Exhibit D



Planet Depos®
We Make It Happen™

HIGHLY CONFIDENTIAL
ATTORNEYS' EYES ONLY
SOURCE CODE

Transcript of Jeffrey Bilmes, Ph.D.

Date: March 6, 2026
Case: Concord Music Group -v- Anthropic

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

1

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

- - - - - - - - - - - - - - -x
CONCORD MUSIC GROUP, INC., ET AL., : Civil Action No.
Plaintiffs, : 5:24-cv-03811-EKL
v. : SVK
ANTHROPIC PBC, :
Defendant. :
- - - - - - - - - - - - - - -x

Videotaped Deposition of JEFFREY BILMES, Ph.D.
* * * HIGHLY CONFIDENTIAL * * *
* * * ATTORNEYS' EYES ONLY - SOURCE CODE * * *
Seattle, Washington
Friday, March 6, 2026
9:00 a.m.

Job No.: 620607
Pages: 1 - 255
Reported By: Connie Recob, CCR 2631, RMR, CRR

2

Deposition of JEFFREY BILMES, Ph.D., held at the offices of:

Cooley LLP
1700 Seventh Avenue, Suite 1900
Seattle, Washington 98101

Pursuant to notice, before Connie Recob, CCR 2631, RMR, CRR, Notary Public in and for the State of Washington.

3

A P P E A R A N C E S

ON BEHALF OF PLAINTIFFS:
COREY MILLER
OPPENHEIM + ZEBRAK, LLP
4530 Wisconsin Avenue, Fifth Floor
Washington, D.C. 20016
(202) 957-5043
corey@oandzlaw.com

TIMOTHY CHUNG (via videoconference)
OPPENHEIM + ZEBRAK, LLP
461 Fifth Avenue, 19th Floor
New York, New York 10017
(212) 951-0521
timothy@oandzlaw.com

ON BEHALF OF DEFENDANT:
TAYLOR GOOCH
WILMERHALE
50 California Street, Suite 3600
San Francisco, California 94111
(628) 235-1026
taylor.gooch@wilmerhale.com

4

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

ON BEHALF OF DEFENDANT:
ELLEN STEPHANO
WILMERHALE
60 State Street
Boston, Massachusetts 02109
(617) 526-6375
ellen.stephano@wilmerhale.com

ALSO PRESENT:
Parker Simpson - Videographer
Ian Chen - Anthropic PBC

Pages 21-24

21

A Not -- so a general purpose LLM, can I ask a clarifying question about that?
Q Yeah.
A So by "general purpose LLM" in that particular case, one as I mentioned in the answer to your previous question, these generally are very large scale with many billions of parameters.
So I have never worked on producing the kinds of models that are at that scale that the companies, especially like OpenAI and Google and others do. So I have never participated in building a model that's as general purpose as those are general purpose only because the size of the model is much smaller that my students and I worked on, and the amounts of training data that my students and I worked on was much smaller. So I can't say that I've worked on things at that scale.
Q Have you built large -- any large language models at a smaller scale?
A It depends precisely on what one means by "build." So what I've done is with my students, I have worked on partial -- so there are models that one can use that are open-weight models, and what my students and I have done is adjusted the

22

weights of those models and adjusted the processing of those models to observe the behavior, but those models are not at the scale as the large commercial models that are widely used today.
Q Have you ever been involved in the process of selecting the training data for a large language model?
A May I ask a clarifying question?
Q Uh-huh.
A So when you say "large language model" in this case, are you referring to the extremely large language models of the form that you would encounter if you were to go to one of the recent companies that commercializes these models or are you referring to the small scale, the small large language model, for lack of a better word, that my students and I have worked on?
Q Well, let me ask it both ways then. Have you ever been involved in the process of selecting the training data for a general purpose large language model on the scale of the ones that have been commercially released by OpenAI, Google and Anthropic?
A No.

23

Q Have you ever been involved in the process of selecting the training data for a large language model on a smaller scale?
A So some of the academic research that we have done involved testing the behavior of changing these smaller scale models and seeing primarily for computational efficiency reasons the effect to which one could improve a small large language model by adjusting it using a smaller amount of data.
That was similar to some of the work that I had mentioned earlier regarding trying to improve the efficiency of training, machine learning models in general, not just large language models.
Q Have you ever built an internet crawler?
A No.
Q Do you have an understanding of what an LLM defense system is?
A I'm sorry. Could you please repeat the question?
Q Do you have an understanding of what an LLM defense system is?
A An LLM defense system, you're asking if I have an understanding of what they are?

24

Q Uh-huh.
A It depends on precisely what you mean by an LLM defense system.
Q What do you understand that to mean?
A Well, I'm not certain what you mean by an LLM defense system.
Q What do those words mean to you?
A Well, it could be a system that defends an LLM from malicious attack of some sort.
Q Have you ever been involved in building a system to defend an LLM against malicious attacks?
A No.
Q Are you a lawyer?
A No.
Q You're not an expert on copyright law?
A No.
Q Do you play any musical instruments?
A I do.
Q What do you play?
A I play bass, jazz bass.
Q Do you have any formal musical training?
A I've had some musical training, but I am not sure I would call it formal.
Q Have you ever taken any university-level classes in music?

Pages 41-44

41

[redacted]

Do you see that?

A I do see that.

[redacted]

A That's right.

Q Is it -- so it's not poss- -- withdrawn.

Is it possible to analyze the final AVRO training data files you reviewed to see whether they contain any of the plaintiffs' works in suit?

ATTORNEY GOOCH: Objection. Form.

42

[redacted]

Now, any encoding of any form of data which includes this form has an encoder and there would be a decoder. If the decoder was available, then it would be possible to convert that back into ASCII tokens as opposed to just the pure byte encoded tokens. But without actually having that decoder, it would not be possible.

BY ATTORNEY MILLER:

Q You're familiar with the concept of hallucination in the context of LLMs?

A I'm familiar with the concept and how people have used it.

Q What is hallucination in the context of LLMs?

A So in the context of hallucination, hallucination is something that is generally undesirable and it is something that most LLM designers would want to minimize as much as possible. What hallucination entails is when the model outputs something that's not factually correct.

Q Why do LLMs hallucinate?

43

A There are many possible reasons, but that is a great question and it is something that is under a lot of study right now, both in the commercial and the academic world.

Q Is there any consensus in the -- in the field about why LLMs hallucinate?

A It is fair to say that there are multiple consensuses, meaning that there's no universal agreements as to why it happens necessarily.

Q How many different consensuses or schools of thought are there on why LLMs hallucinate?

A I don't know the exact counts.

Q What are the main theories?

A The main theories of why hallucination happens, one of them is that some of the models would hallucinate less if they were encouraged during the training process to express uncertainty when they have uncertainty.

So, for example, if a model is encouraged to just guess when it doesn't know something, then a model might guess incorrectly if it doesn't know something versus if a model is trained so that when it doesn't know something, it's willing to say I don't know, which is something that people do when they don't know something, then a good

44

number of people believe that that would reduce hallucinations.

Q Are there other theories of hallucination that have reached some kind of a consensus?

A So the one I just gave to you is the one I'm personally partial to, but there are others, yes.

Q And what are the others?

A I'm -- I can give you maybe one, possibly two more, but I -- it would take me a bit of time to remember all of them.

Q Okay. Why don't we start with the -- those two.

A Or one?

Q Yeah.

A Okay. So others think it might be something related to the fact that LLMs are stochastic engines, that they are inherently generating text stochastically which means every time you use an LLM even with the same prompt, you usually get a different answer when you use an LLM typically.

Q And how does that theory argue that the stochasticity of LLMs leads to -- leads to hallucinations?

Pages 129-132

129

include things like Dolma, which is an open source tool, and other open source tools so it could be the case that they have. I would be surprised if there was no engineer that has never adopted these tools or at least tested them, again assuming they were appropriately licensed, but I don't know for certain precisely which of these tools, which engineer who works or has worked at Anthropic has ever tested, tried out or possibly adopted.

BY ATTORNEY MILLER:

Q In Paragraph 113, as you mentioned, you cite to Anthropic's -- excuse me. You cite to the deposition testimony of Anthropic witnesses; is that right?

A So in Paragraph 113, Paragraph 113 has a relatively large number of footnotes, some of which refer to -- give me a moment -- various different notes and interrogatories and depositions including some from Ben Mann as well as other related documents, for example, a document that explains the concept of what a bag of words, set of features is.

Q Did you do anything to confirm whether Anthropic's witnesses in the deposition testimony you cite in this paragraph have accurately

130

described Anthropic's data cleaning and processing processes?

ATTORNEY GOOCH: Form.

THE WITNESS: Doing such would not have affected my conclusions in any way, and the reason is because of my background in machine learning and artificial intelligence. And my background in artificial intelligence and machine learning that's relevant to me coming to the conclusion that it wouldn't have affected my conclusions in any way is that I know that a lot of the things that are mentioned in Paragraph 113 or things like the open source tools mentioned in Paragraph 112 or just describing the general processes of data cleaning and data preprocessing in Paragraph 111, I just know that those are a necessary element of producing a functional resulting machine learning model.

And given that in this particular case, I know that Anthropic's models are functional general purpose models that are capable of doing an enumerable number of things, it just doesn't seem reasonable to me to think that there would be any chance that they could possibly have achieved what they've achieved without putting significant

131

efforts, significant -- quite significant efforts into data cleaning and data preprocessing.

BY ATTORNEY MILLER:

Q That didn't answer my question at all --

A Apologies.

Q -- so I'll ask you to please answer my question.

Did you do anything this confirm whether Anthropic's witnesses in the deposition testimony you cite in this paragraph have accurately described Anthropic's data cleaning and processing processes?

ATTORNEY GOOCH: Objection. Form. Asked and answered.

THE WITNESS: Yeah, I apologize if that seems like I didn't answer your question, but I want to make sure that I'm providing good and accurate answers.

So my answer is that based on my experience in machine learning and artificial intelligence over many decades and having experienced what it takes in order to produce one of these models, one of these kinds of models, a functional general purpose machine learning model of a variety of different sorts, I know that what

132

is necessary to do that is to have good clean data so that you don't experience the phenomenon of garbage in, garbage out and as a result of that, and a result of my conclusion that it's not reasonable to not do these things, the answer to your question which is -- which is what I was saying a little bit earlier is that I did not -- I do not believe it is necessary to do that because it would not have changed my conclusion.

It wouldn't have been anything that would have provided new information and it wouldn't have ended up changing anything that I concluded in any of these three reports.

And, again, I want to apologize because if it seems like I'm not answering your question I'm happy to -- to try again if you're interested.

BY ATTORNEY MILLER:

Q Did Anthropic filter data for Claude 2 the same way it did for Claude 3.5?

A Those are -- may I ask precisely the years and the dates when those models were active and developed?

Q Do you know when -- the years when those models were active and developed?

ATTORNEY GOOCH: Form.

Pages 137-160

137

many different actual meanings at multiple levels of granularity.

So generally, the notion of concept along with the notion of granularity, along with the notion of the degree of semanticity, namely is it monosemantic, polysemantic, and if it's polysemantic, then how polysemantic is it. All of those things are things that are fairly regularly discussed, not just in the sparse autoencoder literature which is a very small subpart of the neural network literature which is a subpart of the machine learning literature, but it's generally discussed fairly widely.

And so in that particular sentence that you pointed out, the last sentence of Paragraph 116, that sentence is referring specifically to the idea of that redundancy and imbalance is bad. The sentence reads, "No single important concept or linguistic structure should be appreciably overrepresented."

So the idea of overrepresentation is the same idea as duplication. If something is duplicated many times, it's overrepresented. There's too much representation of that concept because -- and we -- the reason why we can say too

138

much it's because it ends up hurting the model.

The example of fruit that we were talking about earlier when we had the duplications of the banana. We had the good -- the two pictures of a banana that were still good and edible and the two pictures of the banana that were bad and should be taken off the shelf. So that would be an example of a concept namely in that particular case at that particular granularity. It's the concept of a fruit. In particular the concept of a banana, and it's not infeasible that in a neural model there could be learned a unit that corresponds to that concept of the banana. It could also be the case that that unit corresponds to aspects of parts of a banana like the concept of yellowness which normally is -- I mean, there are other kinds of bananas, there are red bananas, but the concept of yellowness would normally be something that would be associated with a banana.

So, therefore, it's completely fair in my mind to use the word "concept" when we're referring to various different aspects of machine learning, especially in the case when one -- when one wants to express notions of balance, when one wants to express notions of lack of

139

overrepresentation. When one wants to express notions of diversity and when one wants also to express notions of a dataset not having duplicates in it because, as we've established earlier, duplication generally makes the model perform worse.

So these are all various different ways of referring to certain -- certain things that are fairly similar to each other. But I do want to add that actually that's another good question and I really mean that.

Q In Paragraph 117, you state, "Anthropic employs substantial data filtering and de-duplication techniques to maximize its models ability to generalize beyond their pretraining data and to minimize memorization."

Do you see that?

A Sorry. Give me one moment. I'm sorry.

I see that, yes.

Q And then you have a footnote and you cite the deposition testimony of Dawn Drain.

Do you see that?

A This is Footnote 230 on Page 79?

Q Yes.

A I see that, yes.

140

Q Did you review any of the code for Anthropic's de-duplication processes?

ATTORNEY GOOCH: Form.

THE WITNESS: So as we were talking about earlier and the whole notion of de-duplication is this idea of data cleaning, and in order to produce a machine learning model that functions well, I know that data cleaning, data processing and data de-duplication is absolutely necessary and so therefore as a result, I didn't feel that it would have changed my conclusions at all. In fact, I knew it wouldn't have changed my conclusions at all because these are things that I know from my experience in machine learning are necessary to produce a functional model and so -- because I want to make sure that I -- that I properly answer your question.

So I felt I didn't need to review any specific de-duplication code because it wouldn't have changed my outcomes, my conclusions and so as a result I did not.

BY ATTORNEY MILLER:

Q Okay. Did you compare any of Anthropic's pre de-duped datasets to the purportedly post de-duped datasets to see what the effect of

141

Anthropic's de-duplication processes was?
ATTORNEY GOOCH: Form.
THE WITNESS: In my experience working for -- in machine learning and artificial intelligence for many, many decades, I have experienced firsthand the effect of redundancy removal or diversity enhancement of datasets in many, many different contexts. This includes examples in computer vision. It includes examples in machine -- statistical machine translation. It includes examples in predomics and genetics which are two branches of computational biology. It includes examples of digital pathology which is essentially the problem of trying to diagnose cancer images, H&E stained slides trying to detect cancer and also in many, many other datasets, for example, tabular data, things like housing, price datasets, machine tool wear datasets, financial datasets.
I have looked at many, many, many different kinds of data and in all of those cases over the years, I have found that the effects of de-duplication are highly beneficial. And in general, the effects of redundancy removal and diversity enhancements, the effects of all of that

142

is highly beneficial. That in my mind is a given. It's something that one does when one is developing a machine learning process. One needs to make sure you have good data. One needs to essentially process and analyze and clean, de-dup and --
BY ATTORNEY MILLER:
Q Professor, excuse me. But you're not answering my question.
ATTORNEY GOOCH: Why don't you let him finish.
ATTORNEY MILLER: Because he's not answering the question.
ATTORNEY GOOCH: Why don't you let him finish.
ATTORNEY MILLER: Because he's not answering the question.
BY ATTORNEY MILLER:
Q I will ask the question again which is a simple yes-or-no question. Did you compare any of Anthropic's pre de-duped datasets to the purportedly post de-duped datasets to see what the effect of Anthropic de-duplication processes was?
ATTORNEY GOOCH: Professor, why don't you finish your last answer to the question and then

143

Corey can ask his next question and you can answer that one, but don't feel like you can't finish your question or answer. That's what you're here for.
THE WITNESS: First of all, again, I want to apologize if it seems like I'm not answering your question. But what I honestly am wanting to do is give you a complete, clear and accurate answer to your question and I will come back to that specific point, but I think in order to answer that question accurately, it's necessary to provide context, and so I'm just providing the context in order to finally answer your question. But I will get back to that.
So -- so -- so continuing where -- where I was before, and I'll restate your question just so that you know that I'm thinking about your question. So your question is: Did I compare the effects of the de-duplication process both before and after de-duplication on the models performance; is that correct?
BY ATTORNEY MILLER:
Q Not on the model's performance. Did you look at the datasets both before the de-duplication is run and after de-duplication was

144

run to see what was removed during the de-duplication process?
A Okay. To see --
ATTORNEY GOOCH: Same -- real quick. Same objections.
THE WITNESS: Okay. To see what was removed from the -- Okay. So that question is in my mind.
So continuing what I was saying a little bit before, that -- and I'll just summarize. To my understanding of, and my experience, my firsthand experience, is that one cannot get a functional model unless one does successful proper de-duplication of data. And so it would not have changed my analysis and my conclusions and my opinions if I had looked at exactly what, in the original raw, unclean dataset was removed because of my experience in machine learning and data science and artificial intelligence.
It could not in my mind have been possible to produce a powerful general purpose model that's so good at so many different things if it wasn't the case that proper de-duplication wasn't done because the evidence to me in my experience is that it was done, the strong evidence. And, like

145

I said a little bit earlier, knowing that redundancy is bad in datasets. I can actually describe if you're interested an example of how it's possible to construct a dataset or I should say two different datasets that are very similar in size and even balance but one has redundancy in it and one that does not have redundancy in it, and the one that has redundancy in it produces a model that pretty much fails.

The one that is diverse produces a model that functions quite well even though many of these other characteristics of these two datasets, like their size and some aspects of them as well that say some, for example, machine learning students might think of are sufficient to think of these datasets as being the same, but I could give an example of that of how they would actually functionally be quite different.

So then getting back to your question, it was not necessary for me to form my conclusions in order to do that particular analysis that you're suggesting.

BY ATTORNEY MILLER:

Q In Paragraph 118 in the first sentence, you state, "Anthropic has experimented with

146

different approaches to data de-duplication to improve model generalization and limit memorization including both exact and final level duplication."

Do you see that?

A I'm just getting the context. I'm sorry. One moment. Okay. So this is in Paragraph 118?

Q Yes.

A Okay. And the question was: Do I see the first sentence?

Q Yes.

A I do see the first sentence, yes.

Q Do you know -- you did -- withdrawn.

You did not review the code for any of the de-duplication techniques Anthropic experimented with, did you?

ATTORNEY GOOCH: Asked and answered. Form.

THE WITNESS: So in my experience as a machine learning scientist and AI scientist who's been working in this area for some number of decades, I have come to know that if one doesn't do proper de-duplication of the data and other forms of data processing and cleaning, one gets a resulting model if it's trained on that data that

147

performs very poorly in a variety of different ways.

And so because of that knowledge and that firsthand experience that I've had, I did not think and I should say, I knew that it would not be necessary to look at that code because if it was not done properly, I know that the model would not be functional. It would not be a good general purpose model which is the same behavior that all machine learning models have.

Q Do you know whether the de-duplication techniques Anthropic used for Claude 2 were the same de-duplication techniques Anthropic used for Claude 3.5?

ATTORNEY GOOCH: Objection. Asked and answered.

THE WITNESS: So I know that the machine learning and artificial intelligence field is a rapid and dynamically changing field. Things are changing all the time. There are new techniques that are being introduced. There's new toolkits that are being introduced. There's new open source toolkits that are being introduced, new versions of open source toolkits and new capabilities within each one of those new versions

148

of those open source toolkits. And so what that means is -- is that rapid change is quite regular.

And so it's quite typical, for example, when one model -- when one develops a model, say, in a particular year that the set of tools available or the set of versions of tools available in a particular -- at a particular time that you're developing a model are likely to be different than the set of tools that would be available if one were to develop a model even six months later. And six months does not seem like that long, but it's quite a long time when things are moving so rapidly.

So as a result, I know that those change and so therefore it wasn't something that I felt was necessary to look at to reach my conclusions.

BY ATTORNEY MILLER:

Q Do you know specifically what the differences were, if any, between the de-duplication techniques Anthropic used for Claude 2 and the de-duplication techniques Anthropic used for Claude 3.7?

ATTORNEY GOOCH: Objection. Form.

THE WITNESS: So the specific differences between the de-duplication techniques, let me --

149

sorry. May I have one moment to just look over Paragraph 118?
BY ATTORNEY MILLER:
Q If you need to.
A Thank you.
Yeah, so I'm quite happy to talk about the precise mechanism of the Jaccard similarity. Jaccard similarity is something that's been used for quite some time even as early as the 1990s for measuring similarity between sets of features.
The specific things that were used that were different between these two versions that you just mentioned, what I know is because of the age of the Jaccard similarity metric, which is something that's measured, that's discussed in Paragraph 118, it's quite probable that that was something that was used in both, but I did not think it was necessary to -- and my conclusions would not have changed if I had gone into the specific differences between the de-duplication methods because Jaccard similarity is quite old. It's a quite old method. Now, old doesn't mean bad. Jaccard similarity is still a very powerful and very widely used similarity measure up until today, especially when you have objects that or

150

maybe sets of features that are amenable to being measured using the Jaccard similarity.
Q Did you do any analysis to determine how many copies of Plaintiffs' works in suit that are at issue in this case remained in Claude's training data after Anthropic's de-duplication processes?
ATTORNEY GOOCH: Form.
THE WITNESS: In the reports there was analysis of robots.txt and the robot exclusion protocol, and there were a number of things that that analysis did.
One of the things that it did was to see whether works in suit were available and permissibly crawled web pages.
BY ATTORNEY MILLER:
Q Yeah, I'm not asking about that.
A I promise you I will get -- answer your question appropriately. I -- is it okay if I continue with the context?
Q Please answer the question.
A Okay. So as part of that analysis, that robots.txt analysis -- I have your question in mind and I will swing back to the answer, but will fully answer your question.

151

So as part of the robots.txt analysis, it was used to determine whether the works in suit were available on permissibly crawled websites and that includes whether or not there was an actual robots.txt file present that allowed Anthropic's crawler to crawl as well as when there was no robots.txt present.
So one of the steps of the analysis was to include only websites that had robots.txt files but to see whether each lyric was permissibly crawled, and that analysis found that every one of the 499 lyrics was available on a permissibly crawled website namely one that had robots.txt.
So to get back to your question about de-duplication, it stands to reason that after de-duplication -- after the de-duplication process has happened and the de-duplication process also happened over the entire training data, [redacted]
So it doesn't really therefore matter how many works in suit for each lyric were present in the original pre de-duplicated data. Rather, it's that the de-duplication process doesn't remove the

152

last remaining entity because if it's a sole entity, it's not a duplicate any longer. So as [redacted] **I believe that answered your question.**
Q No, it didn't.
ATTORNEY MILLER: I'm going to move to strike that answer as nonresponsive and I'll ask the question again.
ATTORNEY GOOCH: Okay. There's no judge for you to move to strike anything, so I'm not sure why you're doing that. But since you've said that, I guess I'll have to respond which is that he did completely ask (sic) questions that you've been asking that you frame as a yes-no were clearly not yes-no and require context. The witness is appropriately providing that context and is obviously -- needs to provide that context in order to give you a truthful and honest and complete answer which is what his oath requires and what, I assume, you're not telling the witness not to do.

153

So, yeah, if you want to ask the question again and likely get the same answer, it's your seven hours and it's yours to do with as you want, but you won't be getting more time so be efficient.
BY ATTORNEY MILLER:
Q I'll ask the question a different way: Since you didn't examine any of Anthropic's code regarding its de-duplication processes, how do you [redacted] Claude?
ATTORNEY GOOCH: And I'll just note as the -- as Counsel indicated, he's clearly aware that the question you just asked is identical but asked in a, quote, "different way." So his objection is asked and answered, and I think that Counsel's using his time inefficiently, but he's free to do it however he wants and if he gets the same answer, here we are.
THE WITNESS: So, first of all, let me premise that I honestly truly am trying to answer your questions truly accurately and faithfully. And I honestly -- honestly, I really am saying this truthfully that these are good questions.

154

So how do I know? Let me repeat the question so that you know that I have your question in mind and that I'm thinking about that question in formulating this answer.
So your question is: How do I know that [redacted]? That might not be the exact word, but I believe that's a paraphrase; is that correct?
BY ATTORNEY MILLER:
Q Yes.
A Okay. The reason is because of what I was saying a little bit earlier which is that duplication hurts the model performance, and duplication could mean duplicating it twice or three times or ten times or duplicating it once. When you duplicate something once, it occurs two times. When you duplicate something twice, it occurs three times.
So how do I reason that it should be the case that [redacted] after the de-duplication process? It's, again, falling back on my years of working in machine learning and artificial intelligence where I've come to observe firsthand that duplication, even if it's only a little bit of duplication, has a

155

detrimental effect on the model. And it wouldn't be -- if I were to direct my students and if my students were to come to me and say, I ran that de-duplication process that you gave me but some of the samples still have duplicates in them even though some samples are duplicated twice or duplicated once so that they occur two times, I would tell my students, please go back and fix the bug that you have. This is a machine learning. This is a data science bug. This is an AI bug. I would tell my students, go back and fix the bug and come back after you fixed that bug. Sometimes I would even go in and help them to modify the code to fix that bug if they had such a bug.
So everything in my study in my field tells me that one would not want to have data or a sample or an entity or a lyric or a work or whatever the granularity we're talking about, you know, if it represents something, we would not want it to appear in a duplicate form. It would not make sense to me to want that.
And then secondly, this, you know, look -- to get back to your earlier questions, none of this looking at, say, for example, code, the de-duplication code because it was not necessary

156

to form my final analysis and my final conclusions.
So that is how I can be assured that [redacted]. And I'd be happy if you're interested, I could provide a different version of that answer. I'm quite willing to answer all of your questions.
Q Did you do any analysis where you ran copies of Plaintiffs' works in suit obtained from different sources that are found in Anthropic's base training data through Anthropic's de-duplication techniques to see what the result was?
A I'm sorry --
ATTORNEY GOOCH: Objection. Form. Asked and answered.
THE WITNESS: Would you mind, please, repeating the question?
BY ATTORNEY MILLER:
Q Yeah. Did you do any analysis of applying Anthropic's de-duplication techniques to copies of Plaintiffs' works in suit obtained from different sources that are contained in Anthropic's base training data to see what the result would be?
A So --

157

ATTORNEY GOOCH: Same objections.
THE WITNESS: So in -- in the context of my study of machine learning and artificial intelligence and in my knowledge that duplication of any form is bad and it hurts the model's performance, and knowing that Anthropic is a highly functional general purpose model that can do many, many quite different things, it would not have changed any of my analysis and my conclusion to have done that experiment that you just suggested.
Also, in order to have done that experiment, I would have needed to have access to the de-duplication code that previously I mentioned wasn't necessary to have access to in order to come to the conclusions that I came to in this report.
BY ATTORNEY MILLER:
Q If it's possible to de-duplicate [redacted], is it possible to use -- withdrawn.
Let me try this again. If it's possible to de-duplicate [redacted]

158

[redacted] altogether?
ATTORNEY GOOCH: Form.
THE WITNESS: I'm sorry. Would you mind repeating that again? I want to make sure I've got the question right in my mind.
BY ATTORNEY MILLER:
Q If it's possible to de-duplicate [redacted]
A Okay.
ATTORNEY GOOCH: Same objection.
THE WITNESS: So what I was saying earlier was that because Claude is a highly functional model, highly functional general purpose model that does lots of different things and because of my experience in machine learning and artificial intelligence, and because of my experience in training datasets and testing various different models and seeing firsthand the detrimental

159

effects of redundancy and imbalance in training data, I did not need to look at the code to come to the conclusions that I came to in my report, because the evidence suggests that the code was functioning as it is.
Now, that doesn't mean that necessarily one can as you put it, remove all copies because one would need to know precisely what a copy is. My understanding is for many of the works, for many of the lyrics, there is no definitive copy available. My understanding was that there were requests for definitive copies. I did not receive access to definitive copies, and so therefore, without actually having a definitive copy, first of all, it would not be possible to know.
What I was referring to earlier was the effectiveness of de-duplication. But then it becomes a bit more involved and complicated and the reason is because works, the works, the lyrics occur in many, many different variant forms. The analysis that I did, for example, on the common pile shows that many of the works occurred in fragmented forms. It's quite likely that many of the works occurred in distorted forms, small variations. That includes things like

160

punctuation, word changes, word alterations, instances where there are versions of the lyrics that appear because of phenomenon known as mondegreens which is where people mishear a lyric and perhaps casually themselves type it in on to some web page.
Also, there are instances of discussions of lyrics where parts of lyrics, fragments of lyrics, pieces of lyrics are scattered pretty much everywhere and intertwined and entangled with the rest of the collective body of material that we know as the internet over time.
So the idea of removing all lyrics from training data doesn't seem like at all to me a well-defined problem. In order to do anything computationally or mathematically, the problem needs to be well defined because one needs to know what one needs to do in order to do that.
If it's the case that one doesn't have a definitive copy of lyrics, number one, and if it's the case that one doesn't know precisely what even a lyric is if it occurs in so many different variations and fragments and pieces and bits all over the place and how short of a fragment should be for it to be part of a lyric because if a lyric

Pages 169-172

169

Plaintiffs' copyrighted works for training its LLMs is transformative as transformative is used in the context of fair use law?

ATTORNEY GOOCH: Why don't you point him to his section in his report, Corey?

THE WITNESS: Yeah, I would be very happy to explain a particular section in the report. I am not a lawyer, and so the particular legal definition of transformativeness is not something that I can formally express any expertise in.

I can express expertise in other things, like the effect of duplication -- the adverse effect of duplication on the performance of machine learning models. But in this particular case, it's -- you know, as a -- as a nonlawyer, I -- I can't really express those notions or have an opinion on those notions.

BY ATTORNEY MILLER:

Q On the page right in front of you, do you see there's a big header "Roman VI," and it states, "The transformative nature of generative AI large language models"?

Do you see that?

A Yeah.

Q When you use the word "transformative"

170

there, are you using it to mean the legal concept of transformative as that concept has a meaning in fair use law?

A This is used in its -- in its general sense, and it includes a variety of different sentences. And what the text in the document clarifies is that there are -- it may be in this document. It may be in one of the other documents. It might take me a moment.

But there are instances in these documents where I talk about the various different ways that transformativeness is used.

Q And which way are you using it here?

ATTORNEY GOOCH: And you should feel free to review your report. He's asking you questions about your report, and while he doesn't want to direct you to specific things, Paragraphs 288 through 291.

ATTORNEY MILLER: I'll ask you not to coach the witness, please.

ATTORNEY GOOCH: I'm not coaching the witness. You're asking questions about 600 pages of reports and refusing to point him to any one. This isn't a memory test.

THE WITNESS: So Section 6, the title is

171

just the general notion. And if you look at Section 6A, for example, it talks about how LLM training transforms text into abstract statistical relationships.

So this report is comprehensive and it's meant to convey the general sense in which that word might be used. And from counsel, I came to an understanding of -- from -- as written in Paragraph 286, I believe the definition that you're referring to. But if you look at the same section, Subsection A of that section, for example, it talks about -- in Paragraph 293, it says, "Text data used for LLM training is transformed through heavy computational processing."

It's worth noting that when we say "LLM," what that really means is a transformer. And a transformer is a particular architecture of neural network that includes many, many layers of multiheaded tension layers, followed by multilayer perception layers, various different other forms of normalization layers, as well as a variety of other architecture features, neural architectural features.

And the reason why it's called a

172

transformer is because of its transformative abilities. And I came to know that term certainly long before this case began, and I came to know that term and that term in general, "transformer," and the fact that the transformer, despite its very name is transformative, has the ability to transform.

And the community, thousands of people, tens of thousands of researchers and scientists around the world use this notion of a transformer and how the transformer transforms information and data into new -- generally, in the context of language, statistical abstract, semantic, linguistic forms that, as we were referring to earlier, can transcend some of the information that exists in the original data.

And it's interesting that you say this, because it gets back to this notion of transcend -- transcending the original data. The transformer, arguably, is transformative in its more informal sense, because it can do -- it exhibits behaviors that transcend the behavior of the original training data.

So this is a very -- this is meant to be a complete section, not just a specific narrow