Exhibit 5



Planet Depos®
We Make It Happen™

HIGHLY CONFIDENTIAL
ATTORNEYS' EYES ONLY

Transcript of Benjamin Mann

Date: November 20, 2025
Case: Concord Music Group -v- Anthropic

Planet Depos
Phone: 888.433.3767
Email: transcripts@planetdepos.com
www.planetdepos.com

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
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

Exhibit 208	January 8, 2024, e-mail from Evan Frondorf, Subject: Upcoming ability to intervene on prompts, Bates stamped Anthropic_0000218344	114
Exhibit 209	DRI: Miranda Zhang, last substantial edit June 27, 2023, Bates stamped Anthropic_0000479446 through Anthropic_0000479474	136

* * *

A P P E A R A N C E S

ON BEHALF OF THE PLAINTIFFS, THE MUSIC PUBLISHERS:

Corey Miller, Esq.
Michelle Gomez-Reichman, Esq.
Nicholas C. Hailey, Esq.
OPPENHEIM ZEBRAK LLP
4530 Wisconsin Avenue NW
Fifth Floor
Washington, D.C. 20016
202-596-5043
corey@oandzlaw.com
mgomez-reichman@oandzlaw.com
nick@oandzlaw.com

ON BEHALF OF THE DEFENDANT, ANTHROPIC, PBC:

Sonal N. Mehta, Esq.
WILMER CUTLER PICKERING HALE & DORR, LLP
2600 El Camino Real
Suite 400
Palo Alto, California 94306
650-858-6000
sonal.mehta@wilmerhale.com

- AND -

A P P E A R A N C E S C O N T I N U E D

ON BEHALF OF ANTHROPIC, PBC (CONTINUED):

Joseph Taylor Gooch, Esq.
WILMER CUTLER PICKERING HALE & DORR, LLP
50 California Street
Suite 3600
San Francisco, California 94111
628-235-1002
taylor.gooch@wilmerhale.com

- AND -

Jared V. Grubow, Esq. (Via Zoom)
WILMER CUTLER PICKERING HALEK& DORR, LLP
7 World Trade Center
250 Greenwich Street
New York, New York 10007
212-295-6474
jared.grubow@wilmerhale.com

- AND -

A P P E A R A N C E S C O N T I N U E D

ON BEHALF OF ANTHROPIC, PBC (CONTINUED):

Sara E. Sampoli, Esq.
LATHAM & WATKINS LLP
555 Eleventh Street NW
Suite 1000
Washington, D.C. 20004
202-637-2200
saa.sampoli@lw.com

ALSO PRESENT:

Devon Hanley Cook, Esq.
Anthropic PBC

Philip Astor, Videographer
Planet Depos

* * *

Pages 37-56

37

MS. MEHTA: Corey, I -- I'm not interrupting the deposition. I'm asking you to be respectful of the witness' time and not to just repeat questions that you've already asked twice and you have on the transcript.

Let's move on.

BY MR. MILLER:

Q Have you personally reviewed the contents of the various datasets that Anthropic has acquired from Common Crawl?

MS. MEHTA: Objection to the form.

THE WITNESS: In the process of doing data engineering, it is common to look at a sample of datasets and to do things like generate summary statistics on those datasets to try to understand the broad shape of what might be contained. But I would say it is impossible for any human to review such a large amount of content in -- in totalé. So, yes, I have done some -- some steps of reviewing.

BY MR. MILLER:

Q And have you personally reviewed any metadata for the datasets that Anthropic has acquired from Common Crawl?

MS. MEHTA: Objection to the form.

38

THE WITNESS: Yes. Common Crawl, as an organization, produces metadata that they make available on their website. They also provide metadata that they ship along with the files that they have in their archives, and I have reviewed many samples of that metadata as well as, again, I've generated summary statistics.

BY MR. MILLER:

Q Have you communicated with the individuals at the Common Crawl organization?

A Yes, I have.

Q With whom at Common Crawl did you communicate?

A One of the people was named Sebastian. I don't know if he's still there. And I think there was at least one other person, but I don't recall her name.

Q And what was the subject matter of your communications with Common Crawl?

A I don't recall the details.

Q Do you recall when those communications happened?

A I don't recall.

Q And you don't remember why you were speaking with Common Crawl?

39

A No.

Q Do you know whether the Common Crawl datasets that Anthropic has obtained contain copyrighted content?

MS. MEHTA: Objection to the form.

THE WITNESS: Do I know whether they contain copyrighted content?

BY MR. MILLER:

Q Yes.

A I'm not a lawyer, so what -- what content, specifically, is copyrighted or not, I'm not an expert. I believe it does contain some copyrighted content somewhere in the vast balk of Common Crawl.

Q When did Anthropic begin obtaining datasets from Common Crawl?

MS. MEHTA: Objection to the form.

THE WITNESS: That would have been very soon after we started working on the company that eventually became Anthropic.

BY MR. MILLER:

Q And do you recall whose idea it was to obtain data from Common Crawl?

A I mean, I -- it's not really a thing that needs ideation so I kind of would argue with the

40

form of the question in that it's nobody's idea to, like, breathe, because it's something that, of course, everybody would do. It's just obvious to anybody who works on training large language models that Common Crawl is, like, the first place that anybody looks.

Q Do you recall your testimony this morning [redacted]

MS. MEHTA: Objection to the form. Misstates the testimony.

THE WITNESS: I believe my testimony was [redacted]

BY MR. MILLER:

Q Do you know whether prior authorization was obtained from anyone before Anthropic first obtained datasets from Common Crawl?

MS. MEHTA: Objection to the form.

THE WITNESS: I don't specifically recall.

BY MR. MILLER:

Q Do you know whether Anthropic sought

41

authorization from any copyright-holders before obtaining datasets from Common Crawl?
MS. MEHTA: Objection to the form.
THE WITNESS: I don't specifically recall.
BY MR. MILLER:
Q Has Anthropic used data obtained from Common Crawl to train its commercially released models?
MS. MEHTA: Form.
THE WITNESS: Yes. Every commercial Anthropic model uses data from Common Crawl.
BY MR. MILLER:
Q During what stages -- withdrawn.
Is Anthropic still using data obtained from Common Crawl to train commercially released models?
MS. MEHTA: Objection to the form. Asked and answered.
THE WITNESS: Yes, Anthropic is still using data derived from Common Crawl's data to train models.
BY MR. MILLER:
Q Are you familiar with a dataset known as the Pile?
A Yes, I am.

42

Q What is the Pile?
A The Pile is a dataset that at the time was considered fairly large that was released open source on GitHub by an organization called Eleuther, and at the time was viewed as a pretty standard and foundational dataset for large language model training.
Q Have you reviewed any of the contents of the Pile?
A Yes, I have.
Q Have you reviewed any of the metadata for the Pile?
A Well, each of the sections of the Pile has quite different metadata. And in addition to that, there's the white paper associated with the release. And so I've read the paper, and I have also reviewed some of the subsets of the Piles' metadata, but I would say I have not comprehensively reviewed every document.
Q You referred a couple of times to subsets of the Pile.
What are the subsets of the Pile?
A So within the Pile, there are a number of subdatasets, and that's what I was referring to by "subset." Some examples are Common Crawl, GitHub.

43

Things like that.
Q Do you know what any other subsets contained in the Pile are?
A Yes, I know a few others.
Q What others are there?
A Off the top of my head -- and I could easily tell you specifically if you want to show me a document -- but they include YouTube subtitles, United States patent documents, WebM -- not WebMD, but PubMed, abstracts, I believe, there are certainly a couple of others that I don't remember.
Q I think earlier you mentioned the name of the individual who created the Pile.
MS. MEHTA: Objection to the form.
BY MR. MILLER:
Q Is that right?
A I mentioned the organization that created the Pile, which was called EleutherAI.
Q Okay. The organization.
Have you communicated with anyone at EleutherAI?
MS. MEHTA: Objection to the form.
THE WITNESS: I have communicated with some of its former members, though I don't recall

44

if they were members of the organization at the time that I discussed -- that I talked to them.
BY MR. MILLER:
Q Do you recall what you talked to them about?
A It varied.
Q What were some of the subjects you discussed with them?
A In one case, I was trying to hire the person, but I failed. And in another case, I was trying to convince them that open-source models were not good for the world, and to the extent that they wanted to continue working on that, they should do more to bake safety interventions into their techniques -- their training techniques. There were probably others, but those are the two that I remember.
Q Do you know how the Pile was created?
MS. MEHTA: Objection to the form.
THE WITNESS: So the Pile, as I mentioned, is made up of many sub-datasets, and each of those sub-datasets was constructed in a different manner. And the code for constructing them, some of it is on GitHub in their open source repository, and for others, the code isn't fully

45

indicative or explanatory for how those datasets were created.
So for the parts that are represented by the open source repositories and by the white paper that accompanied the data release, I am aware of the construction methods, but for those parts that were omitted, I am not aware.
BY MR. MILLER:
Q Do you know whether the Pile contains copyrighted content?
MS. MEHTA: Objection to the form.
THE WITNESS: I don't know for sure whether it contains copyrighted content. I guess I would be surprised if it didn't contain some copyrighted content given the large amount of Common Crawl that is in that dataset. But, again, I'm not a lawyer, so I don't know for sure.
BY MR. MILLER:
Q Do you know whether the content in the Pile was all legally obtained?
MS. MEHTA: Objection to the form. Calls for speculation. Also calls for a legal conclusion.
And -- and, Counsel, you know you're not allowed to -- sorry. You know you're not allowed

46

to ask about certain topics, so I don't -- including methods of acquisition pursuant to the orders from the Court.
Did you want to rephrase that question?
MR. MILLER: I'm not asking about how Anthropic obtained the information. I'm talking -- asking about how the dataset itself was assembled.
MS. MEHTA: By the people that assembled the Pile?
MR. MILLER: Yes.
MS. MEHTA: Okay. On that basis, I will allow him answer that, if he knows. But I still object to the form and calling for speculation and legal conclusion.
THE WITNESS: Can you ask your question again?
BY MR. MILLER:
Q Do you know whether the content contained in the Pile was all legally obtained by the -- the pile organization that assembled the dataset?
MS. MEHTA: Form. Speculation. Legal conclusion.
THE WITNESS: I don't know.
BY MR. MILLER:

47

Q Has Anthropic obtained copies of the Pile dataset?
MS. MEHTA: Objection to the form.
THE WITNESS: Yes, we have downloaded the Pile at a -- on at least one occasion.
BY MR. MILLER:
Q Did Anthropic download all of the constituent sub-datasets within the Pile or only some of them?
MS. MEHTA: Form.
THE WITNESS: So there's one complexity here, which is that there was a pre-release version of the Pile, and then a release version of the Pile. And I believe we downloaded both of those. So there may have been other things that I'm not aware of that might be included in your question, but the -- the main dataset that's referenced by the paper, yes, we believe we have all of it.
MR. MILLER: Tab 42, please.
(Plaintiffs' Exhibit 203, January 5, 2021, e-mail from Notion to [redacted] Bates stamped Anthropic_0000458349 through Anthropic_0000458353, was marked and presented for

48

purposes of identification.)
BY MR. MILLER:
Q In the Zoom, you should see what has been marked as Plaintiffs' Exhibit 203, which bears the Bates number Anthropic_458349.
You have -- you have this document in front of you?
A Yes, I believe I have that document.
Q Start at the top.
Do you recognize this document?
A I'm reviewing it.
Yes. So this is a very distorted version of an automated e-mail from Notion that indicates some of the edits that me and Sam McCandlish made to a document that we were working on together called The Science of Data Quality.
Q And at the top, it's dated January 5, 2021, correct?
A That's right.
Q Was that before Anthropic was officially founded?
MS. MEHTA: Form.
THE WITNESS: That was before our incorporation.
BY MR. MILLER:

49

Q Okay. Can you turn to the bottom of the second page, please. I want to direct your attention to the third-to-last paragraph.
Do you see that the document states that, The Pile contains lots of great data sources for us to get started with?
A I see that.
Q Do you know whose words those were?
A They're either my words or Sam McCandlish's words.
Q Would you agree that in 2021, the Pile contained lots of great data sources for what became Anthropic to get started with?
MS. MEHTA: Objection to the form.
THE WITNESS: I think that's a totally reasonable thing to say, yeah.
BY MR. MILLER:
Q Has -- has Anthropic used the Pile or data from -- included in the Pile dataset to train its commercially released AI models?
MS. MEHTA: Form.
THE WITNESS: Yes, we have used subsets of the Pile to train commercially released models.
BY MR. MILLER:
Q Has Anthropic used subsets of the Pile to

50

train all of its commercially released models?
MS. MEHTA: Objection to the form.
THE WITNESS: I don't know about all.
BY MR. MILLER:
Q Does Anthropic still use the Pile to -- or its -- withdrawn.
Does Anthropic still use data from the Pile datasets to train its commercially released models?
MS. MEHTA: Objection to the form.
THE WITNESS: I don't know.
BY MR. MILLER:
Q A little while ago we discussed OpenWebText2.
Do you recall that?
A Yes.
Q Have you personally reviewed any of the subsets of the OpenWebText2 dataset?
A You're referring to the OpenWebText2 dataset that was referenced in the GPT-3 paper?
Q Well, let's take a step back.
There was an OpenWebText -- is there more than one OpenWebText2 dataset?
A Oh, sorry, I had an internal confusion.
So there's WebText2 --

51

Q Mm-hmm.
A -- which was referenced in the GPT-3 paper.
Q Mm-hmm.
A And then there's OpenWebText, but I don't believe there's OpenWebText2, but maybe there is.
Q Okay. It's --
A I don't recall the details.
Q It's possible the confusion is on our -- on my end. So I apologize.
So there's -- are you familiar with a dataset called OpenWebText?
A Yes, I have heard of a dataset that's named something like that.
Q Is there any relationship between the Open WebText dataset and the WebText or WebText2 datasets?
A So when GPT-2 came out and they talked about WebText, there were a number of groups that wanted to reproduce the results in GPT-2. So those groups banded together to produce what was called the OpenWebText. I don't know whether they made an OpenWebText2 or not, but they reproduced the methods described in the GPT-2 and possibly GPT-3 papers to create that dataset.

52

Q Has Anthropic obtained the OpenWebText dataset?
A I don't know, but I'm sure that there are documents furnished for this case that I could use to verify that if you wanted to show them to me.
Q Has Anthropic obtained the WebText2 dataset?
A That would only be possible by agreement with OpenAI, which we certainly don't have.
Q Are you familiar with a YouTube subtitles dataset?
A Yes.
Q What is the YouTube subtitles dataset?
MS. MEHTA: Form.
THE WITNESS: So there are two different YouTube subtitles datasets that I'm familiar with. One is the subset of the Pile that is YouTube site subtitles, and then another one is one that Anthropic acquired. And I believe these datasets contain scraped YouTube subtitles.
BY MR. MILLER:
Q Okay. Is there overlap between the subset of the Pile that contains YouTube subtitles and the separate YouTube subtitles dataset?
MS. MEHTA: Objection to the form.

53

THE WITNESS: There's probably overlap, but I don't recall specifically investigating that.
BY MR. MILLER:
Q So focusing on the -- the separate YouTube subtitles contents, have you reviewed that dataset?
A I'm sure I have spot-checked it at some point in the past.
Q And have you reviewed any of the metadata for that separate YouTube subtitle dataset?
A I probably have spot-checked it here or there.
Q Do you know who created the YouTube Subtitles dataset?
MS. MEHTA: Objection to the form.
THE WITNESS: I don't know.
BY MR. MILLER:
Q Do you know how the YouTube Subtitles dataset was created?
MS. MEHTA: Same objection.
THE WITNESS: I don't know.
BY MR. MILLER:
Q All right. Do you know whether the data in the YouTube Subtitles dataset was all legally

54

obtained?
MS. MEHTA: Objection to the form. Calls for speculation. Lacks foundation. Also calls for a legal conclusion.
THE WITNESS: I don't know.
BY MR. MILLER:
Q All right. Does the YouTube Subtitles content contain copyrighted content?
MS. MEHTA: Same objections.
THE WITNESS: I don't know.
BY MR. MILLER:
Q Are you aware that YouTube contains a lot of music videos?
MS. MEHTA: Objection to the form. Assumes facts not in evidence.
THE WITNESS: I am aware that YouTube contains a lot of music videos.
BY MR. MILLER:
Q And subtitles from music video would contain the -- the lyrics to the songs that are included in those music videos, right?
MS. MEHTA: Objection to the form. Calls for speculation. Assumes facts not in evidence.
THE WITNESS: Well, one thing that I do know about YouTube subtitles is that sometimes the

55

subtitles are hand-generated by humans and then attached or uploaded with the video that they're associated with.
But other time, the subtitles that are attached to videos are machine-generated. And certainly early in the history of YouTube's machine-generated subtitles, they often didn't bear a close resemblance to the words that the humans would have been speaking or singing in those videos.
So without looking at specific data in those files, I wouldn't be able to say for sure whether there were, for example, song lyrics in the YouTube subtitles that we obtained in those datasets.
BY MR. MILLER:
Q Anthropic has obtained a copy of the YouTube Subtitles dataset; is that correct?
MS. MEHTA: Objection to the form.
THE WITNESS: We have a dataset called YouTube Subtitles -- or I believe it's YouTube SRT, or something like that. Yes.
BY MR. MILLER:
Q When did Anthropic obtain that dataset?
A I don't recall --

56

MS. MEHTA: Form.
THE WITNESS: -- specifically.
BY MR. MILLER:
[redacted]
MS. MEHTA: Objection to the form.
[redacted]
BY MR. MILLER:
Q Has Anthropic used the YouTube Subtitles dataset to train its commercially released models?
MS. MEHTA: Objection to the form.
THE WITNESS: We have used that dataset in at least one of our commercial models.
BY MR. MILLER:
Q Do you know which model that is?
A I know of at least one of the models, but I don't know all of them.
Q Which is the model you're aware of for which the training data included the YouTube Subtitles dataset?
MS. MEHTA: Form.
[redacted]

Pages 73-88

73

Q Do you see in row 4, column B, it says, 2 [redacted]

[redacted]

Q Okay. And the header for column B [redacted]

A That's correct.

[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 [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:

[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] to compute a [redacted]
And then once we had that [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

Pages 97-100

97

MS. MEHTA: Objection to the form.

MR. MILLER: Mm-hmm.

98

BY MR. MILLER:

Q Thank you. That's helpful.

MS. MEHTA: Objection to the form. Misstates the testimony.

THE WITNESS: No, there's many different

BY MR. MILLER:

Q Mm-hmm.

99

MS. MEHTA: Objection to the form.

MS. MEHTA: Objection to the form.

THE WITNESS: I would imagine it does, yes.

BY MR. MILLER:

Q And if the dataset -- withdrawn.

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,

BY MR. MILLER:

Q Let me ask a different question.

MS. MEHTA: Objection to the form. Vague. Compound. Calls for speculation.

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.



Planet Depos®
We Make It Happen™

HIGHLY CONFIDENTIAL
ATTORNEYS' EYES ONLY

Transcript of Benjamin Mann, Volume 2

Date: November 21, 2025
Case: Concord Music Group -v- Anthropic

Planet Depos
Phone: 888.433.3767
Email: transcripts@planetdepos.com
www.planetdepos.com

146

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
Friday, November 21, 2025
12:07 p.m. Pacific Standard Time

VOLUME II

Job No. 609873
Pages: 146 - 226
Reported By: April L. Crites, RMR, CRR, CSR, CCR
California Certified Shorthand Reporter

147

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.

* * *

148

I N D E X

EXAMINATIONS CONDUCTED PAGE
BENJAMIN MANN
Cross-Examination by Mr. Miller..... 153

* * *

EXHIBITS

MARKED DESCRIPTION PAGE
Exhibit 210 March 13, 2024, e-mail from Ben Mann to nick@anthropic.com, Subject: [Auto summary] Ben/Nicholas (2024-03-12 14:03 GMT-7)-Transcript, Bates stamped Anthropic_0000218323 156
Exhibit 211 March 13, 2024, e-mail from Ben Mann, Subject: [Auto-summary] Data bottleneck (2024-03-12 15:34 GMT-7)-Transcript, Bates stamped Anthropic_0000012615 165
Exhibit 212 Jack Clark on X post: Technological Optimism and Appropriate Fear, Children in the Dark 173
Exhibit 213 Time Magazine article: Pausing AI Developments Isn't Enough. We need to shut it all down 178

149

Exhibit 214 Prompt, Bates stamped Anthropic_0000431956 214
Exhibit 215 Prompt: Stevie Wonder lyrics, Bates stamped Anthropic_0000524147 216

* * *

Pages 174-177

174

And do you see the post states, Technical -- Technological Optimism and Appropriate Fear, an essay with -- an essay where I grapple with how I feel about the continued, steady march towards powerful AI systems?

MS. MEHTA: Objection to the form. Lacks foundation.

THE WITNESS: Yes, I see that text.

BY MR. MILLER:

Q And then the next sentence states, The world will bend around AI akin to how a black hole pulls and bends everything around itself?

MS. MEHTA: Objection to the form. It lacks foundation.

THE WITNESS: Yes, I see this.

BY MR. MILLER:

Q Do you agree with Mr. Clark that the world will bend around AI akin to how a black hole pulls and bends everything around itself?

MS. MEHTA: Form.

THE WITNESS: I certainly understand where Jack is coming from where he says that. As somebody who has studied physics, I don't know if Jack has, but the way black holes manipulate space and time is actually quite a unique feature of how

175

physics operates.

And I'm not sure I personally would have come up with that analogy, so I'm not sure I would say I agree. But certainly I understand the sentiment.

BY MR. MILLER:

Q What is -- you said you understand where Jack is coming from and you understand the sentiment.

What is the sentiment that's being expressed here that you understand from this post?

MS. MEHTA: Objection to the form.

THE WITNESS: Well, I can't guarantee that my assumptions on Jack's intentions here are 100 percent accurate.

But what I understand this to mean is that AI will transform society, the economy, how we think of ethics, morality, politics. Pretty much every aspect of human civilization will be affected in a very deep way by the advancement of powerful AI systems.

BY MR. MILLER:

Q And do you agree with that?

A Yes, I think that's likely going to happen.

176

Q Do you think AI is going to replace your job?

MS. MEHTA: Objection to the form.

THE WITNESS: I believe that there's no real limits to what AI will or will not be able to do in the -- in the fullness of time. The exact timeline on when various capabilities will be achieved is a matter of forecasting.

But yes, I have no reason to believe that a job like mine couldn't be done by a hypothetical AI system from some point in the future.

BY MR. MILLER:

Q Yesterday, you testified that you worked for a while at the Machine Intelligence Research Institute, also known as MIRI.

Do you remember that?

A Yes, I do.

Q What was the nature of your work at MIRI?

A The specific program that I participated in was one in which they hired a handful of interns, including me, with the goal of helping those interns and simultaneously evaluating them in their ability to transition from a software engineering skill set to a research engineering skill set.

177

Q MIRI was founded by Eliezer Yudkowsky, correct?

THE COURT REPORTER: I'm sorry. Can you repeat the name for me?

MR. MILLER: Eliezer Yudkowsky. I'm not sure I'm pronouncing it correctly, but I'm doing my best.

BY MR. MILLER:

Q Is that correct?

A I might pronounce it Eliezer Yudkowsky. Close enough.

Q And he's the founder of MIRI, right?

A I don't know the exact history of that organization, but I believe he is one of the founders.

Q Do you know whether he still runs MIRI?

A I don't know for sure.

Q Did you work directly with Mr. Yudkowsky at all during your time at MIRI?

A He was certainly in some of the presentations or meetings that I attended, but I wouldn't say that I worked closely with him.

MR. MILLER: Tab 51.

MS. GOMEZ-REICHMAN: Again, Mr. Mann and Addison, please click on tab 51 in the chat.

Pages 210-217

210

In your answer about the bioweapons, you caveated in that the model is engaged in memorization. What if the model is not engaged in memorization, but is still used for that purpose?

MS. MEHTA: Objection.

Sorry. One second. I just want to read the question.

Objection to the form. Incomplete hypothetical. Calls for a legal conclusion.

And then, Mr. Mann, again, if you have a formed view on this independent of advice of counsel, you can share that. Anything that would reflect advice of counsel or would be colored by advice of counsel, you cannot share.

THE WITNESS: I don't think I understand your question.

What do you mean by "not engaged in memorization"?

BY MR. MILLER:

Q So a model -- say someone trains a model on a corpus of scientific texts that describe molecular biology, chemistry, so on and so on, and then uses the AI model based on that scientific knowledge to generate instructions for how to build a bioweapon, but those instructions don't

211

specifically regurgitate -- or -- or any of the copyrighted content on which the model is trained.

Do you think that would be fair use?

MS. MEHTA: Objection to the form. Incomplete hypothetical. Calls for a legal conclusion and speculation.

Again, Mr. Mann, you can testify if you have a formed view on that question independent of advice of counsel, but you cannot share anything that would based on or colored by the advice of counsel.

THE WITNESS: I think we're getting into very specific hypotheticals, and I haven't thought specifically that much about fair use as it pertains to bioweapons, so I'll caveat my answer with -- that I'm sort of reasoning from first principles here. And again, I'm not a lawyer. I'm not an expert on these topics. I'm an interested layperson who knows a little bit about copyright.

If we come back to the definition of copyright, your hypothetical said we have a large corpus of science technology, mathematics-related content. You said something like that, but I'm misquoting a little bit. Assuming that there was

212

no regurgitation, could this be construed as fair use?

I think this hypothetical is hard to judge. Because if I were actually looking at this, I'd want to understand more about the distance between the original input material and the output. For example, there are ways of rephrasing input material that I personally wouldn't call transformative, that nonetheless would not be strict regurgitation.

So I don't think I have enough information in this hypothetical to make a judgment, even a layperson's judgment.

BY MR. MILLER:

Q Do you use Claude personally?

A Yes, I do.

Q Have you created a user account to use Claude?

A I have at least one user account to use Claude, yes.

Q Do you have more than one user account for Claude?

A Yes, I do.

Q How many do you have?

A I have one under my personal e-mail. And

213

I have a couple under my work e-mail in our production service. And for testing purposes, I also have a couple in our staging environment. And it's possible that I've created ephemeral accounts for local testing in the past, but I don't specifically remember.

BY MR. MILLER:

Q When you use Claude personally, do you use Claude to help you write?

A Writing is one of the many purposes that I have used Claude for, yes.

Q Have you ever used Claude to help you write a poem?

A I believe I have.

Q Have you ever used Claude to help you write song lyrics?

A I believe I have.

Q Have you ever used Claude to answer questions about music?

A Yes, I believe I have.

Q In your use of Claude, has Claude ever generated a response to one of your prompts that appeared to you to contain copyrighted content?

MS. MEHTA: Objection. Form.

THE WITNESS: It's possible.

214

BY MR. MILLER:
Q In your use of Claude, has Claude ever generated output in response to one of your prompts containing song lyrics?
MS. MEHTA: Objection to the form.
THE WITNESS: Can you state the question again?
BY MR. MILLER:
Q In your use of Claude, has Claude ever generated an output in response to one of your prompts that contained song lyrics?
MS. MEHTA: You mean original song lyrics or copyrighted song lyrics?
BY MR. MILLER:
Q Copyrighted song lyrics. Existing song lyrics.
MS. MEHTA: Objection to the form.
THE WITNESS: I don't specifically remember.
MR. MILLER: Tab 31, please.
(Plaintiffs' Exhibit 214, Prompt, Bates stamped Anthropic_0000431956, was marked and presented for purposes of identification.)
MR. MILLER: What number are we on again?
MS. GOMEZ-REICHMAN: 214.

215

BY MR. MILLER:
Q In the Zoom, you should see what is being marked for the record as Exhibit 214 and reflects the Bates number Anthropic_431956.
A I see it.
Q Do you recognize what has been marked as Plaintiffs' Exhibit 214?
A It appears to be a conversation between a human and an AI assistant, but I don't specifically recognize it.
Q Did you write the prompt reflected here?
MS. MEHTA: Objection to the form. Foundation.
THE WITNESS: I don't remember writing it.
BY MR. MILLER:
Q Would you be surprised if the metadata for this document indicated that the prompt was written by ben@anthropic.com?
MS. MEHTA: Objection to the form. Are you making a representation to that effect, Counsel?
MR. MILLER: Not yet. I asked the question that I asked.
MS. MEHTA: Objection to the form. Misleading.

216

THE WITNESS: It's certainly possible that I would have written this prompt, but this -- if this date is correct, January 18, 2024, it would have been almost two years ago, and I write a lot of different prompts, so I have no specific recollection of this prompt.
I'll also say that typing in all caps isn't my style, so if it is something that I entered into a system, it's also -- it's still possible that I was not the author.
MR. MILLER: Can we go to tab 34?
(Plaintiffs' Exhibit 215, Prompt: Stevie Wonder lyrics, Bates stamped Anthropic_0000524147, was marked and presented for purposes of identification.)
BY MR. MILLER:
Q You should have in front of you in the Zoom what's being marked as Plaintiffs' Exhibit 214 -- 15, excuse me, 215, which bears Bates numbers Anthropic_524147 and Anthropic_524148.
Do you have that in front of you?
A Yes.
Q Do you recognize what has been marked as Exhibit -- as Plaintiffs' Exhibit 215?

217

MS. MEHTA: Form.
THE WITNESS: I recognize the content but not the presentation.
BY MR. MILLER:
Q Okay. What is the content that you recognize in Plaintiffs' Exhibit 215?
A The content is from a conversation I conducted with Claude in which I transmitted a question from my dad where he said that there was a song called Every Day is a New Day -- or, sorry, that the -- the song included the lyrics, Every day is a new day, but we couldn't remember which song it was. And we were hoping that Claude could find us the name of the song so that we could play it in the music subscription service that we use.
Q Why did you use Claude to try to answer this question?
A I also Googled it. I also asked Gemini. And to the best of my recollection, none of the services gave the correct answer, unfortunately.
Q Have you ever worked in the music industry in any capacity?
A I have not.
Q Do you like music?
A I do like music.