# Testimony of Prof. Gregory Durrett

*U.S. et al. v. Google LLC,* No. 1:20-cv-03010-APM

United States District Court for the District of Columbia

Ex. No. PXRD003 1:20-cv-03010-APM 1:20-cv-03715-APM

April 21, 2025

**Assignment**: Assess whether and how Google Search, Google's Search Index, and User-Side Data may be used to build and improve generative artificial intelligence ("Gen AI") search features and applications

**Rebuttal Assignment 1**: Assess Dr. Allan's assertion that Qualified Competitors could use Large Language Models (LLMs) to "reverse engineer" Google's technology

**Rebuttal Assignment 2**: Assess Dr. Hitt's technical claims regarding (a) the value of data to Gen AI models and (b) the Gemini App's reliance on Google Search

## Google's Gen Al Models and Products integrate Google Search, Google's Search Index, and User-side data

Google's advantages in Search assist it in building high quality Gen Al Models and Products Dr. Allan's assertion that Plaintiffs' proposals would allow competitors to "reverse engineer" Google's technology using LLMs is misleading

Dr. Hitt mischaracterizes the use of data for pre-training and finetuning Gen AI models

Dr. Hitt understates how critically Google's Gemini App relies on Search for grounding



- I. Introduction to Generative AI
- II. Google's advantages in Search assist in building high quality Gen Al Models and Gen Al Search Products
  - A. Al Overviews
  - B. Base Models
  - C. Gemini App
- III. Dr. Allan's misleading "reverse engineering" claim
- IV. Dr. Hitt's flawed technical analysis

## AI, Machine Learning, Gen AI, And LLMs Are Related



#### **AI Overviews**

what are the top tourist destinations in california?

All Images Short videos News Forums Shopping Videos : More

Results for California · Choose area

#### 🔶 Al Overview

California boasts a wide array of popular tourist destinations. Major cities like **San Francisco and Los Angeles**, along with iconic landmarks and natural wonders like Yosemite National Park and the Golden Gate Bridge, draw millions of visitors annually. Other frequently visited areas include San Diego, <u>Redwood National and State Parks</u>, <u>Lake</u> <u>Tahoe</u>, and <u>Palm Springs</u>.



### **Gemini App**

what are the top tourist destinations in california?

California is a vast and diverse state with a wide array of attractions, making it a top tourist destination. Here are some of the most popular and highly-regarded places to visit:

Major Metropolitan Areas:

- Los Angeles: The entertainment capital of the world, home to Hollywood, Universal Studios, Griffith Observatory, the Getty Center, Santa Monica Pier, and vibrant beaches like Venice Beach.
- San Francisco: Known for its iconic Golden Gate Bridge, Alcatraz Island, Fisherman's Wharf, cable cars, diverse neighborhoods like Chinatown and North Beach, and cultural institutions like the Exploratorium and the California Academy of Sciences.
- San Diego: Boasts beautiful beaches like La Jolla and Coronado, Balboa Park (home to the

Language modeling: predict the most likely next token given previous tokens

**Token**: a unit of language, approximately one short word



### Predicting the next token is probabilistic.



These probabilities sum up to 100% over the vocabulary of all tokens.

LMs can predict next tokens repeatedly:

Once upon a time, there was a \_\_\_\_\_ war between two kingdoms

LMs can be applied to carefully crafted prompts:

Q: When was Abraham Lincoln born? A: 1809

The translation of "I like your new hat" into Spanish is

*"Me gusta tu...* 

A "perfect" language model can do many useful tasks if it has seen enough data.

## How Transformer Language Models Work



Most LLMs use **Transformers:** neural networks with many parameters (billions!).

Parameters define mathematical operations to calculate predicted token probabilities.



## LLMs Pre-Train On Data



### Base models are good at text completion

Once upon a time, there was a \_\_\_\_\_.

war between two kingdoms

verdant land known as ...

## LLMs May Be Limited In Functionality After Pre-Training



Base models are good at text completion

Once upon a time, there was a \_\_\_\_\_.

war between two kingdoms

verdant land known as ...

Base models are not always good at other tasks

*Q: When was Abraham Lincoln born? A:* \_\_\_\_.

*Q: When was George Washington born? A:* \_\_\_. *Q: When was John Adams born? A:* \_\_\_ [...]

## LLMs Are Post-Trained To Accomplish Various Tasks



## LLMs May Be Post-Trained To Accomplish More Than One Task



All-purpose model, such as Google Gemini

A model can only predict "*Texas is a* \_\_\_\_\_" is followed by "*state*" if it has seen this information in its training data.

If we ask ChatGPT "Write me a biography of [X]" for various people X, how factual is it?



r/microwavegang ∙1 day ago Bobbydibi

...

MMMMMMMMMM

Mmmmm

m

 $\leftarrow$ 

**DCLM**: open-source LLM pre-training dataset from Common Crawl



**DCLM-Baseline** compared to **FineWeb edu** (dataset from the company Hugging Face)



Suppose an LLM is trained in October 2024.

How can that LLM answer: "who was Kendrick Lamar's musical guest at the 2025 Super Bowl halftime show?"



One solution: add post-Oct 2024 docs to pre-training data and retrain the model? Cons: expensive and time-consuming (pre-training takes weeks or months)

### **Closed-book LLM generation**

Who was Kendrick Lamar's musical guest at the 2025 Super Bowl halftime show?



I don't know who Kendrick Lamar's musical guest was...

### **Retrieval-augmented generation (RAG)**

Who was Kendrick Lamar's musical guest at the 2025 Super Bowl halftime show?



Reformulated query and document text



SZA performed with Kendrick Lamar... 1:20-cv-03010-APM

1:20-cv-03715-APM

SOOG-DO-1-34206171



Pranesh Srinivasan and Anant Nawalgaria Just as humans rely on memory systems as seen in Figure 1 to recall facts, to learn, and to reason, Generative AI models rely on search. While AI models are brilliant at generating content, they need a way to anchor their outputs in reality. Search is what anchors an AI model's outputs in reality. It is a system that acts as a rapid and reliable technique to pull relevant information at lightning speed. Additionally, when these models make a claim, search assists them in verifying the model's accuracy by citing the correct sources and fast-checking any assumptions. This combination of Gen AI and search has propelled technological advancements and reshaped the digital landscape.

"While AI models are brilliant at generating content, they need a way to anchor their outputs in reality. Search is what anchors an AI model's output in reality."

Google

Y CONFIDENTI

### **OpenAI released a dataset called "SimpleQA" with example questions:**

What is the first and last name of the woman whom the British linguist Bernard Comrie married in 1985?

#### Closed-book GPT-40 accuracy: 38.2%

#### **RAG makes this easy:**

Personal life [edit]				
Early life and education [edit]				
Comrie was born in Sunderland, England on 23 May 1947. He eaundergraduate and doctoral degrees in Modern and Medieval Lau Linguistics from the University of Cambridge, <sup>[2][3]</sup> where he also t Linguistics until he moved to the Linguistics Department of the University California. <sup>[4]</sup>				
Personal life [edit]				
He married linguistics professor Akiko Kumahira in 1985. <sup>[5][6]</sup>				

#### Without RAG, contemporary GPT models cannot do this.



If repeatedly queried, the model will give a variety of different and incorrect answers.

## **Search Engines Are Retrieval Tools**

#### LLM

- can generate novel content, which can include mistakes
- retrained infrequently



### Search Engine

- no novel content is generated
- updated continuously



#### RAG

- generates novel content from search results
- updated continuously





- I. Introduction to Generative AI
- II. Google's advantages in Search assist in building high quality Gen AI Models and Gen AI Search Products
  - A. Al Overviews
  - B. Base Models
  - C. Gemini App
- III. Dr. Allan's misleading "reverse engineering" claim
- IV. Dr. Hitt's flawed technical analysis

User-side Data

- Google receives **9x** more queries each day than all of its rivals *combined*
- On mobile, Google receives 19x more queries than all of its rivals put together
- User data is a critical input that directly **improves quality** for general search engines at every stage

28

User data is important for crawling

Search Index

- User data is important for indexingOnly Google and Bing create fulsome web search indexes
- Bing's index is poor quality

Common Crawl captured approximately 27 billion pages in the year 2024. By comparison, Google crawls billion pages *per day*.



- I. Introduction to Generative AI
- II. Google's advantages in Search assist in building high quality Gen Al Models and Gen Al Search Products

## A. Al Overviews

- B. Base Models
- C. Gemini App
- III. Dr. Allan's misleading "reverse engineering" claim
- IV. Dr. Hitt's flawed technical analysis

## **AI** Overviews

- RAG system
- Triggered on % of Google
  Search queries.



#### 🔶 Al Overview

California boasts a wide array of popular tourist destinations. Major cities like **San Francisco and Los Angeles**, along with iconic landmarks and natural wonders like Yosemite National Park and the Golden Gate Bridge, draw millions of visitors annually. Other frequently visited areas include San Diego, Redwood National and State Parks, Lake Tahoe, and Palm Springs.



Al Overviews relies on the Google Search stack to retrieve and generate results



## One Google Search system used by AI Overviews is Fast Search



Generator for AI Overview is "MAGIT"

- Gemini base model fine tuned on query data
- Fined-tuned to generate text in AI Overviews format



wagi	WAGIT-V2.2 [V2-5]
Train	ing Data
NTK/Attorne	y Client Privileged
Self link:	
Authors &	Contributors: Kishan Kolur Louis O'Bryan Vivian Ma
Created: 17	Jul 2024
Updated: 2	2 Jul 2024
Status: Fina	
Context	
This docum	ent describes the data used for training the Magi MAGIT-v2.2 [v2-S].







## Phiroze Parakh

Senior Director

12	Q.	Okay. And are the signals that Tangram
13	uses deri	ved from the Glue model?
14	Α.	Yes.
15	Q.	And the Glue model is user feedback data?
16	Α.	Yes.

Designated Testimony



- I. Introduction to Generative AI
- II. Google's advantages in Search assist in building high quality Gen Al Models and Gen Al Search Products
  - A. Al Overviews
  - **B.** Base Models
  - C. Gemini App
- III. Dr. Allan's misleading "reverse engineering" claim
- IV. Dr. Hitt's flawed technical analysis



- Google pre-trains Gemini models on GCC
- GCC and the Search Index are derived from Docjoins

#### Google Common Corpus (GCC): Proposal

20/100gle-common-corpus-proposal August-November 2021 Author Reviewers Visibility: Confidential (Do not share externally) Status: Reviewed

Executive Summary: this is a proposal to build and give access to a corpus of Docjoins and Imagejoins that are suitable for use outside Search products. The corpus will only contain documents and images considered generally crawlable (i.e., not restricted to Googlebot), with an expected coverage of over of the docjoins and if the imagejoins. It will be implemented as a low-cost, self-hosted solution on top of Core Data Marketplace technology, offering highly-scalable, standard lookup and scanning interfaces.

#### Motivation

The main corpus of <u>Docioins</u> is a large repository of the documents publicly available on the web and visited at least once by Googlebot in the last few months. It currently consists of over

B documents. It is obviously an extraordinarily useful dataset of public content. By comparison, the external <u>Common Crawl</u> corpus is much smaller, with only a bit over B in the latest release. Google's ability to build such a large corpus is to some extent enabled by the *willingness of content publishers to have their content crawled by Googlebot for the purpose of appearing in search* results and other Search products. This means that webmasters may or may not be willing to have this content used for other purposes. At the same time, it would be very wasteful, particularly of the bandwidth and serving costs of web sites, to simply build a separate corpus of Docjoins fetched and kept up-to-date with a separate non-Search crawler. Moreover, we fully expect a large overlap between what is crawlable only with Googlebot (for Search) and what webmaster would be willing to allow other crawlers to retrieve for different purposes. Ideally, we would simply identify Docjoins that are generally available, and make them easily usable for non-Search lines of business at Google without any extra crawling.

"[Docioins] currently consists of over B documents. It is obviously an extraordinarily useful dataset of public content. By comparison, the external Common Crawl corpus is much smaller, with only a bit over 3 B in the latest release."

- Google pre-trains Gemini models on GCC
- GCC and the Search Index are derived from Docjoins

#### Google Common Corpus (GCC): Proposal

go/google-common-corous-proposal August-November 2021 Author: Reviewers: Visibility: Confidential (Do not share externally) Status: Reviewed

Executive Summary: this is a proposal to build and give access to a corpus of Docjoins and Imagejoins that are suitable for use outside Search products. The corpus will only contain documents and images considered generally crawlable (i.e., not restricted to Googlebot), with an expected coverage of over of the docjoins and of the imagejoins. It will be implemented as a low-cost, self-hosted solution on top of Core Data Marketplace technology, offering highly-scalable, standard lookup and scanning interfaces. Besides considering the broad use of documents primarily fetched for building Google's Search index, it is important to be aware that the corpus of Docjoins includes not only document content but also the myriad of Search metadata and signals powering the internals of Google Search. Some of these signals are clearly very sensitive, being derived from aggregated user behavior. Expanding the uses of Docjoins must address both the intent of the webmaster and the need to limit access to only those fields in the Docjoin that are appropriate for non-Search experience products (as defined in <u>go/search-business-data-guidelines</u>).

"[I]t is important to be aware that the corpus of Docjoins includes not only document content but also the myriad of Search metadata and signals powering the internals of Google Search."

## Search signals can be used for filtering

Message	
From:	Fangxiaoyu Feng [ @google.com]
Sent:	6/5/2024 5:14:50 PM
To:	Claire Cui 9google.com]; Paul Haahr @google.com]
CC:	Vahab Mirrokni @google.com]; Andrew Dai @google.com]; Jeff Dean [
	google.com]; Pandu Nayak @google.com]; Urs Hölzle @google.com]; Anca Dragan
	google.com]; Oriol Vinyals [ @google.com]; Jack Rae [ ;oogle.com]; Sebastian Borgeaud
	google.com]; Ethan Dyer @google.com]; Katie Millican @google.com]
Subject:	Re: PageRank for training data?

Hi,

Thanks for the discussion. It has helped me to learn more about the value of Search signals. With <u>@Paul</u> <u>Haahr</u>'s help, we just got approval to use several Search signals to help Gemini pretraining, including QScore, NSR and several low quality page/site signals. Those signals will be very helpful for us to upweighting good, authoritative pages and downweighting the spammy, untrustable ones. We just started conversation with multiple teams in Search today, to discuss how to access those signals.

For those sensitive search signals, we will only allow a few relevant folks to access the raw score. We will experiment different ways to use them (including the upsampling and downsampling) soon, but only bucketing the scores to make them more fine-grained than current Search indexing tiers.

Best, Xiaovu "[W]e just got approval to use several Search signals to help Gemini pretraining... Those signals will be very helpful for us to upweight[] good, authoritative pages and downweight[] the spammy, untrustable ones."

## **Google Has Approved Using Query Data To Build Gemini Base Models**



Designated Testimony



- I. Introduction to Generative AI
- II. Google's advantages in Search assist in building high quality Gen Al Models and Gen Al Search Products
  - A. Al Overviews
  - B. Base Models
  - C. Gemini App
- III. Dr. Allan's misleading "reverse engineering" claim
- IV. Dr. Hitt's flawed technical analysis

- Mobile application or web application
- Allows users to query Gemini models (fine-tuned from Gemini base models)

what are the top tourist destinations in california?

California is a vast and diverse state with a wide array of attractions, making it a top tourist destination. Here are some of the most popular and highly-regarded places to visit:

#### Major Metropolitan Areas:

- Los Angeles: The entertainment capital of the world, home to Hollywood, Universal Studios, Griffith Observatory, the Getty Center, Santa Monica Pier, and vibrant beaches like Venice Beach.
- San Francisco: Known for its iconic Golden Gate Bridge, Alcatraz Island, Fisherman's Wharf, cable cars, diverse neighborhoods like Chinatown and North Beach, and cultural institutions like the Exploratorium and the California Academy of Sciences.
- San Diego: Boasts beautiful beaches like La Jolla and Coronado, Balboa Park (home to the



## Gemini App grounds results





- I. Introduction to Generative AI
- II. Google's advantages in Search assist in building high quality Gen Al Models and Gen Al Search Products
  - A. Al Overviews
  - B. Base Models
  - C. Gemini App
- III. Dr. Allan's misleading "reverse engineering" claim
- IV. Dr. Hitt's flawed technical analysis

Dr. Allan claims that Plaintiffs' data-sharing remedies will allow rivals with access to LLMassisted search engines to "reverse engineer" some Google's technologies

- Reverse engineering some Google technologies does not constitute reverse engineering Google Search



22	I want to tie those back to search. And so	
23	if you built an LLM that was trained on a search	
24	engine task, that model is still not capable of doing	
25	what a search a traditional search engine does;	
1	correct?	
4	THE WITNESS:	
5		
6		
7		
-		Deposition

Deposition Testimony

## Controllable AI

 Search over the years have developed a mature system for triage bad losses and for applying point-fix if necessary. The black-box nature of LLM makes it much harder to have this level of control. I believe this is an area that needs a lot of research. There are many loss-patterns that we don't quite know how to fix, except for engineering lots of data which can be quite time consuming and inefficient.

• A mimicking LLM cannot be analyzed to reveal its internal mechanisms



- Producing a mimicking system with an LLM does not reveal any component substructure of Google's system.
- Dr. Allan does not say that a mimicking system would be as performant as the original system.

## **Dr. Allan's Opinions Are Missing Context**



- I. Introduction to Generative AI
- II. Google's advantages in Search assist in building high quality Gen Al Models and Gen Al Search Products
  - A. Al Overview
  - B. Base Models
  - C. Gemini App
- III. Dr. Allan's misleading "reverse engineering" claim
- **IV.** Dr. Hitt's flawed technical analysis

- Dr. Hitt's technical claim that "there are many sources of accessible, highquality, and substitutable data" for training LLMs is wrong
- Dr. Hitt's analysis is flawed:



- Dr. Hitt does not consider Google's Search Index in his assessment of pre-training data sources
- Dr. Hitt ignores how Google can use User-side data to filter pre-training data for Gemini Base Models
- Dr. Hitt ignores how Google uses User-side data to develop and improve AI Overviews





**Lorin Hitt** Google Economics Expert



result from the model.<sup>215</sup> Further, I understand that currently the Gemini app uses the Google Search API results for grounding its responses to only approximately % of prompts.<sup>216</sup> The

» The Gemini App calls upon the Search API for around percent of prompts, though this figure is fluid.

Expert Report

## Google's Gen Al Models and Products integrate Google Search, Google's Search Index, and User-side data

Google's advantages in Search assist it in building high quality Gen Al Models and Products