Minerva26

Encyclopaedia Britannica, Inc. v. Perplexity AI, Inc. — When the AI Company's Own Database Becomes the Evidence

Meet and Confer Podcast · Kelly Twigger · 🎧 Listen · 📄 Case in Minerva26: *Encyclopaedia Britannica, Inc. v. Perplexity AI, Inc.* · No. 25 Civ. 7546 · September 2, 2026 · United States Magistrate Judge Sarah Cave · U.S. District Court for the Southern District of New York

Hi, and thanks for joining me this week. If you ever need to get discovery out of an AI company, and more of you are going to need to do this than you think, the order we're going to discuss today tells you exactly what to ask for, how far back your request should reach, and who ends up paying for it.

The question came up this week in *Encyclopaedia Britannica, Inc. and Merriam-Webster, Inc. v. Perplexity AI, Inc.*, in a decision from September 2, 2026, out of the Southern District of New York. I picked it because it gives us a third category on the map we've been building this year about how artificial intelligence shows up in discovery. Unlike the other decisions we've looked at previously, this is not a fight over somebody's ChatGPT prompts, and it is not a fight over whether a law firm can use AI to review documents. This is a fight over the AI company's own database and its own logs, the machinery behind the product, treated as the evidence itself.

Welcome back to the Meet and Confer podcast. My name is Kelly Twigger. I'm the CEO and founder of Minerva26 and the Principal at ESI Attorneys. I've been a practicing attorney and discovery strategist for just about 29 years. This is the Case of the Week series, where I choose a recent decision or order on the discovery of ESI and turn it into practical advice for you to use every day.

Here's why this case is important. And I mean this for everybody listening, not just the handful of firms litigating against the Perplexitys and OpenAIs of the world. This order is a plain English education in how a retrieval-based AI system works, what it stores, what it logs, and where the record of its own activity lives. That is exactly the vocabulary and information you need to know the next time any party in any case of yours used an AI tool of any kind, and you need to know what to ask for. So there's a lot of great educational material here to dig into, even if you're thinking, Kelly, I don't represent

these kinds of big companies, I'm never going to be asking for data from a Perplexity. You never know, and more and more smaller companies are starting to engage with AI tools, so you're going to need to know the vocabulary and how to ask for that information.

The Technology, First

Let's talk a little bit about the technology. Perplexity runs what it calls an answer engine. A user asks it a question, a prompt, which we've talked about before, and instead of generating an answer purely from what the underlying model learned during training, Perplexity goes out and grabs current information to ground its answer. If you've seen any advertisements or influencers telling you to use Perplexity for research, this is why. The way it does that is called retrieval-augmented generation, or RAG. Think of RAG as a live lookup step bolted onto the front of an LLM. The system searches an index of web content, called the RAG database, pulls back whatever looks relevant to your question, and hands that retrieved content to the large language model along with your original prompt. The model then writes its response using that retrieved material as the source.

That's the entire theory of the plaintiffs' case. Britannica and Merriam-Webster, both of which, dictionaries and encyclopedias are what they make, say that Perplexity copies their copyrighted encyclopedia and dictionary content into the RAG index at the input stage, and then reproduces or paraphrases it in the answers handed back to users at the output stage. Alongside that, they make a trademark claim that Perplexity's hallucinated content gets falsely attributed to their brands.

The second piece of technology at issue here is the UAL database, short for the User Activity Log. Every time someone uses Perplexity, four things get recorded: what the user asked, how the system decided to retrieve and rank content to answer it, the instructions sent to the underlying language model, and the answer the user ultimately got back. That is Perplexity's own record of its RAG system in action, at scale, across millions of queries. So when Britannica went looking for evidence of infringement, those two systems, the RAG database and the UAL logs, are exactly where that proof lived.

Where This Sits on the Map

So let's talk about how this relates to the cases we've already discussed on AI. In *Warner* and *Morgan*, the AI user was a litigant, someone typing into a chatbot to think through their own case, and the fight was about whether those prompts were protected work product. *Heppner* gave us the exception, take that same use case outside the lawyer relationship, and the protection disappears. In *Conservation Law Foundation v. Shell*, it was an expert's prompt that turned out to be discoverable. That is one region of the map, AI as something a person used, and the question is privilege.

Then in *Cerebras* and in *Schulte*, the AI user became counsel and their vendors, running generative AI as the actual review engine deciding what gets produced. *Cerebras* is the parties agreeing up front to a sweeping disclosure regime for all of that information. *Schulte* is the same technology, but contested, where a court refused to enforce that same transparency after the fact. That's the second region of the map, AI as the tool doing the discovery.

Britannica v. Perplexity, today's case, is a third region entirely. Nobody here is disputing the use of a chatbot or a review tool. The AI company's own operational infrastructure, the database it uses to run its product and the logs it keeps of every interaction, is the alleged instrument of the underlying wrong, and therefore the discovery target in its own right. This is not AI helping with discovery. This is the discovery of AI itself. Now, this is a bit different from the OpenAI fight that we've covered on previous episodes of Case of the Week and that really kicked off our generative AI case law in 2025.

Why This Isn't the OpenAI Fight

The OpenAI cases are a training-time theory. OpenAI allegedly copied millions of articles into the data used to train its models, baked that content into the model's own weights, and the trained model can now reproduce it whenever any user, anywhere, happens to ask the right question. That is why a court ordered OpenAI to produce twenty million ChatGPT logs sampled out of tens of billions it retains, a statistical sample across millions of unrelated users, aimed at OpenAI's fair use defense.

Perplexity's RAG architecture is a retrieval-time theory. Nobody claims that Perplexity trained its model on Britannica's content. The claim is that Perplexity copies the content into its RAG index live, off the open web, and reproduces it whenever a user's question happens to retrieve it. There's no training corpus at issue, which is exactly why this discovery is scoped precisely to Britannica's own copyright registration dates, and not a random sample across millions of users, like in OpenAI. It's targeted, not statistical. And it is also why a RAG claim is proving easier to litigate than a training claim. You can point to the actual article, the index entry that copied it, and the output that reproduced it, a much more direct line than proving a work influenced a model's parameter somewhere inside an opaque training run.

The Facts

Let's talk about the facts of this particular case. Britannica and Merriam-Webster sued Perplexity for copyright infringement and trademark dilution, alleging that Perplexity's answer engine cannibalizes traffic to their sites with AI-generated summaries of their own proprietary content.

This is not Perplexity's first fight over this exact issue, and the order in this case walks through that history in real detail. Dow Jones and the New York Post sued Perplexity in

2024 over the same RAG architecture, in the same district, the Southern District of New York. In an August 2025 opinion denying Perplexity's motion to dismiss, United States District Judge Katherine Polk Failla described the RAG database itself as the mechanism by which the alleged infringement happens, copying protected works in as inputs and reproducing them as outputs, and that is the framework that United States Magistrate Judge Sarah Cave adopts directly here. In the Dow Jones Action, Perplexity had already produced two RAG snapshots, totaling more than 300 terabytes, and over 100 terabytes of UAL data, and after losing a motion to compel, seven more months of UAL data on top of that, with United States District Judge Katherine Failla calling the burden proportional to the needs of the case.

Once Britannica sued, Perplexity did not start by fighting. It produced the same RAG snapshots and made the same UAL data it had already turned over in the Dow Jones Action available to Britannica and Merriam-Webster. No dispute was needed for that baseline. The fight only started when Britannica came back on May 8, 2026, asking for ten more months, August 2025 through May 2026, what the order calls the Additional Data. And even then, Perplexity's first move was not a flat refusal. It proposed a cost-sharing arrangement. And when that did not resolve things, it argued that production would be unduly burdensome and disproportionate to the needs of the case.

One more thing before the holding, because it matters how I built this episode. Everything above comes from the order itself. But Perplexity is fighting versions of this same fight against several publishers at once, in the same district, and a live account of the hearing in this case, reported by Inner City Press, shows a parallel dispute happening at the same conference over whether Perplexity has to produce its source code in native format or a converted one. I'm citing that as reported hearing color, not part of the order, because it's not. You can read it yourself: Matthew Russell Lee's piece for Inner City Press, called "AI Beat: As Britannica Sues Perplexity," at matthewrussellleeicp.substack.com. It's paywalled past the opening section, but we'll put a link to that in case you've got a Substack subscription. Let's talk now about the holding here.

The Holding, in Three Parts

In the Court's own words, while the Additional Data was relevant, requiring Perplexity to produce all of it is not proportional to the needs of the case. So Perplexity would produce one additional RAG snapshot and host six more months of additional UAL data for inspection. That's the ruling in one sentence. Here's how the Court got there, under Rule 26(b)(1) and Rule 26(b)(2)(C).

First, on relevance. The relevant window for this kind of technical production tracks the plaintiffs' copyright registration dates, the same rule that District Judge Katherine Failla applied in the Dow Jones Action. Most of Britannica and Merriam-Webster's registrations became effective between April and December 2025, so that is where the window starts. Your technical discovery window in a case like this is not set by when

you filed suit. It is set by when your copyrights become effective, and the months around that. Because remember, this is a retrieval case, not a training case.

The second issue was proportionality. Britannica wanted ten months of data. The Court granted one additional RAG snapshot and six months of UAL data, not the full ten. Nobody got everything they asked for, and the Court did its own tailoring instead of picking a side wholesale.

The third issue was costs. Ordinarily, under the presumption from *Oppenheimer Fund v. Sanders*, the responding party bears its own production costs. Perplexity asked the Court to shift the cost to Britannica under Rule 26(c)(1)(B) and the *Zubulake* framework. And the Court agreed that there was good cause, but capped Perplexity's relief at $6,000 a month, exactly what Britannica had already offered, not the number that Perplexity asked for.

And here's why. Perplexity had told Britannica informally that hosting this data would run $12,000 to $14,000 a month. In its formal declarations filed with the Court, that number jumped to the hundreds of thousands. Now, given the volumes of data here, a single day of UAL data alone runs about 20 terabytes. That bigger number is not implausible on its face. Extracting and hosting that much data is a real undertaking, but nothing in the record explained why the estimate moved that much, and the Court called the gap “the orders-of-magnitude difference”, viewing the higher figure “with considerable skepticism.” With considerable skepticism, try to say that ten times fast.

The lesson here is not that Perplexity's number, the original quote of $12,000 to $14,000, in terms of hosting content, was wrong. It's that an unexplained jump in your own cost estimate reads as unreliable, whether or not it actually is. I think what happened here, likely, is that the cost of hosting the data was probably $12,000 to $14,000. The cost of getting the data into that hosting platform was not considered. And that's something we're going to talk about in the takeaways. All right, let's talk about those.

Takeaways and What's Next

First, the one for everyone, not just the AI copyright bar. This order is a working vocabulary lesson in how a retrieval-based AI system operates, and that vocabulary is worth having any time an AI tool shows up on the other side of a discovery dispute, regardless of what the case is actually about. You need to understand how these tools work, because they are creating evidence that you're going to want for your case.

If you are litigating against an AI company on a copyright theory specifically, the registration-date rule from the Dow Jones Action gives you your scoping argument before you ever file a motion to compel. Draft your request to that window, and expect a court to hold you to it if you overreach.

If you are defending an AI system and fighting proportionality on cost, the lesson from this case is not that your number has to stay fixed. Costs genuinely change as you learn what a production actually requires. The lesson is that you have to show your work when it does. Perplexity's estimate moved from $13,000 a month to hundreds of thousands without ever explaining what changed, and that silence is what cost it the bigger number, not the jump itself. Put the reason on the record, or the court will assume the worst explanation for you.

And here's the throughline that takes us back to *Cerebras* and *Schulte*. In *Cerebras*, the parties bought predictability by agreeing to a governance framework before anyone fought. In *Schulte*, nobody agreed to anything, and the requesting party lost trying to get the same transparency by motion. Here, nobody agreed to anything either, but this time it is the responding party absorbing the cost of not having settled these terms in advance. Three different situations, three different outcomes. And the common denominator is whether the parties planned the technical discovery terms before the fight, or left a judge to reinvent them mid-case.

This case, to me, the Perplexity cases and this information that's having to be produced, really reminds me of the OpenAI cases, because it's the first time we're seeing this kind of evidence have to be provided. And it's likely that we'll start to see more protocols like *Cerebras* in these kinds of situations, where the parties get out ahead of it and decide that they can share the real-time costs of what that information is and make sure the court has it.

There's a couple things I'm watching following up on this case. Perplexity is fighting versions of this same fight against multiple publishers at once. They're in the same district, the Southern District of New York, and the rulings are already citing each other back and forth, the Dow Jones Action informing Britannica, Britannica likely to inform whatever comes next. The discovery playbook for suing or defending a RAG-based AI company on copyright is being assembled in real time, and it will look different a year from now. If you want to keep following it, you can look up the Generative AI tag on Minerva26, that's where we track every one of these decisions as they come down.

If this episode was useful, please share it with someone in your world who's going to run into an AI company on the other side of a discovery dispute, because that day is coming for more of you than you think. Thanks so much for joining me, and I'll see you on our next episode.