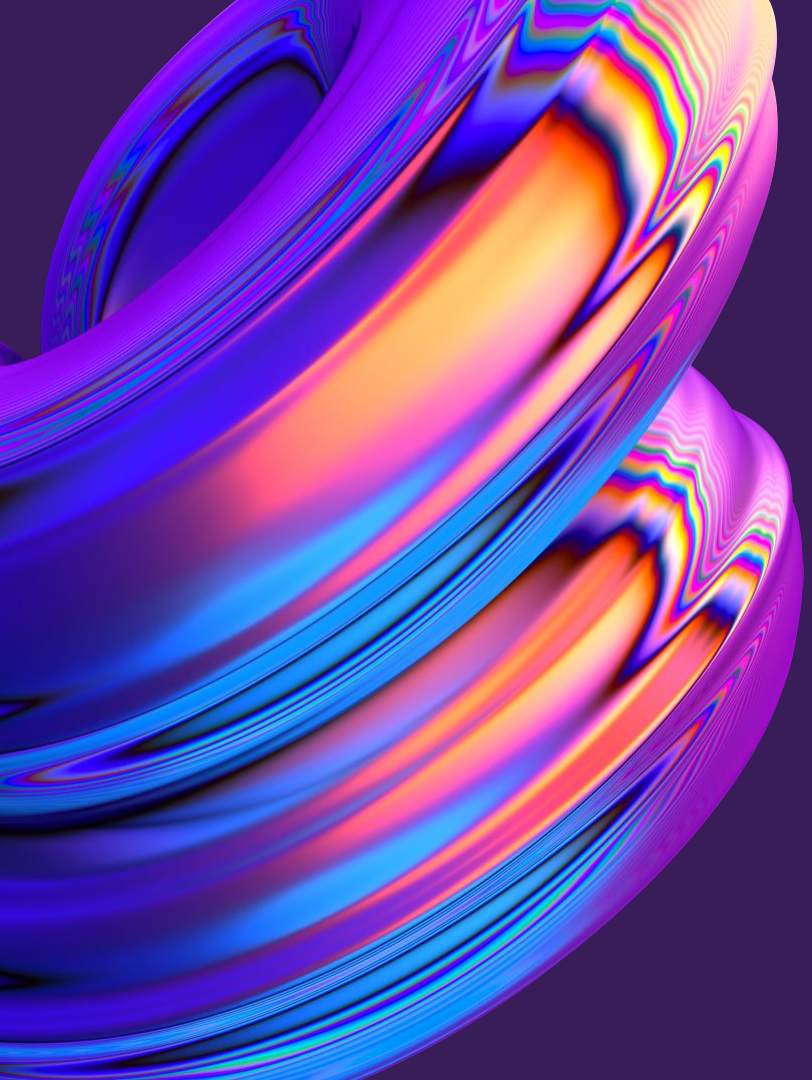


MERANTIX MOMENTUM

# ChatGPT & GPT-5: insights & Future developments

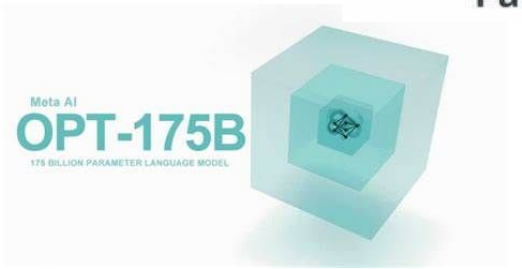
Dr. Johannes Otterbach
Simmons & Simmons Digital Day, 11.05.2023

# What are LLMs

- General purpose NLP models based on Deep Learning
- Typically billions of parameters
- Trained on trillions of words

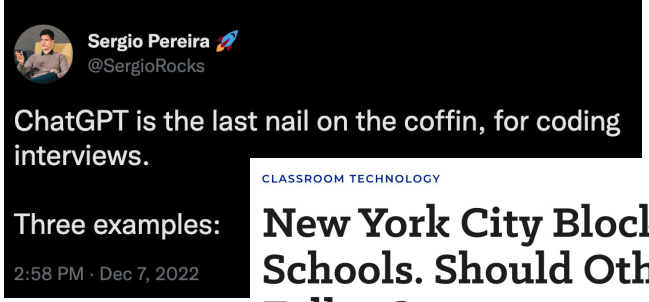



PaLM 540B




# Impact of LLMs

- LLMs are quickly becoming a platform technology
- They have real world impact already




CLASSROOM TECHNOLOGY

## New York City Blocks ChatGPT at Schools. Should Other Districts Follow?

By Alyson Klein –

## Clarification on Large Language Model Policy LLM

We (Program Chairs) have included the following statement in the Call for Papers for ICML represented by 2023:

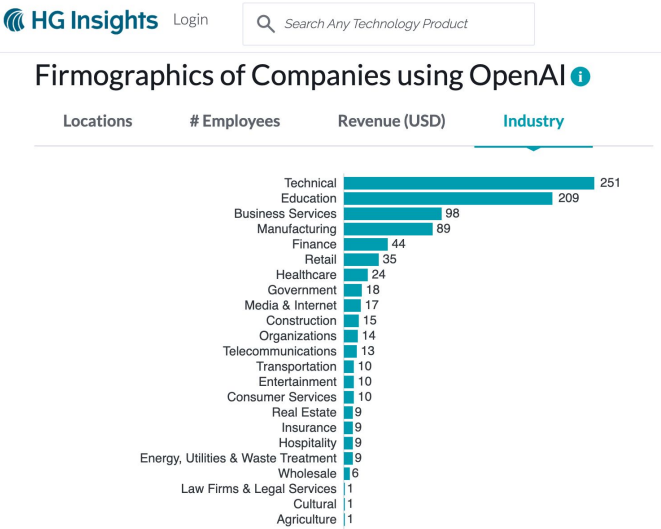
*Papers that include text generated from a large-scale language model (LLM) such as ChatGPT are prohibited unless the produced text is presented as a part of the paper's experimental analysis.*

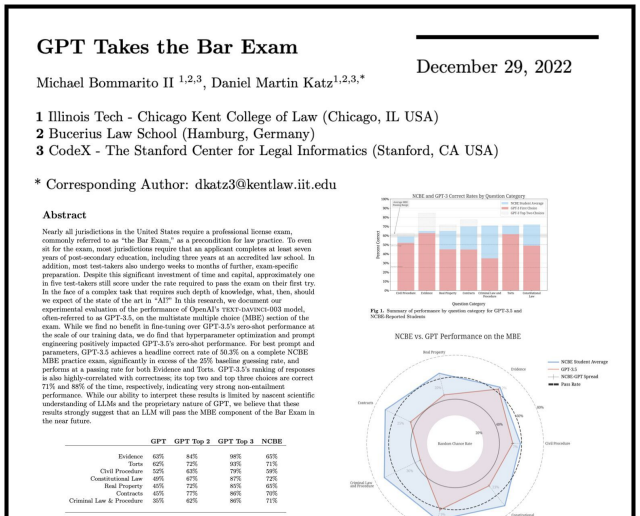
ions from potential authors and
ppreciate your feedback and
e intention behind this statemen
ICML 2023.

for ICML 2023 prohibits text
. This does not prohibit authors
or-written text.

the principle of being
t potential issues of using LLMs

Ve expect this policy may evolve
s and their impacts on scientific



GPT Takes the Bar Exam

December 29, 2022

Michael Bommarito II [1,2,3], Daniel Martin Katz[1,2,3,*]

1 Illinois Tech - Chicago Kent College of Law (Chicago, IL USA)
2 Bucerius Law School (Hamburg, Germany)
3 CodeX - The Stanford Center for Legal Informatics (Stanford, CA USA)

* Corresponding Author: dkatz3@kentlaw.iit.edu

Abstract

Nearly all jurisdictions in the United States require a professional license exam, commonly referred to as "the Bar Exam," as a precondition for law practice. To even sit for the exam, most jurisdictions require that an applicant completes at least seven years of post-secondary education, including three years at an accredited law school. In addition, most test-takers also undergo weeks to months of further, exam-specific preparation. Despite this significant investment of time and capital, approximately one in five test-takers still score under the rate required to pass the exam on their first try. In the face of a complex task that requires such depth of knowledge, what, then, should we expect of the state of the art in "AI"? In this research, we document our experimental evaluation of the performance of OpenAI's TEXT-DAVINCI-003 model, often-referred to as GPT-3.5, on the multistate multiple choice (MBE) section of the exam. While we find no benefit in fine-tuning over GPT-3.5's zero-shot performance at the scale of our training data, we do find that hyperparameter optimization and prompt engineering positively impacted GPT-3.5's zero-shot performance. For best prompt and parameters, GPT-3.5 achieves a headline correct rate of 50.3% on a complete NCBE MBE practice exam, significantly in excess of the 25% baseline guessing rate, and performs at a passing rate for both Evidence and Torts. GPT-3.5's ranking of responses is also highly-correlated with correctness; its top two and top three choices are correct 71% and 88% of the time, respectively, indicating very strong non-entailment performance. While our ability to interpret these results is limited by nascent scientific understanding of LLMs and the proprietary nature of GPT, we believe that these results strongly suggest that an LLM will pass the MBE component of the Bar Exam in the near future.

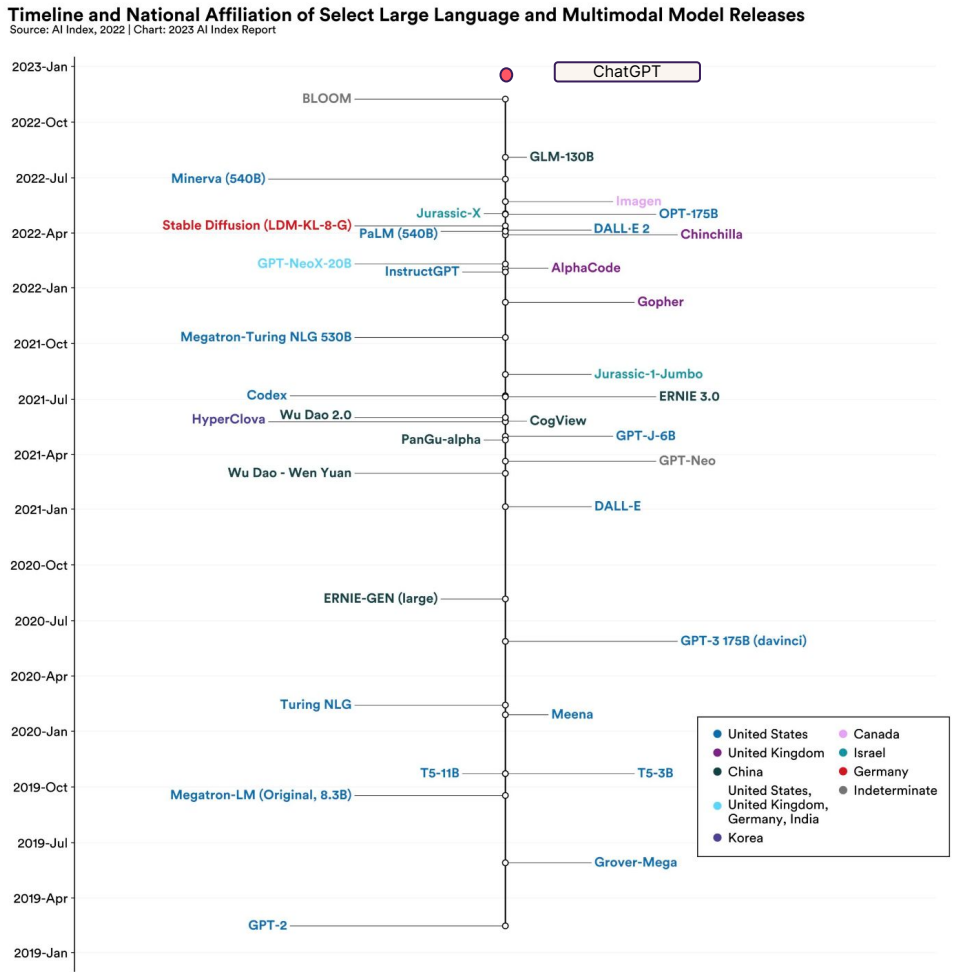
| | GPT | GPT Top 2 | GPT Top 3 | NCBE |
|---|---|---|---|---|
| Evidence | 63% | 84% | 98% | 65% |
| Torts | 62% | 72% | 93% | 71% |
| Civil Procedure | 52% | 63% | 79% | 59% |
| Constitutional Law | 49% | 67% | 87% | 72% |
| Real Property | 45% | 72% | 85% | 63% |
| Contracts | 45% | 77% | 80% | 70% |
| Criminal Law & Procedure | 35% | 62% | 86% | 71% |



MERANTIX MOMENTUM

# Timeline of LLMs / GenAI

- GenAI progress start more than 3 years ago
- Hit the public with ChatGPT in December 2022



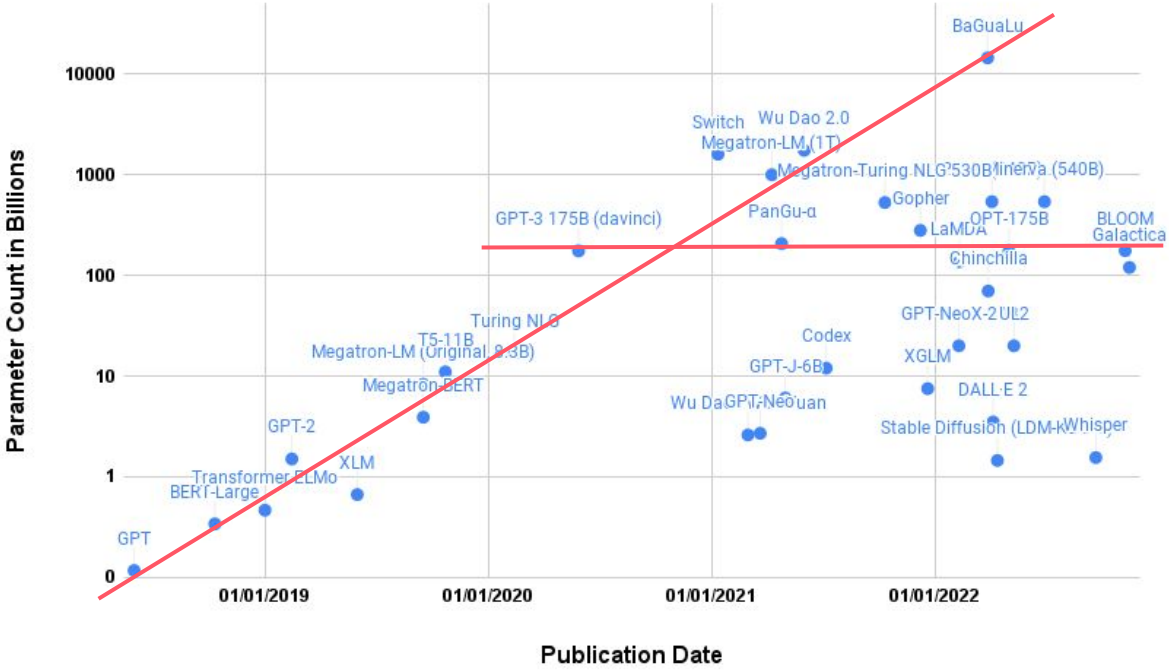

The United States is still ahead in terms of AI conference and repository citations, but those leads are slowly eroding. Still, the majority of the world's large language and multimodal models (54% in 2022) are produced by American institutions.

MERANTIX MOMENTUM

# Model sizes keep increasing

2 Trends:

- Model sizes keep increasing (we don't know the size of the latest LLMs anymore)
- ca. 18-month delay for repeat implementations



# Driver: Predictable Performance Improvements

- Performance **improvements** on all downstream NLP tasks are **predictable** and scale with the **number of parameters**

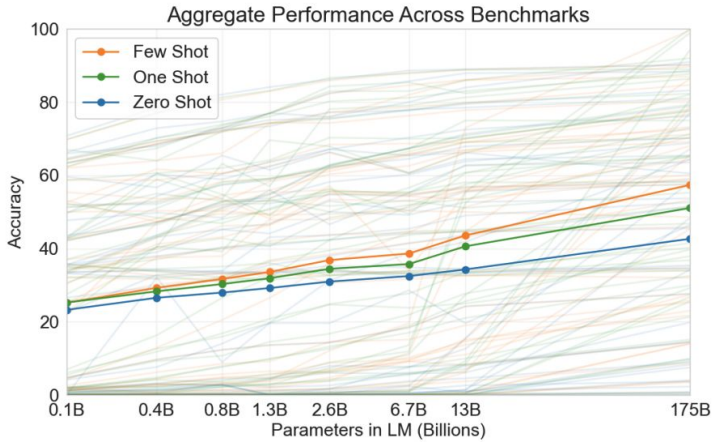
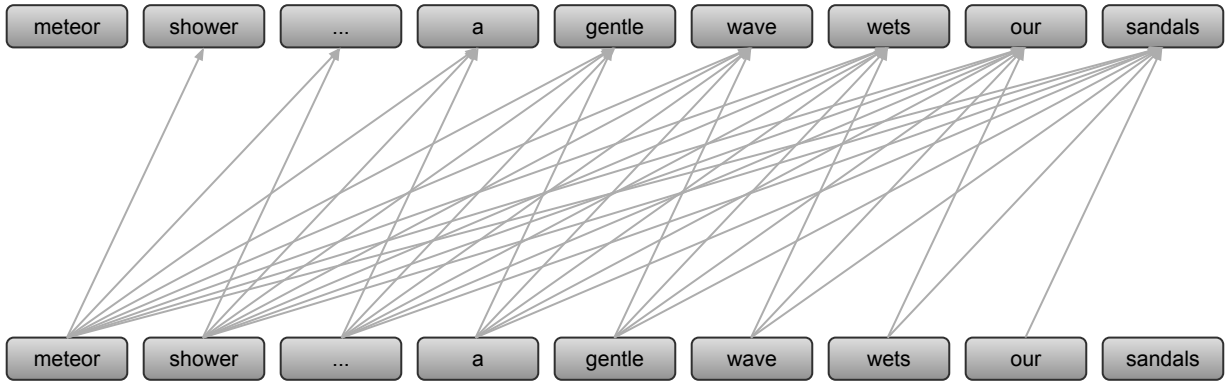



**Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks** While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

# What is a transformer in a nutshell?

- LLMs are based on autoregressive transformer models
- The objective is to predict the next word given all the previous ones

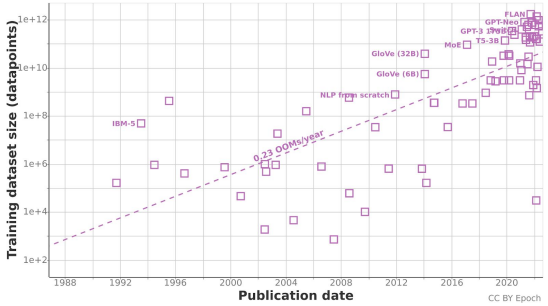





— Haiku by Michael Dylan Welch, *HSA Newsletter* XV:4, Autumn 2000

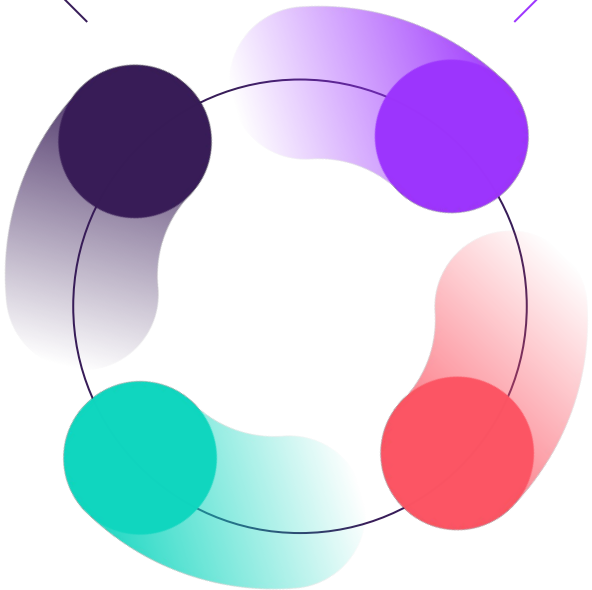
# How to build these models

## Data




## Infrastructure

- Thousands of GPUs
- Tens of thousands of CPUs
- USA declared Chips & AI as **critical infrastructure**



## Benchmarks

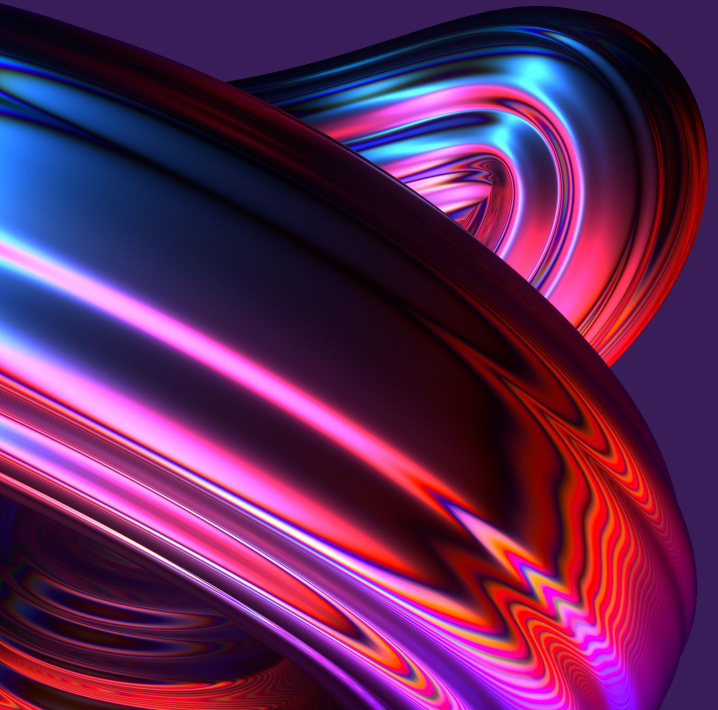
- Measuring AI progress is instrumental
- Many benchmarks are US-English centric
- Extreme bottleneck for developing non-English LLMs

## Talent

- Building SotA AI models requires a plethora of skills
- The European market is significantly lagging behind US and CHN
- We need to break the mentality of “this is just IT”

MERANTIX MOMENTUM

MERANTIX MOMENTUM



# Quo Vadis: GenAI?

# This is just another week in AI

- Each of these points is highly impactful
- The list is not complete, e.g. Google Med-PaLM, OpenAI Whisper API, ...

For reference

- The preceding week we saw LLaMA being leaked, Google USM
- The subsequent week we saw the AI moratorium, OpenAI privacy bug, “Sparks of AGI” paper

**Lior** ⚡ 
@AlphaSignalAI

This might be the most eventful week AI has ever seen:

Monday:
-Stanford Alpaca 7B

Tuesday:
-GPT4
-Anthropic releases Claude
-Google's PaLM API
-AdeptAI raises $350M
-Google adds GenAI to workspaces

Wednesday:
-Pytorch 2.0
-MidjourneyV5

Thursday:
-Microsoft 365 Copilot

6:00 PM · Mar 16, 2023 · **887K** Views

**1,031** Retweets **151** Quotes **4,658** Likes **1,321** Bookmarks

# The list of impactful results is too long to keep track …

- LLMs now squarely arrived in the productization step
- First wave of Foundation Models

- Second wave of Foundation Models
- Computer Vision & Speech Recognition

MERANTIX MOMENTUM

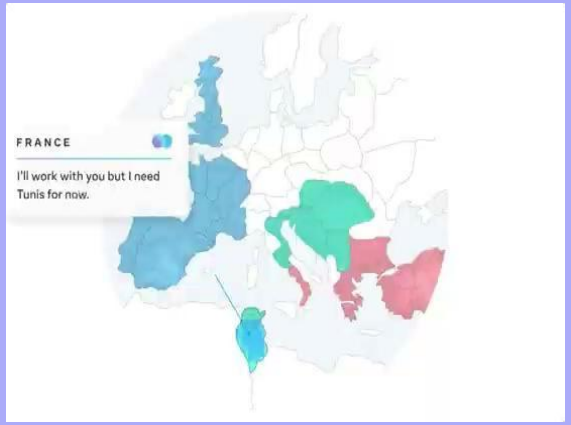
# LLMs are turned into complex applications

Simulacra of Human Behavior

- Multiple AI agents interact in a SIM environment based on natural language
- Emergent human-like behavior appears, e.g. waking up, cooking, going to work and on dates, cooperation, etc.


Diplomacy

- strategic board game set in early 20th century Europe where players negotiate and maneuver to gain control of countries.
- Players rely on diplomacy, alliances, and cunning to achieve victory
- Communication is done purely bilateral and in written text
- Humans were unable to detect AI agents

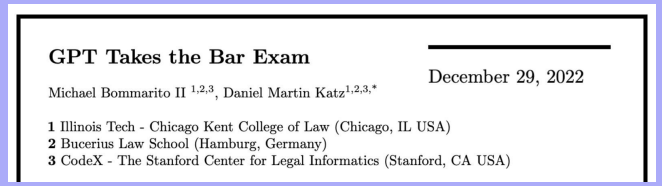


# LLMs are turned into complex applications

More features being worked on right now

- Chain-of-Thought (CoT) reasoning or planning
- “Watermarking” of LLM output
- Detoxification & bias reduction
- Grounding & Factual correctness
- Incorporation of external knowledge
- ...

Imagine the combination of



**GPT Takes the Bar Exam**

December 29, 2022

Michael Bommarito II [1,2,3], Daniel Martin Katz[1,2,3,*]

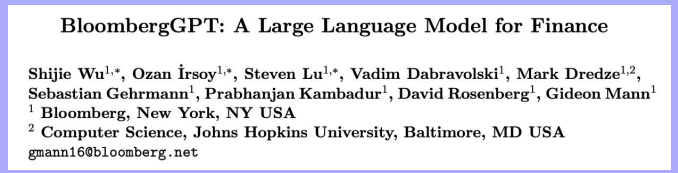
**1** Illinois Tech - Chicago Kent College of Law (Chicago, IL USA)
**2** Bucerius Law School (Hamburg, Germany)
**3** CodeX - The Stanford Center for Legal Informatics (Stanford, CA USA)

simmons + simmons

# Outlook

- The field is moving extremely fast
- Single modalities are now being turned into products
- Future models will combine the modalities. GPT-4 is just a demo of this

- Differences can be made in dataset compostion
- E.g. BloombergGPT

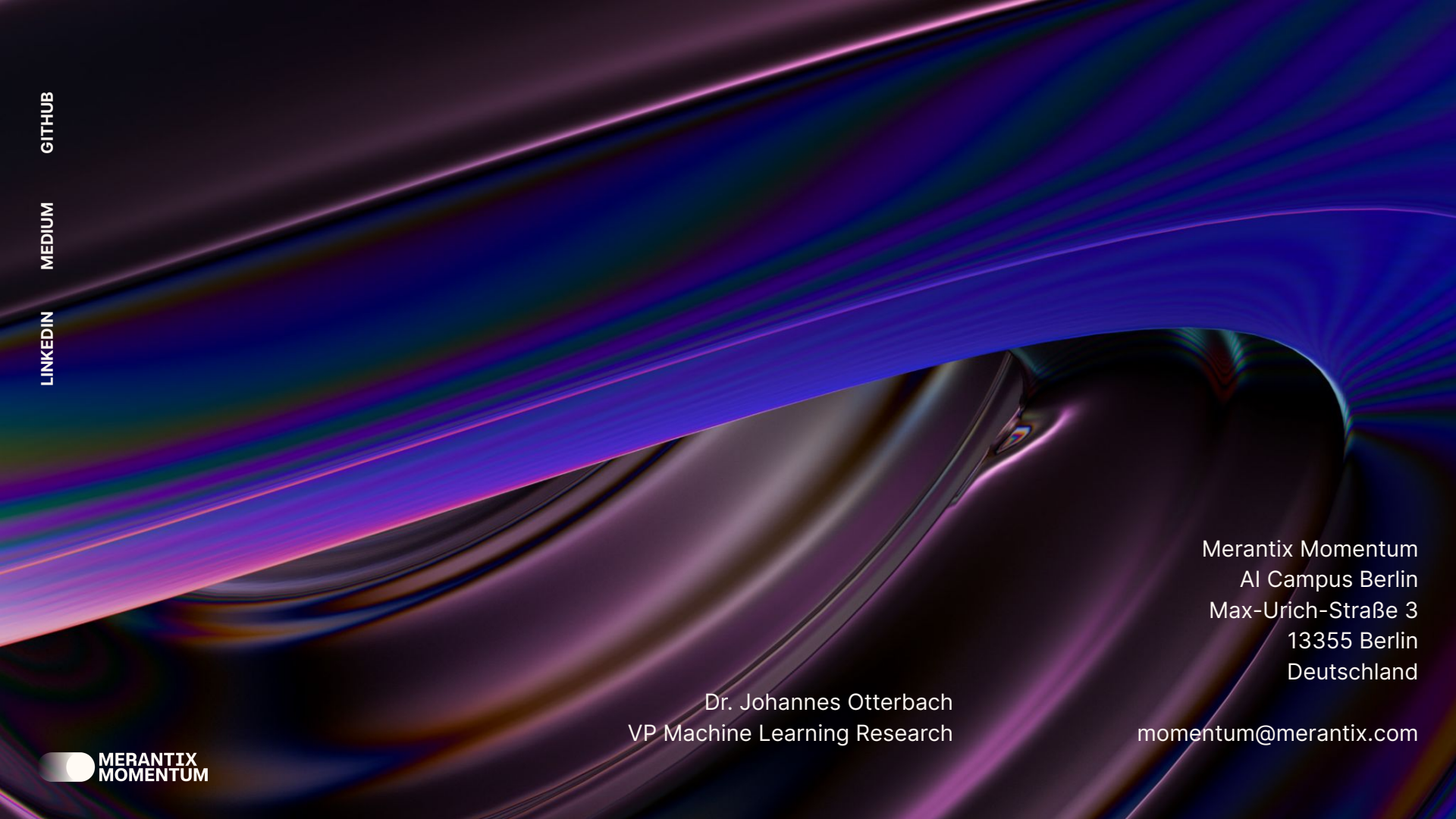


BloombergGPT: A Large Language Model for Finance

Shijie Wu[1,*], Ozan İrsoy[1,*], Steven Lu[1,*], Vadim Dabravolski[1], Mark Dredze[1,2], Sebastian Gehrmann[1], Prabhanjan Kambadur[1], David Rosenberg[1], Gideon Mann[1]
[1] Bloomberg, New York, NY USA
[2] Computer Science, Johns Hopkins University, Baltimore, MD USA
gmann16@bloomberg.net

# Take home message

AI will not replace you, the person using it will.




Merantix Momentum
AI Campus Berlin
Max-Urich-Straße 3
13355 Berlin
Deutschland

Dr. Johannes Otterbach
VP Machine Learning Research

momentum@merantix.com

MERANTIX
MOMENTUM