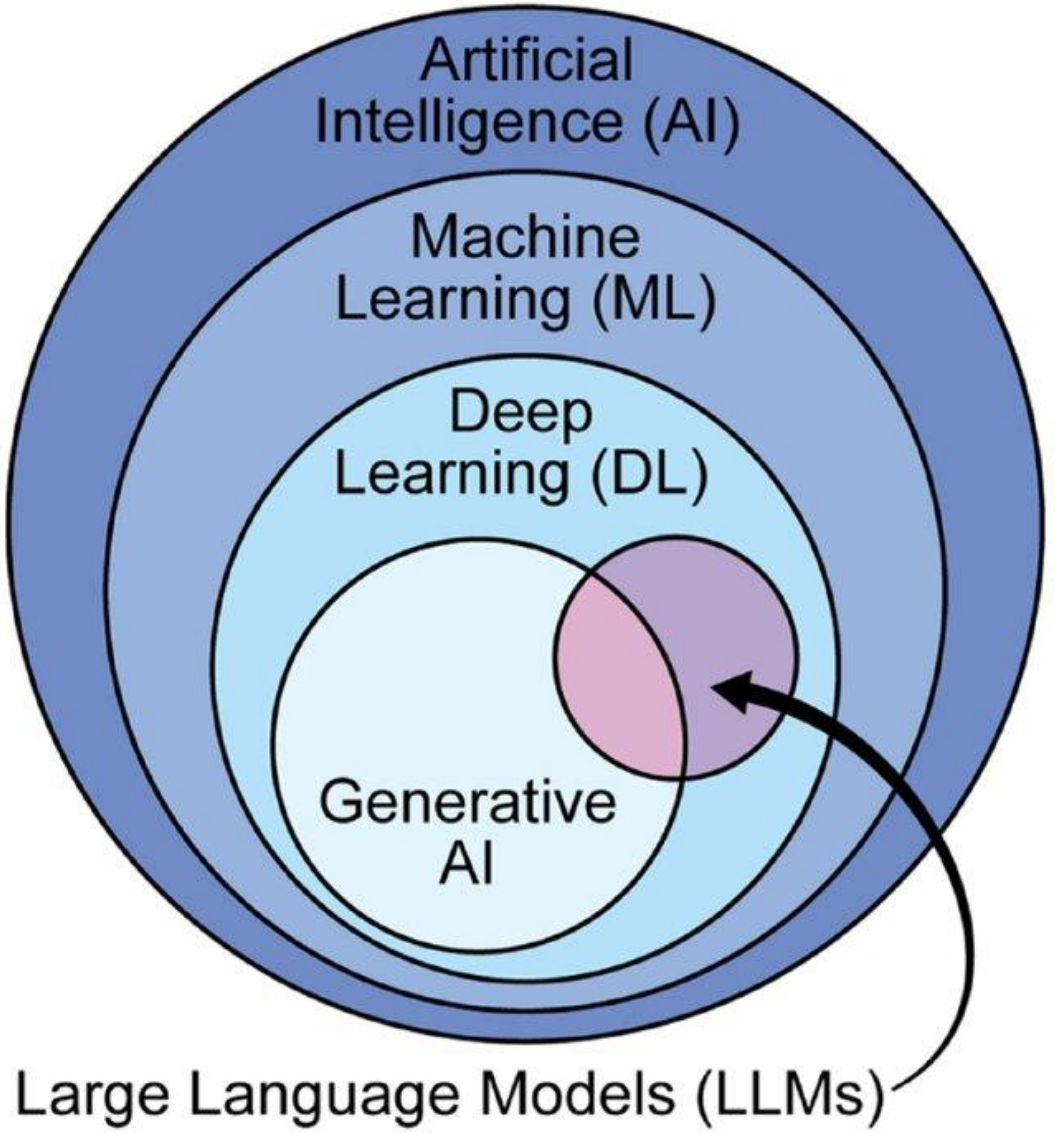
ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Η Προστασία της Ιδιωτικότητας στον Κόσμο της Παραγωγικής Τεχνητής Νοημοσύνης

Βασίλειος Σ. Βερύκιος
Καθηγητής, Σχολής Θετικών Επιστημών & Τεχνολογίας
Αντιπρύτανης Φοιτητικής Μέριμνας
Ελληνικό Ανοικτό Πανεπιστήμιο (ΕΑΠ)
verykios@eap.gr

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ



Βαθιά Μάθηση: Χρησιμοποιεί νευρωνικά δίκτυα πολλών επιπέδων (layers) σε διάφορες εφαρμογές όπως επεξεργασία εικόνας και φυσικής γλώσσας.

Παραγωγική Τεχνητή Νοημοσύνη: ειδικός τομέας της βαθιάς μάθησης που ασχολείται με τη δημιουργία νέου περιεχομένου, όπως κείμενο, εικόνες, μουσική, κλπ

•Πηγή: DOI 10.1177/17562848241227031

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

LLM - Large Language Model

- Ένα μεγάλο γλωσσικό μοντέλο (LLM - Large Language Model) είναι ένα **μοντέλο Τεχνητής Νοημοσύνης** που αποτελείται από ένα **νευρωνικό δίκτυο** με **πολλές παραμέτρους** (συνήθως δισεκατομμύρια), το οποίο **εκπαιδεύεται** σε μεγάλες ποσότητες κειμένου που **δεν έχει επισημανθεί** με ετικέτες χρησιμοποιώντας **αυτοεπιβλεπόμενη μάθηση** (Self-supervised learning).

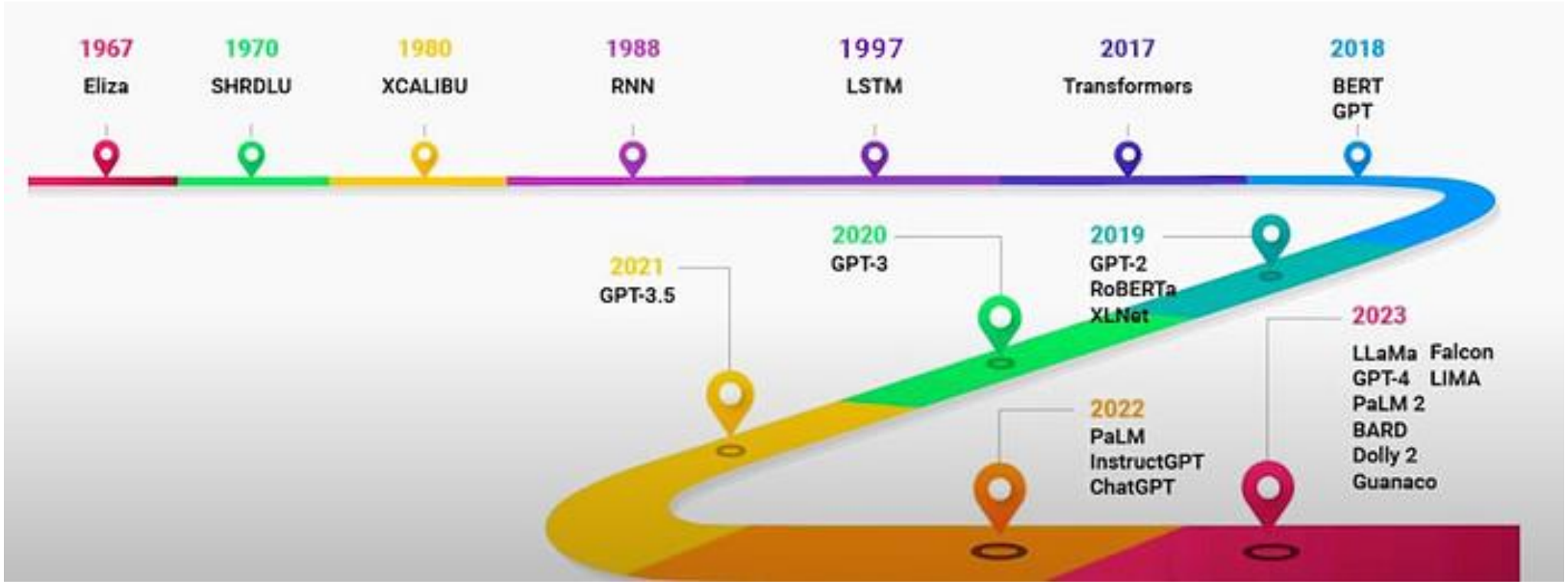
- Τα LLMs εμφανίστηκαν γύρω στο 2017 και αποδίδουν καλά σε μια μεγάλη ποικιλία εργασιών.

- Αυτό έχει μετατοπίσει το επίκεντρο της έρευνας για την **επεξεργασία φυσικής γλώσσας** από το προηγούμενο πρότυπο της εκπαίδευσης εξειδικευμένων **εποπτευόμενων μοντέλων** για συγκεκριμένες εργασίες

Πηγή: https://el.wikipedia.org/wiki

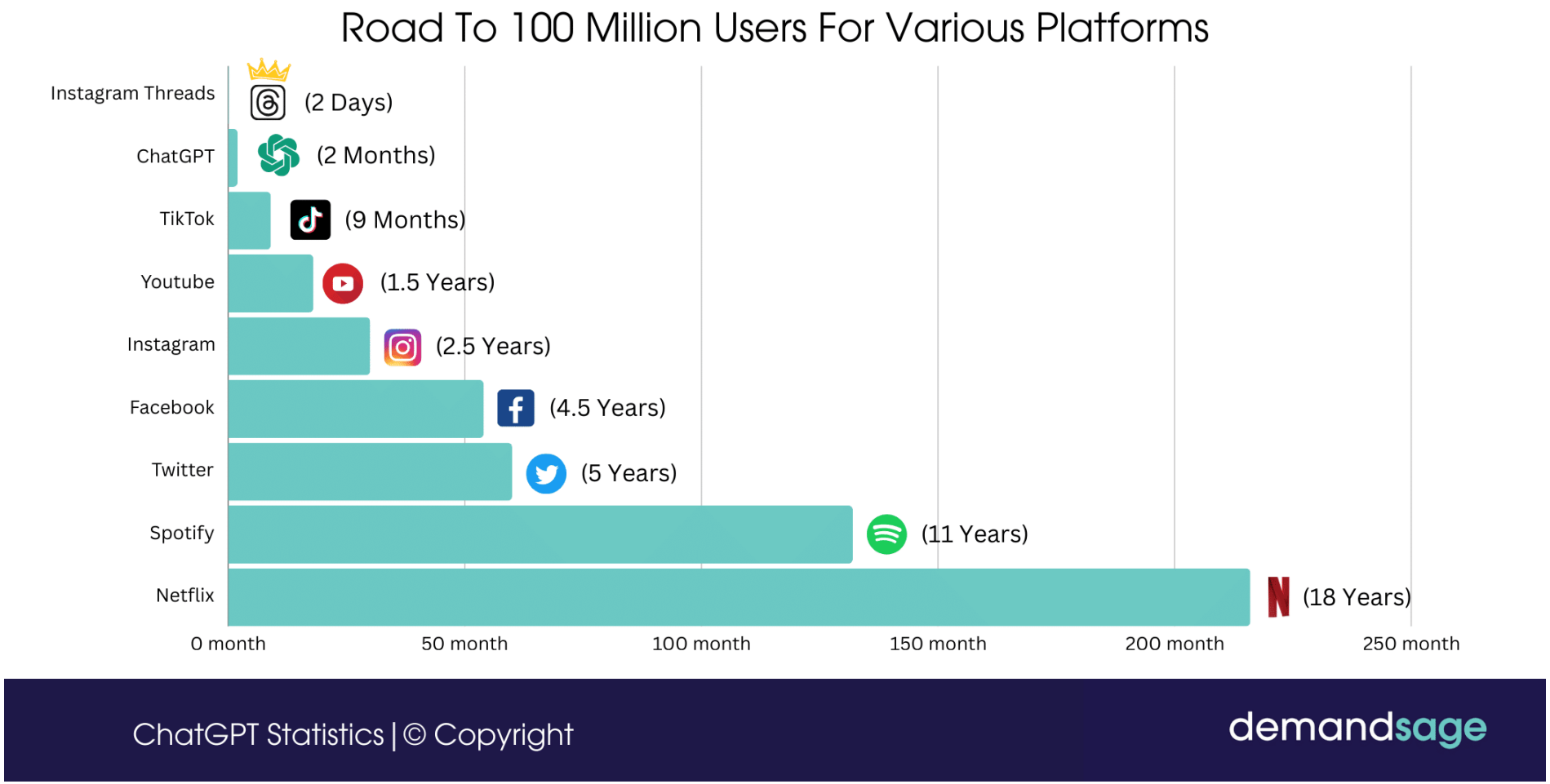
ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Εξέλιξη των LLMs



ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Η Εκρηκτική Άνοδος της χρήσης του ChatGPT



ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Πώς Λειτουργούν τα Μεγάλα Γλωσσικά Μοντέλα (LLMs)

Πρόβλεψη της Επόμενης Λέξης

Τα LLMs εκπαιδεύονται σε τεράστιες ποσότητες κειμένων, αναλύοντας τη ροή και τη λογική της γλώσσας για να προβλέψουν την επόμενη λέξη ή φράση σε ένα κείμενο.

Φανταστείτε ότι γράφετε ένα νομικό έγγραφο, όπως μια σύμβαση ή μια αγωγή.

Καθώς πληκτρολογείτε, τα Μεγάλα Γλωσσικά Μοντέλα μπορούν να "προβλέψουν" ποια είναι η επόμενη λέξη που πιθανώς θα χρησιμοποιήσετε με βάση το κείμενο που ήδη έχετε γράψει και τις εκατομμύρια άλλες συμβάσεις, υπομνήματα και νομικά κείμενα που το μοντέλο έχει "διαβάσει" κατά την εκπαίδευσή του.

Πώς Λειτουργούν τα Μεγάλα Γλωσσικά Μοντέλα (LLMs)

Πώς λειτουργεί στην πράξη;

Ας υποθέσουμε την παρακάτω πρόταση:

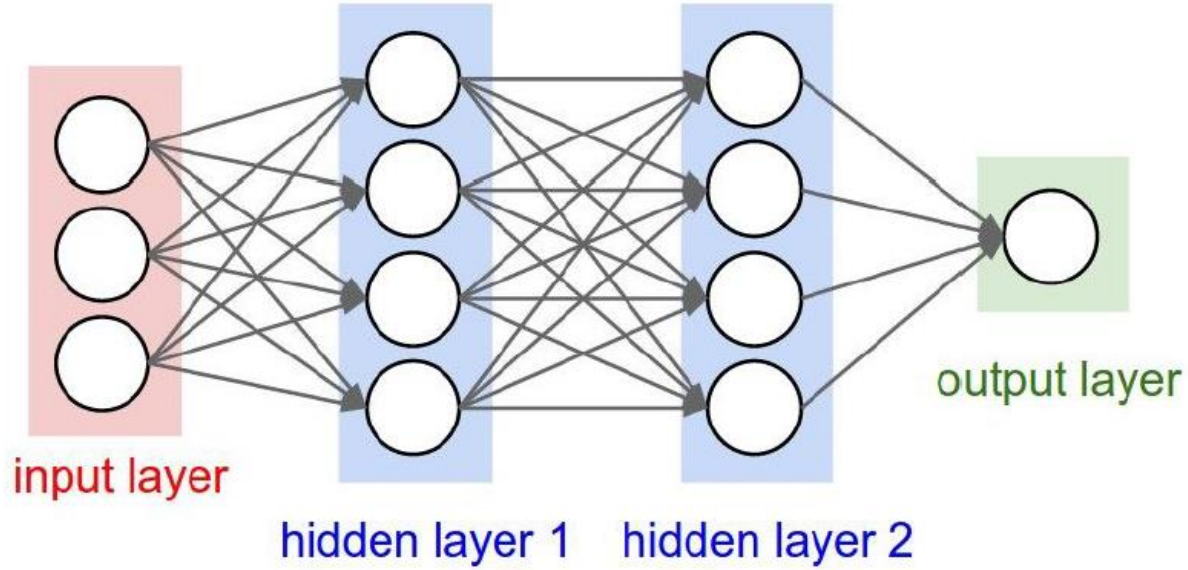
"Ο ενάγων επικαλείται το άρθρο... “

Το μοντέλο μπορεί να αναγνωρίσει ότι η πιο πιθανή επόμενη λέξη είναι "22", αν αναγνωρίσει ότι το άρθρο 22 είναι συχνά αναφερόμενο στο συγκεκριμένο νομικό πλαίσιο που εργάζεστε.

Το LLM δεν γνωρίζει τι ακριβώς σκεφτόσασταν αλλά μπορεί να προβλέψει τη λέξη με τη μεγαλύτερη πιθανότητα βάσει του προηγούμενου κειμένου και της γνώσης του από άλλα νομικά έγγραφα.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Πώς Λειτουργούν τα Μεγάλα Γλωσσικά Μοντέλα (LLMs)




```
# forward-pass of a 3-layer neural network:
f = lambda x: 1.0/(1.0 + np.exp(-x)) # activation function (use sigmoid)
x = np.random.randn(3, 1) # random input vector of three numbers (3x1)
h1 = f(np.dot(W1, x) + b1) # calculate first hidden layer activations (4x1)
h2 = f(np.dot(W2, h1) + b2) # calculate second hidden layer activations (4x1)
out = np.dot(W3, h2) + b3 # output neuron (1x1)
```

Attention Is All You Need

- Το paper **"Attention Is All You Need"**, που δημοσιεύτηκε από ερευνητές της Google και το Πανεπιστήμιο του Τορόντο, παρουσιάζει τον **Transformer**, ένα νέο μοντέλο νευρωνικών δικτύων που βασίζεται α**προσοχής** (atten ανάγκη για **επαν** που ήταν κοινές

[PDF] **Attention is all you need**

A Vaswani - Advances in Neural Information Processing Systems, 2017 - user.phil.hhu.de

Attention is **all you need Attention** is **all you need** ...

☆ Αποθήκευση 99 Παράθεση Γίνεται αναφορά σε 134253 Σχετικά άρθρα »

Attention Is All You Need

Ashish Vaswani* **Noam Shazeer*** **Niki Parmar*** **Jakob Uszkoreit***
Google Research
usz@google.com

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

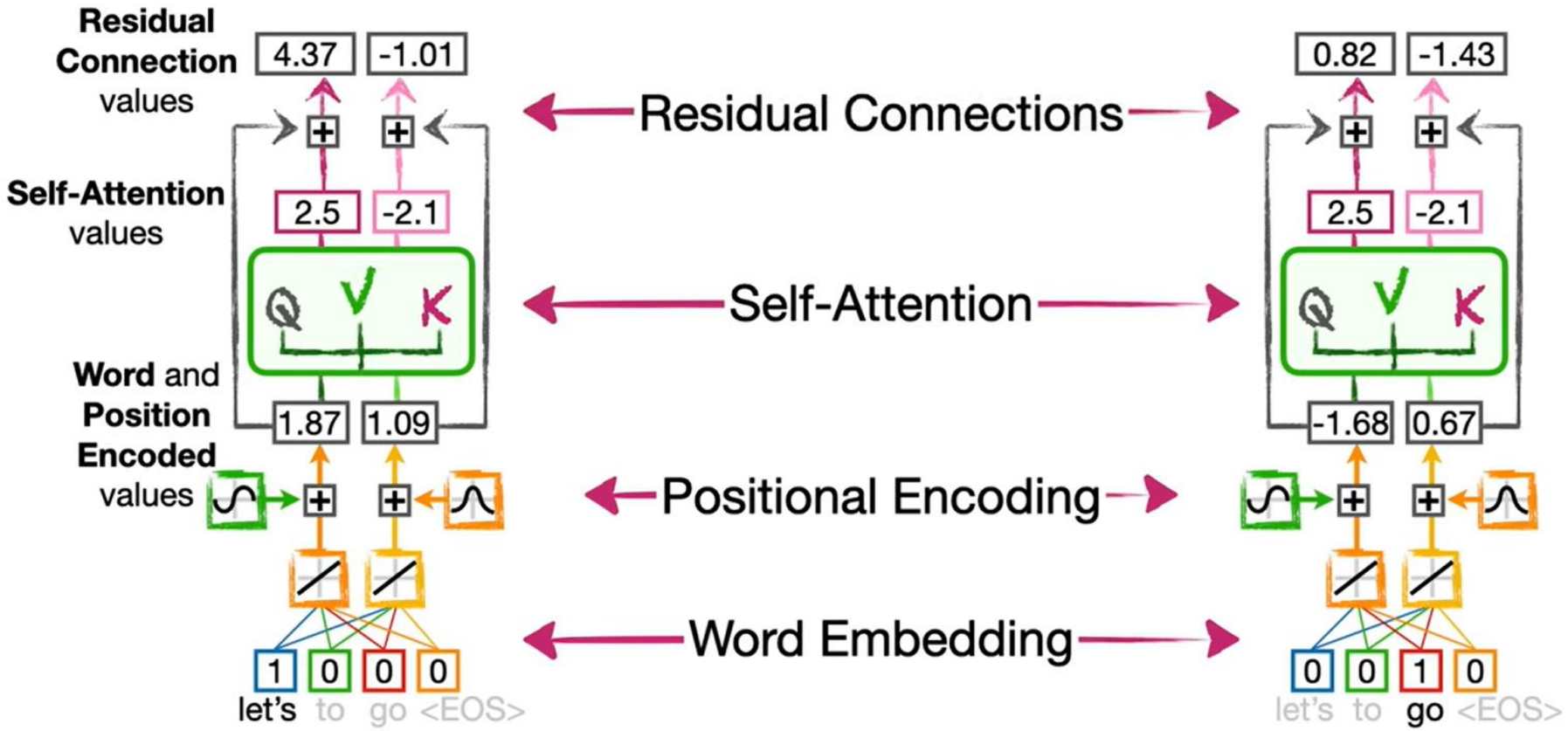
Abstract

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.

- Ο Transformer είναι ιδιαίτερα αποτελεσματικός σε εργασίες **μετατροπής ακολουθιών**, όπως η **μηχανική μετάφραση**, προσφέροντας υψηλή παραλληλία και βελτιωμένη απόδοση.
- Η εργασία αυτή έθεσε τα θεμέλια για τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) όπως το GPT, αποτελώντας σημαντική εξέλιξη στην επεξεργασία φυσικής γλώσσας.

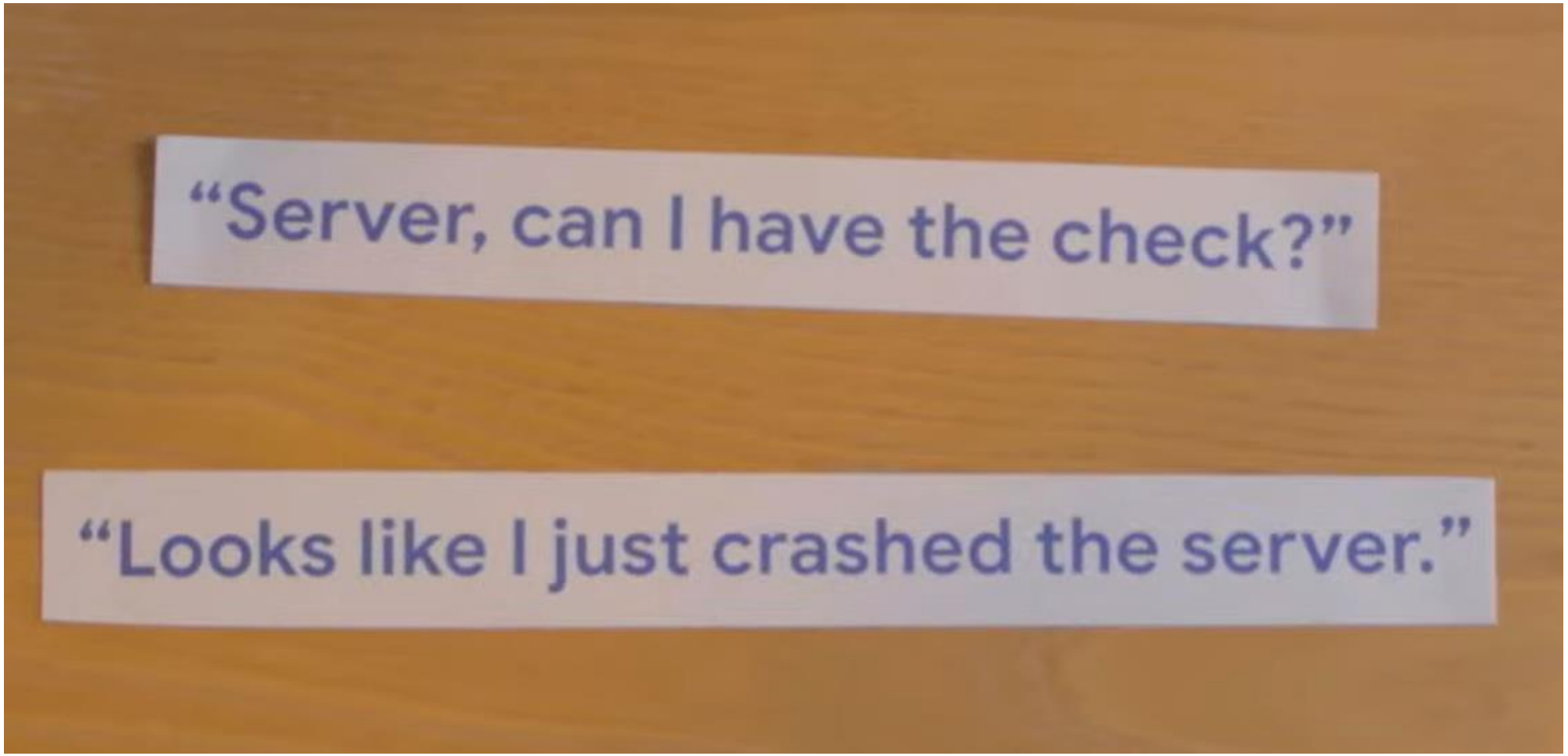
ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Απλοποιημένος Μετασχηματιστής (Transformer) για Κωδικοποίηση Εισόδου



ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Η αυτοπροσοχή επιτρέπει σε ένα νευρωνικό δίκτυο να κατανοήσει μια λέξη στο πλαίσιο των λέξεων που την περιβάλλουν…



ΕΛΛΗΝΙΚΟ
ΑΝΟΙΚΤΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ

How your data is used to improve model performance

Learn more about how OpenAI uses content from our services to improve and train our models.

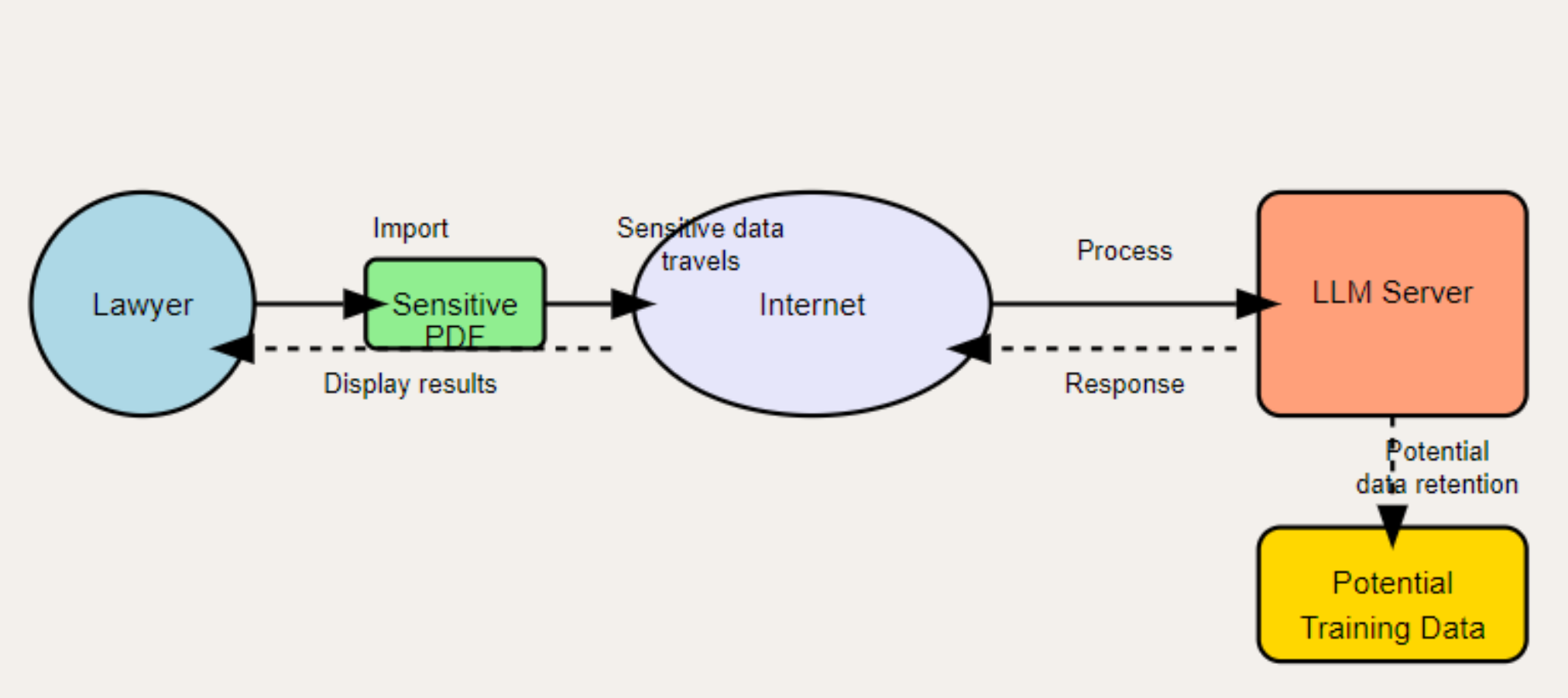
Updated over a week ago

One of the most useful and promising features of AI models is that they can improve over time. We continuously improve our models through research breakthroughs as well as exposure to real-world problems and data. When you share your content with us, it helps our models become more accurate and better at solving your specific problems and it also helps improve their general capabilities and safety. We don't use your content to market our services or create advertising profiles of you—we use it to make our models more helpful. ChatGPT, for instance, improves by further training on the conversations people have with it, unless you opt out.

Πηγή:https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Που και πως μεταφέρεται η πληροφορία;



ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Αυξανόμενες πιέσεις για τη ρύθμιση της AI και την ανάγκη για ένα πλαίσιο προστασίας των χρηστών

NEWS › COMPANY NEWS

OpenAI Faces Lawsuit Alleging Misuse of Internet Users' Data

The AI company behind ChatGPT has been accused of misappropriating millions of users' private data

By MACK WILOWSKI Published June 29, 2023 10:39 AM EDT

CNBC MARKETS BUSINESS INVESTING TECH POLITICS VIDEO INVESTING CLUB JOIN PRO JOIN LIVESTREAM

TECH

Italy became the first Western country to ban ChatGPT. Here's what other countries are doing

PUBLISHED TUE, APR 4 2023•4:48 AM EDT | UPDATED MON, APR 17 2023•1:24 AM EDT

Ryan Browne
@RYAN_BROWNE_

SHARE

KEY POINTS

- Italy last week became the first Western country to ban ChatGPT, the popular AI chatbot.
- ChatGPT has both impressed researchers with its capabilities while also worrying regulators and ethicists about the negative implications for society.
- The move has highlighted an absence of any concrete regulations, with the European Union and China among the few jurisdictions developing tailored rules for AI.
- Various governments are exploring how to regulate AI, and some are thinking of how to deal with general purpose systems such as ChatGPT.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Τα νομικά συστήματα προσαρμόζονται στις νέες προκλήσεις της AI

THE NATIONAL LAW REVIEW

TRENDING ABOUT NLR QUICK LINKS NEWSLETTERS CAREER CENTER SEARCH

TCPAWORLD STORY: Ringba Beats Litigator List On Appeal And the TCPAWorld Cracks Me Up | Colorado AG Proposes Draft

Advertisement

CAUGHT LISTENING?: Google's AI Faces Privacy Law Showdown

by: Blake Landis of Troutman Amin, LLP - *TCPAWorld*

Posted On Wednesday, September 4, 2024

RELATED PRACTICES & JURISDICTIONS

Communications Media Internet

Consumer Protection

Litigation Trial Practice

All Federal

9th Circuit (incl. bankruptcy)

Greetings TCPAWorld!

Get excited! What a time to be alive in the latest AI evolution of legal cases emerging and yet to come. I'm back with a deep dive into a juicy new class action against tech titan Google. The lawsuit, *Barulich v. Google, LLC*, 3:24CV06225, filed in California federal court, claims Google's AI-powered customer service platform is illegally eavesdropping on consumers' calls and violating state privacy laws. WOW.

CURRENT PU

Post Your Public

PUBLIC NOTICE (
THE ASSETS: Ful
Fulcrum BioEnerg

PUBLIC NOTICE (
BUSINESS SALE:

PUBLIC NOTICE (
Presto Automatio

PUBLIC NOTICE (
Ricebran Technol

PUBLIC NOTICE (
TrueNorth Project

PUBLIC NOTICE (
Short Duration Ho

Bloomberg

Live TV Markets Economics Industries Tech Politics Businessweek Opinion More

Technology
AI

Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak

- Employees accidentally leaked sensitive data via ChatGPT
- Company preparing own internal artificial intelligence tools

Gift this article

By Mark Gurman

May 2, 2023 at 3:48 AM GMT+3
Updated on May 2, 2023 at 8:54 AM GMT+3

Save

This article is for **subscribers only**.

Samsung Electronics Co. is banning employee use of popular generative AI tools like ChatGPT after discovering staff uploaded sensitive code to the platform, dealing a setback to the spread of such technology in the workplace.

The Suwon, South Korea-based company notified staff at one of its biggest divisions on Monday about the new policy via a memo reviewed by Bloomberg News. The company is concerned that data transmitted to such artificial intelligence platforms including Google Bard and Bing is stored on external servers, making it difficult to retrieve and delete, and could end up being disclosed to other

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Προστασία των πνευματικών δικαιωμάτων

Reuters

World | US Election | Business | Markets | Sustainability | Legal | Breakingviews | Technology

Meta used copyrighted books for AI training despite its own lawyers' warnings, authors allege

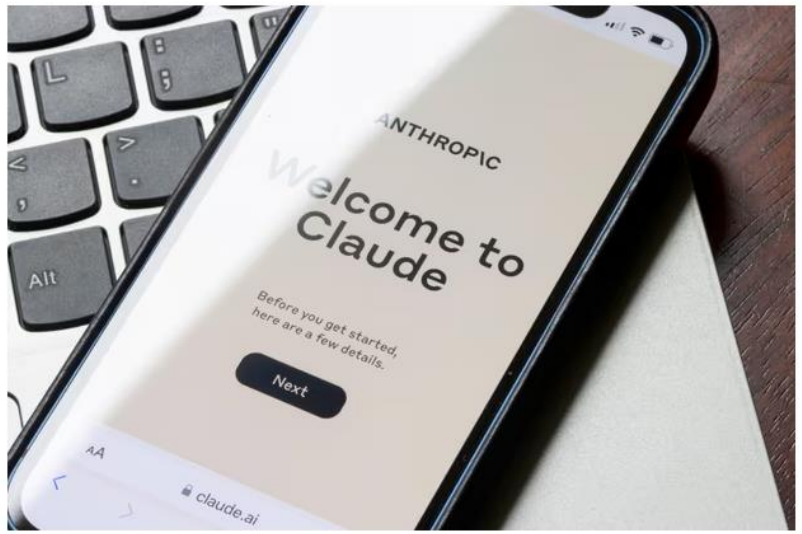
By Katie Paul

December 13, 2023 1:32 AM GMT+2 · Updated 9 months ago

Meta AI logo is seen in this illustration taken September 28, 2023. REUTERS/Dado Ruvic/Illustration/File Photo Purchase Licensing Rights

Authors sue Anthropic for copyright infringement over AI training

Andrea Bartz, Charles Graeber and Kirk Wallace Johnson allege company misused work to teach chatbot Claude



'It is no exaggeration to say that Anthropic's model seeks to profit from strip-mining the human expression and ingenuity behind each one of those works,' the complaint reads. Photograph: Ted Hsu/Alamy

The artificial intelligence company Anthropic has been hit with a class-action lawsuit in California federal court by three authors who say it misused their books and hundreds of thousands of others to train its AI-powered chatbot Claude, which generates texts in response to users' prompts.

The complaint, filed on Monday by writers and journalists Andrea Bartz, Charles Graeber and Kirk Wallace Johnson, said that Anthropic used pirated versions of their works and others to teach Claude to respond to human prompts.

Προκλήσεις Ιδιωτικότητας στα Μεγάλα Γλωσσικά Μοντέλα και Συστήματα ΤΝ

- Τα μεγάλα γλωσσικά μοντέλα εκπαιδεύονται σε τεράστιες ποσότητες δεδομένων που μπορεί να περιλαμβάνουν προσωπικές πληροφορίες.

- Τα μοντέλα έχουν την ικανότητα να παράγουν κείμενο που φαίνεται πειστικό, αλλά μπορεί να περιέχει ανακρίβειες ή πλήρως εσφαλμένες πληροφορίες, φαινόμενο γνωστό ως **"παραίσθηση"** (hallucination).

- Αυτό δημιουργεί **κινδύνους** για την **σωστή πληροφόρηση** και μπορεί να έχει σοβαρές επιπτώσεις όταν χρησιμοποιούνται σε τομείς όπως η Νομική ή η Ιατρική για παράδειγμα.

- Η αντιμετώπιση αυτού του προβλήματος απαιτεί **βελτιώσεις στα μοντέλα** και **αυξημένη επαγρύπνηση από τους χρήστες**.

Προκλήσεις Ιδιωτικότητας στα Μεγάλα Γλωσσικά Μοντέλα και Συστήματα ΤΝ

- Το Άρθρο 16 του GDPR παρέχει στους πολίτες το δικαίωμα να ζητούν τη **διόρθωση ανακριβών προσωπικών δεδομένων** που τους αφορούν.

- Ωστόσο, η εφαρμογή αυτού του δικαιώματος σε μεγάλα AI μοντέλα παρουσιάζει προκλήσεις, καθώς τα δεδομένα ενσωματώνονται στη δομή του μοντέλου κατά την εκπαίδευση.

- Η αδυναμία άμεσης διόρθωσης ή διαγραφής συγκεκριμένων πληροφοριών εγείρει **ερωτήματα σχετικά με τη συμμόρφωση με το GDPR** και απαιτεί την ανάπτυξη νέων τεχνικών και διαδικασιών.

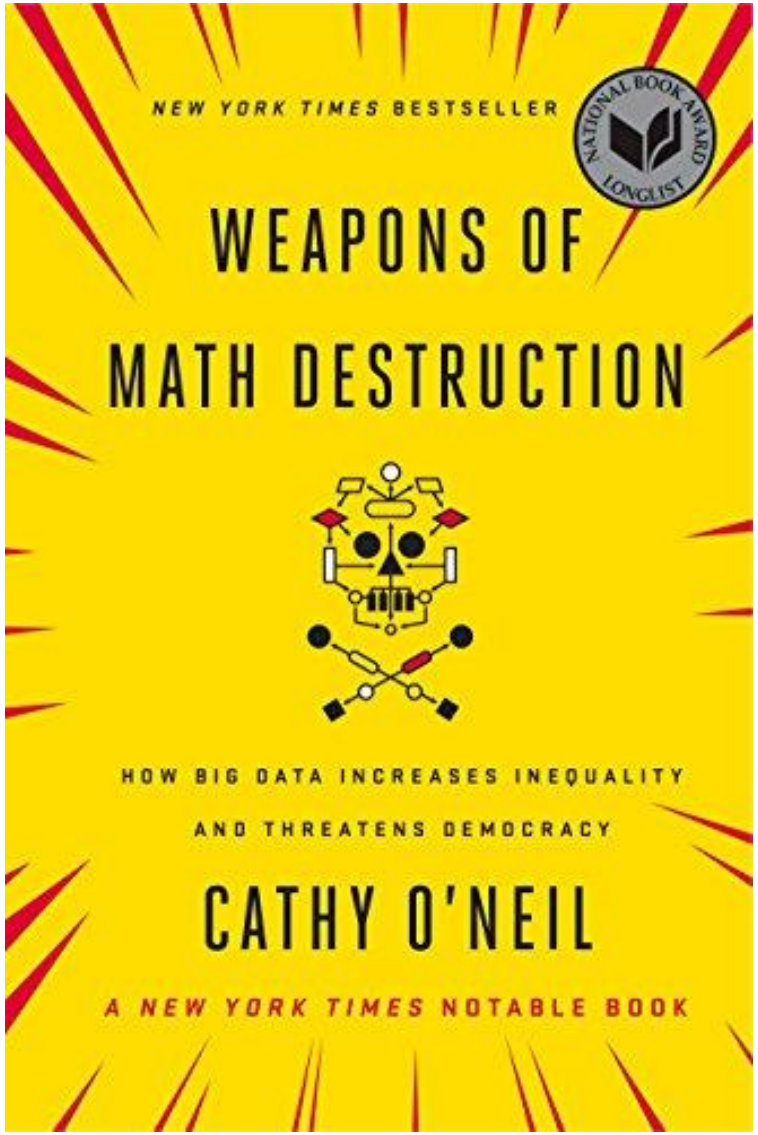
ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Αντιμετώπιση Κοινωνικών Προκαταλήψεων

- Τα μοντέλα ΤΝ μπορεί να αναπαράγουν ή και να ενισχύουν υπάρχουσες **κοινωνικές προκαταλήψεις**, καθώς εκπαιδεύονται σε δεδομένα που περιέχουν αυτές τις προκαταλήψεις.
- Αυτό μπορεί να οδηγήσει σε διακρίσεις και αθέμιτη μεταχείριση συγκεκριμένων ομάδων.
- Η αντιμετώπιση αυτού του ζητήματος απαιτεί **προσεκτική επιλογή** και επεξεργασία των **δεδομένων εκπαίδευσης**, καθώς και την εφαρμογή αλγοριθμικών τεχνικών για τη μείωση των προκαταλήψεων.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Μας αρέσει να θεωρούμε τα δεδομένα ως αμερόληπτα, αλλά η αλήθεια είναι ότι τα δεδομένα μας είναι απλώς μια εικόνα των συστημάτων που έχουμε σε εφαρμογή. Χωρίς να εξετάζουμε πώς οι δικές μας αντιλήψεις επηρεάζουν τη δημιουργία και την εφαρμογή αυτών των δεδομένων, απλώς ενισχύουμε τα υπάρχοντα συστήματά μας.



ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Διαφάνεια

- Η πολυπλοκότητα των μοντέλων ΤΝ καθιστά δύσκολη την κατανόηση του τρόπου με τον οποίο λαμβάνουν αποφάσεις ή παράγουν αποτελέσματα. Η **έλλειψη διαφάνειας** μπορεί να υπονομεύσει την εμπιστοσύνη των χρηστών και να δημιουργήσει νομικά ζητήματα, ιδίως σε περιπτώσεις όπου απαιτείται λογοδοσία.

- Η προώθηση της **επεξηγησιμότητας** και η ανάπτυξη "διαφανών" AI μοντέλων είναι κρίσιμη για την αντιμετώπιση αυτού του προβλήματος.

- Τα LLMs εκπαιδεύονται σε τεράστιους όγκους δεδομένων, συμπεριλαμβανομένων πιθανώς **ευαίσθητων πληροφοριών**. Υπάρχει ο κίνδυνος τα μοντέλα να **"διαρρεύσουν"** αυτές τις πληροφορίες μέσω των απαντήσεών τους. Προκλήσεις όπως η **"επίθεση αναδρομής"** μπορούν να επιτρέψουν την εξαγωγή αρχικών δεδομένων εκπαίδευσης.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Προτεινόμενες Λύσεις για την Προστασία της Ιδιωτικότητας

Χρήση On-Premise λύσεων

Η εγκατάσταση και λειτουργία AI μοντέλων σε **τοπικούς διακομιστές** (on-premise) επιτρέπει στους οργανισμούς να διατηρούν πλήρη έλεγχο των δεδομένων τους.
Με αυτόν τον τρόπο, μειώνεται ο κίνδυνος διαρροής ευαίσθητων πληροφοριών, καθώς τα δεδομένα δεν αποστέλλονται σε εξωτερικούς παρόχους υπηρεσιών.

Εφαρμογή Τεχνικών RAG (Retrieval-Augmented Generation)

Οι τεχνικές RAG συνδυάζουν την παραγωγική τεχνητή νοημοσύνη με εξωτερικές βάσεις δεδομένων, επιτρέποντας στα μοντέλα να ανακτούν και να χρησιμοποιούν επικαιροποιημένες και ακριβείς πληροφορίες.
Αυτό μειώνει την πιθανότητα παραγωγής ανακριβών ή ψευδών δεδομένων, βελτιώνοντας την αξιοπιστία των απαντήσεων.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Προτεινόμενες Λύσεις για την Προστασία της Ιδιωτικότητας

Χρήση API με Ελεγχόμενη Πρόσβαση

Η αλληλεπίδραση με AI μοντέλα μέσω API επιτρέπει τον έλεγχο των δεδομένων που αποστέλλονται και λαμβάνονται.

Με τη ρύθμιση κατάλληλων πολιτικών ασφαλείας και δικαιωμάτων πρόσβασης, οι οργανισμοί μπορούν να προστατεύσουν την ιδιωτικότητα των χρηστών και να διασφαλίσουν τη συμμόρφωση με κανονισμούς όπως το GDPR.

Ανάπτυξη Μηχανισμών Διόρθωσης και Διαγραφής Δεδομένων

Σύμφωνα με το Άρθρο 16 του GDPR, οι οργανισμοί πρέπει να παρέχουν τη δυνατότητα διόρθωσης ή διαγραφής προσωπικών δεδομένων.

Η ανάπτυξη τέτοιων μηχανισμών στα AI μοντέλα είναι κρίσιμη για τη νόμιμη και ηθική χρήση τους.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Προτεινόμενες Λύσεις για την Προστασία της Ιδιωτικότητας

Ενσωμάτωση Privacy by Design στις Αναπτυξιακές Διαδικασίες

Η προσέγγιση "Privacy by Design" ενσωματώνει την προστασία της ιδιωτικότητας από τα αρχικά στάδια σχεδιασμού ενός συστήματος.

Αυτό περιλαμβάνει την ελαχιστοποίηση της συλλογής δεδομένων, την ανωνυμοποίηση και την εφαρμογή ισχυρών μέτρων ασφαλείας καθ' όλη τη διάρκεια ζωής του προϊόντος ή της υπηρεσίας.

Εφαρμογή Διαφορικής Ιδιωτικότητας

Η διαφορική ιδιωτικότητα προσθέτει ελεγχόμενο "θόρυβο" στα δεδομένα ή στα αποτελέσματα, διασφαλίζοντας ότι οι μεμονωμένες εγγραφές δεν μπορούν να αναγνωριστούν.

Αυτό επιτρέπει την ανάλυση και τη χρήση δεδομένων χωρίς να θυσιάζεται η ιδιωτικότητα των ατόμων.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Προτεινόμενες Λύσεις για την Προστασία της Ιδιωτικότητας

Εκπαίδευση AI Μοντέλων με Συνθετικά ή Ανωνυμοποιημένα Δεδομένα

Η χρήση συνθετικών δεδομένων ή δεδομένων που έχουν ανωνυμοποιηθεί μειώνει τον κίνδυνο παραβίασης της ιδιωτικότητας.

Έτσι, τα μοντέλα μπορούν να εκπαιδευτούν αποτελεσματικά χωρίς να χρησιμοποιούν πραγματικά προσωπικά δεδομένα.

Συμμόρφωση με Κανονισμούς και Πρότυπα Ασφαλείας

Η υιοθέτηση και η συμμόρφωση με διεθνείς κανονισμούς, όπως το GDPR, και πρότυπα ασφαλείας εξασφαλίζει ότι οι πρακτικές του οργανισμού είναι σύμφωνες με τις νομικές απαιτήσεις και τις βέλτιστες πρακτικές του κλάδου.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Προτεινόμενες Λύσεις για την Προστασία της Ιδιωτικότητας

Συνεχής Ενημέρωση και Βελτίωση των Μοντέλων για Μείωση Ψευδών Πληροφοριών
Η τακτική ενημέρωση των AI μοντέλων και η βελτίωση των αλγορίθμων τους συμβάλλει στη μείωση της παραγωγής ανακριβών ή "παραισθητικών" δεδομένων, ενισχύοντας την αξιοπιστία τους.

Διαφάνεια και Επεξηγησιμότητα των AI Συστημάτων
Η ανάπτυξη μοντέλων που μπορούν να εξηγήσουν τις αποφάσεις και τις προβλέψεις τους επιτρέπει στους χρήστες να κατανοήσουν και να εμπιστευτούν τα αποτελέσματα. Αυτό είναι ιδιαίτερα σημαντικό σε περιβάλλοντα όπου απαιτείται λογοδοσία.

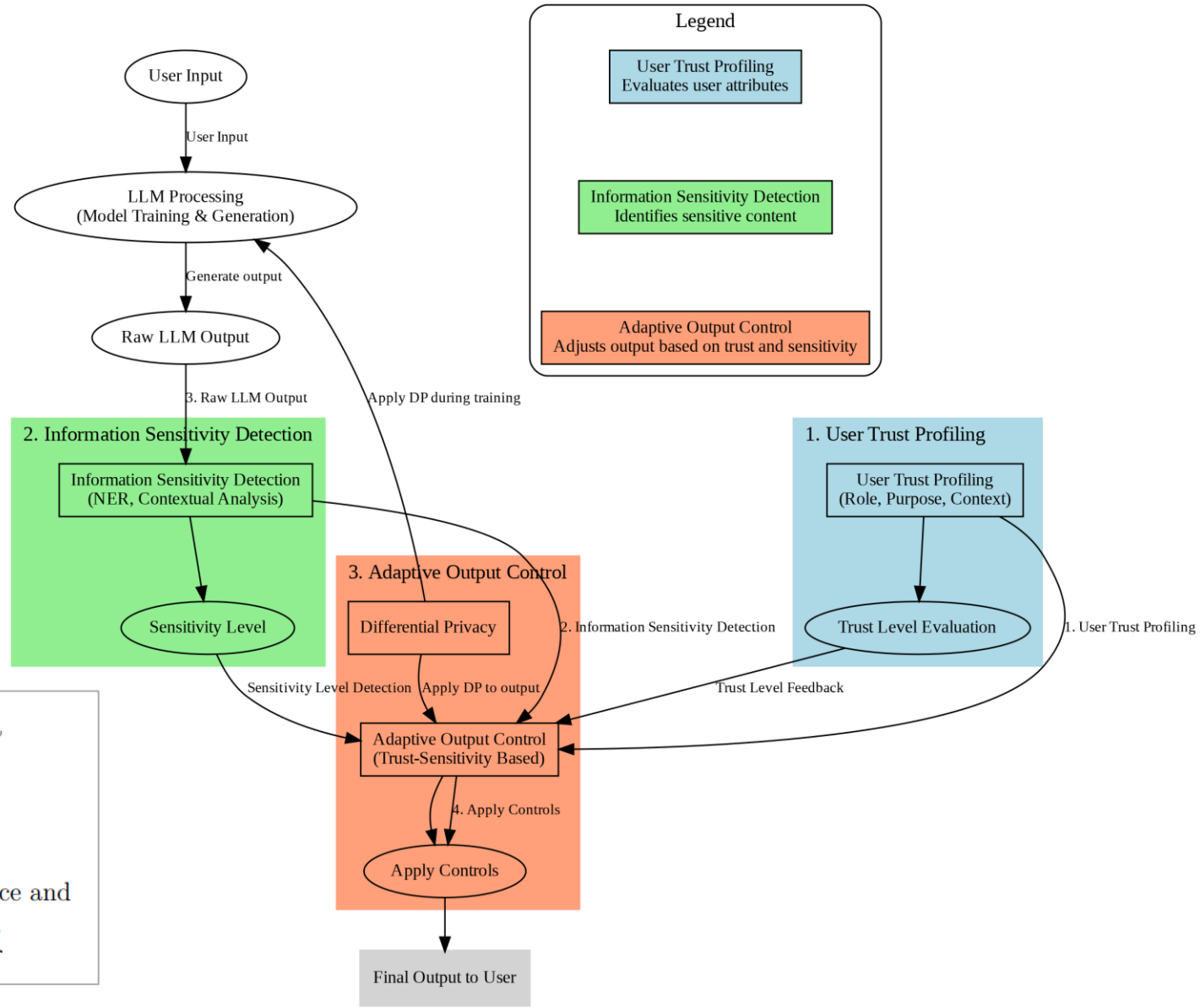
ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Προτεινόμενες Λύσεις για την Προστασία της Ιδιωτικότητας

Εκπαίδευση και Ευαισθητοποίηση Χρηστών και Εργαζομένων
Η ενημέρωση σχετικά με τους κινδύνους και τις βέλτιστες πρακτικές για την ιδιωτικότητα ενισχύει την ικανότητα των ατόμων να προστατεύουν τα προσωπικά τους δεδομένα και να χρησιμοποιούν υπεύθυνα τα AI εργαλεία.

Συνεργασία με Νομικούς και Ειδικούς σε Θέματα Ιδιωτικότητας
Η διαβούλευση με νομικούς συμβούλους και ειδικούς σε θέματα ιδιωτικότητας βοηθά στον εντοπισμό πιθανών προβλημάτων και στην ανάπτυξη στρατηγικών για την αντιμετώπισή τους.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ



Trustworthy AI: Securing Sensitive Data in Large Language Models

Georgios Feretzakis[1] and Vassilios S. Verykios[1]

[1]Big Data Analytics and Anonymization lab, School of Science and Technology, Hellenic Open University, Patras, Greece ,
georgios.feretzakis@ac.eap.gr , verykios@eap.gr

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Συμπεράσματα

- Η **προστασία της ιδιωτικότητας** στον τομέα της παραγωγικής τεχνητής νοημοσύνης είναι ένα **πολυδιάστατο ζήτημα** που απαιτεί συνδυασμό **τεχνικών**, **νομικών** και **ηθικών** προσεγγίσεων.

- Μέσω της **κατανόησης των κινδύνων** και της **εφαρμογής κατάλληλων μέτρων**, μπορούμε να **εκμεταλλευτούμε τα οφέλη** της τεχνητής νοημοσύνης χωρίς να **θυσιάζουμε την ιδιωτικότητα** των ατόμων.

- Ποιος **φέρει την ευθύνη** για τις ενέργειες ή τα αποτελέσματα που προκύπτουν από τη χρήση AI συστημάτων; Αυτό το ερώτημα είναι ιδιαίτερα σημαντικό όταν προκύπτουν ζημίες ή παραβιάσεις δικαιωμάτων.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Συμπεράσματα

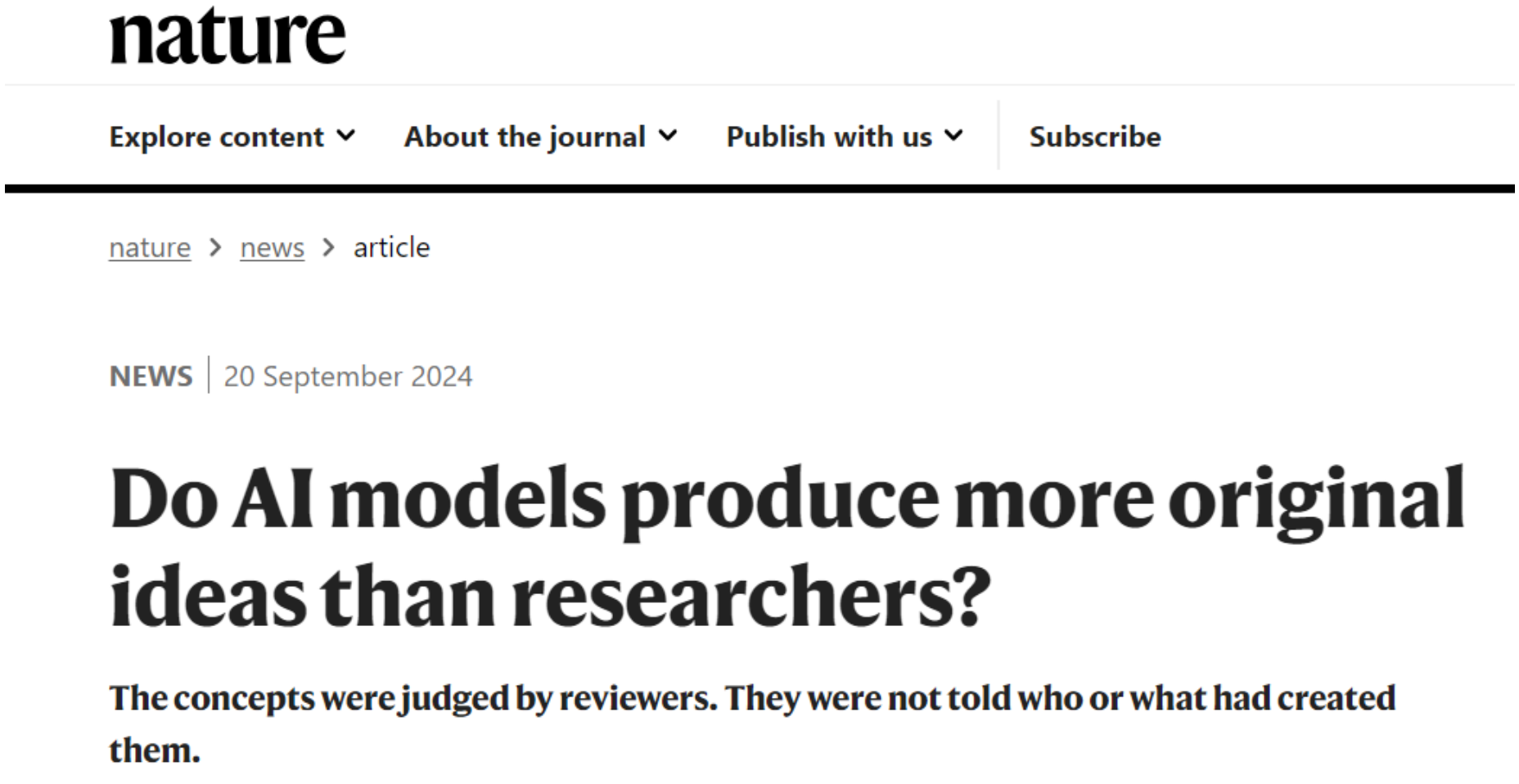
- Η καθιέρωση ενός **σαφούς νομικού πλαισίου** που θα ορίζει τις **ευθύνες** των **προγραμματιστών**, των **χρηστών** και των **εταιρειών** είναι απαραίτητη για την ασφαλή ενσωμάτωση της τεχνητής νοημοσύνης στην κοινωνία.

- Η **ενημέρωση του κοινού** σχετικά με τις **δυνατότητες** και τους **κινδύνους** της τεχνητής νοημοσύνης είναι απαραίτητη για την προώθηση μιας υπεύθυνης χρήσης της τεχνολογίας.

- **Προγράμματα εκπαίδευσης** και **καμπάνιες ευαισθητοποίησης** μπορούν να βοηθήσουν τους πολίτες να κατανοήσουν πώς προστατεύεται η ιδιωτικότητά τους και πώς μπορούν να ασκήσουν τα δικαιώματά τους.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Ευχαριστώ πολύ Ερωτήσεις



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ | Subscribe

nature › news › article

NEWS | 20 September 2024

Do AI models produce more original ideas than researchers?

The concepts were judged by reviewers. They were not told who or what had created them.