |
|---|---|---|---|
| 16비트(bf16) | 138 GB | 4 | - |
| 8비트 양자화(fp8) | 71 GB | 2 | 16비트 대비 99% 수준 |
| 4비트 양자화(int4) | 39 GB | 1 | 16비트 대비 95% 수준 |

* 32k Context Length 기준 추론 과정 KV 캐시 및 Overhead까지 고려한 수치

04

# Llama K 제공 방식

상품 제공 방식

고객 적용 사례

FAQ

# 고객 환경에 적합한 형태를 선택해서 쓸 수 있습니다.

아래의 대표적인 제공 방식 외에도 오픈소스로 공개하여 고객 서비스 환경에 맞는 유연한 Delivery 가능

## CODE

Python, Java 등 범용적인
프로그래밍 언어로 작성된
오픈소스 기반 코드 형태로 제공

### 대상

모델을 직접 커스터마이징 하거나 개발 인력을 보유한 고객, 비용 부담 없이 AI 모델 도입하고자 하는 고객

### 특징

상세 주석과 예제 코드가 함께 제공되어, 내부 환경에 맞게 손쉽게 모델을 수정하고 적용 가능

## Azure Resource

Microsoft Azure 환경을
기반으로 Llama K 모델을
배포/실행 가능하도록 제공

### 대상

이미 Azure 기반 클라우드 환경을 운영 중인 고객

### 특징

배포 자동화, 모니터링, 보안 설정이 Azure와 통합되어 있어 효율적 운영 가능

# Llama K Prompt Structure

아래의 6가지 요소로 구성하여 프롬프트를 구성할 경우, 모델이 보다 정확하고 의미 있는 응답을 생성하는데 도움이 됩니다.

## 필수 요소

### 명령

모델이 수행해야 할 작업이나 지시를 명확하게 전달합니다.

*각 영업 대리점에 KT 상품을 소개하는 AI 챗봇 서비스를 설명하는 자료를 만들어 주세요.*

## 권장 요소

### 맥락

작업의 배경, 상황, 목적 등 이해를 돕는 설명을 제공합니다.

*이 자료는 AI 서비스에 익숙하지 않은 고객과 직원들이 챗봇의 기능과 장점을 이해하고 활용할 수 있도록 돕는 것이 목적입니다.*

### 역할 ※

모델에게 특정 인물이나 전문가의 관점을 부여합니다.

*당신은 KT 영업본부 직원입니다.*

### 예시

모델이 참고할 수 있는 입력 자료나 사례를 제공합니다.

*(샘플) 제목: KT AI 챗봇 서비스 안내 자료*

*. 소개: AI 챗봇 서비스를 통해 현장에서 고객 문의를 빠르게 처리하는데 목적이 있다*

*. 주요 기능: 실시간 고객 문의 응답*

*. 적용 효과: 대기 시간 감소*

*. 활용 방법: 맞춤형 응대 문장 설정 가능*

## 선택 요소

### 포맷

출력 결과의 형식이나 구조를 지시합니다.

*문단 형태의 줄글로 쉽게 보여 주세요.*

### 어조 ※

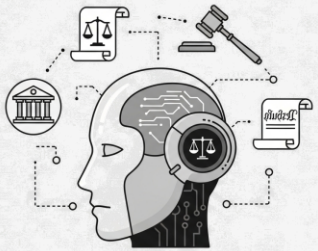
문제, 말투, 표현 수준 등을 지정하여 스타일을 조정합니다.

*AI 서비스에 생소한 고객과 직원을 대상으로 쉽고 재미있게 설명해 주세요.*

※ '역할'과 '어조'는 System Prompt에 반영하는 것이 응답의 일관성과 품질 향상에 도움됨

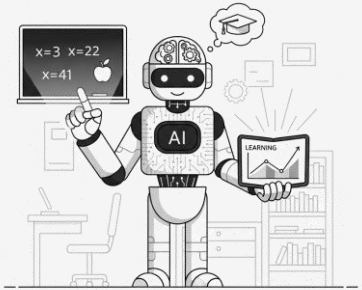
# 다양한 산업에 특화시켜 활용할 수 있습니다.

## 법률 Vertical 모델



법학/법률 분야에 대해
전문가 수준의 법률 및 연관 지식을 바탕으로
다양한 법률 문서(판결/판례/소장 등)를 이해/해석하여
논리적인 법적 추론/비교/예측을 제공합니다.

## 교육 Vertical 모델




초중고교 국/영/수/과/사 과목의
교과/수업내용 이해를 바탕으로
추가 파인튜닝을 하지 않아도
정확하고 상세한 답안을 제공합니다.

## 금융 Vertical 모델

금융업의 용어, 약어, 전문 개념 등을
숙지하고 있으며, 특히 재무제표, 시장 분석
자료 등 표에 대한 높은 이해를 바탕으로
수치 이해 계산 및 해석 능력을 제공합니다.

# Frequently Asked Questions

## Q. 이 모델의 주요 특징은 무엇이며, 기존 글로벌 모델과 비교해 어떤 차별점이 있나요?

Llama K는 Meta의 Llama 3.3 70B모델을 기반으로, 대규모 고품질 한국어 데이터를 추가 학습시켜 한국어 능력을 세계 최고 수준으로 끌어올린 74B 파라미터 규모의 한국어 특화 초거대 언어모델(LLM) 입니다. 강력한 기초 모델의 추론 능력은 그대로 활용하되, 추가된 4B 파라미터는 주로 한국어의 복잡한 문법, 고유한 어휘, 그리고 산업별 특화 용어를 이해하는 데 사용되어, 단순히 한국어를 아는 모델이 아닌 한국 전문가 모델로 재탄생했습니다.

## Q. 저희 회사 데이터에 맞게 모델을 미세조정할 수 있나요?

네, 가능합니다. Llama K는 KT에서 보유한 다른 AI 모델과 마찬가지로 특정 도메인의 데이터나 작업 스타일에 맞게 미세 조정을 지원합니다. 고객사의 고유 데이터를 활용한 미세조정을 통해 해당 도메인에 완벽하게 특화된 우리 회사만의 전문가 AI를 구축할 수 있도록 전문 컨설팅 및 기술 지원을 제공합니다.

## Q. On-Premise 형태로도 제공이 가능한가요?

네, On-Premise상품 제공하고 있습니다. Closed도메인 고객(기업Data의 외부 반출이 어려운 고객) 대상으로 AI E2E서비스(Cloud/GPU인프라, AI모델, 플랫폼, AI Application까지 Full Stack)를 제공합니다.
또한 KT 생성형 AI 모델의 다양한 Task 수행 기능을 간단한 입출력 인터페이스를 통해 기존 서비스와 Integration 할 수 있도록 제공하여 기존 서비스 간의 Integration 작업도 할 수 있습니다.

/ 더 많은 FAQ는 ai.kt.com에서 확인하실 수 있습니다. /

# 감사합니다.