

Θέμα: Ασφάλεια πρακτόρων μηχανικής μάθησης που βασίζονται σε μεγάλα γλωσσικά μοντέλα	
Επιβλέπων: Γεώργιος Σπαθούλας	Στοιχεία επικοινωνίας: gspathoulas@uth.gr
Σκοπός και στόχοι <p>Η παρούσα πτυχιακή εργασία αποσκοπεί στη μελέτη και ανάλυση των ζητημάτων ασφάλειας και ιδιωτικότητας που προκύπτουν από τη χρήση πρακτόρων (agents) βασισμένων σε μεγάλα γλωσσικά μοντέλα (Large Language Models – LLMs), όπως τα GPT, Claude ή Gemini. Στόχος είναι η κατανόηση των βασικών επιπέδων απειλών, η διερεύνηση ευπαθειών (prompt injection, data exfiltration, model manipulation) και η ανάπτυξη ή αξιολόγηση τεχνικών ενίσχυσης της ασφάλειας αυτών των πρακτόρων.</p>	
Αντικείμενο <p>Οι πράκτορες μηχανικής μάθησης που αξιοποιούν LLMs μπορούν να εκτελούν πολύπλοκες εργασίες, να αλληλεπιδρούν με APIs, βάσεις δεδομένων και περιβάλλοντα λογισμικού, καθώς και να λαμβάνουν αυτόνομες αποφάσεις. Ωστόσο, η αυξανόμενη πολυπλοκότητα των αλληλεπιδράσεών τους δημιουργεί νέες επιφάνειες επίθεσης, καθώς οι εισροές φυσικής γλώσσας μπορούν να χρησιμοποιηθούν για να χειραγωγήσουν ή να παρακάμψουν τους μηχανισμούς ασφαλείας. Η εργασία εστιάζει σε τεχνικές όπως:</p> <ul style="list-style-type: none">• Prompt Injection Attacks (ενσωμάτωση κακόβουλων οδηγιών σε εισροές),• Jailbreaking (παράκαμψη πολιτικών ασφαλείας του LLM),• Data Leakage / Exfiltration (αποκάλυψη εμπιστευτικών πληροφοριών μέσω διαλόγου),• Model Exploitation (κατάχρηση της συμπεριφοράς του μοντέλου για ανεπιθύμητες ενέργειες). <p>Θα διερευνηθούν επίσης προσεγγίσεις ανίχνευσης και αποτροπής αυτών των επιθέσεων, όπως η ανάλυση προτροπών (prompt sanitization), τα φίλτρα εισόδου-εξόδου (input/output filters) και η χρήση μοντέλων εποπτείας (guard models).</p>	
Η εργασία περιλαμβάνει <ul style="list-style-type: none">• Ανάλυση των βασικών τύπων ευπαθειών και επιθέσεων σε LLM agents.• Επισκόπηση σύγχρονων ερευνητικών εργασιών στον χώρο του LLM security.• Σχεδιασμό και υλοποίηση απλών σεναρίων επίθεσης (π.χ. prompt injection, jailbreaking).• Ανάπτυξη ή αξιολόγηση τεχνικών άμυνας (prompt filtering, policy enforcement).• Πειραματική μελέτη της αποτελεσματικότητας των προτεινόμενων μέτρων.• Συζήτηση των προκλήσεων ασφάλειας στην ανάπτυξη αυτόνομων πρακτόρων TN.	
Σχετιζόμενα μαθήματα <ul style="list-style-type: none">• Ασφάλεια συστημάτων υπολογιστών• Κρυπτογραφία• Τεχνητή νοημοσύνη	

Προτεινόμενη μεθοδολογία έρευνας

<p>Η εργασία θα βασιστεί στη Design Science Research Methodology (DSRM) και θα περιλαμβάνει:</p> <ul style="list-style-type: none">• Βιβλιογραφική ανασκόπηση σχετικά με LLM security, επιθέσεις prompt injection και τεχνικές μετριασμού.• Ανάλυση και ταξινόμηση των πιθανών απειλών σε LLM-based agents (threat modeling).• Σχεδιασμό και υλοποίηση πρωτοτύπων σεναρίων επιθέσεων και μηχανισμών ανίχνευσης/προστασίας.• Αξιολόγηση της αποτελεσματικότητας των αμυντικών μέτρων με ποσοτικούς δείκτες.• Εξαγωγή συμπερασμάτων και προτάσεις για βελτιστοποίηση πολιτικών ασφαλείας LLM agents.
--

Προσδοκώμενα αποτελέσματα

<ul style="list-style-type: none">• Κατανόηση των τύπων επιθέσεων και κινδύνων σε πράκτορες βασισμένους σε LLM.• Δημιουργία ή αξιολόγηση τεχνικών ενίσχυσης ασφάλειας σε LLM pipelines.• Ανάπτυξη ενός πειραματικού πλαισίου για προσομοίωση και μελέτη επιθέσεων.• Συμβολή στη βελτίωση της ασφάλειας και αξιοπιστίας των συστημάτων τεχνητής νοημοσύνης νέας γενιάς.

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.• A. Zou et al., "Universal and Transferable Adversarial Attacks on Aligned Language Models," arXiv preprint arXiv:2401.10968, 2024• OpenAI, "Model Spec and Safety Classifications for LLM Agents," OpenAI Technical Report, 2025.• J. Wei et al., "Protecting Large Language Models from Prompt Injection," USENIX Security Symposium, 2025.
