

Created by **Midjourney**, Prompt by A. Tagarelli, 2023



21 Febbraio, 2023

AI for Law: sfide e  
opportunità nell'era  
dell'intelligenza  
artificiale

In presenza @Talent Garden Roma



# AI4Law: Migliorare la Comprensione dei Codici di Legge attraverso Modelli di Linguaggio Neurali Pre-addestrati

**Andrea Tagarelli**  
Università della Calabria



UNIVERSITÀ DELLA CALABRIA

DIPARTIMENTO DI  
INGEGNERIA INFORMATICA,  
MODELLISTICA, ELETTRONICA  
E SISTEMISTICA

DIMES



# AI per la Legge: Obiettivi

Created by Midjourney, Prompt by A. Tagarelli, 2023

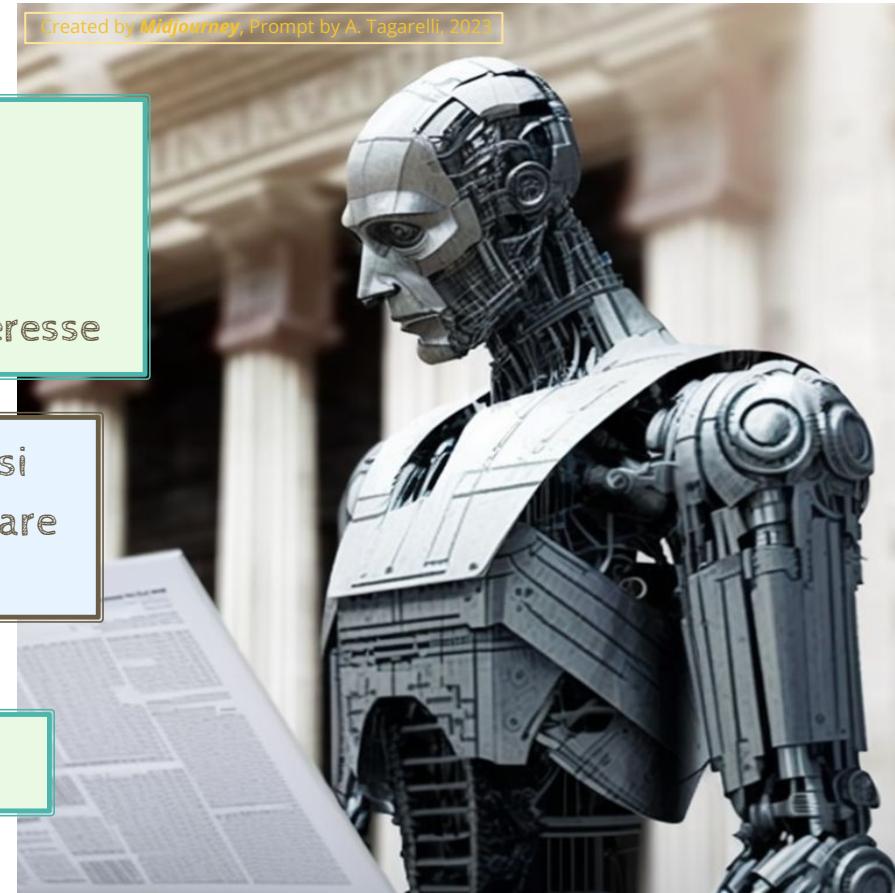
Supporto nella valorizzazione di enormi quantità di documenti legali

- automazione nella ricerca di informazioni
- comprensione (analisi/sintesi) di testi di interesse



Semplificazione di alcuni processi risparmiando tempo per affrontare problemi più complessi

Democratizzazione dell'accesso al diritto



# AI per la Legge: Problemi e Task computazionali

## Legal Search

- Information Retrieval/Ranking (giurisprudenza, codici di legge)
- Textual Entailment
- Question Answering

## Legal Document Review

- Passage retrieval, NER, Similarity, Sentiment Analysis, Summarization, ...

## Legal Outcome Prediction

- “Law as a prediction” (Oliver Wendell Holmes, 1897)
- Legal judgements, precedents, alleged violations, ...

# AI per la Legge: NLP breakthrough

**Modello di linguaggio:** Un modello addestrato su un'enorme quantità di testi, al fine di apprendere la distribuzione delle parole nella lingua

- Oltre 60 anni di ricerca: dall'ipotesi distribuzionale ai modelli stocastici di scoperta dei topic

**Modelli di linguaggio neurali**

- **No feature engineering**
- Rappresentano la semantica del linguaggio e relazioni non lineari tra le parole

**Transformer (Vaswani et al., 2017)**

- Catturano schemi lessicali complessi (struttura sequenziale, dipendenze a lungo termine, etc.) ottenendo una più completa rappresentazione di una sequenza testuale
- Incorporano il meccanismo di **attenzione**: assegna un peso maggiore alle parole in base alla loro maggiore informatività o rilevanza per il task

# AI per la Legge: NLP breakthrough

## Transformer sono modelli self-supervised

- Non richiedono dati etichettati da esperti

## Transformer sono modelli pre-addestrati

- Prima addestrati su grandi corpora,
  - es., prevedere le parole che seguono una frase, o
  - indovinare le parole mascherate dal loro contesto circostante
- Successivamente “riutilizzati” per un particolare task mediante un “addestramento di adattamento” al dominio del task (**fine-tuning**)

## Transformer sono few-shot learner

- Meta-learning che consente a un modello pre-addestrato di generalizzare su nuove classi di dati,
  - ma imparando sulla base di pochi esempi per ognuna delle nuove classi!



*"DO YOU SPEAK  
ENGLISH?"*



*"I DO NOW."*

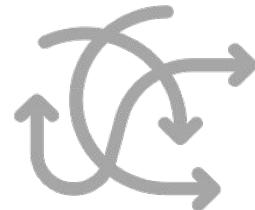
# Law Article Retrieval

Un problema di predizione: *data una richiesta in linguaggio naturale, predire l'articolo o gli articoli più rilevanti*

Legge Italiana



es., Codice Civile



Elevato numero di  
classi  
(centinaia/migliaia)



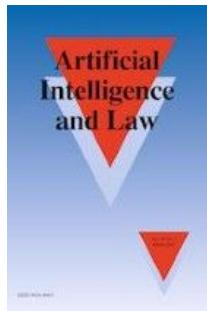
Few-shot learning  
problem



No test query  
benchmarks

# Machine Learning, NLP and Network Science Team

## @Artificial Intelligence and Data Science Lab - DIMES, Università della Calabria

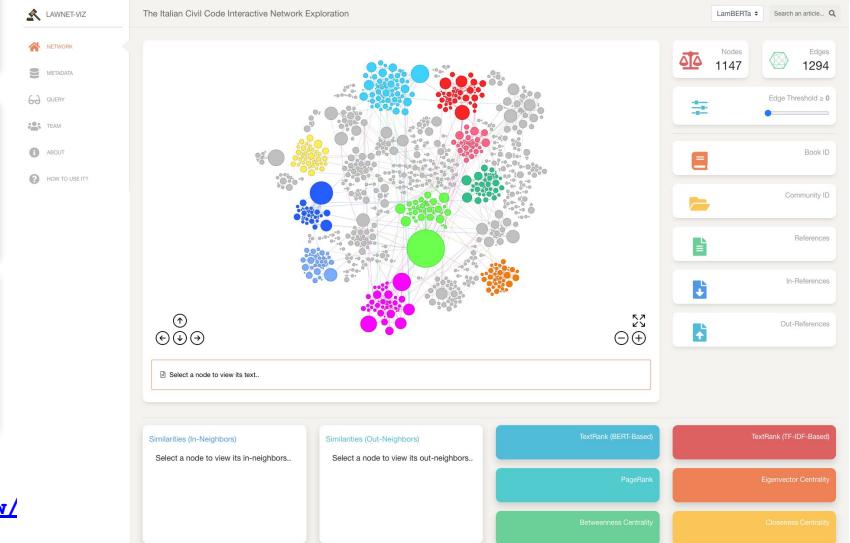


A. Tagarelli, A. Simeri (2022)  
**Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code**  
*Artificial Intelligence and Law*, 30(3):417-473, 2022  
(Online Published: Sep 2021). Springer Nature.  
DOI: 10.1007/s10506-021-09301-8



A. Simeri, A. Tagarelli (2023)  
**Exploring domain and task adaptation of LamBERTa models for article retrieval on the Italian Civil Code**  
*Proc. The 19th Conference on Information and Research science Connecting to Digital and Library science*. Bari, Italy, February 23-24, 2023.

A. Tagarelli, A. Simeri (2022)  
**LamBERTa: Law Article Mining Based on Bert Architecture for the Italian Civil Code**  
*Proc. The 18th Conference on Information and Research science Connecting to Digital and Library science*. Padova, Italy, February 24-25, 2022.



# LamBERTa - Law article mining based on BERT architecture

Fine-tuned Italian BERT per task di sequence classification

Addestramento su singoli libri o intero Codice Civile

Predizione a livello di articolo

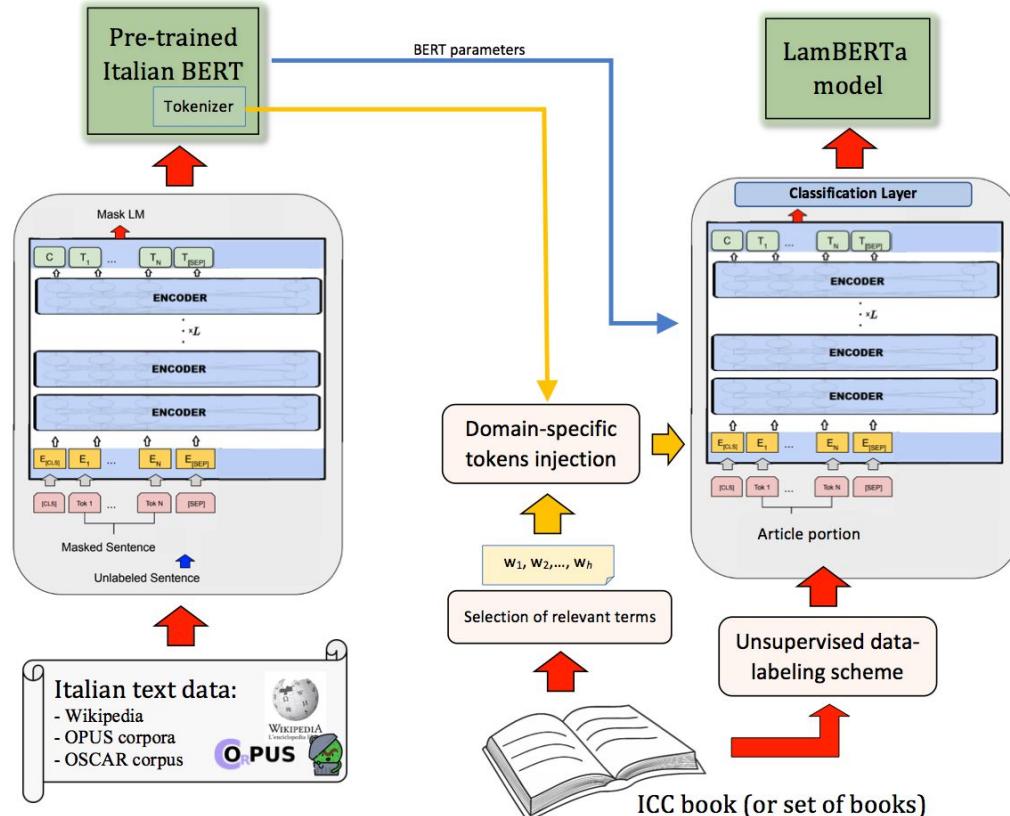
2 elementi chiave:

- Come creare (molte) istanze di addestramento per ogni classe (i.e., articolo)?

➡ UNSUPERVISED ARTICLE LABELING

- Come adattare il modello al lessico legale?

➡ DOMAIN-SPECIFIC TOKEN INJECTION



# LamBERTa - Law article mining based on BERT architecture

## UNSUPERVISED ARTICLE LABELING

N-way K-shot learning (N: #articoli, K: >1!)

Scelto *minK*, metodo round-robin che itera su repliche dello stesso gruppo di training units per articolo

🏆 Unigram with parameterized emphasis on the title

Bilanciamento rispetto alle diverse lunghezze degli articoli

Art. 586.

(Acquisto dei beni da parte dello Stato).

In mancanza di altri successibili, l'eredita' e' devoluta allo Stato. L'acquisto si opera di diritto senza bisogno di accettazione e non puo' farsi luogo a rinuncia.

Lo Stato non risponde dei debiti ereditari e dei legati oltre il valore dei beni acquistati.

## DOMAIN-SPECIFIC TOKEN INJECTION

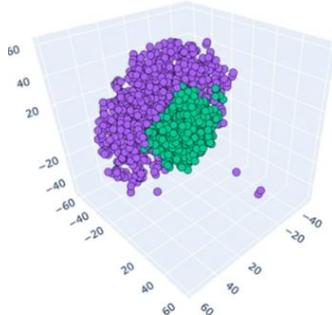
Arricchimento del tokenizer con termini nel Codice Civile che sono out-of-vocabulary per Italian BERT

- Selezione parole maggiormente informative (via stopword removal, cut-off frequency thresholds, TF-IDF feature weighting)
- Incremento: da 2.3% (Libro 2) a 6.6% (Libro 6)

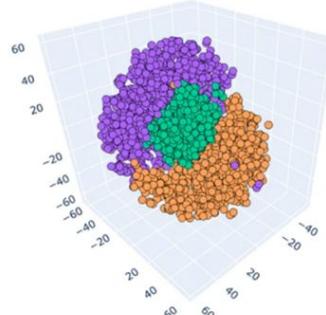


Il metodo di inizializzazione degli embedding dei nuovi token aggiunti ha impatto sulle performance

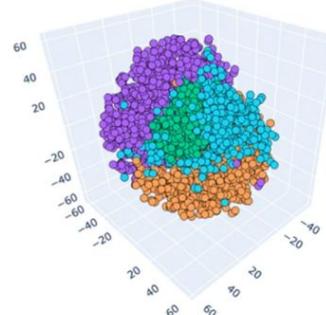
# LamBERTa - Law article mining based on BERT architecture



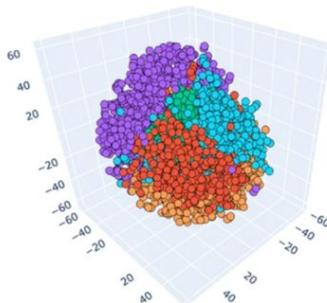
(a) Books 3 and 4



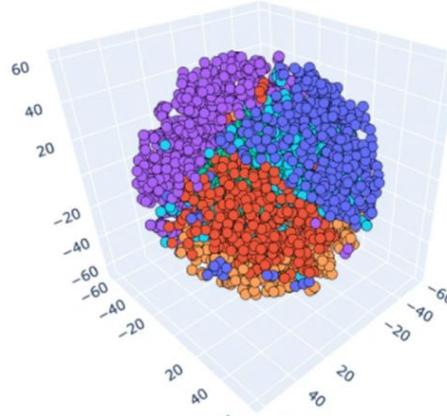
(b) Books 3-5



(c) Books 3-6



(d) Books 2-6



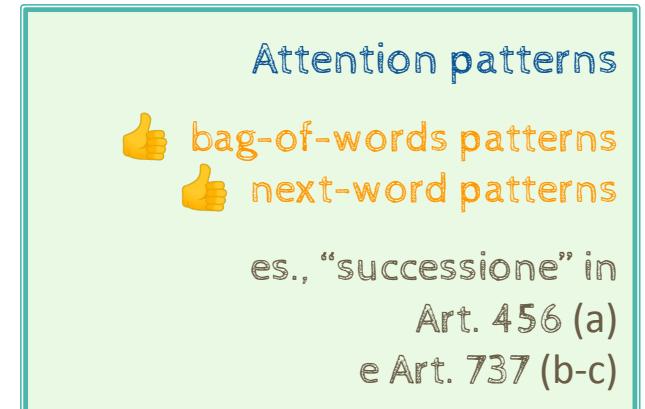
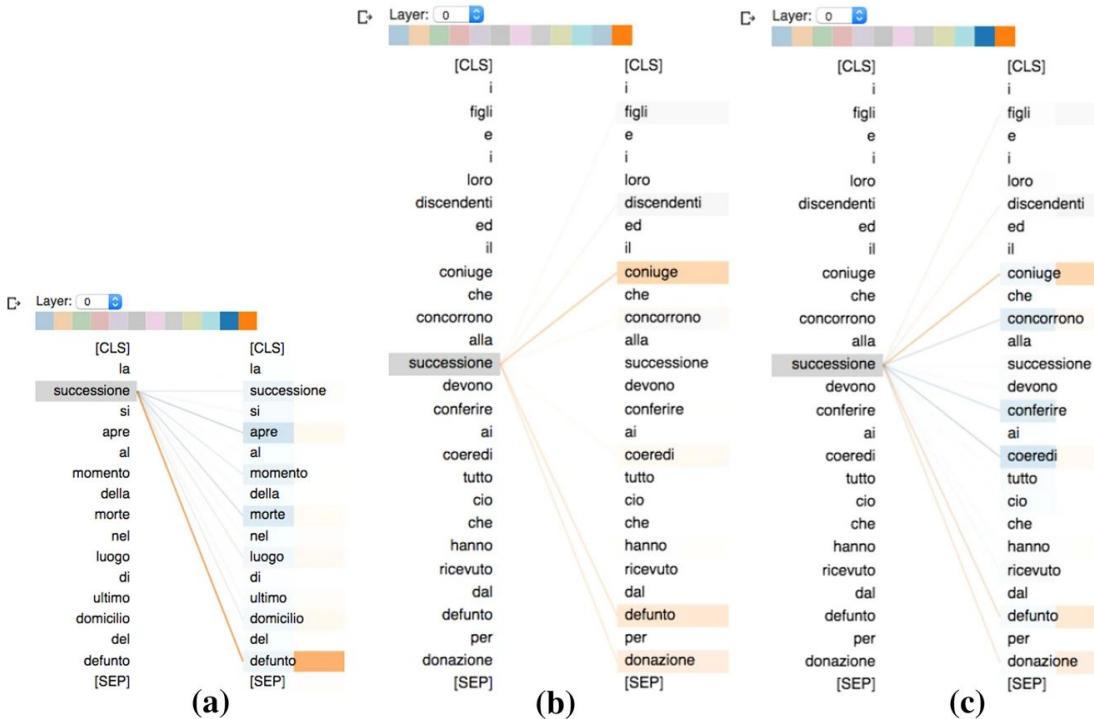
(e) All ICC books

Visualizzazione t-SNE 3D  
degli embeddings  
degli articoli del Codice Civile

Articoli di diversi libri sono  
distribuiti in diversi sottospazi

Coerente con l'organizzazione  
logica del Codice Civile

# LamBERTa - Law article mining based on BERT architecture



# LamBERTa - Law article mining based on BERT architecture

meno  
difficile

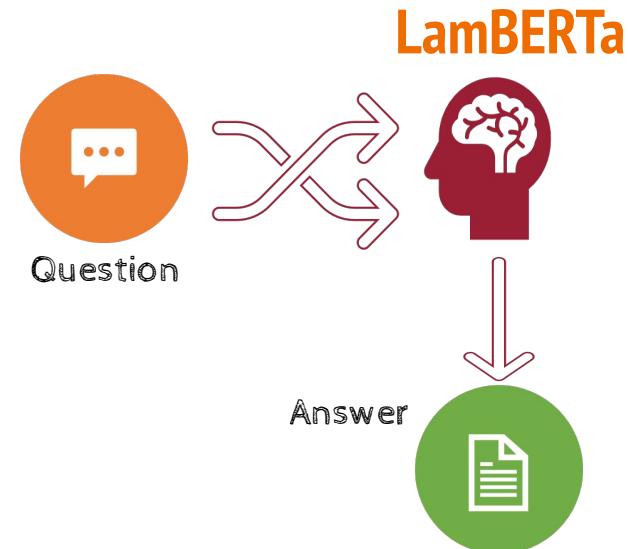
Frasi selezionate casualmente dagli articoli del libro *B*

Parafrasi (traduzione IT-EN-IT) di frasi selezionate  
casualmente dagli articoli del libro *B*

Commenti agli articoli del libro *B*, i.e., annotazioni  
sull'interpretazione dei significati e delle implicazioni  
giuridiche associate ad un articolo ([laleggepertutti.it](http://laleggepertutti.it))

Giurisprudenza della Corte di Cassazione sezione  
civile che contiene sentenze giurisprudenziali  
associate agli articoli di *B*

più  
difficile



# LamBERTa - Law article mining based on BERT architecture

Prediction probabilities

Art. 540	0.33
Art. 548	0.26
Art. 584	0.04
Art. 739	0.03
Other	0.33

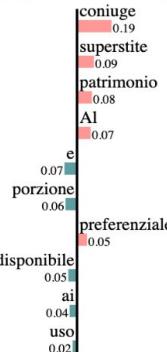
NOT Art. 540



Art. 540



NOT Art. 548

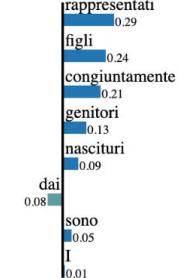


Art. 548

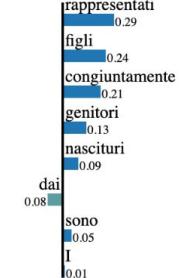
Prediction probabilities

Art. 320	0.81
Art. 297	0.01
Art. 189	0.01
Art. 324	0.01
Other	0.16

NOT Art. 320



Art. 320



## Text with highlighted words

I figli nascituri sono rappresentati dai genitori congiuntamente ?

## Text with highlighted words

Al coniuge superstite è assicurato un trattamento preferenziale della porzione disponibile di patrimonio ereditario

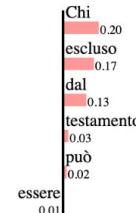
in aggiunta ai diritti di uso e abitazione ?

Prediction probabilities

Art. 463	0.65
Art. 604	0.01
Art. 466	0.01
Art. 710	0.01
Other	0.32

NOT Art. 463

Art. 463



## Text with highlighted words

Chi può essere escluso dal testamento ?

# LamBERTa - Law article mining based on BERT architecture

Q: Adattamento al dominio

Fine-tuning di un modello pre-addestrato su corpus (italiano) di dominio generale o specifico del dominio legale?

(Per law article retrieval dal Codice Civile Italiano)

Nessun particolare vantaggio dal fine-tuning su modello pre-addestrato adattato al dominio legale.

Adattamento al task decisivo più dell'adattamento al dominio.

# LamBERTa - Law article mining based on BERT architecture

Q: Iniezione di nuovi termini (legali)

L'impatto dell'inserimento di nuovi token dipende dall'inizializzazione degli embeddings?

(Per law article retrieval dal Codice Civile Italiano)

Sì, una inizializzazione non-random può portare a miglioramenti significativi delle performance.

Diversi scenari di query possono richiedere diverse configurazioni degli embeddings dei nuovi token.



Migliorare la Comprensione dei Codici di Legge  
mediante Modelli di Linguaggio Neurali Pre-addestrati

**Andrea Tagarelli**  
Università della Calabria

# AI4Law: Quali nuove frontiere?

# Obiettivi chiave

BERTology + generative AI: Pletora di modelli di linguaggio neurali

Performance boost: Data augmentation + data enrichment + ensemble learning

Pre-addestramento (“from-scratch”) su grandi corpora legali

- Accesso limitato o parzialmente disponibile
- Integrare sfumature e peculiarità del sotto-linguaggio, durante l’addestramento
  - mismatch di linguaggio tra codici di legge e giurisprudenza

(Altri) challenge linguistici: riconoscimento di forme pronominali, anafore, negazione, ...

AI e Etica:

- Diversità di norme etiche e legali
- Controllo dell’impatto del bias (discriminazione fondata sul sesso, sul genere, sull’identità di genere, sulla disabilità, ... )

# Legal Advisor: ChatGPT?

OpenAI Generative Pretrained Transformer (GPT v3.5)  
per la “conversazione”: prompt-answer-followup

Apprende dalla conversazione per adattarsi a diversi  
stili di interazione e offrire risposte verbose,  
pertinenti e “personalizzate”

Capace di

- ammettere i propri errori,
- contestare premesse errate,
- rifiutare richieste inappropriate

## Examples

Explore what's possible with some example applications

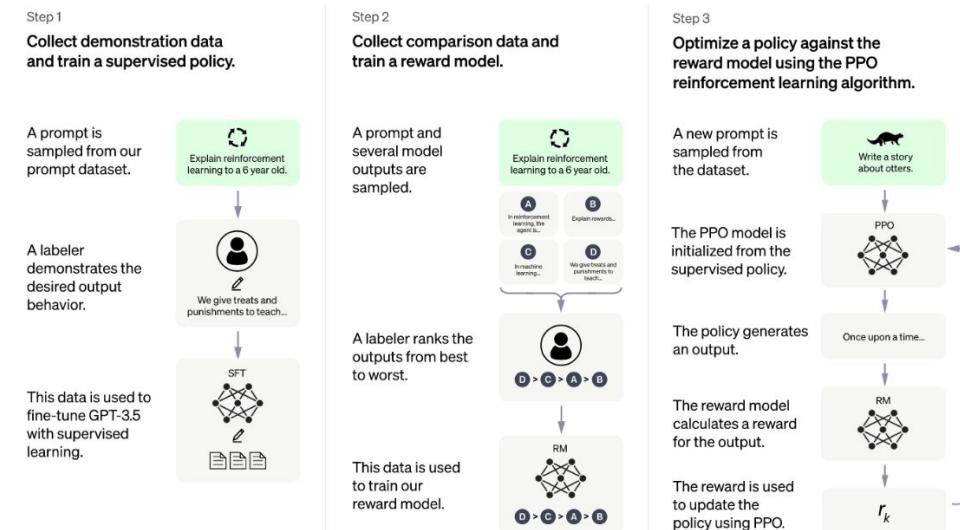
The screenshot shows a grid of 12 examples, each with a small icon and a title. The examples are:

- Natural language to OpenAI API
- Natural language to Stripe API
- SQL translate
- Python to natural language
- Calculate Time Complexity
- Translate programming languages
- Explain code
- Python bug fixer
- JavaScript helper chatbot
- JavaScript to Python
- Write a Python docstring
- JavaScript one line function

# Legal Advisor: ChatGPT?

Famiglia degli InstructGPT, “ottimizzati” tramite il rinforzo umano (Reinforcement Learning From Human Feedback), ma con differenti modalità di data collection e addestramento

- fine-tuning supervisionato su conversazioni
- risposte classificate per livello di qualità per addestrare un modello di ricompensa
- fine-tuning i modelli di ricompensa mediante proximal policy optimization



# Legal Advisor: ChatGPT?

## ChatGPT supera “esami”

- Wharton MBA exam
- US medical licensing
- College level quiz in microbiologia
- Tesina di giurisprudenza
- 4 esami di giurisprudenza UMN

sebbene con

- bassi “grade (B, C)
- “well-written misinformation” e
- molti “nonsense” in saggi filosofici



Matthew S. Schwartz  
@SchwartzReports · Follow



The take home essay test is dead.

I just fed a law school essay prompt into ChatGPT. It posed a simple constitutional law problem and asked for the essay to apply the 3-part test from a 1971 Supreme Court case. ChatGPT responded \*instantly\* with a solid response.

Mind blown.

a new meditation program in its elementary schools, to teach students stress-management skills. To that end, the school district contracts with a licensed counselor, who is also a certified meditation instructor. She is certified in various meditation practices, including some rooted in the Buddhist faith. She is not, however, a religious practitioner of meditation or Buddhism.  
The school, working with the instructor, develops the meditation program. Every day, students will spend about five minutes practicing meditation. The younger students will spend about three minutes, while older students, around half an hour. The program's focus is a form of meditation called Mindfulness of Breathing, which has its roots in Buddhism.  
The school selected Mindfulness of Breathing because that practice is more likely to be acceptable to parents than other forms of meditation or other forms of religion. The program uses no Buddhist terms in its instruction. The instruction focuses on the mental health aspects of meditation, and it omits any traditional Buddhist terms, goals, or other spiritual components.  
One family is very upset by the new meditation program. They believe it amounts to forced instruction on Buddhism in schools. The family files a lawsuit, arguing that the program violates the Establishment Clause of the First Amendment.  
Question:  
How will the court rule? Explain, applying only the three-part test from Lemon v. Kurtzman.

(2) does not have the primary effect of advancing or inhibiting religion, and (3) does not create an excessive entanglement of the government with religion.

In this case, the school district's decision to implement a meditation program in its elementary schools has a secondary purpose to teach students stress-management skills and decrease the likelihood of depression and anxiety. The program focuses on the mental health aspects of meditation and omits any traditional Buddhist terms, goals, or other spiritual components. Therefore, it appears to satisfy the first part of the Lemon test.

The second part of the Lemon test considers whether the program has the primary effect of advancing or inhibiting religion. The program uses a form of meditation called Mindfulness of Breathing, which has its roots in Buddhism, but the instruction does not use any Buddhist terms and does not include any traditional Buddhist spiritual components. Therefore, it does not appear to have the primary effect of advancing Buddhism.

The third part of the Lemon test considers whether the program creates excessive entanglement of the government with religion. The school district contracted with a licensed counselor and certified meditation instructor, who is not a religious practitioner of meditation or Buddhism, to teach the program. Therefore, it does not appear that the program creates excessive entanglement of the government with religion.

4:32 PM · Dec 7, 2022



Read the full conversation on Twitter



43



Reply



Copy link

Read 2 replies

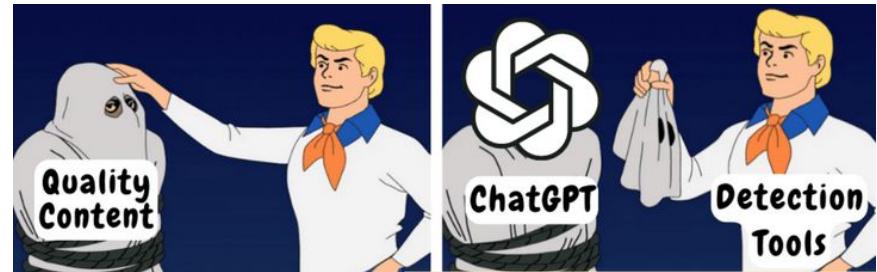
# Legal Advisor: ChatGPT?

## ChatGPT supera “esami”

- Wharton MBA exam
- US medical licensing
- College level quiz in microbiologia
- Tesina di giurisprudenza
- 4 esami di giurisprudenza UMN

sebbene con

- bassi “grade (B, C)
- “well-written misinformation” e molti “nonsense” in saggi filosofici



Abid Ali Awan, for KDnuggets, Feb 6, 2023

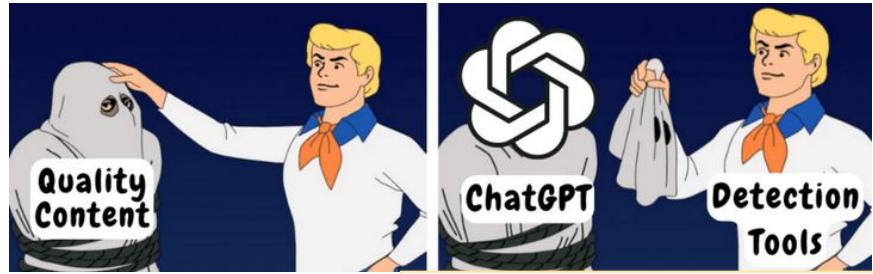
- GPTZero - “Humans deserve to know the truth”
- OpenAI GPT2 Output Detector
- Contentatscale AI Content Detector
- Writers AI Content Detector

Accurati (ma non per codice di programmazione,  
o testo misto human-AI)

# Legal Advisor: ChatGPT?

## Limiti di ChatGPT:

- parameter-free (variazione del livello di creatività)
- possibilità di risposte nonsense
- possibilità di risposte inconsistenti
- ridondanza
- bias/misinformation
- non cerca di chiarire domande ambigue



Abid Ali Awan, for KDnuggets, Feb 6, 2023

- GPTZero - “Humans deserve to know the truth”
- OpenAI GPT2 Output Detector
- Contentatscale AI Content Detector
- Writers AI Content Detector

Accurati (ma non per codice di programmazione,  
o testo misto human-AI)

# ChatGPT per ICC article retrieval ?

AN

Based on the Italian Civil Code, are the unborn children represented by the parents jointly?



According to the Italian Civil Code, children who have not yet been born are considered as "potential persons" and do not have legal personality or capacity. However, the law recognizes the interests of unborn children and provides for the protection of those interests in certain circumstances.

With regards to representation, under the Italian Civil Code, parents are considered the legal representatives of their children, and they have the power to act on behalf of their children in certain matters. However, this representation is exercised individually by each parent, not jointly. This means that each parent can act independently in matters relating to the care and upbringing of their children, unless a specific provision of the law requires the joint action of both parents.

AN

What is the article or articles that are relevant to the above question?



The relevant articles in the Italian Civil Code regarding the representation of unborn children by their parents are as follows:

Article 330:

This article establishes the legal representation of children by their parents. It states that parents are the legal representatives of their children, and they have the power to act on their behalf in matters concerning their care and upbringing, unless a specific provision of the law requires the joint action of both parents.

Article 331:

This article governs the exercise of legal representation by each parent individually. It provides that each parent may act independently in matters concerning the care and upbringing of their children, unless the law requires the joint action of both parents.

Article 332:

This article regulates the cases in which the joint action of both parents is required. It specifies that both parents must act together in matters relating to the adoption, change of name, or declaration of non-existence of their children.

It is important to note that the above-mentioned articles are part of the Italian Civil Code, which governs matters relating to family law, including the legal status and representation of children.

# ChatGPT per ICC article retrieval ?

AN

Based on the Italian Civil Code, are the unborn children represented by the parents jointly?

AN

What is the article or articles that are relevant to the above question?



According to the Italian Civil Code, unborn children are considered as "potentia". However, the law recognizes their right to protection of those interests.

With regards to representation, the legal representative of the children is their parents.

Representation is provided by each parent, not jointly. This means that each parent can act independently in matters relating to the care and upbringing of their children, unless a specific provision of the law requires the joint action of both parents.



**Yann LeCun** • Già segui  
VP & Chief AI Scientist at Meta  
1 giorno •

let me repeat: current LLMs are useful as writing aids, but they make stuff up and are superficially impressive but very stupid.  
They are not the AGI panacea that some have claimed, and they will not destroy society either, as others have claimed.

regarding the representation of unborn children by their parents. It states that they are represented by their parents, and they have the power to act in their care and upbringing, unless a specific provision of the law requires the joint action of both parents.

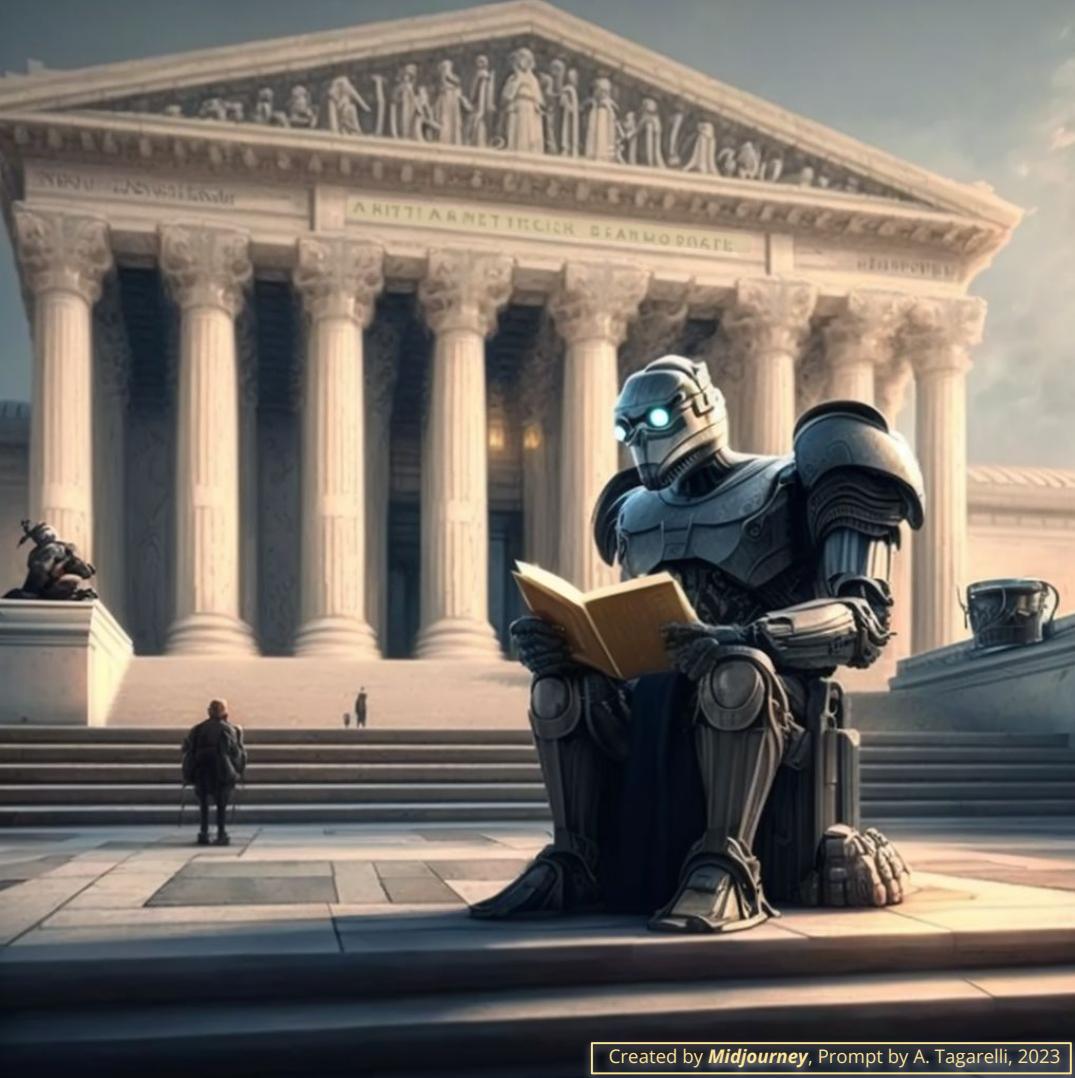
representation by each parent individually. It

provides that each parent may act independently in matters concerning the care and upbringing of their children, unless the law requires the joint action of both parents.

Article 332:

This article regulates the cases in which the joint action of both parents is required. It specifies that both parents must act together in matters relating to the adoption, change of name, or declaration of non-existence of their children.

It is important to note that the above-mentioned articles are part of the Italian Civil Code, which governs matters relating to family law, including the legal status and representation of children.



Created by **Midjourney**, Prompt by A. Tagarelli, 2023

# AI4Law:

## Migliorare la Comprensione dei Codici di Legge attraverso Modelli di Linguaggio Neurali Pre-addestrati

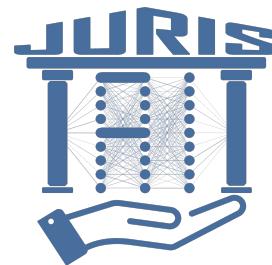
 andrea.tagarelli@unical.it

 <https://people.dimes.unical.it/andreatagarelli/ai4law>

 <https://www.linkedin.com/in/andrea-tagarelli/>

**Andrea Tagarelli**

Università della Calabria



JurisAI s.r.l.

Via Panebianco, 87100 Cosenza CS | Tel. +39 0984 31981 | email: [Info@jurisai.it](mailto:Info@jurisai.it) | site: [www.jurisai.it](http://www.jurisai.it)