Language model 2024-current

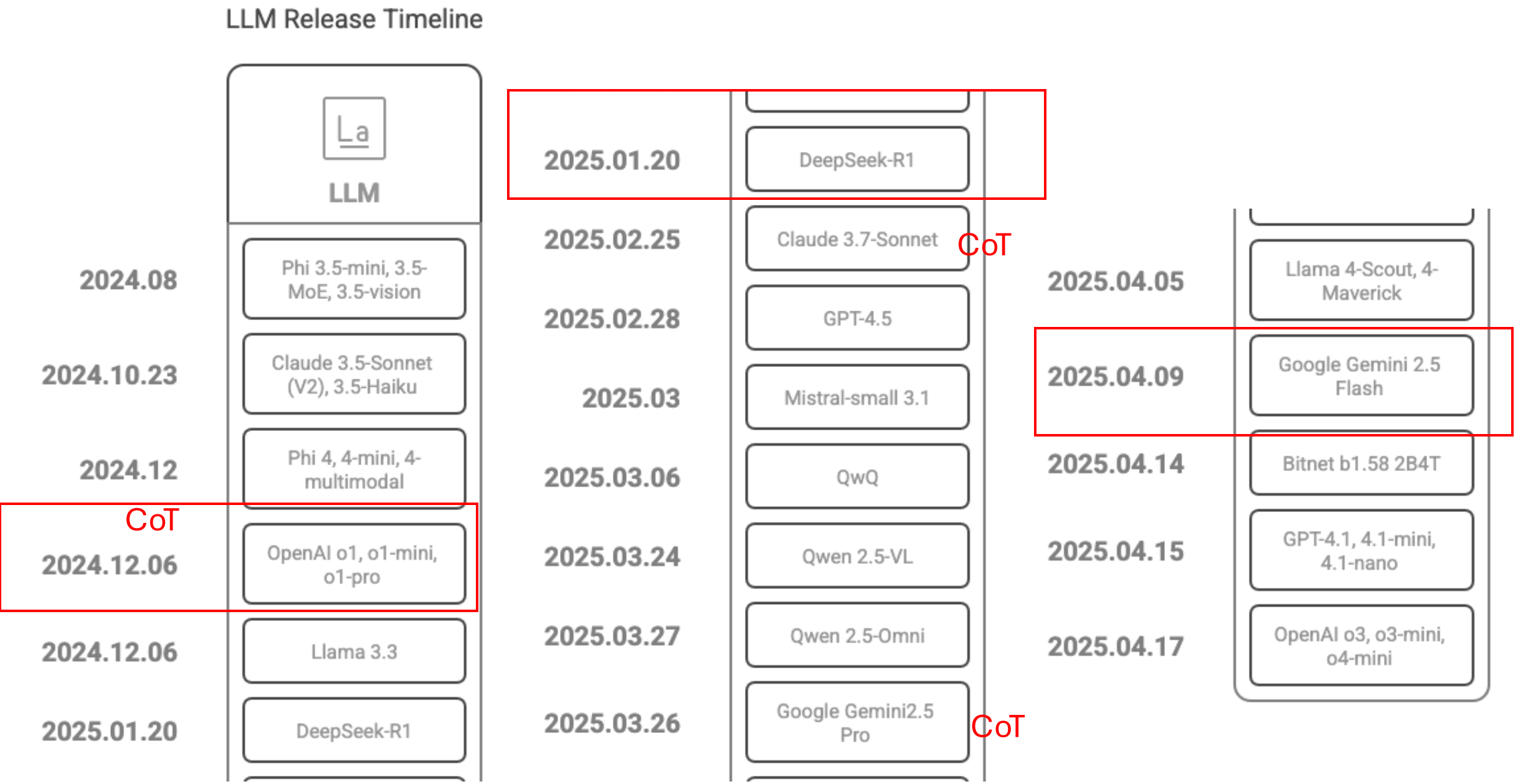
2025.04.18

신재익

- 추론 (Reasoning) 능력
- 사고의 흐름 (Chain of Thought, CoT)

- 멀티모달 모델

- 속도 vs 비용


LLM Release Timeline
LLM
2024.08
Phi 3.5-mini, 3.5-MoE, 3.5-vision
2024.10.23
Claude 3.5-Sonnet (V2), 3.5-Haiku
2024.12
Phi 4, 4-mini, 4-multimodal
CoT
2024.12.06
OpenAI o1, o1-mini, o1-pro
2024.12.06
Llama 3.3
2025.01.20
DeepSeek-R1
2025.01.20
DeepSeek-R1
2025.02.25
Claude 3.7-Sonnet
CoT
2025.02.28
GPT-4.5
2025.03
Mistral-small 3.1
2025.03.06
QwQ
2025.03.24
Qwen 2.5-VL
2025.03.27
Qwen 2.5-Omni
2025.03.26
Google Gemini2.5 Pro
CoT
2025.04.05
Llama 4-Scout, 4-Maverick
2025.04.09
Google Gemini 2.5 Flash
2025.04.14
Bitnet b1.58 2B4T
2025.04.15
GPT-4.1, 4.1-mini, 4.1-nano
2025.04.17
OpenAI o3, o3-mini, o4-mini

OpenAI

- 2024.12.06 : OpenAI o1 , 01-mini , o1-pro
  - CoT 기능 추가한 추론형 멀티모달 모델
- 2025.02.28 : GPT-4.5
  - 마지막 비 추론 모델
- 2025.04.15 : GPT-4.1 , 4.1-mini , 4.1-nano
  - 4.5보다 매우 저렴
- 2025.04.17 : OpenAI o3 , o3-mini, o4-mini

https://platform.openai.com/docs/models

Anthropic Claude

- 2025.10.23 Claude 3.5-Sonnet (V2) , 3.5-Haiku
  - Computer use
- 2025.02.25 : Claude 3.7-Sonnet
  - 추론모델(유료) , vibe coding

https://docs.anthropic.com/en/docs/about-claude/models/all-models

Google Gemini

- 2025.03.26 : Google Gemini 2.5 Pro
  - CoT 내장
- 2025.04.09 : Google Gemini 2.5 Flash
  - Adaptive Thinking

https://ai.google.dev/gemini-api/docs/models

Google Gemma

- 2025.03.12 : Gemma 3 1B , 4B , 12B , 27B
  - 멀티모달 , CoT 내장

https://ai.google.dev/gemma/docs/core
https://huggingface.co/collections/google

Meta Llama

- 2024.12.06 : Llama 3.3
- 2025.04.05 : Llama 4-Scout , 4-Maverick
  - 멀티모달, 최초의 오픈소스 MoE,
  - 밴치마크 cheating 논란

https://www.llama.com/docs/model-cards-and-prompt-formats/
https://huggingface.co/collections/meta-llama/

MS Phi

- 2024.12 : Phi 4 , 4-mini , 4-multimodal
  - 멀티모달 (오디오포함), CoT
- 2024.08 : Phi 3.5-mini , 3.5-MoE , 3.5-vision
  - MoE 지원

- MAI-DS-R1 : post-trained DeepSeek-R1 reasoning model
- 2025.04.14 : Bitnet b1.58 2B4T
  - https://github.com/microsoft/BitNet

https://huggingface.co/collections/microsoft/
https://ai.azure.com/explore/models

Mistral

- 2025.03 : Mistral-small 3.1
  - CoT, Kor, vision,
- 2025.03 : Pixtral
  - 멀티모달

https://docs.mistral.ai/getting-started/models/models_overview/
https://huggingface.co/mistralai

DeepSeek

- 2024.12.26 : DeepSeek-V3
  - Chat
- 2025.01.20 : DeepSeek-R1
  - Reasoner, multi-round

https://api-docs.deepseek.com/quick_start/pricing
https://huggingface.co/collections/deepseek-ai

Alibaba Qwen

- 2025.03.06 : QwQ
  - 추론모델
- 2025.03.24 : Qwen 2.5-VL
  - Vision-language 모델
- 2025.03.27 : Qwen 2.5-Omni
  - 멀티모달

https://qwenlm.github.io/blog/
https://huggingface.co/collections/Qwen

- 모델 발표
- 2024.12.06 : OpenAI o1 , 01-mini , o1-pro
- 2025.02.28 : GPT-4.5
- 2025.04.15 : GPT-4.1 , 4.1-mini , 4.1-nano
- 2025.04.17 : OpenAI o3 , o3-mini, o4-mini
- 2025.10.23 Claude 3.5-Sonnet (V2) , 3.5-Haiku
- 2025.02.25 : Claude 3.7-Sonnet
- 2025.03.26 : Google Gemini2.5 Pro
- 2025.04.09 : Google Gemini 2.5 Flash
- 2024.12.06 : Llama 3.3
- 2025.04.05 : Llama 4-Scout , 4-Maverick
- 2024.12 : Phi 4 , 4-mini , 4-multimodal
- 2024.08 : Phi 3.5-mini , 3.5-MoE , 3.5-vision
- 2025.04.14 : Bitnet b1.58 2B4T
- 2025.03 : Mistral-small 3.1
- 2025.01.20 : DeepSeek-R1
- 2025.03.06 : QwQ
- 2025.03.24 : Qwen 2.5-VL
- 2025.03.27 : Qwen 2.5-Omni