UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

ANDREA BARTZ, ANDREA BARTZ, INC., CHARLES GRAEBER, KIRK WALLACE JOHNSON, and MJ + KJ, INC., individually and on behalf of others similarly situated, Plaintiffs, v. ANTHROPIC PBC, Defendant.	Case No. 3:24-CV-05417-WHA Action Filed: August 19, 2024 **DECLARATION OF JARED KAPLAN IN SUPPORT OF DEFENDANT ANTHROPIC PBC'S MOTION FOR SUMMARY JUDGMENT**

I, Jared Kaplan, hereby declare as follows:

A. Personal Background

1. I am a co-founder and the Chief Science Officer of Defendant Anthropic PBC ("Anthropic") and submit this declaration in support of Anthropic's Motion for Summary Judgment. Unless stated otherwise, all facts stated herein are within my personal knowledge. If called upon, I would be willing to testify as set forth in this declaration.

2. I have a B.S. in physics and mathematics from Stanford University and a Ph.D. in physics from Harvard University.

3. I began my career as a theoretical physicist, with research interests in effective field theory, particle physics, cosmology, scattering amplitudes, and the conformal field theory bootstrap. I have been a professor in the Department of Physics and Astronomy at Johns Hopkins University since 2012.

4. During a sabbatical from my work in theoretical physics, I began to collaborate with computer scientists to conduct research into machine learning and the development of large language models. I have taught courses and published over 60 scholarly articles on a mixture of theoretical physics, machine learning, and large language models ("LLMs").

5. I am one of the founders of Anthropic, which we established in 2021. Since 2023, my title has been Chief Science Officer. I have extensive personal knowledge regarding Anthropic's business operations and practices, including knowledge of how Anthropic's generative artificial intelligence models were created and trained, their technical capacities and processes, our goals for the models, and our business operations related to the models.

B. Anthropic's Background as a Research Company

6. Anthropic is an artificial intelligence ("AI") company based in San Francisco, California, working to develop generative AI models, and, in particular, LLMs. LLMs are text-based generative AI models trained on extremely large volumes of data to develop a functional understanding of how language works, to use it to generate new text.

7. Anthropic is a Public Benefit Corporation and its mission is to build safe, beneficial artificial intelligence. Anthropic focuses on research projects that we believe will advance the progress

of the technology towards those safe and beneficial uses. For example, Exhibit 7 is a document that I prepared in July 2021 (a few months after the company was founded) that sets out many of the early research projects Anthropic was focused on.

8. Anthropic has extensively published the results of our research efforts to help other AI researchers and developers. There are dozens of Anthropic academic publications, which have been frequently cited by other researchers, publicly available at www.anthropic.com/research/.

9. One representative example of the science enabled by Anthropic's LLM research is **interpretability.** Interpretability is the science of understanding how LLMs "think" by studying models in ways similar to how neuroscientists map neural pathways and observe activation patterns in biological brains. While LLMs were historically "black boxes," whose behaviors couldn't be reliably interpreted even by their developers, Anthropic has made breakthrough discoveries concerning interpretability in its analysis of its commercial LLM product, Claude. Two examples of this type of research are found in our widely-cited papers *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*[1] and *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*.[2]

10. The applications of our interpretability research extend beyond general purpose LLMs. For example, biology researchers have reported that they are further building on this work to identify interpretable features in protein language models, which are large language models trained on large databases of protein sequences that help researchers learn the relationship between protein sequences, structures, and functions.[3]

---

[1] Trenton Bricken et al., *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning* (Oct. 4, 2023), https://transformer-circuits.pub/2023/monosemantic-features/index.html (Exhibit 8).

[2] Adly Templeton et al., *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet* (May 21, 2024), https://transformer-circuits.pub/2024/scaling-monosemanticity/ (Exhibit 9).

[3] *Evo Mechanistic Interpretability Visualizer*, Arc Institute https://arcinstitute.org/tools/evo/evo-mech-interp (last accessed Mar. 17, 2025) (Exhibit 10); Garyk Brixi et al., *Genome Modeling and Design Across All Domains of Life with Evo 2* (Feb. 21, 2025) at 1-2, 18, https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1.full.pdf (Exhibit 11); Elana Simon & James Zou, *InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse*

[redacted]

11. Another key research area for Anthropic is **alignment.** Alignment is the science of developing AI systems that reliably follow human values and intentions. For example, Anthropic's December 2022 paper *Constitutional AI: Harmlessness from AI Feedback*[4] has been cited over 1,400 times on Google Scholar[5] and details our research into how models can evaluate and revise their outputs using human-defined principles that are encoded in the models themselves, making them more helpful while reducing harmful or evasive responses. As another example, in our paper *Many-shot Jailbreaking*,[6] Anthropic also identified a novel vulnerability in AI systems called "many-shot jailbreaking." "Jailbreaking," in the generative AI context, refers to attempts to circumvent or bypass the safety guardrails, content policies, and ethical limitations built into AI systems. The publication of our research identified a new vector bad actors could use to bypass LLM safety measures via the technique of many-shot jailbreaking, wherein the user gives the model hundreds of demonstrations of undesirable behavior to steer model behavior.[7]

12. Another key research area for Anthropic is the **societal impact** of LLMs. Our societal impact research focuses on quantifying and analyzing how AI systems affect and interact with human society. For example, in our paper *Evaluating and Mitigating Discrimination in Language Model Decisions*,[8] Anthropic analyzed how prior versions of its LLMs might produce discriminatory outputs in various societal contexts and identified careful interventions for reducing these risks in practice.

Autoencoders (Jan. 28, 2025) at 1-4, 11, https://www.biorxiv.org/content/10.1101/2024.11.14.623630v2.full.pdf (Exhibit 12).

[4] Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (Dec. 15, 2022), https://arxiv.org/abs/2212.08073 (Exhibit 13).

[5] *Google Scholar Search Results for "Constitutional AI: Harmlessness from AI Feedback,"* Google Scholar https://scholar.google.com/scholar?hl=en&as_sdt=0%2C9&q=constitutional+ai%3A+harmlessness+from+AI+feedback&btnG= (last accessed March 24, 2025) (Exhibit 14).

[6] Cem Anil et al., *Many-shot Jailbreaking* (Apr. 2, 2024), https://www-cdn.anthropic.com/af5633c94ed2beb282f6a53c595eb437e8e7b630/Many_Shot_Jailbreaking__2024_04_02_0936.pdf (Exhibit 15).

[7] *Id.* at 1.

[8] Alex Tamkin et al., *Evaluating and Mitigating Discrimination in Language Model Decisions* (Dec. 6, 2023), https://arxiv.org/pdf/2312.03689 (Exhibit 16).

13. While Anthropic's early LLMs were used for research and development purposes, during late 2022 through early 2023, Anthropic's focus evolved to include commercial deployments of our LLMs. Expanding to commercial LLMs enabled Anthropic to demonstrate to the external world – including competitors, policymakers, and others – that it is possible to commercially deploy cutting-edge LLMs safely, which advances its underlying mission of advancing reliable, interpretable, and steerable AI systems. Anthropic focuses on commercial model development alongside safety research because a safe LLM that nobody uses cannot fully demonstrate the potential of reliable, beneficial frontier AI systems.

14. Anthropic has not changed its long-term mission. In fact, since our commercial models were released, we have increased our rate of publishing safety research. For example our paper *Forecasting Rare Language Model Behaviors*,[9] published just a few weeks ago, describes our alignment research concerning forecasting certain dangerous model behaviors (*e.g.*, an LLM providing instructions to a user on how to make dangerous biological or chemical substances).

15. Another example is our paper *Alignment Faking in Large Language Models*, which provided the first empirical evidence that modern AI systems may develop deceptive behaviors to preserve their preferences without explicit instruction to do so. We demonstrated that LLMs will strategically comply with training objectives during training to prevent modification of their behavior when unmonitored. Anthropic is only able to produce important research like this because of the size and scale of its LLMs, which allow us to predict potentially dangerous behaviors that makers of smaller LLMs would not be able to evaluate or test themselves before deployment.

C. <u>Anthropic's LLM Claude and Claude's Uses</u>

16. Anthropic's signature product is a general purpose LLM called Claude. Claude 1, the first commercially released and generally accessible version of Claude, was released in March 2023. Claude 2 was released in July 2023. The Claude 3 model family (Haiku, Sonnet, and Opus) was released on March 4, 2024, with the different versions corresponding to the increasing capabilities of

[9] Erik Jones et al., *Forecasting Rare Language Model Behaviors* (Feb. 24, 2025), https://arxiv.org/pdf/2502.16797 (Exhibit 17).

[redacted]

the model (*i.e.*, Haiku was the least capable model, Sonnet was more capable, and Opus was the most capable).

17. On June 20, 2024, Anthropic released Claude 3.5 Sonnet. On October 22, 2024, it released Claude 3.5 Haiku. On February 24, 2025, Anthropic released Claude 3.7 Sonnet, which is its most recent commercially released and generally accessible LLM.

18. To create Claude, Anthropic has employed hundreds of engineers who have devoted hundreds of thousands of hours of work to its development and we have spent more than [redacted] in computing costs alone related to research and training.

19. Claude.ai (the web interface where consumers access Claude) averages around [redacted] active users on a daily basis. There are also [redacted] of Anthropic business customers that rely on Claude.

20. Claude is a versatile LLM that can be used in many different contexts depending on a user's needs. Claude was trained on existing examples of human work product, but it is designed to create new, original outputs and act as a general purpose assistant. Claude is not designed to memorize, copy, or reproduce its training data. Claude interprets and responds to a vast variety of user queries, known as "prompts," in an intelligent, human-like way. As a result of Anthropic's research, development, and training of the model, Claude's original outputs are produced algorithmically without real-time human input. Claude provides inventive, tailored outputs for users on myriad topics. It allows users to engage in general, open-ended, and back-and-forth contextual conversations involving complex reasoning, creativity, thoughtful dialogue, software coding, or detailed content creation.

21. Depending on the nature of the user prompt, Claude can:

a. Handle basic instructions and logical scenarios, including formatting outputs, following if-then statements, and making a series of logical evaluations in a single prompt.

b. Take direction on tone and "personality" when providing outputs.

c. Speak a variety of common human languages.

d. Provide outputs in a variety of programming languages, including generating code based on natural language descriptions, troubleshooting and adding functionality to existing codebases, and translating code from one programming language to another.

e. Process huge amounts of text from documents, emails, FAQs, chat transcripts, records, charts, or graphs and edit, rewrite, summarize, classify, extract structured data, draft Q&As, and complete other tasks based on the content.

f. Interact with external tools and APIs to perform tasks, manipulate data, and provide dynamic and accurate responses.

g. Provide helpful answers to user queries, with detailed background on technical, scientific, and cultural knowledge.

22. Individual users rely on Claude for a vast array of purposes. Anthropic recently published a research paper titled *Clio: Privacy-Preserving Insights into Real World AI Use*[10] ("Clio"). The Clio paper studies the many use cases of Claude in the real world ("Clio" is shorthand for **Cl**aude **i**nsights and **o**bservations). As detailed in the paper, Anthropic found that the most common consumer use of Claude is to help users write computer code and for other business cases ("We find that coding and business use cases dominate, with significant activity in areas like debugging code, drafting professional emails, and analyzing business data.").[11] A demo video of Claude's code-writing capabilities can be seen at the link below (in which Claude creates an 8-bit video game featuring a crab named "Claw'd" that jumps over seashells, using just a few prompts).[12]

23. Claude is frequently used for a wide variety of content creation writing projects, including editing, rewriting, summarizing, and brainstorming ideas. Claude can assist users in straightforward writing tasks like drafting emails and job descriptions,[13] or in developing more

---

[10] Alex Tamkin et al., *Clio: Privacy-Preserving Insights into Real World AI Use* (Dec. 18, 2024), https://arxiv.org/pdf/2412.13678 (Exhibit 18).

[11] *Id.* at 2.

[12] Anthropic, *Claude 3.5 Sonnet for sparking creativity*, YouTube (Jun. 20, 2024) https://www.youtube.com/watch?v=rHqk0ZGb6qo&list=PLf2m23nhTg1OPBSr9HgYKxhF3wzUpMZP_&index=2 (Exhibit 19).

[13] Introducing Claude, Anthropic (Mar. 14, 2023), https://www.anthropic.com/news/introducing-claude (Exhibit 20).

complex, creative outputs, such as poetry.[14] A demo video of Claude's writing ability can be seen at the link below (in which Claude creates an outline of the plot for a novel about the crab "Claw'd" and then provides further suggestions to refine his story through a genogram and development of a metaphor related to LLMs).

24. Businesses utilize Claude to support their everyday operations and day-to-day needs. I am consistently surprised by the inventive and important uses people find for Claude in these contexts.

25. For example, Anthropic's partner, the University of Sydney, has used Claude to accelerate whale conservation efforts, specifically by using the model to analyze spectrograms, which are visual patterns created from an immense amount of acoustic data of whale sounds collected through a vast network of underwater microphones, to learn about the location and migration of minke whales.[15]

26. Another partner, Magic School, uses Claude to generate teaching tools such as curriculum plans, quizzes, behavior support plans, and report card comments to help teachers save time on administrative tasks so they can focus on interacting with students and avoid teacher burnout.[16]

27. Anthropic has also partnered at the government level to help the European Parliament improve access to over 2.1 million archived official government records dating back to 1952 by providing advanced search functionality, summarizing documents, making documents available in numerous languages, and allowing users to build reports about documents and extrapolate information about them.[17]

---

[14] Anthropic, *Claude 3.5 Sonnet as a writing partner*, YouTube (June 20, 2024), https://www.youtube.com/watch?v=-dWfl7Dhb0o (Exhibit 21); *see also* Anthropic, *Claude 3.5 Sonnet for sparking creativity*, YouTube (June 20, 2024), https://www.youtube.com/watch?v=rHqk0ZGb6qo (Exhibit 19).

[15] *The University of Sydney and Accenture Accelerate Whale Conservation with Claude*, Anthropic https://www.anthropic.com/customers/university-of-sydney (Exhibit 22)

[16] *MagicSchool Transforms K-12 Education for 3 Million Educators and Their Students with Claude*, Anthropic https://www.anthropic.com/customers/magicschool (last accessed Mar. 16, 2025) (Exhibit 23).

[17] *European Parliament expands access to their archives with Claude*, Anthropic https://www.anthropic.com/customers/european-parliament (last accessed Mar. 16, 2025) (Exhibit 24).

28. Novo Nordisk uses Claude as part of its FDA clinical drug trial process, reducing the time to produce reports on clinical trial data from 12 weeks to 10 minutes, which has the effect of speeding up the drug approval process for new treatments.[18]

29. Other examples of customer use cases include searching deposition transcripts,[19] evaluate film scripts for Hollywood studios,[20] accelerating accounting tasks,[21] and powering a chatbot available in English and Spanish to allow city employees and residents access information about health services.[22] Even more use cases are detailed at https://www.anthropic.com/customers.

D. The Goal of LLM Training Is To Give The Model Generative Capability

30. Anthropic's goal is for Claude to be capable of assisting users with a broad range of capabilities. For this reason, when Anthropic trains Claude, it wants the model to develop a generalized understanding of language patterns and relationships—not to memorize specific expressive content from the data on which it is trained.

31. For example, Anthropic's research paper *Scaling Laws and Interpretability of Learning from Repeated Data*[23] (Exhibit 30) details how memorization of training data is a problem that inhibits the ability of an LLM to generalize, rather than an intended feature of LLMs: "[G]eneralization and its internal implementation are disrupted when the model memorizes repeated data – a vivid illustration of the memorization-generalization trade-off." This is sometimes referred to as a model "overfitting" its training data. *Id.* ("Our results show that overfitting on the repeated data results in

---

[18] *Anthropic & AWS: Enterprise AI at Scale*, Anthropic https://assets.anthropic.com/m/319d41d452322da/original/Anthropic-AWS-info-sheet.pdf (Exhibit 25).

[19] *Steno Helps Attorneys Find the Critical Insights in Legal Transcripts with Claude*, Anthropic https://www.anthropic.com/customers/steno (last accessed Mar. 16, 2025) (Exhibit 26).

[20] *Jumpcut Helps Hollywood Find the Next Big Script with Claude*, Anthropic https://www.anthropic.com/customers/jumpcut (last accessed Mar. 16, 2025) (Exhibit 27).

[21] *Campfire Accelerates Accounting with Claude*, https://www.anthropic.com/customers/campfire (last accessed Mar. 16, 2025) (Exhibit 28).

[22] *AWS, Accenture and Anthropic Join Forces to Help Organizations Scale AI Responsibly*, Accenture (Mar. 20, 2024), https://newsroom.accenture.com/news/2024/aws-accenture-and-anthropic-join-forces-to-help-organizations-scale-ai-responsibly (Exhibit 29).

[23] Danny Hernandez et al., *Scaling Laws and Interpretability of Learning from Repeated Data* (May 21, 2022), https://arxiv.org/abs/2205.10487 (Exhibit 30).

worse test loss, and this co-occurs with a disproportionate degradation in the model's [performance]. . . .").

32. This is not a unique feature of Claude. In general, LLM makers seek to avoid overfitting and memorization of training data. *See also* Chiyuan Zhang et al., *Counterfactual Memorization in Neural Language Models* (Sept. 22, 2022)[24] (Exhibit 31), at 1 ("The paradigm of taking a single large-scale pre-trained model and fine-tuning it for many tasks motivates the study of these models' ability to generalize by avoiding memorizing their training data.") (internal citations omitted).

33. Focusing on the ability to generalize allows LLMs like Claude to understand a virtually limitless array of potential user prompts, and to complete a broad range of tasks by creating original outputs.

34. Anthropic has not developed Claude to function as a resource for users to generate or replicate copies of training data in response to queries and doing so would be antithetical to its purpose. Further, as explained in more detail below, if Anthropic detects that a user is attempting to evade to Anthropic's constitutional principle against copyright infringement (described below) and/or otherwise to elicit copyrighted material from Claude, Anthropic implements a variety of mitigations (described below) designed to refuse to provide the user with the requested output. These and other measures effectively prevent Claude from reproducing large portions of a book.

E. Building a Neural Network

35. To build an LLM like Claude, engineers first develop a "**neural network**," which is a computational model capable of learning patterns in language and concepts from enormous sets of data, including text and other data, called the "training corpus." These models are called neural networks because they are structured as an interconnected series of nodes ("**neurons**") in a layered formation that loosely mimics the structure of the human brain. These "nodes" or "neurons" are the model's basic computational elements, which process information throughout the model, passing tokens (representing words or bits of words) along through the "brain" when the model is asked to do something.

[24] https://proceedings.neurips.cc/paper_files/paper/2023/file/7bc4f74e35bcfe8cfe43b0a860786d6a-Paper-Conference.pdf

36. The neural network analyzes text in the context of surrounding language in the training data to learn language patterns and the relationship between the words and phrases in the training corpus. The neural network learns statistical, linguistic, and relationship information through an iterative multi-step training process. Through this iterative training on large amounts of data, the model builds a map of how humans use language without being explicitly programmed with human-defined rules. That is, the model itself learns structure and rules of language, from the myriad examples in the training data. The features of this map are stored within the model in an organized set of numerical values called "**parameters**," which include weights and biases, which effectively map the relationships among words and concepts across the full training corpus—*e.g.*, how strongly a concept like "dog" is associated with concepts like "puppy," instead of something like "smartphone." After training is complete, the model will use its parameters to analyze inputs from users and generate outputs or perform tasks in response.

37. Thus, developing a "generalized, functional understanding of how language works" means that the model has extracted statistical, linguistic, and relationship information from the training data, encoded this information as numerical values in its parameters, and is capable of passing prompt input text through those parameters to process it to generate new text that also comports with that understanding of how language functions.

F. Training Data Volume and Diversity

38. The iterative, multi-step LLM training process begins with the collection of a massive amount of data for training.[25] The training corpus needs to be diverse and vast—scale is necessary for the statistical, linguistic, and relationship information extracted from the data to be generalizable and for patterns to emerge.[26]

---

[25] Yiheng Liu et al., *Understanding LLMs: A Comprehensive Overview from Training to Inference* (Jan. 6, 2024) at 6-8, https://arxiv.org/pdf/2401.02038v2 (Exhibit 32); Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models* (Mar. 29, 2022) at 1, https://arxiv.org/pdf/2203.15556 (Exhibit 33) ("Kaplan et al. (2020) showed that there is a power law relationship between the number of parameters in an autoregressive language model (LM) and its performance. As a result, the field has been training larger and larger models, expecting performance improvements.").

[26] Liu et al., (Exhibit 32) at 6.

[redacted]

39. Anthropic, like other LLM companies, measures the size of its training corpora in "**tokens**." Tokens (described further below) are basic semantic units that can be represented mathematically, and can correspond to words, subwords, or characters. It generally takes 1-2 tokens to express a single word, but this conversion varies based on language, tokenizer design, and a number of other factors. For example, using the most recent version of Anthropic's tokenizer, it takes approximately 1.3 tokens to represent a single English-language word. Many trillions of tokens are required to properly train an LLM like Claude and to ultimately allow the LLM to recognize patterns in language to any useful degree. For example, the Claude 3.7 Sonnet model was trained on approximately [redacted] "sampled" tokens (and approximately [redacted] "base" tokens, i.e., the number of tokens after filtering and deduplication, which does not account for the fact that some tokens are used multiple times in the training process).

40. The more data (and more tokens) used in training LLMs, the better they are able to perform in a wide array of subject areas, reason, generalize beyond the corpus, and improve the quality, utility, and creativity of their outputs. The volume allows the LLM to gain a generalized understanding of how language works, so that it can interpret prompts and produce original outputs. In other words, the more data that the model is exposed to, the broader its "understanding," for lack of a better word, about human language and the world. At the same time, given the scale of the training data, any particular book (or other source of a modest number of tokens) has at best negligible effect on the effectiveness of training.

41. This scale is effectively a technological requirement, at least given the current state of the art: the company simply could not have built Claude using materially less data than it uses. Put differently, without using the order of magnitude of data that Anthropic used to train Claude, Claude literally could not exist.

42. The scale of training data is not all that matters; it is also important that the training data is diverse in subject matter and linguistic style to train the model effectively. LLMs not trained on diverse training data will perform poorly in both specific and general ways: they may be able to perform only a limited number of tasks related to that training data, and will also have a more limited general understanding of facts about the world and about how language works. *See* Brando Miranda

[redacted]

et al., *Beyond Scale: The Diversity Coefficient as a Data Quality Metric for Variability in Natural Language Data*[27] (Aug. 26, 2024) (Exhibit 34), at 2 ("In particular, the richness and variety of data, otherwise known as data diversity, is a key aspect of data quality that plays an important role in researchers' and practitioners' choice of pre-training corpora for general capabilities [citations]. In other words, diverse data is high quality data when your goal is to instill general capabilities in a model.") (internal citations omitted).

43. Thus, while it is important to train on long-form content like books, it is also important to train on short form content (like website chats) for the LLM to develop a complete understanding of the range of uses of language. *See* Priya Khandelwhal et al., *A Guide to Improving Long Context Instruction Following on Open Source Models*, Scale Blog (Sept. 26, 2024)[28] (Exhibit 35) ("High-quality, diverse training data is foundational to fine-tuning long context models. To build a model capable of excelling across different tasks, you need a range of input data that includes both long and short contexts. Task variety is key….When constructing datasets for long context training, pulling data from sources like books, encyclopedias, research papers, and code ensures a robust variety of instruction following tasks."); *see also* Tianyu Gao et al., *How to Train Long-Context Language Models (Effectively)* (Oct. 3, 2024)[29] (Exhibit 36) ("We find that using code repositories and long books as long-context data and mixing them with high-quality short-context data is crucial for both long-context performance and retaining the short-context capabilities of the pre-trained model.").

44. Finally, Anthropic has to rely on datasets composed substantially of writing from humans, as opposed to "synthetic data" (data generated by machines). Due to the impact that low quality synthetic data might have on outputs, it would be impossible to properly train an LLM like Claude entirely or predominantly on synthetic data.

45. **The Pretraining Data Corpus**. The training corpus (or "corpora" plural) used to train Claude is composed of a wide variety of source material—including scientific papers, computer code, and data from the web—of which books datasets are only a small part, in comparison to much larger

[27] https://arxiv.org/pdf/2306.13840

[28] https://scale.com/blog/long-context-instruction-following

[29] https://arxiv.org/pdf/2410.02660

[redacted]

volumes of text from across the entire Internet and computer code. *See, e.g.,* Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku 3* (2024) (Exhibit 37); Anthropic, *Model Card and Evaluations for Claude Models 2* (2023) (Exhibit 38).

46. The size of the training corpora for the various Claude models ranges from approximately [redacted] sampled tokens in Claude 1 to over [redacted] sampled tokens in Claude 3.7 Sonnet (the most recent model), and future models will use even larger training corpora.

47. **Use of Books for Pretraining Data.** Since its founding, Anthropic has used datasets of books as part of its training corpus for its LLMs. As with other types of pretraining data, it is important to obtain vast amounts of books and also to have diverse types of books in the training corpus to create a model with true generative capabilities. For example, while we typically value non-fiction and STEM books more highly than other types of books on an incremental basis, we also want fiction, books in a wide variety of genres that cover factual matters on a wide range of topics, and books in multiple languages. When Anthropic includes book datasets in its training corpus, it generally uses the entire text of the book. This is because of the need for a large volume of data and because using full books provides additional context for long context training that an excerpt and shorter source materials cannot provide.

48. **Internet Books Datasets**. In 2021, when Anthropic first started creating research models, it used certain datasets composed of books that it obtained from the Internet ("Internet books datasets"), as there were no other sources of books datasets available that were suitable to train LLMs. [redacted][30]

49. [redacted]

[30] It is also undoubtedly true that some books or portions of books have also been in other places in the training corpora for commercial models, such as in Common Crawl, which is a free, open repository of web crawl data. The full Common Crawl dataset includes over 250 billion webpages spanning 18 years. It is essentially a periodic copy of much of the Internet.

[redacted]

[redacted]

50. [redacted]

51. [redacted]

52. [redacted]

53. **Full Composition of Various Pretraining Data Mixes.** While I have addressed books specifically above, Anthropic's training corpora are composed of a wide variety of source material, including scientific papers, computer code, and data from the web, of which books datasets constitute only a small part. For example, the below chart illustrates the approximate composition of three training corpora for Claude 1 (the first commercial model) (*Fig. 1*), Claude 3 Opus [redacted] (*Fig. 2*), and Claude 3.5 Haiku [redacted] (*Fig. 3*).[32] The percentages show the approximate number of both "unique tokens" and "sampled tokens." **Unique tokens** show the percentage of the group of data with respect to the full training

[31] [redacted]

[32] Model Data Domain Breakdown (Exhibit 41).

[redacted]

corpus *before pretraining begins*. **Sampled tokens** show the percentage of the group of data with respect to the full training corpus *that were processed during training* (as certain tokens in the training corpus will be processed more than one time and others will not be processed).

Source of Tokens	% Unique Tokens	% Sampled Tokens
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

Fig. 1 - Claude 1 Training Corpus Composition
Total Sampled Tokens = [redacted]

Source of Tokens	% Unique Tokens	% Sampled Tokens
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

Fig. 2 - Claude 3 Opus Training Corpus Composition
Total Sampled Tokens = [redacted]

[33] ClaudeBot is the name of Anthropic's webcrawler that obtains data from the Internet for use in the training corpus.

[redacted]

Source of Tokens	% Unique Tokens	% Sampled Tokens
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

Fig. 3 - Claude 3.5 Haiku Training Corpus Composition
Total Sampled Tokens = [redacted]

G. Pretraining

54. **Data Processing and Tokenization**. After the massive quantity of data is collected and assembled into the training corpus, the data is processed (*e.g.*, by removing duplicates and obvious junk phrases, such as lorem ipsum). Deduplication is key because duplicated datasets can ultimately result in less intelligent outputs. Anthropic engages in deduplication efforts to prevent data points from inadvertently being weighted too heavily by the algorithm during the training process due to their frequency. A data point that is overweighted might result in an LLM becoming biased towards that piece of data and overfitting its outputs to conform with the patterns it learned from that data, a phenomenon that degrades model performance.

55. Another key part of data processing is called "**tokenization**." Tokenization refers to the process of converting human-readable text into binary computer code via a "tokenizer" and then breaking down this code into smaller numerical units called tokens. Tokens are words or portions of words in numerical form. Each token is assigned a number by its tokenizer that corresponds to its number in the model's token dictionary, which is a collection of all the unique tokens that a model recognizes along with their assigned numerical identifiers. Once the materials in the training corpus are transformed from human-readable text into the tokens, the tokens are typically grouped into

[redacted]

sequences, often around [redacted] tokens per group, although this can vary. These sequences are then shuffled to make random batches for training, which is meant to prevent the model from memorizing the corpus' original arrangements, or learning false relationships that only exist because of how the original corpus was assembled.

56. **Vectorization and Weights**. After the tokens are created, they are each converted into "**vectors**," which are mathematical representations of how words relate to other words. The model's "**weights**" and biases — numerical values that determine how the model understands vectors and their related concepts to be related — start off as random. Through the training process, which involves (very roughly) on the order of a million billion billion repeated mathematical calculations, the model adjusts the weights based on the vector inputs, as well as the vectors themselves, as it learns more about how humans use words and concepts in writing. The representation of that learning is a multidimensional matrix of weights that capture nuances in meaning; the relationship of words to one another; various grammatical properties; how words are modified in context with prefixes, suffixes and the like; and essentially all the other characteristics of language that are necessary to approximate the way humans use language, including, *e.g.*, semantic meaning (the meaning of the token in relation to other tokens); contextual relationships (how words relate to each other in context—"bank" as in financial institution versus "bank" as in river edge); syntactic information (grammatical properties like parts of speech and word tense); morphological information (information about token structure, prefixes, suffixes, and other word forms); positional information (where tokens appear in a sequence); cultural and social connotations (subtle associations that words carry based on usage in society); and hierarchical relationships (information about whether concepts are broader or narrower than others), along with other information, all of which are nuanced details beyond the tokens' plain definition. This iterative adjustment of the model's internal weight parameters allows it to learn complex patterns and meaningful concepts regarding language, transforming the initially random vectors into rich, contextual representations of the way language works. The outcome of the process is a set of numerical values (the "weights" and "biases") reflecting what the model has learned from the training corpus. The corpus itself is not stored in the model. And once trained, the model no longer uses the corpus.

[redacted]

57. **Forward pass.** During the "**forward pass**," the vectors go through a process called "matrix multiplication," in which each element of the vector is multiplied by corresponding elements in a weight matrix of numerical values and added to another numerical value called a "**bias**." This matrix multiplication computes a weighted sum for each neuron in the next layer. Functions called "**activation functions**" are then applied to the outputs in between layers. The weighted sum is then processed in the same way through numerous layers, each layer with its own weight and bias information. After passing the data through all of these sequential layers, the model will create a probability distribution to predict which token(s) are most likely to come next in a given sequence of tokens.

58. **Loss Calculation / Backpropagation / Gradient Descent.** After the Forward Pass has generated predictions of what token should come next in the sequence, the LLM training software measures how well the LLM did on the prediction through a technique called "**loss calculation**;" that is, it calculates the difference between the LLM's prediction and the value it expects, in comparison to what is actually in the training set as a whole. After calculating the difference, Anthropic uses a mathematical technique called "**backpropagation**" to determine how each weight in the model contributed to that difference between the LLM's prediction and the valid value from the training set. Then, in a process called "**gradient descent**," the model makes tiny, precise adjustments to the weights so that the next time data is passed through the model, it will make a better prediction. Overall, through this process the model learns meaningful patterns rather than memorizing specific examples. This happens across billions of weights, with each one getting a small nudge in the right direction. The entire process is done billions of times. Once that is complete, Anthropic has a pretrained model.

59. Together with the code that runs the model, it is the final numeric values of the weights resulting from this process that are the "model" or "LLM" itself—not the content of the training corpus that was used to train the model. After training is complete, the training corpus is not stored in the model or used in any way by the model. The model now functions without reference to the training data.

[redacted]

H. Fine-Tuning

60. A pretrained model is capable of functioning, but the quality of the generated outputs from the LLM can be improved—in the sense of making them more useful or relevant to specific tasks—by a process called "**fine-tuning**." Fine-tuning is a secondary training phase during which the model is further trained to achieve specific objectives instead of general ones, using techniques known as "supervised learning" and "reinforcement learning."

61. There are two common modalities for fine-tuning: supervised learning and reinforcement learning.

62. During supervised learning, Anthropic provides the model pairs of prompts and responses so that the model can see what a good response is to a given type of prompt. This data is typically validated by humans who are "supervising" the training by labelling datasets with approved responses. This is described in Anthropic's research paper *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*.[34]

63. During reinforcement learning, Anthropic gives the LLM positive or negative feedback for outputs via a model that evaluates the quality of the outputs. The model used in this stage outputs numerical scores rather than verbal or text outputs. *See* Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment* (Dec. 9, 2021) (Exhibit 43) at 14-15.

64. During fine-tuning, Anthropic implements its "Constitutional AI" principles, which among many other principles of good model behavior, teach Claude to avoid producing outputs that could constitute copyright infringement.

I. Inference

65. "**Inference**" is the process of running the LLM. Users supply inputs, like spreadsheets, photos, and/or questions, and Claude generates outputs. Because the model is generative and not deterministic (*i.e.*, the same input will not always produce the same output), there is no way to predict exactly how Claude will respond to a given input during the inference process. Rather, Claude generates probabilistic outputs based on predictions.

[34] Yuntao Bai et al., *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* (April 12, 2022), https://arxiv.org/pdf/2204.05862 (Exhibit 42).

[redacted]

J. Guardrails - Additional Suppression of Replication

66. Finally, Claude also has auxiliary system guardrails designed to suppress model output that could replicate training data, which helps stop outputs that could constitute copyright infringement (among other undesirable behaviors)—including its "Prompt Shield" and [redacted]. These auxiliary systems sit on top of the model—*i.e.*, they are products or features that are separate from the model's training or fine-tuning itself.

67. Claude's Prompt Shield detects when a user attempts to generate an output that would violate certain elements of Anthropic's Terms of Service (Exhibit 44)[35] including by eliciting outputs that contain copyrighted content. In the case of such prohibited user prompts, Prompt Shield will, behind the scenes, add to the user-provided prompt a request that is invisible to the user, asking Claude to be particularly considerate and careful not to substantially reproduce copyrighted content when responding. Claude will then tailor its output to avoid replication of the training data or refuse to respond to the prompt entirely.

68. [redacted]

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct. Executed on March 27, 2025, in San Francisco, California.

JARED KAPLAN

[35] Consumer Terms of Service, available at https://www.anthropic.com/legal/consumer-terms ("You may not access or use, or help another person to access or use, our Services in the following ways: … To infringe, misappropriate, or violate intellectual property or other legal rights…").

[redacted]

CERTIFICATE OF SERVICE

I, Douglas A. Winthrop, am the ECF user whose identification and password are being used to file the foregoing **DECLARATION OF JARED KAPLAN IN SUPPORT DEFENDANT ANTHROPIC PBC's MOTION FOR SUMMARY JUDGMENT**.

Dated: March 27, 2025

/s/ *Douglas A. Winthrop*