Episode Book

EPISODE #6309

Why DeepMind Systems Live Outside the Gemini API

PUBLISHED	27 October 2026 at 18:45 UTC
DURATION	25:51
LISTEN	myweirdprompts.com/episode/deepmind-models-outside-gemini-api-03211/
HOSTS	Corn, Herman, Dolores, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
BOOK GENERATED	28 October 2026 at 09:53 UTC

SYNOPSIS

Google DeepMind regularly announces breakthrough systems that never appear as simple endpoints in Google AI Studio or the Gemini API. Exploring systems like AlphaEvolve, Co-Scientist, Genie, and AlphaFold reveals why they are kept separate. The restrictions stem from serving costs, workflow requirements, and commercial strategy rather than mere secrecy, requiring specialized cloud setups, research applications, or dedicated spin-outs instead of a standard chat prompt box.

TRANSCRIPT



Corn

Seven tenths of one percent of Google's worldwide compute, recovered by an algorithm most developers cannot simply call. Daniel's prompt is about that peculiar experience of reading a DeepMind announcement, opening AI Studio, and discovering that the exciting thing has no endpoint waiting for you. He wants to know which systems live outside the public Gemini offering, how people actually reach them, and why Google keeps those doors separate.



Herman

That compute saving comes from AlphaEvolve, and it makes the missing endpoint rather more interesting. Google has something that improves the machinery running Google. Selling everyone access is only one possible use of it.



Corn

The pricing page is competing with the electricity bill.



Herman

There are four drivers to separate. Serving cost, safety staging, the shape of the system itself, and commercial strategy. They overlap, but they imply different futures. Expensive inference might become cheaper. An experimental interface might mature. A competitive advantage might remain private because it is still a competitive advantage.



Corn

I also want to distinguish unavailable from unavailable to me. A subscription button, an academic application, and a commercial negotiation are very different obstacles. Only one accepts my card without asking what I intend to accomplish.



Herman

Google doesn't publish a canonical register of restricted DeepMind systems. You piece this together from launch announcements and access programmes. The list is longer than people expect. Genie three, AlphaFold three, the WeatherNext forecasting models, AlphaGenome, the maths systems AlphaProof and AlphaGeometry. Then the three I want to start with, because they sit closest to Gemini. AlphaEvolve, Co-Scientist, also presented



Herman (continued)

through Hypothesis Generation, and the Deep Think reasoning modes.



Corn

Calling all three models already conceals something. AlphaEvolve uses Gemini models inside it. If I can access the ingredients, I haven't necessarily been given the machine.



Herman

Think of what you're missing. AlphaEvolve selects candidate programs, changes them, runs evaluators, keeps promising variants, and repeats. Gemini Flash supplies breadth, while Gemini Pro supplies deeper proposals. The operational system includes the search process and the tests, not merely the language model's weights.



Corn

You are about to tell me how elegant that is, Herman, before mentioning who writes the tests.



Herman

The person with the problem writes or supplies the evaluator. You caught me. That requirement is the hard boundary around the excitement.



Corn

I have lived through your installations. The elegant part generally arrives before the permissions dialogue.



Herman

Your contribution generally arrives when it stops working. Which, to be fair, is useful product testing. Would you call AlphaEvolve ready if the search worked beautifully but an ordinary customer couldn't specify a reliable score?



Corn

Ready for a specialist, perhaps. Not ready for a promise that says, improve my business. If you can automatically measure the answer, you have already done a substantial piece of the intellectual work.



Herman

Co-Scientist has a different structure. Multiple agents generate hypotheses, critique them, compare them, and conduct an idea tournament. Deep Think is closer to what people expect from a reasoning model, but its most capable configurations involve substantial inference resources. Those distinctions matter before we assign a single motive to every access restriction.



Corn

Otherwise we make a secret laboratory out of three different procurement problems.



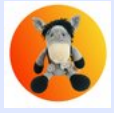
Herman

Some of it really is frontier research. Some of it is a service Google hasn't made self-serve. Some is a product with a narrow customer definition. A DeepMind announcement tells you that something exists. It does not automatically tell you that Google has decided to operate it for arbitrary developers.



Corn

AlphaEvolve's access language is particularly revealing. In May two thousand twenty-five, DeepMind said it was planning early access for selected academic users and exploring broader availability. That leaves a lot of room between a working result and a product I can budget for.



Herman

And that room got filled, but not where Daniel was looking. AlphaEvolve went through a private preview on Google Cloud, and on July tenth this year it reached general availability there, on the Gemini Enterprise Agent Platform. So it is a product now. It is still not a model you pick in AI Studio or call from the Gemini API.



Corn

So the door exists. It's just in the enterprise wing.



Herman

Right. The original Ars Technica coverage described AlphaEvolve as too complex to provide publicly. What changed is not that it became simpler. Google built it a home that fits its shape, inside a cloud account, next to the customer's own evaluation environment.



Corn

I would want the evaluator conversation before the sales conversation anyway. Suppose I ask it to improve a scheduling algorithm. Speed is measurable, but fairness, starvation, and what happens when a machine disappears are constraints. A faster answer that quietly drops difficult jobs is wonderfully efficient until somebody notices.



Herman

You need correctness tests and a score that captures the desired improvement. In practice, multiple constraints. The system can exploit a weak evaluator as readily as it can discover a legitimate shortcut. You don't have to imagine it being malicious. Searching aggressively against an incomplete measurement is enough.



Corn

Finally, software that shares management's interest in the metric.



Herman

Co-Scientist shifts that burden. Its tournament can rank hypotheses and develop proposals, but a proposal winning an argument among agents doesn't make it biologically true. The useful output is a candidate for investigation. Experiments still have to happen.



Corn

That is where I'd resist the agent count as a sales pitch. Ten versions of the same assumption can hold an extremely orderly meeting.



Herman

Correlated errors are a real concern. What would persuade you that the tournament was contributing something, Corn?



Corn

Show me a hypothesis that survives independent checking and gives an experimental team a better next move. I don't need the agents to sound like a department. I need the department to spend less time pursuing rubbish.



Herman

That is also why the access route asks about the research. Hypothesis Generation is one of the Google Labs science tools behind a request form. Researchers from any institution can request access, but requesting is not obtaining. The associated tools include Computational Discovery, built around AlphaEvolve and ERA, and Literature Insights, built around NotebookLM.



Corn

The form is the front door. There may still be a building inspection.



Herman

Deep Think needs a more careful answer than no API. Ultra subscribers can use it in the Gemini app. Selected researchers and companies can obtain API access through an early access programme. The February upgrade was the first API opening, but it was application-gated rather than an ordinary public model selection.



Corn

Meaning the sentence, it has an API, can be true and completely unhelpful to Daniel.



Herman

The programme prioritised energy modelling, new materials, and biomedicine. Those domains give Google demanding problems and partners who can assess the outputs. They also give it a manageable population of users rather than an unknown surge of requests.



Corn

There are other doors that shouldn't be mistaken for secret models. Gemini for Research supplies API credits and higher limits to eligible faculty, staff, and doctoral students. That changes the terms of access. It doesn't, by itself, establish a different underlying model.



Herman

Applications are reviewed monthly, and the credits can be granted or removed at Google's discretion. I would be careful about making a long-running project depend on assuming that subsidy is permanent.



Corn

My least favourite autoscaling event is the funding disappearing.



Herman

The European programme announced this month, through the RAISE initiative, is the science version of that. Up to two hundred researchers, three million dollars in computing credits, and named systems. AlphaFold three, WeatherNext, AlphaGenome, AlphaEvolve and AlphaEarth Foundations. How selection works and when access starts hasn't been spelled out yet.



Corn

That list is interesting because none of those are chat models. Subsidised access to a public model and privileged access to a scientific one are different benefits, and this is mostly the second.



Herman

Subscription prototypes are another route. Gemini Spark went first to trusted testers and then into beta for US Ultra subscribers. That is staged access to an agent experience, rather than a promise of a developer endpoint. These programmes can resemble one another from outside while serving quite different purposes.



Corn

Genie three is the one I'd put in that drawer. The world model, where you type a scene and then walk around inside it.



Herman

Project Genie, since January, for Google AI Ultra subscribers in the US, over eighteen, framed as a research prototype. Sessions are reportedly capped at about sixty seconds, and some of the features from the research announcement aren't in it. There's no API at all. And that one is mostly about cost. It generates frames in real time as you move, so every user is holding a slice of accelerator for the whole session.



Corn

A chat answer ends. A world keeps rendering until you leave.



Herman

AlphaFold three is a different door again. There's AlphaFold Server, free, with a daily cap on predictions. There's code and weights for academic use under restrictive terms. And commercial drug discovery goes through Isomorphic Labs, Alphabet's own spin-out. That's commercial strategy in plain sight. The most valuable customer for protein structure prediction is a pharmaceutical company, and Alphabet built a company to be the one that serves it.



Corn

And the weather models?



Herman

WeatherNext, the line that grew out of GraphCast and GenCast, comes as data and as a cloud product. Earth Engine, BigQuery, Vertex AI. Which makes sense. Nobody wants to chat with a forecast. You want the forecast in your data warehouse every six hours.



Corn

So half of Daniel's list aren't hidden at all. They're just not shaped like a prompt box.



Herman

And AlphaProof and AlphaGeometry are the other end of it. Olympiad maths systems, published as research results, and never offered as products in any form.



Corn

Welcome back to My Weird Prompts, where Daniel supplies the prompt and we follow it. Today, the DeepMind systems missing from the public API, and apparently somebody has brought paperwork.

D

Dolores

Good afternoon. This is the World Podcast Authority, compliance division. Your paperwork is not in order. This episode has no transcript filed in the correct triplicate format, form T X seven, reference number four four one two nine.



Herman

There isn't a World Podcast Authority to file anything with.

D

Dolores

I will carry on regardless. Please note that your podcast licence may be revoked if the corrected triplicate transcript is not filed within fourteen days.



Corn

Podcasting doesn't require a licence either.

D

Dolores

I will make a note of that. The relevant form will follow in the post. Good afternoon.



Herman

The technical distinction is that an API can serve an entire workflow, but it needs the right contract. Long jobs need state, progress reporting, cancellation, and somewhere to put results. AlphaEvolve also needs an execution environment for evaluating candidate code. A standard prompt and completion interface leaves most of that unspecified.



Corn

Then saying these things aren't API-shaped is a little too strong. They aren't shaped like the particular API people are looking in.



Herman

Fair correction. Gemini Deep Research demonstrates the distinction. Its developer route uses the Interactions API rather than the standard content-generation endpoint. Missing from one catalogue doesn't mean incapable of programmatic access. It can mean the operation has a different lifecycle.



Corn

A request that finishes after lunch needs more than a spinning button.



Herman

Compute economics is probably the most load-bearing explanation for Deep Think. Additional reasoning means more backend work per useful answer. If you let arbitrary applications issue arbitrary volumes, the bill and the capacity planning become much harder than a tightly controlled launch.



Corn

Tokens don't reveal the whole cost either. A short final answer could conceal a substantial amount of internal search. People see three paragraphs and wonder why they can't have ten thousand of them immediately.



Herman

There is no reliable public multiplier I can give you for that cost. Different tasks and configurations can use different amounts of work. Application review lets Google understand expected demand, batch work where appropriate, and favour tasks where expensive reasoning might actually be valuable.



Corn

Would you pay for it to handle ordinary household questions?



Herman

Not routinely. I want expensive reasoning when an error costs more than the inference. Although I can hear you preparing to bring up my enthusiasm for elaborate tools.



Corn

You would use a materials programme to choose shelving.



Herman

Load-bearing shelving deserves respect. Still, no, not every task needs the deepest mode. Corn, would a public price solve your objection to gated access?



Corn

It would solve uncertainty about purchasing, not uncertainty about continuity. I can build around an expensive service if the terms are stable. A cheap preview that can vanish leaves me maintaining a dependency on somebody's continuing interest in my project.



Herman

Safety staging addresses a separate concern. Deep Think launches included additional safety evaluation, and its consumer rollout was delayed for further testing. Starting with a smaller, monitored population gives the provider a chance to see what users actually do before widening access.



Corn

I accept that explanation without accepting it as the explanation for everything. A research proposal being approved tells us who has access. It doesn't tell us the evaluator is sound or that a hypothesis is safe to act on.



Herman

Selection is one control, not validation of every output. In biomedicine especially, you need to separate persuasive reasoning from evidence that an intervention works. My interest in the tool doesn't remove that distinction.



Herman

Commercial sequencing gives another reason to choose those users. A materials team can supply a valuable problem, domain expertise, and a route to deployment. Google learns which parts of the workflow need support before trying to sell the same service widely.



Corn

Those early users are helping define the product. That makes the relationship more than a subscription, even when money changes hands. Google can want their expertise as much as their custom.



Herman

Compare that with Gemini four Argon's public introductory pricing. Two dollars per million input tokens, ten per million output tokens. There is a meter a developer can plan around, with the introductory rates due to double after November thirtieth.



Corn

Even the price rise has a calendar appointment.



Herman

Deep Think's reported API prices are less settled. I've seen conflicting output rates, and I wouldn't quote either as a confirmed Google tariff. That contrast is informative without pretending we have a trustworthy public price for every gated service.



Corn

The strongest strategic reason seems to sit inside AlphaEvolve. Google gets an algorithm that improves its infrastructure, then uses that infrastructure to run more AI work. I don't think we need to assume that selling this capability outside is its first priority.



Herman

Google's reported Borg result makes that concrete. The scheduling heuristic has been in production for over a year, recovering an average of seven tenths of one percent of worldwide compute. A small percentage at that scale is a continuing operational gain.



Corn

It earns its keep without acquiring a single customer.



Herman

A Gemini matrix-multiplication kernel improved by twenty-three percent, which translated into a one percent reduction in training time. Those are different levels of measurement. A dramatic improvement in one kernel becomes a smaller, still valuable improvement across the whole training run.



Corn

That distinction saves us from multiplying every good benchmark by the entire company. The kernel doesn't perform all the other work just because the announcement is enthusiastic.



Herman

AlphaEvolve also produced a Verilog rewrite for an upcoming TPU, and up to a thirty-two point five percent improvement in a FlashAttention implementation. Software, chip design, scheduling. There are several places for a useful discovery to compound.



Corn

Would releasing it actually surrender the advantage? Competitors don't get Google's workloads, evaluators, and deployment knowledge with the agent.



Herman

They don't. I shouldn't imply that access transfers the whole advantage. It transfers a potentially useful search capability. Competitors would still need suitable problems, good tests, compute, and engineers who can integrate the results. The moat is partly the system and partly the environment in which Google can use it.



Corn

Which is probably why Google was comfortable taking it to general availability in the cloud. You sell the search, you keep the environment. And it sells as an enterprise agent, not as a cheap endpoint anyone can wire into a weekend project.



Herman

The mathematical results show why researchers want access independently of that business case. It found a method for multiplying four by four complex matrices using forty-eight scalar multiplications, improving on a longstanding construction associated with Strassen.



Corn

Fewer multiplications, rather more application forms.



Herman

Across more than fifty open mathematical problems, it matched the best known solutions in roughly seventy-five percent of cases and improved them in twenty percent. Those are results over a chosen problem set, not a universal probability that it will solve your mathematical problem.



Corn

The evaluator constraint selects the territory. You can search productively when a proposed construction can be checked, even if nobody knows the best possible answer. That is a different promise from asking it to pronounce on any unresolved conjecture.



Herman

It found a configuration of five hundred and ninety-three spheres for the kissing-number problem in eleven dimensions. The important achievement is the verified construction. It doesn't require accepting a fluent explanation on trust.



Corn

Hm.



Herman

Google also describes uses in DNA sequencing error correction, disaster prediction, simulated power-grid stabilisation, and customer work in drug discovery and supply chains. The grid claim needs that word simulated. An improved simulation result and a deployed grid controller are not interchangeable.



Corn

Publicity still benefits from the restriction. Deep Think can post an ARC-AGI-two score of eighty-four point six percent while most developers can't run their own comparison. The headline travels much faster than access.



Herman

Its other announced results include forty-eight point four percent on Humanity's Last Exam without tools and a Codeforces rating of three thousand four hundred and fifty-five. Impressive reported results, but they don't tell a team its latency, cost, or success rate on actual work.



Corn

I would rather see the failure cases on my workload than another decimal place on somebody else's. Access matters because it lets the user ask a question the launch material doesn't answer.



Herman

A public service lets you buy that experiment. A selected programme makes obtaining the experiment a relationship, with a proposal, a reviewer, and an institution attached.



Hilbert

That is quite right. The small correction is that the form permits two hundred words, not fewer than two hundred. Corn, you said fewer when we were discussing the application before recording. I am using all two hundred this month. The Municipal Upholstery Society is applying for the think cupboard. The minutes require an institutional affiliation. I put independent researcher. I am independent and I research. Neville objects because the society has a constitution. Neville objects to most things with a submit button.



Corn

An institutional affiliation seems ambitious for a society still disputing its own independence.



Hilbert

The dispute concerns whether an occupied chair counts as storage. It is in the minutes every month. My research description concerns the material properties of a very old chair. I want to know whether it merits reupholstering. I don't own it. I price it at one hundred and eighty pounds and leave it where it is. The society currently appoints it deputy treasurer so it can remain in the meeting room under the furniture rules.



Herman

The think cupboard is being asked a valuation question with a materials description attached.



Hilbert

It approves the reupholstering. The access approval arrives this week. The chair approval arrives eleven minutes later. Herman, your evaluator explanation earlier takes longer than either, but yours includes the bit about checking the answer. Mine says yes. That is the most useful answer I get from a model this month. Neville now wants the think cupboard to approve the fabric. I tell him the access concerns energy, materials, and biomedicine. Fabric is materials. He insists we record that interpretation as a motion.



Corn

The deputy treasurer should abstain.



Hilbert

It does. Neville records one abstention and orders a fabric sample. The sample costs six pounds.



Herman

Six pounds is a modest first experiment. The larger open question is whether falling inference costs eventually make the think cupboard ordinary infrastructure, or whether the doors stay plural. Consumer prototype, cloud product, science programme, research licence, paper only.



Corn

My expectation is that some doors widen and some stay negotiated. The internal optimisation advantage survives a cheaper compute bill. Thanks to Hilbert Flumingtop for producing the show.



Herman

If you want more of this, try episode fourteen, AGI's Crossroads; and episode twenty, Architectural AI. This has been My Weird Prompts, the human-AI collaboration podcast. If you enjoy following these questions with us, leave a review wherever you listen.



Corn

See you tomorrow.



Herman

Until then.