

# 거대언어모델과

# 저작권

ARTIFICIAL INTELLIGENCE / TECH / COPYRIGHT

## Sarah Silverman is suing OpenAI and Meta for copyright infringement / The lawsuits allege the companies trained their AI models on books without permission.

By [Wes Davis](#), a weekend editor who covers the latest in tech and entertainment. He has written news, reviews, and more as a tech journalist since 2020.

Jul 10, 2023, 3:14 AM GMT+9 | [142 Comments](#) / [142 New](#)



**Sarah Silverman**

**Christopher Golden**

**Richard Kadrey**



Sarah Silverman의 The Bedwetter를 한국어로 요약해줘.



Sarah Silverman의 "The Bedwetter: Stories of Courage, Redemption, and Pee"는 그녀의 개인적인 이야기와 경험을 기반으로 한 자서전입니다. 아래는 주요 내용을 간단하게 요약한 것입니다.

1. **어린 시절:** Silverman은 뉴햄프셔주의 작은 도시에서 유대인 가정에서 자랐습니다. 그녀는 특히 자신의 신체와 스스로를 받아들이는 데 어려움을 겪었으며, 그 중 하나가 오랜 기간 동안 침대에 소변을 누는 문제였습니다.
2. **성장과 발전:** 그녀는 자신의 불편함과 문제점을 유머로 전환하여 자신만의 특별한 스타일을 발전시켰습니다. 그녀는 또한 자신의 문제와 투쟁을 직면하며 성장하는 과정에서 얻은 교훈을 공유합니다.
3. **코미디 케어리어:** Silverman은 코미디언으로서의 그녀의 경력을 통해, 성공, 실패, 그리고 그녀를 독특하게 만든 일들에 관한 내용을 나눕니다. 그녀의 유머 스타일은 대부분 까칠하고 경계를 허무는 것으로 알려져 있습니다.
4. **인간관계:** Silverman은 가족, 친구, 동료와의 관계에 대한 개인적인 이야기도 나눕니다. 그녀의 이야기는 사랑, 헌신, 그리고 유머를 통해 어려움을 극복하는 방법에 대한 교훈을 전합니다.
5. **사회 문제와 사상:** 그녀는 책에서 또한 사회의 여러 문제, 특히 성별, 인종, 종교와 관련된 문제에 대한 그녀의 견해와 생각을 나눕니다.

결론적으로, "The Bedwetter"는 Sarah Silverman의 유머 감각, 사회에 대한 독특한 시각, 그리고 인생의 투쟁과 극복에 관한 깊은 통찰력을 보여줍니다.

# LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron\*, Thibaut Lavril\*, Gautier Izacard\*, Xavier Martinet  
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal  
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin  
Edouard Grave\*, Guillaume Lample\*

Meta AI

## Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community<sup>1</sup>.

## 1 Introduction

Large Languages Models (LLMs) trained on massive corpora of texts have shown their ability to perform new tasks from textual instructions or from a few examples (Brown et al., 2020). These few-shot properties first appeared when scaling models to a sufficient size (Kaplan et al., 2020), resulting in a line of work that focuses on further scaling these

performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance compared to the best existing LLMs. For instance, LLaMA-13B outperforms GPT-3 on most benchmarks, despite being 10× smaller. We believe that this model will help democratize the access and study of LLMs, since it can be run on a single GPU. At the higher-end of the scale, our 65B-parameter model is also competitive with the best large language models such as Chinchilla or PaLM-540B.



ELEutherAI

[About](#) [Research](#) [Papers](#) [Releases](#) [Blog](#)



# Empowering Open-Source Artificial Intelligence Research

# The Pile

An 800GB Dataset of Diverse Text for Language Modeling

## What is the Pile?

---

The Pile is a **825 GiB** diverse, open source language modelling data set that consists of 22 smaller, high-quality datasets combined together.

[Pile Paper \(arXiv\)](#)

## Download

---

The Pile is hosted by [the Eye](#).

[Download Pile](#)

The format of the Pile is [jsonlines](#) data compressed using [zstandard](#).

Have a model that uses or evaluates on the Pile? [Let us know!](#)

*[Submitted on 31 Dec 2020]*

## The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, Connor Leahy

Recent work has demonstrated that increased training dataset diversity improves general cross-domain knowledge and downstream generalization capability for large-scale language models. With this in mind, we present \textit{the Pile}: an 825 GiB English text corpus targeted at training large-scale language models. The Pile is constructed from 22 diverse high-quality subsets -- both existing and newly constructed -- many of which derive from academic or professional sources. Our evaluation of the untuned performance of GPT-2 and GPT-3 on the Pile shows that these models struggle on many of its components, such as academic writing. Conversely, models trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, while improving performance on downstream evaluations. Through an in-depth exploratory analysis, we document potentially concerning aspects of the data for prospective users. We make publicly available the code used in its construction.

Subjects: **Computation and Language (cs.CL)**

Cite as: [arXiv:2101.00027](https://arxiv.org/abs/2101.00027) [cs.CL]

(or [arXiv:2101.00027v1](https://arxiv.org/abs/2101.00027v1) [cs.CL] for this version)

<https://doi.org/10.48550/arXiv.2101.00027> 



“a copy of the contents  
of the **Bibliotik**  
private tracker.”

# “shadow libraries” listed

# legal or illegal?

# stochastic reasoning?

# shadow libraries!

AI가 기존 저작물 기반으로  
창작물을 만든다는 사실만으로도  
저작권 위반 정당화 불가능

원본고 동일하지 않은

새로운 저작물 생성

하기때문

새로운 저작물의

일정 비율이

원본과 동일하다는 것이

입증되어야



음악:

신곡과 기존 곡

유사성 논쟁

**"연속 4-5마디 동일"**

# 생성 AI

## 작동원리 이해

### → 생성 AI 정의(Definition)

# 학습 데이터

→ 패턴 → 가중치

→ Interpolation

by Parameters

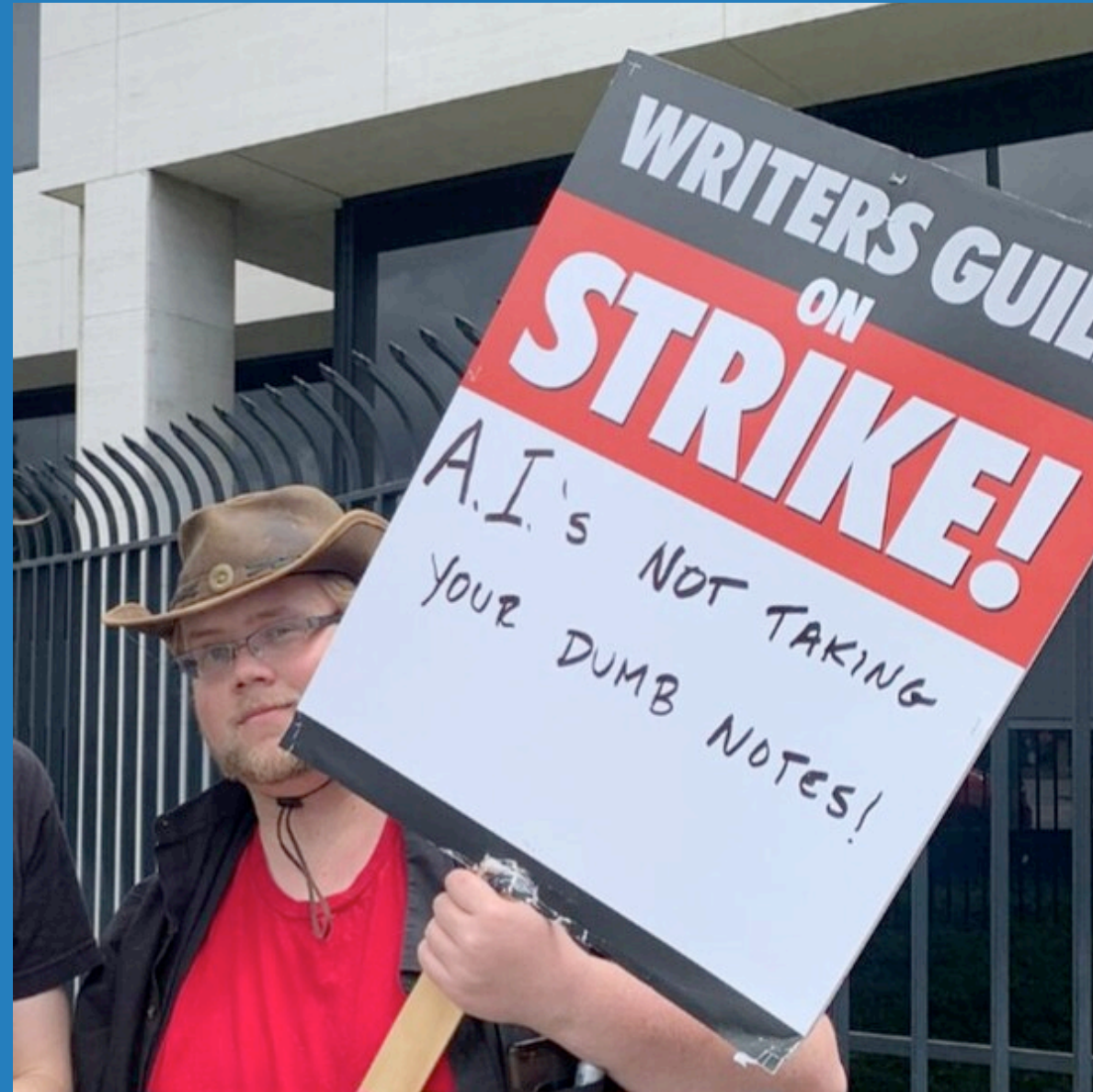
# 분자적 보간

# atomic interpolation

# 분자적 보간 by parameters

# Problem 1:

since 23.05.02





# Writers Guild of America


The New York Times

Hollywood Strikes | Actors' Strike | Writers' Strike | A Looming Deadline | How TV Will Be Affected | Should You Go to the Movies?

## Hollywood Writers Go on Strike, Halting Production

The dispute, which pits 11,500 television and screenwriters against the major studios, has shattered 15 years of labor peace in the entertainment business.

Give this article | Share | Bookmark | 956 Comments



Television and screenwriters picketed outside Netflix's office building in Los Angeles and on the streets of New York City. Mark Abramson for The New York Times



# SAG-AFTRA against AI-Revolution



The screenshot shows the top portion of a news article on The Guardian's website. At the top right is the 'The Guardian' logo. Below it is a navigation bar with categories: News, Opinion, Sport, Culture, Lifestyle. A 'Support us' button with a right-pointing arrow is on the left. Below the navigation bar is a secondary menu with categories: World, UK, Climate crisis, Environment, Science, Global development, Football, Tech, Business. The main image shows a group of people at a protest, with a woman in the foreground holding a sign that reads 'SAG-AFTRA ON STRIKE!'. Below the image is the article title: 'Artificial intelligence (AI) 'Bargaining for our very existence': why the battle over AI is being fought in Hollywood'. The byline reads 'Lois Beckett in Los Angeles and Kari Paul in Oakland'. The date and time are 'Sat 22 Jul 2023 12:00 BST'.

Support us →

The Guardian

News Opinion Sport Culture Lifestyle

World UK Climate crisis Environment Science Global development Football Tech Business



**Artificial intelligence (AI)**  
**'Bargaining for our very existence': why the battle over AI is being fought in Hollywood**

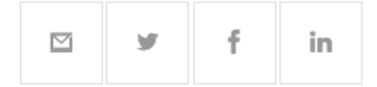
The ramifications of artificial intelligence are of concern to the actors and writers on strike - from big stars to bit players

**Lois Beckett in Los Angeles and Kari Paul in Oakland**

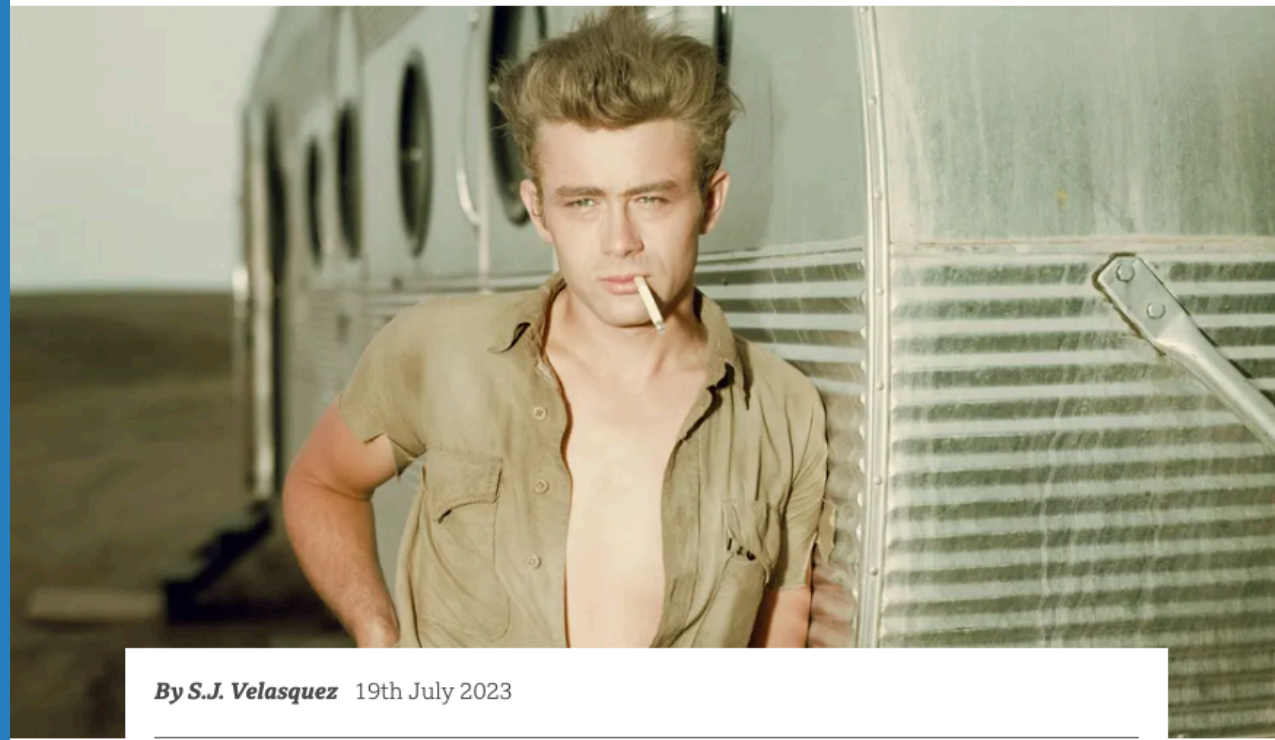
Sat 22 Jul 2023 12:00 BST

FUTURE NOW | ARTIFICIAL INTELLIGENCE

# How AI is bringing film stars back from the dead



(Image credit: Getty Images)



By S.J. Velasquez 19th July 2023

Celebrities such as James Dean can be brought back to life as digital clones thanks to the power of artificial intelligence, but it is raising troubling questions about what rights any of us have after we die.

# Problem 2:

# 과학적 프레임워크

# vs. 상업용 애플리케이션

# Solution 1:

BREAKING

# Grimes Helps Artists Distribute Songs Using Her AI Voice—If They Split Royalties. Here's How It Works.

Antonio Pequeño IV

Forbes Staff

*I cover breaking news.*

Follow

Jun 12, 2023, 05:41pm EDT



# Solution 2:

AI

# Shutterstock expands deal with OpenAI to build generative AI tools

Kyle Wiggers

@kyle\_l\_wiggers / 2:52 AM GMT+9 • July 12, 2023

# OpenAI



# 사회적 공익에 대한

# 예술가와 작가의 기여

# Solution ?



THE  
NEW YORKER

ANNALS OF ARTIFICIAL INTELLIGENCE

## THERE IS NO A.I.

*There are ways of controlling the new technology—but first we have to stop mythologizing it.*

By Jaron Lanier  
April 20, 2023



# Data Dignity

# Data Dignity

**= Data as Labor**

**people could get paid for**

**what they create,**

**even when it is filtered**

**and recombined through big  
models,**



**Implementing data dignity**

**will require**

**technical research**

**and policy innovation.**

# Opening

# the black box

# against UBI

**the goal should be**

**to at least establish**

***a new creative class***

**instead of a new dependent class.**