Episode Book

EPISODE #6257

Where Your Prompts Actually Go

PUBLISHED	24 October 2026 at 16:27 UTC
DURATION	24:22
LISTEN	myweirdprompts.com/episode/ai-inference-geographic-routing-19890/
HOSTS	Daniel, Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.22
SOURCES	13 cited by the research (listed at the end)
BOOK GENERATED	25 October 2026 at 06:19 UTC

SYNOPSIS

Tracing where a model execution happens reveals a stark divide between where data is stored and where GPUs compute answers. The public documentation for Anthropic and OpenAI separates data residency from inference geography, with cloud marketplaces like AWS offering different boundaries than direct APIs. Examining endpoint routing policies, premium pricing for geographic restrictions, and partner offerings demonstrates why latency measurements and reseller logos fail to reveal the physical servers processing each prompt.

TRANSCRIPT



Daniel

When I send a prompt from Jerusalem, where does the actual GPU work happen? I can see the service name, but that doesn't tell me where the model weights are sitting. I'd like to separate what these companies actually disclose from what we're guessing.



Corn

Daniel's asking for the physical geography behind the chat box. Not the address on the invoice, and not just where a conversation gets stored. Where does the machine doing the inference live, and can we trace a request to it?



Herman

The closest public answer is usually a routing policy. You get a menu of permitted geographies, sometimes a regional guarantee, but not a list of buildings. Anthropic's default lets inference run in any available geography for performance and availability.



Corn

An unusually large delivery window.



Daniel

Does that mean the geography is changing between requests, or just that they reserve the right to change it?



Herman

The policy permits it. Global endpoints explicitly route toward available regional capacity, but that doesn't establish that your successive requests actually move. That's our first distinction. What the service can do, what it promises, and what happened to your particular prompt are different pieces of information.



Corn

Herman, before you start putting pins in a map, you haven't got addresses.



Herman

I have regions. I accept that my pins need considerable improvement.



Herman

Anthropic uses multiple cloud providers to process customer data. Its infrastructure subprocessor listings name AWS, Google Cloud, and Microsoft Azure. Cloudflare is listed for traffic routing. Those names describe different roles, though. Encountering a Cloudflare network address doesn't mean Cloudflare has the GPUs generating your answer.



Daniel

Are we talking about Anthropic owning machines inside those clouds, or renting the whole service?



Herman

The public material establishes the cloud relationships, not a complete ownership ledger for every machine. A model developer, a cloud operator, and the company selling you access can all be different parties. The same model can also be offered through products with different operational arrangements.



Corn

You can buy the same intelligence through several checkout counters. The counters don't locate the intelligence.



Herman

AWS's Claude Platform user guide makes that surprisingly explicit. It says, quote, Data may not reside in AWS. Inference may route to Anthropic's primary cloud. Subservices may change without notice.



Daniel

That's a fairly substantial qualification if I've deliberately bought it through AWS.



Corn

The shopping basket stays in AWS.



Herman

Your purchasing channel doesn't, by itself, establish the execution location. Also, that warning is about Claude Platform on AWS. We shouldn't casually transfer every property of that product onto Bedrock, which has its own endpoint and regional arrangements.



Herman

Corn, if you saw an AWS purchase on a company's architecture diagram, would you assume the model execution stayed there?



Corn

I'd expect it, yes. I'd be wrong to treat my expectation as a guarantee, but the expectation isn't unreasonable. A diagram with one cloud logo strongly invites you to stop asking where the next arrow goes.



Herman

That's why the product boundary matters more than the logo. A cloud marketplace relationship, a managed model deployment, and a call forwarded to a provider are not interchangeable architectures.



Corn

Your diagram now needs a second page.



Herman

Probably. Although a carefully labelled arrow would save us both a meeting.



Daniel

What does Anthropic actually let me choose on its own API? Forget the reseller for a moment.



Herman

Two independent settings. Inference geo controls where the model runs. You can set it per request or through a workspace default. Workspace geo controls storage at rest and endpoint processing, including things such as image transcoding and code execution.



Corn

That second category is doing more work than the word storage suggests. If I upload an image, converting it before the model sees it is computation too. It just isn't the model's inference.



Herman

Right... no, let me narrow that. It isn't the particular model execution covered by the inference geography setting. An application can have several computational stages, and these controls divide them rather than assigning the whole application one address.



Daniel

Could I select Europe for inference and somewhere else for storage?



Herman

Not on that direct API menu as currently documented. Inference offers global, which is the default, or US. Workspace geography currently offers only the US, and it can't be changed after the workspace is created.



Corn

A globe with one selectable country.



Herman

Global means inference may run in any available geography. US means it runs only in US infrastructure. There is no European, Asian, or Israeli selection in that inference setting. That says what the control offers. It doesn't say there are no European machines.



Daniel

Does the response tell me which choice it actually used? Could I inspect a request afterwards?



Herman

The usage information includes an inference geography field, but its values are still US or global. It gives you a per-request indication at that coarse level. It doesn't return a city, a cloud region, or a facility identifier.



Corn

Global is a disappointing postcard.



Herman

Particularly because it isn't a geographic observation in the ordinary sense. If it says global, you can't deduce a specific destination. Nor should you assume it necessarily ran outside America. America remains inside the permitted global set.



Corn

That distinction matters. Otherwise somebody will plot every global response as an international transfer and produce an extremely confident map of information they do not possess.



Herman

What would you accept as evidence of a specific location, Corn?



Corn

A provider's explicit execution record would help. A short response time wouldn't. If the answer arrives quickly, I know it arrived quickly. I don't suddenly know the building had a view of the Mediterranean.



Daniel

Could latency at least help narrow it down? Europe seems like an obvious guess from here.



Herman

It can inform a guess, but request duration includes much more than distance. Queuing, prompt processing, output length, and the serving system's load can all matter. The network endpoint you reach may also be an entry point that forwards work elsewhere.



Corn

Measuring the front door doesn't locate the kitchen.



Herman

Capacity is part of the documented global endpoint policy. The service can route toward regions with available capacity. Precisely how it weighs latency against queues, model availability, and other operational factors is not publicly specified in the material we're discussing.



Corn

And the routing layer just picks whichever region feels like taking the request, like a flock of geese deciding where to land.
Gloria: Gloria, Standards and Practices. I need to formally clarify that the phrase about routing requests to whatever server feels like taking them, as though the load balancers



Corn (continued)

were geese choosing a field, drew a complaint from the International Association of Waterfowl. Policy four one seven, governing the imputation of whimsy to poultry. This is the fourth clarification today. Please stop.



Corn

My apologies to the waterfowl.



Herman

The actual capacity decision has economic value. Letting a provider choose among more infrastructure gives it more room to handle demand and outages. Restricting the permitted pool removes some of that flexibility.



Daniel

Is that why choosing the US costs more?



Herman

That's the economic interpretation. The documented fact is a one point one times multiplier for US-only inference on Claude four point six and later. It applies across token categories, including input, output, cache writes, and cache reads. Global is the standard price.



Corn

Even the cached tokens pay for their address.



Herman

The premium purchases a geographic constraint. It doesn't reveal how much that constraint costs Anthropic on each request. We shouldn't pretend the price multiplier is a measurement of infrastructure expense.



Corn

I was about to object to that, Herman. You'd nearly made the surcharge sound like a law of physics.



Herman

A pricing decision, then. The physics department is acquitted.



Daniel

Does that control apply to older models too?



Herman

No. Claude four point six and later support it. The four point five generation and earlier reject that parameter with a four hundred error. That includes Opus, Haiku, and the deprecated Sonnet from that generation.



Corn

Older models decline to discuss their whereabouts.



Herman

Organizations that had already opted out of global routing were migrated to a US-only allowed geography and a US default. That's important because an existing restriction wasn't simply replaced by the new global default.



Daniel

The partner offerings sound broader. Are European customers getting something through a cloud partner that the direct API doesn't offer?



Herman

They can have different regional options. Bedrock and Vertex are listed as covering Asia-Pacific, Canada, Europe, and the US. Microsoft Foundry is listed with US coverage and Europe coming soon. Availability still depends on the particular product and deployment.



Corn

Europe can exist in one catalogue and be absent from another without anybody moving a server. The catalogue is describing what you're allowed to buy.



Herman

The regional picture includes Japan and South Korea, with Singapore, India, and Australia also named. Global endpoints dynamically route toward available regional capacity. Regional endpoints carry a guarantee that storage and model processing remain within the specified region.



Daniel

Could we count those regions and get a rough number of inference locations?



Herman

We'd get a count of named regional offerings, not physical sites. A cloud region can contain multiple facilities. Several products can share infrastructure. A model can also be available in only part of a cloud's overall footprint.



Corn

Counting menu entries won't count ovens.



Herman

Anthropic's broader default policy names select countries across the US, Europe, Asia, and Australia, while saying data is stored in the US. That gives us a broad permitted footprint, not a dispatch record for Daniel's next request.



Daniel

How does OpenAI compare? Is its menu more informative about where the GPU work happens?



Herman

It separates the concepts explicitly. For ChatGPT, data residency concerns storage at rest. Inference residency means GPU execution on the covered customer content happens exclusively within the selected region. Those guarantees have different availability.



Corn

This is where I'd expect people to read the storage promise and mentally add the GPU promise. A saved conversation has an address, therefore everything that touched it must have worked there. That feels intuitive, but it doesn't follow.



Herman

ChatGPT inference residency is offered in Europe, meaning the European Economic Area plus Switzerland, the United States, and the United Arab Emirates. It requires data residency in the same region.



Daniel

Only those places for inference, even though the storage list is longer?



Herman

The published data residency menu covers ten regions. Alongside those inference destinations, it includes the UK and Canada, plus Australia, India, Japan, Singapore, and South Korea. The important comparison is the broader storage menu against the narrower GPU execution menu.



Corn

Your conversation can have more housing options than its arithmetic.



Herman

One more boundary matters here. We're discussing ChatGPT's inference residency description, and then OpenAI's API configuration and pricing. Those are related, but they aren't one undifferentiated product contract.



Daniel

What does the API side require?



Herman

Data residency is a project configuration requiring sales approval. Eligible models released on or after March fifth carry a ten percent uplift. You don't establish eligibility merely by putting a regional preference into an ordinary request.



Corn

Geography comes with a sales representative.



Herman

The UAE inference offering is also feature-limited. Its listed models are GPT five point five and GPT five point six Luna. Image generation and internal search aren't included.



Herman

Other exclusions include ChatGPT Work, improved memory, and GPT Live. Selecting that region doesn't mean the complete feature set automatically follows you there.



Daniel

If the GPU stays inside the selected region, does any of my content still leave it?



Herman

Potentially, yes. OpenAI explicitly says non-GPU processing may still occur globally. Authentication, routing, and analytics may occur elsewhere. Examples of CPU work include extracting text from uploaded PDF or Word files and coordinating steps across services.



Corn

The PDF can travel before the GPU stays put.



Herman

Corn, would you call that misleading, or a reasonably precise division of the service?



Corn

The division is precise. The misunderstanding is predictable. If the headline says inference residency, I'd expect the GPU promise they're giving. If somebody sells it to me as everything stays here, they've enlarged that promise beyond what the documentation says.



Herman

That's my reading too. It makes the covered operation more specific, not less useful. A GPU execution guarantee is valuable. It just doesn't automatically encompass the parsing, authentication, routing, and orchestration around it.



Daniel

What about Google? Since Google has its own huge infrastructure, should we expect a clearer physical picture there?



Herman

Ownership of infrastructure doesn't automatically produce public topology. We have a documented Google-managed Claude offering with data residency in the US, EU, and Asia-Pacific, plus controls for network boundaries and customer-managed encryption keys. That establishes an offering, not a global site inventory.



Corn

Nor does a Claude offering establish Gemini's execution map. The building owner may be the same; the service terms aren't interchangeable.



Herman

I can't give you a confidently established first-party Gemini inference residency menu here. Google's specific picture is the least settled part of this comparison. I'm not going to fill that gap using the Claude product or assume every Google Cloud region serves every Gemini model.



Daniel

Does encryption tell us anything about whether the weights are copied into those regions?



Herman

Not from those controls alone. Customer-managed encryption keys concern a particular data protection arrangement. They don't expose an inventory of model weight replicas, their backups, or the hardware on which they're loaded.



Corn

There's a separate question buried in your original prompt, Daniel. Even a perfect record of where our request ran wouldn't tell us every other place the weights exist.



Herman

Yes. Execution requires access to the model's parameters in some usable form, but the public geographic guarantees don't describe the whole distribution system. They don't enumerate idle copies, recovery copies, or how model material is staged for serving.



Daniel

That's the part I find difficult. These are sensitive weights, yet the service needs enough copies to serve people globally. Do we know how they reconcile that?



Herman

Not at the detail needed to draw that architecture. Distributed serving doesn't necessarily mean handing every location an ordinary downloadable file, either. We shouldn't invent the packaging, access controls, or replication process from the fact that regional execution is offered.



Corn

Secrecy about the weights isn't evidence that there is only one copy. Regional availability isn't evidence that every server holds one. Both shortcuts turn uncertainty into a diagram much too quickly.



Herman

The general cloud pattern is visible in Azure OpenAI's global batch documentation. Stored data remains in its designated geography, while inference may be processed at any Azure OpenAI location. Again, storage location and execution permission are separately described.



Daniel

Bring that back to my request from Jerusalem. What's your actual best guess, without pretending we can trace it?



Herman

For an ordinary direct Anthropic API request with no restrictive setting, the applicable default is global. Its published default footprint spans select countries in the US, Europe, Asia, and Australia. There is no Israel inference selection, and Israel isn't named among those default destinations.



Corn

Where you press send doesn't add a residency instruction.



Herman

My guess would be Europe or the US. Europe is a plausible candidate on proximity grounds; the US is another documented pool. That's an inference, not a measured probability. I can't rank the destinations reliably without knowing capacity and routing conditions.



Daniel

Would you say Europe is most likely?



Herman

I'd call Europe plausible, not proven most likely. Nearness alone doesn't establish the largest available pool for your chosen model. The documents permit other destinations. They don't give us the traffic shares needed to turn that guess into a percentage.



Corn

Our map has acquired a question mark. An improvement on your pins.



Herman

A more defensible cartographic instrument.



Daniel

Then nobody can give me the number of physical inference sites either?



Herman

We don't have a published authoritative count for Anthropic, OpenAI, or Google. Region lists are proxies, and imperfect ones. Someone quoting an exact total needs to establish whether they're counting buildings, cloud regions, deployment products, or actual serving clusters.



Corn

Even a facility count wouldn't tell you how many were serving your particular model. One enormous site and several small deployments can have very different capacity.



Herman

To identify the destination of a particular request, you'd need more than...



Hilbert

A passport for the prompt, Herman. Mine lives beside the router. I own a receipt printer plugged into the little computer under my desk, and every Claude request gets a departure slip. No destination gets printed without evidence. I leave a blank for the arrival stamp.



Corn

You have established an unusually efficient passport office.



Hilbert

It isn't an office. It's a condition of membership in the Prompt Repatriation Society. I chair the arrivals committee. We meet online every Thursday. Our rule is that no prompt gets a second question until its first answer has returned. I currently have eleven thousand blank arrival slips sewn into a curtain. It hangs across the kitchen doorway. My snout catches the bottom row when I carry tea through.



Herman

The global response field won't supply the destination stamp you're waiting for.



Hilbert

I heard you explain that. You said global rather carefully, which usually means Corn is about to object to the noun. I print it in the comments box. The destination stays empty. The minutes require that distinction. Nessa Keane can separate thirty-eight perforated slips in a minute without tearing the border. She insists that entitles her to sign arrivals. It does not. The dispute remains open.



Corn

Her speed won't establish jurisdiction.



Hilbert

Can the US-only setting make one answer arrive in America while the society's account remains registered somewhere else? I need that exact arrangement for the motion about the missing treasurer. I don't need a city. A country will do.



Herman

Execution geography and account registration are different things. The setting constrains inference, but it won't prove where a person is.



Hilbert

Fine. I can leave his address blank too. The curtain is now two metres forty wide. The printer paper costs me six shekels a roll.



Corn

That paper bill exists whether he learns the destination or not. There are costs attached to recording an answer that still doesn't contain the information you wanted.



Herman

More telemetry could improve that, but a regional label would still need a clear meaning. Does it identify GPU execution, the entry point, preprocessing, or storage? Adding a city name without specifying the operation could create another confident misunderstanding.



Daniel

That's probably the distinction I'd have missed. I'd have treated a location returned by the API as the location of the whole request.



Corn

The common misconception is that a data residency promise gives the whole interaction one geographic address.



Herman

Storage, GPU execution, and surrounding processing can have different guarantees. Even an inference residency commitment can leave non-GPU work outside the region. The exact operation named in the promise matters.



Daniel

At least that gives me a grounded answer. We can describe the permitted footprint, but not trace my ordinary prompt to a building.



Corn

The next useful expansion would be execution records precise enough to fill that destination blank, without pretending one location describes every stage.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

For more along these lines, there's episode forty-eight, Renting vs. Building; episode twenty, Architectural AI; and episode eighteen, Beyond the GPU. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoyed the conversation, leave us a review wherever you're listening.



Daniel

Thanks, both. See you soon.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Anthropic, Data residency, Claude Platform docs [primary]**
   https://platform.claude.com/docs/en/manage-claude/data-residency

2. **Anthropic Privacy Center, Where are your servers located? Do you host your models on EU servers? [primary]**
   dated June 15, 2026
   https://privacy.claude.com/en/articles/7996890-where-are-your-servers-located-do-you-host-your-models-on-eu-servers

3. **Anthropic, Regional compliance [primary]**
   https://claude.com/regional-compliance

4. **AWS, Data residency [primary]**
   Claude Platform on AWS
   https://docs.aws.amazon.com/claude-platform/latest/userguide/data-residency.html

5. **AWS, Claude Platform on AWS user guide [primary]**
   https://docs.aws.amazon.com/pdfs/claude-platform/latest/userguide/cpa-ug.pdf

6. **OpenAI, Data controls in the OpenAI platform [primary]**
   https://developers.openai.com/api/docs/guides/your-data

7. **OpenAI Help Center, Data residency and inference residency for ChatGPT [primary]**
updated ~2026-10-21
https://help.openai.com/en/articles/9903489-data-residency-and-inference-residency-for-chatgpt

8. **OpenAI, Expanding data residency access to business customers worldwide [primary]**
https://openai.com/index/expanding-data-residency-access-to-business-customers-worldwide/

9. **Google Cloud Blog, Expanding Claude availability and deployment choice on Google Cloud [primary]**
https://cloud.google.com/blog/products/ai-machine-learning/expanding-claude-availability-on-google-cloud

10. **Microsoft Learn, How to use global batch processing with Azure OpenAI in Microsoft Foundry Models [primary]**
https://learn.microsoft.com/en-us/azure/foundry-classic/openai/how-to/batch

11. **OpenIntel, Anthropic**
Digital Infrastructure Audit
https://openintel.uk/anthropic/v-dig/

12. **Sub-Processors.com, Anthropic Subprocessors**
https://sub-processors.com/company/trust.anthropic.com

13. **Third Party Index, Anthropic**
checked 2026-10-23
https://thirdpartyindex.com/vendors/anthropic