

Θέμα: Prompt Injection Attacks σε LLMs: Πλαίσιο δοκιμών και σύγκρισης διαφορετικών μοντέλων	
Επιβλέπων: Γεώργιος Σπαθούλας	Στοιχεία επικοινωνίας: gspathoulas@uth.gr
Σκοπός και στόχοι Η παρούσα πτυχιακή εργασία αποσκοπεί στη συστηματική μελέτη των επιθέσεων τύπου prompt injection σε μεγάλα γλωσσικά μοντέλα (Large Language Models – LLMs) και στη δημιουργία ενός πλαισίου δοκιμών για την αξιολόγηση και σύγκριση της ανθεκτικότητας διαφορετικών μοντέλων. Στόχος είναι: <ul style="list-style-type: none"> • η κατανόηση των μηχανισμών με τους οποίους οι επιθέσεις prompt injection επηρεάζουν τη συμπεριφορά των LLMs, • η ανάπτυξη τυποποιημένων σεναρίων επίθεσης, • και η συγκριτική αξιολόγηση διαφορετικών μοντέλων ως προς την ασφάλεια και ανθεκτικότητά τους. 	
Αντικείμενο Τα LLMs χρησιμοποιούνται ευρέως σε εφαρμογές όπως chatbots, agents και συστήματα υποστήριξης αποφάσεων. Ωστόσο, είναι ιδιαίτερα ευάλωτα σε επιθέσεις μέσω φυσικής γλώσσας, όπου κακόβουλες οδηγίες ενσωματώνονται στις εισροές (prompts) με σκοπό την αλλοίωση της συμπεριφοράς του μοντέλου. Η εργασία εστιάζει στα εξής: <ul style="list-style-type: none"> • Prompt Injection Attacks, • Jailbreaking τεχνικές για παράκαμψη πολιτικών ασφαλείας, • Context Manipulation (εκμετάλλευση ιστορικού διαλόγου), Κεντρικός άξονας είναι η ανάπτυξη ενός πλαισίου δοκιμών που θα επιτρέπει: <ul style="list-style-type: none"> • την εκτέλεση επιθέσεων σε πολλαπλά LLMs, • τη μέτρηση της επιτυχίας τους, • και τη σύγκριση της συμπεριφοράς των μοντέλων. 	
Η εργασία περιλαμβάνει <ul style="list-style-type: none"> • Ανάλυση των επιθέσεων prompt injection και των παραλλαγών τους. • Επισκόπηση της πρόσφατης βιβλιογραφίας στον χώρο της ασφάλειας των LLMs. • Σχεδιασμό και υλοποίηση ενός πλαισίου δοκιμών (evaluation framework). • Δημιουργία συνόλου επιθέσεων (attack suite) με διαφορετικά επίπεδα πολυπλοκότητας. • Επιλογή και αξιολόγηση διαφορετικών LLMs (π.χ. open-source και commercial APIs). • Ορισμό μετρικών αξιολόγησης (π.χ. attack success rate, robustness score). • Πειραματική σύγκριση των μοντέλων και ανάλυση αποτελεσμάτων. • Προτάσεις για βελτίωση της ασφάλειας και ανθεκτικότητας των LLMs. 	
Σχετιζόμενα μαθήματα <ul style="list-style-type: none"> • Ασφάλεια συστημάτων υπολογιστών • Κρυπτογραφία • Τεχνητή νοημοσύνη 	

Προτεινόμενη μεθοδολογία έρευνας

Η εργασία θα βασιστεί στη Design Science Research Methodology (DSRM) και θα περιλαμβάνει: <ul style="list-style-type: none"> • Βιβλιογραφική ανασκόπηση σε θέματα prompt injection και LLM robustness. • Καταγραφή και ταξινόμηση επιθέσεων (attack taxonomy). • Σχεδιασμό και υλοποίηση ενός framework για testing. • Ανάπτυξη αυτοματοποιημένων σεναρίων επίθεσης. • Εκτέλεση πειραμάτων σε διαφορετικά μοντέλα. • Συλλογή και ανάλυση δεδομένων με ποσοτικές μετρικές. • Εξαγωγή συμπερασμάτων και προτάσεις για βελτιώσεις.
--

Προσδοκώμενα αποτελέσματα

<ul style="list-style-type: none"> • Δημιουργία ενός επαναχρησιμοποιήσιμου πλαισίου αξιολόγησης LLMs. • Συγκριτική ανάλυση της ανθεκτικότητας διαφορετικών μοντέλων. • Κατανόηση των πιο αποτελεσματικών τύπων επιθέσεων. • Προτάσεις για ασφαλέστερο σχεδιασμό LLM-based συστημάτων.

Ενδεικτικές πηγές

<ul style="list-style-type: none"> • S. Perez et al., "Prompt Injection Attacks and Defenses in Large Language Models," IEEE Security & Privacy Magazine, 2024. • Greshake, Kai, et al. "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection." Proceedings of the 16th ACM workshop on artificial intelligence and security. 2023. • J. Wei et al., "Protecting Large Language Models from Prompt Injection," USENIX Security Symposium, 2025. • DeBenedetti, Edoardo, et al. "Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents." Advances in Neural Information Processing Systems 37 (2024): 82895-82920.
--