EXHIBIT 131

(FILED UNDER SEAL)

1

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

CONCORD MUSIC GROUP, INC., ET AL., Plaintiffs, -vs- ANTHROPIC PBC, Defendant.) Case No. 5:24-cv-03811-EKL SVK

HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY

VIDEOTAPED DEPOSITION OF:

SHAWN SHAN, Ph.D.

The videotaped deposition of SHAWN SHAN, Ph.D., called as a witness pursuant to notice and pursuant to the provisions of the Federal Rules of Civil Procedure 30 and the Rules of the United States Supreme Court thereof pertaining to the taking of depositions; taken before ISAIAH ROBERTS, CSR, RPR, a Certified Shorthand Reporter, at Arnold & Porter, 70 West Madison Street, Suite 4200, Chicago, Illinois 60602-4231 on March 3rd, 2026, commencing at 9:00 a.m. Central Time.

2

APPEARANCES:

OPPENHEIM ZEBRAK, LLP
By: MR. KEITH HOWELL
MR. TIMOTHY CHUNG (Appearing remotely)
4530 Wisconsin Avenue
5th Floor
Washington, D.C. 20016
Phone: 202.957.5043
Email: Khowell@oandzlaw.com

Appearing on behalf of the Plaintiffs;

WILMERHALE
By: MS. SONAL MEHTA
MS. ELLEN STEPHANO (Appearing remotely)
2600 El Camino Real
Suite 400
Palo Alto, California 94306
Phone: 628.235.1026
Email: Sonal.mehta@wilmerhale.com

Appearing on behalf of the Defendant.

ALSO PRESENT:

MR. IAN CHEN, Anthropic, In-house Counsel

3

I N D E X

PAGE
WITNESS: SHAWN SHAN, Ph.D.
Examination by Ms. Mehta. 5

E X H I B I T S

PAGE
Shan Exhibit Number 1. 18
(Expert Report of Dr. Shawn Shan)
Shan Exhibit Number 2. 203
(Rebuttal Expert Report of Dr. Shawn Shan)
Shan Exhibit Number 3. 265
(Claude_inconsistency_test.csv)
Shan Exhibit Number 4. 270
(Claude_transcriptions_coldplay_style.csv)
Shan Exhibit Number 5. 275
(Special_character_prompts_responses_bypassed_run1.csv)
Shan Exhibit Number 6. 280
(Claude_extraction_test_full.csv)
Shan Exhibit Number 7. 298
(Reply Report of Dr. Shawn Stan)

4

THE VIDEOGRAPHER: Here begins Media Number 1 in the videotape deposition of Dr. Shawn Shan in the matter of Concord Music Group, Inc., et al., v. Anthropic PBC in the United States District Court, Northern District of California, San Jose Division, case number 24-cv-03811-EKL-SVK.

Today's date is March 3rd, 2026. The time on the video monitor is 9 a.m. Central. The videographer today is Richard Figueroa representing Planet Depos. This video deposition is taking place at Arnold & Porter, Chicago, Illinois.

Would counsel please voice identify themselves and state whom they represent.

MS. MEHTA: Good morning. Sonah Mehta from WilmerHale on behalf of Anthropic. With me on the Zoom is my colleague Ellen Stephano. And Ian Chen, in-house counsel at Anthropic, is also on the line, on the Zoom.

MR. HOWELL: Keith Howell from Oppenheim + Zebrak on behalf of the music publisher plaintiffs.

THE VIDEOGRAPHER: The court reporter today is Isaiah Roberts, representing Planet Depos. The witness will now be sworn in.

MS. MEHTA: Actually sorry. Just one thing.

Is there anyone else on the Zoom that wasn't announced? I think Tim might be on.

5

THE VIDEOGRAPHER: Timothy Chung is on Zoom.
MR. CHUNG: Timothy Chung on behalf of publishers.
MR. HOWELL: There you go.
SHAWN SHAN, Ph.D.,
called as a witness herein, having first been duly sworn, was examined and testified as follows:
DIRECT EXAMINATION
BY MS. MEHTA:
Q. Good morning, Dr. Shan.
A. Good morning.
Q. Have you ever been deposed before?
A. No.
Q. Okay. Let's go through background rules just to make sure you and I are on the same page about how we'll conduct the day today. Okay?
A. Yes.
Q. Okay. The first thing is we have a written transcript of the testimony, so we need to make sure that we are giving verbal responses. Shaking of the head or nodding or uh-huhs or huh-uhs will not get transcribed on the transcript, so I'd ask you to do your best to please make sure you answer verbally. Okay?
A. Okay.
Q. For the same reason, it's important that you

6

and I not speak over one another, and that you not speak over your counsel. So I'm going to do my best to let you finish your answer before I ask a question. I'm going to ask you to do your best to let me finish a question before you answer. Okay?
A. Okay.
Q. Is there any reason that you wouldn't be able to provide your most truthful and accurate testimony here today?
A. No.
Q. You understand you're testifying under oath just as you would in front of a judge and a jury in San Jose?
A. Yes.
Q. If at any point you don't understand my question, please ask me for clarification. If you don't ask for clarification, I'm going to assume you understood the question, you're providing your most truthful and accurate testimony in response. Okay?
A. Okay.
Q. We'll take a break every hour, 90 minutes or so, just whenever we're at a good stopping place. If you need a break in less time than that, let me know and I'll wrap up that line of questioning and then we'll take a break. Okay?

7

A. Okay.
Q. When were you first approached in this matter?
MR. HOWELL: Objection. Vague.
THE WITNESS: I recall September last year or maybe August last year. I got an email from one of the counsels about retaining me for this matter.
BY MS. MEHTA:
Q. And when you say September or August last year, you mean 2025?
A. That's correct.
Q. Okay. And this was counsel from the Oppenheim + Zebrak firm?
A. That's correct.
Q. Do you know how they got your name?
MR. HOWELL: Objection. Form. Speculation.
THE WITNESS: I do not know.
BY MS. MEHTA:
Q. Had you spoken with Dr. Zhao about this case at any point prior to the lawyers from Oppenheim + Zebrak reaching out to you?
A. I do not recall.
Q. You were working with Mr. Zhao last year; correct?
MR. HOWELL: Objection. Form.
THE WITNESS: He is my advisor during my Ph.D.

8

BY MS. MEHTA:
Q. When did you first start working with him?
MR. HOWELL: Objection. Form.
THE WITNESS: Can you clarify what you mean by working with him?
BY MS. MEHTA:
Q. When did you first -- when did he first become your advisor?
A. When I started my Ph.D., which is 2020, September.
Q. Okay. And have you also worked with him on projects outside of his advisory capacity as you did for your Ph.D.?
A. No.
MR. HOWELL: Objection. Form.
BY MS. MEHTA:
Q. Did you work with him on the Glaze project?
A. Yes.
Q. Okay. So when I had just asked you whether you worked with him on projects outside of your -- his advising on your Ph.D., would you be Glaze to be part of his advising on your Ph.D.?
A. Yes.
Q. Okay. When you started working with him, were there any other apps or tools or projects that you

85

the later paragraph you're referring to?
BY MS. MEHTA:
Q. I mean the rest of the section, so I mean your discussion at paragraphs 19 through 23 is intended to summarize how LLMs like Claude work -- or are trained, rather?
MR. HOWELL: Objection. Form. Vague. Compound.
THE WITNESS: Sorry. Can you repeat your question?
MS. MEHTA: Yes.
BY MS. MEHTA:
Q. My question was -- that paragraphs 19 to 23 are intended to summarize how LLMs like Claude are trained; correct?
MR. HOWELL: Objection. Form. Vague. Compound. Misstates document.
THE WITNESS: There are some portion of -- this background is about LLMs and the ones are similar to Claude, but there's also discussion about general data collection, like crawling, other type of AI models which follow very similar data collection, training processes.
BY MS. MEHTA:
Q. Okay. Let's go to paragraph 19.

86

At paragraph 19, you say, "Text generative LLMs like Claude are a class of generative AI that were trained on and mimick natural human language in their written output."
Do you see that?
A. Yes.
Q. That's an accurate statement?
MR. HOWELL: Objection. Form.
THE WITNESS: Yes.
BY MS. MEHTA:
Q. You go on to say that "Broadly speaking" -- same paragraph -- "Broadly speaking, generate AI refers to AI systems that analyze existing data, such as audio, images, or text, and generate content based on learned patterns derived from the datasets they are trained on."
Do you see that?
A. Yes.
Q. Is that an accurate statement?
A. Yes.
Q. And does that statement apply to Claude in so far as it refers to text?
MR. HOWELL: Objection. Form.
THE WITNESS: What do you mean by apply to Claude?

87

BY MS. MEHTA:
Q. Meaning is Claude an generative AI system that analyzes existing data, such as text, and generates content based on learned patterns derived from the datasets that it is trained on?
MR. HOWELL: Objection. Form.
THE WITNESS: The exact training process and the model weights are not produced as part of this inspection, but based on all the public documents and the training data I've inspected, as well as the source code, that is my impression.
BY MS. MEHTA:
Q. Okay. Now, when you said that the model weights and the training data was not made available, the model weights were made available to you, weren't they?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: I do not believe the model weights are made available. The training data is made available. Before, I was referring to the training source codes about how -- the architecture of Claude, as well as the exact optimization that Anthropic used to train each of their models was not produced as part of their inspection.
BY MS. MEHTA:

88

Q. Do you know if the publishers asked for that information that you say was not produced?
MR. HOWELL: Objection. Form.
THE WITNESS: When you say that information, what are you referring to?
BY MS. MEHTA:
Q. Yes. You said that the model source code about the architecture of Claude and the exact optimization that Anthropic used to train each of their models was not produced as part of the inspection.
Do you know if Claude -- if the publishers' counsel actually asked for any of that information?
MR. HOWELL: Objection. Form.
And I'll caution the witness not to reveal attorney-client protected and/or attorney work product protected information, including his conversations with counsel.
THE WITNESS: I do not know.
BY MS. MEHTA:
Q. Could you go to -- back to paragraph 19.
The next sentence says, "Specifically, LLMs like Claude analyze the statistical relationship or conditional probability of tokenized words and phrases and are taught to generate written outputs based on preceding context."

89

Do you see that?
A. Yes.
Q. Is that an accurate statement?
MR. HOWELL: Objection. Form.
THE WITNESS: That is accurate statement based on what I understand about how LLM works.
BY MS. MEHTA:
Q. And is it an accurate statement based on how you understand Claude works? In other words, you understand that Claude analyzes the statistical relationships or conditional probability or tokenized words and phrases and is taught to generate written outputs based on preceding context?
MR. HOWELL: Objection. Form. Compound. Vague.
THE WITNESS: Can you repeat your question?
MS. MEHTA: Yes.
BY MS. MEHTA:
Q. Is that an accurate statement about how Claude works as opposed to how LLMs generally work? In other words, Claude analyzes the statistical relationship or tokenized words or phrases and is taught to generate written outputs based on preceding context?
MR. HOWELL: Objection. Form. Compound.
THE WITNESS: Again, I have not reviewed the

90

exact source codes for Claude, but based on the public papers that are published by Anthropic and my general expertise on how LLMs work in general, I believe Claude does exactly what I detailed right here.
BY MS. MEHTA:
Q. The next sentence, you say, "In the case of LLMs like Claude, the outputs are text that appear natural and coherent generated in response to human inputs."
Do you see that?
A. Yes, I do.
Q. Is that an accurate statement of how Claude works, to your understanding?
MR. HOWELL: Objection. Form. Vague. Misstates document.
THE WITNESS: What do you mean by how Claude works? Because I don't think I make statement about how Claude works in this sentence.
BY MS. MEHTA:
Q. You do. You say, "In the case of LLMs like Claude, the outputs are texts that appear natural and coherent, generated in response to human inputs."
Is that an accurate statement? Let's start with that.
A. Yes. That's an accurate statement.

91

Q. Okay. And does that accurately describe how Claude works? In other words, that Claude has outputs that texts that appear natural and coherent, generated in response to human inputs?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: So in a case when -- from my experience of prompting Claude, when I ask in normal human language, it responds with natural and coherent text.
And I believe that if Claude were to ask, you know, source coding information, for example, you would generate source code, which oftentimes is not natural, coherent to human eyes, but natural and coherent regardless for a machine to understand.
BY MS. MEHTA:
Q. Okay. That's fair.
Would you agree that where Claude is prompted for text the output is text that appears natural, coherent -- natural and coherent?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
MS. MEHTA: As you say here.
MR. HOWELL: Same objections.
THE WITNESS: I have not done the study or a holistic study of how Claude responds to human inputs,

92

so I cannot say definitively that is the case.
BY MS. MEHTA:
Q. Okay. You're not denying what you said in your report though that in the case of LLMs like Claude, the outputs or texts that appear natural and coherent generated are in response to human inputs. You still stand behind that?
MR. HOWELL: Objection. Form.
THE WITNESS: I stand by the fact that given the human English inputs -- or maybe non-English, which I have no expertise in -- the output generally looks natural and coherent, that is the goal of training LLMs. But, of course, there are cases if the model is not well-trained or prompt in certain ways, models can generate gibberish, can reveal sensitive information, and we have seen many literature about those.
BY MS. MEHTA:
Q. With respect to the human inputs that you're referring to, are you referring to prompts made by users of Claude?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: That is correct.
BY MS. MEHTA:
Q. Who controls the prompts that are made -- or the human inputs, to use your word, that are made to

93

Claude?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: What do you mean by control?
BY MS. MEHTA:
Q. Well, when you prompt Claude, does Claude tell you what the prompt should be or do you put in your own prompt?
MR. HOWELL: Objection. Form. Vague. Compound.
THE WITNESS: Can you repeat your question, please.
MS. MEHTA: Yes.
BY MS. MEHTA:
Q. My question is: When you prompt Claude, does Claude tell you or prescribe what the prompt should be or do you put in your own prompt?
MR. HOWELL: Objection. Form. Vague. Compound.
THE WITNESS: I believe how Claude works, based on my inspection, especially related to the code of guardrails, users do have some control over the prompting inputs.
However, a company like Claude when they host these closed source LLMs, they can, you know, reject

94

prompts as I've seen in the code, they can append additional instructions to Claude, and to the best of my knowledge, these are not controlled by user.
BY MS. MEHTA:
Q. Okay. So let's set aside guardrails then. So I -- I'm not asking about guardrails that are put in place.
In terms of the initial prompt to Claude, that is something that you or the user defines; correct?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: The exact comment, if a user were to were Claude.ai chat interface, it is controlled by the user when typing that exact comment. But what Anthropic does to that prompt, probably uses a guardrail, and there may or may not be other things that Anthropic do. I have no information about that.
However, I have seen literature where LLM host services will tweak or rephrase user prompts, make the model potentially generate better response.
BY MS. MEHTA:
Q. Now, you haven't done an analysis yourself of whether or not Anthropic does that? I think you just said that; right?
MR. HOWELL: Objection. Form. Vague.

95

Misstates prior testimony.
THE WITNESS: I have reviewed the source code -- source guard for guardrails, so I understand how specifically prompts show and the CCS were to block or modify responses -- or modify the input prompt.
I have not seen any evidence where Claude will rephrase user's messages, but that is something that can be done do improve the performance so there is a probability that Anthropic uses inputs like that.
BY MS. MEHTA:
Q. Let me break down the question, because I think there's three parts of what you're talking about, and I want to make sure we're being pretty precise.
So let's first start with the actual submission of the prompt. The user and the user alone types in the prompt and actually makes the submission of the prompt to Claude; correct?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: Not always.
BY MS. MEHTA:
Q. Who other than the user would be inputting a prompt into Claude?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.

96

THE WITNESS: So LLMs, one of the use case is, as we describe, a user go to Claude.ai or .com to use the service, but if I understand correctly, that is a fraction of how LLMs are used today. And, you know, Claudes have API access, and for most of those API access, the code script control what prompt is being used.
And there are certain automated systems that can guide those prompts, and Claude is also used as a reasoning systems or an agent, where, you know, user ask original task, but Claude will generate command and prompt itself to do a step-by-step educate -- execution of the task.
And in all of the other cases other than if the user just simply go to Claude.ai, I think users oftentimes do not have control over the exact prompts that are being fed into Claude.
BY MS. MEHTA:
Q. So let's focus on the prompts that you made for purposes of your work in this case.
Before you hit submit, the actual typing of the prompt, that you did yourself; correct?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: So I used both the API and Claude.ai, and for all the Claude.ai prompt, I manually

97

typed. What were some of the song names, I may have copy and past from, you know, some main document as part of the exhibits.

And in terms of API, I used a program to control more holistic tests of the AI system when I need to submit multiple inquiries, and it is the computer that control the actual submission of those things.

BY MS. MEHTA:

Q. Okay. So let's then split it up between Claude.ai and the API.

With respect to Claude.ai, you manually typed or copied and pasted in the prompt before it was submitted to Claude; correct?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: Yes. That is correct. I either typed or copy, paste from, you know, the documents, digitally onto the interface and submitted it to myself.

BY MS. MEHTA:

Q. And you defined what that prompt would be?

MR. HOWELL: Objection. Form.

THE WITNESS: What do you mean by define the prompts?

BY MS. MEHTA:

Q. You made a judgment as to what the text of that prompt should be before you submitted it?

98

MR. HOWELL: Objection --

THE WITNESS: Yes.

MR. HOWELL: Form.

THE WITNESS: Yes.

BY MS. MEHTA:

Q. No one else told you what that prompt should say; correct?

MR. HOWELL: Objection. Form.

Caution regarding any attorney-client work product.

THE WITNESS: I was assigned a task on what I test, and depending on different task, of course, I would use different prompts, but the final prompt is -- I can put into the system, I copy and paste myself.

BY MS. MEHTA:

Q. In other words, you had control over what that final prompt was before you put it into the system?

MR. HOWELL: Objection. Form. Vague. Calls for speculation.

THE WITNESS: Can you repeat the question?

MS. MEHTA: Yes.

BY MS. MEHTA:

Q. You yourself had the final control over those prompts before they went into the system for purposes of

99

your prompt and output analysis?

MR. HOWELL: Objection. Form. Vague. And calls for speculation.

THE WITNESS: I, you know, go onto Claude.ai, I type it out or copy and paste with my house, and I submit it -- submitted it by pressing enter.

BY MS. MEHTA:

Q. Anthropic didn't control what you could and couldn't put into that prompt; correct?

MR. HOWELL: Objection. Form. Vague. Calls for speculation. Misstates prior testimony.

THE WITNESS: What do you mean by Anthropic control what I --

BY MS. MEHTA:

Q. Anthropic didn't tell you what the prompt should say or limit what your prompt should say, did it?

MR. HOWELL: Objection. Form. Vague. Calls for speculation. Misstates prior testimony.

THE WITNESS: So there are cases where the prompt was blocked --

BY MS. MEHTA:

Q. I'm not asking about -- I'm not asking about after you submitted. I'm asking before you hit the submit.

Before you hit the submit button, Anthropic

100

didn't limit what you could or could not type into the prompt and did not tell you the what you had to put into the prompt; isn't that right?

MR. HOWELL: Again, counsel, objection. Form. Vague. Calls for speculation. Misstates prior testimony.

And you have to let the witness finish his responses. That's at least the second time you've cut him off.

Go ahead.

THE WITNESS: Yes, I do not see evidence of Anthropic limit what can be typed into the text box as long as it's English words.

BY MS. MEHTA:

Q. Now, let's go to the API part of this.

So you said you also used the APIs to do some prompting. When you used the APIs to do the prompting, did you define what the commands were in terms of what the API should call for?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: Define prompts that API used.

BY MS. MEHTA:

Q. Meaning, did you -- did you define -- I'm not sure what other word to use.

Did you prescribe what calls the API should

125

record. The time is 11:41 a.m.
BY MS. MEHTA:
Q. Okay. Dr. Shan, could you turn to page 8 of your report and paragraph number 21 again.
A. Did you say 21?
Q. Yes.
A. Okay.
Q. So we had just been going through that, and now I want to get the last sentence in that. It says, "Consequently, LLM developers like Anthropic seek to train LLMs like Claude on a vast training corpus of human written texts often comprising of terabytes or petabytes of data and hundreds of billions or trillions of tokens."
Do you see that?
A. Yes.
Q. Is that an accurate statement?
MR. HOWELL: Objection. Form.
THE WITNESS: Yes. Again, that is based on my understanding of how today's LMs are trained. And this may not be the case, you know, in the future.
BY MS. MEHTA:
Q. Is it based on -- is that an accurate statement as to how Claude was trained based on your review of the information in the record in this case?

126

MR. HOWELL: Objection. Form.
THE WITNESS: Are you asking -- there's two parts to that sentence. Are you asking for both parts or asking for one specific part of that statement?
BY MS. MEHTA:
Q. How about if I ask you the question this way. Is it accurate that Anthropic trained L -- trained Claude on a vast training corpus of human written texts?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: Based off my analysis of the dataset that were produced, yes, I believe so.
BY MS. MEHTA:
Q. Is it accurate that Anthropic trained Claude on terabytes or petabytes bites of data?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: I do not recall the exact numbers, but I believe LLMs like the commercial high performance LLM that we have today are trained on billions of tokens.
BY MS. MEHTA:
Q. So you wouldn't be surprised to hear if -- I wasn't asking about tokens in that question actually. Sorry. I was asking about terabytes or petabytes of data.
A. So based on how --

127

MR. HOWELL: Objection. Form. Vague.
THE WITNESS: Sorry.
Based on how the data was produced to me, it is one remote server. And the dataset from -- as far as I can tell, are not hosted on the same remote server and how it's produced is link or a mount, as technically we call it, to a number of different packets on AWS server.
And given the large scale of the data, I do not have the capacity to count exactly how many terabyte or petabyte of data there is, but based on my analysis of how much I search, how long the script takes to load the data, I believe there is at least terabyte, not hundreds or close to petabyte of data that was produced in the training process.
BY MS. MEHTA:
Q. Okay. And then is it accurate that Anthropic trained Claude on hundreds of billion or trillions of tokens?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: So the same constraint since I reviewed the English text, which is the focus of this case. So, for example, I did not focus my analysis on the source code. I did not focus my analysis on Common Crawl-News, for example.
But based on the amount of training that I

128

have seen and what I know in industry, how the conversion rate between token and word, English words, I believe this is accurate statement that the models are trained on hundreds -- or on billions or trillions of tokens for the different models that Anthropic used commercially.
BY MS. MEHTA:
Q. Would you turn to paragraph 22. And you can -- if you can just read that paragraph to yourself quietly. I don't need you to read it out loud.
A. I have finished.
Q. With respect to paragraph 22 of your opening report, you talk about how developers of LLMs typically source large portions of their training corpora by scraping text from the website -- from web pages and from using large dataset that are available on the internet; is that right?
MR. HOWELL: Objection. Form.
THE WITNESS: This is my understanding based on the data I inspected at Claude as well as publications around training large LLM models.
BY MS. MEHTA:
Q. My understanding is that it's well known that all of the large commercial large language models trained on Common Crawl. Is that your understanding?

177

that Anthropic produced to me, which, if I understand correctly, is the data they collected for the purpose of training Claude.

I understand that there may be deduplications or, you know, data cleaning technologies that Anthropic may or may not use for this process, but from a computer science perspective that -- the state of the art deduping or cleaning method that I'm aware of have very high false positive and false negative rates removing duplicates from the raw cleaning data to the final mix.

And I believe that all of these methods rely on compared data points against data points to see whether data collected from this web page is similar to data on from another web page. If they're similar, they would get deduplicated based on a probabilistic hashing technique.

And in the case where I examined lyrics, what I find is that lyric does appear across different websites, but oftentime the rest of the web pages are very different. There's no reason for me to believe any dedupe method that Anthropic may or may not use will remove these duplicates. And my opinion is that I have found, you know, millions and hundreds of thousands of copies or incidents of these lyric in the raw training sets.

178

And, you know, based on my inspection of the final models, this also shows that Anthropic Claude models can sometimes regurgitate copyright lyrics. And based on, you know, my understanding of this technology and how Dr. Bilmes also discussed, the more times a piece of data appears in your train data mix, the more likely something will get regurgitated.

BY MS. MEHTA:

Q. Right. So I'm going to ask you just to focus on my question because I want to wrap this up so we can take lunch, but I do want to make sure we just wrap this up.

So I'm not asking about whether you think ultimately the training mix would include lyrics or not. We'll talk about that after lunch. I just have a very specific factual question.

The data that you -- the training data that you analyzed in the inspection environment and that you analyzed for your report, that was the raw training data; correct?

MR. HOWELL: Objection. Form. Vague. Misstates prior testimony. Asked and answered.

THE WITNESS: I do not know how Anthropic used each folders, even those indicated may or may not used in the final training models. And as I stated before, I

179

have not seen the final model training code and the final training weights.

So these datasets I only examine as they were produced. I do not have any opinion on what it's used for.

BY MS. MEHTA:

Q. Okay. Is it also fair to say that you don't know whether or not the datasets that you looked at were prior to or after deduplication?

MR. HOWELL: Objection. Form. Vague. Misstates prior testimony.

THE WITNESS: As I stated, these are data produced to me. You may have used dedupe method or you may not. And as I stated about the technology of deduping, at this scale deduping methods is far from perfect. So even if Anthropic were to then dedupe method after the fact, there's high likelihood most or large number of these duplicates are not removed in the training.

BY MS. MEHTA:

Q. Right. I'm not asking about that. I'm, again, asking a simple factual question.

The data that you looked at, you don't know whether that was before or after deduplication; correct?

180

MR. HOWELL: Objection. Form. Vague. Calls for speculation.

THE WITNESS: Again, I simply inspect the data produced on the environments. I do not have evidence or the source code for how these data are collected, stored, used, processed to train the model, but I do believe these are related to different model training.

There's a mapping specific how about, you know, this Common Crawl data was used for this model, the other Common Crawl was used for a different model. So I believe these data are used in some part of building Anthropic's AI technology.

BY MS. MEHTA:

Q. Right. That's not my question. My question was: The actual data that you looked at and that you analyzed, do you know whether that data that was in the inspection environment was pre- or post-deduplication?

MR. HOWELL: Objection. Form. Vague. Calls for speculation. Asked and answered.

THE WITNESS: Again, I look at all the data, and I do not know whether additional process, cleaning, other potential removal of CMI was being applied before the training, and I have not inspect the exact training data.

And, again, I believe this is irrelevant to my

181

analysis because any dedupe methods is not perfect. And if I find a large number of data in general, that's strong evidence that there's -- an overwhelming the many duplicates in the final mix.
BY MS. MEHTA:
Q. Right. Again, I'm not asking about whether you think it's relevant or not relevant. You have your opinion on that. Ultimately the judge and the jury will decide. I'm asking a factual question.
Are you able to say under oath whether the data that you looked at in the inspection environment was before or after deduplication?
MR. HOWELL: Objection. Form. Vague. Asked and answered. And I think the witness asked about lunch about 10 minutes --
MS. MEHTA: As soon as I get an answer to this question, we're going to take a break.
MR. HOWELL: You've asked it several times at this point. The witness has provided an answer.
THE WITNESS: I'm sorry. Can you repeat the question?
BY MS. MEHTA:
Q. Yes. Again, I'm not asking whether you think it's relevant or not. I'm not asking how deduplication works. I'm asking a very simple, factual question.

182

Are you able to say under oath whether the data that you looked at in the inspection environment was before or after duplication?
MR. HOWELL: Objection. Form. Vague.
MS. MEHTA: Deduplication.
MR. HOWELL: Objection. Form. Vague. Asked and answered.
THE WITNESS: I have reviewed data as produced. And from different evidence I see, including the mapping file, it shows that these data are used for different Claude models.
BY MS. MEHTA:
Q. That's not my question. My question is not whether it was used at any point in time. Please focus on the specific language of the question. Once we answer this, I'm going to take the break you asked for for lunch.
My question is: Do you know whether the data that you looked at was before or after the deduplication process?
MR. HOWELL: Objection. Form. Vague. Calls for speculation. Asked and answered.
THE WITNESS: Again, I go in, review all the data produced. And based on my understanding of the case, these are data used to build Anthropic AI

183

technology.
BY MS. MEHTA:
Q. Okay. Again, sir, I need you to focus on the question I'm asking. That is not responsive to the question I'm asking. I'm not asking whether you think the data was used in the process, whether it was used to train or build the Anthropic models. It's a very simple, factual question.
Do you know and can you say under oath whether or not the data was before or after deduplication? Or are you just refusing to answer that question?
MR. HOWELL: Objection. Form. I object to the commentary, which is improper and harassing. It's -- the question has been asked and answered. It calls for speculation.
We've been going about 30 -- at least 30 minutes longer than we had agreed prior to this session so we're going to take a hard stop after the next answer.
MS. MEHTA: No, I'm not going to agree to that. I want an answer to this question.
MR. HOWELL: If you want to sit here while the time runs, you're welcome to do so. We're going to take lunch at -- we're at almost an hour and 15. I think we talked about a 30-minute session before this.

184

BY MS. MEHTA:
Q. Dr. Shan, can you say under oath whether or not the data you inspected was before or deduplication?
MR. HOWELL: Objection. Form. Calls for speculation. Asked and answered.
THE WITNESS: As I said before, I go in; they produce a large number of data. There's evidence about what these data are, and the evidence pointing to these data are used to build AI models. And, again, I have not reviewed specific training code where you can load the data to build a model. That is to extent what I did, what I put in the report.
BY MS. MEHTA:
Q. Okay. So you don't know whether or not, based on the training code, whether this was before or after deduplication, do you?
MR. HOWELL: Objection. Form. Calls for speculation. Asked and answered. Harassing.
THE WITNESS: Again, I reviewed what I was produced with. I do not know whether there's -- specific code was used after the fact. Maybe or maybe not, but it doesn't change my opinion on the report because of all the discussion I had about deduping and explaining methodology and, you know, imperfection of this method at this scale.

225

Q. Great. Is it your testimony that you can definitivelyclaim that in generating a particular output Claude references a particular piece of training data?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: Can you repeat that?
BY MS. MEHTA:
Q. Yes. Can you definitively claim that in generating a particular output Claude references a particular piece of training data?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: As we discussed, I do not know the exact underlying of what weights correspond to what memorized training data.
But based on the empirical results and the research that I cited in paragraph 13, it seems to suggest that given a sequence of word -- Claude models oftentime or sometimes will predict mirror definitely next word in a sequence. And from my tests that -- I showed that we can do this for copyrighted lyric, and the probability and the results of this near verbatim extraction is beyond what can happen by random chance.

226

So I believe the training data during the training process which are used to optimize this model have a significant impact on the model's ability to regurgitate these training analyses.
BY MS. MEHTA:
Q. Okay. My question is different than that.
I'm not asking whether you think that it is -- it's suggested or you can infer it.
I am asking whether you can claim definitively that when Claude outputs a particular output, that it is referencing a particular piece of training data?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: What do you mean by referencing specific training data?
BY MS. MEHTA:
Q. That it is retrieving a specific piece of training data. Can you claim that definitively?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: So given how LLMs are trained or how general AI models are trained, the training process, the model weights are changing. They're influenced by the training data pieces. So each individual training piece have some impact to the model weights.

227

And the training models -- as during generation time, the model weights directly impacts what information was getting generated, getting responded. And because model weights are influenced by training data, it is standard to believe training data has a large impact to the final model predictions.
And there are a lot of research looking at -- for example, including some of my research, looking at, for example, given the model generation. Try to explain why model generate this way. And these research are able to show the specific training data points to explain why a model generate this way.
And oftentime it is because similar training data, it's because, you know, certain anomaly in a training set. So I believe a model generation is impacted by this training data.
BY MS. MEHTA:
Q. You did not analyze the model weights used in the Claude models at issue in this case; correct?
MR. HOWELL: Objection. Form.
THE WITNESS: I was not produced with the models weights found in these models.
BY MS. MEHTA:
Q. Okay. Now, going back to your answer about how the training data can influence -- how the model

228

weights are influenced by the training data and therefore the training data has an impact on the final model predictions. That's -- that's about the training data in general. I'm asking about particular pieces of training data.
So can you say definitively that when the model outputs a particular output, that it is retrieving or referencing a particular piece of training data as opposed to the training data as a whole that ultimately is weighted -- that goes into the training mix that ultimately is weighted that ultimately impacts the generation of the output -- or the prediction of the output?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: So there are a lot of research, as I discussed before, that can what they call trace back or understand the influence function of what is a subset of training data points that impact the model to make certain predictions.
And this is useful to, you know, analyze certain behaviors or model to understand what are the causes. And there are very strong analytical background, analytical support for these type of analyses.

229

And to your question, for LLM generation, oftentime it is not one specific data point impacting entire generation. It is, you know, a portion of all related or nearby-related, you know, words and data to it.

And in the case -- and obviously different training data have different impact to certain generation. For example, if you train on Japanese, it's not going to help too much on generally English and vice versa.

And in this case what I was looking at is generating near-verbatim copy of copyright lyrics. And the only place that the copyright lyrics appeared is in the copyright lyrics and the training data. And based on my inspection, those are, you know, hundreds of thousands of copies in there.

So, yes, I believe these -- the presence of these data cause the model to have certain weights. While we cannot understand exactly how these weights are learned, but it nevertheless shows it's able to generate the near-verbatim copy, which is direct consequence of data being trained into the model.

BY MS. MEHTA:

Q. And then when you say we cannot understand how the models are learned, that is -- you mean the -- how

230

the model weights are used to train the model?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: I'm simply referring to if you show the model weights to someone, it's not human readable; but, again, nevertheless, you know, storing data doesn't need to be in human readable form.

BY MS. MEHTA:

Q. Right. My question was -- oh, actually, never mind. You answered my question. Okay. Thank you. I appreciate that.

Let's go back to paragraph number 10. Let's go -- actually, let's go to paragraph number 20.

A. On page 5?

Q. Yes. And you can just read paragraph 20 to yourself.

A. Okay. I have finished reading paragraph 20.

Q. Okay. Paragraph 20 you talk -- or you respond to Dr. Bilmes' opinion about URLs that contain plaintiff's lyrics, and then you mention or you describe the difference between the absence of a technical block with affirmative legal position -- permission.

Do you see that?

MR. HOWELL: Objection. Form.

MS. MEHTA: Second sentence.

THE WITNESS: Yes, I see that.

231

BY MS. MEHTA:

Q. You're not offering any opinions about what does and does not constitute legal permission to scrape a website, are you?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: I'm not --

MR. HOWELL: Calls for legal conclusion.

THE WITNESS: I'm not offering legal opinions about what constitute as consenting in a legal sense or not.

But as a general person who owns certain intellectual property rights and have talked -- someone who have spoke to many copyright holders, that is my position, that Robots.txt does not -- is not the same or is not the legal copyright mechanism that right's owners use to manage their copyrights, including myself, on the internet.

BY MS. MEHTA:

Q. Do you know what -- actually, strike that. Let me ask the question this way.

Did you see any evidence in the record in this case about the particular lyric's websites that you reference as being a source for training data and whether or not they engaged Robots.txt?

MR. HOWELL: Objection. Form. Vague.

232

Compound.

MS. MEHTA: I can -- I can ask a better question. That wasn't a very clear question.

BY MS. MEHTA:

Q. There are a number of different lyric websites which you contend are a source for the works in suit lyrics in Anthropic's training data.

Are you with me so far?

A. Yes.

MR. HOWELL: Objection. Form.

BY MS. MEHTA:

Q. Did you see any evidence in this case about whether any of those lyrics websites -- the people that run those websites -- either did or did not intend to authorize Anthropic to -- or anyone else to scrape their websites? Not your speculation, but any specific evidence in this case.

MR. HOWELL: Objection. Form. Vague. Calls for speculation.

THE WITNESS: So based on my analysis of how Anthropic collected data, the data collection happens across many years, including sourcing from Common Crawl and the, you know, data sharks, which was Common Crawl -- that Anthropic leveraged [redacted]

[redacted]

233

And I have not and do not have the technical capacity to assess at a time of accessing, whether through Common Crawl or whether through Anthropic, the exact Robots.txt and preference of these lyric websites.

BY MS. MEHTA:

Q. Did you see any deposition testimony or documents from any of the lyrics' websites about whether they intended or did not intend to allow scrapers like Common Crawl or companies like Anthropic to scrape the lyrics on their websites? Any specific testimony or documents?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: I believe Dr. Bilmes did talk about Robots.txt where he retrieved, I believe, from way back -- from some of these sites. But I also believe there are multiple ways for website owners to stop crawling from a website.

This includes Robots.txt as a soft blocks [phonetic] as, you know, we have detail in my report. I do not think that is a good way to manage content online. There's also other ways. For example, there's many CDNs, which, for example, Cloudflare. You host your website through Cloudflare, and Cloudflare will manage blocking of bots, which do not involve Robots.txt. They detect and block these bots.

234

And there are also cases with terms of service. I believe I have seen websites who do not have Robots.txt but make it clear in their terms of service that, you know, any bots cannot scrape or collect data through alternative methods.

BY MS. MEHTA:

Q. Right. My question was about the specific lyric websites that you referenced in your analysis.

Did you see any evidence in the form of documents or deposition testimony that those websites either intended or did not intend to allow scraping of lyrics from their websites?

MR. HOWELL: Objection. Form. Vague. Calls for speculation.

THE WITNESS: I do not know -- have no information about what the preference of these lyrics websites are across the last 15 years.

BY MS. MEHTA:

Q. Did you see evidence that any of the lyrics websites that you reference in your analysis employed a CDN like Cloudflare to block scrapers at any point in time?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: If I recall, there may be one of the website did employee CDN, but the data I have here

235

are biased in the sense that these are oftentime the data that successfully crawled and collected by Anthropic, and I found substantial content on these websites.

So it's entirely possible there are many other websites that blocks -- or use CDN at some point through their history. And I do not have information on them simply because Anthropic didn't successfully crawl or collect from there or Common Crawl didn't successfully collect or crawl those websites.

BY MS. MEHTA:

Q. Okay. Did you identify in your review of the record in this case any evidence that the publishers in this case included contractual provisions in their contracts with lyric websites that required the lyric websites to block scraping?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: I have not reviewed the specific contracts for each of these licensees. And, you know, I believe these licenses can be changed or renewed over time, so I also have no information about what these licenses look like across the last, you know, 15 years.

BY MS. MEHTA:

Q. Have you seen any evidence of your review of the record in this case that the publishers took any

236

steps with respect to the lyric websites to require that the lyric websites use the Robots.txt exclusion profile -- or exclusion preference I should say -- for lyrics, for the publishers' lyrics?

MR. HOWELL: Objection. Form. Vague. Calls for speculation.

THE WITNESS: Again, I have not reviewed the communication between lyric website and from, say, the publishers where they license the data from.

But based on my understanding and the specific user study I did talking to rights owners, most rights owner does not know, have never heard about Robots.txt. So it would not be surprise to me that certain publisher are not aware of such techniques exists and they are not certainly not aware that, you know, this could be applied to AI crawling, you know, 5 to 10 years in the future.

BY MS. MEHTA:

Q. So let -- let me break that down.

The first thing is you did not review any communications or agreements between the publishers and any lyric websites; correct?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: Not that I recall.

BY MS. MEHTA:

245

respect Robots.txt? Specific evidence?
MR. HOWELL: Objection. Form. Vague. Asked and answered.
THE WITNESS: I have not seen evidence in terms of Anthropic engineers' testimony. And the source codes was not produced to me with respect to Anthropic respecting or not respecting Robots.txt.
But based on my understanding of the 1.8 number and my follow-up tests I did where I wrote a color where check Robots.txt and to assume to not respect Robots.txt to look at the percentage of websites online that block Anthropic, I believe the percentage is around 5 percent.
But the discrepancy here is because many study has reported, including mine, that the number of websites has blocked AI crawlers has increased tremendously over the course of last couple of years. So I believe that is a discrepancy between the 1.8 percent in Dr. Bilmes' analysis and the 5ish percent in my analyses, which corresponds to a crawler that do not respect Robots.txt whatsoever.
BY MS. MEHTA:
Q. Could you turn to paragraph number 36.
A. Okay.
Q. Okay. Let's go to the first sentence in

246

paragraph number 36. You say, "Dr. Bilmes acknowledges that Claude's outputs are stochastic, yet he fails to recognize that a stochastic system cannot be proven 'not' to have memorized material based on a small non-exhaustive sample of outputs."
Do you see that?
A. Yes, I do.
Q. Is that sentence accurate?
MR. HOWELL: Objection. Form.
THE WITNESS: Yes.
BY MS. MEHTA:
Q. The next sentence says, "The probabilistic nature of these models ensures that a refusal to generate content in one instance does not preclude its generation in the next."
Do you see that?
A. Yes.
Q. Is that an accurate sentence?
MR. HOWELL: Objection. Form.
THE WITNESS: Yes.
BY MS. MEHTA:
Q. Is it the case that you can put the identical prompt into Claude at the same time from two different computers or phones and receive different outputs?
MR. HOWELL: Objection. Form. Vague. Calls

247

for speculation.
THE WITNESS: It is possible, but if you ask a very easy answer like one plus one equal two, despite where you ask, it will likely give two deterministically despite the underlying prompt.
BY MS. MEHTA:
Q. Right. But there are going to be instances in which when it's asking for something that's not very simple, Claude will output different outputs in response to identical prompts, even at the same time; correct?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: It is possible.
BY MS. MEHTA:
Q. And if you go to the next sentence, you say, "To demonstrate this inconsistency, I conducted an experiment using the Claude 3 Haiku API, prompting the model 10 times with the same request to continue the lyrics of one of the 499 songs at issue."
Do you see that?
A. I do.
Q. And then you reference in the footnote your "claude_inconsistency_test.py."
Do you see that?
A. I do.

248

Q. And then you say, "The results were inconsistent: the model refused seven times citing copyright concerns, but successfully bypassed its own guardrails to generate the requested lyrics in the other three attempts." And then you see a footnote 27.
A. That's correct.
Q. Okay. And then the citation at the end of that sentence is to Appendix A; correct?
A. Yes.
Q. Okay. Now, are the 10 -- the 10 prompts that you delivered to the model as part of this experiment -- the only prompts that you made to Claude, or did you make other prompts besides these 10?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: Are you referring to this specific experiment I talk about in this paragraph or are you generally referring to all the tests that I've done for Claude models?
BY MS. MEHTA:
Q. Let's start with this experiment. Did you try any other prompts besides -- or did you try it more times than just the 10 times?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: I only test this 10 times, and I think result was consistent -- or was supported by

297

THE WITNESS: I do not recall my opinion on Cleo and Dr. Bilmes' analysis on that, but if you, you know, have the specific paragraph, I'm happy to take a look.
BY MS. MEHTA:
Q. Yeah. My question was a little different.
If you had an opinion on that, it would have been in your reply report; correct?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: I believe Dr. Bilmes provided terabytes of data after opening report, and a little bit less, but also terabytes of data for the rebuttal report.
I did not have the time to review every single one or even the majority of the data that was being produced. I did look at the codes and read Dr. Bilmes' analysis on that, but if I find any additional evidence by looking at that data, I reserve my right to change my opinion with respect to that.
BY MS. MEHTA:
Q. The judge will make judgment about whether you can change -- about whether you have the right to change your opinions or make any new opinions.
With respect to -- you mentioned the reply

298

report, so let me switch to the next exhibit, which will be Shan 7.
(Shan Exhibit Number 7 marked.)
I'm handing you what has been marked as Shan Exhibit Number 7.
A. Okay.
Q. Okay. So if you could go to paragraph number 17. And you can just read that to yourself and let me know when you've had a chance.
A. Okay.
Q. Are you able to quantify with precision the number of copies of the works in suit that exist in the final data mixes for any Claude model?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: Can you repeat that question, please.
MS. MEHTA: Yes.
BY MS. MEHTA:
Q. Are you able to quantify with precision the number of copies of the lyrics to the works in suit that exist in the final data mixes for any commercial Claude model, if any?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: So my task and what I did analysis was searching the list of 499 songs across the

299

data that's being provided to me by Anthropic in the environment. And my analysis to my best knowledge is a vast undercount of the number of lyrics which exist.
As I talk in my report, given the limits of computation that we have discussed before -- and so I have no precise numbers. I can't say about exact of times each work in suit appears in any part of the training data, but I did find hundreds of thousands to millions of copies of these in the raw training data that's produced to be in these machines.
BY MS. MEHTA:
Q. In terms of the exact number of times, if any, that the works in suit appear in the final training mix, are you able to quantify that, yes or no?
MR. HOWELL: Objection. Form. Vague. Asked and answered.
THE WITNESS: Which final training mix are you referring to?
MS. MEHTA: For any Claude commercial mode.
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: Are you referring to the one by Dr. Bilmes in -- I think I cite this somewhere else in my report as well.
So, again, my analysis is looking at all of the data produced by Anthropic. And Dr. Bilmes claimed

300

in his report that either with a name -- was a name [redacted] -- number .avro file is the final mix that you're referring to?
BY MS. MEHTA:
Q. Any training mix for any final Claude model, are you able to quantify the precise number of instances of the lyrics to the works in suit in that training mix, if any?
MR. HOWELL: Objection. Form. Vague. Asked and answered.
THE WITNESS: I do not know which data is final mix. I know that the data -- all of the data collected by Anthropic to train their different versions of the AI models and there is a mapping file detail which data set was being used for which model. And there may again or may not be deduplication of this data.
And to extent -- as Dr. Bilmes also conceded -- these dedupe methods are not perfect. They are far from perfect. And my analysis, despite these limits -- I've searched only a fraction of the dataset -- found a large number of -- were consumed in the environment that was -- that was was produced in.
BY MS. MEHTA:
Q. Again, I understand all of those points that

313

And I also want to point out that for this analysis, I also made sure that the CMI is associated with the correct song. For example, if, say, a Coldplay song -- I'm not sure which publisher it was, just say it was Universal -- I will only search the name Universal on a match of Coldplay.

So I believe these steps I've taken minimize the false positives for this case, but I do believe because of these strict requirements and the fact that I'm only counting five publishers out of many, this is an undercount of what's the, you know, correct number of CMI attached.

BY MS. MEHTA:

Q. Are you able to quantify how much a false positives and how many false negatives are associated with your [redacted] instances of CMI?

MR. HOWELL: Objection. Form. Vague. Compound.

THE WITNESS: So generally when developing an automated system, whether it's machine learning or whether it is detecting of these methods, what a developer typically do is find a threshold, find a methodology, review the test data. In this case, I did review the test data of a dozen examples and all of the examples I reviewed are correct version of that case.

314

So based on the validation, I believe the false positive is zero. However, was a false positive in -- generally when deployed in the wild, whether it's AI system or whether it is this system, I do not believe there is a methodology where you can tell definitively what is the real world false positive.

BY MS. MEHTA:

Q. Is it your testimony that based on manual review of a dozen of the [redacted] instances of CMI, you were able to rule out false positives?

MR. HOWELL: Objection. Form. Vague. Misstates prior testimony.

We also have to take a break, Counsel. It's been an hour and 15.

THE WITNESS: I believe based on standard evaluation processes, false positive is one calculated when developer and engineer develop an algorithm. And there's also the false positive in the real world when you deploy. Data may change, and systems that work well today may not work on data tomorrow.

And based on my testing of my internal data, which, again, is a random sample of the surmise, I have not seen false positives, and I do not have an opinion on -- while I do not know the quantities -- exact quantity of false positives in running through terabytes of data and -- yes.

315

MS. MEHTA: Let's take a break.

THE VIDEOGRAPHER: Stand by. We are going off the record. The time is 4:35 p.m.

(4:35 p.m., off the record.)

(4:47 p.m., on the record.)

THE VIDEOGRAPHER: We are going back on the record. The time is 4:47 p.m.

BY MS. MEHTA:

Q. If you could look at paragraph number 30 of your reply report.

Oh, actually, I'm sorry. Before you look at that, let me ask you another question.

In your -- either your original or your rebuttal or your reply CMI analysis, did you quantify the number of web pages that include lyrics to the works in suit that have no CMI?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: I believe I tested once -- it does include CMI. From what I -- you know, I'm not giving legal opinions here, but from what I understand about licensing generally if you license lyrics or books from publishers, you generally are required to attach CMI. And this also probably applied to sublicensees, which some of the websites are from.

316

So to the best of my knowledge, when something's posted on a website without CMI attached, I would -- my best guess would be those are unlicensed copies, which I have seen cases of, you know, lyrics on Reddit or random blogs where people post these things, and I suspect those would be the cases where CMI are not attached.

BY MS. MEHTA:

Q. So my question was not whether you think that might exist or why that might exist.

My question was: Did you actually count, did you actually quantify the number of pages that include lyrics to the works in suit that do not have CMI?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: Again, I do not -- I did not count specifically for the ones you just mentioned, and, again, I believe to the best of my knowledge that those are likely to be unlicensed copies that people post online.

BY MS. MEHTA:

Q. And then in terms of your opinions about, you know, the best of your knowledge where websites have licenses to lyrics and their requirements associated with CMI, to be clear, you have not reviewed any of the license agreements between the publishers and any lyric

317

websites as part of your assignment in this case; isn't that correct?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: I have not reviewed the terminal service or the license agreements between licensees and the publishers or sublicenses with licensees, and some of these may touch on privilege information with my communications with the counsel. So I'm not going to talk about those.
BY MS. MEHTA:
Q. Okay. You're not -- the stuff that you said about privileged information, communications with counsel, just to be very clear, you are declining to answer on the basis of work product or privilege communications you've had with counsel about the licensed terms between the publishers and lyric websites?
MR. HOWELL: Objection. Form.
THE WITNESS: That is correct.
BY MS. MEHTA:
Q. Let's go to paragraph number 30.
A. Okay. I'm there.
Q. And with respect to paragraph number 30, you say, "If CMI is present in the base training data and then the model outputs those same lyrics without the

318

associated CMI, it reasonably follows that the CMI must have been removed at some point in the training process."
Do you see that?
A. I do see that paragraph.
Q. Okay. What I don't see here is any instance in which you found the lyrics for a particular work in suit in the base training data and then model output with those lyrics without the associated CMI.
That's not cited here, is it?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: Can you repeat your question?
MS. MEHTA: Yes.
BY MS. MEHTA:
Q. Yeah. You have not cited any example or instance in which the lyrics for a particular work in suit were present in the base training data and then the model output those lyrics without the CMI; correct?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: So I believe the argument I'm trying to make here and the evidence is that based on my initial analysis as detailed in my opening report, the majority or the vast majority of the lyric I find, whether it is part of the work in suit or some other lyric I saw, have CMI attached right after the lyric.

319

And it seems to be my general experience on the internet as well, lyric has CMI attached.
And in my -- in different parts of these reports output, I have -- as you already discussed some of them -- I have repeatedly asked Claude to generate lyric, to generate sometimes verbatimly, sometimes new songs. In none of these cases have seen a single instance where CMI is generated out.
And the fact that CMI are overwhelmingly present in the training data and plus the fact that I did not see any CMI throughout different parts of my analysis, it shows to me that the CMI are either removed before training into the final model or the model outputs, there is algorithm to detect possible generating of CMI either got trimmed off or processes that instruct the model not to generate CM. And I have not reviewed the model inference and training with respect to this prompt.
BY MS. MEHTA:
Q. Okay. Is it fair to say that because you did not review the model inference and training code with respect to that processing data, you cannot identify any code in which CMI is affirmatively removed?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: I was not produced with specific

320

code to remove CMI, and one evidence I have with respect to this is I believe from one of the documents I've reviewed -- I forgot which one -- it talks about Anthropic engineers talking about using this technology using Newspaper3k to process their data.
To the -- to what extent Anthropic actually used this technology for the model, I do not know. But I have strong evidence to show that it it may have used because engineers talked about such technology.
BY MS. MEHTA:
Q. Did you receive or review any code associated with inference or training where there's specific code for removal of CMI? Yes or no?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: I'm sorry. Can you say that again?
MS. MEHTA: Yes.
BY MS. MEHTA:
Q. Did you receive or review any code where there was specific code associated with removal of CMI? Yes or no?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: Well, if there are lots of evidence, as I have discussed before, where CMI's overwhelmingly present in the data collected and how it

321

is not present in the final training output -- and I believe when you train the LLMs, you know, LLM doesn't understand when the lyrics stop, when the CMI start.

So based on the -- you know, predict next token, training methodology the LLM rely on, it is -- it should be likely -- very likely for CMI to be predict out from the model once it detects the lyric's sending, which is not the case from my inspection from different part of this about how Claude model reports -- how Claude works.

And -- and my analyze -- my understanding that Anthropic may use Newspaper3k, which is a method that's specifically designed to take a white page and only extract the main content of the white page, which include removes removing the footer, the headers, and potentially any additional information that is not part of the main contents.

And, you know, whether CMI is deliberately removed or accidentally removed because of the use of Newspaper3k, I do not know. I have not done an independent analysis of the case.

BY MS. MEHTA:

Q. And you haven't seen any code in the code that was made available to you in the inspection environment that confirms that Anthropic actually used Newspaper3k;

322

correct?

MR. HOWELL: Objection to form. Vague.

MS. MEHTA: Your analysis -- your testimony is based on the document you saw in that, not on the source code; correct?

MR. HOWELL: Objection. Vague.

THE WITNESS: So I believe the document I saw is -- if I remember this correctly, was a conversation between Anthropic engineers about using this tool. And if I remember correctly again, that the conclusion was affirmative on using this as a tool. And I believe those are leadership or, you know, senior engineers in the team that would have contributed to the source code.

And I was not produced the exact line-by-line usage of Newspaper3k in my inspection environment, and I was not, again, produced with any code to the best of my knowledge to analyze or clean the data before training.

BY MS. MEHTA:

Q. Does Anthropic's tokenizer have a token that includes the copyright symbol?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: Again, I do not know, and that is -- you would -- that would depend on the specific version that was being used, the specific data set that was being used.

323

But based on my understanding of LLMs and based on my understanding that Anthropic used or did not completely reinvented the view of the tokenizer, I believe most of the public tokenizer does encode special characters, does encode copyright sign.

BY MS. MEHTA:

Q. Could you go to paragraph 48.

Okay. At paragraph 48 you talk about some additional experimentation that you did in the context of this reply report; correct?

MR. HOWELL: Objection. Form.

THE WITNESS: Are you referring to 48?

MS. MEHTA: Yes.

THE WITNESS: Can you repeat your question, please.

MS. MEHTA: Yes.

BY MS. MEHTA:

Q. You conducted some additional experiments with respect to the guardrails in the context of this reply report; correct?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: I have done additional experiments as detailed in 48 paragraph on my reply report. These are new experiments I did for the reply report.

324

BY MS. MEHTA:

Q. Okay. And in that experiment you tested prompts for 10 works in suit, correct -- sample of 10 works in suit?

MR. HOWELL: Objection. Form.

THE WITNESS: I believe that is the case.

BY MS. MEHTA:

Q. How did you pick that sample of 10 works in suit?

MR. HOWELL: Objection. Form.

THE WITNESS: I randomly picked these songs from the list of songs that were produced to me.

BY MS. MEHTA:

Q. Your testimony is that you randomly picked these songs?

MR. HOWELL: Objection. Form.

THE WITNESS: That is what I recall.

BY MS. MEHTA:

Q. Did counsel suggest to you what songs you should test?

MR. HOWELL: Objection. Form. Asked and answered.

THE WITNESS: I recall that I chose these songs randomly by myself.

BY MS. MEHTA:

341

THE WITNESS: Yes. As I stated, these tools exist, and they are not -- they are downloaded and forked by hundreds of -- or tens of -- or thousands of users sometimes. And people make changes to these tools.

And based on my understanding of how software development works, there are -- lots of people use, you know, these prompts. And also from my experiment with research, a lot of researchers from different academic groups, from internal Anthropic are using these prompts to test their system, publish these papers. So I believe lots of people are reading this.

BY MS. MEHTA:

Q. I'm not asking for your speculation. I'm not asking about what you think is happening from the research papers. I'm asking about actual users of Claude -- not AI generally -- but Claude itself.

Let me ask the question more specifically.

Can you point me to a single document or piece of data in this case that suggests that any user of Claude not associated with the publishers has ever prompted Claude using the prompt structures that you used for your experiments? Specific documents or data to evidence that.

MR. HOWELL: Objection. Form. Vague. Asked

342

and answered.

THE WITNESS: Again, I think for -- well, I guess clarification question.

Do you narrow the scope to only documents and people involved or retained by -- in this case, or are you referring to in general?

BY MS. MEHTA:

Q. No. I'm asking -- I want you to exclude the people that are working for the publishers in this case.

A. Okay.

Q. So I'm not asking about the lawyers or their investigators or their experts.

Outside of the people working for the publishers in this case, can you identify any specific example from the documents and the data in this case that you reviewed of any user actually using these prompt formats?

MR. HOWELL: Objection. Form. Vague. Asked and answered.

THE WITNESS: When you say documents in this case, do you mean all of the confidential documents or do you also mean potential public documents on the cloud or online research papers related to this case?

BY MS. MEHTA:

Q. Any specific example of a prompt that

343

used -- how about this, I can narrow it.

The prompt formats that you used, can you identify any instance of anyone associated with the publishers ever using those prompt -- anyone not associated with the publishers ever using those prompt formats to search for lyrics or the works in suit specifically?

MR. HOWELL: Objection. Form. Vague. Asked and answered.

THE WITNESS: I believe we talked about this earlier in the conversation where I have seen a lot of academic papers and I have met some of these people at conferences.

And based on the papers they've tested their methodology, which are very similar to some of mine. And that tested in general corpuses -- or large corpuses of internet data. And, again, based on my inspection, there are large amount of work in suit as part of internet data, so I believe they was test based on what I understand in their research papers and publishing to include the code and sometimes data as part of the results.

BY MS. MEHTA:

Q. I'm asking about the outputs specifically, not training data.

344

Can you identify any output not associated with the publishers in which the prompt was based on the prompts you used, the same prompt format you used, for lyrics or for anything related to the works in suit?

MR. HOWELL: Objection. Form. Vague. Asked and answered.

THE WITNESS: Again, a lot of the research papers, a lot of blogs talk about very similar techniques. General internet data I do not -- I have not done independent analysis on their exact results.

But based on my understanding that these work in suit appear frequently on the internet data and there are many people with different backgrounds who are using similar technique, I believe there are incidents where work in suit will be tested using similar type of prompts.

BY MS. MEHTA:

Q. Can you identify any specific example of that under oath today?

MR. HOWELL: Objection. Form. Vague. Asked and answered.

MS. MEHTA: I'm not asking for speculation about what some other researcher did. Any specific example you can identify under oath today.

MR. HOWELL: Objection. Form. Vague. Asked

349

songs. I believe the analysis, based on what I reviewed in the research paper were -- which were from respected institutes, respected researchers, they did comprehensive analysis of the internet data.

And based on my internal review of the likelihood of Claude songs appears in the internet data -- and, again, I believe it would be an understatement to say that some what I've tested is likely that many people across these analysis has tested multiple times during their analysis of the internet data of similar type of prompts for verbatim extraction.

BY MS. MEHTA:

Q. For the 499 works in suit?

MR. HOWELL: Thank you, Dr. Shan, for your time.

MS. MEHTA: You're not going to let him answer that question just to -- I'm clarifying the answer he gave.

MR. HOWELL: I think we're three minutes over time now. I've given you a lot of follow-ups.

The deposition's over at this point.

Thank you, Dr. Shan.

MS. MEHTA: Okay. Are you -- are you going to have any questions?

MR. HOWELL: No, I don't have any questions.

350

Thank you for your time, Dr. Shan.

MS. MEHTA: Before we go off the record, I want to mark the transcript highly confidential, attorneys' eyes only.

MR. HOWELL: That's a good point.

And I -- obviously we'll reserve the right to have Dr. Shan review the transcript and make any edits.

THE STENOGRAPHER: I'm sorry. What was that?

MR. HOWELL: Oh, I reserve the right for Dr. Shan to review the transcript and make any edits that -- to correct any statements of his testimony.

THE VIDEOGRAPHER: Stand by. This marks the end of the deposition. We're going off the record. The time is 5:30 p.m.

(5:30 p.m., deposition concluded.)

351

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

CONCORD MUSIC GROUP, INC., ET AL.,
Plaintiffs,
-vs-
ANTHROPIC PBC,
Defendant.
Case No. 5:24-cv-03811-EKL SVK

I, SHAWN SHAN, Ph.D., state that I have read the foregoing transcript of the testimony given by me at my deposition on March 3rd, 2026, and that said transcript constitutes a true and correct record of the testimony given by me at said deposition except as I have so indicated on the errata sheets provided herein.

SHAWN SHAN, Ph.D.

No corrections (Please initial)
Number of errata sheets submitted (pgs.)
SUBSCRIBED AND SWORN to before me this day of , 2026.

NOTARY PUBLIC

352

CERTIFICATE OF REPORTER

I, ISAIAH ROBERTS, a Certified Shorthand Reporter (IL) and Registered Professional Reporter, being empowered to administer oaths and affirmations remotely pursuant to Section 30(f)(1) of the Federal Rules of Civil Procedure, do hereby certify:

That the foregoing proceedings were taken remotely before me at the time and place herein set forth; that any witness in the foregoing proceedings, prior to testifying, were placed under oath; that a verbatim record of the proceedings was made by me using machine shorthand which was thereafter transcribed under my direction; further, that the foregoing is an accurate transcription thereof.

I further certify that I am neither financially interested in the action nor a relative or employee of any attorney or any of the parties.

Further, that if the foregoing pertains to the original transcript of a deposition in a Federal Case, before completion of the proceedings, review of the transcript was requested.

IN WITNESS WHEREOF, I have this date subscribed my name.

Isaiah Roberts, CSR, RPR
Illinois CSR #084-00489

No. 621066

Re: Deposition of Shawn Shan
Date: 3/3/2026
Case: Concord Music Group -v- Anthropic
Return to: transcripts@planetdepos.com

Page	Line	Correction/Change & Reason
40	14	From "venerated" to "generated"; typo
43	12	From "verify the accuracy" to "verify with accuracy"; transcription error
45	15	From "something is accurate compared to" to "something is AI generated accurately compared to"; transcription error
72	22	From "date" to "data"; typo
82	24	From "ecstatic" to "aesthetic"; transcription error
95	3	From "source guard" to "source code"; transcription error
95	8	From "be done do improve" to "be done to improve"; transcription error
99	5	From "house" to "mouse"; transcription error
107	14	From "have integrated into perfect model" to "have integrated that into a model"; transcription error
109	10	From "inspected those date" to "inspected those data"; typo
118	14	From "me of the inspections" to "me on the inspections"; typo
147	6	From "incidence" to "incidents"; transcription error
161	2	From "doe" to "code"; typo
163	24	From "even have the exact date" to "even having the exact data"; transcription error
187	22	From "a common message" to "a commit message"; transcription error
189	13	From "Git common message" to "Git commit message"; transcription error
189	14	From "put Git common" to "put Git commit"; transcription error
215	17	From "ever" to "every"; typo
223	3	From "[redacted] typo
244	4	From "correctly checks" to "correctly checked"; transcription error
245	10	From "color" to "code"; transcription error
255	14	From "glasses" to "classes"; typo
262	13	From "hear" to "here"; typo
291	14	From "Cleo" to "Clio"; transcription error
291	18	From "no" to "not"; typo
297	2	From "Cleo" to "Clio"; transcription error
299	10	From "produced to be" to "produced to me"; typo
302	4	From "denotes" to "denoted"; typo
312	12	From "it's rarely have false" to "it rarely has"; transcription error
312	14	From "copyrighted" to "copyright"; typo
313	22	From "developer typically do" to "developer will typically do"; transcription error
317	4	From "terminal" to "terms of"; transcription error
319	16	From "CM" to CMI"; typo

321	15	From "include removes removing" to "include removing"; transcription error
327	5	From "that are the cases I seen" to "those are the cases I've seen"; transcription error
345	6	From "nd the fact" to "And the fact"; typo

4/10/2026
(Date)

(Signature)

No. 621066

Re: Deposition of **Shawn Shan, Ph.D.**
Date: 3/3/2026
Case: Concord Music Group -v- Anthropic
Return to: transcripts@planetdepos.com

ACKNOWLEDGMENT OF DEPONENT

I, Shawn Shan, Ph.D., do hereby acknowledge that I have read and examined the foregoing testimony, and the same is a true, correct and complete transcription of the testimony given by me and any corrections appear on the attached Errata sheet signed by me.

4/10/2026 _____

(Date) (Signature)