EXHIBIT 7

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

173

training mix, I would have definitely paid attention to it.
BY MS. MEHTA:
Q. Is there any other nonprivileged testimony that you want to offer to explain what you did to look for the final training mixes?
MR. HOWELL: Objection. Form. Vague.
THE WITNESS: No.
BY MS. MEHTA:
Q. Did you ask for any help? You said there were points in time where Anthropic engineers helped you during the inspection. Did you ever ask any Anthropic engineer, either directly or indirectly, to help you find the final training mixes in the inspection environment?
MR. HOWELL: Objection. Form. Vague.
Same caution.
THE WITNESS: So for Anthropic engineers, all the conversation is constrained to accessing the machine or technical difficulty of -- of accessing machine and installing packages that we already talked about.
BY MS. MEHTA:
Q. Okay. Could you go to paragraph number 45.
A. I'm sorry. You said 45?

174

Q. Yes.
A. Okay.
Q. So in paragraph number 45 you talk about your inspection of Anthropic's training data repositories and the fact that you say you identified several distinct categories of data used to train Anthropic's Claude model.
Do you see that?
MR. HOWELL: Objection. Form.
THE WITNESS: Which part of this? It's a pretty long paragraph.
MS. MEHTA: First sentence.
THE WITNESS: Oh, okay.
Yep.
BY MS. MEHTA:
Q. Okay. And then you say, "I navigated the file directory structure of Anthropic's training data repositories and observed that the folder names in those repositories identified in general terms the source and nature of the data contained within."
Do you see that?
A. Yes.
Q. Is that accurate?
MR. HOWELL: Objection. Form.
THE WITNESS: So as I stated before, there are

175

many datas that I recognize, whether it's Common Crawl or Common Crawl-News or the Pile. And for those datasets I performed the validation. As I said, I look at 10. Sometimes I look at more than 10 samples from these folders.
But then there is also some of dataset that is labeled that doesn't mean anything to me. And for those cases I took a look at some of these data, but there are too many folders for me to look at, you know, every single folder.
BY MS. MEHTA:
Q. Right. You say in the next sentence, "To verify the reliability of these labels, I opened and examined 10 random samples from folders with names that indicate hosting possible lyrics."
Do you see that?
A. Yes.
Q. And then you say, "In every instance the content of the files was consistent with the dataset label on the folder."
Do you see that?
A. Yes.
Q. Is that what you mean when you say you opened files -- opened the folders with names you recognize, like Common Crawl or the Pile?

176

MR. HOWELL: Objection. Form.
THE WITNESS: For the initial inspection, yes. I look at, you know, different folder. For example, if there is source codes, I did not really pay attention to some of those folders.
BY MS. MEHTA:
Q. Okay. Is it fair to say that you did not do any analysis of the final training mixes for the Claude models at issue in this case?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: Before I answer, can we do a lunch break after this line?
MS. MEHTA: Yeah. Let me -- I'll wrap up in just a minute, and then we'll take the lunch.
THE WITNESS: Okay. Can you repeat that question?
MS. MEHTA: Yes.
BY MS. MEHTA:
Q. Fair to say that you did not do an analysis of the final training mixes for the Claude models at issue in this case?
MR. HOWELL: Objection. Form. Vague. Calls for speculation.
THE WITNESS: So my inspection is the data

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

179

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.

301

you want to make. My question is actually just a specific, factual question.

Are you able to quantify for the judge and the jury in this case for any final training mix for any commercial Claude model the exact number of instances of lyrics to the works in suit, if any?

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

THE WITNESS: As I stated, I do not know or was not produced which final mix was used, and I found millions in the data in the environment that was produced. And based on my knowledge and background of the case, I believe most of the data are collected by Anthropic to train their models.

BY MS. MEHTA:

Q. How about we go about it this way, let's pick a model. Claude 3.7. How many -- I'm asking for an exact number -- how many, if any, documents or instances -- let me start that question again.

For Claude 3.7, can you quantify how many instances, if any, of the lyrics to the works in suit exist in the final training mix for that model?

MR. HOWELL: Objection to form. Vague. Asked and answered.

THE WITNESS: For Claude 3.7, I was produced a

302

set of data and it marked as used for training the model. I found -- I do not recall the exact numbers for 3.7 specifically, but I found a large number of works in suit in the data that marks -- or as denotes as 3.7 used for training.

BY MS. MEHTA:

Q. Okay. And you don't know whether or not there was deduplication after that; right?

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

THE WITNESS: Again, I reviewed the data as produced. I do not know any processing that may or may not be on the data set, but I believe these are the data as denoted by the inspection department used to build, to train, to optimize these models.

There may be, as we discussed, tokenization and processing involved, which are standard process, and I have found, again, hundreds of thousands of copies of these lyrics across all different models. And I do not know the exact numbers for 3.7. And even if -- even for the numbers I put in my report, I am confident these are undercounts or conservative estimates because of the constraint I had with inspection.

BY MS. MEHTA:

Q. Now, the numbers you put in your reports,

303

those are based on the data that you looked at in the inspection environment; correct?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: It is based on the search script that I detailed in my opening report. I -- it is the output and the number of the search script.

BY MS. MEHTA:

Q. And you don't know whether -- if there were further processing or deduplication after the data that you looked at, the number would go up or down? You don't know that, do you?

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

THE WITNESS: I believe I find a huge number of copies even when searching a fraction of the data, given the constraint I have.

And there's a chance that certain data points may get removed, but as Dr. Bilmes, as I discussed, deduping at this scale across websites for lyrics is very challenging. And given the fact that my analysis is already an undercount, I believe there is a large number of copies of work in suit in the final training mix for the one that was used to train for Claude models.

BY MS. MEHTA:

304

Q. Can you quantify with specificity how much data was removed from the data that you looked at to the final training mixes for any model of commercial Claude?

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

THE WITNESS: Again, I was not produced this information with respect to the cleaning of Anthropic data. I do not know whether the data produced to me is all of the final mixes or there are collected for the purpose of training these models.

And, again, I -- based on the data I was provided, I was able to find a large number of copies.

BY MS. MEHTA:

Q. Did you -- actually, let me ask the question this way: We talked earlier about LLMs generally. Would you agree with me that LLMs generally are trained on tokenized text, not human readable text?

MR. HOWELL: Objection. Form. Vague.

THE WITNESS: For the LLMs that we have today, I believe there is a tokenization process, and there's many ways to do this. And some of the other models can, in fact, read -- or read to the extent as computer sites of English text.

But I believe the ones we -- in this case are trained on tokenized data based on the information

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