# Mid-Training Design, Open Model Coalitions, and Inference Hardware Lead the Week

AI High Signal Digest

2026-03-22

## Mid-Training Design, Open Model Coalitions, and Inference Hardware Lead the Week

*By AI High Signal Digest • March 22, 2026*

PRISM supplied unusually concrete evidence that mid-training choices shape what later RL can unlock, while NVIDIA and Huawei made consequential moves in open models and inference hardware. The rest of the cycle brought notable advances in video learning, robotics, agent infrastructure, and AI compliance.

### Top Stories

*Why it matters:* The most consequential developments this cycle were about the infrastructure behind AI progress: how models are trained, how open ecosystems are organized, what hardware can lower inference costs, and how general robot models are being pushed toward precise control.

#### 1) PRISM turns mid-training into a measurable design problem

PRISM frames **mid-training** as a distinct stage between pretraining and RL, where targeted high-quality data mixtures build reasoning foundations. The project ran controlled experiments on roughly **27B** tokens across **7 models**, **4 families**, and **3B-24B** parameters, spanning dense Transformers and attention-Mamba hybrids, while measuring changes in performance, weights, representations, and downstream RL [1].

> "The single biggest lever in mid-training design is Data Composition." [2]

Across those ablations, math-only improved math, math+code improved math and code, and math+code+science produced the best overall results while most improving GPQA-Diamond during later RL [2]. The authors also reported that adding science during mid-training unlocked **+17 to +28 points** on GPQA-Diamond once RL was added later, while changing the RL data mix itself moved results by **less than 2 points** [2].

A separate timing result on Granite-4 Micro found that mid-training **after** long-context pretraining gave the largest gains in math, code, and science while preserving general reasoning; doing it at **8K** context hurt long-context ability, though much of that could be restored with a brief extension phase and model merging [3]. One practitioner summary distilled the practical upshot as **3-4x larger gains during later RL** when mid-training is tuned well beforehand, while other practitioners emphasized the work's value as a comprehensive disambiguation of a stage many teams already use [4, 5, 6]. Resources: project and paper [1].

**Impact:** PRISM makes mid-training look less like hidden craft knowledge and more like a controllable stage that determines what later RL can actually amplify [2, 3].

### 2) NVIDIA is trying to industrialize open model development with the Nemotron Coalition

NVIDIA announced the **Nemotron Coalition** with Black Forest Labs, Cursor, LangChain, Mistral AI, Perplexity, Reflection AI, Sarvam AI, and Thinking Machines Lab to develop the open-source **Nemotron** family of foundation models [7]. NVIDIA's stated idea is to build shared high-end base models that outperform what any single company could build alone, then let partners specialize them for different applications [7].

The first project is pretraining **Nemotron 4 base** with Mistral, with later post-training involving more partners. NVIDIA also outlined expected roles including multimodal work from Black Forest Labs, agent systems expertise from LangChain, evaluation datasets and real-world performance requirements from Cursor, and applied-system feedback from Perplexity [7].

**Impact:** This is a coordinated attempt to make open foundation models into shared industrial infrastructure rather than one-off lab releases [7].

### 3) Huawei is pushing an inference-focused hardware response with Atlas 350

Huawei launched the **Atlas 350** accelerator card, powered by its **950PR AI chip**, at the Ascend AI Partner Summit on March 20. According to the cited report, Huawei says the card delivers **2.87x** the single-card compute performance of NVIDIA's H20 and is currently the only product in China supporting **FP4** low-precision inference [8].

The same report lists **112GB HBM**, **60%** higher multimodal generation throughput, **4x** better memory-access efficiency for small operators, **1.56 PFLOPS** at FP4 precision, **1.4 TB/s** of memory bandwidth, and **600W**

TDP [8]. One expert note added that FP4 support matters especially for staying competitive in inference, even without native FP4 training [9].

**Impact:** The significance here is not just raw chip specs. It is whether domestic Chinese hardware can materially improve inference cost and throughput at a time when deployment efficiency matters more and more [8, 9].

### 4) Physical Intelligence's RL tokens target the precision gap in robotics

Physical Intelligence introduced **RL tokens** as compact snapshots of robot state that let a small model quickly learn and refine actions in real time [10]. The company argues the bottleneck for general-purpose robot models is often the **"last millimeter"** of precision, where broad competence is not enough [10].

Its method compresses high-dimensional VLA embeddings into a low-dimensional token, trains that token with a reconstruction objective, and then uses a small actor-critic module to learn residual action corrections directly on the robot through trial and error [10]. Reported results were robots that are up to **3x faster**, make fewer mistakes, can beat human teleoperation in some cases, and learn with as little as **15 minutes** of real-world practice. Full research: pi.website/research/rlt [10, 11].

**Impact:** The design separates general policy generation from fast local correction, which could be an important pattern for getting broad robot models to reliable task execution [10].

## Research & Innovation

*Why it matters:* The strongest research signals were about better use of depth, data, memory, and embodiment—areas that often move production systems more than a single benchmark headline.

- **Depth and information reuse: Attention Residuals** replaces fixed residual weights with attention over preceding layer outputs to reduce hidden-state dilution; in a **48B** model trained on **1.4T** tokens, the authors report better gradient distribution and consistent downstream gains [12]. **MoDA** tackles a similar problem by letting attention read key/value states from preceding layers, while keeping **97.3%** of FlashAttention-2 efficiency at **64K** context; in **1.5B** models it improved perplexity by **0.2** and downstream task scores by **2.11%** with a **3.7%** FLOP increase [13].
- **State-space sequence models: Mamba-3** combines discretized SSM recurrence, complex-valued state updates, and a multi-input/multi-output formulation. At **1.5B** parameters, it improved average accuracy by **1.8 points** over Gated DeltaNet while using half the state size of Mamba-2 [14].
- **Video and visual reasoning: V-JEPA 2.1** adds dense predictive loss, hierarchical self-supervision, and multimodal tokenizers, with reported

**20-point** gains in action anticipation and robotic grasping and new SOTA results on Ego4D, EPIC-KITCHENS, and TartanDrive [15]. **HopChain**, from Qwen and Tsinghua LeapLab, synthesizes chained visual-reasoning data for RLVR; added to Qwen3.5 VL training, it improved **20 of 24** benchmarks and topped **50 accuracy points** in the ultra-long-CoT regime [16].

- **Cheaper image generation:** Apple researchers' **Feature Auto-Encoder** trains diffusion models on compressed embeddings from a pretrained vision model, with up to **7x faster** training while keeping image quality comparable to state-of-the-art diffusion systems [17].
- **Memory and planning: GradMem** writes context into compact memory states by optimizing memory tokens at test time with a reconstruction loss, rather than only encoding context in a forward pass [18]. **Temporal Straightening** adds a curvature regularizer that makes latent trajectories more locally straight, aligning Euclidean and geodesic distances and improving goal-reaching success [19].
- **Evaluating scientific taste:** A paper on **Reinforcement Learning from Community Feedback** trained a "Scientific Judge" on **700,000** citation-matched paper pairs to predict research impact, then used it as a reward model for a "Scientific Thinker" that proposed higher-impact ideas than baselines [20].

## Products & Launches

*Why it matters:* Product teams kept translating model progress into working systems—faster agent infrastructure, more enterprise control, more local deployment, and new interfaces that treat existing software as the substrate.

- **OpenAI agent infrastructure:** OpenAI said agent workflows can now spin up containers for skills, shell, and code interpreter about **10x faster**. The change comes from a **container pool** in the Responses API that reuses warm infrastructure instead of creating a full container for each session; OpenAI also published a hosted shell quickstart [21].
- **Enterprise agent stack:** LangChain launched an enterprise agent platform built with NVIDIA AI. The stack supports AI-Q plus Deep Agents for enterprise search, shallow and deep research agents using Nemotron and frontier LLMs, LangSmith tracing, and connections to internal data through NeMo Agent Toolkit; LangChain linked a full guide [22].
- **Vision-native software control:** Mat Velloso's **Unswitch** prototype uses vision to operate existing software "more like a person does." He says prompts are a last resort, and demos show multi-tab research compiled into documents or slides, screenshots turned into formatted Excel sheets with formulas, and spatial organization across files, calendars, contacts, and email without replacing the underlying apps [23, 24, 25, 26, 27, 28]. The prototype runs natively on Mac and Windows and was built without JS or Python [29].

- **Offline local AI stack: Project N.O.M.A.D.** packages local AI models via Ollama + Open WebUI, full Wikipedia via Kiwix, offline maps, and a browser-based management UI into a system that runs without internet or telemetry after install. The project says it can be installed with one curl command on Debian-based systems and accessed across a local network as a headless server [30].
- **Agent skills as open source:** MiniMax open-sourced an official **skills** repository for agents, with curated skills for iOS and Android development, Office file editing, and GLSL visual effects [31].

## Industry Moves

*Why it matters:* Corporate moves this cycle point to the next layer of competition: monetization, leadership, sector-specific deployment, and the training infrastructure other labs quietly standardize on.

- **OpenAI monetization:** Reuters reported that OpenAI will begin showing ads to users of the **free** and **Go** versions of ChatGPT in the United States in the coming weeks [32, 33].
- **DeepMind leadership:** Google DeepMind appointed **Jas Sekhon** as chief strategy officer; Demis Hassabis highlighted Sekhon's prior role as Bridgewater's chief scientist and head of AI when introducing the hire [34].
- **AI in agriculture: Halter** reached a **$2B valuation**. Its product is AI-powered collars that let ranchers herd cattle from their phones using sound and vibration cues, and **Founders Fund** is leading the round [35].
- **Training stack standardization:** Multiple labs are reportedly using **Megatron** for training. Reflection AI and Periodic Labs were both cited, and one practitioner summarized the situation bluntly: for training MoEs, Megatron is "the only game in town" [36, 37, 38].

## Policy & Regulation

*Why it matters:* The legal and compliance edge of AI keeps moving from abstract debate to concrete distribution rules: authorship, app-store boundaries, and the operating cost of monitoring agents at scale.

- **Authorship:** A legal explainer emphasized that under U.S. law, AI-generated art without human authorship does **not** get copyright protection; brands building on AI art were urged to understand that ownership position clearly [39].
- **Platform rules for AI coding apps:** Replit said its App Store coding app has kept the same core generate-code, server-side compile, and webview-preview workflow for **4 years**, and that Apple eventually acknowledged it was not violating guidelines [40]. Follow-on commentary argued that the distinction between remotely hosted code and locally

downloaded-and-run code may become important if Apple tightens rules around AI coding webviews [41].
- **Compliance cost:** Fiddler's new TCO guide argues that evaluating agents with external LLMs creates a "Trust Tax" that can reach roughly **$2.6M per year**, because every trace adds external API cost on top of tooling fees [42].

## Quick Takes

*Why it matters:* These smaller updates give a useful read on where deployment is heading: cheaper local models, practical agent evaluation, developer ergonomics, and lighter-weight coding stacks.

- **Local deployment:** PinchBench results on **Qwen3.5 27B** using UnslothAI K_XL quantizations showed little degradation in best results; **Q4_K_XL** averaged about **84%** with thinking enabled, **Q3 KXL** remained viable at **14.5GB**, and a later non-thinking run made Q3 KXL the top performer for speed-conscious settings. One follow-up said this makes OpenClaw usable on a **16GB** card with decent reliability [43, 44, 45].
- **Autonomous research, reality check:** Karpathy's `autoresearch` package aims to let agents iterate on training code while humans iterate on prompts. In a real-scale test, Mikhail Parakhin ran **103** distributed experiments over a week and found one improvement, calling it a worse batting average than personal experimentation but still a "free" gain [46, 47].
- **Frontend generation:** OpenAI published frontend guidance for **GPT-5.4** after one developer said the model can produce "pretty great frontends" when used with enough thought and intentionality [48, 49].
- **Agent monitoring:** LangChain published a conceptual guide arguing that agent observability needs a distinct production playbook because natural-language input is unbounded, prompts are sensitive to small changes, and multi-step reasoning is hard to anticipate in development [50].
- **Memory footprint:** T3 Code claimed significantly lower RAM usage than Claude Code in one comparison—**350.9 MB** versus **635.5 MB**—and said its Electron app was **2x** more efficient than a Bun CLI in that setup [51, 52].
- **Model release watch: MiniMax-M2.7-highspeed** was spotted inside OpenCode without specs yet [53, 54], and **GLM-5.1** was teased as an incoming release [55].
- **Hiring signal:** One engineer said interview loops are already changing in light of LLMs, with less weight on LeetCode-style screening [56].

---

### Sources

1. X post by @bharatrunwal2
2. X post by @bharatrunwal2
3. X post by @bharatrunwal2
4. X post by @cwolferesearch
5. X post by @code_star
6. X post by @cwolferesearch
7. X post by @TheTuringPost
8. X post by @jukan05
9. X post by @teortaxesTex
10. X post by @TheTuringPost
11. X post by @TheTuringPost
12. X post by @TheAITimeline
13. X post by @TheAITimeline
14. X post by @TheAITimeline
15. X post by @TheAITimeline
16. X post by @ShenzhiWang_THU
17. X post by @DeepLearningAI
18. X post by @yurakuratov
19. X post by @TheAITimeline
20. X post by @TheAITimeline
21. X post by @OpenAIDevs
22. X post by @LangChain
23. X post by @matvelloso
24. X post by @matvelloso
25. X post by @matvelloso
26. X post by @matvelloso
27. X post by @matvelloso
28. X post by @matvelloso
29. X post by @matvelloso
30. X post by @godofprompt
31. X post by @MiniMax_AI
32. X post by @Reuters
33. X post by @kimmonismus
34. X post by @demishassabis
35. X post by @TheRundownAI
36. X post by @eliebakouch
37. X post by @_lewtun
38. X post by @TheZachMueller
39. X post by @LearnOpenCV
40. X post by @amasad
41. X post by @paul_cal
42. X post by @TheTuringPost
43. X post by @TheZachMueller
44. X post by @TheZachMueller

45. X post by @TheZachMueller
46. X post by @karpathy
47. X post by @MParakhin
48. X post by @sherwinwu
49. X post by @gdb
50. X post by @LangChain
51. X post by @theo
52. X post by @theo
53. X post by @ivanfioravanti
54. X post by @MiniMax_AI
55. X post by @kimmonismus
56. X post by @fleetwood___