

THE MODEL PULSE

ISSUE 22 · WEEK 38 OF 2026

THE BIG READ

Voice models stopped waiting for tools, and the architecture decision moved to concurrency, metering, and trust boundaries

KEY SIGNALS THIS PERIOD

2

Models added to the tree

3

Vendors shipped frontier-class

3

Architectural patterns crossed multiple vendors

2

Benchmark moves

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

September 19, 2026

Issue 22 · Weekly read

WEB

brianletort.ai/industry/model

The Model Pulse archive

THE BIG READ

The thesis this issue defends.

Independent industry analysis. Compiled from public model cards, vendor blogs, and leaderboards. The body sets the read for the rest of the brief.

Google's Gemini 3.8 Live and Live Extended Thinking are the only new closed-frontier model releases verified in the September 14-19 window. The architectural change is not simply better speech: the live model can keep listening and responding while tools run in the background, while Extended Thinking can reason and narrate progress at the same time. Google prices audio input at \$0.005 per minute and output at \$0.018 per minute, so an architect can now model the conversational meter separately from the cost and latency of the tools behind it.

Salesforce supplied the other consequential model event by post-training NVIDIA Nemotron 3 Super into Koa, a CRM reasoning model whose weights and inference remain inside Salesforce's trust boundary. That is a domain model in pilot, not a new foundation model, and its three-times-fewer-errors result is vendor-reported on Salesforce's own CRM benchmark. The open-weight side was quieter inside the window: Shanghai AI Laboratory published the technical report for Atria Dawn, a 744B Mixture-of-Experts agentic derivative whose weights appeared immediately before the window. The procurement read is to evaluate concurrency, meter composition, and deployment boundary alongside benchmark rank; none of those properties is captured by a single model score.

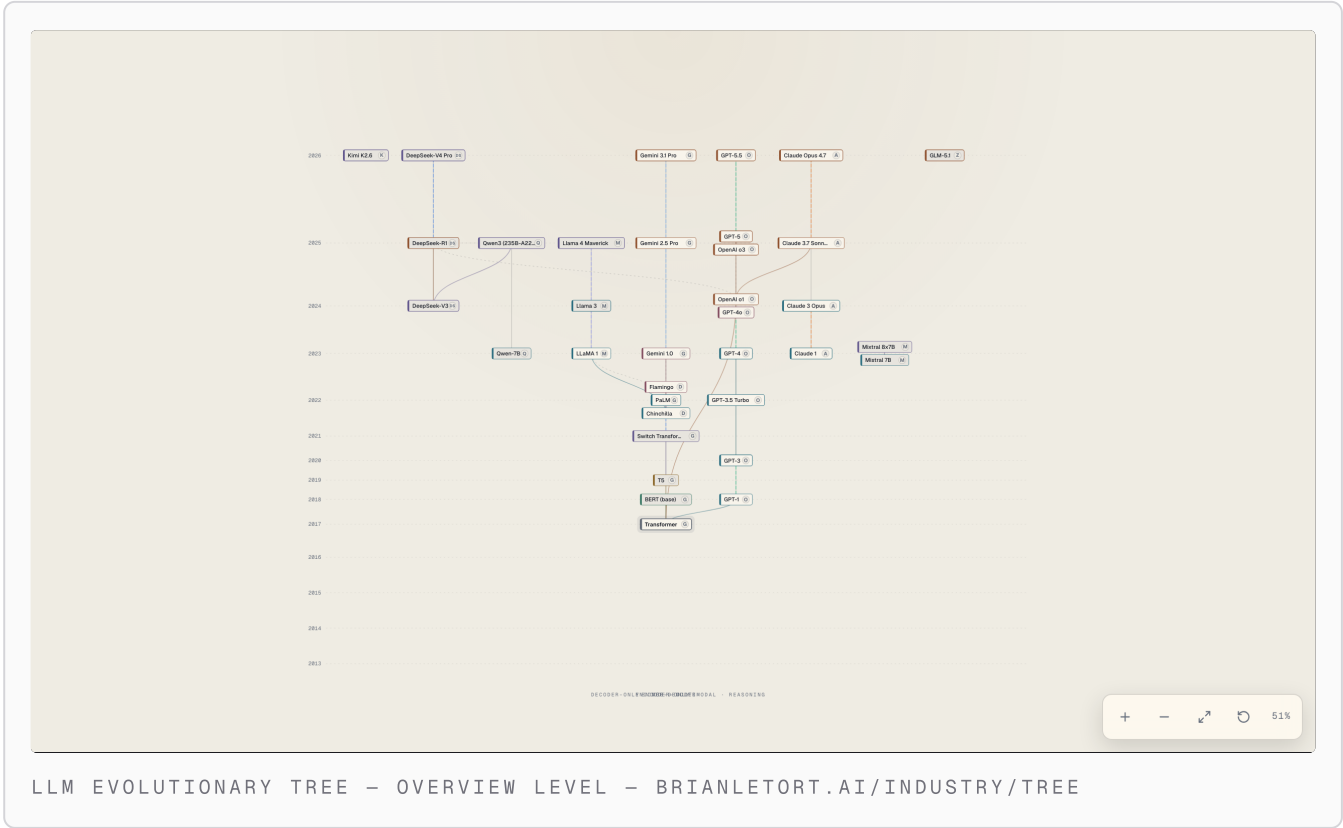
THE BAR

A model release matters when it changes one of three things: which workload runs on-prem, which tier of license is viable for a deployment, or which capability ceiling the closed frontier enforces. Single-vendor releases that change none of these are noise.

TREE DELTA

What changed in the tree.

Two rows added for Google's live model pair. They share a voice-first lineage but separate low-latency scale from deeper concurrent reasoning.



ADDED (2)

gemini-3-8-live

gemini-3-8-live-extended-thinking

UPDATED (0)

No updates to existing rows this period.

Koa is covered as a domain post-training event but is not added to the general LLM tree this week; Salesforce has not published model parameters, a model card, or weights. Atria Dawn's paper landed in-window, while its weights appeared before the Monday cutoff, so it is covered without being counted as a W38 tree addition.

FRONTIER MOVEMENTS

Flagship-class releases.

3 releases this period. Vendor-stated frontier capability. The releases that reset the closed-source ceiling.

2026-09-15

RELEASE 01

Google DeepMind

FRONTIER

MULTIMODAL

Gemini 3.8 Live

Native speech-to-speech model keeps dialogue active while visual grounding and tools run in the background

The procurement change is concurrency: the model can acknowledge a request and continue the conversation while an API call finishes, rather than forcing a silent turn boundary. Google publishes separate audio input and output rates, making the conversational layer measurable, but buyers still need to add backend tools and any downstream model charges to get the workflow cost.

SOURCE: GOOGLE

2026-09-15

RELEASE 02

Google DeepMind

REASONING

REASONING

Gemini 3.8 Live Extended Thinking

Concurrent deeper reasoning narrates progress without ending the live conversation

Google separates a higher-complexity live model that reasons and speaks simultaneously, including progress narration during multistep work. That is operationally useful for long tool chains, but enterprises receive it only in private preview initially, so production plans should distinguish developer API availability from enterprise support and governance.

SOURCE: GOOGLE

2026-09-15

RELEASE 03

Salesforce and NVIDIA

SPECIALIST

REASONING

Koa

CRM reasoning model post-trained from Nemotron 3 Super stays inside Salesforce's trust boundary

Koa is a domain-specialized derivative moving into select pilots, not a frontier foundation release. Salesforce controls the weights and performs post-training and inference within its own infrastructure, which is the procurement advantage; its three-times-fewer-errors benchmark claim remains vendor-run and should be reproduced on the buyer's own CRM actions.

SOURCE: SALESFORCE

OPEN WEIGHTS

Open-frontier and open-source drops.

1 releases this period. Open-weights drops that change procurement options. Pull these into pilot when score parity meets license parity.

2026-09-14

RELEASE 01

Shanghai AI Laboratory

OPEN FRONTIER

MOE

Atria Dawn Preview

744B Mixture-of-Experts agentic derivative publishes its technical report under an MIT-weight release

The report date falls inside the window, while the repositories appeared immediately before it, so this is a documentation event rather than a clean W38 weight launch. The permissive license and OpenAI-, Anthropic-, and Responses-compatible hosted interfaces lower switching friction, but a 744B system still carries a substantial serving floor and requires independent evaluation.

SOURCE: SHANGHAI AI LABORATORY TECHNICAL REPORT AND MODEL REPOSITORIES

ARCHITECTURE WATCH

Patterns to track.

3 architectural patterns that crossed multiple vendors this period. Each pattern names the trend, the exemplar releases, and what it changes for deployment, cost, or capability.

PATTERN 01 Conversation and execution become parallel planes

Gemini 3.8 Live

Gemini 3.8 Live Extended Thinking

The live model no longer has to stop speaking while a tool or deeper reasoning path runs. That improves perceived latency but creates concurrent failure and cost domains, so architecture reviews need explicit behavior for tool timeout, stale narration, cancellation, and a user interrupt that arrives while background work is still active.

SOURCE: GOOGLE

PATTERN 02 Domain post-training moves behind the system-of-record boundary

Salesforce Koa

NVIDIA Nemotron 3 Super

Salesforce post-trained an open NVIDIA base on synthetic CRM-shaped data, retained control of the resulting weights, and runs inference in its own trust boundary. That gives an incumbent a middle route between a general frontier API and a self-hosted model, but it also concentrates evaluation, benchmark design, and deployment control with one vendor.

SOURCE: SALESFORCE

PATTERN 03 Compatibility becomes an open-model distribution feature

Atria Dawn Preview

Atria Dawn pairs downloadable weights with hosted interfaces compatible with several dominant API shapes. The practical switching cost therefore moves away from request syntax and toward evaluation, tool semantics, safety policy, and the infrastructure needed to serve a 744B Mixture-of-Experts model.

SOURCE: SHANGHAI AI LABORATORY REPOSITORIES

BENCHMARK MOVES

Where the leaderboard moved.

2 benchmark deltas that change a procurement read. Scores reflect public leaderboards or vendor model cards as of publication.

ARTIFICIAL ANALYSIS SPEECH-TO-SPEECH LEADERBOARD

MOVE 01

Google reports Gemini 3.8 Live Extended Thinking at number one on the leaderboard at launch

Gemini 3.8 Live Extended Thinking	#1
Gemini 3.8 Live	Released for lower-cost scale

SOURCE: GOOGLE CITING ARTIFICIAL ANALYSIS

SALESFORCE CRM BENCHMARK

MOVE 02

Koa is reported to match or exceed leading models on CRM actions with three times fewer errors

Koa	3x fewer errors, vendor-reported
Leading general models	Comparison set not published

SOURCE: SALESFORCE

TIER SCORECARD

Who leads, who pushes.

Leader-vs-challenger by tier, useful for procurement shortlists when matching workload to model class. As of Sep 19, 2026.

TIER	LEADER	CHALLENGER	READ
Closed frontier	Claude Fable 5.1	GPT-6 Astra	No general frontier model displaced the W37 ordering; Reuters reported competitive pressure but no Anthropic release.
Open frontier	DeepSeek-V4.1-Flash	Atria Dawn Preview	DeepSeek remains the more documented serving proposition; Atria Dawn brings permissive agentic weights at a much larger serving floor.
Reasoning	Claude Fable 5.1	Gemini 3.8 Live Extended Thinking	The challenger leads a speech-specific leaderboard but has not displaced the standing general reasoning leader.
Coding	GPT-6 Astra	Claude Fable 5.1	No coding benchmark moved comparably inside the window, so W37's order is carried forward.
Multimodal	Gemini 3.8 Live Extended Thinking	Gemini 3.8 Live	The live pair owns this week's multimodal movement through native audio, visual grounding, and concurrent tools.
Edge / small	Nex-N2.5 mini	Desert Ant on-device model set	No in-window edge release supplied stronger verified evidence, so the prior ordering remains.

VENDOR SIGNALS

Pricing, gating, depreciation.

3 non-release moves that shift vendor risk — pricing, depreciations, gating decisions, license changes — each with a one-line procurement read.

2026-09-15

SIGNAL 01

Google

Live audio priced as separate input and output minute meters

At \$0.005 per input minute and \$0.018 per output minute, the voice layer becomes independently budgetable. Buyers must still add tool, reasoning, media infrastructure, and idle-conversation costs before comparing it with a cascaded architecture.

SOURCE: GOOGLE

2026-09-15

SIGNAL 02

Salesforce

Domain model offered in select pilots with US general availability expected in winter 2026

Koa is not generally available and its benchmark is vendor-controlled. Procurement should use the pilot to measure error rates on the organization's own CRM actions and preserve a general-model fallback.

SOURCE: SALESFORCE

2026-09-19

SIGNAL 03

Anthropic

Possible model release remains a reported deliberation rather than a product event

Reuters reported that Anthropic was considering a release amid investor scrutiny, but no model, date, price, or benchmark was announced. Treat it as competitive context and keep it out of production roadmaps.

SOURCE: REUTERS

WATCHLIST

On the radar next.

4 model-side catalysts in the next 7–30 days that would change the read materially. Watching these tells us whether the canopy is widening or thinning.

SEP 21–30

WATCH 01

Independent Gemini Live speech and interruption measurements

The published rate card needs latency, interruption, and tool-timeout measurements before it supports architecture selection.

OCT 2026

WATCH 02

Koa pilot benchmark disclosure

A public task set, comparison roster, and absolute error rates would determine whether the reported multiple is reproducible.

WINTER 2026

WATCH 03

Koa US general availability

General availability, regional scope, and model-catalog controls decide whether the specialist model moves beyond evaluation.

NEXT 30 DAYS

WATCH 04

Anthropic model decision

A release, delay, or explicit cancellation would turn this week’s reported deliberation into a procurement-relevant event.

METHODOLOGY AND AUTHORSHIP

How this brief is built.

Compiled from public model cards, vendor blogs, leaderboards, and official lab announcements. The publication is anchored to the LLM Evolutionary Tree at brianletort.ai/industry/tree, which the brief annotates each period.

WHAT EACH SECTION IS FOR

TREE DELTA	Model rows added or updated in content/llm-tree/models.yaml since the prior issue. Every id is real and clickable on the web view.
FRONTIER AND OPEN WEIGHTS	Releases that reset the closed ceiling or move the open-frontier line. Each card cites a single primary source.
ARCHITECTURE WATCH	Patterns that crossed multiple vendors in the period. Named once, exemplified by recent releases.
BENCHMARK MOVES	Public leaderboards and vendor model cards. Bars reflect the score range across the rows shown, not zero-baselined.
TIER SCORECARD	Leader vs challenger by tier. A snapshot for procurement shortlists; refreshed every issue.

AUTHORSHIP

Brian Letort

BrianLetort.AI · Independent analysis. All sources public. Not investment guidance.

THIS ISSUE

Issue 22

Week 38 of 2026 · September 19, 2026

WEB

brianletort.ai/industry/model

The Model Pulse archive, the LLM Evolutionary Tree, and the AI Stack Weekly companion publication.