Qwen 3.5 small models go multimodal-on-edge as U.S. AI procurement tightens and Codex Spark ramps speed

AI High Signal Digest

2026-03-03

Qwen 3.5 small models go multimodal-on-edge as U.S. AI procurement tightens and Codex Spark ramps speed

By AI High Signal Digest • March 3, 2026

A packed cycle led by Alibaba's Qwen 3.5 small multimodal models (and the architecture tricks enabling long context on consumer hardware), major U.S. procurement moves impacting Anthropic and OpenAI's DoW contract language, and a new speed push in coding models via GPT-5.3 Codex "Spark." Also included: key research on on-device training, attention mechanisms, agent frameworks, and a roundup of notable product launches.

Top Stories

1) Qwen 3.5 "small" models push multimodal + long-context onto consumer/edge hardware

Why it matters: The latest open(-ish) model drop isn't just about smaller checkpoints—it's about **making multimodal agents practical on constrained devices** and keeping long-context usable without falling off a performance cliff.

Alibaba Qwen released the **Qwen 3.5 Small Model Series** (0.8B, 2B, 4B, 9B) and said base models are included as well [1]. The pitch: "more intelligence, less compute," with **native multimodal** capability, an improved architecture, and scaled RL [1].

A technical breakdown highlights **Gated DeltaNet hybrid attention**: a 3:1 mix of linear attention layers to full attention layers, intended to keep memory flat while preserving quality—cited as enabling even the **0.8B** model to support

a **262K-token context window** [2]. The models handle text/images/video natively, with a vision encoder using **3D convolutions** to capture motion in video and merging features from multiple layers (no post-hoc adapter) [2].

Practical distribution signal: Qwen 3.5 small is already broadly runnable via common local stacks—**Ollama** (tool calling/thinking/multimodal) [3], **LM Studio** (Qwen3.5-9B, ~7GB to run locally) [4, 5], and **MLX** [6].

2) U.S. government AI procurement escalates: Treasury ends Claude use, while OpenAI updates DoW surveillance language

Why it matters: "AI policy" is being expressed through **department-by-department procurement decisions** and **contract amendments**—and the exact wording is now under intense scrutiny.

- **U.S. Treasury** said it is terminating all use of Anthropic products, including Claude, within the department [7, 8].
- Separately, Sam Altman said OpenAI worked with the "Department of War" (DoW) to add explicit language that the AI system **"shall not be intentionally used for domestic surveillance of U.S. persons and nationals"**, including a stated prohibition on deliberate tracking via **commercially acquired personal/identifiable information** [9].
- Altman also said the DoW affirmed OpenAI services will not be used by DoW intelligence agencies (e.g., NSA) without a follow-on contract modification [9]. Another thread summarizes that deployment to NSA/DoW intelligence agencies will be withheld "for now" to allow time to address potential surveillance loopholes through democratic processes [10].

3) Coding-model speed race intensifies: GPT-5.3 Codex "Spark" preview lands on Plus for heavy Codex users

Why it matters: For coding agents, **throughput and latency** increasingly determine what workflows are viable (e.g., multi-step refactors, large code migrations, parallel execution).

OpenAI's Codex team is rolling out **GPT-5.3-Codex-Spark** to "most engaged Codex subs" on **ChatGPT Plus**, describing it as their **fastest model yet** at **well over 1K tokens/sec** [11]. Another post echoes: "Spark's the fastest model we've ever made," rolling out to the heaviest Codex users on Plus [12]. The rollout is positioned as a **temporary preview at no extra cost through March 20** [11].

4) Anthropic's consumer + developer momentum continues amid product shipping burst

Why it matters: Distribution (app ranking) and workflow features (memory, remote control, scheduled tasks) can translate into durable adoption—even while government procurement turns volatile.

An Anthropic insider described a week featuring **Claude hitting #1 on the App Store**, record sign-ups, and servers "melting," alongside shipping "a ton of new stuff" quickly [13]. Separately, Claude's **Memory** is now available on the **free plan**, with easier import and the ability to export memories anytime [14].

Research & Innovation

Why it matters: This cycle's research signal clusters around two themes: (1) **making models cheaper to run/train locally**, and (2) **agent architecture + evaluation maturity**.

Training and inference efficiency

- **Reverse-engineering Apple's Neural Engine (ANE) for training**: A researcher built a transformer training loop that runs forward/backward passes directly on ANE hardware via undocumented APIs, bypassing CoreML [15, 16]. Reported metrics include M4 ANE at **6.6 TFLOPS/W vs 0.08 for an A100 (80× more efficient)** and a claim that "38 TOPS" marketing corresponds to **19 TFLOPS FP16** throughput [16].
- **Nebius "LK Losses" for speculative decoding**: new objectives that directly optimize speculative decoding acceptance rate, reporting **8–10% gains** over KL minimization across multiple draft architectures and target models [17].
- **MLRA (Multi-Head Low-Rank Attention)**: described as supporting **native 4-way tensor parallelism**, achieving **2.8× decoding speedup** over MLA at 2.9B scale while improving perplexity and zero-shot commonsense benchmarks [18].

Agent architecture, evaluation, and coordination

- **Auton Agentic AI Framework (Snapchat paper)** proposes standardized patterns for integrating reasoning, memory systems, tool usage, and planning to reduce brittleness and improve reproducibility of agentic systems [19].
- Research on **hierarchies emerging in multi-agent systems** studies how flat cooperative groups can transition into hierarchical organizations, identifying mechanisms and conditions that drive the shift [20].

Formal methods and theorem proving

- **Minimal agent architecture for theorem proving**: research describes a deliberately streamlined Lean-interfacing agent that aims for competitive proof generation performance with improved reproducibility and accessibility [21].
- **Large Lean formalization milestone**: mathematics.inc reported a **~200K LOC** formalization (using Gauss) of Maryna Viazovska's 2022

Fields Medal theorems on optimal sphere packing in dimensions 8 and 24, described as the only Fields Medal-winning result from this century to be completely formalized and the largest single-purpose Lean formalization in history [22].

Products & Launches

Why it matters: Product releases are converging on a "full-stack agent workflow": model access, orchestration, observability, and deployable endpoints.

Local model availability (Qwen 3.5)

- **Ollama**: Qwen 3.5 small models are available with native tool calling/thinking/multimodal support [3].
- **LM Studio**: Qwen3.5-9B is available for local download/run (~7GB) and supports image input + tool calling [4, 5].
- **On-device**: A demo shows Qwen 3.5 running on-device on an iPhone 17 Pro; the 2B 6-bit model uses MLX optimized for Apple Silicon [23].

Document parsing / OCR

- **FireRed-OCR-2B**: reported as #1 on OmniDocBench v1.5 with **92.94%** end-to-end document parsing, and top results on OCRBench TextRec (**93.5**) [24]. The post claims it can run in **4–5GB VRAM** (bfloat16) on a single RTX 3090 and is **Apache 2.0** licensed for commercial use [24].

Video and media tooling

- **Runway Gen-4.5**: introduced as a frontier video model emphasizing motion quality/prompt adherence/visual fidelity [25]; in Video Arena's Text-to-Video leaderboard it scores **1218** and is tied for **#15** [26].
- **fal.ai "Video Depth Anything"**: optimized, production-ready version with 3 model sizes, up to 1080p, side-by-side comparison outputs, and raw depth export [27].
- **Grok Imagine**: video extension is now available in-app (extend generated videos) [28].

Developer observability and workflow

- **GitHub Copilot**: OpenTelemetry support is coming, aimed at observability into the agent loop [29, 30].
- **nCompass_tech VSCode extension v0.1.0**: unifies profiling + trace collaboration + analysis, with an AI agent integrated into Cursor / Claude Code; they report improving a Hopper GEMM kernel from **30% slower to 3% faster** than a near-optimal CUTLASS kernel within a day [31].

Industry Moves

Why it matters: Competitive dynamics are now shaped as much by **distribution + platform positioning** as by raw model capability.

- **MiniMax earnings (first as HKEX public company)**: 2025 results include **$79M revenue (+159% YoY)**, **70%+ international**, gross margin **12.2% → 25.4%**, **236M+ users** across 200+ countries, and **214K enterprise clients & developers** [32]. MiniMax says 2026 focus is evolving from a model company into an **AI platform** focused on coding, workplace productivity, and multimodal creation [32].
- **China AI adoption (QuestMobile, Dec 2025)**: top five active user bases listed as Doubao **226M**, DeepSeek **135M**, Tencent Yuanbao **41M**, Alibaba Ant Afu **27M**, Alibaba Qianwen **25M** [33].
- **Google DeepMind hiring**: NiJinjie joined as a research scientist working on **Gemini scaling and RL** under Yi Tay and Quoc Le [34].

Policy & Regulation

Why it matters: Contract clauses and legal interpretations are becoming operational constraints (or loopholes) for AI deployment—especially for national security use.

OpenAI–DoW contract language dispute: "lawful purposes" and what "applicable law" means

- Commentary highlights the contract language: "The Department of War may use the AI System for all lawful purposes" [35].
- OpenAI publicly claimed the agreement references surveillance and autonomous weapons laws/policies "as they exist today," implying standards stay aligned even if laws change [36].
- OpenAI's Head of National Security Partnerships said their intention is that "applicable law" means law applicable **at the time the contract is signed** [37, 38].
- A legal counter-analysis argues "lawful purposes" is inherently **ambulatory** (law at time of performance), citing longstanding doctrine and Supreme Court precedent, and says OpenAI's "lock in current standards" claim is not supported by the excerpted language [39, 40, 41].

Ongoing debate about surveillance loopholes in the amended language

Even after the added clause, critics argue terms like "intentional" and "deliberate" may still allow "incidental" loopholes [42]. Another critique questions how terms apply to metadata/identifiers and non-surveillance uses of derived outputs [43].

Musk vs OpenAI "AGI" claims circulate as a legal narrative

Posts about the Musk vs OpenAI case claim court documents indicate OpenAI leaders internally considered **GPT-4o** to be AGI [44]. Another thread claims Musk is arguing GPT-4o constitutes AGI and that this would void Microsoft's exclusive license and require public availability under OpenAI's founding agreement [45].

Quick Takes

Why it matters: These are smaller signals that often foreshadow where product and research effort is about to concentrate.

- **Liquid AI** released **LFM2.5-1.2B-Thinking**, a 1.17B-parameter reasoning model reported to run under **900MB RAM** and operate about twice as fast as similar models, designed for small-device agents and local workflows without cloud compute [46].
- **Cognition** previewed **SWE-1.6**, reporting **950 tok/s** and "exceeds top open-source models" on SWE-Bench Pro, while noting overthinking/self-verification issues in the preview [47].
- **Telegram**: all chatbots can now stream responses in real time, positioned as helpful for AI assistants [48].
- **Grok**: Elon Musk announced **Grok 4.20 Beta 2** is out with release notes [49].
- **BullshitBench v2**: new version adds 100 questions across domains and tests 70+ model variants; claims include Anthropic models scoring exceptionally well and reasoning having a negative effect on BS detection [50].

Sources

1. X post by @Alibaba_Qwen
2. X post by @LiorOnAI
3. X post by @ollama
4. X post by @Alibaba_Qwen
5. X post by @lmstudio
6. X post by @Alibaba_Qwen
7. X post by @SecScottBessent
8. X post by @DeItaone
9. X post by @sama
10. X post by @polynoamial
11. X post by @ah20im
12. X post by @jeffintime
13. X post by @alexalbert___
14. X post by @claudeai
15. X post by @LiorOnAI

16. X post by @AmbsdOP
17. X post by @HuggingPapers
18. X post by @papers_anon
19. X post by @dair_ai
20. X post by @dair_ai
21. X post by @omarsar0
22. X post by @mathematics_inc
23. X post by @adrgrondin
24. X post by @ModelScope2022
25. X post by @runwayml
26. X post by @arena
27. X post by @fal
28. X post by @grok
29. X post by @pierceboggan
30. X post by @pierceboggan
31. X post by @adityaraja0
32. X post by @MiniMax_AI
33. X post by @Sino_Market
34. X post by @NiJinjie
35. X post by @jeremyphoward
36. X post by @jeremyphoward
37. X post by @jeremyphoward
38. X post by @natseckatrina
39. X post by @jeremyphoward
40. X post by @jeremyphoward
41. X post by @jeremyphoward
42. X post by @NickEMoran
43. X post by @David_Kasten
44. X post by @Seltaa_
45. X post by @kexicheng
46. X post by @DeepLearningAI
47. X post by @cognition
48. X post by @durov
49. X post by @elonmusk
50. X post by @petergostev