

LLM Knowledge Cutoff Dates (Every Major Model Fact-Checked, June 2026)

Metehan Yesilyurt

AI Search & Information Retrieval Researcher · metehan.ai

Published: September 06, 2026 · Canonical: <https://metehan.ai/articles/llm-knowledge-cutoff-dates/>

Ask any model when its knowledge stops and you often get a wrong answer. Qwen maintainers have said openly that self-reported cutoff dates are unreliable because models are not trained to identify themselves. Blog posts copy each other. Aggregators lag new releases.

This guide fixes that. I fact-checked **LLM knowledge cutoff dates** across OpenAI, Anthropic, Google Gemini, Qwen, Kimi, and MiniMax using official API docs, model cards, and provider system prompts first, then cross-checked against OpenRouter and aggregator data. Every row notes the source tier so you know what to trust.

If you track [AI visibility](#)^[1] or run [AEO](#)^[2], cutoff dates matter. A model that stopped learning in January 2025 cannot reliably describe your product launch from March 2026 unless it browses or you feed it context.

TL;DR

- Freshest western frontier cutoffs: **GPT-5.5** (December 1, 2025) and **Claude Fable 5 / Opus 4.8** (January 2026 reliable knowledge).
- Biggest surprise: **Gemini 3.1 Pro** and **Gemini 3.5 Flash** both list **January 2025** in official Google docs despite May 2026 releases. Third-party blogs claiming January 2026 for 3.5 Flash are wrong.
- Chinese labs vary: **Qwen 3.5+** uses a **2026** cutoff in DashScope system prompts; **Kimi K2.6** is **April 2025**; **MiniMax M3** is **January 2026**.
- Never trust a model's self-reported cutoff. Check the API docs or this table.

How we fact-checked

Tier 1 (official): Provider API documentation, model overview pages, HuggingFace chat templates with embedded cutoff strings, Anthropic Help Center.

Tier 2 (secondary): OpenRouter model metadata, aiknowledgecutoff.com, llm-bento.com, technical reviews citing vendor specs.

Tier 3 (unverified): Model self-reports when asked "what is your knowledge cutoff?" We excluded these from the tables unless no Tier 1 or 2 source exists.

Two cutoff types (Anthropic only): Anthropic publishes both a **reliable knowledge cutoff** (where knowledge is most extensive) and a **training data cutoff** (broader date range). Other providers usually publish one date.

Release vs cutoff gap: A model released in June 2026 can still have a January 2025 cutoff. New release does not mean fresh training data.

Checked on **June 17, 2026**.

Master comparison: freshest model per family

Family	Flagship model	Knowledge cutoff	Release date	Source tier	Live data option
OpenAI	GPT-5.5	December 1, 2025	April 23, 2026	Tier 1	Web search, file search
Anthropic	Claude Fable 5	January 2026	June 9, 2026	Tier 1	Tool use, web (via tools)
Google	Gemini 3.5 Flash	January 2025	May 19, 2026	Tier 1	Search grounding
Qwen	Qwen3.7 Max	2026 (year-level)	May 2026	Tier 1 (API prompt)	Web search tools
Moonshot	Kimi K2.6	April 2025	April 20, 2026	Tier 2	Tool use (no native browse)
MiniMax	MiniMax M3	January 2026	June 1, 2026	Tier 1 (HF template)	Tool use

OpenAI / ChatGPT knowledge cutoff dates

OpenAI publishes exact dates on each model page at developers.openai.com/api/docs/models^[3].

Model	API ID	Knowledge cutoff	Release	Context	Notes
GPT-5.5	gpt-5.5	December 1, 2025	Apr 23, 2026	1.05M	Current OpenAI flagship
GPT-5.5 Pro	gpt-5.5-pro	December 1, 2025	Apr 24, 2026	1.05M	Higher reasoning tier
GPT-5.4	gpt-5.4	August 31, 2025	Mar 5, 2026	1.05M	Frontier workhorse
GPT-5.4 Pro	gpt-5.4-pro	August 31, 2025	Mar 5, 2026	1.05M	Pro variant
GPT-5.4 mini	gpt-5.4-mini	August 31, 2025	Mar 17, 2026	400K	Fast, cheaper
GPT-5.4 nano	gpt-5.4-nano	August 31, 2025	Mar 17, 2026	400K	Lowest cost GPT-5.4
GPT-5.3 Chat	gpt-5.3-chat	August 31, 2025	Mar 3, 2026	—	ChatGPT conversation model
GPT-5.3 Codex	gpt-5.3-codex	August 31, 2025	Feb 24, 2026	—	Agentic coding
GPT-5.2	gpt-5.2	August 31, 2025	Dec 11, 2025	400K	Previous frontier
GPT-5.2 Pro	gpt-5.2-pro	August 31, 2025	Dec 10, 2025	—	Pro variant
GPT-5.2 Chat	gpt-5.2-chat	August 31, 2025	Dec 10, 2025	—	Chat variant
GPT-5.1	gpt-5.1	September 30, 2024	Nov 12, 2025	—	Older generation
GPT-5	gpt-5	September 30, 2024	Aug 7, 2025	—	First GPT-5 release
GPT-4.1	gpt-4.1	June 1, 2024	Apr 14, 2025	1M	Still available via API
GPT-4o	gpt-4o	October 1, 2023	May 13, 2024	128K	Retired from ChatGPT Feb 2026

Fact-check note: GPT-5.5 adds roughly four months of knowledge over GPT-5.4. If ChatGPT defaults to GPT-5.5 for your tier, anything after December 2025 needs browsing or pasted context.

Anthropic / Claude knowledge cutoff dates

Anthropic maintains tiered training corpora across its Claude model generations, distinguishing between base parametric training snapshots and real-time tool augmentation. Below is the verified knowledge cutoff schedule across Claude model classes:

Anthropic distinguishes **reliable knowledge cutoff** from **training data cutoff**. Source: [Models overview](#)^[4] and [Help Center](#)^[5].

Claude Fable 5 and Mythos class (above Opus)

Model	API ID	Reliable cutoff	Training cutoff	Release	Notes
Claude Fable 5	claude-fable-5	January 2026	January 2026	Jun 9, 2026	Mythos-class, GA. Default in Claude Code Pro/Max
Claude Mythos 5	claude-mythos-5	January 2026	January 2026	Jun 9, 2026	Limited availability (Project Glasswing)

Fable 5 sits above the Opus line. Same January 2026 cutoff as Opus 4.8, with safety classifiers that can refuse requests (Opus 4.8 does not).

Opus family

Model	API ID	Reliable cutoff	Training cutoff	Release
Claude Opus 4.8	claude-opus-4-8	January 2026	January 2026	May 28, 2026
Claude Opus 4.7	claude-opus-4-7	January 2026	January 2026	Apr 16, 2026
Claude Opus 4.6	claude-opus-4-6	May 2025	August 2025	Feb 5, 2026
Claude Opus 4.5	claude-opus-4-5	May 2025	August 2025	Nov 24, 2025
Claude Opus 4.1	claude-opus-4-1	January 2025	March 2025	Aug 5, 2025 (deprecated)

Sonnet and Haiku (current)

Model	API ID	Reliable cutoff	Training cutoff	Release
Claude Sonnet 4.6	claude-sonnet-4-6	August 2025	January 2026	Jan 17, 2026
Claude Haiku 4.5	claude-haiku-4-5	February 2025	July 2025	Oct 15, 2025
Claude Sonnet 4.5	claude-sonnet-4-5	January 2025	July 2025	Sep 29, 2025

Fact-check note: Fable 5 and Opus 4.8 share the freshest Anthropic cutoff at January 2026. Sonnet 4.6 has a newer training data cutoff (January 2026) but a lower reliable cutoff (August 2025), which means it may know recent facts unevenly.

Google Gemini knowledge cutoff dates

Google publishes cutoff dates on each model page. Source: [Gemini API models](#)^[6].

Model	Model ID	Knowledge cutoff	Release	Context	Notes
Gemini 3.5 Flash	gemini-3.5-flash	January 2025	May 2026 (GA)	1M	Flagship Flash, thinking on by default
Gemini 3.1 Pro Preview	gemini-3.1-pro-preview	January 2025	Feb 19, 2026	1M	Also gemini-3.1-pro-preview-customtools
Gemini 3.1 Flash Lite	gemini-3.1-flash-lite	January 2025	Mar 3, 2026	1M	Cost-optimized
Gemini 3 Flash Preview	gemini-3-flash-preview	January 2025	Dec 17, 2025	1M	Preview predecessor to 3.5 Flash
Gemini 3 Pro Image Preview	gemini-3-pro-image-preview	January 2025	Nov 20, 2025	65K	Image generation
Gemini 2.5 Pro	gemini-2.5-pro	January 2025	Mar 25, 2025	1M	Previous Pro generation
Gemini 2.5 Flash	gemini-2.5-flash	January 2025	Mar 25, 2025	1M	Previous Flash generation

Fact-check correction: Several third-party articles claim Gemini 3.5 Flash has a **January 2026** cutoff. Official Google documentation and the [Gemini 3.5 Flash FAQ](#)^[7] both state **January 2025**. The entire Gemini 3.x line shares the same static cutoff regardless of release date.

For post-January 2025 facts, Google expects you to use **Search Grounding**, not fresher weights.

Qwen knowledge cutoff dates

Alibaba does not publish a single cutoff page. Cutoffs vary by model generation and are embedded in DashScope API system prompts. Qwen maintainers warn that models [hallucinate their own cutoff dates](#)^[8] when asked directly.

Model / series	Example API IDs	Knowledge cutoff	Source	Notes
Qwen3.7 Max	qwen3.7-max, qwen3.7-max-2026-05-20	2026 (year-level)	Tier 1 (DashScope prompt)	Flagship, May 2026
Qwen3.7 Plus	qwen3.7-plus	2026	Tier 1 (DashScope prompt)	Cost-effective tier
Qwen3.6 Max Preview	qwen3.6-max-preview	2026	Tier 1 (DashScope prompt)	Apr 2026 preview
Qwen3.6 Plus	qwen3.6-plus	2026	Tier 1 (DashScope prompt)	Apr 2026
Qwen3.6 Flash	qwen3.6-flash	2026	Tier 1 (DashScope prompt)	Fast tier
Qwen3.5 Plus	qwen3.5-plus, qwen3.5-plus-2026-02-15	2026	Tier 1 (DashScope prompt)	Feb 2026
Qwen3.5 Flash	qwen3.5-flash, qwen3.5-flash-2026-02-23	2026	Tier 1 (DashScope prompt)	Vision-language Flash
Qwen3.5 397B A17B	qwen3.5-397b-a17b	2026	Tier 1 (DashScope prompt)	Open-weight flagship
Qwen3 Max	qwen3-max, qwen3-max-2026-01-23	June 30, 2025	Tier 2 (OpenRouter)	API tier, Sep 2025
Qwen3 Next 80B (Instruct / Thinking)	qwen3-next-80b-a3b-*	September 30, 2025	Tier 2 (OpenRouter)	MoE next-gen
Qwen3 2507 series	qwen3-235b-a22b-instruct-2507, qwen3-30b-a3b-*-2507	June 30, 2025	Tier 2 (OpenRouter)	2507 = July 2025 release tag
Qwen3 Coder Plus / Flash	qwen3-coder-plus, qwen3-coder-flash	June 30, 2025	Tier 2 (OpenRouter)	Coding-focused
Qwen3 Coder 480B A35B	qwen3-coder-480b-a35b	June 30, 2025	Tier 2 (OpenRouter)	Largest coder model
Qwen3 VL series	qwen3-vl-*	March 31, 2025	Tier 2 (OpenRouter)	Vision-language
Qwen3 dense / MoE base	qwen3-235b-a22b, qwen3-32b, qwen3-30b-a3b	March 31, 2025	Tier 2 (OpenRouter)	Original Qwen3 launch
Qwen2.5 series	qwen2.5-*	June 30, 2024	Tier 2 (OpenRouter)	Previous generation

How to read Qwen version tags: 2507 in a model name means a July 2025 release snapshot, not a July 2025 knowledge cutoff. The 2507 Instruct models carry a June 30, 2025 cutoff per OpenRouter metadata.

Fact-check note: Qwen 3.5 and above claim 2026 knowledge in official API system prompts, making them among the freshest open models. Older Qwen3 variants stop at March or June 2025. Do not ask the model what it knows. Check the API ID you are calling.

Kimi (Moonshot AI) knowledge cutoff dates

Moonshot does not publish a centralized cutoff table. Kimi model docs at platform.moonshot.ai^[9] list capabilities but omit cutoff dates.

Model	API ID	Knowledge cutoff	Release	Source	Notes
Kimi K2.7 Code	kimi-k2.7-code, kimi-k2.7-code-highspeed	Not published	Jun 12, 2026	—	Coding-focused, thinking always on
Kimi K2.6	kimi-k2.6	April 2025	Apr 20, 2026	Tier 2 (vendor specs via reviews)	Multimodal, 256K context, agent swarm
Kimi K2.5	kimi-k2.5	Not published	Jan 27, 2026	—	Visual agentic model
Kimi K2 Thinking	kimi-k2-thinking	Not published	Nov 6, 2025	—	Open reasoning variant
Kimi K2 0905	kimi-k2-0905	December 31, 2024	Sep 2025	Tier 2 (OpenRouter)	Original K2 release
Kimi K2 0711	kimi-k2-0711	December 31, 2024	Jul 2025	Tier 2 (OpenRouter)	Earlier K2 snapshot

Fact-check note: K2.6 specs cited in multiple technical reviews (training on 15.5T tokens, April 2025 cutoff) come from Moonshot release materials, not a standalone docs page I could fetch. OpenRouter still lists the original K2 snapshots at December 2024. Some API directories incorrectly list K2.6 as January 2025. I side with **April 2025** for K2.6 based on consistent vendor-spec reporting.

K2.5 and K2.7 Code cutoffs are **unconfirmed** as of June 2026. Treat as April 2025 or newer until Moonshot publishes official dates.

MiniMax knowledge cutoff dates

MiniMax embeds cutoff dates in HuggingFace chat templates for some models. The [release notes](#)^[10] do not list cutoffs.

Model	API ID	Knowledge cutoff	Release	Source	Notes
MiniMax M3	MiniMax-M3	January 2026	Jun 1, 2026	Tier 1 (HF chat template)	1M context, multimodal
MiniMax M2.7	MiniMax-M2.7, MiniMax-M2.7-highspeed	Not published	Mar 18, 2026	—	Self-evolving agent model
MiniMax M2.5	MiniMax-M2.5, MiniMax-M2.5-highspeed	January 2025	Feb 12, 2026	Tier 2 (llm-bento, reviews)	SWE-bench 80.2%
MiniMax M2.1	MiniMax-M2.1, MiniMax-M2.1-highspeed	June 2025	Dec 23, 2025	Tier 2 (OpenRouter, langmart)	Polyglot coding
MiniMax M2	MiniMax-M2	September 2025	Oct 27, 2025	Tier 2 (llm-bento)	First M2 open release
MiniMax M1	MiniMax-M1	June 30, 2024	Aug 2025	Tier 2 (OpenRouter)	Long-context reasoning
MiniMax-01	MiniMax-01	March 31, 2024	Jan 15, 2025	Tier 2 (OpenRouter)	Text + vision, 4M context

Fact-check note: M3 is the only MiniMax model with a Tier 1 cutoff (January 2026) baked into its official chat template. M2.5 at January 2025 means it is stale relative to M3 despite strong coding benchmarks. M2.7 cutoff is unpublished; likely between M2.5 and M3 based on release timing.

What the cutoff gap means in practice

Provider	Flagship	Cutoff	Release	Gap (months)
OpenAI	GPT-5.5	Dec 2025	Apr 2026	~4
Anthropic	Claude Fable 5	Jan 2026	Jun 2026	~5
Google	Gemini 3.5 Flash	Jan 2025	May 2026	~16
Qwen	Qwen3.7 Max	2026	May 2026	~0 to 5
Moonshot	Kimi K2.6	Apr 2025	Apr 2026	~12
MiniMax	MiniMax M3	Jan 2026	Jun 2026	~5

Gemini has the largest gap between release hype and static knowledge. GPT-5.5 and Claude Fable 5 are the freshest western frontier models. Qwen 3.5+ and MiniMax M3 compete on cutoff freshness among Chinese labs.

When cutoff dates matter for SEO and AEO

Training data vs live answers: ChatGPT with browsing, Gemini with Search Grounding, and Perplexity with native search can answer post-cutoff questions. The cutoff still matters for default answers, API calls without tools, and which pages get cited from memory versus retrieval.

Brand visibility: If your product launched after a model's cutoff and the model names a competitor instead, the fix is usually content plus retrieval, not waiting for retraining. See [how to improve brand visibility in AI search engines^{\[11\]}](#).

Do not trust one run: AI answers are non-deterministic. A model may know a pre-cutoff fact on one run and miss it on the next. Measure trends, not snapshots.

Frequently asked questions

Here are direct, authoritative answers to frequently asked questions about large language model knowledge cutoff dates, model retraining schedules, and retrieval augmented generation:

What are LLM knowledge cutoff dates?

LLM knowledge cutoff dates mark the point after which a model's training data ends. The model has no built-in knowledge of events, products, or news after that date unless it uses browsing, search grounding, or you provide context in the prompt.

Which LLM has the most recent knowledge cutoff?

As of June 2026, Claude Fable 5, Claude Opus 4.8, Qwen 3.5+, and MiniMax M3 claim January 2026 or full-year 2026 cutoffs. GPT-5.5 is close behind at December 1, 2025. Gemini 3.5 Flash and 3.1 Pro remain at January 2025 despite recent releases.

What is ChatGPT 5.5's knowledge cutoff?

GPT-5.5 has a knowledge cutoff of **December 1, 2025**, confirmed on the [official OpenAI model page](#)^[12]. GPT-5.4 and GPT-5.2 use August 31, 2025.

What is Claude Fable 5's knowledge cutoff?

Claude Fable 5 has a reliable knowledge cutoff of **January 2026** and a training data cutoff of **January 2026**, per Anthropic's models overview and Help Center. It released June 9, 2026.

What is Gemini 3.5 Flash's knowledge cutoff?

Gemini 3.5 Flash has a knowledge cutoff of **January 2025**, per [official Google documentation](#)^[13]. Blog posts claiming January 2026 are incorrect. Use Search Grounding for recent facts.

What is Gemini 3.1 Pro's knowledge cutoff?

Gemini 3.1 Pro Preview (`gemini-3.1-pro-preview`) has a knowledge cutoff of **January 2025**, same as the rest of the Gemini 3.x family.

Do Qwen models know their own cutoff date?

Usually not reliably. Qwen maintainers state models are not trained for self-identification and will give inconsistent cutoff answers. Use the API model ID and provider documentation instead of asking the model.

What is Kimi K2.6's knowledge cutoff?

Kimi K2.6 is widely reported at **April 2025** based on Moonshot release specifications cited in technical reviews. Moonshot has not published a standalone cutoff page. Original Kimi K2 snapshots (0711, 0905) use December 31, 2024 per OpenRouter.

What is MiniMax M3's knowledge cutoff?

MiniMax M3 has a knowledge cutoff of **January 2026**, embedded in its official HuggingFace chat template. MiniMax M2.5 uses January 2025 and MiniMax M2 uses September 2025 per secondary sources.

How often do knowledge cutoffs update?

Major western providers update roughly once or twice per year per model generation. Chinese labs release faster but publish cutoff data less consistently. A new model name does not guarantee a newer cutoff.

LLM knowledge cutoff dates decide which version of your brand ChatGPT, Perplexity and Gemini can actually cite. AEO and GEO work that ignores cutoffs is built on sand - if your AI citations depend on a fact the model does not know, no amount of answer engine optimization will surface it.

REFERENCES & SOURCES

- [1] **AI visibility** — <https://metehan.ai/articles/best-ai-visibility-tools>
- [2] **AEO** — <https://metehan.ai/articles/best-aeo-geo-tools>
- [3] **developers.openai.com/api/docs/models** — <https://developers.openai.com/api/docs/models>
- [4] **Models overview** — <https://platform.claude.com/docs/en/about-claude/models/overview>
- [5] **Help Center** — <https://support.claude.com/en/articles/8114494-how-up-to-date-is-claude-s-training-data>
- [6] **Gemini API models** — <https://ai.google.dev/gemini-api/docs/models>
- [7] **Gemini 3.5 Flash FAQ** — <https://ai.google.dev/gemini-api/docs/whats-new-gemini-3.5>
- [8] **hallucinate their own cutoff dates** — <https://github.com/QwenLM/Qwen3/issues/1442>
- [9] **platform.moonshot.ai** — <https://platform.moonshot.ai/docs/introduction>
- [10] **release notes** — <https://platform.minimax.io/docs/release-notes/models>
- [11] **how to improve brand visibility in AI search engines** — <https://metehan.ai/articles/how-to-improve-brand-visibility-in-ai-search-engines>
- [12] **official OpenAI model page** — <https://developers.openai.com/api/docs/models/gpt-5.5>
- [13] **official Google documentation** — <https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash>