EXHIBIT 124

(FILED UNDER SEAL)

1

IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA
SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., et al., : Civil Action No. 5:24-cv-03811-EKL-SVK

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

HIGHLY CONFIDENTIAL
ATTORNEYS' EYES ONLY

VIDEOCONFERENCED (HYBRID) AND VIDEO-RECORDED

DEPOSITION OF BENJAMIN MANN

PURSUANT TO FEDERAL RULE OF CIVIL PROCEDURE 30(b)(1)

San Francisco, California

Thursday, November 20, 2025

4:06 p.m. Pacific Standard Time

VOLUME I

Job No. 609867

Pages: 1- 145

Reported By: April L. Crites, RMR, CRR, CSR, CCR

California Certified Shorthand Reporter

2

DEPOSITION OF BENJAMIN MANN

HELD AT THE OFFICES OF:

WILMER CUTLER PICKERING HALE and DORR, LLP

50 California Street

Suite 3600

San Francisco, California 94111

(628) 235-1000

* * *

Pursuant to notice, and pursuant to the Federal Rules of Civil Procedure, before April L. Crites, RMR, CRR, CSR, CCR, and California Certified Shorthand Reporter No. 14493.

* * *

3

I N D E X

EXAMINATIONS CONDUCTED PAGE

BENJAMIN MANN

Cross-Examination by Mr. Miller..... 10

* * *

EXHIBITS

MARKED	DESCRIPTION	PAGE
Exhibit 87	Execution: Copyfairy as a sampling stage, Database, Bates stamped Anthropic_0000429077 through Anthropic_0000429094	106
Exhibit 183	[redacted] Launch signoff (Product, Models, T&S, Comms, Bates stamped Anthropic_12437 through Anthropic_0000012449	124
Exhibit 200	Benjamin Mann's LinkedIn web page	12
Exhibit 201	ANTHROP\C org chart, Bates stamped Anthropic_0000000168	23
Exhibit 202	arXiv:2005.14165v4 [cs.CL] 22Jul2020, Language Models are Few-Shot Learners, OpenAI	26

4

Exhibit 203	January 5, 2021, e-mail from Notion to [redacted] Bates stamped Anthropic_0000458349 through Anthropic_0000458353	47
Exhibit 204A	Excel spreadsheet: The Below are TBD Pending Pretraining Team Discussions Post Scan Launch, Bates stamped Anthropic_0000458418	65
Exhibit 204B	Excel spreadsheet, "Note: Ben has a similar sheet here; dataset name, Bates stamped Anthropic_0000510804	70
Exhibit 205	Slack communication, dated June 4, 2021, Bates stamped Anthropic_0000014096 through Anthropic_0000014106	88
Exhibit 206	January 4, 2024, e-mail from Sam Martin to [redacted] Subject: Dataset for copyright filter checks, with attachment, Bates stamped Anthropic_0000218330	108
Exhibit 207	January 4/5, 2024, e-mail from [redacted] to Sam Martin and Ben Mann, Subject: Re: Dataset for copyright filter checks, Bates stamped Anthropic_0000218341 through Anthropic_0000218342	113

9

THE VIDEOGRAPHER: Here begins media number 1 in the videotaped deposition of Benjamin Mann, individual, in the matter of Concord Music Group v. Anthropic, in the United States District Court, Northern District of California, San Jose Division, case number 24-CV-03811-EKL.

Today's date is November 20, 2025. The time on the video monitor is 16:06.

The videographer today is Philip Astor, representing Planet Depos.

This video deposition is taking place at Wilmer Cutler Pickering Hale & Dorr, LLP, San Francisco, California.

Will counsel please voice-identify themselves and state whom they represent?

MR. MILLER: Corey Miller from Oppenheim + Zebrak, LLP, on behalf of the plaintiff music -- plaintiff music publishers. And also here with me from Oppenheim + Zebrak is Michelle Gomez-Reichman.

MS. MEHTA: Good afternoon. Sonal Mehta from WilmerHale on behalf of Anthropic and the witness. With me today is my colleague Taylor Gooch. We also have on the Zoom our colleague Jared Grubow. We have with us in-house counsel

10

from Anthropic in the room, Devon Hanley Cook, and we also have Sara Sampoli, our co-counsel, from Latham & Watkins.

THE VIDEOGRAPHER: The court reporter today is April Crites, representing Planet Depos.

The witness will now be sworn.

BENJAMIN MANN

of lawful age, Witness herein, having been first duly cautioned and sworn, as hereinafter certified, was examined and said as follows:

CROSS-EXAMINATION

BY MR. MILLER:

Q Good afternoon, Mr. Mann.

Do you understand that you are under oath?

A Yes, I do.

Q And that's the same oath to tell the truth you would take if we were in a courtroom before a judge and jury?

A Yes, I understand that.

Q Is there any reason you cannot testify today in accordance with that oath?

A There's no such reason.

Q Okay. Have you been deposed before, Mr. Mann?

A Yes, I have.

11

Q How many times?

A Is the question how many cases or how many instances of depositions?

Q How many instances of depositions.

A And a deposition on one day is different from a deposition on -- on another day, or is that counted as one?

Q You can count those as one.

A Okay. I think it's twice before.

Q Twice before.

And what were the cases in which you were deposed?

A I believe it was both for the same case, but one in my personal capacity as a former employee of Anthropic -- sorry -- of OpenAI, and then again on behalf of Anthropic, both in my personal capacity, and as a 30(b)(6) witness. And in both cases, the -- it was -- Bartz was the case.

Q That's the Bartz v. Anthropic case?

A It must have involved OpenAI, as well, so maybe I'm mistaken and it was actually multiple cases. But I believe it was the same plaintiff in both cases.

Q Other than the -- the Bartz litigation and

12

the OpenAI litigation, have you been deposed in any other cases?

MS. MEHTA: I assume you're excluding this morning's deposition?

MR. MILLER: Yes, not counting this morning.

BY MR. MILLER:

Q Have you been deposed in any other cases?

A And that doesn't include any potential other ongoing lawsuits that I haven't sat deposition for yet?

Q No. If you haven't sat for deposition, I'm not -- yet, I'm not asking you about those.

A Okay. Yeah, then that would be it.

Q Okay. Have you ever testified in court?

A No.

Q Was the testimony you gave in your previous depositions truthful and accurate?

A To the best of my knowledge, yes.

MR. MILLER: Tab 28.

(Plaintiffs' Exhibit 200, Benjamin Mann's LinkedIn web page, was marked and presented for purposes of identification.)

MR. MILLER: What number are we up to?

BY MR. MILLER:

73

Q Do you see in row 4, column B, it says, 2 point -- [redacted], correct?
A Specifically, in B4 it says [redacted]
Q Okay. And the header for column B says billion tokens, correct?
A That's correct.
Q So the [redacted] is that correct?
MS. MEHTA: Objection to the form.
THE WITNESS: It appears that's what the author of this spreadsheet had in mind when they wrote that, yes.
BY MR. MILLER:
Q And you see in row 7, there's a reference to LibGen PDFs?
A Yes.
Q Does that suggest that LibGen PDFs are were used in a -- in any training mix for this model?
MS. MEHTA: Objection. Calls for speculation. Lacks foundation. Asked and answered.
THE WITNESS: That might suggest this. But if I recall, I know that I've thought about it a bit more, [redacted]

74

[redacted]
BY MR. MILLER:
Q When did that occur?
MS. MEHTA: Form.
THE WITNESS: It would have been in the same year as the spreadsheet was created, I would feel fairly confident saying.
BY MR. MILLER:
Q [redacted]?
MS. MEHTA: I'm going to object to the form and also caution the witness you can answer that to the extent you have an answer independent of any advice of counsel, but please don't share any privileged communications with counsel.
THE WITNESS: I don't think I can answer that question without divulging attorney-client privilege.
BY MR. MILLER:
Q Okay. Are you familiar with the term synthetic data?
A Yes, I am.
Q What do you understand synthetic data to be?

75

MS. MEHTA: Objection. Asked and answered.
THE WITNESS: In general, synthetic data refers to data that wasn't produced, quote/unquote, naturally, which means that, for example, a person created it or it was the byproduct of something that happened in the real world; instead, it is generated by a software program. It can be derived from natural data in some cases, but typically, it is, yeah, generated by a software artifact.
BY MR. MILLER:
Q And you used the term "natural data." Is natural data the opposite of synthetic data?
MS. MEHTA: Objection to the form.
THE WITNESS: It's not a term that's typically used, but I think that would be fair to use in this case.
BY MR. MILLER:
Q Is there a term that's more typically used?
MS. MEHTA: Form.
THE WITNESS: I don't think so.
BY MR. MILLER:
Q Does Anthropic use synthetic data to train

76

its commercially released models?
MS. MEHTA: Objection to the form.
THE WITNESS: I mean, in some ways you could view all of reinforcement learning as a form of synthetic data, in that there is a series of prompts that have been generated, maybe from something in the model pipeline, and then from there, all of the data is generated by the model. There is no humans in the loop anywhere. So I would say yes.
BY MR. MILLER:
Q Yeah. Does Anthropic use synthetic data in pre-training for any of its commercially released models?
A I think the line gets a little bit fuzzier [redacted]
Q Who at Anthropic selects the data to be used in a particular data mix for a particular training run for a model that Anthropic's developing?
MS. MEHTA: Objection to the form.

77

THE WITNESS: Are you asking about for current model runs or what's the time period we're discussing?
BY MR. MILLER:
Q Let's start with currently.
MS. MEHTA: Same objection.
THE WITNESS: I don't know specifically who the lead is on that, but I'm sure that there is a lead that -- I think there are people who we call run captains who are responsible for this kind of thing.
BY MR. MILLER:
Q Are you familiar with data filtering?
A Yes.
Q What is data filtering?
A Data filtering is the process of taking some existing dataset and removing either whole or partial documents from that dataset before other use of that data.
Q Is data filtering different from data cleaning?
MS. MEHTA: Objection to the form.
THE WITNESS: It can be, yeah, because -- well, yeah, it can be.
BY MR. MILLER:

78

Q What is the difference between data filtering and data cleaning?
MS. MEHTA: Objection to the form.
THE WITNESS: As I understand these terms, there are at least -- at least one big difference comes to mind, which is that filtering tends to be only removal of content or removal of -- and that content could be whole documents or partial documents.
In data cleaning, it may involve filtering, but it may also involve manipulation or augmentation of data that makes it easier for the model to learn. I think that's a pretty good definition.
Q How does data cleaning involve the manipulation or augmentation of data?
MS. MEHTA: Objection to the form.
THE WITNESS: So to take a fairly trivial example, there are all these Kaggle competitions to train models with sort of old-school techniques, like XGBoost.
And in these cases, the datasets are often tabular. One way of augmenting a dataset or cleaning it would be to combine multiple columns into one column so that the modeling technique had

79

an easier time. And that kind of combination might happen through all sorts of different techniques. It could be concatenation of strings. It could be summing columns or dividing them, if they're numeric.
That's -- that's one example.
BY MR. MILLER:
Q Does -- when Anthropic is -- withdrawn.
Does Anthropic both filter and clean data that it uses for training?
MS. MEHTA: Objection to the form.
THE WITNESS: I would say yes.
BY MR. MILLER:
Q Does one of those processes occur before the other one?
MS. MEHTA: Objection to the form.
THE WITNESS: Yes. Typically, we do those two things in at least two different steps, possibly more. I think our data processing pipelines have grown extremely complicated since I last worked on them. But I am roughly familiar with what they were at the time that I worked on them.
BY MR. MILLER:
Q Okay. And as far as you're familiar with

80

them, what is the sequence of steps?
MS. MEHTA: Objection to the form.
THE WITNESS: My recollection is that first we deduplicate the data, and that's a form of filtering, and then we do a quality filter step, which you could consider data cleaning.
Actually, I guess before even the deduplication step, we [redacted]
We also have [redacted] in which we might manipulate those [redacted] in terms of the [redacted] that they're at, throwing away what seem to be low-quality [redacted] based on some kind of summary statistics that we might use.
Many, many different kinds of steps that could be involved, depending on the dataset.
BY MR. MILLER:
Q Does filtering typically involve the removal of -- of data that Anthropic wants to exclude from the -- from the training data?
MS. MEHTA: Objection to the form.
THE WITNESS: So typically when we do

81

things like deduplication --
MR. MILLER: Mm-hmm.
THE WITNESS: -- which is a form of filtering, yes, we don't want so many copies of that data to appear in the final outputs, but that process is agnostic to the value of a single copy or a small number of copies of that data. So I would say there are some filtering techniques that we use that don't bear on whether, quote/unquote, Anthropic wants that data.
Q Are there some filtering techniques Anthropic uses that do bear on whether Anthropic wants that data?
MS. MEHTA: Objection to the form.
THE WITNESS: So you can think of quality filtering as a form of data filtering. And for that, again, we usually [redacted]. So the way it works is we get [redacted]. But back in the day, we used the [redacted].
And then once we had that score between [redacted] we would decide [redacted] whether to keep or not keep that

82

document.
And we would bias towards keeping more of the documents that [redacted] high quality and fewer of the documents [redacted] were low quality, but there was always a small chance that we would keep documents that we thought were low quality and throw out documents we thought were high quality, because we had uncertainty -- or, rather, I would say [redacted]
And if we were to throw out every document that the [redacted] we would likely significantly skew our data distribution in a way that damaged model performance. And we empirically validated that hypothesis, as well.
BY MR. MILLER:
Q Are you familiar with something called an extractor tool?
MS. MEHTA: Objection to the form.
THE WITNESS: There are many tools that we have used over the years that I would classify as extractors.
BY MR. MILLER:
Q What is an extractor?

83

A So, typically, an extractor refers to a piece of software that goes from a more voluminous or uncompressed or otherwise raw form of some data down to a more training-ready form of that data.
So, for example, we would have PDF extractors that took PDF data and turned it into a plain-text representation.
BY MR. MILLER:
Q Has Anthropic created its own extractors, or does it use third-party-created extractor tools?
MS. MEHTA: Objection to the form. Foundation. Speculation.
THE WITNESS: So we have certainly created a lot of extraction software, but it is often built on top of open source technology. But even then, many of these open source libraries contain many different knobs and dials that can be used to tune an extraction process.
And at that point, it's up to us -- empiricism, intuition, spot-checking of the data -- to decide what kind of extraction we should be doing. And sometimes we'll even use multiple different extraction techniques for the same type of content to provide variety.

84

BY MR. MILLER:
Q You mentioned using extractors on PDFs. Has Anthropic also used extractors on HTML documents?
MS. MEHTA: Objection to the form. Foundation.
THE WITNESS: Yes, I think it would be fair to say that.
BY MR. MILLER:
Q For what purpose has Anthropic used extractors on HTML documents?
MS. MEHTA: Objection to the form.
THE WITNESS: Well, a natural HTML document as would you find it on the web contains a large amount of boilerplate -- for example, ad-tracking libraries that are in JavaScript embedded in the page, different kinds of invisible content that's been rendered to a user, site navigation, all kinds of content that won't really help the model learn.
And as we've discussed before, we have pretty strong evidence that duplicated data in model training causes the model to be less capable after training.
And so for all these reasons, it's very

85

valuable for us to get rid of as much of that extraneous content as possible before training the model.
BY MR. MILLER:
Q Does the extraneous content that Anthropic seeks to get rid of also include copyright notices on a web page?
MS. MEHTA: Objection to the form. Calls for speculation.
THE WITNESS: It depends strongly on the web page itself and how that web page is structured. The way most of these content extractors work on HTML is they look for things like

tags inside the main body of the HTML

document, and so if a copyright notice were in the body in a

tag, then I find it highly unlikely

that we would remove it.
BY MR. MILLER:
Q Do you know whether -- withdrawn.
What is a

tag?

A In HTML -- HTML stands for hypertext markup language.
Q Mm-hmm.
A And what that's referring to is there's this structure, language, that contains many

86

different kinds of tags. I mentioned before there's the body tag, there's a

tag

tags

generally are meant to semantically represent paragraphs of text, but they are also used just for delineating one piece of content from another in a web page.
There is a very, very wide distribution of how these kinds of tags are used on the Internet, so it's pretty hard to generalize. But that's sort of what the specification intended

tags

to be used for.
MR. MILLER: Tab 2? Do you remember what number this was marked?
MS. GOMEZ-REICHMAN: 187.
BY MR. MILLER:
Q In the Zoom, you should see what we marked earlier this morning as Plaintiffs' Exhibit 187, which has the Bates number Anthropic_349548.
Do you remember this document?
A Yes, it looks familiar.
Q And I want to you go all the way to the end, very bottom of the last page.
Do you see at the bottom of the page, there's a bunch of links?
A Yes.

87

Q And then, are those links the kinds of things Anthropic would try to remove during the filtering process?
MS. MEHTA: Objection to the form. Calls for speculation.
THE WITNESS: Well, this is generally site navigation, so that is the kind of thing that would typically be removed by our HTML extraction libraries.
BY MR. MILLER:
Q And then underneath the site navigation links, in the bottom left corner, do you see it says Copyright 2025 Anthropic PBC?
A Yes, I see that.
Q Is that sort of copyright notice something Anthropic would typically try to remove using an extractor in the filtering process?
MS. MEHTA: Objection to the form.
THE WITNESS: It's not something that we would specifically try to remove. But these extractors would potentially do it. I -- without trying it on this page, I wouldn't be able to tell you what the effect would be.
BY MR. MILLER:
Q Are you familiar with the Newspaper

88

extractor?
A Yes, I've used it.
Q What is the Newspaper extractor?
A Newspaper is an HTML-to-plain-text converting library. I believe it has a wide variety of other potential uses, but that's how we use it.
Q Okay. Are you familiar with the Readability extractor?
A Yeah.
Q And what is the Readability extractor?
A So there's at least one library called Readability that is maintained by Mozilla, and that also takes web pages and converts them into a simplified HTML or a plain-text representation.
Q Are you familiar with a tool called jusText?
A Yes.
Q And What is jusText?
A It's in the same class of software as Newspaper and Readability.
Q Okay.
MR. MILLER: Okay. Tab 37.
(Plaintiffs' Exhibit 205, Slack communication, dated June 4, 2021, Bates stamped

97

[redacted]
MS. MEHTA: Objection to the form.
THE WITNESS: [redacted]
MR. MILLER: Mm-hmm.
THE WITNESS: [redacted]

98

[redacted]
BY MR. MILLER:
Q Thank you. That's helpful.
[redacted]
MS. MEHTA: Objection to the form. Misstates the testimony.
THE WITNESS: [redacted]
BY MR. MILLER:
Q Mm-hmm.
[redacted]

99

[redacted]
MS. MEHTA: Objection to the form.
THE WITNESS: [redacted]
BY MR. MILLER:
[redacted]
MS. MEHTA: Objection to the form.
THE WITNESS: I would imagine it does, yes.
BY MR. MILLER:
Q And if the dataset -- withdrawn.
[redacted]
MS. MEHTA: Objection to the form. Vague. Calls for speculation.

100

THE WITNESS: I'd be happy to answer any questions about specific datasets, but it's hard to generalize since there are so many datasets, [redacted]
BY MR. MILLER:
Q Let me ask a different question.
[redacted]
MS. MEHTA: Objection to the form. Vague. Compound. Calls for speculation.
THE WITNESS: [redacted]
BY MR. MILLER:
Q You talked several times today about how Anthropic's goal is to build models that are helpful, harmless, and honest. There's a tension between helpfulness and -- excuse me, did I say harmful? I meant harmless. No, I had it right. It's been a long day.

CERTIFICATE OF REPORTER

I, April L. Crites, RPR, RMR, CRR, CSR, the officer before whom the foregoing deposition testimony was taken, do hereby certify:

That before being examined, BENJAMIN MANN was by me duly sworn to testify to the truth, the whole truth, and nothing but the truth;

That the foregoing transcript is a true and correct record of the testimony given;

That said testimony was taken by me stenographically and thereafter reduced to typewriting under my direction;

That reading and signing was requested;

And that I am neither counsel for, related to, nor employed by any of the parties to this case and have no interest, financial or otherwise, in its final disposition.

IN WITNESS WHEREOF, I have here unto subscribed my name this 25th of November, 2025.

My commission expires May 31, 2026:

April L. Crites, RPR, RMR, CRR, CSR
CA Certified Shorthand Reporter No. 14493