EXHIBIT A



Planet Depos®
We Make It Happen™

HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY

Transcript of Ben Zhao, Ph.D.

Date: February 20, 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.,
Plaintiffs,
v.
ANTHROPIC PBC,
Defendant.
: Case No. 5:24-cv-03811-EKL-SVK
- - - - - - - - - - - - - - - - x

Highly Confidential
Attorneys' Eyes Only
Pursuant to the Protective Order

Videotaped Deposition of BEN ZHAO, Ph.D.
Chicago, Illinois
Friday, February 20, 2026
9:06 a.m. CST

Job No.: 621063
Pages: 1 - 462
Stenographically Reported By:
Michelle M. Yohler, CSR, RMR, CRR

2

Deposition of BEN ZHAO, Ph.D., held at the location of:

ARNOLD & PORTER
70 West Madison Street, Suite 4200
Chicago, Illinois

Pursuant to notice before Michelle M. Yohler, a Certified Shorthand Reporter, Registered Professional Reporter, Registered Merit Reporter, Certified Realtime Reporter, and a Notary Public, CSR No. 84-4531, in and for the State of Illinois.

3

A P P E A R A N C E S

ON BEHALF OF THE PLAINTIFFS:
MR. KEITH HOWELL
MR. TIMOTHY CHUNG (Remotely)
OPPENHEIM & ZEBRAK, LLP
4530 Wisconsin Avenue, NW, Fifth Floor
Washington, D.C. 20016
202.480.2999
khowell@oandzlaw.com

ON BEHALF OF THE DEFENDANT:
MS. SONAL N. MEHTA
WILMER HALE
2600 El Camino Real, Suite 400
Palo Alto, California 94306
650.858.6000
sonal.mehta@wilmerhale.com
-and-
MR. JARED V. GRUBOW (Remotely)
WILMER HALE
250 Greenwich Street
New York, New York 10007
212.230.8800
jared.grubow@wilmerhale.com
(Continued)

4

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

ALSO PRESENT:
Mr. Ian Chen,
In-House Counsel for Anthropic
Mr. James Porter, Videographer

Pages 129-52

129

course of your work in this case?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I believe I have some screenshots from a while ago, but I don't believe that they are necessarily complete.
BY MS. MEHTA:
Q Did you provide the screenshots that you did take to counsel?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I don't recall whether I told them about the results, or copy-pasted, or shared screenshots.
BY MS. MEHTA:
Q Other than the -- the personal prompts and outputs that you put in -- some of which you may have record of and some of which you may not have records of -- did you do any other testing to assess the functionality of Anthropic's guardrails?
MR. HOWELL: Objection. Form. Vague.
BY THE WITNESS:
A I'm sorry. Could you repeat the first part of that question, the -- the "except for"

130

portion?
BY MS. MEHTA:
Q Yes. And, actually, let me -- I'll actually break that down a --
A Okay.
Q -- little bit.
You mentioned that you did your -- some personal experiments with prompts and outputs. You may have records of some of that; you may not have some records of all of it.
Did you include any of that testing in your expert reports in this case?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I believe I referred to some of that in one of my earlier reports, perhaps the one from 2023, where I talked about my interactions with Anthropic. I don't recall exactly whether I referred to it again in the later reports.
BY MS. MEHTA:
Q Okay. Now, let's set aside those prompts and outputs.
Did you do any other testing to assess the efficacy of Anthropic's guardrails?
MR. HOWELL: Objection. Form. Vague.

131

BY THE WITNESS:
A I didn't see it necessary in more recent times to do my own personal testing because to do so at a -- at a more comprehensive level would have been more time-consuming.
Instead, I relied on published research from the academic community with which -- you know, including a fair number of well-structured empirical tests by different university groups that, you know, tested Anthropic's guardrails far more explicitly than I have.
BY MS. MEHTA:
Q Okay. You said "I didn't see it necessary," so I can -- I can interpret that to mean that you did not, in fact, do that; is that fair?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I did not repeat individual ad hoc tests on Anthropic's guardrails since my first -- I believe since my first report I may have done some more interactions with Anthropic, but not in a way that was necessarily targeting the efficacy of their guardrails.
And, again, to me, far more important was

132

the fact that there have been a number of studies that have studied these guardrails and their efficacy from well-regarded research groups, and many of those reports are published and online, which inform some of my opinion about their current efficacy.
BY MS. MEHTA:
Q Okay. I understand your testimony about the research, and we're going to talk about that. I'm not asking about the academic research.
The question I'm asking is: Since your 2023 declaration, did you personally do any empirical testing or evaluation of the efficacy of Anthropic's guardrails?
MR. HOWELL: Objection. Form. Vague.
BY THE WITNESS:
A Again, it's a -- you know, when you say "empirical testing," perhaps I don't view it as a specific test, but I did interact with Anthropic, and particularly, you know, probing it for specific questions and responses. I believe in some cases in preparation, for example, for some of my lectures, trying to understand, you know, what current models do, and current models behave.
So, yeah, some of that did involve, you

133

know, direct interactions with Anthropic, which, of course, means direct interaction with its guardrails.
BY MS. MEHTA:
Q Okay. I'm not asking about your use of Anthropic in the ordinary course of your work; I'm asking -- and whether or not when you did that there were -- the guardrails may have been implemented. That's not what I'm getting at. I'm not asking about whether just in the ordinary use of Claude the guardrails were implemented.
I'm asking a very specific question: Since your 2023 declaration, did you do any empirical testing or evaluation of the efficacies of Anthropic's guardrails yourself?
MR. HOWELL: Objection. Form. Vague. Asked and answered.
BY THE WITNESS:
A What I'm trying to make clear is that this is the course of my normal work. I teach a class on average machine learning, which, by its definition, covers issues of what models can and cannot do, what their behavior is. In fact, you know, matters of architecture and all these related issues are part of the class.

134

So -- and, in fact, to your more specific question, yes, I did actually test Anthropic by giving specific inputs that would, in some cases, trigger guardrail responses as part of my interactions when I prepare for classes like that.
BY MS. MEHTA:
Q Okay. Now, those interactions in which you prompted Anthropic and you received the guardrails, that's not something that you disclosed as part of your expert opinions in this case, correct?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I don't believe so, not in the more recent round of reports.
BY MS. MEHTA:
Q Are there other instances in which -- actually, let me ask the question this way: Other than the instances in which for your class or your research you've interacted with Anthropic and you've prompted Anthropic's guardrails, for purposes of this work in this case, have you done anything to empirically test or evaluate the adequacy of Anthropic's guardrails since your original declaration in 2023?

135

MR. HOWELL: Objection. Form. Vague. Asked and answered.
BY THE WITNESS:
A Given what I do and the research that I do and the classes I teach on the subject matter, there is no clear line between when I do something to test a particular system whether I have a label on it that is whether it belongs to a particular legal case or a class or my general research purposes.
So did I, at the time of interacting with Anthropic, say this is my consulting time and, therefore, this is -- belongs in a legal matter? I did not.
However, given what I do, nearly all of my research, especially since the arrival of generative AI, has been about these related issues. And so I have not explicitly partitioned off which of my opinions have been through the course of my research or teaching experiences versus which explicit actions were taken on behalf of any stakeholder in a -- in a court case.
BY MS. MEHTA:
Q We talked earlier about how you endeavored to put all of the bases of your opinions in your

136

expert reports, right?
A Yes.
Q In your expert reports, you did not include any empirical test or evaluation of the efficacy of Anthropic's guardrails that you yourself conducted, correct?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A With respect to specific instances where I interacted with Claude during the course of research or during the course of preparation for class, no, I did not indicate those in -- as part of my report.
BY MS. MEHTA:
Q Okay. And where you interacted personally with Claude in the context of the reports -- so I'm not talking about your research or the class, but specifically for the reports -- you also didn't include any empirical study of the efficacy of Claude's guardrails; isn't that right?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I'm sorry. I'm a little bit confused. We had already established that I had already covered that in my earlier reports in 2023.

137

BY MS. MEHTA:
Q I'm sorry. I'm -- that's fair. I mean since the 2023 declaration, and let me clarify that.
With respect to, not the research you do for your class or for your research generally, but specifically with respect to your work in this case, you have not personally done any empirical study or testing of the efficacy of Claude's guardrails since your original declaration in 2023, fair?
MR. HOWELL: Objection. Form. Asked and answered.
BY THE WITNESS:
A Again, I -- yeah, I do believe we've already covered this question before, and my answer is the same.
I do not put explicit labels on my interactions with language models or AI systems, so whether it's in the course of a research project, whether it's in the course of preparation for a class, whether it's in the course of explicit preparation for counsel or a deposition or a statement, there is -- I do not put labels on these interactions with models and say: This is

138

for purpose X.
So when you say I did not -- you're asking me a question about did I specifically do it for this particular court case, my answer is all my things, all the activities that I do research and teaching are relevant to this case, and I don't draw explicit boundaries when it comes to did I do only this action for a legal case.
BY MS. MEHTA:
Q That's not what I'm asking.
What I'm asking you is whether -- what I'm getting at is what you disclosed in your expert reports. In your expert reports post the original 2023 declaration, you have not disclosed any empirical testing or evaluation of the efficacy of Claude's guardrails that you yourself conducted, fair?
MR. HOWELL: Objection. Form. Asked and answered.
BY THE WITNESS:
A In my more recent reports, I do not recall including examples of my interactions with Claude that were -- that took place other than explicitly labeled time with -- for the court case.
This was largely because of the fact that

139

we had substantially more evidence about the efficacy of these mechanisms from multiple sources, including Dr. Shan's report, including published research that was elsewhere that I consider far more exhaustive and comprehensive than my own interactions.
BY MS. MEHTA:
Q And you did not design any experiment or study to test the efficacy of Claude's guardrails for purposes of your expert opinions in this case, correct?
MR. HOWELL: Objection. Form. Vague.
BY THE WITNESS:
A I think your question is asking if I designed experiments for the purposes of this testimony, and the answer is no because there has been already substantial work in designing experiments and tests by the research community about these types of guardrails.
So, no, I did not design any new experiments for the purposes of this testimony.
BY MS. MEHTA:
Q And I think this is a straightforward question, but you did not conduct any experiment or study to test the efficacy of Claude's

140

guardrails for purposes of your expert opinions in this case, correct?
MR. HOWELL: Objection. Form. Misstates prior testimony. Asked and answered.
BY THE WITNESS:
A So not having designed new experiments does not mean that I did not conduct experiments that were relevant to prior published research about how to test and evaluate these types of guardrails.
You know, in my interactions with Anthropic, I have indeed leveraged techniques that are commonly known in the community as to how to test and how to evaluate some of these guardrail systems.
BY MS. MEHTA:
Q You did not disclose the results of any experiment or study designed to test the efficacy of Claude's guardrails in your expert reports in this case; isn't that right?
MR. HOWELL: Objection. Form. Vague. Misstates prior testimony.
BY THE WITNESS:
A I believe we've already covered this particular topic, and I'll restate the answer

141

again.
I have not disclosed in this report additional experimentation with re- -- regard -- conducted by me personally with regard to these guardrails because there were far more substantial and comprehensive data that we have from research reports and from Dr. Shan on these same questions.
BY MS. MEHTA:
Q If the jury and the judge wanted to know what experiments you -- or studies you conducted with respect to the efficacy of the guardrails through your research, in preparation for your class, or otherwise that was not included in your expert reports, how would we go about understanding that?
MR. HOWELL: Objection. Form. Calls for speculation.
BY THE WITNESS:
A Yeah, I -- again, this is entirely speculative. These things were not in my report. They were not the main bases for my opinions in this report. So I guess I would have to see what kind of requests are made and discuss with counsel on how to go about addressing those requests.

142

BY MS. MEHTA:
Q Have you ever -- well, I'll give you a time frame. Since post your original report -- so for purposes of the 2025 and 2026 expert reports, did you ever prompt Claude for copyrighted material and receive a notification that Claude could not give you that information?
MR. HOWELL: Objection. Form. Vague.
BY THE WITNESS:
A I'm not sure what you mean by Claude -- the last bit of your answer -- of your question was that you wanted to know if I had instances where Claude refused to answer; is that the question?
BY MS. MEHTA:
Q Yes, refused to answer on the basis of a copyright guardrail.
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I'm not sure -- certain, but there might have been.
BY MS. MEHTA:
Q What percentage of times that you prompted Claude for copyrighted material since your original report -- so let's say 2024, 2025, and

143

2026 -- did Claude give you a refusal to answer on the basis of a copyright guardrail?
MR. HOWELL: Objection. Form. Vague.
BY THE WITNESS:
A I guess I would say a minority of the time. So maybe 20, 30 percent.
Of course this is entirely dependent on what kind of questions I ask. But there's hallucinations, of course, there's, you know, incorrect answers. So there's a variety of different responses.
You're asking about refusals to answer on the basis of a copyright; I would actually say that's a -- not a significant portion.
BY MS. MEHTA:
Q When you specific -- have you ever specifically prompted Claude for copyrighted information post your 2023 declaration?
A Yes.
Q How many times -- what percentage of times in response to a specific request for copyrighted material did you get a copyright guardrail?
MR. HOWELL: Objection. Form. Vague.
BY THE WITNESS:
A Yeah, I guess the issue here with this

144

question is you're asking for specific requests, and I don't know how specific means, you know, if -- I don't believe I've asked Claude much to the order of, you know: "Give me Page 1 for this book." So that would be a small instance of that. I don't recall me asking -- forming much of those prompts.
BY MS. MEHTA:
Q Have you ever asked Claude for copyrighted material and received back a complete or substantial regurgitation of copyrighted material since your original declaration?
MR. HOWELL: Objection. Form. Vague.
BY MS. MEHTA:
Q Actually, let me -- I'll be even more precise with the question --
A Okay.
Q -- this might help.
Since 2024, have you ever prompted Claude for copyrighted material and received back complete or substantial regurgitation of copyrighted material?
MR. HOWELL: Objection. Form. Vague.
Object to the extent it calls for a legal conclusion.

145

BY THE WITNESS:
A I honestly don't recall. I think my tests with guardrails were not solely focused on copyrighted work, so I -- I honestly -- I honestly couldn't tell. I don't recall.
BY MS. MEHTA:
Q Can -- can you identify a single instance since 2024 in which Claude gave you a complete or substantial regurgitation of copyrighted material in response to any prompt?
MR. HOWELL: Objection. Form. Vague. And to the extent it calls for a legal conclusion.
BY THE WITNESS:
A I'm sorry. I honestly don't recall.
Yeah, these were fairly ad hoc tests that I would do when I have a particular idea and just trying to get a sense of how the model reacts. So I -- these were not, you know, well-planned-out written comprehensive tests, and so I -- I really couldn't tell you. Sorry.
BY MS. MEHTA:
Q Okay. That's -- that's fine.
In all the work that you -- you said you were originally signed up for this case in 2023; is that right?

146

A Yeah, sometime in that time frame. It's been a few years.
Q In all the work that you've done on this case in 2024, 2025, and now 2026, can you identify any instance in which you prompted Claude and received a complete or a substantial regurgitation of copyrighted material in response to a prompt?
MR. HOWELL: Objection. Form. Vague. Calls for a legal conclusion.
BY THE WITNESS:
A I didn't take it upon me to form my opinion based on my own experimentation with Claude.
You know, that type of experimentation I consider much less reliable than published results of people who are designing well-formed, you know, meticulous tests that target a specific behavior.
So I don't recall using anything like that as a basis -- as a main basis for my opinions in these reports. Nor did I feel like I needed to re-evaluate the systems, given that I had experiences with them in 2023 and I felt like those experiences were informative.
And, yeah, I didn't feel like it would be informative to go back and repeat the same test again and again.
BY MS. MEHTA:
Q Was there anything stopping you from conducting the same tests that you did in 2023 subsequently in 2024, 2025, or 2026?
MR. HOWELL: Objection. Form. Calls for speculation.
BY THE WITNESS:
A Nothing in particular other than the fact that I think I've already seen what was capable and what was happening and -- you know, so I didn't go back and repeat the same test.
BY MS. MEHTA:
Q Now, you understand that Claude has changed since 2023, right?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I understand that these models are changing almost -- fairly frequently and with lots of under-the-cover things I -- I may or may not be aware of.
BY MS. MEHTA:
Q And you also understand that Anthropic has implemented lots of new guardrails since 2023, right?

147

148

MR. HOWELL: Objection. Form.
BY THE WITNESS:
A I understand that many new guardrails have gone up since this lawsuit was filed, yes.
BY MS. MEHTA:
Q Have you done anything to test yourself the efficacy of the guardrails since the new guardrails have been implemented post this lawsuit?
MR. HOWELL: Objection. Form. Asked and answered.
And I'll note we're over the hour mark, so if we can find a natural stopping point.
BY THE WITNESS:
A You know, I think we can come back to one of my earlier answers, which is that, yes, I have tested some of the guardrails in more, sort of, ad hoc informal settings.
But, again, when I did those interactions with the models, I did not intentionally repeat the same tests that I did back in 2023. For those type of tests, I relied on existing research literature and Dr. Shan, but my interactions were primarily focused more on other questions with guardrails.

149

BY MS. MEHTA:
Q Would you have the ability to design an experiment to test the guardrails and the efficacy of the guardrails if you so choose?
MR. HOWELL: Objection. Form.
BY THE WITNESS:
A Would I be able to? Yes, of course. I mean, I don't see a particular need to. There's a lot of published reports that give very specific guidance on how to form, you know, what are almost benchmarks for these types of guardrails. And there's well-known methodology that, you know, researchers are taking when they assess the capabilities of these kind of guardrails.
So I don't see a need to reinvent the wheel, but, you know, many of the methodologies are fairly familiar to me. They're, you know, common knowledge in the research literature.
So, yeah, I don't see a need to invent new ones, but, you know, the existing methodologies certainly all make sense and it's things that I understand fairly well, yes.
BY MS. MEHTA:
Q Right. So the question I was getting at is -- I'll ask it a slightly different way.

150

Was there anything stopping you from conducting your own test of the efficacy of Anthropic's guardrails post 2023?
MR. HOWELL: Objection. Form. Vague. Calls for speculation. Asked and answered.
BY THE WITNESS:
A I believe there were higher priority things that demanded my time. There is plenty of research from people that I know and respect in the field who have done fairly exhaustive tests who have demonstrated the -- the, sort of, ineffectiveness of some of these guardrails.
And so, no, I did not see it necessary to repeat their tests because I honestly, you know, have higher priorities.
BY MS. MEHTA:
Q Did you -- you mentioned the -- a number of people in the field that you respect that engaged in these tests. You understand that a number of the tests have been characterized or criticized as involving adversarial prompting, right?
MR. HOWELL: Objection. Form. Vague.
BY THE WITNESS:
A Not sure "criticize" is the right word.

151

When we ascertain the efficacy of any system, you don't test it with all the most naive and basic conditions because then you don't learn anything.
So, you know, whether it's computer security or safety or any of these fields, the common practice is to understand where the limitations of systems are, and if you just give them the most common and basic test, that's not going to accomplish the purpose.
BY MS. MEHTA:
Q Are you aware of any test of the efficacy of Anthropic's guardrails with respect to copyright regurgitation that did not involve adversarial prompting or jailbreaking?
MR. HOWELL: Objection. Form. Vague. Compound.
BY THE WITNESS:
A I believe Dr. Bilmes in his report mentioned his own tests, which seemed to be largely non-adversarial.
But certainly I believe in most of these reports that are out there, the reports tend to cover a range of prompts starting from the most naive and basic to techniques that are increasingly specialized or, you know,

152

experimental.
BY MS. MEHTA:
Q Setting aside Dr. Bilmes' opinions, which I know you disagreed with. We'll get into that later.
With respect to the academic literature and the research that you relied upon for your opinions in this case, are you aware of any research or study that has tested Anthropic's guardrails with respect to copyright regurgitation and successfully elicited copyrighted material without jailbreaking or using adversarial prompting?
MR. HOWELL: Objection. Form. Vague. Asked and answered.
BY THE WITNESS:
A That's a very specific question. I mean, if I had time, I could go back and revisit the conditions for a bunch of these research reports and answer that more concretely.
Offhand I would say probably because I remember especially some of these older reports, you know, when people were first starting to look at both hallucinations and memorization, you know, many of the initial tests were more