ChatGPT

LLM 모델 이름의 수식어 의미와 특성 분석

Gemma-Ko-7B

- **'Ko'** – Gemma 모델 이름에 붙은 "Ko"는 해당 모델이 **한국어 특화** 버전임을 나타냅니다. 원래 Gemma는 Google의 연구를 기반으로 만들어진 경량 LLM 계열로 영어 등 다국어 버전과 지시 수행(Instruct) 버전 등이 공개되어 있는데 [1], Gemma-Ko는 한국어 데이터로 추가 학습되어 **한국어 이해 및 생성에 최적화**된 변형입니다. 즉, 한국어 어휘와 문장 구조에 맞게 어휘집(vocabulary)을 구성하고 한국어 말뭉치로 추가 학습함으로써 한국어 질의에 대한 유창성과 정확도를 높인 모델입니다.

- **7B** – 숫자는 **모델의 파라미터 수(규모)**를 뜻하며, 7B는 약 70억 개의 파라미터를 가진 소형 모델임을 의미합니다. 이는 대형 모델 대비 성능은 낮을 수 있으나, **경량화로 인해 필요한 VRAM 등 자원이 적고 응답 속도가 빠른 장점**이 있습니다 [1]. Gemma-Ko-7B는 **베이스(Base) 모델**로서 사전 학습만 거친 상태이며, 별도의 지시문 응답 튜닝이나 RLHF 없이 공개되었습니다 [2]. 따라서 일반적인 텍스트 생성 능력은 갖추었지만 **지시 수행 최적화가 되어 있지 않으므로** 사용자 질문에 바로 친절하게 답변하기보다는, 프롬프트를 잘 구성하거나 추가 파인튜닝(예: LoRA)을 통해 활용 성능을 높일 수 있습니다.

A.X-4.0-Light (SKT)

- **'Light'** – SK텔레콤의 A.X 4.0 모델에서 "Light"는 **경량화 버전**임을 나타냅니다. A.X 4.0은 원본 72억 파라미터짜리 **표준 모델(72B)**과 약 **7억 파라미터(7B)** 규모의 **경량 모델** 두 가지로 공개되었는데 [3], 이 중 Light는 소형(7B) 모델입니다. 경량 버전은 **토큰 처리 효율을 높이고 배포 유연성을 확보**하기 위해 제공된 것으로, 예를 들어 A.X 4.0은 한국어 입력에 대해 GPT-4 대비 33% 적은 토큰을 사용하면서도 뛰어난 성능을 보이며, 대용량(72B)과 경량(7B) 모델 둘 다 맥락길이를 크게 확장한 것이 특징입니다 [3]. Light 모델은 **성능은 낮지만 메모리 요구량과 추론 지연이 크게 줄어들어** 기업 환경에서 빠른 응답이 필요한 서비스나 한정된 자원에서의 실험에 적합합니다. (참고로 A.X 4.0 Light의 한국어 이해 성능은 같은 7B급 모델인 Qwen2.5-7B, 카카오 KoGPT 계열 등보다도 높은 것으로 보고되었습니다 [4].) **지시 수행** 능력도 기본적으로 포함하고 있어, Light라도 바로 QA/대화용으로 활용 가능하며, 다만 대형 모델 대비 복잡한 문제에서는 한계가 있을 수 있습니다.

- **버전 4.0** – A.X 뒤의 버전 번호는 **모델 세대 및 업그레이드**를 의미합니다. A.X 4.0은 SKT가 2018년부터 개발해 온 한국어 특화 LLM "A.(에이닷)"의 최신 세대 모델로, **개방형 중국 모델 Qwen 2.5를 기반**으로 한국어 데이터를 대규모 추가 학습하여 실제 비즈니스 환경 성능을 끌어올렸습니다 [5]. 버전 4.0에서는 **맥락 길이 131k 토큰 지원**, 한국어 문화/상식에 대한 심층 이해 향상, 함수 호출 등 도구 사용 능력 등이 통합되었으며, Standard와 Light 두 가지로 나뉘어 배포되었습니다 [3].

Mi:dm-2.0-Base-Instruct (KT)

- **'Base'** – Mi:dm 2.0 모델의 "Base"는 **기본이 되는 대형 모델(11.5B)**을 가리킵니다. KT가 개발한 Mi:dm 2.0은 한국어 문화적 맥락과 상식을 깊이 반영한 "한국어 중심 AI" 언어모델로, **11.5억 파라미터의 Base 모델과 2.3B 파라미터의 Mini 모델** 두 가지로 공개되었습니다 [6]. Base는 **성능과 모델 크기의 균형**을 추구하여 실제 응용에 충분한 성능을 내면서도 너무 크지 않게 설계되었으며, Depth-up Scaling(DuS) 기법으로 8B 규모 모델을 확장한 구조를 가지고 있습니다 [7]. 반면 Mini는 Base를 **가지치기(pruning)와 지식 증류(distillation)**로 축소해 만든 소형 버전으로, 디바이스나 제한된 자원 환경에 최적화되었습니다 [8].

- **‘Instruct’** – 모델 이름에 "Instruct"가 붙은 것은 **지시 문장/질문에 따라 답변하도록 추가 튜닝**되었음을 의미합니다. 즉, Mi:dm 2.0 Base-Instruct는 **Base 모델을 기반으로 사용자 지시에 응하는 대화형으로 파인튜닝**한 버전입니다. 이러한 지시 특화 튜닝을 통해 모델은 질문에 대한 직접적인 답변이나 설명형 응답 등 **LLM 어시스턴트 스타일의 출력**을 생성할 수 있습니다. 예를 들어 Base-Instruct 모델은 대화 프롬프트 포맷을 인지하고 시스템/사용자 역할을 구분하여 답변하도록 학습되었으며 [9], 별도의 Prompt Engineering 없이도 "질문: ~에 답해줘" 형태의 입력에 적절히 응답하는 경향을 보입니다. 반면 **Base 모델은 사전학습 상태 그대로라**서 이러한 친절한 응답보다는 확률적으로 다음 토큰을 생성하는 언어모델 본연의 출력을 내놓기 때문에, 사용자의 질의에 대한 정확한 답변보다는 훈련 코퍼스 상 연관 텍스트를 이어가는 식의 응답이 나타날 수 있습니다. 따라서 **QA 성능을 높이려면 Base 모델은 LoRA 등을 통한 추가 파인튜닝이나 세심한 프롬프트 예시 제공이 필요**하며, Base-Instruct는 이미 어느 정도 QA에 맞춰져 있어 바로 활용하기에 적합합니다.

HyperCLOVA X SEED-Think-14B (NAVER)

- **‘Think’** – 이름에 "Think"가 붙은 HyperCLOVA X 모델은 **고차원 문제를 단계별 추론으로 해결하는** 능력이 강화된 **추론 특화 모델**임을 뜻합니다. 네이버는 HyperCLOVA X 시리즈 중에서 **복잡한 문제 해결을 위해 단계적 사고(chain-of-thought)**를 할 수 있는 모델들을 "Think"라는 명칭으로 구분하였습니다. “'Think'는 모델이 단계적 추론으로 고난도 문제를 풀 수 있는 추론 능력을 지녔음을 의미한다”고 네이버는 설명하고 있습니다 [10]. 이러한 Think 모델은 일반 모델과 비교해 **응답을 산출하기 전에 내부적으로 생각의 흐름을 전개**하며 (예: 수식 문제를 풀 때 중간 계산과정을 텍스트로 생성), 그 결과 복잡한 질문에 대한 논리적 정확도가 향상됩니다.

- **추론 방식 및 시간 영향** – HyperCLOVA X SEED Think 14B 모델은 **추론 모드**와 **비추론 모드**를 나누어 운영할 수 있습니다. 프롬프트 상에 `<|im_start|>assistant/think`와 같은 토큰을 넣으면 모델이 우선 **숨겨진 ‘생각’ 단계의 출력을 충분히 생성한 후** 최종 답변을 내놓는데 [11] [12], 이러한 **체인-오브-소트(Chain-of-Thought) 생성**으로 인해 **추론 시간이 늘어날 수 있습니다**. 즉, 모델이 곧장 답을 산출하지 않고 수십~수백 토큰 분량의 추론 과정을 거치므로 **전체 생성 토큰 수가 증가**하고 그만큼 시간이 더 걸립니다. 또한 HyperCLOVA X Think 모델은 맥락 길이가 32k 토큰으로 매우 긴데 [13], 긴 컨텍스트를 다루기 위한 연산(예: 슬라이딩 윈도우 또는 희소 Attention 등)이 추가되어 **토큰당 계산량**이 일반 14B 모델보다 많을 수 있습니다. 다만 이는 추론 정확도를 높이기 위한 설계로, **디코딩 방식 자체가 특이하다기보다는** 다단계 출력을 생성하도록 한 **튜닝 기법(RLVR: 검증형 보상 강화학습 등)**의 영향입니다 [14]. 요약하면 Think 모델은 내부적으로 한 번 더 생각하고 답하는 구조이므로 응답이 신중하고 정확한 반면, **추론 단계로 인해 응답 지연(latency)이 증가**하는 트레이드오프가 있습니다.

- **‘SEED’** – HyperCLOVA X 모델명 중 "SEED"는 **경량화·효율화된 공개 버전**임을 나타냅니다. 2023년 공개된 HyperCLOVA 내부 거대 모델(HyperCLOVA X)은 수백억~수천억 파라미터 규모였으나, 2024~2025년에 네이버는 자체 기술로 경량화한 **HyperCLOVA X Seed 시리즈** (예: 3종)를 공개하였습니다 [15]. SEED-Think-14B는 그 중 하나로, **대형 모델 HyperCLOVA X Think의 성능을 유지하면서도 파라미터를 147억 개로 줄인** 모델입니다 [16]. 이는 **불필요한 파라미터를 가지치기하고 지식 증류**를 통해 압축했으며, 거기에 **최신 강화학습 기반 추론 능력 강화기법**(SFT + RLVR + LC + RLHF 조합)을 적용해 **작은 모델로 큰 모델 수준의 추론 성능**을 달성하고자 한 결과물입니다 [14]. 그 결과 학습 비용은 동급 해외 모델 대비 1% 수준으로 낮추고도, 한국어 이해 등 여러 벤치마크에서 유사 규모 최고 성능을 보였습니다 [17].

Kanana-1.5-15.7B-A3B-Instruct (KAKAO)

- **‘A3B’** – 카카오의 Kanana-1.5-15.7B-A3B 모델에서 "A3B"는 **Mixture-of-Experts (MoE)** 기반의 특수 아키텍처를 의미합니다. 구체적으로 157억 파라미터 중 **약 30억 파라미터(3B)만 활성화**되어 동작함을 가리키는데 [18], 이는 곧 **토큰 생성 시마다 일부 전문가(neuron expert)만 사용**되어 연산량을 줄이는 **스파스(sparse) 모델**임을 뜻합니다. Kanana 15.7B-A3B 모델은 **15.7B 규모의 파라미터를 가지고도 매 토큰 추론에는 3B 상당의 계산만 수행**하므로, 동일한 8B~16B급 밀집(dense) 모델 대비 **약 1/3 수준의 FLOPs로 유사한 성능**을 발휘합니다 [19]. 실제로 “전체 157억 파라미터 중 약 30억만 활성화되어 동작하므로, 연산 비용을 크게 줄이면서도 높은 성능을 유지”한다고 설명되고 있습니다 [18]. 이러한 A3B MoE 구조 덕분에 **추론 효율과 응답 속도가 향**

상되며, 특히 대용량 모델을 실제 서비스에 투입할 때 **운영 비용을 절감**할 수 있는 장점이 있습니다. (참고: A3B 모델은 파라미터 전체를 메모리에 적재는 하지만, 연산시 부분만 사용하기에 **GPU 메모리 대역폭 및 연산 자원 사용이 경량화**됩니다. 다만 여러 전문가 간의 게이팅 연산이 추가되므로 구현에 따라 약간의 오버헤드는 있을 수 있습니다.)

- **'Instruct'** – Kanana 1.5 A3B 모델의 Instruct 버전은 **사람 지시나 질문에 최적화된 지시 따라하기 튜닝 모델**입니다. 카카오는 Kanana 시리즈에서 **사람 선호도 반영 학습(인간 피드백 강화)**과 지식 증류, on-policy distillation 등을 거쳐 **사용자 질문에 대한 답변 정확도와 친절도를 높인** Instruct 모델을 함께 공개했습니다 [20] [21]. 따라서 Kanana-1.5-15.7B-A3B-Instruct는 동일한 MoE 구조의 **Base 모델을 기반으로 추가적인 지시형 미세튜닝 및 강화학습(RLHF)**을 통해, 질의응답이나 대화에서 **더 자연스럽고 협조적인 답변 스타일**을 보여줍니다. 예를 들어, 이 모델은 일반적인 Kanana Base 모델에 비해 KoMT-Bench, IFEval 등 **한국어 지시 수행 평가에서 더 높은 점수**를 보이며 [22], 코드 및 수학 문제 등에서도 사용자 질문 의도를 잘 해석해 단계적 풀이를 내놓습니다. (다만 흥미롭게도, 평가 결과에 따르면 Kanana 15.7B A3B Instruct가 일부 지표에서 동급 8B 밀집 모델 대비 약간 낮은 점수를 보이는데 [22], 이는 MoE 특성상 지식 분산으로 인한 답변 스타일 편차나 RL 튜닝 세부 차이 때문으로 추정됩니다.)

- **기타 특성 ('1.5' 버전)** – Kanana 모델명의 "1.5"는 **모델 버전**으로, 2024년 공개된 Kanana 1.0 대비 **향상된 두 번째 세대**임을 나타냅니다. Kanana 1.5 시리즈는 전반적으로 **성능/효율 업그레이드 및 멀티모달 확장**이 이루어졌는데, 예를 들어 동일 발표에서 Kanana-1.5-v-3B (비전+텍스트 겸용 30억 모델)도 함께 공개되었습니다 [23] [24]. 이처럼 Kanana 1.5 세대에서는 **모델 구조 혁신(MoE 도입)**과 **향상된 학습 기법**으로 한국어/영어 능력을 높이고, Apache 2.0 라이선스로 공개하여 국내 생태계 활성화를 도모하고 있습니다 [18] [25].

EXAONE-4.0-32B (LG)

- **'4.0' 통합형 LLM** – LG AI연구원의 EXAONE 4.0 모델은 **비추론 모드(Non-reasoning)와 추론 모드(Reasoning)를 하나로 통합**한 차세대 LLM입니다 [26]. 이전 세대 EXAONE 3.5가 빠른 응답 및 뛰어난 사용성(Instruction-following)에 초점을 맞추고, 별도로 거대 모델 EXAONE Deep이 고도의 추론 능력을 보유했었던 것에 비해, **버전 4.0에서는 두 가지 모드를 한 모델에 결합**함으로써 **일상적인 질문에 대한 신속하고 정확한 답변**부터 **고차원 문제에 대한 심층 추론**까지 한 모델이 수행할 수 있게 되었습니다 [26]. 이를 통해 사용 편의성과 논리적 사고력을 모두 갖춘 **하이브리드 AI 모델**을 지향하며, 함수 호출 등의 **에이전트 도구 사용 능력**도 기본 통합되었습니다 [27] [28]. 또한 4.0 버전에서는 **스페인어를 추가 지원**하여 다국어 범위를 넓혔고, **하이브리드 어텐션 구조**(국소주의+전체주의 결합) 도입과 **QK-Reorder-Norm** 등의 새로운 아키텍처 개선을 적용해 성능을 끌어올렸습니다 [29]. 요약하면, EXAONE 4.0 = EXAONE 3.5(사용성 좋은 챗 모델) + EXAONE Deep(강력 추론 모델)의 장점을 합쳐 업그레이드한 버전이라 할 수 있습니다.

- **32B** – EXAONE 4.0 시리즈는 **32억 파라미터(mid-size)** 모델과 **1.2B 파라미터 소형 모델** 두 가지로 제공됩니다 [30]. 이 중 EXAONE-4.0-32B가 주력 **전문가 수준 모델**로, 한국어·영어·스페인어에 모두 능통하며 각종 벤치마크에서 동급 30B급 공개 모델들을 상회하는 최고 성능을 입증했습니다 [26]. 32B 모델은 **64레이어 Transformer에 131k 토큰 문맥길이**를 지원하고, **하이브리드 어텐션(국소+글로벌 3:1 비율)**으로 긴 입력 처리 효율을 개선하였습니다 [31] [32]. 또한 **Reasoning 모드** 활성화 시 `<think>` 태그를 이용해 체인-of-Thought를 수행하고 [33] [34], Non-reasoning 모드에서는 곧바로 답변을 생성하는 등 **질문 난이도에 따른 유연한 대응**이 가능합니다. EXAONE 4.0은 애초에 **지시형 데이터로 충분히 튜닝**되어 나왔기 때문에 별도의 Instruct라는 수식어를 붙이지는 않았지만, 사실상 **챗봇 어시스턴트 스타일**로 동작하며, 필요 시 Reasoning 모드로 **추론 과정을 내부적으로 거친 후 답변하는 기능**까지 갖춘 것이 특징입니다. 단일 모델로 이러한 모드 전환을 지원하기 때문에, **일반 질의응답에서는 빠른 응답을, 어려운 문제에서는 깊이 있는 풀이를** 자동으로 수행하여 사용자로서는 일관된 경험을 얻을 수 있습니다.

LLM QA 활용 시 수식어별 고려사항

- **Base vs. Instruct 모델 선택:** 질문 answering 성능을 높이기 위해서는 가급적 **지시형(Instruct)으로 튜닝된 모델**을 활용하는 것이 유리합니다. Instruct 모델은 이미 인간 질문-답변 양식에 맞춰 미세조정되어 있어, 별다른 프롬프트 기교 없이도 비교적 **정확하고 간결한 답변**을 생성합니다 [10] [9]. 반대로 **Base (사전학습만 된) 모델**은 prompt engineering이 성능에 크게 영향을 미치므로, few-shot 예시 제공이나 체계적인 질문 서술 등 **프롬프트 설계에 신경**써야 합니다. 또한 Base 모델을 자체 데이터로 추가 파인튜닝(예: **QLoRA 활용 미세조정**)할 경우, 처음부터 Instruct 모델을 미세조정하는 것보다 **Base 모델을 사용하는 편이 안정적**인 경우가 많습니다. 왜냐하면 Instruct 모델은 이미 일반적인 지시 팔로우로 최적화되어 있어 특정 도메인으로 재튜닝하면 **기존의 광범위한 지식이 부분적으로 희석**되거나, RLHF로 인한 보수적 응답 경향 때문에 **학습 신호에 덜 민감**할 수 있기 때문입니다. 반면 Base는 보다 **가공되지 않은 상태이므로 LoRA 등을 통한 새로운 태스크 학습률이 높고**, 이후에도 원하는 스타일로 Instruct化를 직접 수행할 수 있는 장점이 있습니다.

- **경량(Light/Mini) vs. 대형 모델:** 자원 제약과 응답 속도 측면에서 Light/Mini 모델과 대형 모델의 트레이드오프를 고려해야 합니다. **경량 모델**(예: 7B)은 메모리 부담이 적고 **양자화**(quantization)까지 하면 일반 PC에서도 빠르게 추론이 가능하지만, 지식 양과 논리적 추론 능력이 제한적일 수 있습니다 [35]. **대형 모델**(예: 32B, 72B)은 QA 정답율이나 복잡한 질문 대응에서 우수하지만, **VRAM 요구량이 커서 8bit/4bit 양자화를 해도 배포가 어려울 수 있고** 응답 시간도 느립니다. 따라서 주어진 환경에서 **QLoRA로 미세조정을 수행**할 때에도, **경량 모델은 학습 시간과 메모리 측면에서 유리**하여 빠른 실험 사이클을 돌릴 수 있지만 최고 성능이 낮고, **대형 모델은 미세조정 비용이 크지만 얻을 수 있는 성능 잠재치가 높음**을 감안해야 합니다. 실제로 금융 QA같이 정확도가 중요한 대회라면 가능하면 상위 규모 모델(Instruct 튜닝됨)을 선택하고, 자원 내에서 4bit 양자화하여 배포하는 전략이 유효합니다. 반면 실시간 답변 서비스처럼 속도가 중요하고 약간의 성능 저하는 허용된다면 경량 모델(또는 MoE 모델)을 사용하는 편이 나을 것입니다.

- **'Think'(추론) 모드 활용:** HyperCLOVA X의 Think나 EXAONE 4.0의 reasoning 모드는 **복잡한 질문에 대한 성능 향상**을 위해 제공됩니다. 체인-오브-소트를 유도하면 수학 문제 풀이 등에서 정답률이 높아질 수 있으나, **응답 지연이 커지고** 경우에 따라 불필요하게 장황한 출력을 낼 가능성도 있습니다. 그러므로 **QA 시스템에서 모든 질의에 일괄적으로 추론 모드를 켜기보다는, 난이도를 판단**해 복잡한 문제에만 Think 모드를 사용하거나, "간단히 답할 수 있는 질문은 바로 답하도록" 프롬프트를 설정하는 등 **선별적 적용 전략**이 필요합니다. 또한 Think 모델을 양자화할 때에는 내부 추론 단계까지 quantize 연산을 거치므로 정확도 손실 가능성이 있어, **적절한 모델 교정(예: 중간 계산은 FP16 유지)** 등의 고려도 할 수 있습니다.

- **Prompt Engineering과 수식어 특성:** Instruct 모델의 경우 프롬프트에 질문만 써도 대부분 적절히 답하지만, Base 모델은 보통 **질문 앞에 시스템 역할 지시**나 **예시 Q&A**를 넣어야 제대로 답을 합니다. 예를 들어 Base에는 "질문: ___ 답변:" 형태로 맥락을 주는 것이 효과적입니다. 반면 **Think 모델**은 '... 생각하고 답변해' 같은 문구나 전용 토큰이 필요할 수 있고 [11] [12], EXAONE처럼 **enable_thinking=True** 옵션으로 토크나이징해야 체인-오브-소트가 시작됩니다 [33]. 따라서 **모델의 수식어에 따라 최적 프롬프트 형태도 달라지므로**, 실험 시 해당 모델 카드의 지침(프롬프트 템플릿, 토큰 등)을 참고하여야 합니다. 특히 **MoE 모델(A3B)**은 사용자에게 출력되는 답변 스타일은 일반 모델과 동일하지만, 내부 동작이 다르므로 프롬프트 엔지니어링에 직접적인 영향은 없으나, **일관성** 면에서 간혹 전문가 선택에 따른 응답 편차가 있을 수 있어(예: 같은 질문을 여러 번 하면 답변이 미세하게 달라질 수 있음) 이를 평균화하거나 검토하는 전략이 필요합니다.

요약: 모델 이름에 포함된 수식어들은 해당 LLM의 **아키텍처(예: MoE, Hybrid Attention)**, **학습 방식(예: 지시 튜닝, RLHF, 지식 증류)**, **응답 스타일(예: 단계별 풀이 vs. 즉각 답변)**, **효율 및 규모(경량화 여부)** 등을 나타냅니다. 사용자는 이 정보를 바탕으로 **과제와 환경에 맞는 모델**을 선택해야 합니다. 예를 들어, **상세한 설명이 필요한 금융 QA**에는 **대형 Instruct 모델**이나 **Think 모델**이 유리하겠지만, **실시간 대화 서비스**에는 **경량 모델이나 A3B 모델**이 적합할 수 있습니다. 또 **추가 파인튜닝**을 고려한다면 Base 모델 (또는 Base-Instruct)을 택해 **QLoRA로 도메인 특화 튜닝**을 하고, 추론 시에는 필요에 따라 Think 모드를 켜 정확도를 높이는 식의 전략을 활용할 수 있습니다. 각각의 수식어가 시사하는 특성을 이해함으로써, 제한된 자원 내에서 **최적의 성능 대 효율 균형점을 찾는 것**이 가능합니다.

참고 자료: 모델 공개 페이지 및 보도자료 등에서 각 수식어의 의미를 확인할 수 있습니다 (예: KT Mi:dm 2.0 설명 [6], Naver HyperCLOVA X-Think 설명 [10], Kakao Kanana MoE 설명 [18], SKT A.X 4.0 Light 설명 [3], LG EXAONE 4.0 기술 보고 [26] 등). 이러한 자료들은 모델 선택과 활용 전략 수립에 유용한 정보를 제공합니다.

---

[1] [2] beomi/gemma-ko-7b · Hugging Face
https://huggingface.co/beomi/gemma-ko-7b

[3] [4] [5] [35] skt/A.X-4.0-Light · Hugging Face
https://huggingface.co/skt/A.X-4.0-Light

[6] [7] [8] [9] K-intelligence/Midm-2.0-Base-Instruct · Hugging Face
https://huggingface.co/K-intelligence/Midm-2.0-Base-Instruct

[10] [15] [17] Naver unveiled its own lightweight AI inference model for free, saying, "We will expand the base of .. - MK
https://www.mk.co.kr/en/it/11374203

[11] [12] [13] [14] [16] naver-hyperclovax/HyperCLOVAX-SEED-Think-14B · Hugging Face
https://huggingface.co/naver-hyperclovax/HyperCLOVAX-SEED-Think-14B

[18] [20] [23] [24] [25] Kakao Becomes First in Korea to Open-Source Advanced AI Models
https://koreatechtoday.com/kakao-becomes-first-in-korea-to-open-source-advanced-ai-models/

[19] [21] [22] kakaocorp/kanana-1.5-15.7b-a3b-instruct · Hugging Face
https://huggingface.co/kakaocorp/kanana-1.5-15.7b-a3b-instruct

[26] [2507.11407] EXAONE 4.0: Unified Large Language Models Integrating Non-reasoning and Reasoning Modes
https://arxiv.org/abs/2507.11407

[27] [28] [29] [30] [31] [32] [33] [34] LGAI-EXAONE/EXAONE-4.0-32B · Hugging Face
https://huggingface.co/LGAI-EXAONE/EXAONE-4.0-32B