

Apple Rebuilds Siri on Google's 1.2-Trillion-Parameter Gemini Model—$1 Billion Annual Deal Announced June 8 at WWDC 2026

Apple just paid Google $1 billion per year to make Siri intelligent. The company that built its brand on privacy now routes your voice queries through its biggest rival's servers.

The Announcement: What Apple Actually Said

On June 8, 2026, at WWDC in Cupertino, Apple revealed that the new Siri runs on a custom 1.2-trillion-parameter Google Gemini model. The deal, confirmed in a joint statement from both companies, commits Apple to approximately $1 billion annually under a multi-year licensing agreement.

This isn't a minor feature update. Apple replaced Siri's entire reasoning layer with Google's infrastructure. The new system ships with iOS 27 and transforms Siri from a glorified web search into what Apple calls an "always-on copilot" capable of multi-step task execution

across emails, notes, on-screen content, and web sources.

The technical specs matter here: 1.2 trillion parameters puts this custom Gemini variant among the largest models deployed at consumer scale. For context, GPT-4 was estimated at roughly 1.7 trillion parameters across its mixture-of-experts architecture. Apple negotiated a purpose-built model, not an off-the-shelf API.

According to Tech Insider's WWDC coverage, the most significant architectural change is the new Extensions system. Users can explicitly choose between Google Gemini, OpenAI's ChatGPT, Anthropic's Claude, or xAI's Grok through iOS 27's "Search or Ask" panel. Apple also replaced automatic routing to Google or ChatGPT with an Apple-built AI web search integrated directly into Siri.

> The old Siri returned links. The new Siri takes actions. That's not iteration—that's a category change.

Why Apple Made This Deal

Apple's AI strategy has been in crisis for at least three years. While OpenAI shipped ChatGPT and Google integrated Gemini across its products, Siri remained embarrassingly limited. The gap wasn't closing—it was widening.

Building a competitive foundation model from scratch would have taken Apple 3-5 years and billions in compute costs with no guarantee of success. The talent war for ML researchers has been brutal, and Apple's reputation for secrecy made recruiting harder. Google, OpenAI, and Anthropic had multi-year head starts in training infrastructure, data pipelines, and model optimization.

The $1 billion annual fee sounds massive, but Apple's services revenue exceeded $85 billion in fiscal 2025. This deal represents roughly 1.2% of that segment's revenue to solve their most visible AI deficiency. The math made sense.

More strategically, the Extensions architecture reveals Apple's actual play: they're positioning iOS as the AI orchestration layer, not the AI provider. By offering users choice

between Gemini, ChatGPT, Claude, and Grok, Apple becomes the platform through which AI providers reach 1.5 billion active devices.

The Privacy Calculus

Apple's privacy positioning made this deal politically complicated internally. RedShark News reported that the announcement emphasized Apple-controlled preprocessing on-device before queries reach Gemini's servers. Sensitive information like health data, financial details, and intimate conversations supposedly gets filtered locally.

The architecture works like this: Siri performs initial intent classification on the Neural Engine, determines whether the query requires cloud reasoning, strips identifiable metadata, then routes to Gemini with differential privacy noise added. Results return through Apple's Private Relay infrastructure.

Does this actually protect user privacy? Partially. The preprocessing catches obvious cases—asking about symptoms, financial transactions, personal relationships. But Gemini still processes the actual reasoning request. If you ask Siri to summarize your emails about a business deal, that context reaches Google's servers in some form.

Apple traded absolute privacy for competitive capability. They're betting users care more about Siri finally working than about theoretical data exposure.

Technical Architecture: How the New Siri Works

The 1.2-trillion-parameter model isn't running on your iPhone. Apple uses a tiered architecture that balances latency, capability, and cost.

On-Device Layer

iOS 27 includes an updated 3-billion-parameter on-device model running on the Neural Engine. This handles:

- Intent classification (determining what type of request you're making)
- Simple completions (setting timers, playing music, toggling settings)
- Privacy filtering (identifying and redacting sensitive content)

- Context extraction (pulling relevant details from on-screen content)

Apple claims 40% of Siri queries complete entirely on-device without any cloud communication. This handles the command-and-control use cases where the old Siri was actually competent.

Cloud Reasoning Layer

Complex queries route to Apple's infrastructure, which then calls the custom Gemini model via a dedicated API. The interesting technical detail here: Apple negotiated for dedicated inference capacity, not shared compute. This means predictable latency and no queuing during peak usage.

The custom Gemini variant was fine-tuned on Apple-specific tasks: calendar management, Apple ecosystem integration, multi-app workflows. It understands concepts like Handoff, Universal Clipboard, and Focus modes natively.

Response times for cloud queries average 1.2 seconds for simple reasoning tasks and 3-4 seconds for multi-step agentic workflows. Not instant, but fast enough that users perceive Siri as responsive rather than waiting.

The Extensions Framework

The Extensions architecture is where Apple's platform strategy becomes clear. Technical breakdowns of the WWDC announcement show that third-party AI providers integrate through a standardized API that Apple controls.

Each Extension provider (OpenAI, Anthropic, xAI) must:

- Accept Apple's privacy preprocessing requirements
- Return responses in Apple's structured format
- Support Apple's citation and attribution standards
- Pass Apple's content policy review

Users can set a default AI provider or choose per-query. The "Search or Ask" panel presents all four options with one tap to switch. Apple takes a revenue share (reportedly

15-30%) on any subscriptions initiated through Extensions.

> Apple isn't building AI. Apple is building the App Store for AI—and taking the same cut.

What Most Coverage Gets Wrong

The tech press has focused almost exclusively on the privacy angle. "Apple surrenders to Google" makes a compelling headline, but it misses the strategic reality.

This Isn't Surrender—It's Platform Arbitrage

Apple has never built core technology when they could tax others for access to their users. They didn't build the music labels that power Apple Music. They didn't build the apps that make the iPhone valuable. They didn't build the payment networks behind Apple Pay.

Apple builds platforms that intermediate between users and providers, taking margin in the middle. The Gemini deal extends this model to AI.

Google gets $1 billion annually and prominent placement as Siri's default. In exchange, Apple gets:

- A competitive AI assistant without R&D risk
- Platform control over how AI reaches iOS users
- Revenue share from competing AI providers
- The ability to switch providers if better options emerge

That last point matters. The Extensions architecture means Apple isn't locked into Google. If Anthropic ships a model that outperforms Gemini, Apple can change defaults with a software update. Google's $1 billion buys prominent placement, not permanent exclusivity.

The Gemini Model Is Custom—Not Generic

Most coverage treats this as Apple licensing standard Gemini Ultra. The 1.2-trillion-

parameter specification suggests a purpose-built variant. For comparison, Gemini Ultra launched at around 1 trillion parameters with mixture-of-experts scaling.

Apple likely co-developed this model with specific training objectives: Apple ecosystem knowledge, iOS application integration, privacy-aware reasoning patterns. This isn't Gemini-in-a-trenchcoat; it's a joint development effort that took 18+ months based on the January 2026 joint statement timeline.

The custom nature explains why Apple didn't simply adopt the ChatGPT partnership route Microsoft chose. Apple wanted architectural control, not just API access. The $1 billion annual fee buys engineering collaboration, not just inference compute.

Privacy Isn't Dead—It's Differentiated

The hot take that Apple abandoned privacy fundamentally misunderstands their market position. Apple's privacy stance was always relative to competitors, not absolute.

Google's business model requires data collection for advertising. Apple's business model requires hardware margins. When Apple says "privacy," they mean "we won't monetize your data the way Google does." That distinction still holds.

The new Siri routes queries through Google's infrastructure, but Apple controls the data pipeline:

- Apple preprocesses on-device before transmission
- Apple adds differential privacy noise to reduce identifiability
- Apple routes through Private Relay to mask IP addresses
- Apple contractually prohibits Google from using Siri data for advertising

Is this as private as purely on-device processing? No. Is it materially different from using Google Search directly? Yes.

Apple's privacy position was always pragmatic, not purist. This deal makes that pragmatism explicit.

Practical Implications for Engineering Teams

If you're building products in the Apple ecosystem, the new Siri creates both opportunities and risks.

SiriKit Is Dead—Extensions Are the Future

The old SiriKit framework, which let apps expose limited voice intents, is effectively deprecated. The new Extensions architecture offers far deeper integration but requires significant implementation work.

Apps that previously used SiriKit for basic commands (ordering food, sending payments, controlling smart home devices) should begin migrating to the Extensions API immediately. Apple has announced a 24-month deprecation timeline for SiriKit, with iOS 29 removing legacy support entirely.

The Extensions API differs fundamentally:

- Apps describe capabilities in natural language rather than predefined intents
- Siri can compose multi-step workflows across multiple apps
- Context from other apps flows into your app's responses
- You implement a reasoning callback, not a command handler

This is more powerful but also more complex. Budget 2-3 engineering months for a significant Extensions integration.

Agentic Workflows Create New Product Categories

The "always-on copilot" framing isn't marketing fluff. The new Siri can execute multi-step tasks that previously required human attention at each stage.

Example workflow that now works: "When I get an email from any @acme.com address about the project timeline, add the mentioned dates to my calendar, create a note summarizing the key points, and draft a reply confirming I received it."

This creates opportunities for apps that:

- Expose rich context to Siri (more data enables smarter automation)
- Support fine-grained permissions (users trust granular control)
- Implement reliable action primitives (Siri handles orchestration, you handle execution)

Apps that fully embrace Extensions can become default components in user-created workflows. Apps that don't become invisible to Siri-first users.

Multi-Provider Testing Is Now Required

Your app needs to work with Gemini, ChatGPT, Claude, and Grok. Each AI provider has different capabilities, reasoning patterns, and failure modes.

Gemini handles structured data well but sometimes hallucinates on ambiguous requests. ChatGPT excels at natural conversation but struggles with precise numerical operations. Claude is more cautious and may decline requests other models accept. Grok takes a more aggressive interpretation of user intent.

Build test suites that exercise your Extensions integration against all four providers. The same user request may produce different action sequences depending on their chosen AI. Your app needs to handle all variants gracefully.

Privacy Architecture Affects Feature Scope

Apple's preprocessing layer strips data it considers sensitive before routing to AI providers. This affects what information Siri can use when interacting with your app.

If your app handles health data, financial transactions, or intimate communications, some context may be redacted before Siri reasons about it. This can cause unexpected behavior: Siri may take actions that seem wrong because it's working with filtered information.

Apple's developer documentation includes a privacy classification guide that explains what data categories get preprocessed. Review this carefully—it affects what your Extensions integration can accomplish.

Who Wins and Who Loses

Winners

Google gets $1 billion annually plus default placement on 1.5 billion devices. Gemini becomes the reasoning layer for the world's most valuable user base. Even if users switch to competitors, Google establishes Gemini as the baseline.

OpenAI, Anthropic, and xAI get distribution they couldn't buy. Being one tap away in iOS 27's Extensions panel is worth more than any marketing campaign. The revenue share hurts margins, but volume compensates.

Enterprise app developers can now build agentic workflows that cross application boundaries. The integration complexity shifts from developers to Apple's orchestration layer. If you build good action primitives, Siri handles the composition.

Users who gave up on Siri finally get an assistant that works. The capability gap between Siri and competitors closes meaningfully for the first time since ChatGPT launched.

Losers

Apple's ML research teams just watched the company buy what they couldn't build. Internal morale will suffer. Expect departures to competitors who value and fund foundational AI research.

Voice assistant startups competing on iOS face an existential challenge. Why download a separate app when Siri does the same thing with deeper OS integration?

Privacy maximalists who chose Apple specifically to avoid Google now face an uncomfortable choice. The platform they trusted now routes data through the company they avoided.

Samsung and Android OEMs lose their differentiation window. Google Gemini powered their assistants first, but Apple just made that advantage table stakes.

The Next Twelve Months

Q3-Q4 2026: iOS 27 Launch and Extension Scramble

iOS 27 ships in September. The first three months will be chaotic as developers race to implement Extensions before competitors. Early movers will capture user defaults that prove sticky.

Expect aggressive pricing from AI providers trying to become users' preferred Extension. OpenAI has reportedly prepared a 50% discount on ChatGPT Pro for iOS users who set it as default. Anthropic and xAI are likely planning similar campaigns.

The quality gap between providers will become obvious quickly. Users will discover which AI works best for their specific tasks and lock in preferences.

Q1-Q2 2027: Agentic Features Mature

Apple's initial agentic capabilities are conservative—executing predefined workflow patterns with human confirmation. By mid-2027, expect more autonomous operation.

The roadmap reportedly includes:

- Background task execution without explicit user triggers
- Cross-device workflow continuation
- Proactive suggestions based on learned patterns
- Third-party automation marketplace (think Shortcuts App Store)

Enterprise adoption accelerates as IT departments realize Siri can now handle legitimate business workflows, not just consumer novelties.

Late 2027: The Contract Renegotiation

The Gemini deal is multi-year but not eternal. Apple will have 18 months of data on which AI provider users actually prefer when given choice. That information becomes leverage.

If Gemini remains dominant, Google keeps favorable terms. If users migrate to ChatGPT or

Claude, Apple renegotiates aggressively. The Extensions architecture gives Apple optionality that makes them a formidable counterparty.

Google's best strategy is making Gemini so good that users never bother trying alternatives. Expect major Gemini updates timed to iOS release cycles.

The Strategic Lesson

Apple's Gemini deal illustrates a principle that applies far beyond consumer AI: **platform power beats technology power.**

Google built better AI. Apple controls the platform where users experience AI. Apple's platform position let them acquire Google's technology lead without building it themselves.

This pattern repeats across tech history. Microsoft controlled the PC platform and commoditized application software. Google controlled search and commoditized content publishers. Apple controlled mobile and commoditized app developers.

Now Apple is applying the same playbook to AI. They don't need the best model—they need the best model integration. And integration is what Apple does better than anyone.

For CTOs and founders watching this unfold, the lesson is clear: if you're building AI technology, think hard about who controls the platform where your users live. Technology advantages evaporate when platform owners decide to intermediate.

The companies that thrive will either own their own platforms or build technology so differentiated that platforms can't replicate it. The middle ground—good technology on someone else's platform—is where margin compression happens.

Apple just paid $1 billion to prove that AI capability is increasingly a commodity. The sustainable advantage lies in controlling where and how that capability reaches users.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting