Episode Book

EPISODE #6146

Grounding Is Not Solved: DIY vs Gateway vs Native Search

PUBLISHED	19 October 2026 at 19:56 UTC
DURATION	25:54
LISTEN	myweirdprompts.com/episode/grounding-diy-gateway-native-search-49268/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	20 October 2026 at 07:46 UTC

SYNOPSIS

Grounding used to mean wiring up a search API, a scraper, and a lot of glue. Now there's a middle tier — gateways like OpenRouter and Vercel that bolt search onto any model — plus models that search on their own. This episode walks through all three routes: the five moving parts of DIY, how gateway routing actually decides which engine to use, and what native grounding really costs once token weight and hidden dials enter the picture. Along the way: per-search versus per-prompt billing, encrypted result blocks you have to pass back, domain-filter limits that range from zero to a hundred, and a worked thousand-question cost comparison with prices as published this week.

TRANSCRIPT



Corn

The line you hear everywhere right now is that grounding is solved. You tick a box, the model reads the live web, citations show up, everyone goes home. I'd love that to be true, mostly because I'd love to go home.



Herman

It's half true, and the half that's false is the part with the invoice. The plumbing didn't disappear. It moved to someone else's bill.



Corn

Which is Daniel's question, more or less. Getting a model to answer from today's web used to mean a search API, a scraper, and you pasting the results into the prompt yourself. Now there are two other ways. A gateway like OpenRouter bolts search onto any model you pick, or you call a model that searches on its own. He wants examples of each, how they stack up against building it yourself, and whether outsourcing always costs a steep premium.



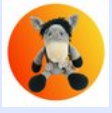
Herman

We did an episode in the spring on why native grounding keeps failing, and I leaned hard toward the bolt-on search services. Prototype-only was my phrase.



Corn

Confident. You said it like a man who'd never been wrong in front of a microphone.



Herman

I was allowed one. And some of it has aged badly, because the interesting thing now isn't native versus bolt-on. It's that a middle tier grew up between them, and the prices across all three have moved.



Corn

Start with the one everybody already knows. Do it yourself.



Herman

Five moving parts, and people forget how many there are. Something decides the question needs the web. Something rewrites the question into a search query, because users type nonsense. A search API returns links and snippets. You fetch and clean the pages, if snippets aren't enough. Then everything goes into the prompt and the model answers, which is a second model call on top of the one that decided to search.



Corn

So DIY already has two model calls before you've shipped anything.



Herman

Often, yes. On the search line, the list prices are cheap. Exa's standard search is seven dollars per thousand requests. Tavily is eight dollars per thousand on pay as you go, down to five on its bigger plans. Brave is five.



Corn

And the discount end?



Herman

Serper, the Google results reseller. Third-party price lists put big packs around thirty cents per thousand, though I couldn't read Serper's own pricing page, so treat that as a rumour with a number attached. Perplexity also sells raw search without the model attached, five dollars per thousand, or one dollar for what they call fast search.



Corn

Seven dollars per thousand. That's under one cent a search. I've paid more for a stamp.



Herman

That's why the search fee is almost never the story. Hold that thought, because the cost segment is where it bites.



Corn

Then the middle tier. What does a gateway actually do?



Herman

Take OpenRouter. You add a suffix, colon online, to any model name, or you add a web plugin to the request. The mechanics are the part I find neat. If you don't choose an engine, OpenRouter checks whether the model's own provider has native search, and that covers OpenAI, Anthropic, Google, Perplexity and xAI. If it does, it uses that. If it doesn't, it falls back to Exa.



Corn

So the same flag means two different things depending on the model.



Herman

Yes. You can also force an engine: native, Exa, Firecrawl, Parallel, or Perplexity. The default is five results, and they come back as citation annotations attached to the answer. Model tokens are billed on top, same as always. Everything I just said is in OpenRouter's own docs. What I couldn't find is the per-context-size price when it uses the provider's native search, and whether any platform fee sits on top of your credits.



Corn

So there's a gap in the middle of the price list. Naturally.



Herman

Here's the part I did confirm. When it routes through Exa, the search fee is seven dollars per thousand. Which is exactly Exa's own direct price. No visible markup on the search line.



Corn

Hold on. A middleman charging the same as the shop?



Herman

On that line, yes. If there's a premium it's elsewhere, and I'll say where.



Corn

And the others?



Herman

Vercel's AI Gateway does something similar but as tools you hand any model. Perplexity search at five dollars per thousand, Exa at seven, a couple of others, and it also passes through the native tools from Anthropic, OpenAI, Google and xAI. The docs were updated in September, and some of those engines had promotional free periods, so the number you pay can differ from the list.



Corn

Free periods. Pricing page with a loyalty card.



Herman

LiteLLM is the odd one. It's a proxy you run yourself. When a model emits a web search tool call, LiteLLM catches it, runs the search against a backend you've configured, then makes the follow-up request with the results. Backends include Perplexity, Tavily, Exa, Brave, even a self-hosted SearXNG. You bring your own key and pay the backend directly.



Corn

Then it isn't outsourced at all. It's DIY with the plumbing already installed.



Herman

That's how I'd class it. Gateway shaped, DIY priced. As for Cloudflare's AI Gateway, Requesty and Portkey, what I found was routing, caching, fallbacks. I didn't find built-in web search. I can't tell you it doesn't exist, only that I didn't find it.

L

Larry

Close your eyes. Breathe in. Imagine a morning with no edges. Silk. Warm stone. The hush of a thousand citations, falling like snow on a distant, expensive hill. Somewhere, a door opens onto a room you have never been in, and yet you know exactly where the light switch is. Introducing Verité Grounding. Not a search. Not a result. A feeling. A single drop, applied to the wrist of your language model, and suddenly it knows things. Things about today. Things about you. Things about the weather in a city it has never visited. Verité Grounding. The scent of knowing. Whisper it. Whisper... Verité. Available in two sizes: one for the home, and one for the home with a larger kitchen. Verité Grounding is a mop. Results may include a clean floor. Please do not apply to models.



Corn

Right.



Herman

Every few weeks, it seems.



Corn

Back to the web. What about the models that do it themselves?



Herman

Gemini first, because its billing flipped. Grounding with Google Search on Gemini three and later gives you five thousand search requests a month free, then fourteen dollars per thousand. But it's billed per search query the model chooses to run, not per prompt. On the older two point five models it was per prompt.



Corn

So one question that makes the model search three times is three charges.



Herman

A research-flavoured prompt can cost a multiple of the headline rate. The five thousand free pool covers a small app completely, though. OpenAI's version is ten dollars per thousand calls on the reasoning models, plus the search content billed as tokens at the model's rate. The old search preview models, the ones with search in the name, shut down on July twenty-third, so anything still pointing at them is broken.



Corn

And Anthropic?



Herman

Also ten dollars per thousand searches plus standard tokens, and the results count as input tokens, in that turn and in later turns. There's a cap you can set on the number of searches per request. Failed searches aren't billed. Simple questions typically take one to three searches. Anthropic's own documentation example, one search, came to just over six thousand input tokens.



Corn

Six thousand tokens to find out who won a football match.



Herman

I'm going to ask you something, because I don't know which way you lean. Is that the part that bothers you, the token weight, or the dial being hidden?



Corn

The dial. If the model decides how many searches to run, then I've given it a credit card and told it to use its judgement.



Herman

Fair, and the controls vary a lot. Anthropic exposes the query it ran and you can cap uses. OpenAI lists every URL it consulted, which is more than the citations it shows. For domain filtering, OpenAI takes up to a hundred domains, xAI takes five, Anthropic lets you allow or block but not both at once, and Google's native search through OpenRouter has no filtering at all.



Corn

Then there's Perplexity.



Herman

Where search is the product. Sonar models search every time. I'll flag this one as shaky: Perplexity's documentation no longer seems to list Sonar rates. The figure I have is from a third-party page dated June. It puts Sonar token costs around one dollar per million and a request fee between five and twelve dollars per thousand depending on how much context you ask for. So don't plan a budget on it.



Corn

And xAI?



Herman

Their old Live Search was shut off on January twelfth. Now it's web search and X search as agent tools, five dollars per thousand calls. Anything that still calls the old endpoint gets an error.



Corn

Before the churn, one thing I want to know. If the model does the search for me, what do I owe it back?



Herman

More than people expect. Google's terms say you display their search suggestions with the answer. OpenAI wants the inline citations visible and clickable, and its sources field lists every page the model consulted, which is more than the ones it cites. Anthropic returns the results as encrypted blocks that you have to pass back on the next turn, so you can't quietly edit what the model saw.



Corn

So I rent the search and I'm also contractually a billboard.



Herman

A little, yes. xAI's X search has its own twist: it's billed per item returned, posts at five dollars per thousand and profiles at ten, and OpenRouter says it's passing per-item billing through from September twenty-first. A single question that pulls fifty posts is fifty items, not one call.



Corn

Then the headline rate is a polite suggestion.



Herman

On that one, closer to a lower bound. On the Perplexity agent API, I'd note, search is a quarter of a cent per call, and page fetches half of a tenth of a cent, so the raw pieces are getting very cheap. That's a reason the fee isn't the story.



Corn

That's four things in a year that a DIY builder simply wouldn't have noticed. Search previews shut off, billing flipped, endpoint killed, Anthropic's tool revised more than once.



Herman

That's the part people undercount. If you build against a search API and your own glue, the vendor churn hits your search vendor and nobody else. If you build against native tools, it hits you every time a model provider changes its mind. A gateway absorbs some of it.



Corn

Now the bit Daniel actually asked. Does anyone get fleeced?



Herman

Let me do a worked example with round assumptions. A thousand questions. Each triggers one search. The prompt plus injected results comes to about six thousand input tokens, and the answer is five hundred tokens out. Same model for everything, GPT five, at one dollar twenty-five per million in and ten per million out. Prices as published this week, October nineteenth. They will change.



Corn

Go.



Herman

Tokens first, because they're identical across the DIY and gateway routes. Six thousand in times a thousand questions is six million input tokens, so seven dollars fifty. Five hundred out is half a million output, five dollars. Twelve dollars fifty in tokens.



Corn

Before any search fee.



Herman

Before any search fee. Now add the search line. DIY with Tavily at pay as you go, eight dollars, so twenty dollars fifty. DIY with Brave, five dollars, so seventeen fifty. Exa direct, seven, nineteen fifty.



Corn

And through OpenRouter's Exa engine?



Herman

Nineteen dollars fifty. The same. If you pick Parallel's fast tier through OpenRouter, it's a dollar for the search, so thirteen fifty.



Corn

So the gateway beat Tavily.



Herman

On this setup, yes, and the cheap Serper packs would land around thirteen fifty as well, though you'd also be fetching and cleaning pages yourself. Native OpenAI is ten dollars for the search line, so twenty-two fifty. That's the real premium on this setup, two or three dollars a thousand over the cheaper DIY routes.



Corn

Two dollars a thousand questions. I'll have to sit down.



Herman

Sit down for Gemini. Say Gemini three Flash. The tokens are cheaper, four dollars fifty for the same load. If you're inside the free allowance, that's your whole bill. Pay for the search at fourteen dollars per thousand and it's eighteen fifty. But have the model run two queries per question and the search line doubles to twenty-eight, so thirty-two fifty.



Corn

So the free tier makes it the cheapest option in the building, and the paid tier is the second-priciest.



Herman

And xAI with Grok four point three comes out around thirteen seventy-five, Perplexity's Sonar at the low context setting roughly six fifty, though that's the shaky figure, and that one searches internally so your token bill only covers your prompt and answer.



Corn

I notice that the search fee ranges from a dollar to fourteen, and the token bill is twelve fifty before anyone's searched anything.



Herman

That's the hidden cost. The search results land in your input tokens. On GPT five the injected results alone are seven dollars fifty per thousand questions, close to the entire search fee on the Exa route. And it's the same on every route. The model you pick matters more than whether you chose DIY, gateway or native.



Corn

Test it. Swap GPT five for the small model and see whether your argument survives.



Herman

Good test. GPT five mini is twenty-five cents in and two dollars out per million. Same six thousand in, five hundred out, and the token bill falls to two dollars fifty per thousand questions.



Corn

So the fee you were shrugging at...



Herman

Is now most of the bill. Tavily's eight dollars against two fifty in tokens. On a cheap model the search line is three-quarters of what you pay, and that's exactly where the Parallel engine at a dollar, or a big Serper pack, starts to matter. Both effects are real. They just take turns depending on the model.



Corn

And Anthropic?



Herman

Ten dollars per thousand for the search, then Claude's token rates, which I haven't got current figures for, so I won't invent a total. The shape is the same as OpenAI's.



Corn

So the premium on the search line is real in one place.



Herman

Native, mainly. Ten to fourteen dollars per thousand against one to seven for the raw APIs. The gateway line is basically list price. The big cost everybody shares is token inflation, and the vendors know it. Anthropic has dynamic filtering, where code trims results before they reach the context, and OpenAI has a token budget setting on returned content.



Corn

Now the part you'll defend and I'll doubt. DIY is cheaper on paper, so why isn't it just cheaper?



Herman

Because the invoice isn't the cost. Who writes the query rewriting loop? Who keeps the scraper alive when a site changes its markup? Who handles rate limits, which on some of these tiers are very low? Brave's answers product is two queries per second. And you're running the extra model turn. I can't put a sourced number on engineering time, and I'd be making one up.



Corn

Make one up, I'm asking. What would you do if you were building this on a Tuesday?



Herman

Gateway... no, wait. Gateway, for the first version, because switching engines is a string change. DIY only once I'd seen enough traffic to know what I was paying for.



Corn

You've just argued yourself out of your own spring episode.



Herman

Partly. I still think native search is the weakest on quality control, and I stand by that. But on price, I was wrong to treat it as the dear option across the board.



Corn

You say that now. Give it a month and you'll be back to calling it a prototype.



Herman

Possible. Here's a real question for you, Corn, because I can't settle it myself. Would you let a model run its own searches unattended, with nobody capping the number?



Corn

With a spending cap, yes. Without one, no. I don't let myself near the snack drawer unsupervised, and I've got better judgement than a language model.



Herman

Your snack drawer has a lock, though. A max uses setting is that lock, and half these tools don't ship with it switched on.



Corn

Then I want the lock on by default and the key in my pocket. Slowly in my pocket, but in it.



Herman

Noted. Keys in pockets, locks on drawers.



Corn

The thing that interests me is the free tier on Gemini. Your cost per question drops to the token bill, so a hobby project is basically free.



Herman

Which is the reason it exists, and I'd be careful about it. The moment the app grows past five thousand searches a month, it's the second most expensive line on the sheet. That's the part nobody budgets for, the step between free and fourteen dollars.



Corn

It's a cliff with a welcome mat on it.



Herman

The same pattern, in the other direction, with the older numbers. Under the two point five scheme it was something like thirty-five dollars per thousand grounded prompts, according to one forum post I can't date. So the headline rate fell, and what you get charged for changed.



Corn

Something nags me about the token side. Those pages sit in the conversation after the answer, so the second question pays for the first question's pages again, doesn't it?



Herman

On Anthropic's tool, yes, the results count in the search turn and in every later turn. Every follow-up drags the same pages along and you pay for them again, the way you'd keep paying rent on a...



Hilbert

Storage unit. Nine dollars a week, and he'll tell you it's eight and then there's the key.



Herman

Who does?



Hilbert

Brannock. He does my printing. Fenwick Row, number forty. Eleven years I've bought envelopes off him, the long ones, and he has never once said a price before ringing it up. You made us wait for the price earlier, Herman, and you were right to, because it's the same thing, the number's never where you expect it. I'm skipping the API this week. I've accepted everything you said about the glue and the maintenance, so I'm having him do the search. Five results per question, that's the default, so Brannock prints each one on yellow stock, one sheet per result, and I read them aloud to the model into the desk microphone. I've bought a lectern. Ninety-four dollars.



Herman

The five sheets match the default result count, so the setup is at least faithful to how the gateways work.



Hilbert

Now the part neither of you got to. The pages ride along in every later turn. So I read them again. Every turn. By my ninth question I'm reading forty-five sheets. Forty-four. The one with the coffee ring doesn't count.



Corn

That's a lot of reading for one microphone.



Hilbert

Corn, the lectern's built for a person with a chin. I read with my snout against the lamp, and the lamp is Brannock's. He bills me for it by the hour. He won't print a thing unless his goose is in the shop. Pauline. She's on the payroll at eleven dollars and forty cents an hour, and she has a pension. I've seen the paperwork.



Herman

A goose on payroll is a fixed cost, so it behaves like the engineering time we were talking about. It doesn't scale with the number of queries.



Hilbert

She sits on the guillotine cutter during the whole print run because it's the only warm part of the shop. There's a second goose, Dennis, but Dennis only comes in for Christmas cards. Pauline decides the stock. Yellow. He's never offered me white.



Corn

So the bottleneck in the system is a goose with a pension.



Hilbert

And Thursdays. He does one job on a Thursday, whatever the job is, and this Thursday it's legal pads for the chiropodist next door. I asked for a Wednesday twice. He said he'd put it to Pauline. Four years on the stamp counter at the sorting office, I know what a stamp is worth, and he takes them off me at face value and gives them back as change.



Herman

Because the pages are re-read on every turn, your total grows faster than the number of questions. It's closer to the square of the turns than a straight line.



Hilbert

I worked it out on the back of one of his envelopes. Three hundred and twelve dollars for the first week. Pauline has Fridays off.



Corn

Three hundred and twelve for a week of goose and yellow paper. Next time someone says a search API is dear, I know what to hold it against.



Herman

There's a plain version of that in the real numbers. The expensive part of the second question isn't the second search, it's re-buying the first. So on any of these routes, watch how long the conversation runs, and whether the tool lets you drop old results out of the context.



Corn

Right. Daniel asked for a decision, so make it. Who gets which tier?



Herman

Native first, for anyone who wants it working this afternoon at small volume. Gemini's free pool is a legitimate reason all by itself. So is one API call that comes back with citations attached. And so is needing something only the native tool has, like Google's index or X posts through xAI.



Corn

When does native stop being enough?



Herman

When you need control, or when the bill is driven by searches you didn't choose to run. A model that decides how many queries to make is a model holding your credit card. Anthropic's cap on uses helps. OpenAI's domain filter helps. Google's native search through a gateway has no filter at all.



Corn

So I'd go gateway first.



Herman

Gateway when you run more than one model, or when you want to swap search engines without rewriting code. It absorbs the churn. If a provider kills its search model, that's a string change. On the search line it sits at list price, at least for Exa, and with the caveat that I couldn't see the price when it routes to a provider's own search. Check your first invoice. And LiteLLM if you want gateway behaviour with your own search contracts.



Corn

And DIY is for people who enjoy their weekends being taken from them.



Herman

DIY is for when search quality is the product. If you need to rank sources, fetch whole pages, dedupe, reach sites the APIs skip, or keep the results out of the context window, you want the control, and you pay for it in glue and maintenance rather than in the fee.



Corn

You'd pick which?



Herman

Gateway for the first version. Then I'd measure. You?



Corn

Gateway, and I'd go back to sleep after the invoice. The thing nobody tells you about cheap is that it still has an invoice.



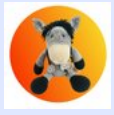
Herman

You'd have said native a year ago.



Corn

A year ago I'd have said a lot of things. Most of them were available as previews.



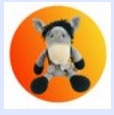
Herman

The shutdown dates taught me to read the fine print twice.



Corn

The belief I'd put money on is that outsourced grounding is always the pricey option. Someone else does the work, so someone else adds a margin.



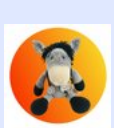
Herman

On the search line, the gateway sits at list price, the native tools run a few dollars per thousand above the raw APIs, and the cost that actually moves your bill, the injected tokens, lands on all three routes the same. Pick a cheaper model and you save more than you'd save by switching architecture.



Corn

The number I'll carry out of this is the one nobody budgets: how many old pages ride along in every turn. Whoever is charging for the re-reading, it usually isn't the search.



Herman

Thanks to Hilbert Flumingtop for producing.



Corn

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Herman

See you tomorrow.