Προστασία της ιδιωτικότητας και μεγάλα γλωσσικά μοντέλα: Η επίθεση περί «Συμπεράσματος Συμμετοχής Δεδομένων» (Membership Inference Attack)

Βασιλική Διαμαντοπούλου, Πανεπιστήμιο Αιγαίου, ΑΔΑΕ
Στέφανος Γκρίτζαλης, Πανεπιστήμιο Πειραιώς, ΑΔΑΕ, ΕΕΔΑ

01/10/2025 – Διαδικτυακή Ημερίδα

Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs) (1/2)

- Τα LLMs είναι προηγμένα μοντέλα βαθιάς μάθησης που έχουν σχεδιαστεί για να επεξεργάζονται και να παράγουν ανθρώπινη γλώσσα
- Βασίζονται στην αρχιτεκτονική μετασχηματιστών (transformer architecture), η οποία χρησιμοποιεί μηχανισμούς προσοχής (attention mechanisms) για την κατανόηση του πλαισίου και των σχέσεων μεταξύ των λέξεων
- Τα μοντέλα αυτά – μοντέλα γενικής χρήσης – εκπαιδεύονται σε εκτεταμένα σύνολα δεδομένων, τα οποία συχνά περιλαμβάνουν:
  - Δημόσια διαθέσιμο περιεχόμενο
  - Ιδιόκτητα σύνολα δεδομένων
  - Εξειδικευμένα δεδομένα για συγκεκριμένους τομείς (sector-specific)

Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs) (2/2)

- Εκπαιδεύονται σε έναν τεράστιο όγκο δεδομένων
- Συνεισφέρουν σε διάφορους τομείς μέσω:
  - Δημιουργίας και σύνοψης κειμένου
  - Μετάφρασης ερωτήσεων-απαντήσεων
- Χρήση:
  - Ανάλυση γλωσσικών προτύπων
  - Λογική και μαθηματική συλλογιστική
  - Παραγωγή κώδικα
- Δυνατότητες:
  - Προσαρμογή σε διαφορετικά πλαίσια
  - Ολοκληρώνουν νέες εργασίες χωρίς περαιτέρω εκπαίδευση

Μεγάλα Γλωσσικά Μοντέλα – Φάσεις (1/3)

Η ανάπτυξη των LLMs διακρίνεται σε διαφορετικά στάδια:

1. Εκπαίδευση: Κατασκευή του Μοντέλου
   i. Συλλογή συνόλου δεδομένων
   ii. Προ-επεξεργασία δεδομένων
   iii. Μετασχηματισμός
   iv. Βρόγχος εκπαίδευσης/ανάδρασης και βελτιστοποίηση

Μεγάλα Γλωσσικά Μοντέλα – Φάσεις (2/3)

Η ανάπτυξη των LLMs διακρίνεται σε διαφορετικά στάδια:

2. Συνεχής Βελτίωση – Ευθυγράμμιση Μοντέλων (μετά την εκπαίδευση)

   Διαδικασία που περιλαμβάνει χρήση τεχνικών όπως:

   - η εποπτευόμενη ρύθμιση σε δεδομένα που αφορούν συγκεκριμένους τομείς
   - η Ενισχυτική Μάθηση με Ανθρώπινη Ανατροφοδότηση (RLHF)

Μεγάλα Γλωσσικά Μοντέλα – Φάσεις (3/3)

Η ανάπτυξη των LLMs διακρίνεται σε διαφορετικά στάδια:

3. Συμπεράσματα: Δημιουργία αποτελεσμάτων
   A. Είσοδος:
      i. Υποβολή ερωτήματος από τον άνθρωπο
      ii. Tokenization και ενσωμάτωση
   B. Επεξεργασία:
      i. Μέσω της αρχιτεκτονικής μετασχηματιστή, όπου οι μηχανισμοί προσοχής και τα επίπεδα αποκωδικοποιητή προβλέπουν τα επόμενα tokens στην ακολουθία
      ii. Ο αποκωδικοποιητής παράγει διάνυσμα βαθμολογιών για κάθε λέξη στο λεξιλόγιο
      iii. Στη συνέχεια, αυτές οι βαθμολογίες μετατρέπονται σε πιθανότητες
      iv. Το μοντέλο επιλέγει το πιο πιθανό token ως την επόμενη λέξη στην ακολουθία, διασφαλίζοντας ότι το δημιουργημένο κείμενο είναι συνεκτικό και σχετικό με τα συμφραζόμενα
   C. Έξοδος: Το μοντέλο παράγει πιθανότητες για πιθανές επόμενες λέξεις, επιλέγοντας τις πιο πιθανές επιλογές με βάση την είσοδο και τα συμφραζόμενα. Αυτές οι προβλέψεις συνδυάζονται για να δημιουργήσουν συνεκτικές και σχετικές απαντήσεις

Μεγάλα Γλωσσικά Μοντέλα και Κίνδυνοι για την Ιδιωτική Ζωή

- Κίνδυνοι για την προστασία της ιδιωτικής ζωής κατά τη διαδικασία ανάπτυξης/λειτουργίας του μοντέλου:
  - Φάση 1 – κατά τη συλλογή δεδομένων: Το σύνολο εκπαίδευσης, δοκιμών και επικύρωσης θα μπορούσε να περιέχει προσωπικά δεδομένα
  - Φάση 2 – κατά τη διαδικασία βελτίωσης του μοντέλου, όπου ενδέχεται να χρησιμοποιηθούν κείμενα ή γνωσιακές βάσεις που μπορεί να περιέχουν προσωπικά δεδομένα
  - Φάση 2 – κατά τη διαδικασία της ανατροφοδότησης, όπου οι αλληλεπιδράσεις των χρηστών ενδέχεται να αποθηκεύονται χωρίς επαρκείς δικλείδες ασφαλείας.
  - Φάση 3 – κατά την παραγωγή/δημιουργία αποτελεσμάτων: Τα αποτελέσματα θα μπορούσαν να αποκαλύψουν προσωπικά δεδομένα ή να περιέχουν παραπληροφόρηση

Εφαρμογές των Μεγάλων Γλωσσικών Μοντέλων (1/3)

- Τα LLMs φέρνουν την υπόσχεση της επανάστασης στον τρόπο που εργάζονται οι άνθρωποι, αυτοματοποιώντας τις εργασίες και βελτιώνοντας την εμπειρία του χρήστη, την παραγωγικότητα και τη λειτουργική αποδοτικότητά τους
- Οι εφαρμογές τους είναι ποικίλες, από τη δημιουργία και τη σύνοψη κειμένου έως την υποστήριξη κωδικοποίησης, την ανάλυση συναισθημάτων και πολλά άλλα
- Ορισμένα LLM είναι *πολυτροπικά*, ικανά να επεξεργάζονται και να δημιουργούν πολλαπλές μορφές δεδομένων, όπως εικόνα, ήχο ή βίντεο.

Εφαρμογές των Μεγάλων Γλωσσικών Μοντέλων (2/3)

- **Chatbots και βοηθοί AI:** Τα LLMs ενδυναμώνουν εικονικούς βοηθούς, κατανοούν και επεξεργάζονται τη φυσική γλώσσα, ερμηνεύουν την πρόθεση του χρήστη και δημιουργούν απαντήσεις
- **Δημιουργία περιεχομένου:** Τα LLMs βοηθούν στη δημιουργία άρθρων, αναφορών και υλικού μάρκετινγκ, δημιουργώντας κείμενο, βελτιστοποιώντας έτσι τις διαδικασίες δημιουργίας περιεχομένου
- **Μετάφραση γλώσσας:** Τα προηγμένα LLMs διευκολύνουν τις υπηρεσίες μετάφρασης σε πραγματικό χρόνο
- **Ανάλυση συναισθήματος:** Οι επιχειρήσεις χρησιμοποιούν τα LLMs για να αναλύουν τα σχόλια των πελατών και το περιεχόμενο των μέσων κοινωνικής δικτύωσης, αποκτώντας γνώσεις για το δημόσιο αίσθημα και ενημερώνοντας για στρατηγικές αποφάσεις.

Εφαρμογές των Μεγάλων Γλωσσικών Μοντέλων (3/3)

- **Δημιουργία κώδικα και εντοπισμός σφαλμάτων:** Οι προγραμματιστές αξιοποιούν τα LLMs για να δημιουργούν αποσπάσματα κώδικα και να εντοπίζουν σφάλματα, ενισχύοντας την αποτελεσματικότητα της ανάπτυξης λογισμικού
- **Εργαλεία εκπαιδευτικής υποστήριξης:** Τα LLMs χρησιμοποιούνται στην εξατομικευμένη μάθηση, δημιουργώντας εκπαιδευτικό περιεχόμενο, εξηγήσεις και απαντώντας σε ερωτήσεις μαθητών
- **Επεξεργασία νομικών εγγράφων:** Εξέταση και σύνοψη νομικών κειμένων, εξάγοντας σημαντικές πληροφορίες
- **Υποστήριξη πελατών:** Αυτοματοποίηση απαντήσεων σε ερωτήματα πελατών και ανάθεση σύνθετων υποθέσεων σε ανθρώπους
- **Αυτόνομα οχήματα:** Οδήγηση αυτοκινήτων με δυνατότητες λήψης αποφάσεων σε πραγματικό χρόνο

Επιπτώσεις από τις Ευπάθειες των Μεγάλων Γλωσσικών Μοντέλων

- Οι ευπάθειες ασφάλειας και ιδιωτικότητας ενδέχεται να εγείρουν σοβαρές ανησυχίες και κινδύνους κατά την ανάπτυξή τους στον πραγματικό κόσμο, με ποικίλες επιπτώσεις σε διαφορετικούς τομείς εφαρμογών:
  - Περίπλοκη ανθρώπινη αλληλεπίδραση
  - Ψευδή αποτελέσματα/παραίσθηση (hallucination), παραπληροφόρηση και διάδοση παραπληροφόρησης:
  - Κυβερνοέγκλημα και Κοινωνικά Ζητήματα
  - Μεταφορές
  - Υγειονομική περίθαλψη και ιατρική
  - Εκπαίδευση
  - Κυβέρνηση
  - Επιστήμη

Ευπάθειες Μεγάλων Γλωσσικών Μοντέλων

- Τα μεγάλα γλωσσικά μοντέλα είναι ευπαθή σε επιθέσεις που έχουν στόχο να πλήξουν την **ασφάλεια** των **πληροφοριών** και την **προστασία** των **δεδομένων**
- **Ασφάλεια Πληροφοριών:** Σκοπός είναι η θωράκιση του συστήματος, υπακούοντας στις απαιτήσεις CIA, περιλαμβάνοντας αποφυγή:
  - Μη εξουσιοδοτημένης πρόσβασης
  - Τροποποίησης
  - Δυσλειτουργίας
  - Άρνησης υπηρεσίας σε εξουσιοδοτημένους χρήστες
- **Ιδιωτικότητα Χρηστών:** Προστασία προσωπικών πληροφοριών, ελέγχοντας την πρόσβαση σε συστήματα που επεξεργάζονται προσωπικές πληροφορίες

Επιθέσεις σε Μεγάλα Γλωσσικά Μοντέλα

Prompt Injection

Gradient Leakage Attacks

Data Poisoning Attacks

Jailbreaking Attacks

Bias Exploitation

Model Inversion

Membership Inference Attacks

Backdoor Attacks

Denial of Service

Extraction/stealing

PII Leakage Attacks

Unintended Memorisation

Προκλήσεις κατά της Ιδιωτικότητας στα Μεγάλα Γλωσσικά Μοντέλα

- Οι κίνδυνοι για την ιδιωτικότητα προκύπτουν από την ικανότητα των LLMs να επεξεργάζονται και να δημιουργούν κείμενο βάσει του μεγάλου όγκου και ποικιλίας δεδομένων εκπαίδευσης
- Τα μοντέλα ενδέχεται να καταγράφουν και να αναπαράγουν ευαίσθητες πληροφορίες που υπάρχουν στα δεδομένα εκπαίδευσης

- → Ανησυχία για την ιδιωτικότητα κατά τη διαδικασία δημιουργίας κειμένου

Προκλήσεις κατά της Ιδιωτικότητας στα Μεγάλα Γλωσσικά Μοντέλα

- Βασικές προκλήσεις:
  - Απομνημόνευση δεδομένων
  - Διαρροή δεδομένων
  - Πιθανή αποκάλυψη εμπιστευτικών πληροφοριών ή PII

Η επίθεση περί «Συμπεράσματος Συμμετοχής Δεδομένων» (Membership Inference Attack – MIA) (1/4)

- Κατά την επίθεση αυτή, ο επιτιθέμενος προσπαθεί να διαπιστώσει εάν ένα συγκεκριμένο δείγμα δεδομένων αποτέλεσε, ή όχι, μέρος του συνόλου εκπαίδευσης του μοντέλου
- Όταν ένα συγκεκριμένο δεδομένο είναι πλήρως γνωστό στον επιτιθέμενο, το γεγονός ότι γνωρίζει πως το δεδομένο αυτό χρησιμοποιήθηκε για να εκπαιδευτεί ένα συγκεκριμένο μοντέλο, είναι ένδειξη διαρροής πληροφορίας μέσω του μοντέλου
- Στόχος: Παραβίαση ιδιωτικότητας και αποκάλυψη ευαίσθητων πληροφοριών

Παράδειγμα: Γνωρίζοντας ότι τα κλινικά δεδομένα ενός συγκεκριμένου ασθενή που χρησιμοποιήθηκαν για να εκπαιδευτεί ένα μοντέλο, συσχετίστηκαν με μία ασθένεια (προκειμένου να καθοριστεί η κατάλληλη δόση φαρμάκου ή να μελετηθεί η γενετική βάση της προς μελέτη ασθένειας), αποκαλύπτουν ότι ο ασθενής πάσχει από την εν λόγω ασθένεια

Η επίθεση περί «Συμπεράσματος Συμμετοχής Δεδομένων» (Membership Inference Attack – MIA) (2/4)

- Τα μοντέλα συχνά παρουσιάζουν διαφορετική συμπεριφορά:
  - Σε δεδομένα εκπαίδευσης: μεγαλύτερη εμπιστοσύνη / καλύτερη απόδοση
  - Σε νέα δεδομένα → λιγότερη εμπιστοσύνη
- Ο επιτιθέμενος εκμεταλλεύεται αυτές τις διαφορές, υποβάλλοντας ερωτήματα στο μοντέλο και αναλύοντας τις αποκρίσεις

Η επίθεση περί «Συμπεράσματος Συμμετοχής Δεδομένων» (Membership Inference Attack – MIA) (3/4)

1. Επιλογή στόχου

   Ο επιτιθέμενος επιλέγει έναν στόχο με συγκεκριμένα δεδομένα (μέσω κειμένου ή εικόνας) για το οποίο θέλει να ελέγξει αν συμπεριλαμβάνεται στο σύνολο εκπαίδευσης του μοντέλου

2. Αποστολή ερωτήματος

   Ο επιτιθέμενος στέλνει συγκεκριμένα ερωτήματα (queries) στο μοντέλο, τα οποία υποβάλλονται ως είσοδος (input)

3. Καταγραφή απαντήσεων

   Ο επιτιθέμενος λαμβάνει την έξοδο με μορφή προβλεφθεισών πιθανοτήτων, logits, ή κειμένου

Η επίθεση περί «Συμπεράσματος Συμμετοχής Δεδομένων» (Membership Inference Attack – MIA) (4/4)

4. Ανάλυση Απόκρισης
   - Ο επιτιθέμενος εξετάζει χαρακτηριστικά όπως:
     - Πιθανότητες / διακυμάνσεις (confidence) - θέτοντας όρια (thresholds) στα αποτελέσματα για να αποφασίσει αν το αντικείμενο ενδιαφέροντος είναι μέλος του συνόλου δεδομένων ή όχι
     - Σφάλματα του μοντέλου στο αντικείμενο ενδιαφέροντος
   - Κατόπιν, συγκρίνει την απόκριση με άλλα δείγματα που γνωρίζει ότι δεν αποτελούν μέρος του συνόλου δεδομένων εκπαίδευσης ή με άλλα βοηθητικά μοντέλα (shadow models)
   - Αναλύει αν η απόκριση μοιάζει με «εκπαιδευμένο» ή «νέο» δείγμα δεδομένων
5. Εξαγωγή συμπεράσματος

   Το αντικείμενο ενδιαφέροντος συμπεριλήφθηκε ή όχι στο σύνολο δεδομένων εκπαίδευσης

Επίθεση μέσω Συμπεράσματος Συμμετοχής Δεδομένων

- Όσον αφορά τον **στόχο της επίθεσης**, η MIA επιχειρεί να αποκαλύψει τη σχέση μεταξύ του δείγματος και του πραγματικού ιδιωτικού συνόλου εκπαίδευσης
- Όσον αφορά την πηγή πληροφοριών, η MIA βασίζεται στο διάνυσμα πιθανότητας που σχετίζεται με το δείγμα εισόδου

Σενάρια Επίθεσης

- Μοντέλο ***άμεσα*** διαθέσιμο:
  - Black-box scenario: ο επιτιθέμενος μπορεί να τροφοδοτήσει είσοδο στο μοντέλο και να λάβει τα αποτελέσματα του της επεξεργασίας
  - White-box scenario: ο επιτιθέμενος γνωρίζει τον τύπο και την αρχιτεκτονική του μοντέλου μηχανικής μάθησης και του αλγορίθμου εκπαίδευσης
- Μοντέλο ***έμμεσα*** διαθέσιμο:
  - Ένας προγραμματιστής εφαρμογών μπορεί να χρησιμοποιήσει μια υπηρεσία μηχανικής μάθησης για να κατασκευάσει ένα μοντέλο από τα δεδομένα που συλλέγονται από την εφαρμογή και να ζητήσει από την εφαρμογή να κάνει κλήσεις API στο μοντέλο που προκύπτει. Σε αυτήν την περίπτωση, ο αντίπαλος θα παρείχε δεδομένα εισόδου στην εφαρμογή (αντί να τα έδινε απευθείας στο μοντέλο) και θα λάμβανε τα δεδομένα εξόδου της εφαρμογής (τα οποία βασίζονται στα δεδομένα εξόδου του μοντέλου). Οι λεπτομέρειες της εσωτερικής χρήσης του μοντέλου ποικίλλουν σημαντικά από εφαρμογή σε εφαρμογή.

Αντίκτυπος στην Ιδιωτικότητα των Φυσικών Προσώπων

- Έκθεση/διαρροή προσωπικών δεδομένων
- Παραβίαση εμπιστευτικότητας σε κρίσιμους τομείς (υγεία, οικονομία)
- Υπονόμευση εμπιστοσύνης των χρηστών σε υπηρεσίες που βασίζονται σε LLMs και σε συστήματα AI, εν γένει

Τρόποι Άμυνας έναντι στην επίθεση μέσω Συμπεράσματος Συμμετοχής Δεδομένων (1/3)

- Dropout και Model Stacking, ή και συνδυασμός αυτών

  Η τυχαία διαγραφή ενός ορισμένου ποσοστού νευρωνικών συνδέσεων, μπορεί να μετριάσει την υπερπροσαρμογή (overfitting), η οποία αποτελεί παράγοντα που συμβάλλει στη MIA
- Differential Privacy: μπορεί να μειώσει τη διαρροή δεδομένων που σχετίζονται με την ιδιωτικότητα των χρηστών, εξασφαλίζοντας παράλληλα συγκρίσιμη χρησιμότητα του μοντέλου σε περιβάλλον εκτός DP
- MemGuard: μηχανισμός προσθήκης θορύβου που επιδρά στην προβλεπόμενη βαθμολογία εμπιστοσύνης του μοντέλου-στόχου
- Adversarial Regularization: αναφέρεται σε ένα σύνολο τεχνικών που χρησιμοποιούνται για την αποτροπή της υπερπροσαρμογής και τη βελτίωση της απόδοσης γενίκευσης ενός μοντέλου ML

Τρόποι Άμυνας έναντι στην επίθεση μέσω Συμπεράσματος Συμμετοχής Δεδομένων (2/3)

- InferDPT: πρόσφατο πλαίσιο, που έχει προταθεί για την αξιοποίηση των LLM μαύρου κουτιού για τη διευκόλυνση της εξαγωγής συμπερασμάτων που διατηρούν την ιδιωτικότητα, το οποίο ενσωματώνει αποτελεσματικά την DP σε εργασίες δημιουργίας κειμένου
- Pruning αφαιρούνται περιττές παράμετροι από το ίδιο το μοντέλο, χωρίς σημαντική απώλεια στην απόδοσή του, επιτυγχάνοντας ταχύτερους χρόνους εξαγωγής συμπερασμάτων (inference), χαμηλότερη κατανάλωση μνήμης και υπολογιστικής ισχύος
- Knowledge Distillation: τεχνική όπου «διδάσκεται» ένα μικρότερο μοντέλο (ο μαθητής) από ένα μεγαλύτερο, πιο ικανό μοντέλο (τον δάσκαλο), μεταφέροντας τη γνώση του δασκάλου στον μαθητή

Τρόποι Άμυνας έναντι στην επίθεση μέσω Συμπεράσματος Συμμετοχής Δεδομένων (3/3)

- Οι περισσότερες από τις υπάρχουσες τεχνικές άμυνας έχουν πειραματιστεί σε σχετικά μικρά LM, όπως οι ταξινομητές δοκιμών (test classifiers), οι οποίοι δεν αξιολογούνται για LLMs
- Επιπλέον, οι άμυνες που βασίζονται σε DP, παρέχουν ισχυρή προστασία αλλά ενδέχεται να επηρεάσουν τη χρησιμότητα του μοντέλου
- Η σχετική βιβλιογραφία του πεδίου αναφέρει ότι υπάρχει πιεστική ανάγκη για περαιτέρω ερευνητικές μελέτες για την ανάπτυξη αποτελεσματικών τεχνικών άμυνας κατά της MIA σε LLMs

Συμπεράσματα (1/2)

- Τα μοντέλα μηχανικής μάθησης αποκαλύπτουν πληροφορίες για τα σύνολα δεδομένων εκπαίδευσης που χρησιμοποιούν (training datasets), με υψηλά ποσοστά ακρίβειας
- Η επίθεση περί «Συμπεράσματος Συμμετοχής Δεδομένων» αποτελεί σοβαρή απειλή για την ιδιωτικότητα
  - Ο αντίκτυπος επηρεάζει τόσο την προστασία των φυσικών προσώπων όσο και την ασφάλεια των χρηστών

Συμπεράσματα (2/2)

Η έννοια Dalenius desideratum (1977), γνωστή για τον έλεγχο στατιστικής αποκάλυψης, δηλώνει ότι *τίποτα για ένα άτομο δεν πρέπει να είναι δυνατόν να μαθευτεί από τη βάση δεδομένων που δεν μπορεί να μαθευτεί χωρίς πρόσβαση στη βάση δεδομένων.*

Αυτό που προκύπτει είναι ότι αυτό δεν μπορεί να επιτευχθεί με κανένα χρήσιμο μοντέλο

Βιβλιογραφία

- [1] Pan, X., hang, M., Ji, S., & Yang, M. (2020, May). Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 1314-1331). IEEE.
- [2] Sideri, M., & Gritzalis, S. (2025). Gender Mainstreaming Strategy and the Artificial Intelligence Act: Public Policies for Convergence. Digital Society, 4(1), 1-22.
- [3] Neel, S., & Chang, P. (2023). Privacy issues in large language models: A survey. arXiv preprint arXiv:2312.06717.
- [4] Xin, Y., Li, Z., Yu, N., Backes, M., & Zhang, Y. (2022). Membership leakage in pre-trained language models.
- [5] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 100211.
- [6] Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., & Berg-Kirkpatrick, T. (2023). Membership inference attacks against language models via neighbourhood comparison. arXiv preprint arXiv:2305.18462.
- [7] Tong, M., Chen, K., Zhang, J., Qi, Y., Zhang, W., Yu, N., ... & Zhang, Z. (2025). InferDPT: Privacy-preserving Inference for Black-box Large Language Models. IEEE Transactions on Dependable and Secure Computing.
- [8] Nasr, M., Shokri, R., & Houmansadr, A. (2018, October). Machine learning with membership privacy using adversarial regularization. In Proceedings of the 2018 ACM SIGSAC conference on computer and communications security (pp. 634-646).
- [9] Jia, J., Salem, A., Backes, M., Zhang, Y., & Gong, N. Z. (2019, November). Memguard: Defending against black-box membership inference attacks via adversarial examples. In Proceedings of the 2019 ACM SIGSAC conference on computer and communications security (pp. 259-274).
- [10] Li, X., Tramer, F., Liang, P., & Hashimoto, T. (2021). Large language models can be strong differentially private learners. arXiv preprint arXiv:2110.05679.