

The Foundation Model Frontier: Voice AI, Shopping Agents & Enterprise Shifts

Executive Summary

This week saw major leaps in foundation-model capabilities across tech. Google launched Gemini 3.1 Flash Live – a low-latency, voice-first model – and used its Gemini AI to power a new dark-web monitoring service for enterprise security ([www.androidcentral.com \[1\]](https://www.androidcentral.com/apps-software/ai/gemini-3-1-flash-live-is-a-massive-boon-to-the-ai-real-time-assistance-for-you-and-me#:~:text=the%20launch%20of%20a%20new%2C,Gemini%2C%20you%20can%20probably%20guess)) ([www.itpro.com \[2\]](https://www.itpro.com/security/google-just-launched-a-new-gemini-powered-dark-web-monitoring-service#:~:text=The%20new%20dark%20web%20intelligence,Wood%20and%20product%20marketing%20manager)). OpenAI updated ChatGPT with image-based shopping and set a 4% checkout fee for Shopify sales ([www.techradar.com \[3\]](https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage#:~:text=OpenAI%20has%20announced%20a%20new,side%20comparisons)) ([www.techradar.com \[4\]](https://www.techradar.com/pro/what-openai-4-percent-checkout-fee-means-for-the-future-of-commerce#:~:text=OpenAI%20recently%20introduced%20a%204,step%20in%20claiming%20that%20ground)). Microsoft and Salesforce debuted new AI agent tools (Copilot Cowork with Claude and an AI “Foundry”) to automate workflows. Meanwhile, Google’s TurboQuant breakthrough cuts model memory demand 6× and boosts speed 8× ([www.tomshardware.com \[5\]](https://www.tomshardware.com/tech-industry/artificial-intelligence/googles-turboquant-compresses-llm-kv-caches-to-3-bits-with-no-accuracy-loss#:~:text=Google%20Research%20published%20TurboQuant%20on,keys%2C%20while%20reducing%20KV%20cache)), pointing to much cheaper compute. For business leaders, these shifts mean planning for richer AI features and new economics in the coming year.

References:

- [1] [www.androidcentral.com](https://www.androidcentral.com/apps-software/ai/gemini-3-1-flash-live-is-a-massive-boon-to-the-ai-real-time-assistance-for-you-and-me#:~:text=the%20launch%20of%20a%20new%2C,Gemini%2C%20you%20can%20probably%20guess) — <https://www.androidcentral.com/apps-software/ai/gemini-3-1-flash-live-is-a-massive-boon-to-the-ai-real-time-assistance-for-you-and-me#:~:text=the%20launch%20of%20a%20new%2C,Gemini%2C%20you%20can%20probably%20guess>
- [2] [www.itpro.com](https://www.itpro.com/security/google-just-launched-a-new-gemini-powered-dark-web-monitoring-service#:~:text=The%20new%20dark%20web%20intelligence,Wood%20and%20product%20marketing%20manager) — <https://www.itpro.com/security/google-just-launched-a-new-gemini-powered-dark-web-monitoring-service#:~:text=The%20new%20dark%20web%20intelligence,Wood%20and%20product%20marketing%20manager>
- [3] [www.techradar.com](https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage#:~:text=OpenAI%20has%20announced%20a%20new,side%20comparisons) — <https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage#:~:text=OpenAI%20has%20announced%20a%20new,side%20comparisons>
- [4] [www.techradar.com](https://www.techradar.com/pro/what-openai-4-percent-checkout-fee-means-for-the-future-of-commerce#:~:text=OpenAI%20recently%20introduced%20a%204,step%20in%20claiming%20that%20ground) — <https://www.techradar.com/pro/what-openai-4-percent-checkout-fee-means-for-the-future-of-commerce#:~:text=OpenAI%20recently%20introduced%20a%204,step%20in%20claiming%20that%20ground>
- [5] [www.tomshardware.com](https://www.tomshardware.com/tech-industry/artificial-intelligence/googles-turboquant-compresses-llm-kv-caches-to-3-bits-with-no-accuracy-loss#:~:text=Google%20Research%20published%20TurboQuant%20on,keys%2C%20while%20reducing%20KV%20cache) — <https://www.tomshardware.com/tech-industry/artificial-intelligence/googles-turboquant-compresses-llm-kv-caches-to-3-bits-with-no-accuracy-loss#:~:text=Google%20Research%20published%20TurboQuant%20on,keys%2C%20while%20reducing%20KV%20cache>

New Model Features and Multimodal AI

Tech giants are rolling out specialized model variants and multimodal AI upgrades. For example, on Mar 27 Google introduced **Gemini 3.1 Flash Live**, a lightweight, low-latency voice model billed as its “highest-quality audio and voice model” ([www.androidcentral.com \[1\]](https://www.androidcentral.com/apps-software/ai/gemini-3-1-flash-live-is-a-massive-boon-to-the-ai-real-time-assistance-for-you-and-me#:~:text=the%20launch%20of%20a%20new%2C,Gemini%2C%20you%20can%20probably%20guess)). By focusing on a “voice-first” pipeline, Gemini 3.1 delivers faster, more natural responses and can even follow twice the dialogue context of its predecessors. In business terms, expect voice-driven assistants (in phone systems, apps, etc.) to become far more capable.

Anthropic made smaller tweaks to its coding assistant: Claude Code gained an “auto mode” that lets it grant itself safe permissions without user approval ([www.techradar.com \[2\]](https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage#:~:text=OpenAI%20has%20announced%20a%20new,side%20comparisons)). In practice this reduces frequent stops for developer sign-off on mundane tasks (file opens, analysis, etc.), smoothing long coding sessions. It’s an incremental change, but it hints at future AI tools that handle whole processes with minimal human prompts.

OpenAI is pushing into commerce with ChatGPT. In late March it unveiled a new **shopping interface**: users can upload images of products, set budgets, and get side-by-side visual comparisons ([www.techradar.com \[3\]](https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage#:~:text=OpenAI%20has%20announced%20a%20new,side%20comparisons)). ChatGPT is effectively becoming a visual shopping assistant.

Alongside the UI, OpenAI began charging a **4% checkout fee** for Shopify vendors through ChatGPT (www.techradar.com [4]). Although cheaper than Amazon's referral rates (8–15%), this fee signals a bigger strategy: control of conversational commerce. Retailers and brands should note that customers may soon browse and buy directly via chat interfaces, and platform fees will reshape e-commerce economics.

References:

- [1] [www.androidcentral.com — https://www.androidcentral.com/apps-software/ai/gemini-3-1-flash-live-is-a-massive-boon-to-the-ai-real-time-assistance-for-you-and-me#:~:text=the%20launch%20of%20a%20new%20C,Gemini%2C%20you%20can%20probably%20guess](https://www.androidcentral.com/apps-software/ai/gemini-3-1-flash-live-is-a-massive-boon-to-the-ai-real-time-assistance-for-you-and-me#:~:text=the%20launch%20of%20a%20new%20C,Gemini%2C%20you%20can%20probably%20guess)
- [2] [www.techradar.com — https://www.techradar.com/pro/anthropic-gives-claude-code-new-auto-mode-which-lets-it-choose-its-own-permissions#:~:text=Anthropic%20has%20launched%20a%20new,allowing%20safe%20actions%20automatically%20but](https://www.techradar.com/pro/anthropic-gives-claude-code-new-auto-mode-which-lets-it-choose-its-own-permissions#:~:text=Anthropic%20has%20launched%20a%20new,allowing%20safe%20actions%20automatically%20but)
- [3] [www.techradar.com — https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage#:~:text=OpenAI%20has%20announced%20a%20new,side%20comparisons](https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage#:~:text=OpenAI%20has%20announced%20a%20new,side%20comparisons)
- [4] [www.techradar.com — https://www.techradar.com/pro/what-openais-4-percent-checkout-fee-means-for-the-future-of-commerce#:~:text=OpenAI%20recently%20introduced%20a%204,step%20in%20claiming%20that%20ground](https://www.techradar.com/pro/what-openais-4-percent-checkout-fee-means-for-the-future-of-commerce#:~:text=OpenAI%20recently%20introduced%20a%204,step%20in%20claiming%20that%20ground)

AI Agents in Enterprise Workflows

Beyond models themselves, companies are embedding AI as autonomous assistants. Microsoft announced wider availability of **Copilot Cowork**, integrating Anthropic's Claude to automate tasks for frontline teams (www.itpro.com [1]). Cowork offers plug-ins in areas like data analysis, marketing and legal, letting employees delegate tasks to AI agents. Microsoft's blog boasts Copilot Cowork "makes it easy to delegate work" and can "complete tasks, run workflows and do work on your behalf" (www.itpro.com [2]). In practical terms, expect routine research or document tasks to be handed off to AI pilots in tools like Teams or Excel. Salesforce is taking a similar tack: it launched an **AI Foundry** to fast-track "agentic" features from research to product (www.itpro.com [3]). This means Salesforce's CRM and analytics apps will soon include more autonomous features (like AI-written reports or lead scoring done by bots) – a downstream benefit for business users.

AI is also being applied to security operations. Google rolled out an AI-driven **dark web monitoring** service as part of its Threat Intelligence suite (www.itpro.com [4]). It uses Gemini to continuously scan hacker forums and breach data for any mentions of a customer's organization. According to Google, this creates an "organizational profile" specific to your business that updates as your operations change. For enterprises, this can streamline threat intel: instead of manually searching leaks, an AI watches the web. Security leaders should prepare to integrate such tools into their SOCs, while also noting Google's own report that attackers are using AI themselves for reconnaissance (so the arms race continues).

Many core business functions will increasingly wrap foundation-model AI in specialized guises. At Oracle's AI World Tour, for example, the company unveiled 22 new "Fusion Agentic Applications" (sales forecasting, supply-chain coordination, etc.) built on generic AI agents (www.itpro.com [5]). Oracle's pitch is a reliable cloud that lets customers pick any model and data platform (www.itpro.com [6]) – a nod to the emerging hybrid ecosystem. In essence, enterprises need to view new AI abilities as platform capabilities: easy plugins for workflows, rather than standalone experiments.

References:

- [1] [www.itpro.com — https://www.itpro.com/technology/artificial-intelligence/microsoft-is-rolling-out-copilot-cowork-to-more-customers#:~:text=Microsoft%20has%20announced%20that%20Copilot,of%20tasks%20and%20boost%20productivity](https://www.itpro.com/technology/artificial-intelligence/microsoft-is-rolling-out-copilot-cowork-to-more-customers#:~:text=Microsoft%20has%20announced%20that%20Copilot,of%20tasks%20and%20boost%20productivity)
- [2] [www.itpro.com — https://www.itpro.com/technology/artificial-intelligence/anthropics-claude-cowork-tool-is-coming-to-microsoft-copilot#:~:text=In%20a%20blog%20post%20announcing,he](https://www.itpro.com/technology/artificial-intelligence/anthropics-claude-cowork-tool-is-coming-to-microsoft-copilot#:~:text=In%20a%20blog%20post%20announcing,he)
- [3] [www.itpro.com — https://www.itpro.com/technology/artificial-intelligence/salesforce-ramps-up-agentic-ai-research-with-new-foundry-project#:~:text=Salesforce%20has%20announced%20the%20launch,strategic](https://www.itpro.com/technology/artificial-intelligence/salesforce-ramps-up-agentic-ai-research-with-new-foundry-project#:~:text=Salesforce%20has%20announced%20the%20launch,strategic)
- [4] [www.itpro.com — https://www.itpro.com/security/google-just-launched-a-new-gemini-powered-dark-web-monitoring-service#:~:text=The%20new%20dark%20web%20intelligence,Wood%20and%20product%20marketing%20manager](https://www.itpro.com/security/google-just-launched-a-new-gemini-powered-dark-web-monitoring-service#:~:text=The%20new%20dark%20web%20intelligence,Wood%20and%20product%20marketing%20manager)
- [5] [www.itpro.com — https://www.itpro.com/cloud/public-cloud/oracle-leans-hard-on-one-size-fits-all-appeal-of-oci-for](https://www.itpro.com/cloud/public-cloud/oracle-leans-hard-on-one-size-fits-all-appeal-of-oci-for)

enterprises#:~:text=At%20Oracle%20AI%20World%20Tour,operations%2C%20and%20automating%20cash%20flow
[6] www.itpro.com — <https://www.itpro.com/cloud/public-cloud/oracle-leans-hard-on-one-size-fits-all-appeal-of-oci-for-enterprises#:~:text=Sicilia%20argued%20that%20only%20with,Sicilia%20said>

Cost, Infrastructure and Ecosystem Trends

The economic and operational backdrop is shifting to make frontier models more accessible. A key example is Google's **TurboQuant** research. By quantizing the LLM's key-value cache to just 3 bits, TurboQuant cuts memory usage by roughly 6× without hurting accuracy (www.tomshardware.com [1]). On modern GPUs (Nvidia H100), this translates to up to an 8× speedup in attention computation. For businesses, this implies running large (or custom) models will require far less hardware. Cloud providers and on-premises AI clusters could see costs fall dramatically.

Cloud strategy itself is in flux. Industry discussions stress that “proprietary vs open is not a thing” in AI – companies will mix both (www.pcgamer.com [2]). We already see that in practice. Google Cloud and Azure offer strong first-party models (Gemini, OpenAI) integrated with their platforms, while hyperscalers like Oracle emphasize model-agnostic services. Oracle's CEO explicitly argued that only a scalable, reliable cloud plus customer choice of model will enable AI success (www.itpro.com [3]). Meanwhile, AWS has partnered with third-party labs (like Cohere) rather than building a new model from scratch, reflecting a recognition that open-source or partner models will coexist with big vendors. The takeaway: enterprises should choose infrastructure that lets them switch or combine models as needs evolve, rather than locking in one supplier.

Data and business models around these AIs are also changing. Microsoft announced that, starting April 24, **GitHub Copilot** will use user interactions (inputs, code, outputs) to further train its models (opt-out available) (www.windowscentral.com [4]). This cyclical training means developer data feeds back into better AI coding tools. Executives should be aware of the governance side: IP from company code may end up in model training unless carefully managed. On the revenue side, OpenAI's 4% fee is a small stat with big meaning (www.techradar.com [5]). It shows platform vendors aiming to capture a slice of downstream transactions, not just API calls. Teams should factor these fees (and potential data-sharing terms) into partnership decisions.

Finally, leadership must not forget ROI and ethics amidst the hype. As one analyst points out, boards now care about each AI use case's cost, accuracy and carbon footprint (www.techradar.com [6]). Trust, governance and people remain vital (weaning away from “sprint-into-agentic” without a plan) (www.techradar.com [7]). In practice, that means pilots should be focused on industry-specific problems (where reliability matters most) and include metrics and oversight. For example, an AI agent to schedule meetings must work 99% correctly or it won't save time. Businesses should ensure any new AI tool passes these cost/benefit and compliance filters before wide roll-out.

References:

- [1] www.tomshardware.com — <https://www.tomshardware.com/tech-industry/artificial-intelligence/googles-turboquant-compresses-llm-kv-caches-to-3-bits-with-no-accuracy-loss#:~:text=Google%20Research%20published%20TurboQuant%20on,keys%2C%20while%20reducing%20KV%20cache>
- [2] www.pcgamer.com — <https://www.pcgamer.com/software/ai/nvidia-ceo-jensen-huang-says-open-is-not-a-thing-its-proprietary-and-open-and-sparks-talk-of-ai-orchestras/#:~:text=Nvidia%20CEO%20Jensen%20Huang%20says,Mar%2026%2016%3A25%3A14%202026%20UTC>
- [3] www.itpro.com — <https://www.itpro.com/cloud/public-cloud/oracle-leans-hard-on-one-size-fits-all-appeal-of-oci-for-enterprises#:~:text=Sicilia%20argued%20that%20only%20with,Sicilia%20said>
- [4] www.windowscentral.com — <https://www.windowscentral.com/software-apps/microsofts-github-is-going-to-start-using-copilot-interactions-to-train-ai-models-and-its-starting-soon#:~:text=In%20an%20update%20to%20the,From%20the%20GitHub%20announcement%20post>
- [5] www.techradar.com — <https://www.techradar.com/pro/what-openais-4-percent-checkout-fee-means-for-the-future-of-commerce#:~:text=OpenAI%20recently%20introduced%20a%204,step%20in%20claiming%20that%20ground>
- [6] www.techradar.com — <https://www.techradar.com/pro/from-hype-to-value-the-ai-trends-set-to-shape-2026#:~:text=Over%20the%20past%20two%20years%2C,cost%20and%20carbon%20footprint%20of>
- [7] www.techradar.com — <https://www.techradar.com/pro/from-hype-to-value-the-ai-trends-set-to-shape-2026#:~:text=That%20means%20accepting%20the%20realities,and%20outcomes%20matter%20more%20than>

Key Statistics

- OpenAI charges a 4% checkout fee for ChatGPT purchases (vs Amazon's typical 8–15%) ([www.techradar.com](https://www.techradar.com/pro/what-openais-4-percent-checkout-fee-means-for-the-future-of-commerce#:~:text=OpenAI%20recently%20introduced%20a%204,step%20in%20cl aiming%20that%20ground))
- Google's TurboQuant compresses LLM cache ~6× (to 3 bits) with no accuracy loss, yielding up to 8× faster inference ([www.tomshardware.com](https://www.tomshardware.com/tech-industry/artificial-intelligence/googles-turboquant-compresses-llm-kv-caches-to-3-bits-with-no-accuracy-loss#:~:text=Google%20Research%20published%20TurboQuant%20on,keys%2C%20while%20reducin g%20KV%20cache))
- Oracle unveiled 22 new AI "Fusion Agentic Applications" for tasks like sales and supply-chain optimization ([www.itpro.com](https://www.itpro.com/cloud/public-cloud/oracle-leans-hard-on-one-size-fits-all-appeal-of-oci-for-enterprises#:~:text=At%20Oracle%20AI%20World%20Tour,operations %2C%20and%20automating%20cash%20flow))
- GitHub Copilot will use user inputs/outputs for model training starting April 24, 2026 (opt-out available) ([www.windowscentral.com](https://www.windowscentral.com/software-apps/microsofts-github-is-going-to-start-using-copilot-interactions-to-train-ai-models-and-its-starting-soon#:~:text=In %20an%20update%20to%20the,From%20the%20GitHub%20announcement%20post))

KEY TAKEAWAY

Major AI capabilities are becoming baseline business tools. This week, leaders like Google and OpenAI added multimodal assistants (voice chatbots, visual shopping) while automating work processes and dropping costs (TurboQuant shows 6× memory cuts for 8× speedup ([www.tomshardware.com](https://www.tomshardware.com/tech-industry/artificial-intelligence/googles-turboquant-compresses-llm-kv-caches-to-3-bits-with-no-accuracy-loss#:~:text=Google%20Research%20published%20TurboQuant%20on,keys%2C%20while%20re ducing%20KV%20cache))). Over the next 6–18 months these features become table-stakes: CEOs should accelerate integration of AI-driven customer experiences (voice support, shopping chatbots) and operations (AI agents), re-examine vendor deals (model fees, data ownership) and invest in talent and governance to turn these advances into value.

Sources

[Google just launched a new Gemini-powered dark web monitoring service](https://www.itpro.com/security/google-just-launched-a-new-gemini-powered-dark-web-monitoring-service)

<https://www.itpro.com/security/google-just-launched-a-new-gemini-powered-dark-web-monitoring-service>

[Anthropic gives Claude Code new 'auto mode' which lets it choose its own permissions](https://www.techradar.com/pro/anthropic-gives-claude-code-new-auto-mode-which-lets-it-choose-its-own-permissions)

<https://www.techradar.com/pro/anthropic-gives-claude-code-new-auto-mode-which-lets-it-choose-its-own-permissions>

[Shopping in ChatGPT is about to get a lot more interactive – users promised 'improved speed, relevance and product coverage'](https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage)

<https://www.techradar.com/pro/shopping-in-chatgpt-is-about-to-get-a-lot-more-interactive-users-promised-improved-speed-relevance-and-product-coverage>

[What OpenAI's 4% checkout fee means for the future of commerce](https://www.techradar.com/pro/what-openais-4-percent-checkout-fee-means-for-the-future-of-commerce)

<https://www.techradar.com/pro/what-openais-4-percent-checkout-fee-means-for-the-future-of-commerce>

[Microsoft is rolling out Copilot Cowork to more customers](https://www.itpro.com/technology/artificial-intelligence/microsoft-is-rolling-out-copilot-cowork-to-more-customers)

<https://www.itpro.com/technology/artificial-intelligence/microsoft-is-rolling-out-copilot-cowork-to-more-customers>

[Anthropic's Claude Cowork tool is coming to Microsoft Copilot](https://www.itpro.com/technology/artificial-intelligence/anthropics-claude-cowork-tool-is-coming-to-microsoft-copilot)

<https://www.itpro.com/technology/artificial-intelligence/anthropics-claude-cowork-tool-is-coming-to-microsoft-copilot>

[Salesforce ramps up agentic AI research with new foundry project](#)

<https://www.itpro.com/technology/artificial-intelligence/salesforce-ramps-up-agentic-ai-research-with-new-foundry-project>

"A lot of other cloud vendors have been let off the hook": Oracle leans hard on one-size-fits-all appeal of OCI for enterprises

<https://www.itpro.com/cloud/public-cloud/oracle-leans-hard-on-one-size-fits-all-appeal-of-oci-for-enterprises>

From hype to value: The AI trends set to shape 2026

<https://www.techradar.com/pro/from-hype-to-value-the-ai-trends-set-to-shape-2026>

Microsoft's GitHub will use Copilot interactions to train AI models

<https://www.windowcentral.com/software-apps/microsofts-github-is-going-to-start-using-copilot-interactions-to-train-ai-models-and-its-starting-soon>

Google's TurboQuant reduces LLM KV cache to 3 bits with no accuracy loss (8× speedup)

<https://www.tomshardware.com/tech-industry/artificial-intelligence/googles-turboquant-compresses-llm-kv-caches-to-3-bits-with-no-accuracy-loss>

Gemini 3.1 Flash Live is a massive boon to the AI's real-time assistance for you and me

<https://www.androidcentral.com/apps-software/ai/gemini-3-1-flash-live-is-a-massive-boon-to-the-ais-real-time-assistance-for-you-and-me>

