Μεγάλα Γλωσσικά Μοντέλα (LLMs) και Ιδιωτικότητα

Προκλήσεις, Κίνδυνοι και Ρυθμιστικές Προσεγγίσεις

Βασίλειος Βερύκιος

Καθηγητής, Σχολή Θετικών Επιστημών και Τεχνολογίας

Ελληνικό Ανοικτό Πανεπιστήμιο

1η Ημέρα Διαλόγου με την Ερευνητική Κοινότητα

Αρχή Προστασίας Δεδομένων

1 Οκτωβρίου 2025

Περίγραμμα Παρουσίασης

1. Εισαγωγή στα LLMs
2. Προκλήσεις Ιδιωτικότητας
3. Κίνδυνοι για την Ιδιωτικότητα
4. Ρυθμιστικό Πλαίσιο
5. Στρατηγικές Μετριασμού
6. Μελέτες Περίπτωσης
7. Μελλοντικές Κατευθύνσεις
8. Συμπεράσματα

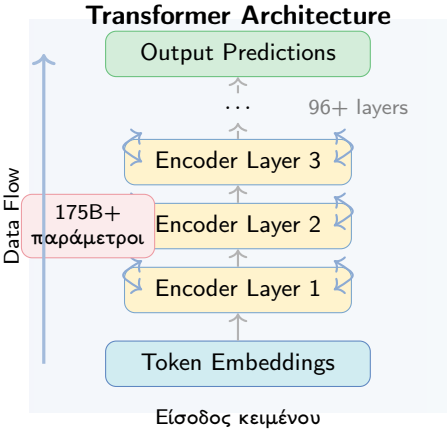
Τι είναι τα Μεγάλα Γλωσσικά Μοντέλα;

Ορισμός

Νευρωνικά δίκτυα βαθιάς μάθησης με δισεκατομμύρια παραμέτρους που εκπαιδεύονται σε τεράστιους όγκους κειμένου

Βασικά Χαρακτηριστικά:

- Transformer αρχιτεκτονική
- Petabytes δεδομένων εκπαίδευσης
- Δισεκατομμύρια παράμετροι
- Πολυγλωσσική κατανόηση

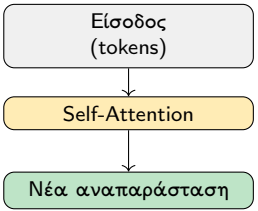


Πώς δουλεύει ο Transformer: Self-Attention

Κεντρική Ιδέα

Κάθε λέξη *προσέχει* άλλες λέξεις στο ίδιο πλαίσιο για να παραγάγει καλύτερη αναπαράσταση.

- Διανύσματα Queries (Q), Keys (K), Values (V)
- Βαθμοί προσοχής: $\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$
- Multi-Head Attention: πολλαπλές *κεφαλές* για διαφορετικές σχέσεις
- Residuals & Layer Norm: σταθερότητα εκπαίδευσης



Γιατί μας νοιάζει για ιδιωτικότητα·

Η attention μπορεί να ενισχύσει την απομνημόνευση σπάνιων ακολουθιών (π.χ. PII), εάν δεν ληφθούν αντίμετρα.

Παραδείγματα LLMs

Εμπορικά Μοντέλα

- GPT-5, o3, o1 (OpenAI)
- Claude 3.5 Sonnet (Anthropic)
- Gemini 1.5 Pro/Flash (Google)
- Llama 3.1 405B (Meta)
- Mistral Large 2 (Mistral)
- Grok-2 (xAI)

Χαρακτηριστικά Μεγέθους

- 7B - 405B+ παράμετροι
- Multi-modal δυνατότητες
- 128K - 2M tokens context
- Εκατοντάδες εκατομμύρια χρήστες
- API, chat & agents
- Real-time reasoning

Κρίσιμη Παρατήρηση

Η ραγδαία υιοθέτηση των LLMs δημιουργεί πρωτόγνωρες προκλήσεις για την προστασία προσωπικών δεδομένων

Tokens, Context Window & περικοπή (truncation)

Τι είναι ένα token·

Μονάδα κειμένου (υπο-λέξη/σύμβολο). 1 ελληνική λέξη $\approx$ 1–3 tokens συνήθως.

Context window

Μέγιστος αριθμός tokens που «χωράει» το μοντέλο σε μία κλήση (π.χ. 128K–2M). Μεγάλο context $\neq$ πλήρης μνήμη.

Κίνδυνοι για ιδιωτικότητα

- Διαρροές σε μεγάλα prompts/αρχεία αν δεν γίνει φιλτράρισμα PII.
- Truncation: κρίσιμα κομμάτια (όροι, προειδοποιήσεις) κόβονται αθόρυβα.

Πρακτικές

PII redaction πριν την αποστολή, επικύρωση μήκους, σύνοψη ευαίσθητων τμημάτων αντί για ωμή αποστολή.

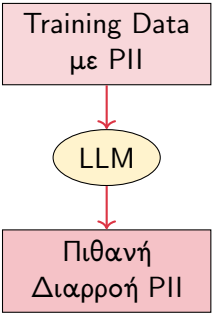
Απομνημόνευση και Διαρροή Δεδομένων

Το Φαινόμενο

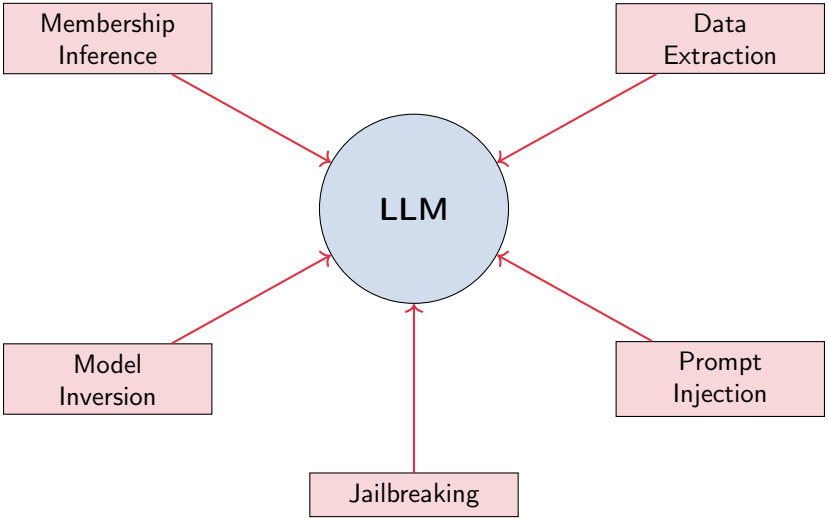
Τα LLMs μπορούν να απομνημονεύσουν και να αναπαράγουν ακριβή αντίγραφα από τα δεδομένα εκπαίδευσης

Τύποι Διαρροής:

- △! Προσωπικά στοιχεία (PII)
- △! Ιατρικά δεδομένα
- △! Οικονομικές πληροφορίες
- △! Κωδικοί πρόσβασης



Τύποι Επιθέσεων στα LLMs



Prompt Injection vs Jailbreaking: τι και πώς·

Prompt Injection

- Κακόβουλες οδηγίες σε δεδομένα/ιστοσελίδες/αρχεία
- Στόχος: διαρροή μυστικών ή εκτέλεση μη ασφαλών ενεργειών
- Άμυνα: content isolation, input/output filters, tool-use allowlists

Jailbreaking

- Χειρισμός συστημικών οδηγιών για παράκαμψη περιορισμών
- Συνήθως καθαρά διαλογικό (χωρίς εξωτερικά δεδομένα)
- Άμυνα: ισχυρά system prompts, safety tuning, αξιολόγηση & adversarial training

Γιατί ξεχωρίζει·

Το injection εκμεταλλεύεται «έμπιστη» είσοδο από εργαλεία/browsing, άρα αυξάνει τον κίνδυνο πραγματικών ενεργειών (π.χ. αποστολή email/λήψη αρχείων).

Ανάλυση Επιθέσεων και Επίπεδα Κινδύνου

Υψηλού Κινδύνου

Data Extraction

- Εξαγωγή training data
- Αποκάλυψη PII

Prompt Injection

- Χειραγώγηση εξόδου
- Παράκαμψη filters

Jailbreaking

- Άρση περιορισμών
- Κακόβουλη χρήση

Μέτριου Κινδύνου

Membership Inference

- Εντοπισμός training samples
- Παραβίαση privacy

Model Inversion

- Ανακατασκευή δεδομένων
- Αποκάλυψη patterns

Σημείωση

Ο συνδυασμός επιθέσεων μπορεί να αυξήσει δραματικά την επικινδυνότητα

Κατηγοριοποίηση Κινδύνων

Άμεσοι Κίνδυνοι

- Διαρροή προσωπικών δεδομένων
- Αποκάλυψη ευαίσθητων πληροφοριών
- Re-identification attacks
- Παραβίαση εμπιστευτικότητας

Έμμεσοι Κίνδυνοι

- Profiling και tracking
- Αυτοματοποιημένες αποφάσεις
- Συμπεράσματα από prompts
- Αλγοριθμικές διακρίσεις

Παραδείγματα Κινδύνων στην Πράξη

Περίπτωση 1: Εξαγωγή PII

Ερευνητές κατάφεραν να εξάγουν emails και τηλέφωνα από το GPT-2 χρησιμοποιώντας targeted prompts

Περίπτωση 2: Membership Inference

Εντοπισμός αν συγκεκριμένα δεδομένα χρησιμοποιήθηκαν στην εκπαίδευση

Περίπτωση 3: Prompt Injection

Χειραγώγηση του μοντέλου για παράκαμψη περιορισμών ασφαλείας

GDPR και LLMs: Προκλήσεις Συμμόρφωσης

Αρχή GDPR	**Πρόκληση για LLMs**
Νομιμότητα (Art. 6)	Ασαφής νομική βάση
Περιορισμός σκοπού (Art. 5.1.b)	Πολλαπλές χρήσεις
Ελαχιστοποίηση (Art. 5.1.c)	Massive datasets
Ακρίβεια (Art. 5.1.d)	Hallucinations
Διαφάνεια (Art. 12-14)	Black box
Δικαίωμα διαγραφής (Art. 17)	Τεχνικά δυσχερής για προ-εκπαιδευμένα μοντέλα

Κρίσιμο Ερώτημα

Πώς εφαρμόζουμε το 'δικαίωμα στη λήθη' σε ένα προ-εκπαιδευμένο μοντέλο δισεκατομμυρίων παραμέτρων;

AI Act: Νέες Υποχρεώσεις

Κατηγοριοποίηση LLMs

- General Purpose AI Models (GPAI)
- Systemic Risk Models ($'10^{25}$ FLOPs)*

**Τεκμήριο συστημικού κινδύνου - υπό αναθεώρηση*

Υποχρεώσεις Παρόχων GPAI

Εφαρμογή από 2 Αυγούστου 2025

1. Τεχνική τεκμηρίωση
2. Πολιτική συμμόρφωσης με copyright
3. Επαρκώς λεπτομερής περίληψη δεδομένων εκπαίδευσης
4. Model evaluation (για systemic risk)
5. Adversarial testing
6. Incident reporting

AI Act για GPAI: χρονοδιάγραμμα & «συστημικός κίνδυνος»

Χρονοδιάγραμμα (ενδεικτικά ορόσημα)

- Δημοσίευση: 12 Ιουλ. 2024 (Reg. (EU) 2024/1689)
- Υποχρεώσεις GPAI αρχίζουν: 2 Αυγ. 2025
- Κατευθυντήριες Επιτροπής: Ιούλ. 2025 (πλαίσιο συμμόρφωσης/περιλήψεων δεδομένων)

Τεκμήριο «συστημικού κινδύνου»

- Ενδεικτικό όριο υπολογιστικής κλίμακας (π.χ. $> 10^{25}$ FLOPs) ως τεκμήριο
- Επιπλέον απαιτήσεις: αξιολόγηση μοντέλου, red-teaming, αναφορές συμβάντων

Τι σημαίνει για οργανισμούς·

Ακόμη και χωρίς «συστημικό κίνδυνο», οι πάροχοι GPAI οφείλουν τεχνική τεκμηρίωση, πολιτική copyright, και «επαρκώς λεπτομερή» περίληψη συνόλων εκπαίδευσης.

Διεθνές Ρυθμιστικό Τοπίο

Χώρα/Οργανισμός	**Ρυθμιστική Προσέγγιση**
ΕΕ	GDPR + AI Act
ΗΠΑ	Executive Order on AI + State laws
Ηνωμένο Βασίλειο	AI White Paper + Data (Use and Access) Act 2025 + ICO Guidance
Καναδάς	Ομοσπονδιακό πλαίσιο σε εκκρεμότητα
Κίνα	Personal Information Protection Law
ΟΟΣΑ	AI Principles
ISO	ISO/IEC 23053, 23894

Παρατήρηση

Παρατηρείται διεθνής κινητοποίηση για τη ρύθμιση των LLMs με διαφορετικές προσεγγίσεις ανά περιοχή, αλλά με κοινούς στόχους προστασίας

Κοινές Ρυθμιστικές Προτεραιότητες

Βασικές Αρχές

- **Διαφάνεια**
  Κατανόηση λειτουργίας AI
- **Explainability**
  Ερμηνεία αποφάσεων
- **Risk-based approach**
  Αναλογικότητα μέτρων
- **Human oversight**
  Ανθρώπινος έλεγχος
- **Accountability**
  Σαφείς ευθύνες

Σημεία Σύγκλισης

- Προστασία θεμελιωδών δικαιωμάτων
- Ασφάλεια και αξιοπιστία
- Πρόληψη διακρίσεων
- Προστασία προσωπικών δεδομένων
- Ηθική χρήση AI

Πρόκληση

Η εναρμόνιση των διαφορετικών ρυθμιστικών πλαισίων για global AI systems παραμένει ανοιχτό ζήτημα

Τεχνικές Προστασίας Ιδιωτικότητας

Κατά την Εκπαίδευση

- Differential Privacy
- Federated Learning
- Data sanitization
- Synthetic data
- PII detection/removal

Κατά την Χρήση

- Input/Output filtering
- Privacy-preserving prompting
- Secure enclaves
- Homomorphic encryption
- Access controls

Best Practice

Συνδυασμός πολλαπλών τεχνικών για 'defense in depth'

Differential Privacy: Μαθηματική Προστασία

Ορισμός:
Ένας μηχανισμός $\mathcal{M}$ ικανοποιεί (ϵ, δ)-DP αν:

Μαθηματική Εγγύηση

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta$$

Παράμετροι:

- ϵ: Privacy budget
- δ: Failure probability

Πλεονεκτήματα

- Μαθηματικές εγγυήσεις
- Προστασία από attacks
- Plausible deniability

Μειονεκτήματα

- Trade-off με accuracy
- Υπολογιστικό κόστος
- Δυσκολία εφαρμογής

DP-SGD στην πράξη (για LLMs)

Μηχανισμός

1. Αποκοπή (clipping) επιμέρους βαθμίδων σε όριο C
2. Προσθήκη θορύβου $\mathcal{N}(0, \sigma^2 C^2 I)$ στις ομαδοποιημένες βαθμίδες
3. Λογιστική ιδιωτικότητας: παρακολούθηση ϵ, δ ανά εποχή

- Επιλογές: μέγεθος παρτίδας, C, λόγος θορύβου σ
- Αντίκτυπος: πτώση ακρίβειας/κόστους, αλλά ισχυρές εγγυήσεις

Καλές πρακτικές

- Pretraining χωρίς PII + DP στο fine-tuning
- Συνδυασμός με ανίχνευση/απομάκρυνση PII

Όρια

Η αυστηρή DP σε πλήρη προεκπαίδευση μεγάλων μοντέλων παραμένει υπολογιστικά δαπανηρή.

Οργανωτικά Μέτρα

1. **Privacy by Design**
   - Ενσωμάτωση privacy από την αρχή
   - Privacy Impact Assessments (PIAs)
2. **Governance Framework**
   - AI Ethics Board
   - Clear policies and procedures
   - Regular audits
3. **Transparency Measures**
   - Model cards
   - Data sheets
   - User notifications
4. **Εκπαίδευση Προσωπικού**
   - Privacy awareness
   - Responsible AI practices

Case Study: ChatGPT και Ιταλική DPA (Garante)

Timeline

Μάρ. 2023	Data breach - Redis bug
31 Μαρ. 2023	Προσωρινή απαγόρευση
Απρίλιος 2023	Διαπραγματεύσεις
28 Απρ. 2023	Επαναλειτουργία

Απαιτήσεις Συμμόρφωσης:

- Age-gating (13+)
- Opt-out για training data
- Διαφάνεια επεξεργασίας
- Ενημέρωση χρηστών (Privacy Notice)
- Νομική βάση (legitimate interest)

Μάθημα

Η proactive συμμόρφωση είναι προτιμότερη από την reactive

Αποτέλεσμα

Βελτίωση πρακτικών privacy σε global επίπεδο

Case Study: Samsung Data Leak

Το Πρόβλημα (Απρίλιος 2023)

- Υπάλληλοι ανέβασαν proprietary code στο ChatGPT
- Διαρροή εμπιστευτικών πληροφοριών
- Κίνδυνος για intellectual property

Αντίδραση Εταιρείας

- Απαγόρευση χρήσης public LLMs
- Ανάπτυξη internal AI tools
- Εκπαίδευση προσωπικού
- Νέες πολιτικές ασφαλείας

Κρίσιμο Μάθημα

Η εκπαίδευση χρηστών είναι εξίσου σημαντική με τα τεχνικά μέτρα

Αναδυόμενες Τεχνολογίες και Λύσεις

Τεχνολογικές Καινοτομίες

- Machine Unlearning
- Selective Forgetting
- Privacy-preserving fine-tuning
- Secure multi-party computation
- Local/Edge LLMs

Ερευνητικές Προτεραιότητες

- Explainable AI
- Privacy-Utility Trade-offs
- Adversarial Robustness
- User Control Mechanisms
- Legal Tech Integration

Ερευνητική Πρόκληση

Πώς επιτυγχάνουμε ισορροπία μεταξύ utility και privacy χωρίς να θυσιάζουμε την απόδοση;

Προτάσεις για τη Βιομηχανία

1. **Adopt Privacy-First Design**
   - Default to privacy-preserving configurations
   - Minimize data collection
2. **Implement Robust Governance**
   - Clear accountability structures
   - Regular assessments
3. **Invest in Privacy Tech**
   - R&D για privacy-preserving ML
   - Collaboration με academia
4. **Collaborate with Regulators**
   - Proactive engagement
   - Regulatory sandboxes
5. **Educate Stakeholders**
   - Users, developers, management
   - Continuous training programs

Κύρια Συμπεράσματα

- △! Τα LLMs παρουσιάζουν **μοναδικές προκλήσεις** για την ιδιωτικότητα
- ≡ Το υπάρχον ρυθμιστικό πλαίσιο χρειάζεται **προσαρμογή**
- □ Απαιτείται **πολυεπίπεδη προσέγγιση** (τεχνική + οργανωτική + νομική)
- ∩ Η **συνεργασία** μεταξύ stakeholders είναι κρίσιμη
- ◦ Η **έρευνα και καινοτομία** στο privacy-preserving AI είναι απαραίτητη

Το Μέλλον

Η επιτυχής ενσωμάτωση των LLMs στην κοινωνία εξαρτάται από την ικανότητά μας να διασφαλίσουμε την ιδιωτικότητα και να οικοδομήσουμε εμπιστοσύνη

Ανοιχτές Προκλήσεις

1. **Machine Unlearning**
   Πώς διαγράφουμε δεδομένα από εκπαιδευμένα μοντέλα;
2. **Privacy-Utility Balance**
   Ποιο είναι το βέλτιστο trade-off;
3. **Compliance Verification**
   Πώς ελέγχουμε black-box models;
4. **Liability Attribution**
   Ποιος ευθύνεται για privacy violations;
5. **Cross-Border Governance**
   Πώς ρυθμίζουμε global AI services;

Πρόκληση για την Ερευνητική Κοινότητα

Αυτά τα ερωτήματα απαιτούν διεπιστημονική συνεργασία και συνεχή έρευνα

Ευχαριστώ για την Προσοχή σας!

Στοιχεία Επικοινωνίας

verykios@eap.gr

www.eap.gr

linkedin.com/in/vassilios-verykios-b6988598

1η Ημέρα Διαλόγου με την Ερευνητική Κοινότητα

Αρχή Προστασίας Δεδομένων • 1 Οκτωβρίου 2025

Βιβλιογραφία

- Carlini, N., et al. (2021). "Extracting Training Data from Large Language Models." USENIX Security Symposium.
- Brown, T., et al. (2022). "What Does it Mean for a Language Model to Preserve Privacy?" ACM FAccT.
- European Data Protection Board (2024). Opinion 28/2024 on certain data protection aspects related to AI models; and Report of the ChatGPT Taskforce (23 May 2024).
- EU AI Act (Regulation (EU) 2024/1689). Official Journal of the EU, 12 July 2024.
- Abadi, M., et al. (2016). "Deep Learning with Differential Privacy." ACM CCS.
- Shokri, R., et al. (2017). "Membership Inference Attacks Against Machine Learning Models." IEEE S&P.
- Thudi, A., et al. (2022). "Unrolling SGD: Understanding Factors Influencing Machine Unlearning."
- Zhang, C., et al. (2023). "Privacy-Preserving Machine Learning: Methods and Applications."

Επιπλέον Πόροι

Οργανισμοί:

- EDPB/EDPS
- AI Now Institute
- Partnership on AI
- Future of Privacy Forum
- Αρχή Προστασίας Δεδομένων

Guidelines:

- GDPR Guidelines
- AI Act Compliance
- ISO/IEC 23053
- NIST AI RMF

Research Groups:

- Stanford HAI
- MIT CSAIL
- Oxford FHI
- DeepMind Safety
- OpenAI Safety

Tools & Frameworks:

- TensorFlow Privacy
- PyTorch Opacus
- PySyft
- Microsoft SEAL
- Google DP Library