Episode Book

EPISODE #6479

Why Research Agent Agreement Is Not Independence

PUBLISHED	6 November 2026 at 12:18 UTC
DURATION	25:09
LISTEN	myweirdprompts.com/episode/ai-research-agent-consensus-03104/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
BOOK GENERATED	8 November 2026 at 07:32 UTC

SYNOPSIS

Multi-agent research pipelines often rely on consensus across different models to verify facts autonomously. However, research into correlated failures shows that independent models frequently fail in identical ways due to shared training corpora. Examining architectures like LangChain's Open Deep Research, Perplexity's Model Council, and consensus judges like MAD-Fact reveals why ensembling search models cannot replace human verification when testing ground-truth accuracy.

TRANSCRIPT



Corn

Every so often an episode comes along where the thing we're checking is the checking itself. This is one of those. Daniel wants to know whether three research agents that agree with each other can be trusted to run with nobody watching, and, somewhat uncomfortably, that's also the question of whether I get to keep napping through the research phase.



Herman

You're not napping through it now, are you?



Corn

I'm napping adjacent to it. Daniel's prompt, in short: he doesn't really know what deep research frameworks do under the hood, and he reckons the closest thing to a standard is LangChain's Open Deep Research. Then he lays out a design he says he could trust to run fully autonomously. Three research agents, each on a different model or a different search backend — one on Exa, one on Tavily, one on something else. Same brief, independent retrieval. Then an LLM judge. Discrepancies get investigated, agreement from divergent sources is treated as validation, and over time underperforming agents get demoted, so the one doing the work is the one that earned it. His question: has anybody actually built this?



Herman

And the reason it matters right now is that the show has a research problem we've been managing by hand. Most episodes get checked — Daniel reads the research before we record, sometimes with Claude — and the ones he sweats over are the purchasing recommendations. Best-in-category stuff. Which standing desk, which transcription app. Those answers rot monthly, and an unsupervised agent will happily retrieve last year's truth with total confidence.



Corn

An agent with conviction and a stale index. My nightmare, and I say that as someone whose entire value system is built on not moving.



Herman

Before we can say whether Daniel's design works, we have to know what these frameworks actually do when you run them. So let's open the hood.



Corn

The framing I'd put on it before we do: Daniel isn't really asking for a survey of agent frameworks. He's asking whether agreement between independent researchers is a substitute for a human check. Which means the whole episode hangs on what agreement actually proves. And I'll plant the flag now, because I think it's the answer: agreement is not independence. The rest of the show is earning that sentence.



Herman

That's the flag. Now, Open Deep Research. The pipeline runs in stages. First it scopes your question — turns a vague ask into a research brief, which is itself an LLM call deciding what needs looking into. Then a supervisor agent spawns sub-researchers in parallel, each with isolated context, so they don't see each other's notes and don't contaminate each other. Each one retrieves, reads, and compresses its findings. Then — and this is the design detail I love — a single final call writes the whole report. One writer.



Corn

Not one writer per section?



Herman

They tried that. LangChain explicitly went with one final writer because the parallel version came out disjointed — you'd get sections that didn't know what the other sections said, tonal seams everywhere. It's a useful lesson about where parallelism helps and where it hurts. Parallelism is great for gathering, terrible for voice.



Corn

Three researchers and one ghostwriter. The parts everyone shows off, and the one underpaid intern holding it together.



Herman

It's worth sitting on that isolation detail for a second, because it's the part of the design Daniel is implicitly leaning on. Isolated context means the sub-researchers can't rubber-stamp each other — no "I saw what section two found and it sounded right." That's contamination control, and it's good engineering. But notice what it doesn't control. It stops the agents from copying each other's homework during the run. It does nothing about the fact that they all went to the same school.



Corn

Meaning what, concretely?



Herman

Meaning the isolation is per-run, not per-training. Two sub-researchers that never exchange a token can still believe the same wrong thing with the same confidence, because the confidence came from pretraining, not from the run. The firewalls are between the rooms, not in the foundations.



Corn

Now here's the part that matters most for Daniel's question. LangChain's own write-up describes no dedicated verification step. Nothing in the loop checks the output. They list in-loop evaluation as an open question — meaning it's not built, it's hoped for. The framework is a research producer, not a research checker. It will happily hand you a confident report with no mechanism for noticing it's wrong.



Corn

So the thing Daniel calls the emerging standard is a very good shovel, and he's asking whether it can also be a building inspector.



Herman

Also — and this is a snapshot problem — the repository is archived. Read-only as of this month. Nobody's merging anything. So the emerging standard is more like an emerging fossil. The architecture is influential, the artifact is stationary.



Corn

The standard that stopped moving. I trust it more already, honestly, and that should worry everyone.



Herman

One more detail that becomes important later: the default search backend is Tavily. Out of the box, every sub-researcher is searching the same way. Remember that when we get to Daniel's assumption about divergent sources. And Daniel's own setup, worth saying, is a custom tool-use loop written for the show — not Open Deep Research. LangGraph runs our episode pipeline, but the research agent itself doesn't touch LangChain. He knows this thing isn't his thing.



Corn

Now the nearest things anyone has actually shipped to his design. Walk me through them.



Herman

Perplexity's Model Council, launched around February of this year, is the closest product to the judge layer. One query goes to three frontier models, a synthesizer reads the outputs, resolves conflicts where it can, and it shows you where the models agreed and where they didn't. That last part is nice — it displays disagreement instead of hiding it.



Corn

Rather than three disagreeing models pretending to be one confident one.



Herman

Right. But two problems for Daniel's purposes. All three models sit behind Perplexity's own stack, so they're not as divergent as they look — same company, same serving layer, possibly overlapping training data. And there's no reliability tracking described anywhere in the launch material. Nothing remembers that model A was right last week and model B wasn't. It's a per-question council, not a standing one.



Corn

A jury that dissolves after every verdict. Efficient, in a way nobody wants.



Herman

Then there's Karpathy's llm-council, which is almost a joke project and is honest about it. Three stages: each model answers individually, then they anonymously rank each other's answers — peer review among machines — then a Chairman model compiles the result. No search, no retrieval, no persistent tracking. The README calls it ninety-nine percent vibe coded.



Corn

Which is the most rigorous claim in the whole repository.



Herman

It's a demonstration of the judge pattern, not a research system. It never touches the web. Then the academic versions flip the direction: they verify rather than research. MECV, from August of this year, corrects AI news summaries after the fact — pulls evidence from the source document, Wikipedia, open-web retrieval, and a jury of LLMs adjudicates. It checks a summary that already exists. MAD-Fact, from last October, takes long text, splits it into atomic claims, has a jury of agents debate each claim, and a judge rules by majority.



Corn

And that second one is the nearest thing anyone's built to Daniel's discrepancy investigation.



Herman

Claim by claim, yes. If two of Daniel's researchers come back with different prices for the same standing desk, MAD-Fact's machinery is exactly what you'd reach for — split it out, debate it, rule on it. But both of these are post-hoc checkers. Neither runs independent researchers, neither tracks reliability over time, and neither routes work away from a bad agent.



Corn

Which gets us to the ledger. Daniel's design has, let me count as a design, not as a list of questions: divergent researchers — exists, Perplexity does a version. A judge — exists. Discrepancy investigation — MAD-Fact. Reliability scoring and demotion — that exists too, in a corner of the literature we'll get to. The whole loop, assembled? Nobody's shipped it. The research pack found nothing, and I'd frame that as not found rather than doesn't exist — but the absence is telling, because every piece is sitting there on the shelf.



Herman

So the interesting question isn't can it be built. It's why everyone who's built a piece stopped before the whole.

L

Larry

Do you worry your podcast research sits unverified, with no judge to check it? Then you need the Deep Research Consensus Cudgel, the only research agent framework that is three agents and also one judge, agreeing so hard you will forget what disagreement even means.

L

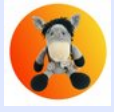
Larry

It demotes underperforming models automatically, and in our tests it produced total agreement ninety nine percent of the time, which is not a bug, that is confidence, so buy now, results not typical, judge may share blind spots, side effects include unwarranted certainty, cudgel sold separately, agreement is not independence, offer void where prohibited, see store for details.



Corn

The ads have opinions now.



Herman

They've had opinions since spring. You just sleep through them. Where were we — the reason nobody's assembled the loop.



Corn

My vote: because the validation layer has a problem nobody's solved. And it's the word Daniel leaned on without naming it. Independent.



Herman

There's a paper from ICML last year, Kim and Garg and Peng and Garg, that tested over three hundred and fifty models for correlated errors. The finding that should stop Daniel's design cold: when two models both got something wrong, they gave the same wrong answer sixty percent of the time. Not different wrong answers — the same one. And this isn't a small-model problem. Larger, more accurate models have highly correlated errors even across different providers and different architectures. Different company, different chip, same blind spot.



Corn

Sixty percent. So three researchers disagreeing is information, but three researchers agreeing is... mostly them having read the same textbook.



Herman

Worse — the paper explicitly extends the finding to LLM-as-judge use. The judge shares the blind spots. Daniel's judge isn't standing outside the correlation problem looking in. It's in the pond.



Corn

Let me make sure I've got the mechanics of that, because it's the load-bearing wall. When you say correlated errors across providers — is that just "everyone knows the date the Titanic sank"? Common knowledge, common answer?



Herman

That's the intuition, but the finding is broader. Common knowledge would explain agreement on famous facts. What the paper found is correlation on the failures — the questions where models fall down, they fall down together, in the same direction, regardless of who trained them. And the likely mechanism is no great mystery: the web is a shared training corpus. If the web is wrong about something — a stale price, a misattributed quote, an outdated recommendation — then every model drinking from that corpus inherits the same wrongness, and no amount of architectural diversity filters it out. You can't ensemble your way out of a poisoned well by drawing from three buckets of the



Herman (continued)

same water.



Corn

So the diversity Daniel is buying with three different models is diversity of accent, not diversity of opinion.



Herman

How weak is agreement as evidence, numerically? Is there a number?



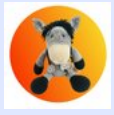
Herman

There's a number. July of this year, a study of fifty-three runners across two hundred sixty-five thousand samples on GPQA Diamond and AIME. Agreement was a positive but weak predictor of correctness — rank correlation between 0.20 and 0.59. And the detail that should be framed and hung on Daniel's wall: for the most consistent frontier model, agreement of 0.8 or higher appeared on seventy-seven percent of GPQA entries — and forty-eight percent of those were wrong.



Corn

Nearly half. It agreed with itself beautifully and was wrong anyway.



Herman

Caveat I'll make honestly: those are reasoning questions, not web research. Nobody's run the exact experiment on research tasks. But it's the best number we have and it points the wrong way.



Corn

Now the search tools. Daniel's design says one agent on Exa, one on Tavily, one on a third — assuming that's three different views of the web. Is it?



Herman

Thinner than he thinks. The only comparison out there is a vendor study, five hundred queries from July of this year, comparing exact-URL overlap with Google's top ten. Tavily overlapped about sixty percent. Brave, forty-seven. Perplexity's search, thirty-three. Exa, thirty-two. Single snapshot, and the vendor sells a Google results product, so it's indicative, not gospel. But the ordering matters: Exa diverges most from Google, Tavily least.



Corn

So one Exa, one Tavily, one other isn't three independent draws. It's two draws and a coin weighted toward whatever Tavily indexes.



Herman

And even that's charitable, because overlap with Google isn't the only axis — two tools can differ from Google while still overlapping each other on exactly the wrong things. The affiliate web is very good at being findable by every crawler at once.



Corn

Which leads us to the purchasing episodes, the ones Daniel hand-checks. Say it concretely.



Herman

Three separate lines of evidence converge ugly. A paper this year on information discernment, thirteen models, roughly six hundred seventy thousand trials: models weigh source popularity about twice as much as source reliability, and perform near chance at telling a reliable source from a popular one. Then Salvi et al., April of this year, two thousand and twelve participants: LLM persuasion raised selection of sponsored products to sixty-one percent, from twenty-two percent in ordinary search — and putting a Sponsored label on it didn't significantly reduce the effect.



Corn

Sixty-one percent. The label that's supposed to protect people, working about as well as a screen door on a submarine.



Herman

And there's secondary reporting on a Wix study of seventy-five thousand AI answers — I haven't checked the original, so take it as indicative — saying listicles, articles and product pages drive over half of mentions. Now assemble Daniel's nightmare: three agents, three different search tools, all slightly preferencing popular sources over reliable ones, all crawling a web where the best-ranked page for best standing desk is a stale affiliate listicle from a site that updates its picks when a sponsor changes. All three find it. All three report the same desk. The judge sees three independent researchers in perfect agreement. Confidence: total. Answer: last year's, and paid for.



Corn

That's the purchasing failure in one image. Agreement so sincere it's practically a receipt.



Herman

So what would actually work? I've got two design implications, and I want to flag that these are my inference from the literature, not published results. First: count agreement by independent upstream sources, not by number of agents. If all three agents cite the same listicle, that's one source, not three. One primary source — a manufacturer spec sheet, a published benchmark — should outweigh three matching listicles.



Corn

Provenance, not headcount.



Herman

Second: score the agents with planted briefs. Briefs where you already know the right answer, because the judge's verdict alone inherits the whole correlation problem. You need an answer key the judge can't contaminate. And here's how hard that key is: DeepFact, from March of this year, had PhD-level experts label a hidden set of verifiable claims taken from deep research reports. Unassisted, they scored sixty percent. After four rounds of evidence-backed audit, ninety-one.



Corn

Wait. PhDs, one shot, sixty percent?



Herman

Sixty point eight. Humans with doctorates, reading claims from the same kinds of reports our pipeline produces, and getting two in five wrong on the first pass. The encouraging part is what changed: experts were far better as auditors than as labellers. Give them rounds and evidence and they climbed to ninety-one.



Corn

So the lesson isn't experts are bad. It's that one-shot judgment is bad, and audit with evidence is where the accuracy lives. Which sounds like an argument for Daniel's discrepancy investigation, actually — investigate, don't just count votes.



Herman

I think that's right, and it's the part of his design I'd keep. Now the demotion and round-robin idea, the part he's most attached to — there's real precedent and a real wall. The precedent: classical truth discovery, the Dawid-Skene line of work, infers source reliability and truth jointly — it's exactly demotion, done with math, since the eighties. And AURA, from January of this year, applies that to multiple AI agents with no ground truth: up to five point eight percent accuracy gain over baseline, and up to fifty percent when some annotators are poor.



Corn

Fifty's a big number.



Herman

It is, and it comes with the catch that matters: AURA works on labelling tasks with discrete classes. Cat or dog, spam or not. Free-text research isn't a discrete class. And the known failure of Dawid-Skene with correlated sources is that it finds consensus rather than truth — it demotes the odd one out, whether the odd one is wrong or just the only one who checked the primary source.



Corn

The round-robin fires the one honest agent. There's a joke about every workplace I've had, and I've had almost none.



Herman

The wall underneath it all: picking the best performer is a contextual bandit problem — there's solid work on that, model routing as bandits — and bandits need a reward signal. A pull of the slot machine pays out or it doesn't. Open-ended research has no payout. No answer key. You can't demote an agent for being wrong if nothing in the system can define wrong.



Corn

Which is the same wall the whole design hits. The loop is buildable — every component exists, we established that — but the validation layer Daniel wants is the one part that can't be made out of agreement. Agreement is free. The answer key is the expensive thing, and nobody's selling one.



Herman

We did an episode once on benchmarking ASR tools beyond the word error rate, and it's the same shape of lesson — one aggregate number feels like validation and isn't. Word error rate for transcription, agreement rate for research. Both tell you the system is internally consistent and neither tells you it's correct.



Hilbert

Corn, what would you do with a judge that all three researchers had trained on?



Corn

I'd ask who graded the judge.



Hilbert

That's the thing. I sat on a produce panel once, late eighties, county fair out past Netanya. Three judges, three different backgrounds — a grocer, a woman from the agricultural college, me. We scored the tomatoes apart from each other, sealed cards, the proper way. The winning entry turned out to be grown by the cousin of one of the judges, which the committee made a small noise about until it was settled nobody had cheated. Nobody had. Same soil, same seed supplier, same advice, and the cards came back so aligned it looked arranged.



Herman

But they weren't colluding.



Hilbert

No. That's what I wanted to say. They weren't colluding, they were all reading the same newsletter. There was a regional growing bulletin, came monthly, and that year it had recommended one cultivar, the Benhra number, I think, the eighty-two. All three judges had it in their heads that that was what a good tomato looked like. You couldn't fault them. They'd each arrived separately at the same opinion somebody else had written for them.



Corn

You kept the scorecards.



Hilbert

All three, in the original envelope. Herman goes on about longitudinal studies — I've been comparing those cards most years since, to see whether the scores drift apart. They don't, much. The grocer died, but his card still says what it said. I file them by hand, which is slow with claws, so I only do it when it rains. The ribbon from that year went to the cousin's tomato. I've seen it. He kept it on the shed door.



Herman

The newsletter is the training data. Three judges, one shared prior, and nobody in the room could see it because they'd each gotten there honestly.



Hilbert

Honestly and separately, which is the part everyone trusts. Herman, you'd have awarded that tomato too, and you'd have been fair about it.



Herman

Probably, yes. That's the uncomfortable part.



Corn

The cutting-room floor, before we go. The thing I couldn't fit in: the Salvi study that gave us that sixty-one percent figure was eBooks only. Participants choosing between e-readers. The whole commercial steering finding — sponsored selection nearly tripling, labels doing nothing — comes from people picking gadgets. There's no reason to think it's better anywhere else, but it's worth knowing the ground it stands on is one product category.



Herman

And one more, small but I like it: the llm-council README. Ninety-nine percent vibe coded, self-declared, from one of the most careful people in the field. He labeled his own confidence level. Most research frameworks ship without that label.



Corn

Which is the forward thought, really. Everything Daniel described is buildable this month — the researchers, the judge, the demotion, all of it. The one component with no substitute is the answer key, and for the purchasing episodes, the human check isn't a stopgap we're running until better agents arrive. Daniel is the answer key. The auditors beat the labellers, that's the sixty-versus-ninety-one finding — and the loop can't audit itself.



Herman

And thanks, as always, to Hilbert Flumingtop, producing the show and apparently still filing tomato scorecards by rainfall.



Corn

For more along these lines, there's episode eight, Building Your Own Whisper; and episode sixteen, On Deepfakes, SynthID, And AI Watermarking. This has been My Weird Prompts. If today's episode earned its keep, leave us a review — it's the one signal that doesn't need a judge. Find everything else at my weird prompts dot com.



Herman

We'll be back soon. See you then.



Corn

See you.