EXHIBIT 188

PageVault

Document title:	Scaling Laws and Interpretability of Learning from Repeated Data \ Anthropic
Capture URL:	https://www.anthropic.com/research/scaling-laws-and-interpretability-of-learning-from-repeated-data
Captured site IP:	178.63.171.13
Page loaded at (UTC):	Wed, 22 Jan 2025 16:53:50 GMT
Capture timestamp (UTC):	Wed, 22 Jan 2025 16:53:51 GMT
Capture tool:	3.2.0
Collection server IP:	54.174.78.137
Browser engine:	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/131.0.0.0 Safari/537.36
Operating system:	linux x64 (Node v20.18.0)
Capture ID:	ncjSZ98WJdUHPXxNKGibFC

EXHIBIT 136 tabbies

PDF REFERENCE #: gBgTttiGYrCGnNHe3yu6BU

Interpretability Research

Scaling Laws and Interpretability of Learning from Repeated Data

May 21, 2022

Read Paper

Abstract

Recent large language models have been trained on vast datasets, but also often on repeated data, either intentionally for the purpose of upweighting higher quality data, or unintentionally because data deduplication is not perfect and the model is exposed to repeated data at the sentence, paragraph, or document level. Some works have reported substantial negative performance effects of this repeated data. In this paper we attempt to study repeated data systematically and to understand its effects mechanistically. To do this, we train a family of models where most of the data is unique but a small fraction of it is repeated many times. We find a strong double descent phenomenon, in which repeated data can lead test loss to increase midway through training. A predictable range of repetition frequency leads to surprisingly severe degradation in performance. For instance, performance of an 800M parameter model can be degraded to that of a 2x smaller model (400M params) by repeating 0.1% of the data 100 times, despite the other 90% of the training tokens remaining unique. We suspect there is a range in the middle where the data can be memorized and doing so consumes a large fraction of the model's capacity, and this may be where the peak of degradation occurs. Finally, we connect these observations to recent mechanistic interpretability work - attempting to reverse engineer the detailed computations performed by the model - by showing that data repetition disproportionately damages copying and internal structures associated with generalization, such as induction heads, providing a possible mechanism for the shift from generalization to memorization. Taken together, these results provide a hypothesis for why repeating a relatively small fraction of data in large language models could lead to disproportionately large harms to performance.

Authors

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Jared Kaplan

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Jared Kaplan

Related

See All

Case Study

Brian Impact Foundation powers their search for the next generation of social innovators with Claude

Case Study

Perplexity delivers factual and relevant answers with Claude

Case Study

Pulpit AI turns sermons into multiple pieces of content with Claude

Footer

© 2025 Anthropic PBC