





**Baker  
McKenzie.**

# Legal Tech and AI Impact on the Legal Ecosystem

 **IFSWF**  
International Forum of Sovereign Wealth Funds





Bucerius Law School, Hamburg, Germany





**CODEX**  
The Stanford Center for Legal Informatics



SLS

PROJECTS

# Legal Complexity Science

Social, economic and political complexity have manifested in increasing levels of legal complexity. While legal systems have to find ways to handle this increase, technology and data science can help further the understanding of their performance and provide them with much-needed tools.

The legal data science project follows a quantitative approach in the analysis of law. Its activities fall into three categories: the creation of data sets, their analysis and the communication and application of their results. The theoretical foundation is provided by complexity science, as the legal system is understood as a complex adaptive system (CAS). As such, its properties can be measured with methods from computer science, physics and mathematics, such as network science, text analytics and data mining. Developing these methods and applying them to various datasets is at the core of the project's activities. However, many of these datasets need to be built, enriched, documented and open-sourced before any analysis can be conducted. As a last step, the resulting research needs to be communicated to other quantitative and normative legal scholars and turned into actionable steps for policymakers and practitioners.

The goal of this project is an enhanced and robust understanding of the legal system and its dynamics. This is achieved through publications in interdisciplinary, peer-reviewed general scientific journals and presentations at both research and practitioner conferences.

RELATED ORGANIZATIONS  
CodeX



**Dirk Hartung**  
CodeX Non-Residential Fellow

[VIEW PROFILE](#)



**Daniel Katz**  
CodeX Affiliated Faculty  
Professor of Law, Illinois Tech – Chicago Kent College  
Director, The Law Lab, Illinois Tech

[VIEW PROFILE](#)



**Michael Bommarito**  
CodeX Non-Residential Fellow

[VIEW PROFILE](#)

*The descriptions of current and past projects of CodeX non-residential fellows are provided to illustrate the kind of work our non-residential fellows are carrying out. These projects are listed here for informational purposes only and are not endorsed by CodeX, Stanford Law School, or Stanford University.*





**Christian Veith**



**Digitalization and the Market for Legal Services**



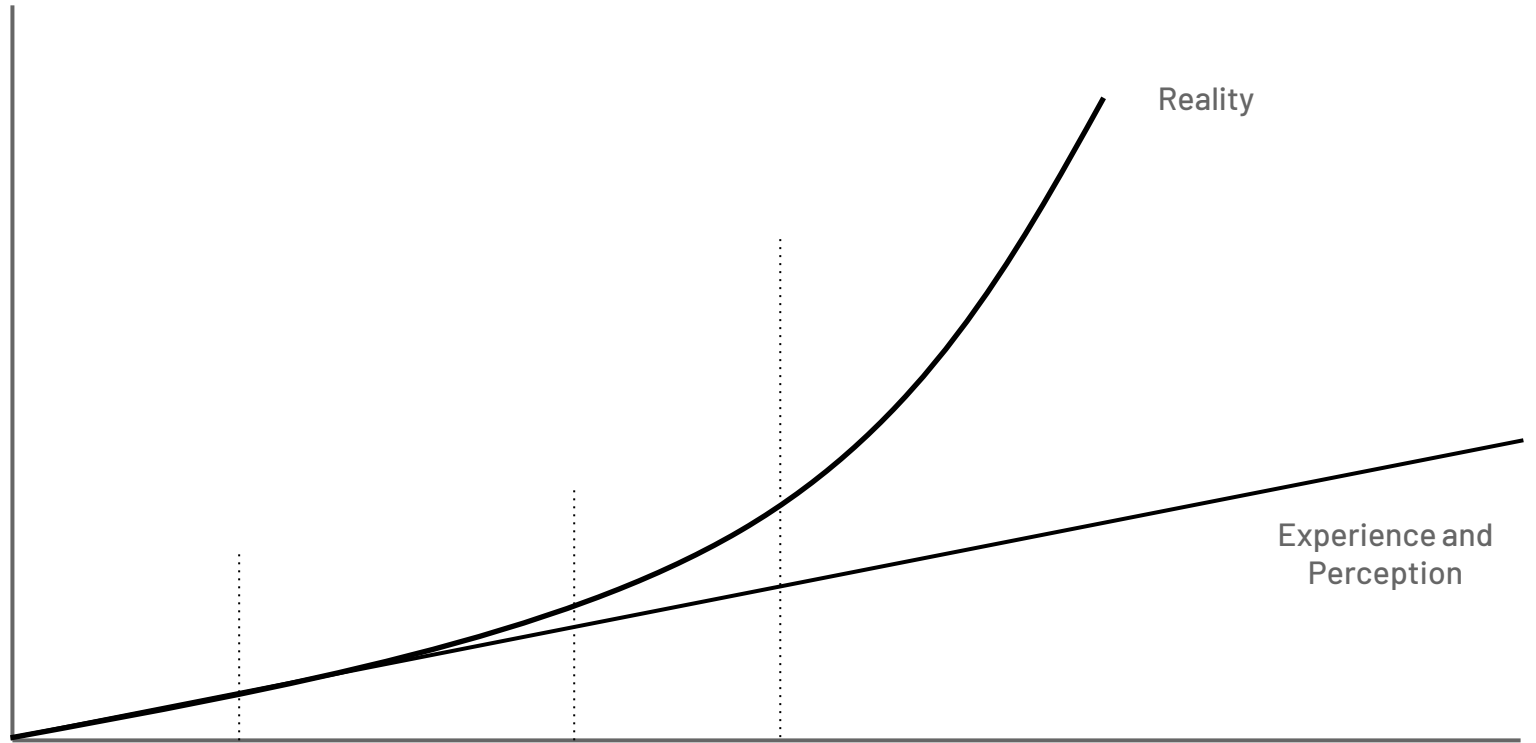
**Dirk Hartung**








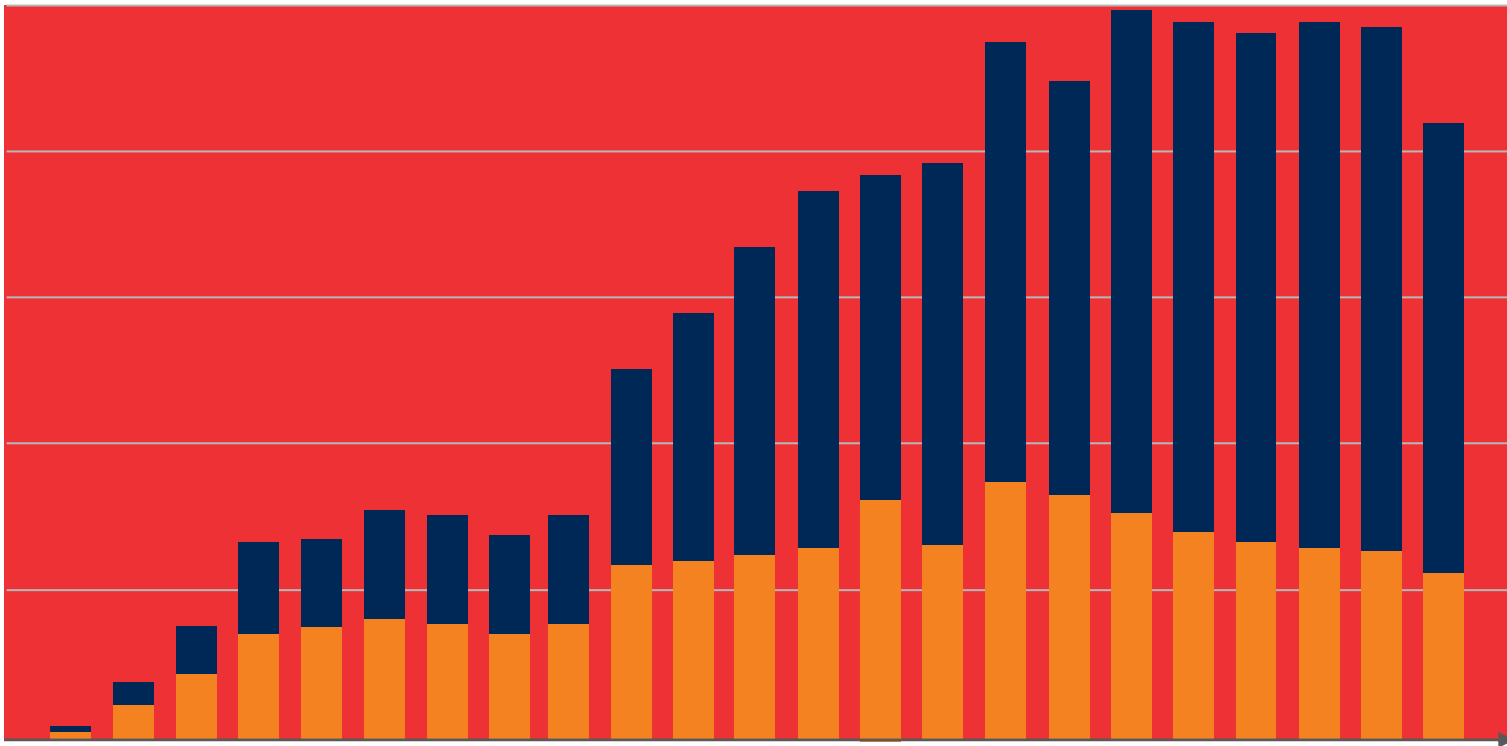
**Manage legal risk  
and handle legal  
complexity.**



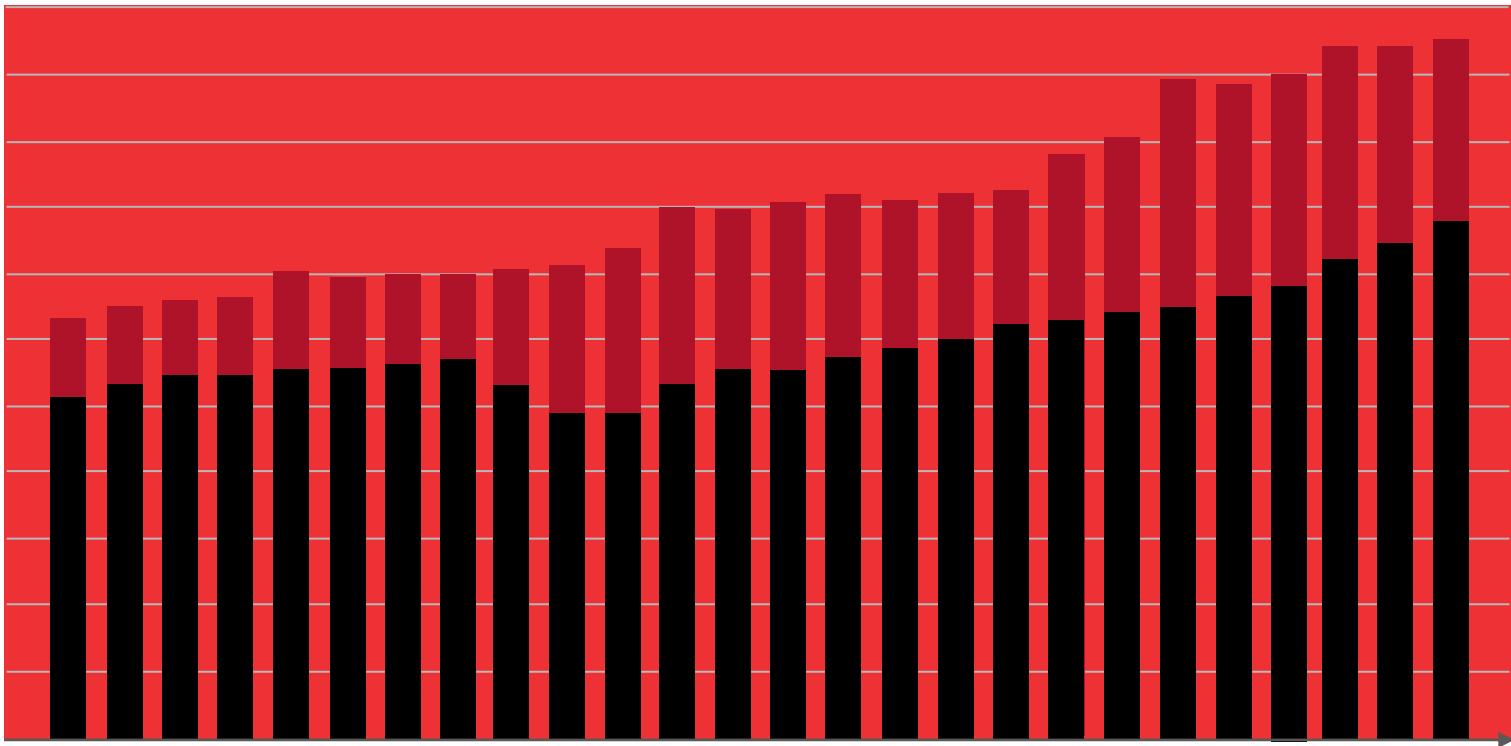




# WHAT IS Legal Complexity?







nature

SCIENTIFIC  
REPORTS



(a) United States (1994)



(b) Germany (1994)



(c) United States (2018)

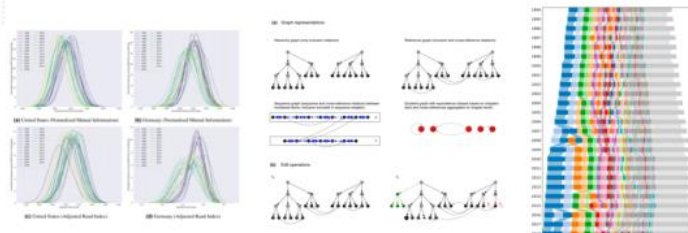


(d) Germany (2018)

# Complex societies and the growth of the law

Daniel Martin Katz<sup>1,2,5</sup>, Corinna Coupette<sup>3</sup>, Janis Beckedorf<sup>4</sup> & Dirk Hartung<sup>2,5</sup>

While many informal factors influence how people interact, modern societies rely upon law as a primary mechanism to formally control human behaviour. How legal rules impact societal development depends on the interplay between two types of actors: the people who create the rules and the people to which the rules potentially apply. We hypothesise that an increasingly diverse and interconnected society might create increasingly diverse and interconnected rules, and assert that legal networks provide a useful lens through which to observe the interaction between law and society. To evaluate these propositions, we present a novel and generalizable model of statutory materials as multidimensional, time-evolving document networks. Applying this model to the federal legislation of the United States and Germany, we find impressive expansion in the size and complexity of laws over the past two and a half decades. We investigate the sources of this development using methods from network science and natural language processing. To allow for cross-country comparisons over time, based on the explicit cross-references between legal rules, we algorithmically reorganise the legislative materials of the United States and Germany into cluster families that reflect legal topics. This reorganisation reveals that the main driver behind the growth of the law in both jurisdictions is the expansion of the welfare state, backed by an expansion of the tax state. Hence, our findings highlight the power of document network analysis for understanding the evolution of law and its relationship with society.





## Buch 3. Sachenrecht

Abschnitt 1.<sup>1)</sup> Besitz

**§ 854. Erwerb des Besitzes.** (1) Der Besitz einer Sache wird durch die Erlangung der tatsächlichen Gewalt über die Sache erworben.

(2) Die Einigung des bisherigen Besitzers und des Erwerbers genügt zum Erwerb, wenn der Erwerber in der Lage ist, die Gewalt über die Sache ausüben.

**§ 855. Besitzdiener.** Übt jemand die tatsächliche Gewalt über eine Sache für einen anderen in dessen Haushalt oder Erwerbsgeschäft oder in einem ähnlichen Verhältnis aus, vermöge dessen er den sich auf die Sache beziehenden Weisungen des anderen Folge zu leisten hat, so ist nur der andere Besitzer.

**§ 856. Beendigung des Besitzes.** (1) Der Besitz wird dadurch beendet, dass der Besitzer die tatsächliche Gewalt über die Sache aufgibt oder in anderer Weise verliert.

(2) Durch eine ihrer Natur nach vorübergehende Verhinderung in der Ausübung der Gewalt wird der Besitz nicht beendet.

**§ 857. Vererblichkeit.** Der Besitz geht auf den Erben über.

**§ 858. Verbotene Eigenmacht.** (1) Wer dem Besitzer ohne dessen Willen den Besitz entzieht oder ihn im Besitz stört, handelt, sofern nicht das Gesetz die Entziehung oder die Störung gestattet, widerrechtlich (verbotene Eigenmacht).

(2) <sup>1</sup>Der durch verbotene Eigenmacht erlangte Besitz ist fehlerhaft. <sup>2</sup>Die Fehlerhaftigkeit muss der Nachfolger im Besitz gegen sich gelten lassen, wenn er Erbe des Besitzers ist oder die Fehlerhaftigkeit des Besitzes seines Vorgängers bei dem Erwerb kennt.

**§ 859. Selbsthilfe des Besitzers.** (1) Der Besitzer darf sich verbotener Eigenmacht mit Gewalt erwehren.

(2) Wird eine bewegliche Sache dem Besitzer mittels verbotener Eigenmacht weggenommen, so darf er sie dem auf frischer Tat betroffenen oder verfolgten Täter mit Gewalt wieder abnehmen.

(3) Wird dem Besitzer eines Grundstücks der Besitz durch verbotene Eigenmacht entzogen, so darf er sofort nach der Entziehung sich des Besitzes durch Entsetzung des Täters wieder bemächtigen.

(4) Die gleichen Rechte stehen dem Besitzer gegen denjenigen zu, welcher nach § 858 Abs. 2 die Fehlerhaftigkeit des Besitzes gegen sich gelten lassen muss.

<sup>1)</sup> Wegen des für das Gebiet der ehem. DDR geltenden Übergangsrechts zu §§ 854 bis 872 beachte Art. 233 § 1 EGBGB; Nr. 21.

BGB

Buch 3

Abschnitt 1

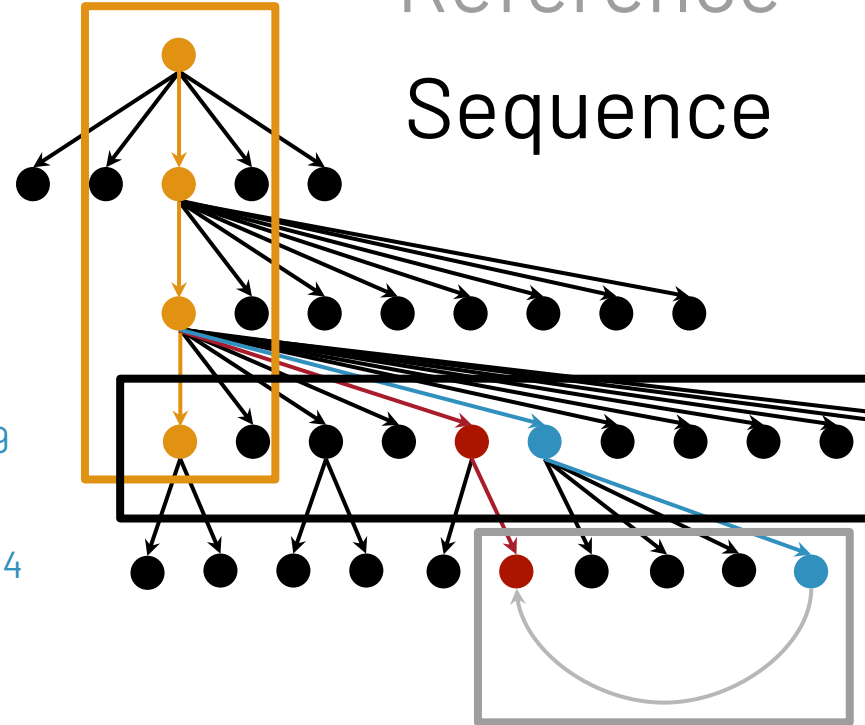
§ 858 § 859

Abs. 2 Abs. 4

## Hierarchy

## Reference

## Sequence



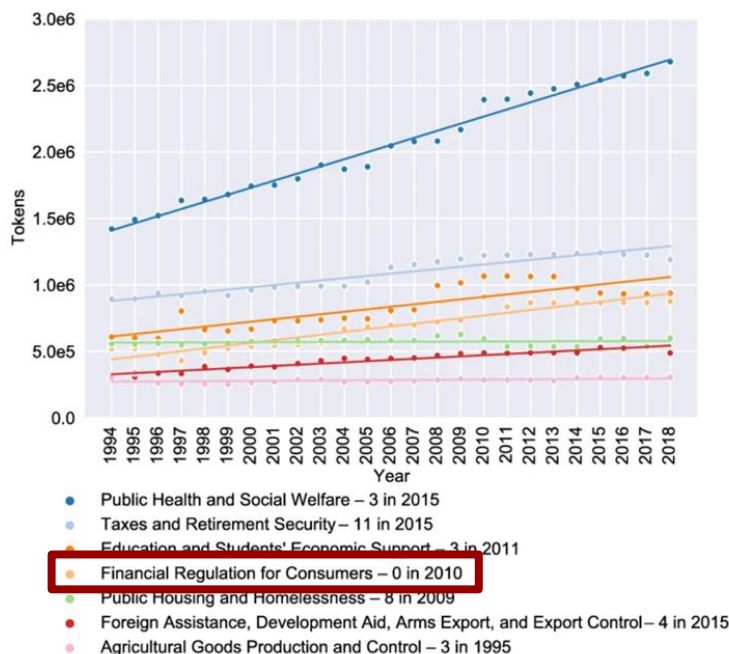
# Summary Statistics (2018)

	United States			Germany		
	1994	2018	$\Delta$	1994	2018	$\Delta$
Tokens	14.0 M	21.2 M	51 %	4.5 M	7.4 M	64 %
Structures	452.4 K	828.1 K	83 %	120.6 K	161.4 K	34 %
References	58.0 K	88.6 K	53 %	76.9 K	139.1 K	81 %

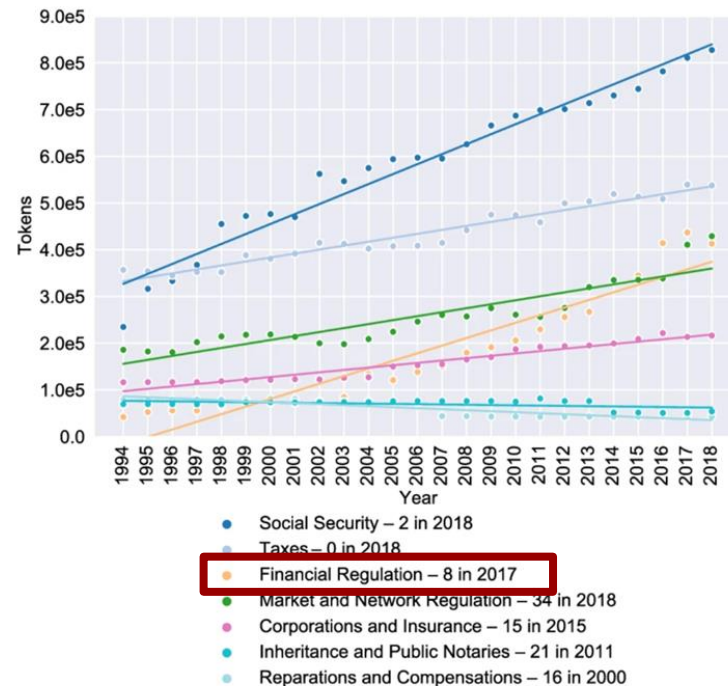
**Table 1.** Federal legislation in the United States and Germany: descriptive statistics (1994 and 2018).

# Figure 6

From: [Complex societies and the growth of the law](#)



(a) United States



(b) Germany

Federal legislation in the United States and Germany: growth statistics by cluster family for selected cluster families (1994–2018). The legends are sorted by the y-values of the regression lines in 2018. The colours are comparable across countries, i.e., *same colour*  $\Leftrightarrow$  *(roughly) same topic*.

# Measuring Law Over Time: A Network Analytical Framework with an Application to Statutes and Regulations in the United States and Germany

Corinna Coupette<sup>1†</sup>, Janis Beckedorf<sup>2†</sup>, Dirk Hartung<sup>3,4\*</sup>, Michael Bommarito<sup>5</sup> and  
Daniel Martin Katz<sup>3,4,5</sup>

<sup>1</sup>Max Planck Institute for Informatics, Saarbrücken, Germany, <sup>2</sup>Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany,  
<sup>3</sup>Center for Legal Technology and Data Science, Bucerius Law School, Hamburg, Germany, <sup>4</sup>CodeX – the Stanford Center for  
Legal Informatics, Stanford Law School, Stanford, CA, United States, <sup>5</sup>Illinois Tech – Chicago Kent College of Law, Chicago, IL,  
United States

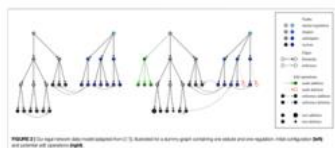
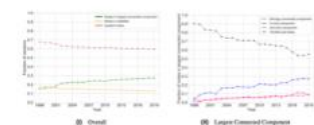
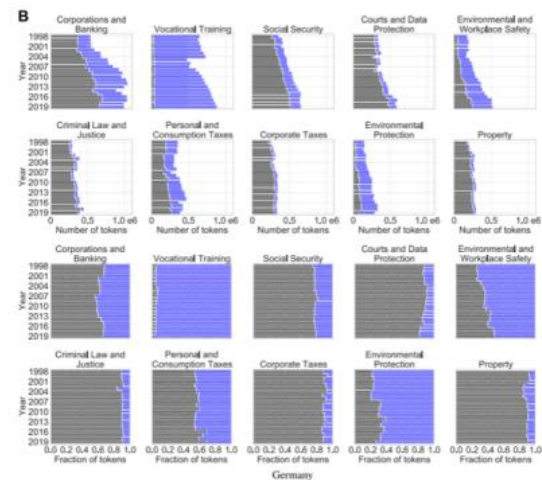
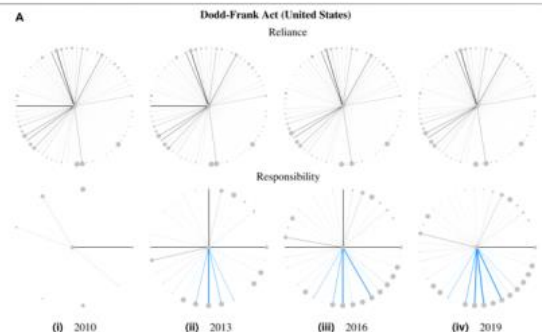
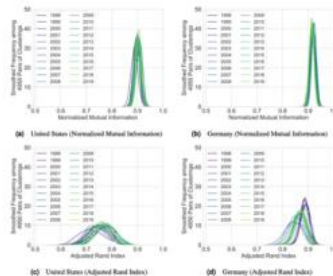


FIGURE 2 | The legal network also with hierarchical (H) and (L) clustered to a binary graph sampling via static prior (negative) value (negative) and  
and control and question (grey).



# Growth in size, organization and inter- connectedness

	Statutes			Regulations		
	1998	2019	$\Delta$	1998	2019	$\Delta$
<b>Tokens</b>	15.2 M	21.4 M	41	43.9 M	84.3 M	92
<b>Structures</b>	516.2 K	838.8 K	63	1.4 M	2.7 M	91
<b>References</b>	80.1 K	112.1 K	40	134.6 K	348.4 K	159

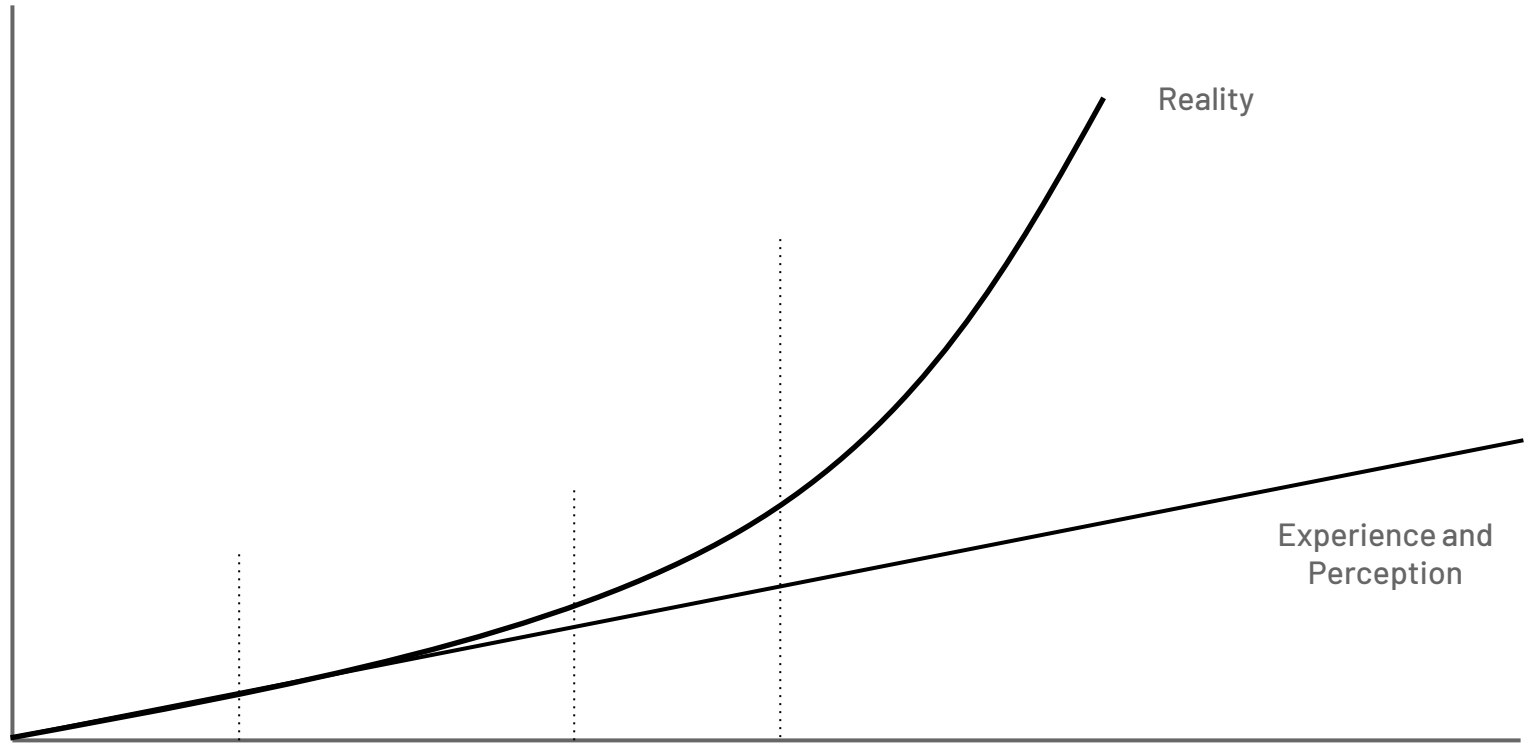
(a) United States

	Statutes			Regulations		
	1998	2019	$\Delta$	1998	2019	$\Delta$
<b>Tokens</b>	5.0 M	7.7 M	54	3.9 M	5.4 M	39
<b>Structures</b>	130.6 K	166.0 K	27	87.9 K	113.7 K	29
<b>References</b>	86.4 K	144.6 K	67	33.5 K	47.1 K	41

(b) Germany

**Table (2).** (Rounded) size of the national legal systems of the United States (top) and Germany (bottom) as measured by the tokens, structural elements, and references in their codified law in 1998 and 2019, including the total percentage change between these years ( $\Delta$ ).









# **2010: A decade of legal technology and innovation**

The Atlantic Popular Latest Sections Magazine



## Rise of the Robolawyers

How legal representation could come to resemble TurboTax

JASON KESSELER | APRIL 2017 ISSUE | TECHNOLOGY

COLOR: JESSICA BROWN BUSINESS 12:00 PM 12:00 AM

## TECH WILL FORCE LAWYERS TO DO MORE FOR THOSE BILLABLE HOURS



PLAY #SILICON VALLEY #STARTUPS #TECH

## How artificial intelligence is transforming the legal profession

POSTED APRIL 11, 2016 10:45 AM EDT BY JULIE SCHWABE

Over decades, an endorsement with a background in advertising and media, some on the list of the 100 most powerful people in the world, who worked as a television attorney, passed over for the job of representing the legal profession.




It's not just technology, it's also the fact that technology can do a job that a human can't. The future is now.

## A.I. Is Doing Legal Work. But It Won't Replace Lawyers, Yet.

By STEVE LOHR | MARCH 19, 2017

The New York Times

The Clickbait Version



Sorry, a Robot Is Not About to Replace Your Lawyer

The legal profession relies more and more on automation. But fears that it will be automated out of existence are overblown, researchers say. For now.

NYTIMES.COM

TECH

## Why Hire a Lawyer When a Robot Will Do?

11 SEP 22, 2016 3:00 AM EDT

By Elaine Ou

BloombergView



## FINANCIAL TIMES

Law Add to myFT

### Artificial intelligence closes in on the work of junior lawyers

'Lawtech' sifts and summarises data with speed and precision to replace routine tasks



FUTURE OF WORK

CNBC

## Lawyers could be the next profession to be replaced by computers

Der Spiegel | @\_Dorffinger | Friday, 17 Feb 2017 1:55 PM ET



April 11, 2016 7:23 am

## Technology: Breaking the law

Michael Spector

FT Financial Times

The legal profession hasn't changed in decades. Here's the case for the Uberisation of it on.ft.com/1S2OHYP

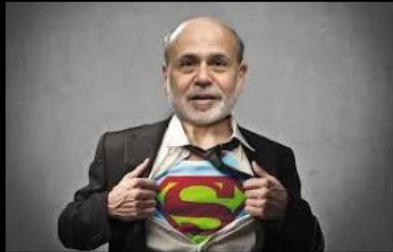


FT FINANCIAL TIMES

# 2008 - FINANCIAL CRISIS



**LIVE SHOT OF BEN  
BERNANKE'S DESK  
9/15/08**





## RAPIDLY EXPANDING LEGAL ECOSYSTEM, CIRCA 2020

### Smart Contracts



### Communication & Collaboration



### Document Analysis



### Workflow Tools



### eDiscovery



### Transaction Management



### Business Intelligence



### Legal Research



### Time & Billing



### Document Management



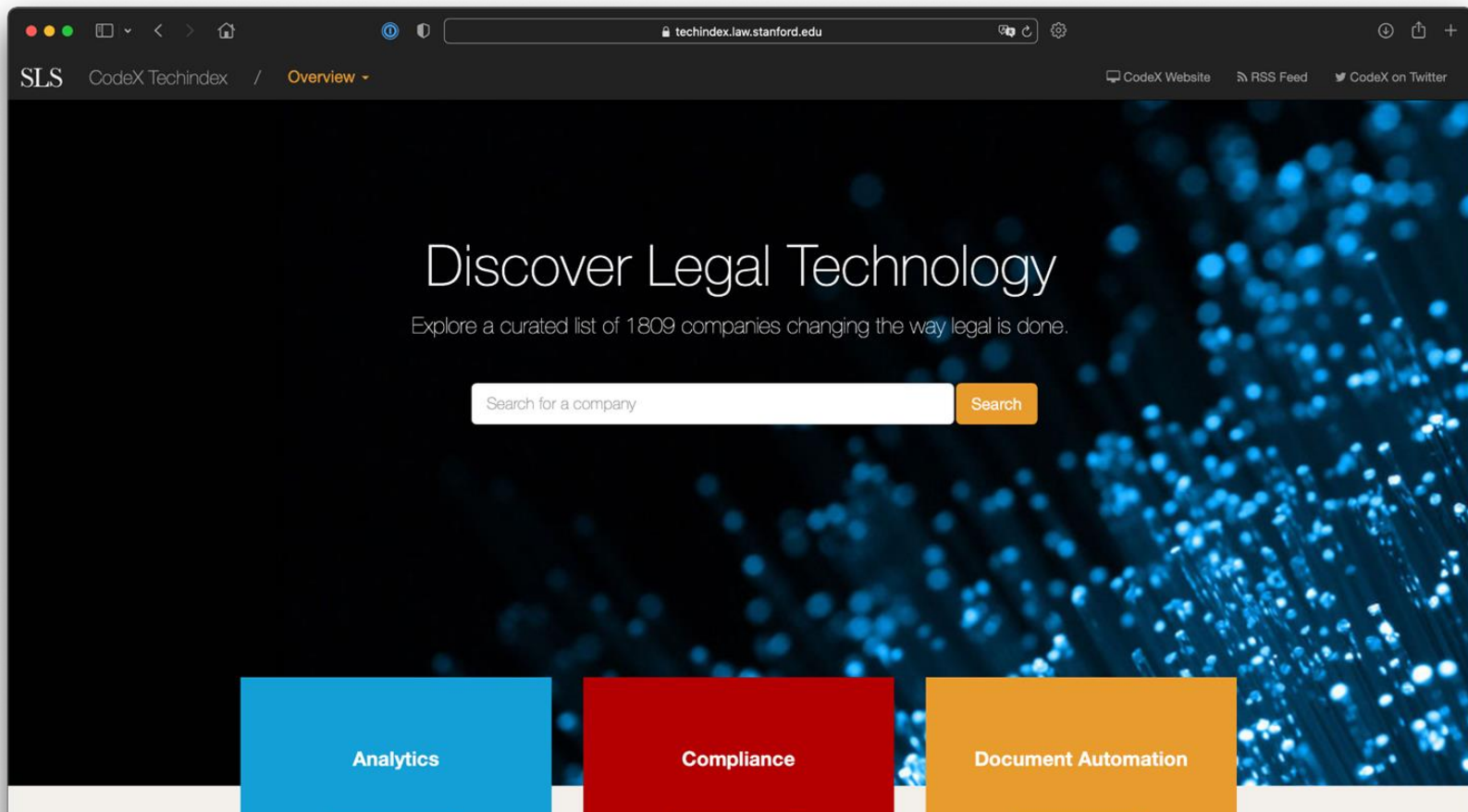
### Case Management



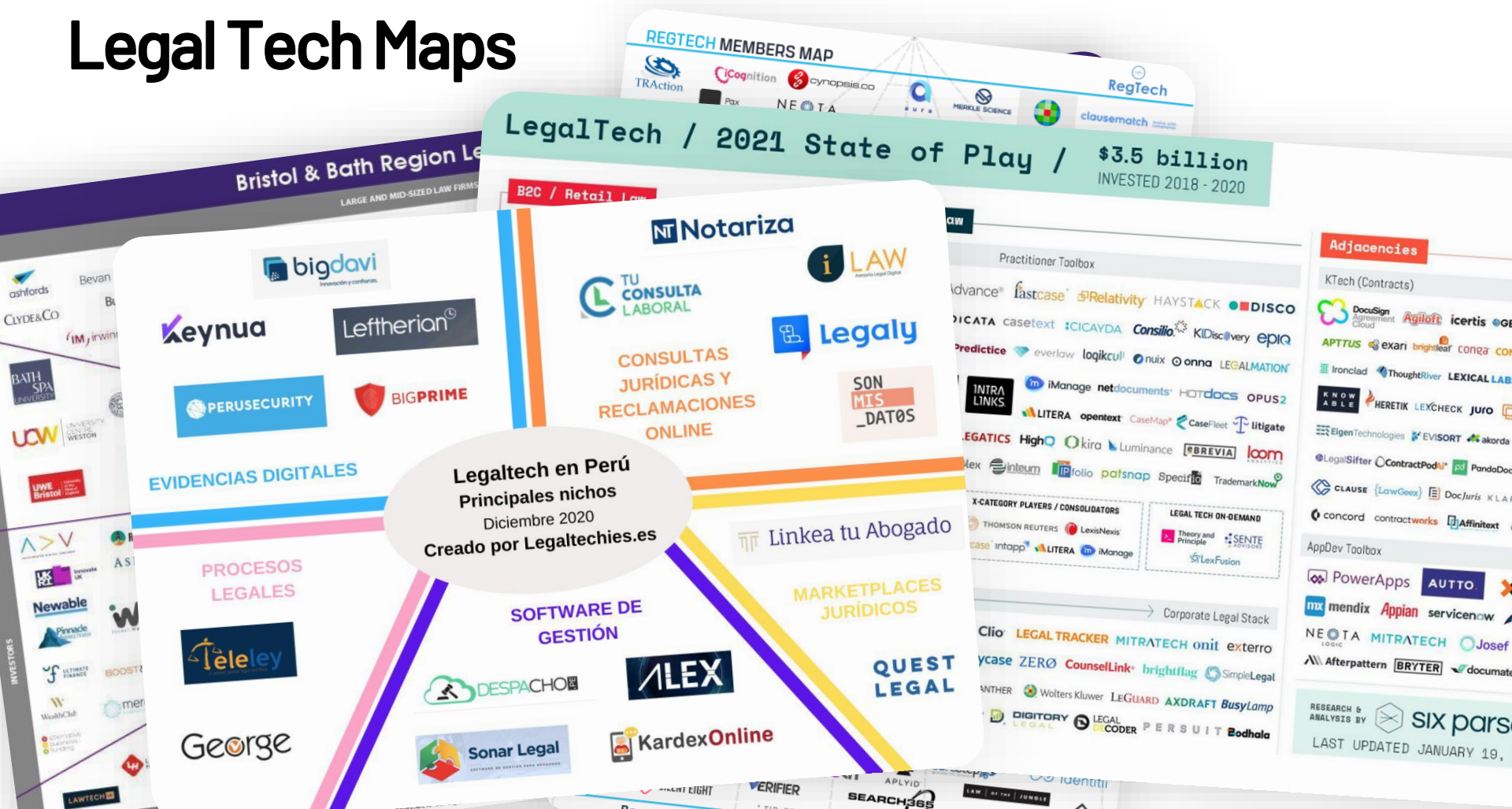
### Document Automation

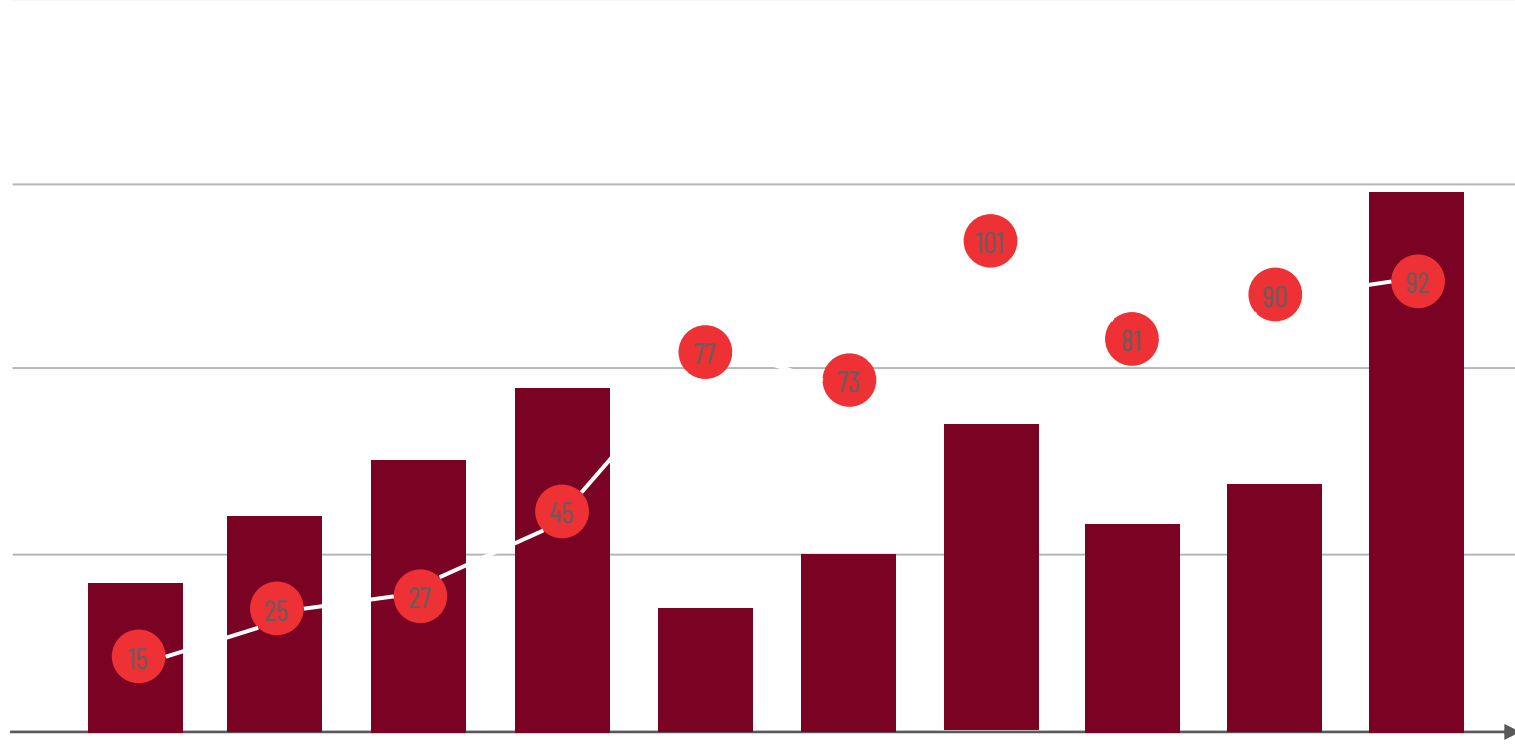


# CodeX TechIndex



# Legal Tech Maps







# Investments(U.S.)

LegalZoom  
\$ 811 M

Ironclad  
\$ 334 M

DISCO  
\$ 233 M

Checkr  
\$ 559 M

Everlaw  
\$ 298 M

Notarize  
\$ 213 M

Clio  
\$ 386 M

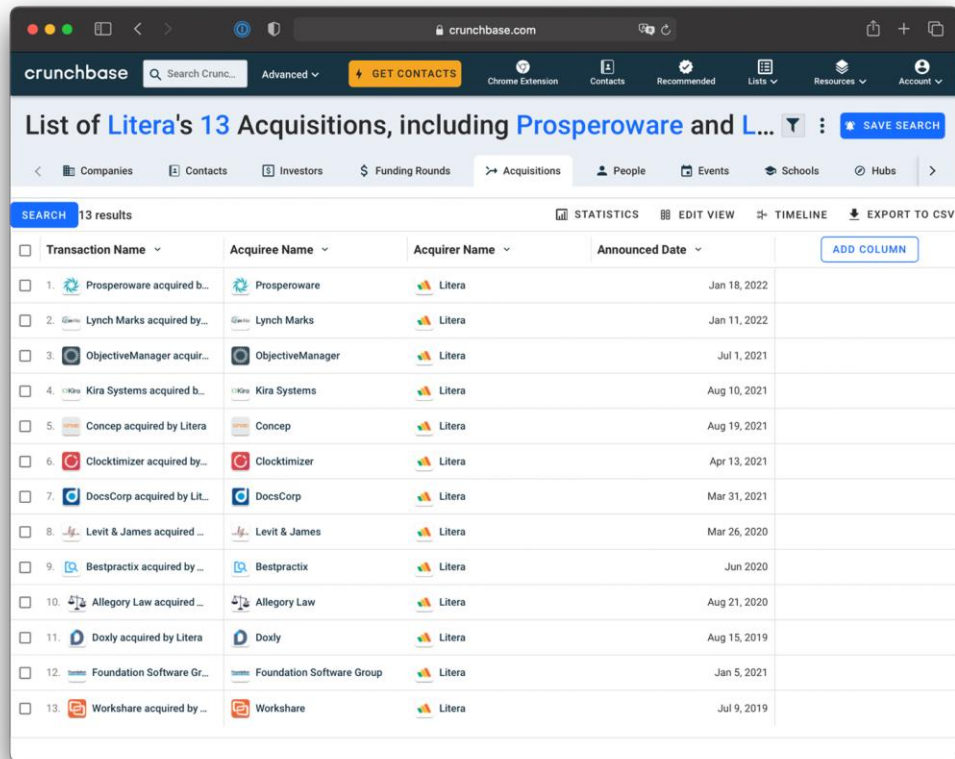
Dye&Durham  
\$ 613 M

RocketLaw  
yer  
\$ 291 M

Ontra  
\$ 240 M



# Rolling up the Market: Litera



crunchbase

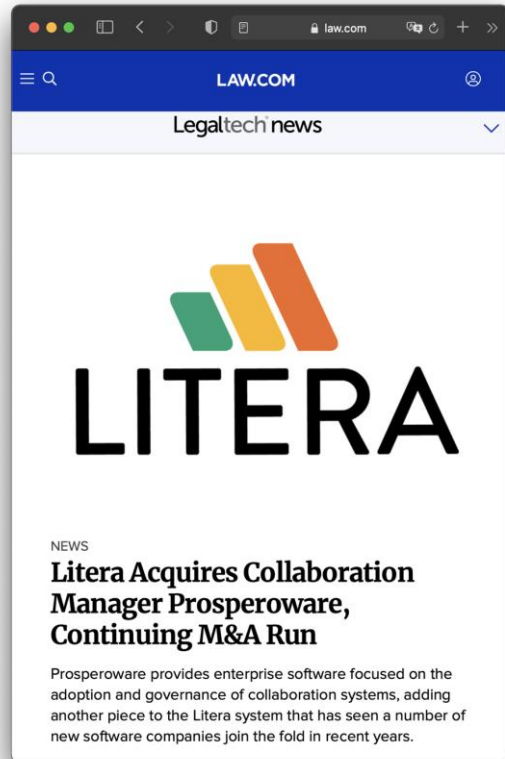
Search Crunchbase... Advanced GET CONTACTS Chrome Extension Contacts Recommended Lists Resources Account

### List of Litera's 13 Acquisitions, including Prosperoware and L...

Companies Contacts Investors Funding Rounds Acquisitions People Events Schools Hubs

SEARCH 13 results STATISTICS EDIT VIEW TIMELINE EXPORT TO CSV

Transaction Name	Acquiree Name	Acquirer Name	Announced Date
1. Prosperoware acquired by...	Prosperoware	Litera	Jan 18, 2022
2. Lynch Marks acquired by...	Lynch Marks	Litera	Jan 11, 2022
3. ObjectiveManager acquired...	ObjectiveManager	Litera	Jul 1, 2021
4. Kira Systems acquired by...	Kira Systems	Litera	Aug 10, 2021
5. Concep acquired by Litera	Concep	Litera	Aug 19, 2021
6. Clocktimizer acquired by...	Clocktimizer	Litera	Apr 13, 2021
7. DocsCorp acquired by LIT...	DocsCorp	Litera	Mar 31, 2021
8. Levit & James acquired ...	Levit & James	Litera	Mar 26, 2020
9. Bestpractix acquired by ...	Bestpractix	Litera	Jun 2020
10. Allegory Law acquired ...	Allegory Law	Litera	Aug 21, 2020
11. Doxly acquired by Litera	Doxly	Litera	Aug 15, 2019
12. Foundation Software Gr...	Foundation Software Group	Litera	Jan 5, 2021
13. Workshare acquired by...	Workshare	Litera	Jul 9, 2019



# Rolling up the Market: Elevate

**Bloomberg Law** WELCOME LOGIN

Business & Practice

**Law Company Elevate Raises \$25M, Aiming for 2021 Public Listing**

BY SAM SKOLNIK AND ROY STROM

June 17, 2019, 4:00 PM

- Elevate is a leading provider of services to law firms, legal departments
- Company sets goal of U.K. public listing by 2021, expects revenue to reach nearly \$200 million by 2023

Elevate Services has received \$25 million in funding from a private equity firm as the Los Angeles-based legal services business aims internally for a public stock market listing in 2021.

Elevate has also predicted its revenue will climb to \$76 million in 2019—and to more than twice that amount by 2023, according to a presentation obtained by Bloomberg Law.

Elevate's growth and drive toward a public listing shows how fast-growing the market for nontraditional legal service models has become. Elevate, which bills itself as a "law company," provides consulting, technology and other services to law firms and law departments.

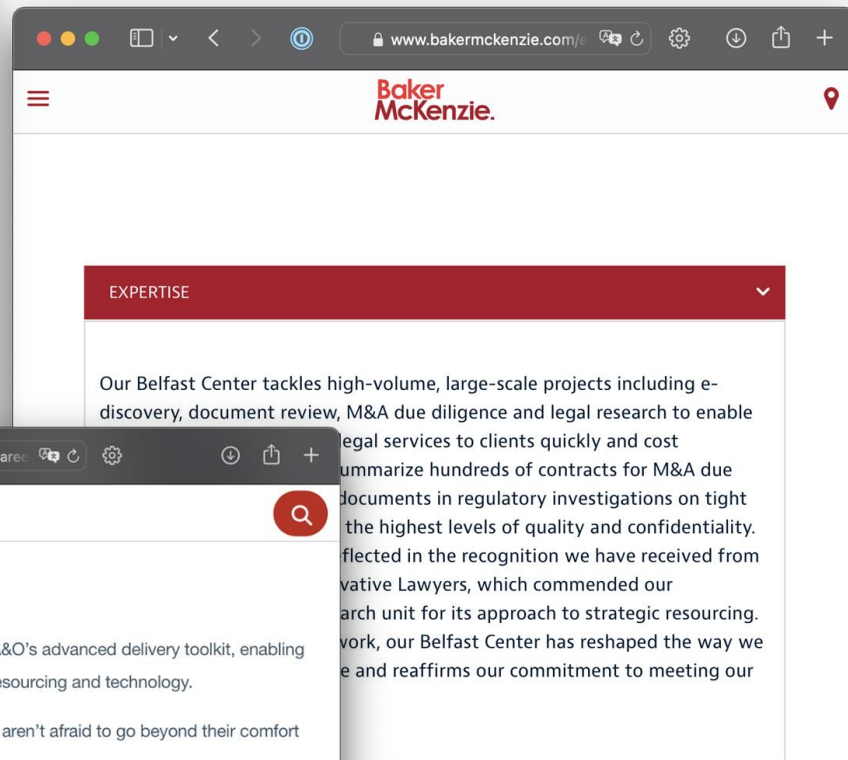
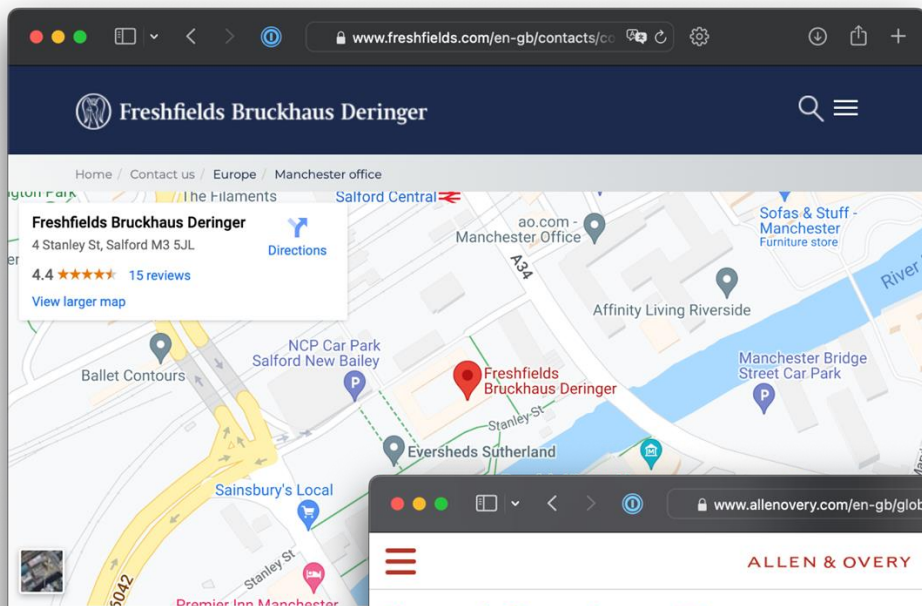
"Elevate is addressing a critical need in a sizable market, with a new business model, just as change is beginning to take hold in the legal sector," said Leon Chen, a partner with Kayne Partner Funds who has joined Elevate's board of directors as part of the Kayne investment deal.

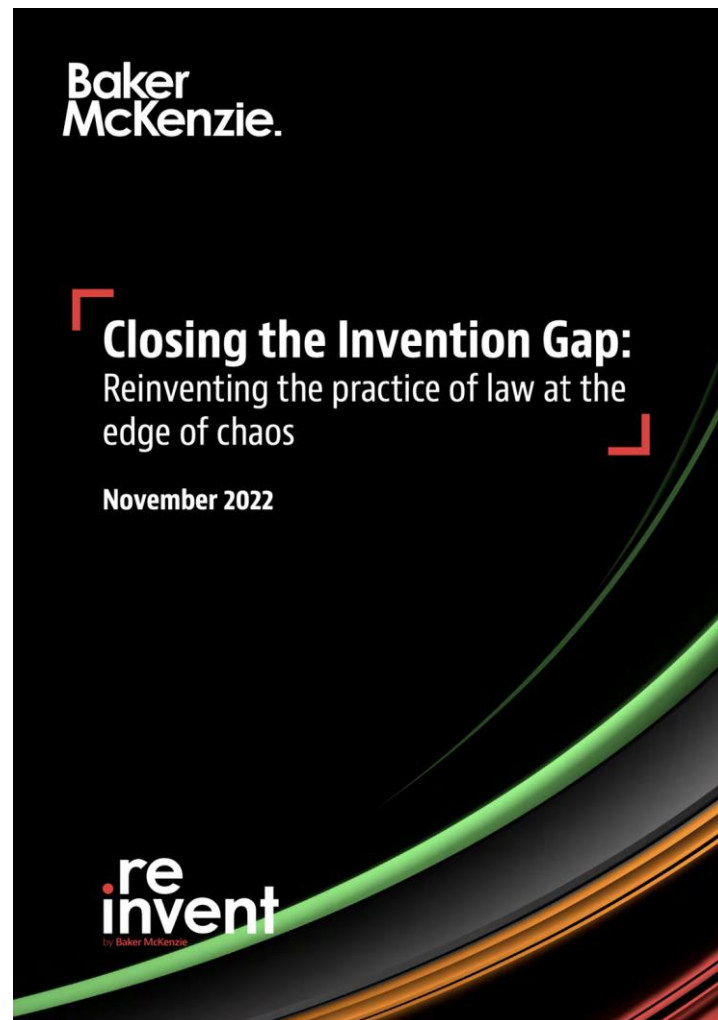
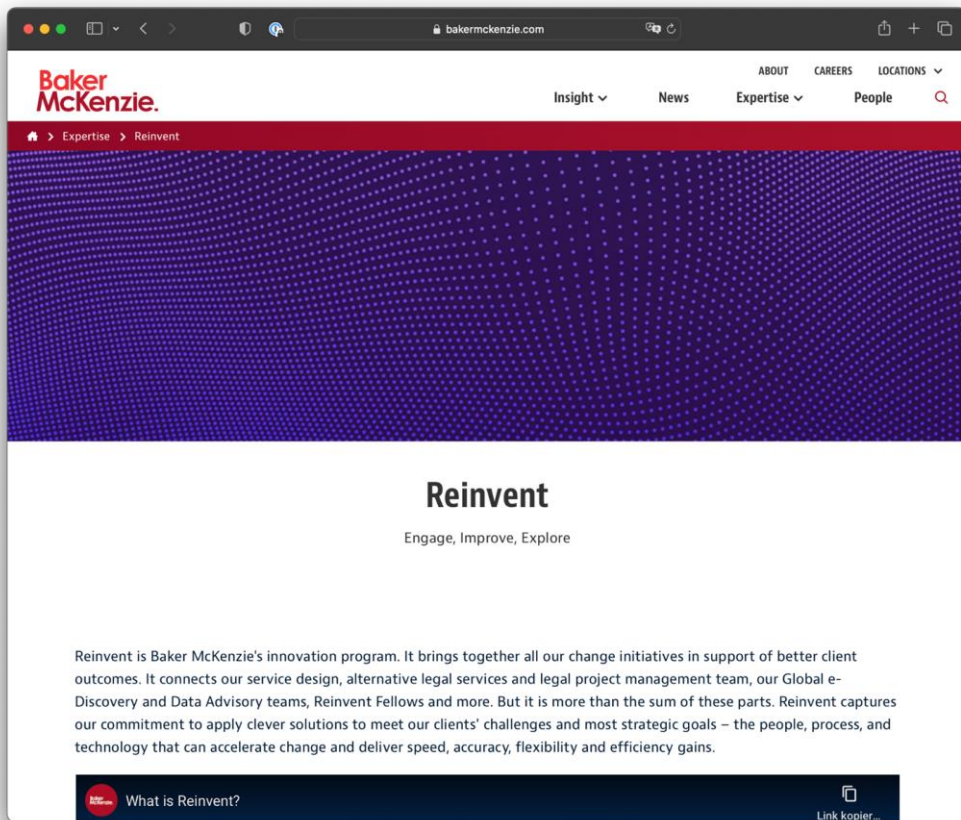
## M&A activity

		Consulting	Services	Technology
LEXPREDICT	Nov '18	+		+
sumati	Dec '18		+	+
HALEBURY	Dec '18		+	
terra SOLUTIONS	Jan '19	+	+	
Cognatio	Jan '19		+	
	Future?			

- AI
- Expertise
- Capabilities
- Practice of law
- Management
- Scale
- Footprint
- Customers

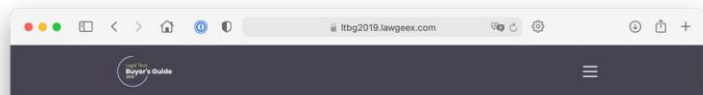
# Legal Services Centers for Auxiliary Services







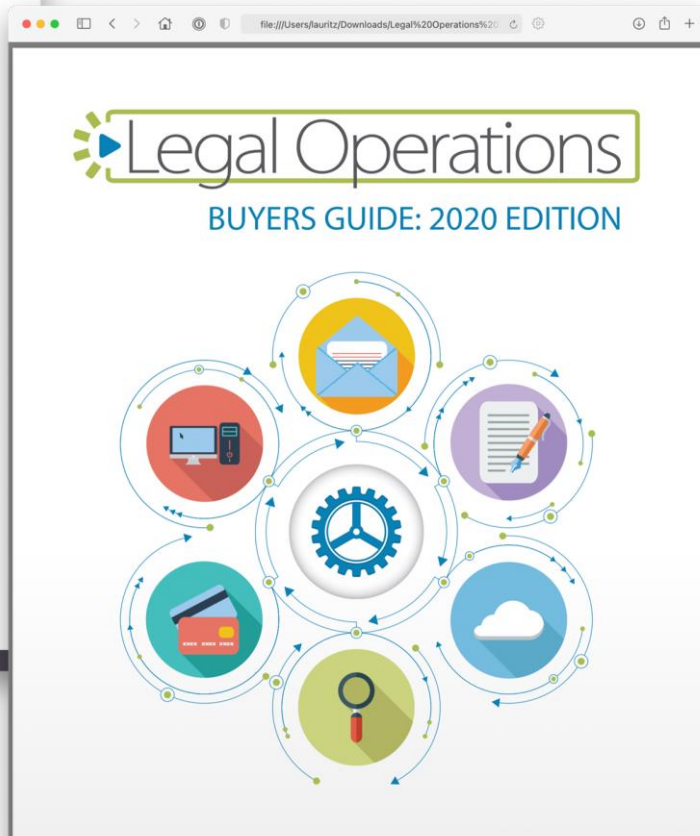
# Legal Tech Consulting



## 2019 Legal Tech Buyer's Guide

This LegalTech Buyer's Guide provides in-depth analysis of leading players across 12 categories of legal technology, including Contract Review Automation, Contract Management, eDiscovery, Legal Research, Communications, matter management, and e-Discovery. It covers over 130 top technology solutions, and offers jargon-free explanations of legaltech "buzzwords" — including artificial intelligence, legal operations, blockchain, and the cloud.

Start Reading!





# Law Companies



# Captive ALSP Tech Subsidiaries



Global directory Product logins Support & training Contact us

THOMSON REUTERS

## Captive law firm ALSPs provide potential growth engines

LEGAL MARKETPLACE

William Josten, Manager of Strategic Enterprise Thought Leadership at Thomson Reuters

17 Mar 2021

Although it's the smallest segment, captive in-house ALSPs have proven themselves mighty, achieving the highest percentage growth rate of any ALSP segment

To dissect some of the key findings of Thomson Reuters recent **2021 Alternative Legal Service Providers** report, it makes some sense to start small, meaning starting with the smallest segment of alternative legal service providers (ALSPs) — those owned and run as captive entities within law firms.

Why start with the smallest segment for analysis? Because though small, these businesses have proven mighty, achieving the highest percentage growth rate of any segment of ALSPs.

In 2015, nascent captive ALSPs accounted for only about \$150 million in global annual revenue; by 2019, that is estimated to have grown to roughly \$480 million. The overall market for ALSPs increased by about \$5 billion in that same time period. From 2017-'19 alone, ALSP market share saw estimated compound annual growth of 15%, more than twice the compound

### Solutions

#### Partner Summit 2021

For tax and accounting leaders, in person

Select an event >

THOMSON REUTERS

#### Partner Summit 2021

Partner Summit events announced: Over the past decade, Thomson Reuters Partner Summits have given firm leaders a unique opportunity to learn how to build a progressive, advisory services program. Become part of this exceptional group at one of this year's events.

#### Featured event

MAR 25, 2021

#### Divergent Design: Assessing the 2021 Global Alternative Legal Service Providers Report

Alternative legal service providers (ALSPs), otherwise known as "New Law" entrants, continue to gain traction in our modern legal services ...

Event details

#### Related posts



# Legal Operations

Professionalization

Labor Arbitrage

## Activities: WHAT?

## Levers: HOW?

## Impact: WHY?



**Vendor Management**



**Risk Management**



**Knowledge Mgmt.**



**Financial Management**



**Technology Mgmt.**



**Data Analytics**



**Technology**



**Inter-/Intrafunctional  
Collaboration**



**Process Improvement**

**Better Resource Efficiency,  
Cost Effectiveness**

**Better Work**

**Better Talent Retention**

# Legal Operations

The Atlantic Popular Latest Sections Magazine



## Rise of the Robolawyers

How legal representation could come to resemble TurboTax

JASON KESSELER | APRIL 2017 ISSUE | TECHNOLOGY

COLOR: JESSICA BROWN BUSINESS 12:00 PM 12:00 AM

## TECH WILL FORCE LAWYERS TO DO MORE FOR THOSE BILLABLE HOURS



PLAY #SILICON VALLEY #STARTUPS #TECH

## How artificial intelligence is transforming the legal profession

POSTED APRIL 11, 2016 10:45 AM EDT BY JULIE SCHWABE

Over decades, an endorsement with a background in advertising and media, some on the list of the 100 most powerful people in the world, who worked as a television attorney, passed over for the job of representing the legal profession.




It's a bold statement, and it's not just about the technology that can do a job that's been done for decades.

## A.I. Is Doing Legal Work. But It Won't Replace Lawyers, Yet.

By STEVE LOHR | MARCH 19, 2017

The New York Times

The Clickbait Version



Sorry, a Robot Is Not About to Replace Your Lawyer

The legal profession relies more and more on automation. But fears that it will be automated out of existence are overblown, researchers say. For now.

NYTIMES.COM

TECH

## Why Hire a Lawyer When a Robot Will Do?

11 SEP 22, 2016 3:00 AM EDT

By Elaine Ou

BloombergView



## FINANCIAL TIMES

Law + Add to myFT

### Artificial intelligence closes in on the work of junior lawyers

'Lawtech' sifts and summarises data with speed and precision to replace routine tasks



FUTURE OF WORK

CNBC

## Lawyers could be the next profession to be replaced by computers

Der Spiegel | @\_Dorffinger | Friday, 17 Feb 2017 1:55 PM ET



April 11, 2016 7:23 am

## Technology: Breaking the law

Michael Spector

FT Financial Times

The legal profession hasn't changed in decades. Here's the case for the Uberisation of it.

on.ft.com/1S2OHYP



FT FINANCIAL TIMES



**WE WOULD HAVE ~2000 LEGAL TECH  
COMPANIES ACROSS THE WORLD**

**LEGAL OPS WOULD GROW  
EXPONENTIALLY**

**A BUNCH OF LAW FIRMS WOULD  
HAVE TECH INCUBATORS**

**A NUMBER OF LAW SCHOOLS WOULD  
HAVE PROGRAMS, CERTIFICATES AND  
DEGREES IN LEGAL INNOVATION**

**A TOP 25 BRITISH LAW FIRM  
WOULD BE LISTED**







# Legal NLP—Breaking the Legal Language Barrier?

Dirk Hartung & Daniel Martin Katz

A LEGAL COMPLEXITY PICTURE

# LAW LAW LAND

Featuring

NATURAL LANGUAGE and  
DOMAIN-SPECIFIC JARGON

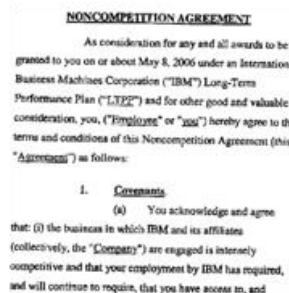


Where Natural Language is the Coin of the Realm

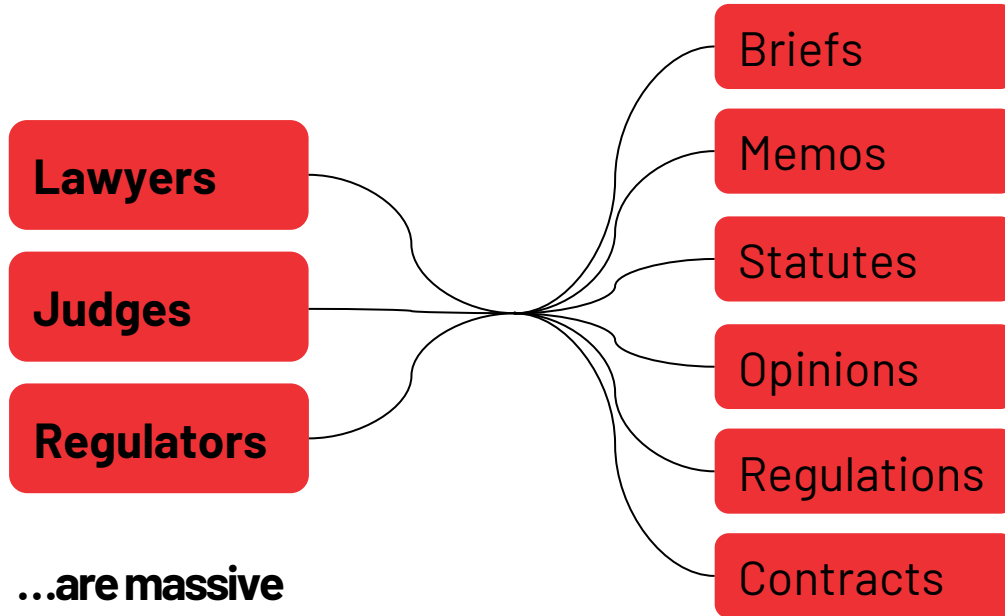
...

# But Law is Not Just About the Consumption of Natural Language

Law / Lawyering is (in part) an exercise in linguistic construction and interpretation



# Text Production at a Massive Scale



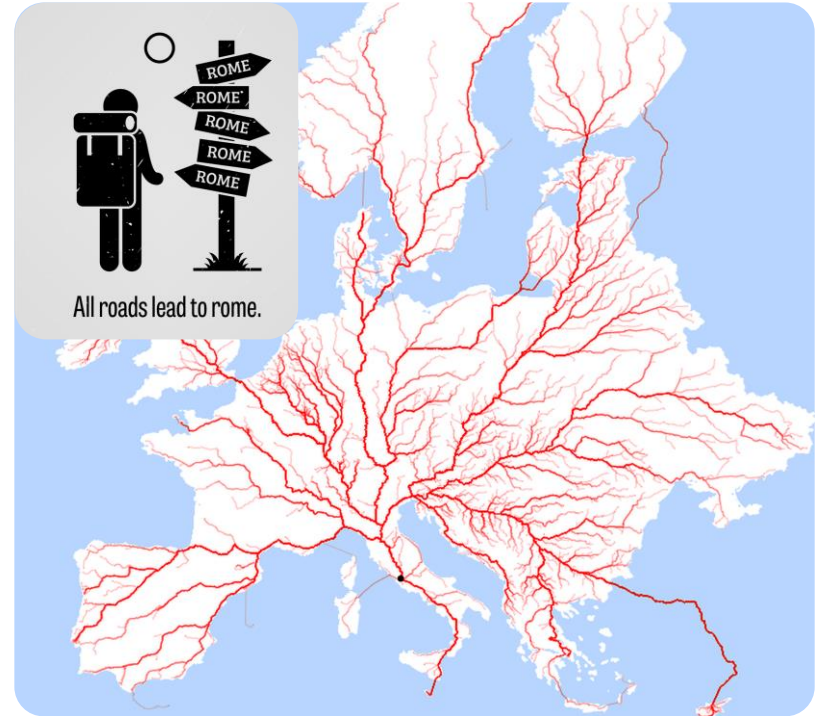
**...are massive  
producers of  
text**

**these are just some of the  
legal work product being  
produced on a daily basis  
across the world's various  
legal systems**

# All Most Roads in Law Lead to a Document

**In Law, most roads lead to a document ...**

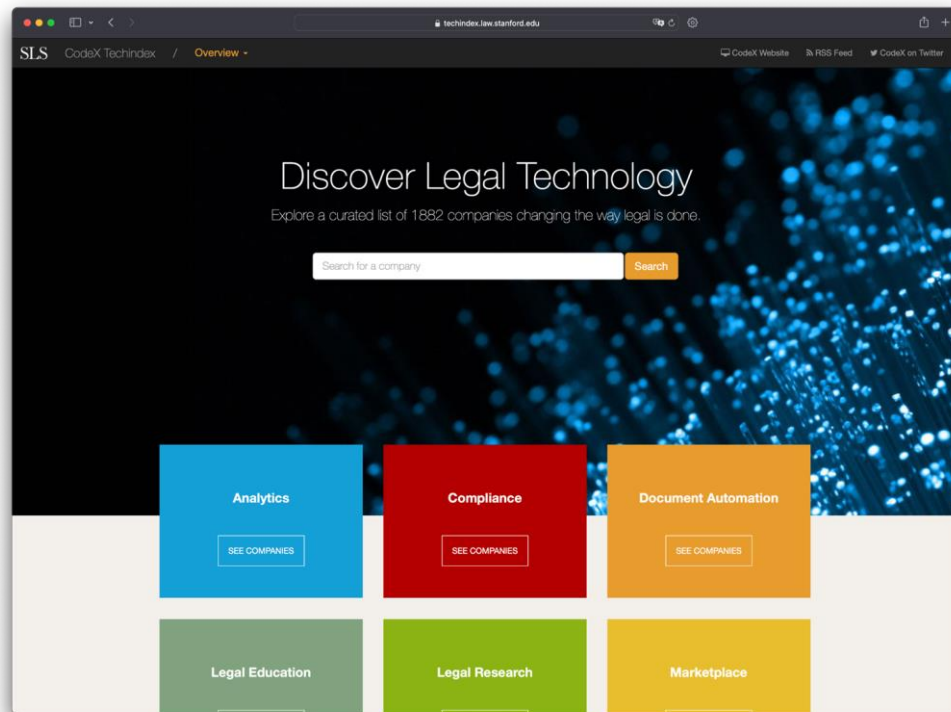
**And that document is very likely to be expressed in natural language ...**



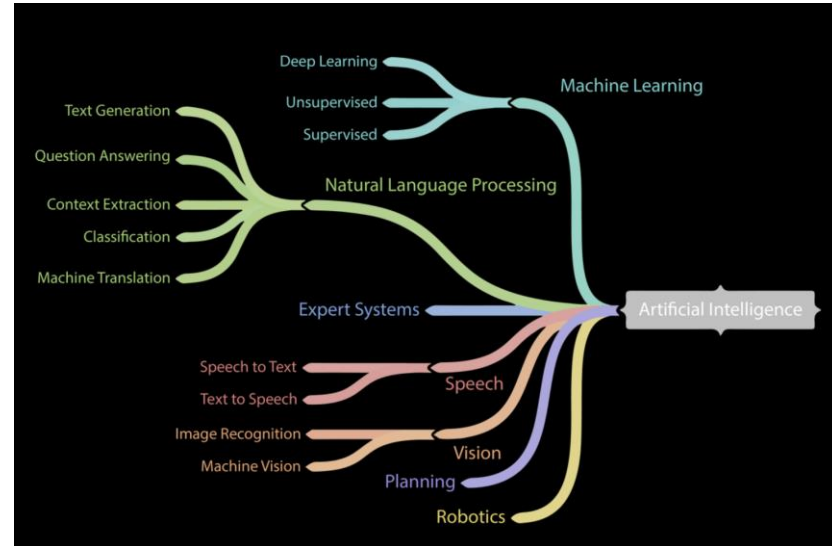
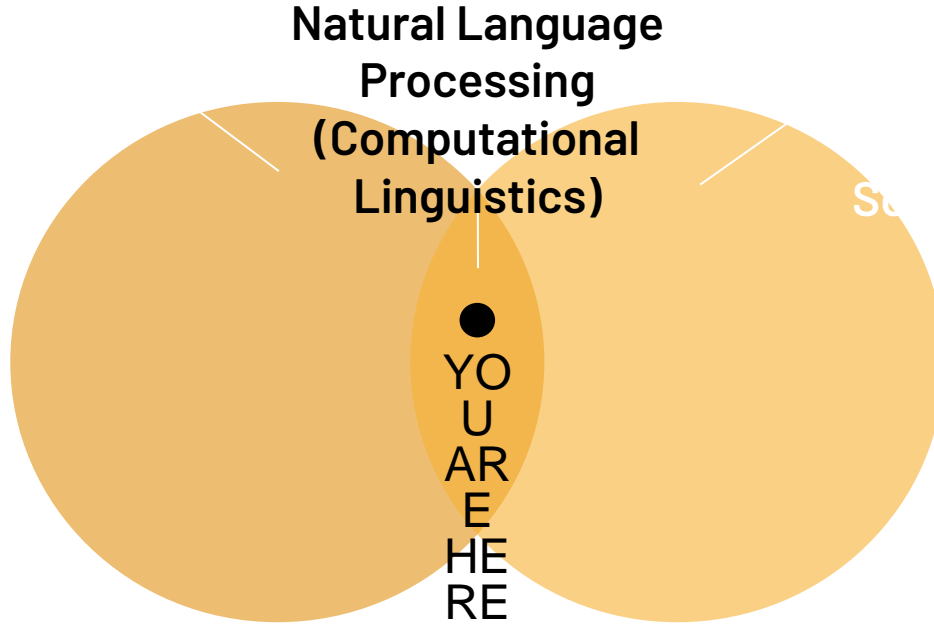


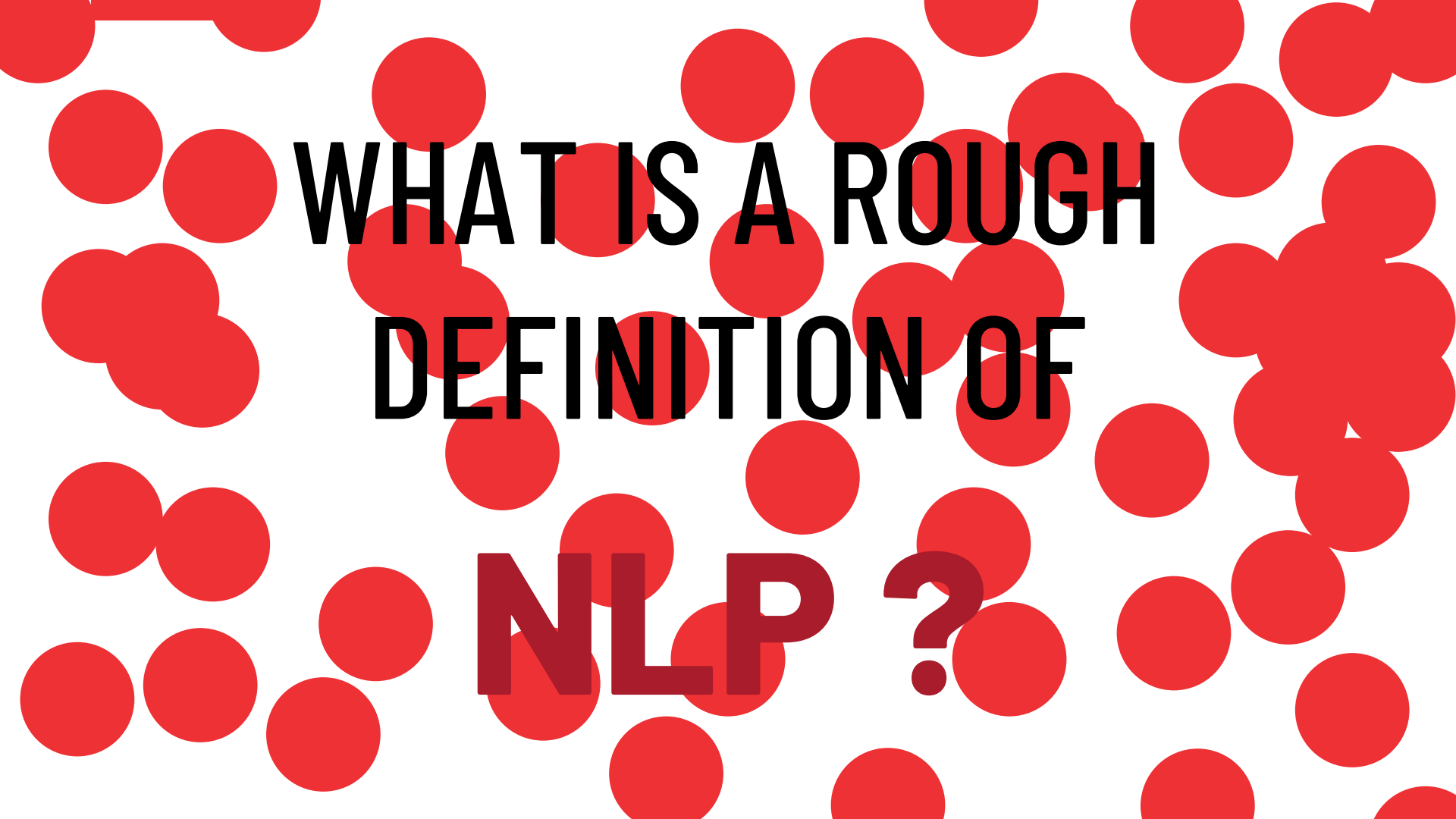
# Significant Growth in Document Technology

[techindex.law.stanford.edu](http://techindex.law.stanford.edu)



# NLP as a Branch of AI





**WHAT IS A ROUGH  
DEFINITION OF  
NLP ?**



**It is the Statistical  
Representation of  
Language ...**



# Historically, Big Divide between Semantics and Syntax

**Syntax Methods**  
(Fairly Easy)

**Semantic Methods**  
(Fairly Difficult)



**There have been a series of  
clever approaches to  
backdoor into semantics\* ...**

(\*while also being scalable)



# Historically, Big Divide between Semantics and Syntax

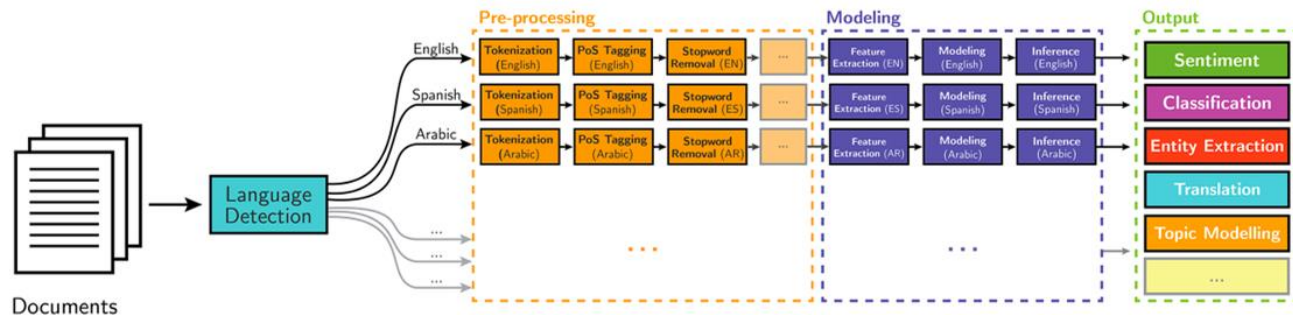
**Syntax Methods**  
(Fairly Easy)

Quasi-Semantic  
Methods

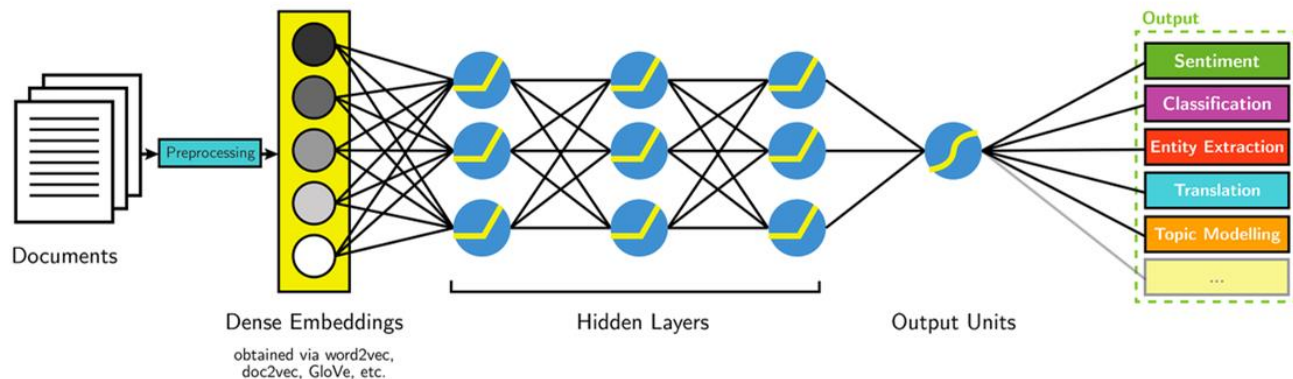
**Semantic Methods**  
(Fairly Difficult)

# The Age of 'Neural' NLP

## Classical NLP



## Deep Learning-based NLP



# Word2Vec (2013)



arXiv:1301.3781v3 [cs.CL] 7 Sep 2013

## Efficient Estimation of Word Representations in Vector Space

**Tomas Mikolov**

Google Inc., Mountain View, CA  
tmikolov@google.com

**Kai Chen**

Google Inc., Mountain View, CA  
kaichen@google.com

**Greg Corrado**

Google Inc., Mountain View, CA  
gcorrado@google.com

**Jeffrey Dean**

Google Inc., Mountain View, CA  
jeff@google.com

### Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

### 1 Introduction

Many current NLP systems and techniques treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary. This choice has several good reasons - simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data. An example is the popular N-gram model used for statistical language modeling - today, it is possible to train N-grams on virtually all available data (trillions of words [3]).

However, the simple techniques are at their limits in many tasks. For example, the amount of relevant in-domain data for automatic speech recognition is limited - the performance is usually dominated by the size of high quality transcribed speech data (often just millions of words). In machine translation, the existing corpora for many languages contain only a few billions of words or less. Thus, there are situations where simple scaling up of the basic techniques will not result in any significant progress, and we have to focus on more advanced techniques.

With progress of machine learning techniques in recent years, it has become possible to train more complex models on much larger data set, and they typically outperform the simple models. Probably the most successful concept is to use distributed representations of words [10]. For example, neural network based language models significantly outperform N-gram models [1, 27, 17].

#### 1.1 Goals of the Paper

The main goal of this paper is to introduce techniques that can be used for learning high-quality word

# Attention (2015)



STANFORD  
UNIVERSITY

arXiv:1508.04025v5 [cs.CL] 20 Sep 2015

## Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong Hieu Pham Christopher D. Manning  
Computer Science Department, Stanford University, Stanford, CA 94305  
{lmthang, hyhieu, manning}@stanford.edu

### Abstract

An attentional mechanism has lately been used to improve neural machine translation (NMT) by selectively focusing on parts of the source sentence during translation. However, there has been little work exploring useful architectures for attention-based NMT. This paper examines two simple and effective classes of attentional mechanism: a *global* approach which always attends to all source words and a *local* one that only looks at a subset of source words at a time. We demonstrate the effectiveness of both approaches on the WMT translation tasks between English and German in both directions. With local attention, we achieve a significant gain of 5.0 BLEU points over non-attentional systems that already incorporate known techniques such as dropout. Our ensemble model using different attention architectures yields a new state-of-the-art result in the WMT'15 English to German translation task with 25.9 BLEU points, an improvement of 1.0 BLEU points over the existing best system backed by NMT and an  $n$ -gram reranker.<sup>1</sup>

### 1 Introduction

Neural Machine Translation (NMT) achieved state-of-the-art performances in large-scale translation tasks such as from English to French (Luong et al., 2015) and English to German (Jean et al., 2015). NMT is appealing since it requires minimal domain knowledge and is conceptually simple. The model by Luong et al. (2015) reads through all the source words until the end-of-sentence symbol  $\langle\text{eos}\rangle$  is reached. It then starts

<sup>1</sup>All our code and models are publicly available at <http://nlp.stanford.edu/projects/nmt>.

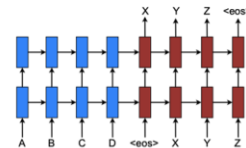


Figure 1: Neural machine translation – a stacking recurrent architecture for translating a source sequence A B C D into a target sequence X Y Z. Here,  $\langle\text{eos}\rangle$  marks the end of a sentence.

emitting one target word at a time, as illustrated in Figure 1. NMT is often a large neural network that is trained in an end-to-end fashion and has the ability to generalize well to very long word sequences. This means the model does not have to explicitly store gigantic phrase tables and language models as in the case of standard MT; hence, NMT has a small memory footprint. Lastly, implementing NMT decoders is easy unlike the highly intricate decoders in standard MT (Koehn et al., 2003).

In parallel, the concept of “attention” has gained popularity recently in training neural networks, allowing models to learn alignments between different modalities, e.g., between image objects and agent actions in the dynamic control problem (Mnih et al., 2014), between speech frames and text in the speech recognition task (?), or between visual features of a picture and its text description in the image caption generation task (Xu et al., 2015). In the context of NMT, Bahdanau et al. (2015) has successfully applied such attentional mechanism to jointly translate and align words. To the best of our knowledge, there has not been any other work exploring the use of attention-based architectures for NMT.

In this work, we design, with simplicity and ef-

# Transformer Architecture (2017)



Google AI

arXiv:1706.03762v5 [cs.CL] 6 Dec 2017

## Attention Is All You Need

Ashish Vaswani\*  
Google Brain  
avaswani@google.com

Noam Shazeer\*  
Google Brain  
noam@google.com

Niki Parmar\*  
Google Research  
niki@google.com

Jakob Uszkoreit\*  
Google Research  
usz@google.com

Llion Jones\*  
Google Research  
llion@google.com

Aidan N. Gomez\*<sup>†</sup>  
University of Toronto  
aidan@cs.toronto.edu

Lukasz Kaiser\*  
Google Brain  
lukaszkaiser@google.com

Illia Polosukhin\*<sup>‡</sup>  
illia.polosukhin@gmail.com

### Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

### 1 Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

\*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

<sup>†</sup>Work performed while at Google Brain.

<sup>‡</sup>Work performed while at Google Research.

# ELMo (2018)



**W** PAUL G. ALLEN SCHOOL  
OF COMPUTER SCIENCE & ENGINEERING

arXiv:1802.05365v2 [cs.CL] 22 Mar 2018

## Deep contextualized word representations

Matthew E. Peters<sup>†</sup>, Mark Neumann<sup>†</sup>, Mohit Iyyer<sup>†</sup>, Matt Gardner<sup>†</sup>,  
{matthewp, markn, mohiti, mattg}@allenai.org

Christopher Clark<sup>\*,</sup>, Kenton Lee<sup>\*,</sup>, Luke Zettlemoyer<sup>1\*</sup>  
{csquared, kentonl, lsz}@cs.washington.edu

<sup>†</sup>Allen Institute for Artificial Intelligence

<sup>\*</sup>Paul G. Allen School of Computer Science & Engineering, University of Washington

### Abstract

We introduce a new type of *deep contextualized* word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). Our word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus. We show that these representations can be easily added to existing models and significantly improve the state of the art across six challenging NLP problems, including question answering, textual entailment and sentiment analysis. We also present an analysis showing that exposing the deep internals of the pre-trained network is crucial, allowing downstream models to mix different types of semi-supervision signals.

### 1 Introduction

Pre-trained word representations (Mikolov et al., 2013; Pennington et al., 2014) are a key component in many neural language understanding models. However, learning high quality representations can be challenging. They should ideally model both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). In this paper, we introduce a new type of *deep contextualized* word representation that directly addresses both challenges, can be easily integrated into existing models, and significantly improves the state of the art in every considered case across a range of challenging language un-

guage model (LM) objective on a large text corpus. For this reason, we call them ELMo (Embeddings from Language Models) representations. Unlike previous approaches for learning contextualized word vectors (Peters et al., 2017; McCann et al., 2017), ELMo representations are deep, in the sense that they are a function of all of the internal layers of the biLM. More specifically, we learn a linear combination of the vectors stacked above each input word for each end task, which markedly improves performance over just using the top LSTM layer.

Combining the internal states in this manner allows for very rich word representations. Using intrinsic evaluations, we show that the higher-level LSTM states capture context-dependent aspects of word meaning (e.g., they can be used without modification to perform well on supervised word sense disambiguation tasks) while lower-level states model aspects of syntax (e.g., they can be used to do part-of-speech tagging). Simultaneously exposing all of these signals is highly beneficial, allowing the learned models select the types of semi-supervision that are most useful for each end task.

Extensive experiments demonstrate that ELMo representations work extremely well in practice. We first show that they can be easily added to existing models for six diverse and challenging language understanding problems, including textual entailment, question answering and sentiment analysis. The addition of ELMo representations alone significantly improves the state of the art in every case, including up to 20% relative error



# BERT (2019)



arXiv:1810.04805v2 [cs.CL] 24 May 2019

## BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

Google AI Language

{jacobdevlin, mingweichang, kentonl, kristout}@google.com

### Abstract

We introduce a new language representation model called **BERT**, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

### 1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the re-

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from Transformers. BERT alleviates the previously mentioned unidi-

# The GPT Trilogy



2018

---

## Improving Language Understanding by Generative Pre-Training

---

Alec Radford	Karthik Narasimhan	Tim Salimans	Ilya Sutskever
OpenAI	OpenAI	OpenAI	OpenAI
<a href="mailto:alec@openai.com">alec@openai.com</a>	<a href="mailto:karthikn@openai.com">karthikn@openai.com</a>	<a href="mailto:tim@openai.com">tim@openai.com</a>	<a href="mailto:ilyasu@openai.com">ilyasu@openai.com</a>

### Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

# GPT-1

# The GPT Trilogy



2019

---

## Language Models are Unsupervised Multitask Learners

---

Alec Radford <sup>\*1</sup> Jeffrey Wu <sup>\*1</sup> Rewon Child <sup>1</sup> David Luan <sup>1</sup> Dario Amodei <sup>\*\*1</sup> Ilya Sutskever <sup>\*\*1</sup>

### Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models ([Lake et al., 2017](#)), reading comprehension systems ([Jia & Liang, 2017](#)), and image classifiers ([Alcorn et al., 2018](#)) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE ([Wang et al., 2018](#)) and decaNLP ([McCann et al., 2018](#)) to begin studying this.

Multitask learning ([Caruana, 1997](#)) is a promising framework for improving general performance. However, multitask training in NLP is still nascent. Recent work reports modest performance improvements ([Yogatama et al., 2019](#)) and the two most ambitious efforts to date have

# GPT-2

# The GPT Trilogy



2020

---

## Language Models are Few-Shot Learners

---

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan†	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess		Jack Clark	Christopher Berner	
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei	

OpenAI

GPT-3

# Big Bird



Google AI

2021

arXiv:2007.14062v2 [cs.LG] 8 Jan 2021

## Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Avinava Dubey,  
Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham,  
Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed  
Google Research  
(manzilz, gurug, avinavadubey)@google.com

### Abstract

Transformers-based models, such as BERT, have been one of the most successful deep learning models for NLP. Unfortunately, one of their core limitations is the quadratic dependency (mainly in terms of memory) on the sequence length due to their full attention mechanism. To remedy this, we propose, BIGBIRD, a sparse attention mechanism that reduces this quadratic dependency to linear. We show that BIGBIRD is a universal approximator of sequence functions and is Turing complete, thereby preserving these properties of the quadratic, full attention model. Along the way, our theoretical analysis reveals some of the benefits of having  $O(1)$  global tokens (such as CLS), that attend to the entire sequence as part of the sparse attention mechanism. The proposed sparse attention can handle sequences of length up to 8x of what was previously possible using similar hardware. As a consequence of the capability to handle longer context, BIGBIRD drastically improves performance on various NLP tasks such as question answering and summarization. We also propose novel applications to genomics data.

### 1 Introduction

Models based on Transformers [91], such as BERT [22, 63], are wildly successful for a wide variety of Natural Language Processing (NLP) tasks and consequently are mainstay of modern NLP research. Their versatility and robustness are the primary drivers behind the wide-scale adoption of Transformers. The model is easily adapted for a diverse range of sequence based tasks – as a seq2seq model for translation [91], summarization [66], generation [15], etc. or as a standalone encoders for sentiment analysis [83], POS tagging [65], machine reading comprehension [93], etc. – and it is known to vastly outperform previous sequence models like LSTM [37]. The key innovation in Transformers is the introduction of a self-attention mechanism, which can be evaluated in parallel for each token of the input sequence, eliminating the sequential dependency in recurrent neural networks, like LSTM. This parallelism enables Transformers to leverage the full power of modern SIMD hardware accelerators like GPUs/TPUs, thereby facilitating training of NLP models on datasets of unprecedented size. This ability to train on large scale data has led to surfacing of models like BERT [22] and T5 [75], which pretrain transformers on large general purpose corpora and transfer the knowledge to downstream tasks. The pretraining has led to significant improvement in low data regime downstream tasks [51] as well as tasks with sufficient data [101] and thus have been a major force behind the ubiquity of transformers in contemporary NLP.

The self-attention mechanism overcomes constraints of RNNs (namely the sequential nature of RNN) by allowing each token in the input sequence to attend independently to every other token in the sequence. This design choice has several interesting repercussions. In particular, the full self-attention have computational and memory requirement that is quadratic in the sequence length. We note that while the corpus can be large, the sequence length, which provides the context in many applications is very limited. Using commonly available current hardware and model sizes, this requirement

# InstructGPT



# 2022

arXiv:2203.02155v1 [cs.CL] 4 Mar 2022

## Training language models to follow instructions with human feedback

Long Ouyang\* Jeff Wu\* Xu Jiang\* Diogo Almeida\* Carroll L. Wainwright\*  
Pamela Mishkin\* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray  
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens  
Amanda Askell<sup>†</sup> Peter Welinder Paul Christiano\*<sup>†</sup>  
Jan Leike\* Ryan Lowe\*  
OpenAI

### Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models *InstructGPT*. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

### 1 Introduction


Large language models (LMs) can be “prompted” to perform a range of natural language processing (NLP) tasks, given some examples of the task as input. However, these models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions (Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021; Weidinger et al., 2021; Tamkin et al., 2021; Gehman et al., 2020). This is because the language modeling objective

\*Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads. Corresponding author: Lowe@openai.com.

<sup>†</sup>Work done while at OpenAI. Current affiliations: AA: Anthropic; PC: Alignment Research Center.

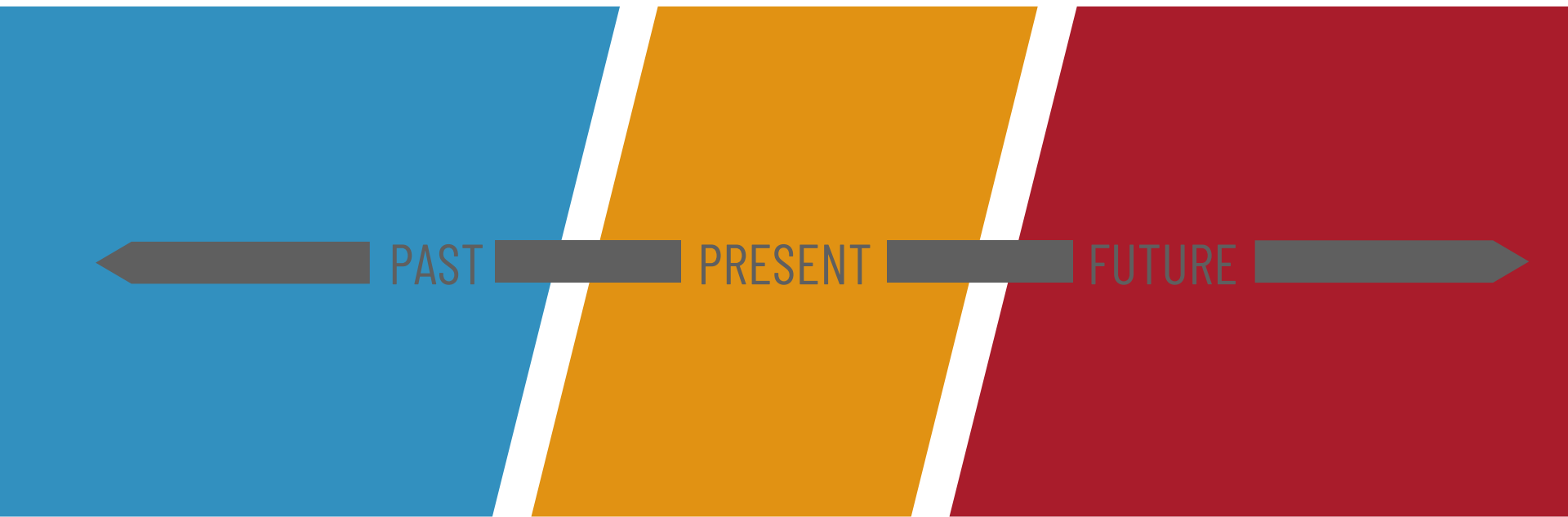






“ Okay that is general  
NLP but what about  
‘LEGAL NLP’ ... ?

# LEGAL NLP



# Natural Language Processing in the Legal Domain

Daniel Martin Katz<sup>1,2,3,4,†,\*</sup>, Dirk Hartung<sup>2,3,†</sup>, Lauritz Gerlach<sup>2</sup>, Abhik Jana<sup>5</sup>, and Michael J. Bommarito<sup>2,3,4</sup>

<sup>†</sup> These authors have contributed equally to this work and share first authorship.

<sup>1</sup>Illinois Tech - Chicago Kent College of Law, USA

<sup>2</sup>Bucerius Law School, Germany

<sup>3</sup>CodeX, Stanford University, USA

<sup>4</sup>273 Ventures, USA

<sup>5</sup>Universität Hamburg, Germany

\*e-mail - dkatz3@kentlaw.iit.edu

February 22, 2023

Version 2.01

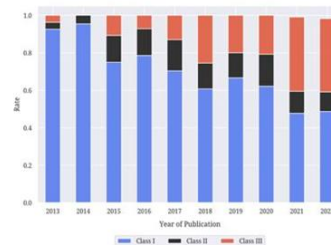
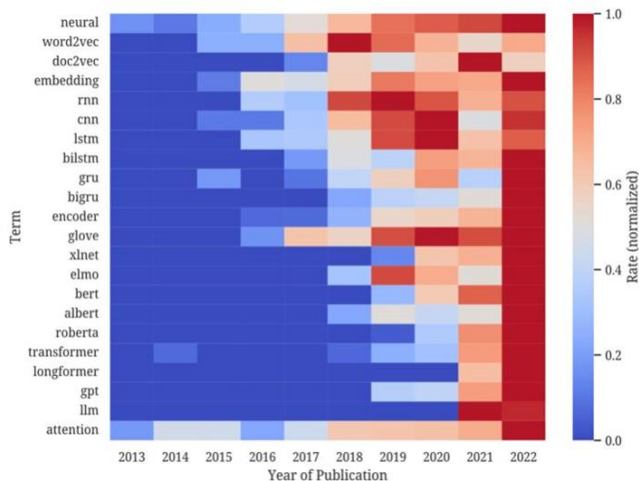
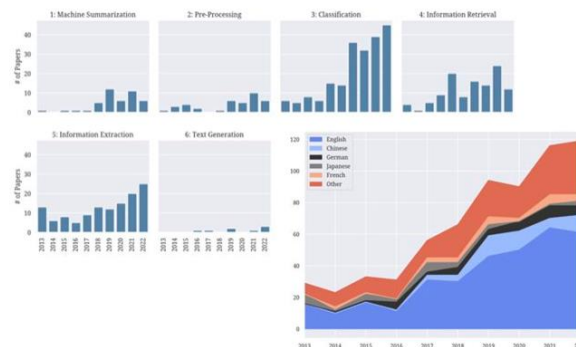
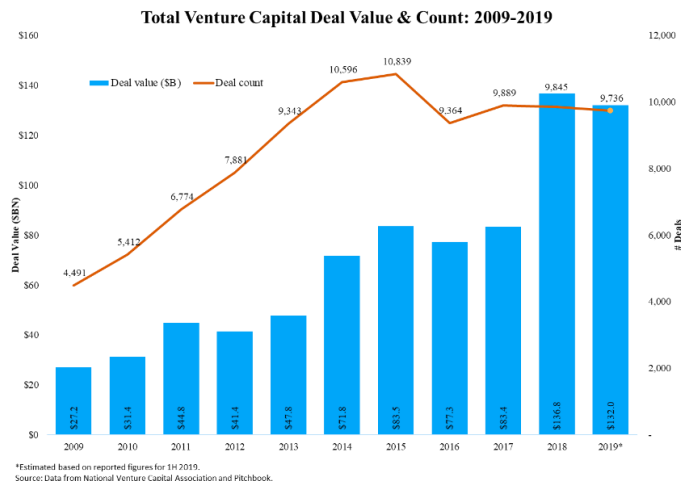


Figure 6. Replication Material Availability as a Function of Time



# The 2010's is the Decade Where the Academic and Commercial Worlds Began to Really Collide ...



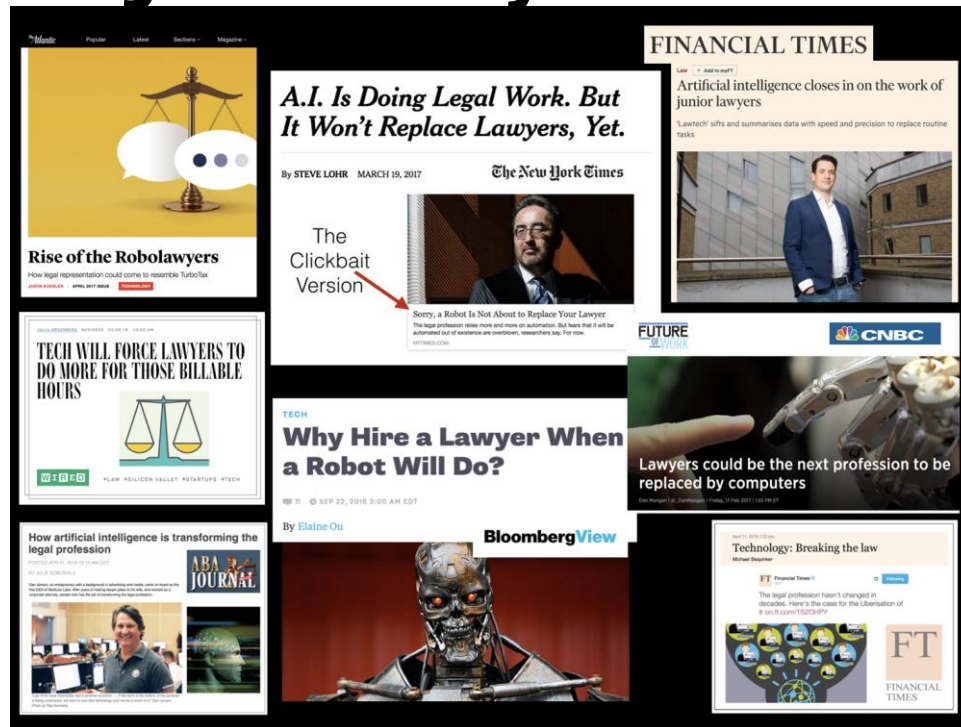
## TECHNOLOGY


### At \$1.2 Billion, It's Already A Record Year For Legal Tech Investment

Will this investment trend continue? At this point, there is no turning back.

By ROBERT AMBROGI

Sep 16, 2019 at 4:16 PM





**But if we look at both the academic and commercial sphere, we still observe a fairly thin account for legal language ...**

**Certainly as compared to humans and expert lawyers ...**

**But this not an uncommon issue across the NLP world**





## Amazon Comprehend Medical

Extract information from unstructured medical text accurately and quickly

Get Started with Amazon Comprehend Medical

8.5M characters of text  
free for the first month  
with the [AWS Free Tier](#)

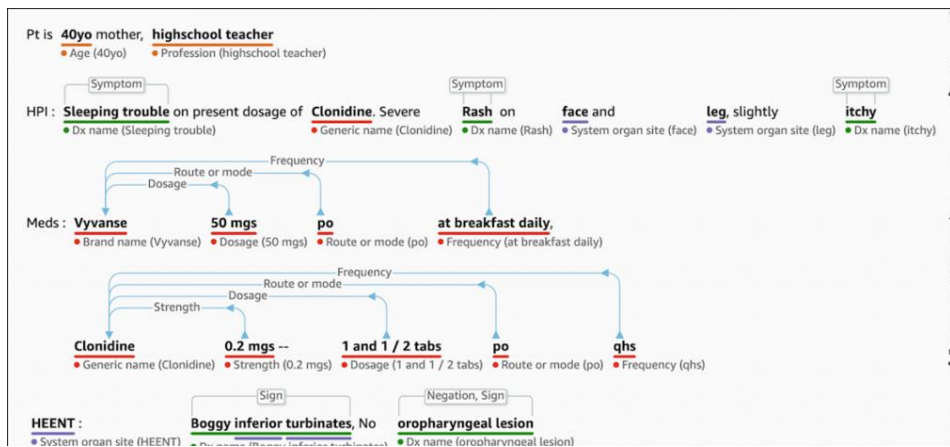
Extract medical information from unstructured medical text like doctors' notes, clinical trial reports, or radiology reports

Identify relationships among extracted health information and link to medical ontologies like ICD-10-CM, RxNorm, and SNOMED CT

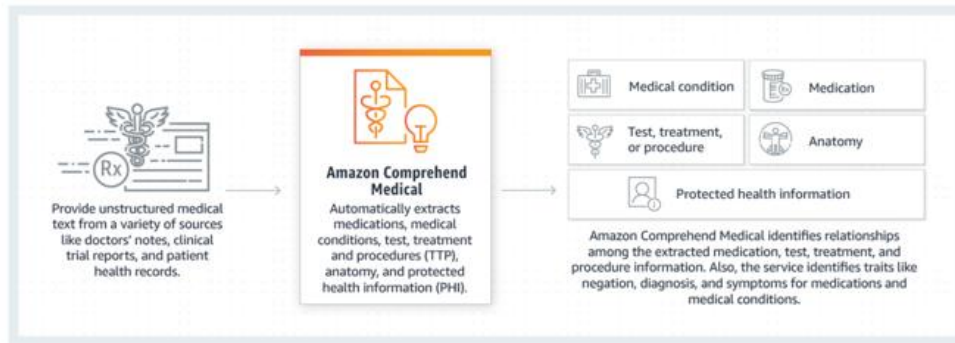
Automate and lower the cost of processing and coding unstructured medical text with easy to use APIs

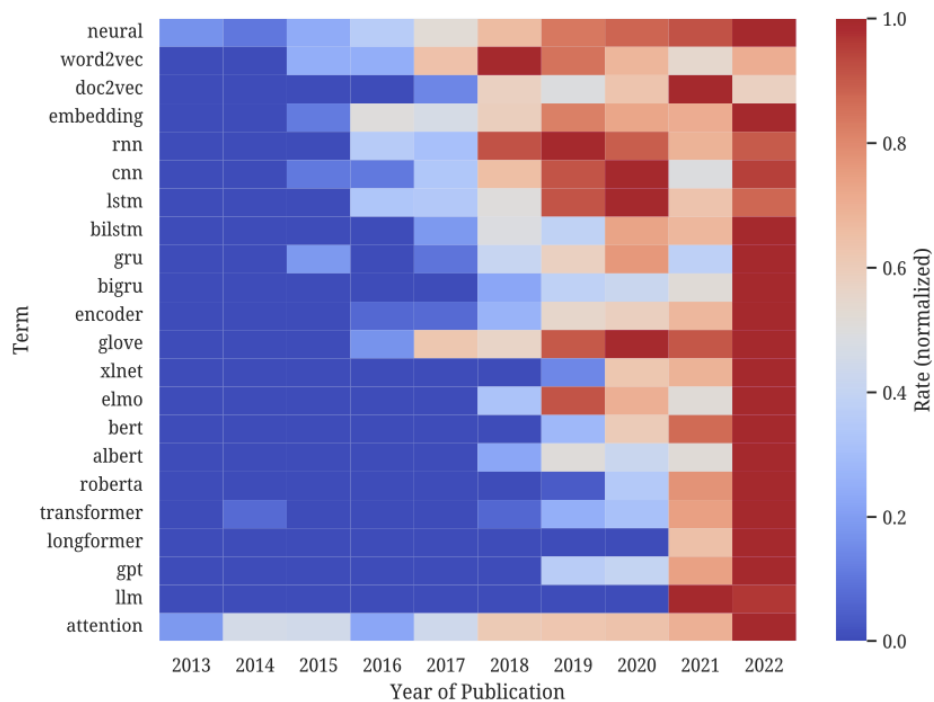
Implement patient data privacy solutions and identify protected health information (PHI) with a HIPAA-eligible service

# The need to understand Sub-Dialects of English is a familiar problem ...




## How it works





**Figure 4.** Relative Rate of Term Usage over Time. Normalization is per-term relative to the maximum annual rate of mentioning papers.



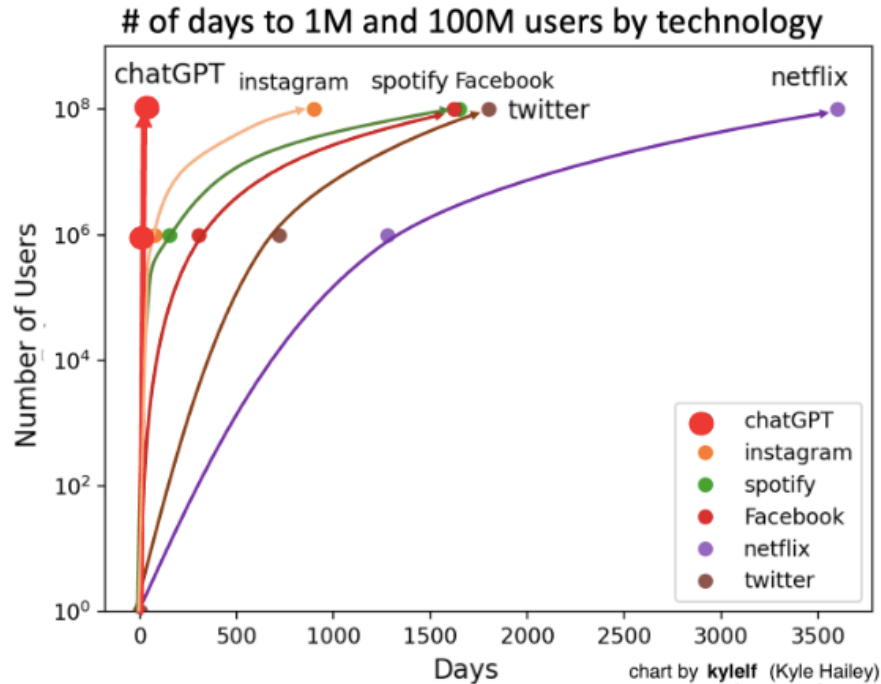
So the Scientific / Engineering task at hand was to improve the performance of Legal NLP Models ...

By further breaking down the legal language barrier

By grafting broader NLP developments to Domain Specific Needs in Law



# Nov 30, 2022



## ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users



\* one million backers \*\* one million nights booked \*\*\* one million downloads

Source: Company announcements via Business Insider/LinkedIn



statista

# GPT Takes the Bar Exam

7 Pages • Posted: 31 Dec 2022 • Last revised: 3 Jan 2023

Michael James Bommarito

273 Ventures; Licensio, LLC; Bommarito Consulting, LLC; Michigan State College of Law; Stanford Center for Legal Informatics

Daniel Martin Katz

Illinois Tech - Chicago Kent College of Law; Bucerius Center for Legal Technology & Data Science; Stanford CodeX - Center for Legal Informatics; 273 Ventures

Date Written: December 29, 2022

## GPT Takes the Bar Exam

Michael Bommarito II<sup>1,2,3,\*</sup>, Daniel Martin Katz<sup>1,2,3,\*</sup>

December 29, 2022

**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: dkatz30@kentlaw.iit.edu

### Abstract

Many of the professions in the United States require a professional license exam, commonly referred to as "the Bar Exam," as a prerequisite for law practice. To pass all five of the states' last professional exams, one must spend at least a full year, many years of post-secondary education, including three years at an accredited law school. In addition, most law schools also require entry to a number of further, non-academic preparations. Despite this significant investment of time and capital, approximately one in five law school graduates will even make the state required to pass the exam on their first try. In the law 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 context, we document our experimental evaluation of the performance of OpenAI's TEXT-DAVINCI-003 model, also referred to as GPT-3.5, on the multistate multiple choice (MBC) section of the exam. While we find no benefit to 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 partially improved GPT-3.5's zero-shot performance. The best prompt and hyperparameter (GPT-3.5) achieved a baseline correct rate of 30.2% on a complex 100-item MBC practice exam, approximately in line with the 20% baseline passing rate, and performed at a similar rate for both Evidence and Torts. GPT-3.5's accuracy of response is also highly correlated with the number of top two and top three choices we report 71% and 88% of the time, respectively, indicating very strong non-random performance. While our ability to interpret these results is limited by limited statistical understanding of LLMs and the proprietary nature of GPT, we believe that these results strongly suggest that an LLM will pass the MBC component of the Bar Exam in the near future.

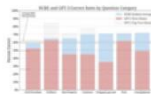


Fig. 1. Percent of questions answered correctly by GPT-3.5 and the baseline passing rate.

MBC vs. GPT Performance on the MBC

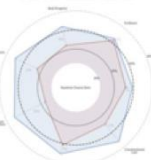


Fig. 3. Accuracy by Question Category for GPT and Average Test-Takers

	GPT-3.5	GPT-3.5	GPT-3.5	GPT-3.5
Evidence	45%	45%	45%	45%
Torts	45%	45%	45%	45%
Contract Law	45%	45%	45%	45%
Constitutional Law	45%	45%	45%	45%
Criminal Law & Procedure	45%	45%	45%	45%

Table 3. Summary of performance by question category for GPT-3.5 and the baseline passing rate.

## GPT takes the Bar Exam

Michael J Bommarito II<sup>1,2,3,\*</sup> and Daniel Martin Katz<sup>1,2,3,\*</sup>

<sup>1</sup>Illinois Tech - Chicago Kent College of Law, Chicago, IL, USA, <sup>2</sup>Bucerius Law School, Hamburg, DE and <sup>3</sup>CodeX - The Stanford Center for Legal Informatics, Palo Alto CA, USA

\*To whom correspondence should be addressed: dkatz30@kentlaw.iit.edu

### Abstract

We experimentally evaluate OpenAI's TEXT-DAVINCI-003 (GPT-3.5) and related models on the multistate multiple choice (MBC) section of the Bar Exam. Despite the fact that humans with seven years of post-secondary education and exam-specific training only answer 68% of questions correct, TEXT-DAVINCI-003 is able to achieve a correct rate of 50.3% for best prompt and parameters and achieved passing scores in the Evidence and Torts sections. In addition to exceeding the baseline passing rate of 25%, the model's top two and top three choices are correct 71% and 88% of the time, respectively, indicating strong non-random performance. We find no benefit to fine-tuning over GPT-3.5's zero-shot performance at the scale of our training data, we document our observations related to prompt engineering, hyperparameters, and older GPT models as well. While our ability to interpret these results is limited by nascent scientific understanding of LLMs and the proprietary nature of GPT models, the trend in improvement for recent GPT models strongly suggests that an LLM will pass the MBC component of the Bar Exam in the near future.

**Key words:** natural language processing (NLP), artificial intelligence (AI)

### Introduction

The legal system is becoming increasingly complex [1][2][3], leading to a need for technology to assist with the quantity, quality, and accessibility of legal services demanded by society. As is often discussed, artificial intelligence and process engineering have promised help for decades to both non-professional and professional users of legal systems [4][5][6]. Significant research and development effort has been devoted to use cases such as automated negotiation or brief construction, pre- and post-negotiation contract process automation, AI-assisted due diligence and e-discovery, the statistical prediction of judicial decision making, and technology-aided search and legal aid for laypeople [7][8][9][10]. However, the complexity of legal language and volume of legal knowledge has made it historically difficult to develop systems that understand the nuances of legal tasks, and many systems have failed to deliver desired results or reach adoption.

Naturally, law is heavily reliant on the use of language, producing massive volumes of textual data [11][12]. Documents such as briefs, motions, statutes, regulations, contracts, patents, and judicial decisions are continuously authored by lawyers, judges, and regulators [2]. To make matters even more difficult, legal language is notoriously complex; lawyers and other legal professionals undertake many years of education and professional training to understand and generate it.

Why is this language so "complex"? Why do so many predicted users of natural languages struggle with legal

documents such as contracts, statutes and regulations, even in their native tongue, to the point that descriptors like "legalese" or "lawyer speak" have become common parlance? The answer is likely twofold. First, for both technical and cultural reasons, the grammar of legal language is significantly different than the grammar of natural language, featuring both highly-stylized sentence and particularly precise phrasing. The resulting sentence structures are typically much longer and more complex than natural language; as the number of clauses and "clauses" (or "clauses") that clauses are connected exceeds the working memory of both human and non-human readers. Second, by the very nature of common law and precedent, legal language is full of semantic nuance and history. Words like "securities" that have common meaning in natural language often have different, context-specific meanings in legal language. Many words that do not occur at all in natural language, like "negotiate" or "indemnify," occur regularly in legal contexts. This semantic depth and breadth traditionally required systems that interact with legal text to embed a large amount of domain-specific knowledge. Viewed from this perspective, legal education and training is required to teach humans to understand and produce this very particular type of language, and it is no surprise that traditional models in NLP struggled in general legal task assessments.

In recent years, however, developments in natural language processing and computing have led to significant advances in state of the art performance. Leveraging advances in neural

V1.01 - December 29, 2022

V2.01 - January 3, 2023



[Research](#) ▾[Product](#) ▾[Safety](#)[Company](#) ▾

# GPT-4

March 14, 2023

## Simulated exams

	<b>GPT-4</b> estimated percentile	<b>GPT-4 (no vision)</b> estimated percentile	<b>GPT-3.5</b> estimated percentile
Uniform Bar Exam (MBE+MEE+MPT) <sup>1</sup>	<b>298/400</b> ~90th	<b>298/400</b> ~90th	<b>213/400</b> ~10th
LSAT	<b>163</b> ~88th	<b>161</b> ~83rd	<b>149</b> ~40th
SAT Evidence-Based Reading & Writing	<b>710/800</b> ~93rd	<b>710/800</b> ~93rd	<b>670/800</b> ~87th
SAT Math	<b>700/800</b> ~89th	<b>690/800</b> ~89th	<b>590/800</b> ~70th
Graduate Record Examination (GRE) Quantitative	<b>163/170</b> ~80th	<b>157/170</b> ~62nd	<b>147/170</b> ~25th
Graduate Record Examination (GRE) Verbal	<b>169/170</b> ~99th	<b>165/170</b> ~96th	<b>154/170</b> ~63rd
Graduate Record Examination (GRE) Writing	<b>4/6</b> ~54th	<b>4/6</b> ~54th	<b>4/6</b> ~54th
USABO Semifinal Exam 2020	<b>87/150</b> 99th~100th	<b>87/150</b> 99th~100th	<b>43/150</b> 31st~33rd
USNCO Local Section Exam 2022	<b>36/60</b>	<b>38/60</b>	<b>24/60</b>
Medical Knowledge Self-Assessment Program	<b>75%</b>	<b>75%</b>	<b>53%</b>
Codeforces Rating	<b>392</b> below 5th	<b>392</b> below 5th	<b>260</b> below 5th
AP Art History	<b>5</b> 86th~100th	<b>5</b> 86th~100th	<b>5</b> 86th~100th
AP Biology	<b>5</b> 85th~100th	<b>5</b> 85th~100th	<b>4</b> 62nd~85th
AP Calculus BC	<b>4</b> 43rd~59th	<b>4</b> 43rd~59th	<b>1</b> 0th~7th



UBE Component	Total UBE Points	Questions	Time	Time Per Question
Multistate Bar Exam (MBE)	200 Points	200 Questions (Multiple Choice)	6 Hours	1 min 48 sec
Multistate Essay Exam (MEE)	120 Points	6 Questions (3-4 Subquestions)	3 Hours	30 minutes
Multistate Performance Test (MPT)	80 Points	2 Questions (3-4 Subquestions)	3 Hours	90 minutes

**Table 1.** Summary of Uniform Bar Exam (UBE) Components

### Sample MBE Question

A plaintiff domiciled in State A brought a federal diversity negligence action in State A against a defendant domiciled in State B. The action was based on an accident that had occurred in State C. The defendant was personally served with process at her office in State B, which is located 50 miles from the State A federal courthouse. The defendant travels to State A once each year for a week long vacation but has no other State A contacts.

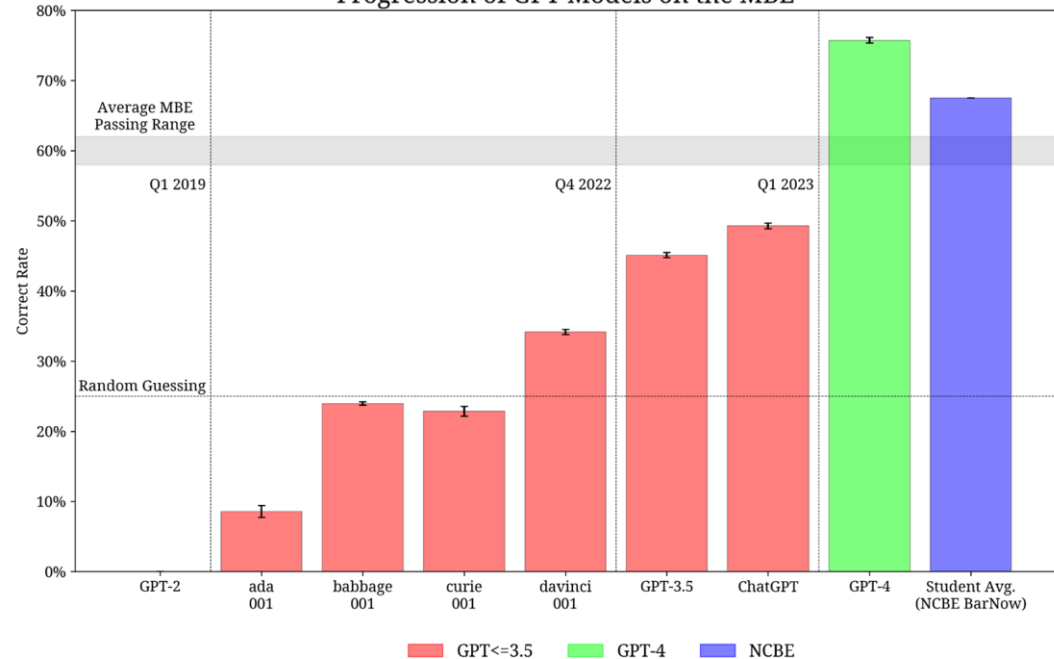
The defendant answered, denying all allegations. One week later, the defendant filed an amended answer, denying all allegations and including the defense of lack of personal jurisdiction. State A has a long-arm statute that permits personal jurisdiction to the constitutional limit.

The defendant has moved for an order dismissing the action based on the personal-jurisdiction challenge asserted in the amended answer.

Should the court issue the order?

- (A) No, because the defendant waived the challenge to personal jurisdiction by failing to include it in her original answer.
- (B) No, because the defendant was personally served with process within 100 miles of the federal courthouse where the action is pending.
- (C) Yes, because the defendant lacks minimum contacts with State A.
- (D) Yes, because service was not delivered to the defendant at her home.

### Progression of GPT Models on the MBE



## July 2022 MEE Evidence Question

Four months ago, Victim was shot and seriously wounded in City. Defendant has been charged with attempted murder. The prosecution's theory is that Victim and Defendant were both members of a criminal street gang called "The Lions," which engages in drug dealing, robbery, and murder in City. The prosecutor alleges that the shooting was the result of a gang dispute.

Defendant has brought a pretrial motion objecting to the prosecutor's introducing the following anticipated evidence:

(A) Testimony by a City detective who will be offered as an expert in gang identification, gang organizational structure, and gang activities generally and as an expert on particular gangs in City. The detective is expected to testify as follows:

I have been a detective on the police force for six years. Throughout that time, my primary assignment has been to investigate gangs and criminal activity in City. I have also worked closely with federal drug and firearm task forces as they relate to gangs. Prior to becoming a detective, I was a corrections officer in charge of the gang unit for City's jail for three years, and my duties included interviewing, investigating, and identifying gang members.

Throughout my career, I have attended training sessions providing education and information on gang structure, membership, and activities. As I've gained experience and knowledge in this area, I've frequently been asked to lead such sessions. I would estimate that I've taught more than 75 such training sessions over the past three years.

Street gangs generally engage in a wide variety of criminal activities. They usually have a clear leadership structure and strict codes of behavior. Absolute loyalty is required and is enforced through violent acts. Members of particular gangs can be identified by clothing, tattoos, language, paperwork, or associations.

I am quite familiar with "The Lions." It is one of City's most violent and feared criminal gangs. Members of The Lions can be identified by tattoos depicting symbols unique to the gang.

(B) Testimony by a former leader of The Lions concerning a photograph of Defendant's tattooed arm. After the photograph is authenticated as a photograph of Defendant's arm, the witness is expected to testify in part as follows:

I am certain that this is a Lions tattoo. I had a similar one removed. You'll notice that it has a shield containing the numbers for the police code for homicide, and Lions' members frequently include police codes in their tattoos to indicate crimes the gang has committed. The tattoo also has a shotgun and sword crossed as an "X," and a lion. Those are symbols frequently used by The Lions. This tattoo indicates to me, based on my experience, that Defendant is a member of The Lions gang.

(C) Testimony by Victim, who is expected to testify for the prosecution in part as follows:

I got into an argument with a gang boss at a meeting of The Lions. I said I wouldn't participate in an attack that was planned on another gang because my cousin was in that gang. The boss looked at Defendant and nodded to him. Next thing I knew, after the meeting, Defendant pulled a gun on me and shot me. I'm sure he did it because of that argument.

The jurisdiction has adopted rules of evidence identical to the Federal Rules of Evidence.

Defense counsel's motion raises the following objections to the evidence described above:

1. The detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony.
2. The photograph of Defendant's tattoo and the former gang leader's anticipated testimony about it is inadmissible character evidence.
3. Victim's anticipated testimony that Defendant shot him because of a gang dispute is irrelevant.

How should the trial court rule on each objection? Explain. (Do not address constitutional issues.)

## GPT-3.0 - July 2022 MEE Evidence Answer)

**1. How should the trial court rule on the defendant's motion that the detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony? Explain. (Do not address constitutional issues.)**

The trial court should rule on the defendant's motion that the detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony. The trial court should find that the detective's testimony is inadmissible because it is not based on sufficient facts or expertise. The trial court should also find that the testimony is unfairly prejudicial because it will likely result in the jury considering irrelevant evidence.

**2. How should the trial court rule on the defendant's motion that the photograph of Defendant's tattoo and the former gang leader's anticipated testimony about it is inadmissible character evidence? Explain. (Do not address constitutional issues.)**

The trial court should rule on the defendant's motion that the photograph of Defendant's tattoo and the former gang leader's anticipated testimony about it is inadmissible character evidence. The trial court should hold that the photograph is not relevant to the issues at trial and that the testimony is inadmissible character evidence.

The trial court should hold that the photograph is not relevant to the issues at trial because it is not evidence that Defendant committed the crime. The trial court should also hold that the testimony is inadmissible character evidence because it is not evidence that Defendant has a disposition to commit the crime. The trial court should note that the character evidence is not relevant to any issue at trial, including whether Defendant is guilty of the attempted murder charge.

The trial court should also hold that the character evidence is inadmissible because it is more prejudicial than probative. The trial court should note that the character evidence is not relevant to any issue at trial, including whether Defendant is guilty of the attempted murder charge. The trial court should find that the probative value of the character evidence is substantially outweighed by the danger of unfair prejudice, and the trial court should exclude the evidence.

**3. How should the trial court rule on the defendant's motion that the Victim's anticipated testimony that Defendant shot him because of a gang dispute is irrelevant? Explain. (Do not address constitutional issues.)**

The trial court should rule on the defendant's motion that Victim's anticipated testimony that Defendant shot him because of a gang dispute is irrelevant by determining whether the evidence is relevant to the crime with which Defendant has been charged. Under Rule 401 of the Federal Rules of Evidence, evidence is relevant if it has any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence. Here, the prosecutor intends to introduce evidence that Defendant shot Victim because of a gang dispute. This evidence is relevant to the crime with which Defendant has been charged. Therefore, the trial court should not rule on the defendant's motion that Victim's anticipated testimony is irrelevant.

July 2022 MEE Evidence Question

Four months ago, Victim was shot and seriously wounded in City. Defendant has been charged with attempted murder. The prosecution's theory is that Victim and Defendant were both members of a criminal street gang called "The Lions," which engages in drug dealing, robbery, and murder in City. The prosecutor alleges that the shooting was the result of a gang dispute.

Defendant has brought a pretrial motion objecting to the prosecutor's introducing the following anticipated evidence:

(A) Testimony by a City detective who will be offered as an expert in gang identification, gang organizational structure, and gang activities generally and as an expert on particular gangs in City. The detective is expected to testify as follows:

I have been a detective on the police force for six years. Throughout that time, my primary assignment has been to investigate gangs and criminal activity in City. I have also worked closely with federal drug and firearm task forces as they relate to gangs. Prior to becoming a detective, I was a corrections officer in charge of the gang unit for City's jail for three years, and my duties included inte

Throughout  
and activities  
estimate that

Street gangs  
codes of bet  
identified by

I am quite fa  
identified by

(B) Testimor  
authenticates

I am certain  
the police co  
has committe  
The Lions. I

(C) Testimor

I got into an  
another gang  
the meeting,

The jurisdic

Defense counsel's motion raises the following objections to the evidence described above:

- 1. The detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony.
- 2. The photograph of Defendant's tattoo and the former gang leader's anticipated testimony about it is inadmissible character evidence.
- 3. Victim's anticipated testimony that Defendant shot him because of a gang dispute is irrelevant.

How should the trial court rule on each objection? Explain. (Do not address constitutional issues.)

GPT-3.0 - July 2022 MEE Evidence Answer)

1. How should the trial court rule on the defendant's motion that the detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony ? Explain. (Do not address constitutional issues.)

The trial court should rule on the defendant's motion that the detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony. The trial court should find that the detective's testimony is inadmissible because it is not based on sufficient facts or expertise. The trial court should also find that the testimony is unfairly prejudicial because it will likely result in the jury considering irrelevant evidence

GPT-3.0 - July 2022 MEE Evidence Answer)

1. How should the trial court rule on the defendant's motion that the detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony ? Explain. (Do not address constitutional issues.)

The trial court should rule on the defendant's motion that the detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony. The trial court should find that the detective's testimony is inadmissible because it is not based on sufficient facts or expertise. The trial court should also find that the testimony is unfairly prejudicial because it will likely result in the jury considering irrelevant evidence.

because of a gang dispute is irrelevant by determining whether the evidence is relevant to the crime with which Defendant has been charged. Under Rule 401 of the Federal Rules of Evidence, evidence is relevant if it has any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence. Here, the prosecutor intends to introduce evidence that Defendant shot Victim because of a gang dispute. This evidence is relevant to the crime with which Defendant has been charged. Therefore, the trial court should not rule on the defendant's motion that Victim's anticipated testimony is irrelevant.

o and  
1. (Do

id the  
should  
dence.

re that  
racter  
should  
of the

l than  
luding  
due of  
should

y that  
sues.)

ot him



## GPT-4 - July 2022 MEE Evidence Answer (Part 1 of 3)

**1. How should the trial court rule on the defendant's motion that the detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony ? Explain. (Do not address constitutional issues.)**

The trial court should deny the defendant's motion and allow the detective's anticipated testimony about gang identification, organization, and activities as proper expert testimony. Under Federal Rule of Evidence 702, a witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if: (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.

In this case, the detective is qualified as an expert by his knowledge, experience, training, and education in the area of gang identification, organization, and activities. He has been a detective on the police force for six years, with a primary assignment of investigating gangs and criminal activity in City. He has also worked closely with federal drug and firearm task forces as they relate to gangs. Prior to becoming a detective, he was a corrections officer in charge of the gang unit for City's jail for three years, and his duties included interviewing, investigating, and identifying gang members. Additionally, he has attended training sessions providing education and information on gang structure, membership, and activities, and has frequently been asked to lead such sessions.

The detective's testimony will help the trier of fact to understand the evidence and to determine a fact in issue, namely, whether Defendant is a member of The Lions gang and whether the shooting was the result of a gang dispute. The detective's testimony is based on sufficient facts or data, as he is familiar with The Lions and has knowledge of their symbols, codes, and activities. The testimony is also the product of reliable principles and methods, as the detective has gained his knowledge and experience through his work as a detective and corrections officer, as well as through his attendance and leadership of training sessions. Finally, the detective has reliably applied the principles and methods to the facts of the case, as he is able to identify the symbols and codes used by The Lions and can explain their significance.

Therefore, the trial court should deny the defendant's motion and allow the detective's anticipated testimony about gang identification, organization, and activities as proper expert testimony.

## ChatGPT - July 2022 MEE Evidence Answer (Part 1 of 3)

**1. How should the trial court rule on the defendant's motion that the detective's anticipated testimony about gang identification, organization, and activities is improper expert testimony ? Explain. (Do not address constitutional issues.)**

The trial court should deny the defendant's motion to exclude the detective's anticipated testimony about gang identification, organization, and activities as improper expert testimony.

Under Federal Rule of Evidence 702, expert testimony is admissible if it is (1) based on sufficient facts or data, (2) the product of reliable principles and methods, and (3) the expert has applied the principles and methods reliably to the facts of the case. The trial court has broad discretion in determining the admissibility of expert testimony, and its decision will only be overturned if it is an abuse of discretion.

In this case, the detective's anticipated testimony satisfies all three prongs of the Rule 702 test. First, the detective's testimony is based on his extensive experience investigating gangs and criminal activity in City, as well as his training and education on gang structure, membership, and activities. He has also interviewed, investigated, and identified gang members in his prior role as a corrections officer. Therefore, the detective has sufficient facts and data to support his testimony.

Second, the detective's testimony is based on reliable principles and methods. He has attended numerous training sessions on gang structure, membership, and activities, and has frequently been asked to lead such sessions. His experience and knowledge in this area make him qualified to testify as an expert on gang identification, organization, and activities. Additionally, his testimony is consistent with the general understanding of gangs and their activities.

Finally, the detective has reliably applied his principles and methods to the facts of the case. He has identified Defendant as a member of The Lions based on his tattoo, which contains symbols unique to the gang. He has also provided general information about gang behavior and activities that is relevant to the prosecution's theory of the case. Therefore, the detective's anticipated testimony is admissible under Rule 702.

The defendant may argue that the detective's testimony is prejudicial, and that its probative value is outweighed by the risk of unfair prejudice. Under Federal Rule of Evidence 403, relevant evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice. However, the detective's testimony is highly probative of Defendant's alleged gang membership and the prosecution's theory of the case. Any potential prejudice can be mitigated by the trial court through limiting instructions to the jury.

In conclusion, the trial court should deny the defendant's motion to exclude the detective's anticipated testimony about gang identification, organization, and activities as improper expert testimony. The detective's testimony satisfies the requirements of Rule 702 and is highly probative of Defendant's alleged gang membership and the prosecution's theory of the case.



## July 2022 MPT-1 Instructional Memo

To: Examinee  
From: Marianne Morton  
Date: July 26, 2022  
Re: Walter Hixon matter

We represent Walter Hixon in connection with complications of his marital status. Mr. Hixon married Joan Prescott in 1986 in the State of Columbia. Several years later they separated. Mr. Hixon believed that Ms. Prescott died in 2001.

In 2012, he married Frances Tucker in the State of Columbia. They purchased a house together in Columbia early in the marriage. A few years ago, Mr. Hixon moved to Franklin for a job opportunity; Ms. Tucker remained in Columbia.

Last month, Mr. Hixon learned that Joan Prescott is still alive. He has informed Ms. Tucker of that fact. He wants to divorce Ms. Prescott, end his purported marriage with Ms. Tucker, and work out shares in the residential property that he and Ms. Tucker own.

I need you to write a memorandum to me addressing the following questions:

1. Does Columbia or Franklin law govern the grounds for annulling Mr. Hixon's marriage to Ms. Tucker?
2. Must Mr. Hixon file a lawsuit to annul his second marriage, and if yes, would he be able to obtain an annulment under the applicable law?
3. If Mr. Hixon files an annulment action in Franklin, would a Franklin court have jurisdiction to annul the marriage and to dispose of the parties' property?
4. Should we advise Mr. Hixon to file in Columbia or in Franklin?

Do not prepare a separate statement of facts, but be sure to incorporate the relevant facts into your analysis and state the reasons for your conclusions and recommendation. Do not address either Mr. Hixon's ending his marriage to Ms. Prescott or the risks of criminal prosecution he may face for bigamy; another associate will research those issues.

## Multistate Performance Exam (MPT)

10-15 Pages of Materials

THE FILE = The Facts  
THE LIBRARY = The Law

~5000 token inputs

MEMORANDUM  
To: Marianne Morton  
From: Examinee  
Date: July 26, 2022  
Re: Walter Hixon matter

**1. Does Columbia or Franklin law govern the grounds for annulling Mr. Hixon's marriage to Ms. Tucker?**

Analysis:

The first step in determining which state's law governs the grounds for annulling Mr. Hixon's marriage to Ms. Tucker is to determine which state has the most significant relationship to the spouses and the marriage. Under the Restatement (Second) of Conflict of Laws § 283, the validity of a marriage will be determined by the local law of the state which, with respect to the particular issue, has the most significant relationship to the spouses and the marriage under the principles stated in § 6.

In this case, both Mr. Hixon and Ms. Tucker were residents of Columbia at the time of their marriage, and they purchased a house together in Columbia early in the marriage. These facts suggest that Columbia has a significant relationship to the spouses and the marriage. However, Mr. Hixon moved to Franklin a few years ago for a job opportunity, while Ms. Tucker remained in Columbia. This fact suggests that Franklin may also have a significant relationship to the spouses and the marriage.

To determine which state has the most significant relationship, we must consider the factors stated in Restatement § 6. These factors include the relevant policies of other interested states and the relative interests of those states in the determination of the particular issue, the protection of justified expectations, certainty, predictability, and uniformity of result, and ease in the determination and application of the law to be applied.

With respect to the relevant policies of other interested states and the relative interests of those states in the determination of the particular issue, both Columbia and Franklin have legitimate policy interests in defining how a relationship as fundamental as marriage can be initiated and ended. However, the fact that Columbia and Franklin recognize different reasons for annulling a marriage indicates the strength of the policy interests involved. In this case, Columbia law permits annulment on the ground of fraudulent inducement, while Franklin law does not. This difference in the law suggests that Columbia has a stronger policy interest in the determination of the particular issue.

With respect to the protection of justified expectations, Mr. Hixon and Ms. Tucker married in Columbia, lived the majority of their married life there, and owned property together there. These facts strongly suggest that the parties had a justified expectation that Columbia law would govern the terms on which the marriage ended.

With respect to certainty, predictability, and uniformity of result, people often move between states, creating the need for a system of well-defined rules to govern which state's laws apply to the creation and termination of marriages. In this case, the fact that Mr. Hixon and Ms. Tucker were residents of Columbia at the time of their marriage and purchased a house together in Columbia early in the marriage suggests that Columbia law should apply to the creation and termination of their marriage.

With respect to ease in the determination and application of the law to be applied, all the important events in this marriage occurred in Columbia. Considerations of ease and administrative efficiency strongly suggest Columbia as the appropriate forum.

Conclusion:

Based on the factors stated in Restatement § 6, it appears that Columbia law should govern the grounds for annulling Mr. Hixon's marriage to Ms. Tucker. Columbia has the most significant relationship to the spouses and the marriage, and the relevant policies of Columbia, the protection of justified expectations, certainty, predictability, and uniformity of result, and ease in the determination and application of the law to be applied all support the application of Columbia law. Therefore, I recommend that we advise Mr. Hixon that Columbia law will likely govern the grounds for annulling his marriage to Ms. Tucker.

MEMORANDUM  
To: Marianne Morton  
From: Examinee  
Date: July 26, 2022  
Re: Walter Hixon matter

**1. Does Columbia or Franklin law govern the grounds for annulling Mr. Hixon's marriage to Ms. Tucker?**

I. Brief Answer

Franklin law governs the grounds for annulling Walter Hixon's marriage to Frances Tucker.

II. Discussion

Under the Restatement (Second) of Conflict of Laws § 6, a court should apply the law of the state with the most significant relationship to the spouses and the marriage under the principles stated in § 6. If a state has no such relationship, that state must apply the law of the state that does. The factors relevant to the choice of the applicable rule of law include the relevant policies of other interested states and the relative interest s of those states in the determination of the particular issue, the protection of justified expectations, certainty, predictability, and uniformity of result, and ease in the determination and application of the law to be applied.

In this case, the parties were married in Columbia, but Mr. Hixon currently resides in Franklin, and the couple purchased a house in Columbia. However, the fact that Mr. Hixon currently resides in Franklin is not determinative of the choice of law issue. Instead, the relevant factors to consider include the policies of the interested states, the relative interests of those states, the protection of justified expectations, and ease of determination and application of the law.

The policies of the interested states are relevant, and the fact that the grounds for annulment differ between Columbia and Franklin is indicative of the strength of the policy interests involved. Columbia law permits an annulment on the ground that the spouse of either party was living and the marriage with that spouse was then in force and that spouse was absent and not known to the party commencing the proceeding to be living for a period of five successive years immediately preceding the subsequent marriage for which the annulment decree is sought. By contrast, Franklin law provides that certain marriages shall be void, including all marriages between parties where either party is lawfully married to another person. While both states have a policy interest in regulating the institution of marriage, the fact that the grounds for annulment differ between the two states suggests that the policy interests of each state are strong.

The relative interests of the states are also relevant. In this case, the parties were married in Columbia, but they purchased a house in Columbia and lived there together. Mr. Hixon moved to Franklin for a job opportunity, but Ms. Tucker remained in Columbia. The fact that the couple purchased a house in Columbia and lived there together suggests that Columbia has a significant relationship to the marriage.

However, the protection of justified expectations also weighs in favor of applying Franklin law. Mr. Hixon currently resides in Franklin, and the couple purchased a house in Columbia together. While the fact that the couple purchased a house in Columbia suggests that they had a justified expectation that Columbia law would apply, the fact that Mr. Hixon currently resides in Franklin suggests that he had a justified expectation that Franklin law would apply.

Finally, ease of determination and application of the law also weighs in favor of applying Franklin law. The parties' marriage was entered into in Columbia, but the parties currently reside in different states, and the property at issue is located in Columbia. Applying Franklin law would provide a clear and predictable rule for determining the grounds for annulment.

III. Conclusion

Given the relevant factors, Franklin law governs the grounds for annulling Walter Hixon's marriage to Frances Tucker.

UBE Component	GPT-4	ChatGPT
Multistate Bar Exam (MBE)	157 Points	116 Points
Multistate Essay Exam (MEE)	84 Points	60 Points
Multistate Performance Test (MPT)	56 Points	37 Points
Overall Score	297 Points	213 Points

**Table 7.** Summary of Overall Performance on Uniform Bar Exam (UBE)

State	UBE Threshold
Alabama, Minnesota, Missouri, New Mexico, North Dakota	260
Indiana, Oklahoma	264
Connecticut, District of Columbia, Illinois, Iowa, Kansas, Kentucky Maryland, Montana, New Jersey, New York, South Carolina, Virgin Islands	266
Michigan	268
Alaska, Arkansas, Colorado, Maine, Massachusetts, Nebraska New Hampshire, North Carolina, Ohio, Oregon, Rhode Island, Tennessee, Texas, Utah, Vermont, Washington, West Virginia, Wyoming	270
Idaho, Pennsylvania	272
Arizona	273

**Table 8.** Uniform Bar Exam (UBE) Minimum Passing Score by State.

# Benchmarks



**The term benchmark, originates from the chiseled horizontal marks that surveyors made in stone structures, into which an angle-iron could be placed to form a "bench" for a leveling rod, thus ensuring that a leveling rod could be accurately repositioned in the same place in the future.**



# Benchmarks: LexGLUE and derivatives



## LexGLUE: A Benchmark Dataset for Legal Language Understanding in English

Ilias Chalkidis<sup>a\*</sup> Abhik Jana<sup>β</sup> Dirk Hartung<sup>γ δ</sup> Michael Bommarito<sup>γ δ</sup>  
 Ion Androutsopoulos<sup>ε</sup> Daniel Martin Katz<sup>γ δ ζ</sup> Nikolaos Aletras<sup>η</sup>

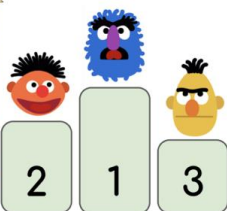
<sup>a</sup> University of Copenhagen, Denmark <sup>β</sup> Universität Hamburg, Germany

<sup>γ</sup> Bucerius Law School, Hamburg, Germany <sup>δ</sup> CodeX, Stanford Law School, United States

<sup>ε</sup> Athens University of Economics and Business, Greece <sup>η</sup> University of Sheffield, UK

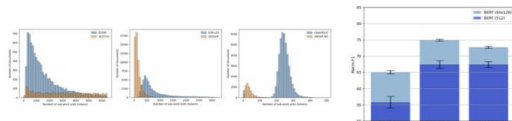
<sup>ζ</sup> Illinois Tech – Chicago Kent College of Law, United States

### THE LEGAL NLP BENCHMARK



Method	ECtHR (A*)	ECtHR (B*)	SCOTUS*	EUR-LEX	LEDGAR	UNFAIR-ToS	CaseHOLD
TFIDF-SVM	64.5	51.7	74.6	65.1	78.2	69.5	71.3
BERT	71.2	63.6	79.7	73.4	68.3	56.3	71.4
RoBERTa	69.2	59.0	77.3	68.9	71.5	62.0	71.9
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	72.1
Longformer	69.9	64.7	79.4	71.7	72.9	64.0	71.6
BigBird	70.0	62.9	78.9	70.9	72.5	62.0	71.5
Legal-BERT	70.0	64.0	80.4	74.7	76.4	66.5	72.1
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7

Table 1: Test results for all examined models across LexGLUE tasks. In standard datasets, we use the hierarchical variant of each model, except for Longformer and BigBird, discussed in Section 4.2.



[buceri.us/lexglue](https://buceri.us/lexglue)

# Results per Data Set

Method	ECtHR (A)*		ECtHR (B)*		SCOTUS*		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub> / m-F <sub>1</sub>
TFIDF+SVM	64.5	51.7	74.6	65.1	<b>78.2</b>	<b>69.5</b>	71.3	51.4	87.2	82.4	<b>95.4</b>	78.8	n/a
BERT	<b>71.2</b>	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	<b>57.9</b>	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	<b>72.1</b>	57.4	88.2	<b>83.1</b>	95.5	80.3	72.6
Longformer	69.9	<b>64.7</b>	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
BigBird	70.0	62.9	78.8	70.9	72.8	62.0	71.5	56.8	87.8	82.6	95.7	81.3	70.8
Legal-BERT	70.0	64.0	<b>80.4</b>	<b>74.7</b>	76.4	66.5	<b>72.1</b>	57.4	88.2	83.0	<b>96.0</b>	<b>83.0</b>	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7	56.6	<b>88.3</b>	83.0	<b>96.0</b>	82.3	<b>75.4</b>

Table 3: Test results for all examined models across LexGLUE tasks. In starred datasets, we use the hierarchical variant of each model, except for Longformer and BigBird, discussed in Section 4.2.

# Overall Aggregated Scores

Method	A-Mean		H-Mean		G-Mean	
	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>	$\mu$ -F <sub>1</sub>	m-F <sub>1</sub>
BERT	77.8	69.5	76.7	68.2	77.2	68.8
RoBERTa	77.8	68.7	76.8	67.5	77.3	68.1
DeBERTa	78.3	69.7	77.4	68.5	77.8	69.1
Longformer	78.5	70.5	77.5	69.5	78.0	70.0
BigBird	78.2	69.6	77.2	68.5	77.7	69.0
Legal-BERT	<b>79.8</b>	<b>72.0</b>	<b>78.9</b>	<b>70.8</b>	<b>79.3</b>	<b>71.4</b>
CaseLaw-BERT	79.4	70.9	78.5	69.7	78.9	70.3

Table 4: Test scores aggregated over tasks: arithmetic (A), harmonic (H), and geometric (G) mean.



# ChatGPT may Pass the Bar Exam soon, but has a Long Way to Go for the LexGLUE benchmark

Ilias Chalkidis

Department of Computer Science, University of Copenhagen  
ilias.chalkidis[at]di.ku.dk

Domain  
Specificity  
Scalability  
Cost  
Alignment  
Other Languages

## Abstract

Following the hype around OpenAI's ChatGPT conversational agent, the last straw in the recent development of Large Language Models (LLMs) that demonstrate emergent unprecedented zero-shot capabilities, we audit the latest OpenAI's GPT-3.5 model, 'gpt-3.5-turbo', the first available ChatGPT model, in the LexGLUE benchmark in a zero-shot fashion providing examples in a templated instruction-following format. The results indicate that ChatGPT achieves an average micro-F1 score of 49.0% across LexGLUE tasks, surpassing the baseline guessing rates. Notably, the model performs exceptionally well in some datasets, achieving micro-F1 scores of 62.8% and 70.1% in the ECtHR B and LEDGAR datasets, respectively. The code base and model predictions are available for review on [https://github.com/coastalcph/zeroshot\\_lexglue](https://github.com/coastalcph/zeroshot_lexglue).

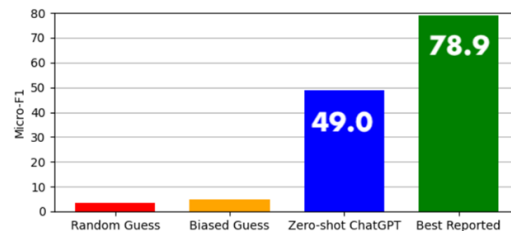
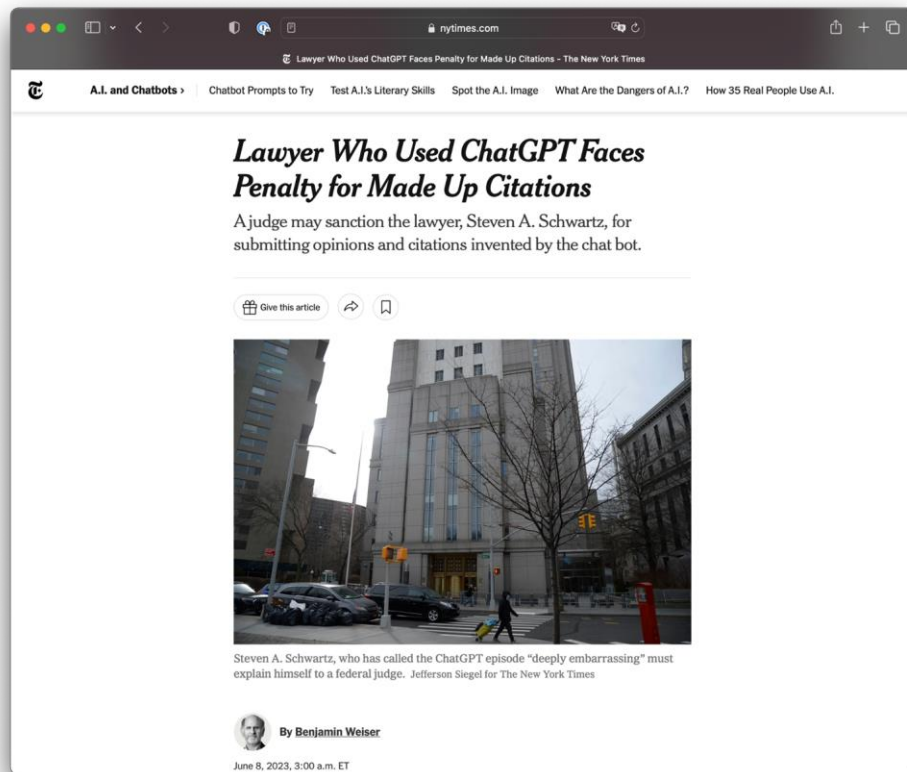


Figure 1: Averaged performance on LexGLUE.

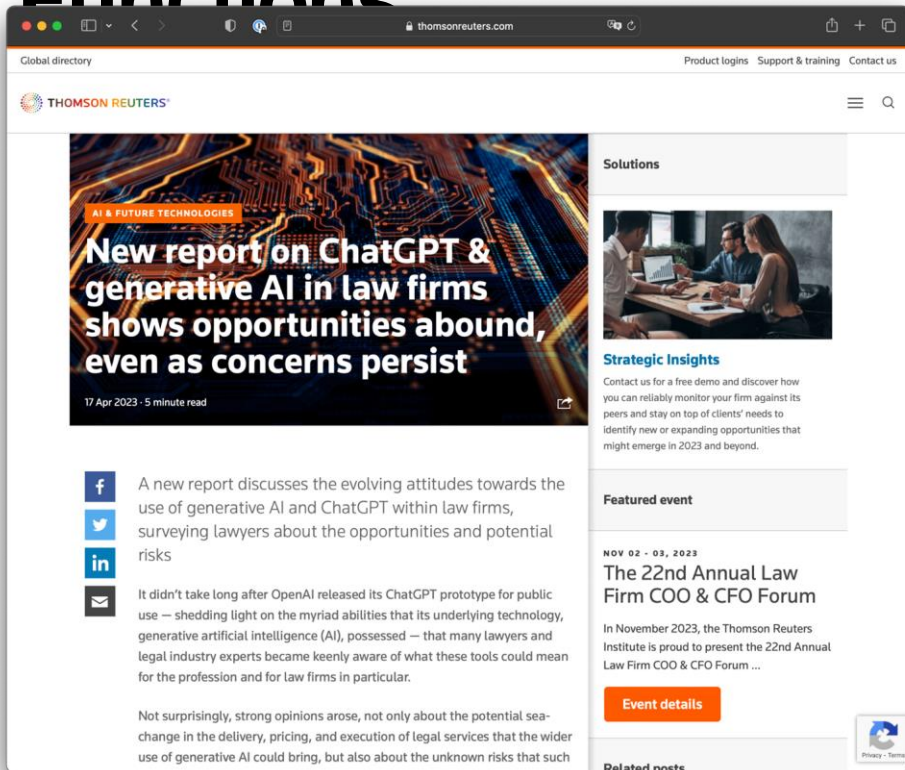
capabilities that cover common sense knowledge, but also extend to specialized domains such as problem solving, programming/debugging, and law, as presented by many users in the web.

Recently, [Bommarito and Katz \(2022\)](#) audited several variants of OpenAI's GPT 2/3/3.5 models in legal bar exam questions, and found that the most advanced -at the time- model ('text-davinci-003') achieves an accuracy of 50.3% on a complete practice exam, significantly in excess of the 25% baseline guessing rate, while

# Don't be fooled



# Law Firms: Research and Business Functions



The screenshot shows the Thomson Reuters website. The main headline is "New report on ChatGPT & generative AI in law firms shows opportunities abound, even as concerns persist" dated 17 Apr 2023. The article text discusses the evolving attitudes towards generative AI and ChatGPT within law firms, mentioning that many lawyers and legal industry experts became keenly aware of what these tools could mean for the profession and for law firms in particular. It also notes strong opinions arose about the potential sea-change in the delivery, pricing, and execution of legal services, as well as the unknown risks that such

**Solutions**

**Strategic Insights**

Contact us for a free demo and discover how you can reliably monitor your firm against its peers and stay on top of clients' needs to identify new or expanding opportunities that might emerge in 2023 and beyond.

**Featured event**

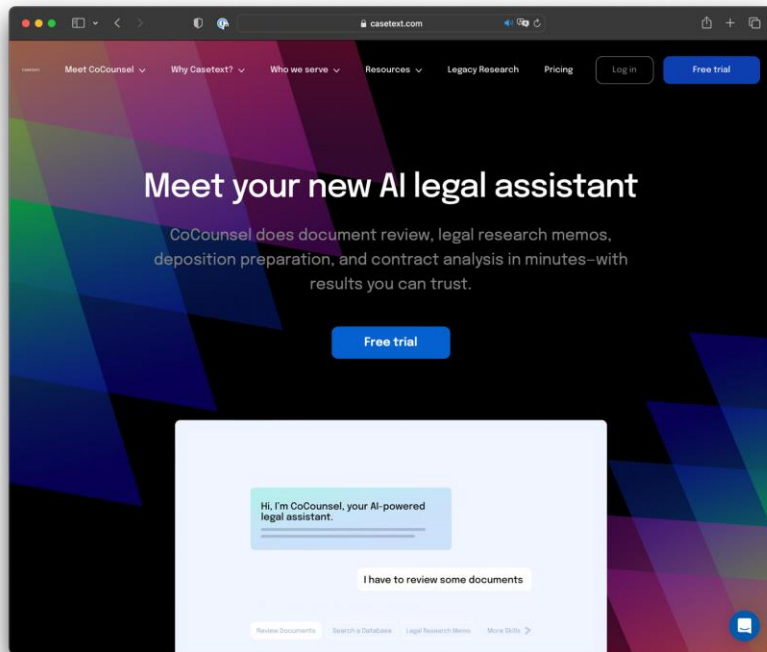
NOV 02 - 03, 2023

**The 22nd Annual Law Firm COO & CFO Forum**

In November 2023, the Thomson Reuters Institute is proud to present the 22nd Annual Law Firm COO & CFO Forum ...

**Event details**

**Related posts**



The screenshot shows the CoCounsel website. The main headline is "Meet your new AI legal assistant". The text below states: "CoCounsel does document review, legal research memos, deposition preparation, and contract analysis in minutes—with results you can trust." There is a "Free trial" button. Below this is a chat interface with a message from the AI assistant: "Hi, I'm CoCounsel, your AI-powered legal assistant." and a user input field with the text "I have to review some documents".

**Meet your new AI legal assistant**

CoCounsel does document review, legal research memos, deposition preparation, and contract analysis in minutes—with results you can trust.

**Free trial**

Hi, I'm CoCounsel, your AI-powered legal assistant.

I have to review some documents

Review Documents Search a Database Legal Research Memos More Tools >

# ESG Rating and Transparency



# Ironclad

ironcladapp.com

Ironclad Digital Contracting Product Customers Learn Careers About

Sign In Request Demo

WATCH State of Digital Contracting

## Business contracts for the digital age

See how the #1 CLM for innovators can help your company transform its contracts from blockers to enablers—making business faster, more collaborative, and more controlled.

Request Demo →

Counterparty replied with a new version  
DOCX REDLINES

Update to Governing Law detected

DIAMOND CORP'S  
MSA DOCX V3

GOVERNING LAW

STANDARD

APPROVERS

SIGNATURES