

Gemini 4 Argon: Built to solve your hardest problems

Datum 13.10.2026
Subjekt Google Italia
Autor Koray Kavukcuoglu
Region Itálie
Kategorie IT a komunikace, Technologie, Bezpečnost

We're giving developers, enterprises, scientists, and consumers access to our latest frontier model.

Gemini 4 Argon is now globally available today for developers and enterprises via the API in AI Studio and Gemini platform, for Google AI Ultra subscribers in Google Antigravity, and Google AI Ultra and Pro subscribers in the Gemini app and AI Mode in Search.

Argon excels at deep work in complex, real-world tasks across domains including coding, finance, legal, cybersecurity, and science. Its depth and precision across multi-step tasks leads Argon to outperform other frontier models on benchmarks across domains, ranking #1 on the Vals Index, which measures model performance across critical economic sectors like finance, coding, legal, and tax work.

The model launches today at an introductory price ¹ of \$2 per million input tokens and \$10 per million output tokens (with cached input tokens at 95% off).

Teams at Google, across enterprises, and in the research community have already been using Argon to land significant impact across fields:

- To solve major engineering problems like large codebase refactoring and optimizations
- To optimize business workflows across domains like legal, finance, and science
- To find and patch critical cybersecurity vulnerabilities
- To enable novel breakthroughs in quantum computing, nuclear fusion, and life sciences

Solving complex engineering tasks

At Google, we've been using Argon on our hardest engineering problems. Over the past few weeks, software engineers have been shipping high-quality code faster than ever. Argon in Antigravity has helped teams increase agentic code submissions by more than 35% while reducing code rollbacks by nearly a third.

- Untangling browser feature code in Chrome: Argon helped the Chrome team separate core webpage handling from a ~6 million line legacy folder of browser features held within Chromium's 40 million-line codebase. Piping multiple clean code changes at a time, Argon helped the team finish an architectural refactor across hundreds of files, completing months of work in weeks. For Chrome's billions of global users, this will unlock the ability for live webpages to move fluidly between tabs, side panels, and floating windows without reloading, eliminating an entire category of browser crashes.
- Optimizing a production Python codebase in hours: Argon in Antigravity is helping the team

behind Google's open-source Agent Development Kit (ADK), a framework powering enterprise agents, push their performance engineering further. The team gave Argon their .agents/skills rulebook and a one-paragraph goal to eliminate framework overhead without breaking existing apps. Working autonomously overnight, Argon surfaced optimizations, now live in `google/adk-python`, that cut in-memory session prep time by 40% in long sessions and sped up cold imports by 21%. Read the engineering deep dive to learn more.

- Quantum chemistry open-source package optimization: xTB (Semiempirical Extended Tight-Binding) is one of the most-cited open-source quantum chemistry packages for molecules. Scientists use it to predict molecules' 3D shapes and to screen large numbers of drug and materials candidates. Using Argon in Antigravity, our researchers made xTB up to 2.3x faster at predicting molecular structures and 1.8x faster at conformer search. The changes have been contributed to the official xTB repository on GitHub.
- Making our open Gemma models nearly three times as fast: Using Argon, we achieved a 2.85x end-to-end inference speedup on an A100 via vLLM (accelerating throughput from ~37,900 to over 108,000 tokens per second on Gemma 4 E2B) with 100% token parity and zero quality drop in just a few days, a process that would usually take months. Check out the upstream pull requests (optimizing everything from attention tuning to speculative decoding) to see the exact changes.
- Delivering complex 3D visualizations: Argon brings major improvements in spatial reasoning, mathematical precision, and visual coding compared to previous Gemini models. Using slash commands like `/goal` and prompting for multi-agent loops with custom skills in Antigravity, Googlers are using Argon to prototype complex 3D visualizations, high-fidelity games, and interactive animations in vanilla code that can be rendered directly in the browser.

Argon built a highly detailed model of a Pagoda using multi-agent visual critique loops in Antigravity. The model autonomously sculpted a multi-building temple complex with photorealistic architecture, foliage and lighting in Blender.

Argon turned pictures of a real desk into a playable miniature racing game using a multi-agent loop in Antigravity. The desk objects, room and race track were modeled and collision-checked using Blender MCP, textured with Nano Banana, and wired with physics, custom shaders, and car engine audio from Gemini Omni Flash in Three.js. Try it now in Google AI Studio.

Argon in Antigravity wrote Python scripts to procedurally sculpt, rig, animate, and PBR-bake a complex 3D game using multi-agent visual critique loops then wired it for the web in Three.js with GLSL shaders, shatter physics, and a single-player AI. Try it now in Google AI Studio.

With zero external libraries and 100% pure Python, Argon wrote its own 3D ray-tracing engine with a custom video tool and audio mixer to build and document a sunlit library from scratch, producing, narrating, and assembling this video with Gemini 3.8 TTS voiceover and music by Lyria 3.5. View the source code.

Starting from a single Nano Banana image in Antigravity, Argon drives Blender MCP in an autonomous loop, aligning camera grids, sculpting Gothic architecture, and staging audio-reactive lighting, to create a finished 3D scene.

Transforming critical business and professional workflows

Argon combines frontier reasoning, deep visual perception, expanded domain knowledge, and stronger writing capabilities across legal, finance, science, and coding to help with high-stakes professional workflows. It communicates clearly, keeping you in the loop, and follows style instructions better than any of our previous Gemini models. Argon answers questions more directly

with less hedging, doesn't repeat itself more than necessary, and skips clichés throughout the responses.

Enterprise customers and startups in our early access program are already seeing strong wins with Argon.

Using its advanced reasoning and multimodal capabilities while maintaining strict grounding, Argon excels in deep multimodal document comprehension, longitudinal data analysis, and interactive analytical sandboxes.

See how Argon synthesizes ~50 GB of Airborne LiDAR scans, GIS room vectors, Library of Congress CAD drawings, and NPS archives to reconstruct 3D models of Chaco Canyon, turning fragmented archaeological records into explorable digital heritage sites. Try it now in Google AI Studio.

In Gemini App, Argon takes an unstructured folder full of project context and formats it into a clean proposal deck for a client using a custom skill. Gemini then drafts and emails the deck to a client while creating an automated workflow for all future incoming client requests with a scheduled task.

Watch Argon synthesize data from 3 open-source repositories into a medical learning tool where you can peel back anatomical layers, inspect clinical metadata, and trigger biomechanics in a 3D exploration of the human body.

Argon using Antigravity in the Gemini macOS app generates design iterations for new legs for a sofa using Nano Banana then generates a 3d printable file for one. Starting from a sketch of the measurements and a photo of the couch Argon builds the model to the correct size to fit the sofa.

At Isomorphic Labs, the team is using Argon as a partner to accelerate their drug discovery and design research workflows. Argon is able to quickly synthesize complex PDFs and patents to extract pertinent information, reducing a process that typically takes weeks into hours.

Accelerating real-world impact with researchers and domain experts

Researchers, engineers, and security specialists are already using Argon to push the boundaries of cyber defense, science, math, and robotics. It's helping cyber defenders and researchers test hypotheses, autonomously audit critical codebases and formal proofs, and catch subtle zero-day vulnerabilities and logic flaws.

Bringing state-of-the-art cyber capabilities to more trusted defenders

Starting today, we're beginning expansion to hundreds of additional Fairwind Program partners.

Wiz is already using Argon for its "Scan for Good" initiative, which protects critical infrastructure for the public good and for free. With early access to the model, Wiz has demonstrated Argon's real-world impact across multiple high-stakes sectors:

- **Securing sensitive government data:** As part of a long-standing cyber resilience program and in collaboration with a federal agency, Argon helped to proactively identify and remediate a gap that risked exposing sensitive data and government infrastructure. This effort helped prevent against the exposure of critical security sector development details and unauthorized access to specialized models.
- **Protecting critical healthcare infrastructure:** Wiz also used Argon to protect vital healthcare

infrastructure. By chaining complex vulnerabilities, the model helped the team discover and quickly fix a public file download bypass that exposed a master key and granted administrator access. Argon went beyond spotting a vulnerability to proactively safeguarding hundreds of thousands of private patient records.

Unlocking new scientific breakthroughs

Argon is already having real-world impact across the scientific community — enabling novel discovery for scientists and experts across multiple domains.

- **Quantum computing:** Argon is helping our researchers accelerate progress across both quantum error correction and algorithm design. To help error decoders keep pace with quantum hardware in real-time, Argon translated an internal version of AlphaQubit, our ML quantum error correction decoder, into a single, highly optimized hardware kernel — nearly doubling its inference throughput so we can deploy larger, more capable models and scale to larger quantum devices, while continuing to meet strict real-time thresholds. Argon is also reducing spacetime costs for quantum subroutines. For *arcsin*, a building block in recent quantum simulation algorithms, an implementation written by Argon and formally verified in Lean cuts spacetime cost by more than 100x compared to a published quantum CORDIC baseline, a step toward a standard library of optimized quantum circuits.
- **Nuclear fusion:** Stellarators are incredibly complex machines that represent a leading candidate for limitless clean fusion energy, but optimizing their 3D magnetic fields is famously difficult. Since 1967, a long-standing hypothesis known as “Grad’s conjecture” assumed that finding exact mathematical solutions for these fields was impossible, until recent counterexamples were found, including by Professor Matt Landreman (University of Maryland). Now, a team at Google alongside Professor Landreman has discovered a broader class of counterexamples to Grad’s conjecture. These new solutions give the fusion community reliable benchmarks to test their stellarator simulations codes, which are valuable tools for future designs.
- **Rare disease research:** Researchers at Gamow Labs used Argon to analyze de-identified genomes from newborns with rare developmental lung disorders. In one case, Argon pinpointed a chromosomal rearrangement that routine software missed for years, and explained why it went undetected. On private Gamow Labs evals derived from real clinical and lab data, Argon topped on RareBench (57% on assistance with rare-disease diagnosis) and LabBench (52.2% on wet-lab reasoning). These advances point to faster diagnosis for families facing rare conditions. See [blog²](#) for more details.
- **Enzyme design:** Engineering enzymes with precise reactivity is notoriously slow, often demanding months of bespoke model tuning and candidate screening. Serine esterases are a class of enzymes that cut ester bonds on targeted molecules, widely used in recycling, and the manufacture of pharmaceuticals, biofuels and detergents. Our scientists worked with Argon to design serine esterases, targeting the benchmark molecule 4MU-acetate that glows under UV light once its ester bonds are broken. In a matter of hours, Argon assembled relevant data, performed model tuning and identified multiple biochemistry-inspired filters to get to a final set of five novel designs. Subsequent laboratory testing confirmed robust catalytic activity across these designs. The speed and ease of successful enzyme generation demonstrates the potential for rapid acceleration in enzyme engineering in fields such as biomanufacturing, green chemistry, and drug development.

To further advance the pace of research, we’re offering 10,000 scientists the highest usage tier in Google Antigravity free of charge for 6 months (onboarded on a rolling basis), subject to eligibility, capacity, and selection. Applications are now open.

Speeding up robotics development

By bringing together advanced visual-semantic reasoning, deep physical understanding, and autonomous code generation, Argon unlocks new workflows across the robotics stack and accelerates progress toward physical AGI. This leap in capability is reflected in Argon's performance on the Embodied Reasoning (ERQA) benchmark, which tests a model's ability to understand and solve physical, spatial problems in 3D environments. Relying on pure reasoning, Argon outperforms all other frontier models, with an 87% score.

To demonstrate these capabilities in action, we tested Argon's ability to learn an out-of-distribution physical skill simply by watching. In an internal test, Argon watched a raw video of a human playing cornhole, mapped the spatial coordinates of the environment, and calculated the required whole-body joint movements. It then autonomously built the simulation code. When transferred onto Aptronik's Apollo humanoid, the robot quickly sank seven consecutive shots in a real-world trial.

This "Video-to-Sim-to-Real" pipeline compresses weeks of manual simulation modeling into a few hours, transforming ordinary videos into a scalable, infinite training ground for robotics. Learn more about how this accelerates the pace of development in robotics in our developers blog.

Developing Argon responsibly and safely

Safety and security are foundational to model development at Google. Because Argon is a highly capable model, we collaborated with a variety of third-party partners on measurement of certain model capabilities and safeguards. Gemini 4 Argon represents our most comprehensive safety and security stack yet:

- **Misuse & jailbreak defenses:** The same coding and scientific capabilities that help defenders patch hospital software or accelerate bioengineering also require strong safeguards against malicious actors. Under our Frontier Safety Framework, we evaluated Argon's enhanced capabilities across several domains, and deployed a multi-layer defense stack combining jailbreak resistance, monitoring, and enforcement. This greatly improved our robustness — for instance, external bio red teaming could only unlock less than 1% of bio sub-tasks we test for. Further, when it comes to dangerous requests — from writing exploit code to helping with dangerous biological materials — Argon is significantly better at not fulfilling harmful requests than our previous models. On an adversarial set meant to stress test the model, violations reduced nearly 2-fold — also significantly better than previous Gemini models.
- **Agent alignment & real-time monitoring:** As agents become more capable problem solvers, we have to ensure they don't get overly eager to finish a task by taking unauthorized shortcuts when they hit a roadblock, like a permission error or a failing test.
 - First, we trained Argon to respect system boundaries, report negative findings honestly, and pause for human confirmation before taking irreversible steps. On coding benchmarks where models have no incentive to hold back, shortcut attempts dropped to just 2% on Gemini 4 Argon, our best performance compared to previous Gemini models.
 - Second, as a defense-in-depth layer, we built an automated misalignment monitor that reads the agent's step-by-step reasoning and actions to catch and block unintended workarounds in real time. We used this monitor across internal deployments, capability evaluations, and reinforcement learning training runs — while strictly avoiding penalizing the model's thoughts during training so its reasoning remains transparent and monitorable.
- **Prompt injection resistance:** When agents browse webpages, read emails, or inspect code, they can encounter hidden malicious instructions designed to hijack their task. Through adversarial

training and automated red teaming, Argon leads the Gray Swan Indirect Prompt Injection (IPI) benchmark, cutting attack success across 15 adaptive attempts to 0.7%.

For more information, see the Gemini 4 Argon Model Card.

Get started today

Argon works longer and harder to solve highly complex scientific, coding, and agentic problems. This is enabled by an industry-leading 1M unified input-output token limit via the Gemini API. Read through our documentation for developers and enterprises to learn about the new continuation protocol and ensure the appropriate setup for your needs. Argon is rolling out today globally:

- For developers: Build with the Gemini API paid tier and if you're a Google AI Ultra subscriber, you can try it in Google Antigravity, AI Studio, and Android Studio.
- For enterprises: Try it in Gemini for business, the API, and in Google Antigravity.
- For everyday users: Try it starting today in the Gemini app and AI Mode in Search if you're a Google AI Ultra or Pro subscriber.
- For scientists & researchers: Apply for a chance to get access to Google Antigravity's highest usage tier free of charge for 6 months (onboarded on a rolling basis), subject to eligibility, capacity, and selection. Applications are now open.
- For cybersecurity defenders: Apply to join the Fairwind Program to get access to Argon without cyber guardrails on a rolling basis.

Gemini 4 Argon is just the beginning. We're looking forward to bringing you the rest of the Gemini 4 family.

<https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-4-argon>