ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

«Ανάλυση και αξιολόγηση διαφορετικών γλωσσικών μοντέλων σε πολλαπλά σενάρια χρήσης»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Γεώργιου Καπριδάκη

Επιβλέπουσα: Διαμαντοπούλου Βασιλική, Μόνιμη Επίκουρη Καθηγήτρια
Μέλη εξεταστικής επιτροπής: Καβαλλιεράτος Γεώργιος, Ερευνητής
Μαλιάτσος Κωνσταντίνος, Μόνιμος Επίκουρος Καθηγητής

Σάμος, Μάρτιος 2026

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πρόλογος και ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσοι συνέβαλαν στην ολοκλήρωση της παρούσας διπλωματικής εργασίας.

Αρχικά, απευθύνω ιδιαίτερες ευχαριστίες στην επιβλέπουσα καθηγήτρια μου, κυρία Διαμαντοπούλου Βασιλική και τον κύριο Καβαλλιεράτο Γεώργιο για την πολύτιμη καθοδήγηση, τις ουσιαστικές συμβουλές και την αμέριστη υποστήριξη καθ' όλη τη διάρκεια της εκπόνησης της εργασίας. Η συμβολή τους υπήρξε καθοριστική τόσο στην επιστημονική προσέγγιση όσο και στην ανάπτυξη των ιδεών που παρουσιάζονται.

Ιδιαίτερη μνεία αξίζει στην οικογένειά μου για την αμέριστη στήριξη, την υπομονή και την ενθάρρυνση που μου προσέφεραν σε όλη τη διάρκεια των σπουδών μου. Χωρίς τη δική τους βοήθεια και κατανόηση, η ολοκλήρωση αυτής της προσπάθειας θα ήταν πολύ πιο δύσκολη.

Τέλος, ευχαριστώ τους φίλους μου για την ηθική υποστήριξη και την εμψύχωση που μου προσέφεραν σε κάθε στάδιο της διαδικασίας.

© 2026

του

Καπριδάκη Γεώργιου

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1 Εισαγωγή 1

1.1 Περιγραφή του Προβλήματος και Στόχοι της Διπλωματικής 1

1.2 Σημασία της Αξιοπιστίας στην Τεχνητή Νοημοσύνη 2

1.3 Σύντομη Περιγραφή των Γλωσσικών Μοντέλων και της εμπλοκής τους με την Αξιόπιστη ΤΝ 2

1.4 Παρουσίαση των βασικών πλαισίων αξιολόγησης 3

1.5 Διάρθρωση της Εργασίας 4

2 Θεωρητικό Υπόβαθρο 6

2.1 Γλωσσικά Μοντέλα 6

2.1.1 Εισαγωγή-Βασικοί Τύποι 6

2.1.2 Αναδυόμενες ικανότητες 9

2.1.3 Κύριες Τεχνικές 11

2.2 Έμπιστή Τεχνητή Νοημοσύνη 13

2.2.1 Εισαγωγή 13

2.2.2 Βασικές Έννοιες Έμπιστης Τεχνητής Νοημοσύνης 14

2.2.3 TAI και Διατυπωμένη Εμπιστοσύνη 17

2.2.4 TAI και Αυθεντική Εμπιστοσύνη 18

2.2.5 TAI και Βασική Εμπιστοσύνη 19

2.3 Τομείς Υψηλής Κρισιμότητας σύμφωνα με την Οδηγία NIS2 21

3 Σχετική Έρευνα και Εργαλεία Αξιολόγησης 24

3.1 Παρουσίαση και Ανάλυση της Υπάρχουσας Βιβλιογραφίας και των Σχετικών Surveys για την Trustworthy AI 24

3.1.1 Θεωρητικά Πλαίσια και Αρχές 24

3.1.2 Εργαλεία Επεξηγησιμότητας 25

3.1.3 Σχολιασμός των Κριτηρίων και Μεθόδων Αξιολόγησης από τις Έρευνες 25

3.2 AI Risk Management Framework 26

3.2.1 Τι είναι το AI RMF της NIST 26

3.2.2 Κατηγορίες κινδύνων 27

3.2.3 Χαρακτηριστικά Αξιόπιστων Συστημάτων Τεχνητής Νοημοσύνης 27

3.2.4 Κύριες Λειτουργίες 27

4 Γλωσσικά Μοντέλα που θα Αξιολογηθούν 29

4.1 GPT -4 29

4.1.1 Εισαγωγή 29

4.1.2 Δυνατότητες *29*
4.1.3 Τρόπος Λειτουργίας *30*
4.1.4 Πεδίο εφαρμογής *31*
4.2 Gemini 1.5 Flash 32
4.2.1 Εισαγωγή *32*
4.2.2 Χαρακτηριστικά *32*
4.2.3 Πεδίο εφαρμογής *34*
5 Αξιολόγηση Μοντέλων 35
5.1 Κατηγορία 1: Εξειδικευμένη Συμβουλευτική 35
5.1.1 Ερώτηση 1: Τι είναι το μοντέλο "Gradient Boosting" στη μηχανική μάθηση; *35*
5.1.2 Ερώτηση 2: Πώς λειτουργούν τα ηλιακά πάνελ; *36*
5.1.3 Ερώτηση 3: Τι σημαίνει "πληθωρισμός" και ποια είναι τα αίτιά του; *37*
5.1.4 Ερώτηση 4: Ποια είναι τα συμπτώματα της έλλειψης βιταμίνης D; *38*
5.2 Κατηγορία 2: Δημιουργικότητα 40
5.2.1 Ερώτηση 1: Δημιούργησε μια σύντομη ιστορία για έναν ταξιδιώτη που ανακαλύπτει έναν χαμένο πολιτισμό. *40*
5.2.2 Ερώτηση 2: Δημιούργησε στίχους για ένα τραγούδι εμπνευσμένο από τον χειμώνα *41*
5.3 Κατηγορία 3: Επεξήγηση και Διάγνωση Προβλημάτων 41
5.3.1 Ερώτηση 1: Ποια είναι τα πλεονεκτήματα και μειονεκτήματα της χρήσης μεγάλων συνόλων δεδομένων για εκπαίδευση AI μοντέλων; *41*
5.3.2 Ερώτηση 2: Γιατί ένας υπολογιστής μπορεί να λειτουργεί αργά; *42*
5.3.3 Ερώτηση 3: Ποια είναι η διαφορά μεταξύ κβαντικών υπολογιστών και κλασικών υπολογιστών; 43
5.4 Αξιολόγηση Γλωσσικών Μοντέλων βάσει Τομέων Υψηλής Κρισιμότητας (NIS2) 45
5.4.1 Τομέας Ενέργειας (Energy) *45*
5.4.2 Τομέας Μεταφορών (Transport) *47*
5.4.3 Τομέας Τραπεζικών Υπηρεσιών (Banking) *48*
5.4.4 Χρηματοπιστωτικές Υποδομές *50*
5.4.5 Υγεία *52*
5.4.6 Πόσιμο νερό *53*
5.4.7 Λύματα *55*
5.4.8 Ψηφιακές Υποδομές *57*
5.4.9 Υπηρεσίες ΤΠΕ *59*
5.4.10 Δημόσια Διοίκηση *60*
5.4.11 Διάστημα *62*
6 ΑνάλυσηΑπόδοσης Μοντέλων 65

6.1 Εισαγωγή 65
6.2 Απόδοση Μοντέλων στις Κατηγορίες Χρήσης 65
6.2.1 Εξειδικευμένη Συμβουλευτική 65
6.2.2 Δημιουργικότητα 66
6.2.3 Επεξήγηση και Διάγνωση Προβλημάτων 66
6.2.4 Συγκριτική Ανάλυση 66
6.3 Συγκριτική Ανάλυση της Απόδοσης των Γλωσσικών Μοντέλων στο Πλαίσιο των Απαιτήσεων της Οδηγίας NIS2 67
7 Συμπεράσματα 69
Βιβλιογραφία 71
Παράρτημα-Screenshots από τις απαντήσεις των 2 LLM's βάσει των 11 Τομέων Υψηλής Κρισιμότητας 74

Λίστα Εικόνων

Εικόνα 1: Κύρια στοιχεία AI RMF.....4

Εικόνα 2: Γενικό Πλαίσιο της Έμπιστης Τεχνητής Νοημοσύνης (Πηγή: https://www.semanticscholar.org/paper/Survey-of-Trustworthy-AI%3A-A-Meta-Decision-of-AI-Wu-Li/5bfc22ae81d47e70fc71b02c84d907c3381975ea/figure/2).....16

Εικόνα 3: Αρχιτεκτονική GPT-4 [33].....31

Εικόνα 4: Συγκριτικά στοιχεία μέσων χρόνων [36].....33

Λίστα Πινάκων

Πίνακας 1: Τομείς Υψηλής Κρισιμότητας της Οδηγίας NIS2.....23
Πίνακας 2: Συγκριτικά συγκεντρωτικά στοιχεία μοντέλων.....67
Πίνακας 3: Συγκριτικός Πίνακας Απόδοσης.....68

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με την αξιολόγηση της αξιοπιστίας των γλωσσικών μοντέλων στην τεχνητή νοημοσύνη (ΤΝ), εστιάζοντας στα GPT-4 και Gemini Flash 1.5. Στο πρώτο κεφάλαιο εισάγεται το πρόβλημα, οι στόχοι της εργασίας και η σημασία της αξιοπιστίας στην ΤΝ, με ιδιαίτερη έμφαση στις αρχές και τα πλαίσια που τη διέπουν. Το δεύτερο κεφάλαιο παρουσιάζει το θεωρητικό υπόβαθρο, περιγράφοντας τα γλωσσικά μοντέλα, τις αναδυόμενες ικανότητές τους, τις κύριες τεχνικές και τη σύνδεση με την έμπιστη τεχνητή νοημοσύνη, καθώς και τις σχετικές ηθικές θεωρίες.

Στο τρίτο κεφάλαιο αναλύεται η υπάρχουσα βιβλιογραφία και τα εργαλεία αξιολόγησης, περιλαμβάνοντας το AI Risk Management Framework του NIST. Εξετάζονται οι κατηγορίες κινδύνων και τα χαρακτηριστικά αξιόπιστων συστημάτων ΤΝ, με έμφαση στις κύριες λειτουργίες τους. Το τέταρτο κεφάλαιο εστιάζει στα γλωσσικά μοντέλα GPT-4 και Gemini Flash 1.5, περιγράφοντας τις δυνατότητές τους, τον τρόπο λειτουργίας τους και τα πεδία εφαρμογής τους.Η αξιολόγηση των μοντέλων πραγματοποιείται στο πέμπτο κεφάλαιο, όπου εξετάζεται η απόδοσή τους σε τρεις κατηγορίες: εξειδικευμένη συμβουλευτική, δημιουργικότητα και επεξήγηση προβλημάτων. Στο έκτο κεφάλαιο παρουσιάζεται η ανάλυση της απόδοσης των μοντέλων, προσφέροντας συγκριτική αξιολόγηση στις κατηγορίες χρήσης. Η εργασία ολοκληρώνεται με τη βιβλιογραφία που υποστηρίζει τη θεωρητική και πρακτική βάση της μελέτης.

Συνολικά, η εργασία προσφέρει μια ολοκληρωμένη προσέγγιση για την κατανόηση της αξιοπιστίας στην ΤΝ, αξιολογώντας πρακτικά παραδείγματα και συνδέοντας τα αποτελέσματα με ηθικές και τεχνικές διαστάσεις. Τα ευρήματα συμβάλλουν στη συνεχιζόμενη έρευνα για την ανάπτυξη έμπιστων συστημάτων ΤΝ, αναδεικνύοντας τις προοπτικές βελτίωσης στην επεξηγησιμότητα, τη διαφάνεια και τη διαχείριση προκαταλήψεων.

Λέξεις Κλειδιά: *Τεχνητή Νοημοσύνη, Γλωσσικά Μοντέλα, Αξιοπιστία, Έμπιστη ΤΝ, Αξιολόγηση Απόδοσης, Ηθική της ΤΝ*

Abstract

This thesis focuses on evaluating the reliability of language models in artificial intelligence (AI), with a particular emphasis on GPT-4 and Gemini Flash 1.5. The first chapter introduces the problem, the objectives of the study, and the importance of reliability in AI, with particular attention to the principles and frameworks that govern it. The second chapter presents the theoretical background, describing language models, their emerging capabilities, key techniques, and their connection to trustworthy AI, as well as relevant ethical theories.

The third chapter analyzes the existing literature and evaluation tools, including the NIST AI Risk Management Framework. It examines risk categories and the characteristics of reliable AI systems, with an emphasis on their core functions. The fourth chapter focuses on the GPT-4 and Gemini Flash 1.5 language models, describing their capabilities, functionality, and areas of application.

The evaluation of the models is conducted in the fifth chapter, examining their performance across three categories: specialized consulting, creativity, and problem explanation. The sixth chapter presents the analysis of the models' performance, offering a comparative evaluation within the usage categories. The thesis concludes with the bibliography that supports the theoretical and practical foundation of the study.

Overall, this thesis provides a comprehensive approach to understanding reliability in AI, evaluating practical examples, and connecting the findings to ethical and technical dimensions. The results contribute to ongoing research on the development of trustworthy AI systems, highlighting opportunities for improvement in explainability, transparency, and bias management.

Keywords: *Artificial Intelligence, Language Models, Reliability, Trustworthy AI, Performance Evaluation, AI Ethics*

1

Εισαγωγή

1.1 Περιγραφή του Προβλήματος και Στόχοι της Διπλωματικής

Καθώς τα συστήματα τεχνητής νοημοσύνης (ΤΝ), ιδιαίτερα τα γλωσσικά μοντέλα, ενσωματώνονται ολοένα και περισσότερο σε διάφορες εφαρμογές της κοινωνίας — από εικονικούς βοηθούς μέχρι την παραγωγή περιεχομένου και την αυτόματη μετάφραση — ανακύπτει η ανάγκη για αξιόπιστα και δίκαια συστήματα. Ωστόσο, η ταχεία ανάπτυξη και διάδοση τέτοιων μοντέλων επιφέρει εγγενείς προκλήσεις. Τα γλωσσικά μοντέλα, όπως το GPT, το BERT και το T5, είναι γνωστό ότι παράγουν συχνά προκατειλημμένο, λανθασμένο ή ακόμη και επιβλαβές περιεχόμενο. Αυτά τα ζητήματα αναδεικνύουν ένα κρίσιμο κενό στην ανάγκη για ασφάλεια, υπευθυνότητα και ευθυγράμμιση των συστημάτων ΤΝ με ανθρώπινες αξίες.

Ο στόχος αυτής της διπλωματικής εργασίας είναι η αξιολόγηση και ανάλυση γλωσσικών μοντέλων μέσα από το πρίσμα της «Αξιόπιστης ΤΝ». Εφαρμόζοντας αναγνωρισμένα πλαίσια όπως το AI Risk Management Framework (AI RMF) του National Institute of Standards and Technology (NIST), η μελέτη αποσκοπεί στο να καθορίσει σε ποιο βαθμό αυτά τα μοντέλα πληρούν τα βασικά κριτήρια για αξιοπιστία, όπως η δικαιοσύνη, η διαφάνεια, η ασφάλεια και η υπευθυνότητα. Η εργασία επίσης επιδιώκει να εντοπίσει πιθανές αδυναμίες στα τρέχοντα γλωσσικά μοντέλα και να προτείνει βελτιώσεις για την ενίσχυση της συνολικής αξιοπιστίας τους. Ποιο το επίπεδο ακρίβειας των αποτελεσμάτων των γλωσσικών μοντέλων και πόσο αξιόπιστες είναι οι πληροφορίες;

Τα γλωσσικά μοντέλα αντιμετωπίζουν σημαντικές αδυναμίες που σχετίζονται με προκατάληψη, διαφάνεια, ακρίβεια και ασφάλεια. Πώς επηρεάζει ο τύπος των ερωτήσεων την απόδοση των γλωσσικών μοντέλων; Επειδή εκπαιδεύονται σε τεράστιες ποσότητες δεδομένων, συχνά ενσωματώνουν και αναπαράγουν κοινωνικές προκαταλήψεις, γεγονός που μπορεί να οδηγήσει σε μεροληπτικές ή ανισόρροπες απαντήσεις. Επιπλέον, η λειτουργία τους ως "μαύρα κουτιά"

δυσκολεύει την επεξήγηση των αποφάσεών τους, ενώ η τάση τους να δημιουργούν ανακριβείς ή επινοημένες πληροφορίες (hallucinations) τα καθιστά επικίνδυνα για κρίσιμες εφαρμογές, όπως η ιατρική και η νομική. Παράλληλα, υπάρχει ο κίνδυνος χρήσης τους για τη διάδοση παραπληροφόρησης, τη δημιουργία επιβλαβούς περιεχομένου ή τη διενέργεια κακόβουλων επιθέσεων που εκμεταλλεύονται τις αδυναμίες τους. Τα αποτελέσματα των μοντέλων πόσο αξιόπιστα μπορεί να είναι;

Εκτός από τα ζητήματα δικαιοσύνης και ασφάλειας, τα γλωσσικά μοντέλα παρουσιάζουν προκλήσεις ως προς την ευθυγράμμισή τους με ανθρώπινες αξίες και τη βιωσιμότητα της λειτουργίας τους. Σε ποιο βαθμό τα μοντέλα ευθυγραμμίζονται με τις ανθρώπινες αξίες; Δεν έχουν αληθινή κατανόηση της γλώσσας αλλά στηρίζονται σε στατιστικές συσχετίσεις, γεγονός που μπορεί να οδηγήσει σε ασυνέπειες ή ανεπιθύμητες απαντήσεις. Επιπλέον, η εκπαίδευση και η λειτουργία τους απαιτούν τεράστια υπολογιστική ισχύ, με σημαντικό περιβαλλοντικό αντίκτυπο. Αυτές οι προκλήσεις καθιστούν αναγκαία τη συνεχή έρευνα και ανάπτυξη για τη βελτίωση της διαφάνειας, της δικαιοσύνης και της συνολικής αξιοπιστίας των συστημάτων τεχνητής νοημοσύνης.

1.2 Σημασία της Αξιοπιστίας στην Τεχνητή Νοημοσύνη

Η αξιοπιστία στην ΤΝ είναι κρίσιμη, καθώς αποτελεί τη βάση για την εμπιστοσύνη των χρηστών και την ευρύτερη κοινωνική αποδοχή. Ακρίβειες ή απρόβλεπτες προκαταλήψεις στα αποτελέσματα ΤΝ μπορούν να έχουν σημαντικές συνέπειες, ιδιαίτερα σε ευαίσθητους τομείς, όπως η υγειονομική περίθαλψη, τα οικονομικά και ο τομέας της δικαιοσύνης. Για παράδειγμα, εάν ένα γλωσσικό μοντέλο ΤΝ που χρησιμοποιείται για διαγνωστικούς σκοπούς παρέχει λανθασμένες πληροφορίες, θα μπορούσε να οδηγήσει σε κακές αποφάσεις με πιθανές επιπτώσεις για τους ασθενείς. Παρομοίως, ένα προκατειλημμένο μοντέλο σε μια εφαρμογή έγκρισης δανείων θα μπορούσε να συμβάλλει σε κοινωνικές ανισότητες, αδικώντας συγκεκριμένες κοινωνικές ομάδες.

Καθώς τα γλωσσικά μοντέλα συνεχίζουν να διαμορφώνουν τις αλληλεπιδράσεις ανθρώπου-μηχανής, η αξιοπιστία τους γίνεται ηθική, νομική και τεχνική ανησυχία. Τα αξιόπιστα συστήματα ΤΝ όχι μόνο ενισχύουν την εμπιστοσύνη, αλλά και ευθυγραμμίζονται με τις ρυθμιστικές προσδοκίες και τις κοινωνικές αξίες. Επιπλέον, η διασφάλιση ότι τα συστήματα ΤΝ σχεδιάζονται για να είναι διαφανή και υπεύθυνα συμβάλλει στη δημιουργία συστημάτων που οι άνθρωποι αισθάνονται άνετα να χρησιμοποιούν. Η Αξιόπιστη ΤΝ στοχεύει στην αντιμετώπιση αυτών των αναγκών, ενσωματώνοντας αρχές που βοηθούν στη μείωση των κινδύνων και στην ενίσχυση της εμπιστοσύνης των χρηστών.

1.3 Σύντομη Περιγραφή των Γλωσσικών Μοντέλων και της εμπλοκής τους με την Αξιόπιστη ΤΝ

Τα γλωσσικά μοντέλα είναι υπολογιστικά συστήματα που επεξεργάζονται και παράγουν ανθρώπινη γλώσσα. Εκπαιδεύονται σε τεράστια σύνολα δεδομένων και μπορούν να δημιουργούν κείμενο που εμφανίζεται συνεκτικό, κατάλληλο για το περιεχόμενο και συχνά δυσδιάκριτο από ανθρώπινο κείμενο. Αυτή η ικανότητα έχει μεταμορφώσει βιομηχανίες και εφαρμογές, επιτρέποντας προόδους σε τομείς όπως η συνομιλιακή ΤΝ, η αυτόματη συγγραφή και η πολυγλωσσική μετάφραση.

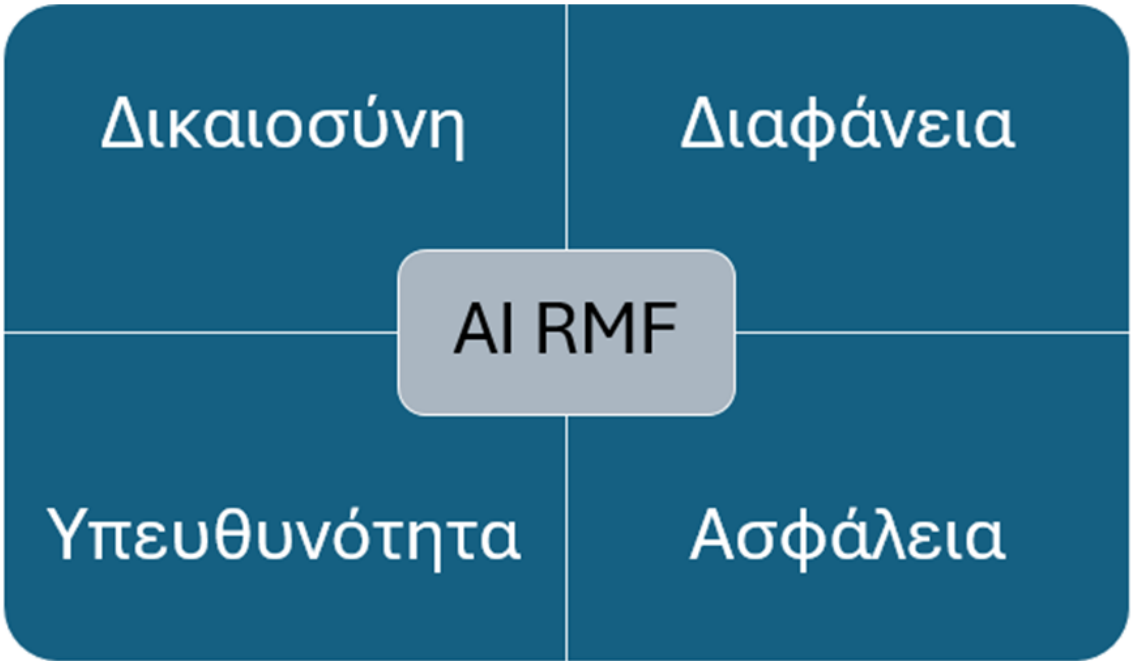
Παρά τις επιτυχίες αυτές, τα γλωσσικά μοντέλα δημιουργούν μοναδικές προκλήσεις σε ό,τι αφορά την αξιοπιστία. Μπορεί να παράγουν προκατειλημμένο ή επιβλαβές περιεχόμενο, να παρουσιάζουν έλλειψη διαφάνειας στη διαδικασία λήψης αποφάσεων και να είναι ευάλωτα σε κακόβουλη παρέμβαση. Αυτά τα ζητήματα περιπλέκουν τις προσπάθειες για καθιέρωση εμπιστοσύνης στα συστήματα ΤΝ. Η αντιμετώπιση αυτών των περιορισμών είναι κρίσιμη για την προώθηση της Αξιόπιστης ΤΝ, η οποία δίνει προτεραιότητα σε ηθικές σκέψεις παράλληλα με την τεχνική απόδοση.

Τα τελευταία χρόνια, παρατηρείται αυξανόμενο ενδιαφέρον για την ανάπτυξη προτύπων και πλαισίων για την Αξιόπιστη ΤΝ, με το NIST AI RMF να αναδεικνύεται ως ένα από τα κορυφαία πρότυπα. Το πλαίσιο αυτό στοχεύει στο να βοηθήσει τους προγραμματιστές να αξιολογούν τα συστήματα ΤΝ με κριτήρια όπως η δικαιοσύνη, η διαφάνεια, η υπευθυνότητα και η ασφάλεια. Εφαρμόζοντας αυτά τα πρότυπα στα γλωσσικά μοντέλα, μπορούμε να αποκτήσουμε πολύτιμες γνώσεις σχετικά με την ευθυγράμμισή τους με τις αρχές της Αξιόπιστης ΤΝ και να εντοπίσουμε περιοχές που χρειάζονται βελτίωση.

1.4 Παρουσίαση των βασικών πλαισίων αξιολόγησης

Για να αξιολογηθούν τα γλωσσικά μοντέλα ως προς την αξιοπιστία τους, αυτή η διπλωματική χρησιμοποιεί το AI Risk Management Framework (AI RMF) του National Institute of Standards and Technology. Το AI RMF έχει σχεδιαστεί ώστε να προσφέρει μια δομημένη προσέγγιση για τον εντοπισμό, την αξιολόγηση και τη μείωση των κινδύνων στα συστήματα ΤΝ, με έμφαση σε παράγοντες που επηρεάζουν την εμπιστοσύνη των χρηστών. Κύρια στοιχεία του AI RMF περιλαμβάνουν:

- **Δικαιοσύνη**: Διασφάλιση ότι τα αποτελέσματα του μοντέλου δεν εμφανίζουν συστηματική προκατάληψη κατά συγκεκριμένων ομάδων.
- **Διαφάνεια**: Η κατανόηση της διαδικασίας λήψης αποφάσεων του μοντέλου από τους προγραμματιστές και τους τελικούς χρήστες.
- **Υπευθυνότητα**: Καθιέρωση μηχανισμών που καθιστούν το μοντέλο και τους δημιουργούς του υπεύθυνους για τα αποτελέσματα και τις πιθανές επιπτώσεις.
- **Ασφάλεια**: Προστασία του μοντέλου από μη εξουσιοδοτημένη πρόσβαση ή κακόβουλη παρέμβαση, ιδιαίτερα σε κρίσιμες εφαρμογές.



Εικόνα 1: Κύρια στοιχεία AI RMF

Εκτός από το AI RMF, άλλα πλαίσια, όπως το NIST Cybersecurity Framework (CSF), προσφέρουν πολύτιμες γνώσεις, ειδικά όσον αφορά την ασφάλεια και την ανθεκτικότητα. Αυτή η διπλωματική διερευνά το πώς αυτά τα πλαίσια, συνολικά, προσφέρουν μια σταθερή βάση για την αξιολόγηση των γλωσσικών μοντέλων. Χρησιμοποιώντας αυτά τα πλαίσια, επιτυγχάνεται μια δομημένη και ολοκληρωμένη αξιολόγηση, βοηθώντας στη γεφύρωση του χάσματος μεταξύ τεχνικών δυνατοτήτων και αξιόπιστης και ηθικής ΤΝ.

1.5 Διάρθρωση της Εργασίας

Τα επόμενα κεφάλαια αυτής της διπλωματικής δομούνται ως εξής:

- **Κεφάλαιο Δεύτερο**: **Θεωρητικό Υπόβαθρο** - Το κεφάλαιο αυτό εισάγει τα γλωσσικά μοντέλα και την εξέλιξή τους με την πάροδο του χρόνου, παρουσιάζοντας τα βασικά στοιχεία της Αξιόπιστης ΤΝ. Επίσης, παρουσιάζει τις θεμελιώδεις αρχές των πλαισίων NIST AI RMF και Cybersecurity Framework, αναλύοντας πώς αυτά επηρεάζουν την αξιολόγηση των συστημάτων ΤΝ.
- **Κεφάλαιο Τρίτο**: **Σχετική Έρευνα και Εργαλεία Αξιολόγησης** - Αυτή η ενότητα εξετάζει τη σχετική βιβλιογραφία για την Αξιόπιστη ΤΝ και αναλύει κριτήρια και μεθόδους αξιολόγησης. Αναπτύσσει συνδέσεις μεταξύ θεωρητικών αρχών και πρακτικών πλαισίων, προετοιμάζοντας το έδαφος για την αξιολόγηση των γλωσσικών μοντέλων.
- **Κεφάλαιο Τέταρτο**: **Γλωσσικά Μοντέλα που Αναλύονται** - Εξετάζονται συγκεκριμένα γλωσσικά μοντέλα (π.χ., GPT, BERT, T5), με περιγραφή της αρχιτεκτονικής και των βασικών χαρακτηριστικών τους. Επιπλέον, αναλύονται οι προκλήσεις αξιοπιστίας που αντιμετωπίζουν αυτά τα μοντέλα.
- **Κεφάλαιο Πέμπτο**: **Ανάλυση και Αξιολόγηση σε Πολλαπλά Σενάρια Χρήσης** - Παρουσιάζονται διαφορετικά σενάρια χρήσης, όπως η κατανόηση φυσικής γλώσσας, η

παραγωγή κειμένου, η μετάφραση και η συνομιλία. Αξιολογείται η απόδοση των μοντέλων με βάση τα πλαίσια του NIST και συγκρίνεται η απόδοση τους σε κριτήρια όπως η δικαιοσύνη, η ασφάλεια και η διαφάνεια.

- **Κεφάλαιο Έκτο**: **Αποτελέσματα και Ευρήματα** - Παρουσιάζονται και αναλύονται τα αποτελέσματα της αξιολόγησης των μοντέλων. Εξετάζονται συγκριτικά τα μοντέλα ως προς την αξιοπιστία τους σε διαφορετικά κριτήρια.
- **Κεφάλαιο Έβδομο**: **Συμπεράσματα** - Το κεφάλαιο αυτό συνοψίζει τα κύρια συμπεράσματα και αποτιμά την απόδοση των γλωσσικών μοντέλων σύμφωνα με τα πρότυπα της Αξιόπιστης ΤΝ. Κλείνει με προτάσεις για μελλοντική έρευνα και ανάπτυξή.

2

Θεωρητικό Υπόβαθρο

2.1 Γλωσσικά Μοντέλα

2.1.1 Εισαγωγή-Βασικοί Τύποι

Η γλώσσα είναι μια εξέχουσα ικανότητα των ανθρώπων να εκφράζονται και να επικοινωνούν, η οποία αναπτύσσεται κατά την πρώιμη παιδική ηλικία και εξελίσσεται σε όλη τη διάρκεια της ζωής τους. Οι μηχανές, ωστόσο, δεν μπορούν να κατανοήσουν φυσικά τις ικανότητες κατανόησης και επικοινωνίας στη μορφή της ανθρώπινης γλώσσας, εκτός αν εξοπλιστούν με ισχυρούς αλγορίθμους τεχνητής νοημοσύνης (AI). Η επίτευξη αυτού του στόχου αποτελεί μακροχρόνια ερευνητική πρόκληση, για να μπορέσουν οι μηχανές να διαβάζουν, να γράφουν και να επικοινωνούν όπως οι άνθρωποι.

Τεχνικά, η γλωσσική μοντελοποίηση (LM) [1] είναι μία από τις κύριες προσεγγίσεις για την προαγωγή της γλωσσικής νοημοσύνης των μηχανών. Γενικά, η LM στοχεύει στη μοντελοποίηση της γενετικής πιθανότητας των ακολουθιών λέξεων, ώστε να προβλέπει τις πιθανότητες μελλοντικών (ή ελλειπόντων) στοιχείων. Η έρευνα για τη LM έχει λάβει εκτεταμένη προσοχή στη βιβλιογραφία, η οποία μπορεί να χωριστεί σε τέσσερα κύρια στάδια ανάπτυξης:

- **Στατιστικά γλωσσικά μοντέλα (SLM)**: Τα SLM [2] αναπτύχθηκαν με βάση τις στατιστικές μεθόδους μάθησης που προέκυψαν τη δεκαετία του 1990. Η βασική ιδέα είναι η κατασκευή ενός μοντέλου πρόβλεψης λέξεων με βάση την υπόθεση του Markov, δηλαδή την πρόβλεψη της επόμενης λέξης με βάση το πιο πρόσφατο πλαίσιο. Τα SLM με σταθερό μήκος πλαισίου n ονομάζονται επίσης n-gram γλωσσικά μοντέλα, όπως τα bigram και trigram γλωσσικά μοντέλα. Τα SLM έχουν εφαρμοστεί ευρέως για την ενίσχυση της απόδοσης σε εργασίες ανάκτησης πληροφορίας και επεξεργασίας φυσικής γλώσσας. Ωστόσο, συχνά υποφέρουν από την "κατάρα της διάστασης": είναι δύσκολο να εκτιμηθούν με ακρίβεια τα μοντέλα υψηλότερης τάξης, καθώς ένας εκθετικός αριθμός πιθανοτήτων μετάβασης πρέπει να εκτιμηθεί. Έτσι, εισήχθησαν ειδικά σχεδιασμένες στρατηγικές εξομάλυνσης για να αντιμετωπιστεί το πρόβλημα της έλλειψης δεδομένων.
- **Νευρωνικά γλωσσικά μοντέλα (NLM)**: Τα NLM [3] χαρακτηρίζουν την πιθανότητα ακολουθιών λέξεων μέσω νευρωνικών δικτύων, όπως τα πολυεπίπεδα αντιληπτικά και τα

επαναληπτικά νευρωνικά δίκτυα (RNN). Μια σημαντική συμβολή ήταν η εισαγωγή της έννοιας της κατανεμημένης αναπαράστασης των λέξεων και η κατασκευή της συνάρτησης πρόβλεψης λέξεων που εξαρτάται από τα συγκεντρωμένα χαρακτηριστικά του πλαισίου (δηλαδή, των κατανεμημένων διανυσμάτων λέξεων). Μέσω της επέκτασης αυτής της ιδέας, αναπτύχθηκε μια γενική προσέγγιση νευρωνικού δικτύου για τη δημιουργία μιας ενιαίας, ολοκληρωμένης λύσης για διάφορες εργασίες επεξεργασίας φυσικής γλώσσας. Επιπλέον, η λέξη 2vec προτάθηκε για να δημιουργήσει ένα απλό ρηχό νευρωνικό δίκτυο για την εκμάθηση κατανεμημένων αναπαραστάσεων λέξεων, οι οποίες αποδείχθηκαν πολύ αποτελεσματικές σε διάφορες εργασίες επεξεργασίας φυσικής γλώσσας. Αυτές οι μελέτες εισήγαγαν τη χρήση των γλωσσικών μοντέλων για μάθηση αναπαράστασης, με σημαντική επίδραση στον τομέα της επεξεργασίας φυσικής γλώσσας.

- **Προεκπαιδευμένα γλωσσικά μοντέλα (PLM):** Ως πρώιμη προσπάθεια, το ELMo προτάθηκε για τη σύλληψη των αναπαραστάσεων λέξεων με βάση τα συμφραζόμενα, εκπαιδεύοντας πρώτα ένα δίκτυο biLSTM και στη συνέχεια τροποποιώντας το δίκτυο biLSTM σύμφωνα με συγκεκριμένες καθοδικές εργασίες. Επιπλέον, με βάση την πολύ παραλληλοποιήσιμη αρχιτεκτονική Transformer με μηχανισμούς αυτοπροσοχής, το BERT προτάθηκε με την προεκπαίδευση αμφίδρομων γλωσσικών μοντέλων σε μεγάλης κλίμακας αδήλωτα σύνολα δεδομένων. Αυτές οι προεκπαιδευμένες αναπαραστάσεις λέξεων με βάση τα συμφραζόμενα είναι πολύ αποτελεσματικές ως χαρακτηριστικά γενικής χρήσης για την κατανόηση του νοήματος, αυξάνοντας σημαντικά την απόδοση των εργασιών επεξεργασίας φυσικής γλώσσας. Αυτή η μελέτη ενέπνευσε έναν μεγάλο αριθμό επακόλουθων εργασιών, καθορίζοντας το παράδειγμα μάθησης "προεκπαίδευση και τροποποίηση". Ακολουθώντας αυτό το παράδειγμα, αναπτύχθηκε ένας μεγάλος αριθμός μελετών για τα PLM [4], εισάγοντας είτε διαφορετικές αρχιτεκτονικές είτε βελτιωμένες στρατηγικές προεκπαίδευσης. Σε αυτό το παράδειγμα, απαιτείται συχνά η προσαρμογή του PLM για να προσαρμόζεται σε διαφορετικές καθοδικές εργασίες.
- **Μεγάλα γλωσσικά μοντέλα (LLM):** Οι ερευνητές διαπίστωσαν ότι η κλιμάκωση του PLM (π.χ., αύξηση του μεγέθους του μοντέλου ή του μεγέθους των δεδομένων) συχνά οδηγεί σε βελτιωμένη ικανότητα του μοντέλου σε καθοδικές εργασίες, ακολουθώντας τον νόμο κλιμάκωσης. Αρκετές μελέτες [5] έχουν εξερευνήσει το όριο απόδοσης εκπαιδεύοντας ένα ολοένα και μεγαλύτερο PLM. Παρόλο που η κλιμάκωση πραγματοποιείται κυρίως στο μέγεθος του μοντέλου, αυτά τα μεγάλα PLM παρουσιάζουν διαφορετικές συμπεριφορές από τα μικρότερα PLM και δείχνουν εκπληκτικές ικανότητες σε μια σειρά από σύνθετες εργασίες. Για παράδειγμα, το GPT-3 μπορεί να επιλύει εργασίες με λίγα παραδείγματα μέσω της μάθησης "in-context", ενώ το GPT-2 δεν τα καταφέρνει εξίσου καλά. Έτσι, η ερευνητική κοινότητα υιοθετεί τον όρο "μεγάλα γλωσσικά μοντέλα (LLM)" για αυτά τα μεγάλα PLM, τα οποία προσελκύουν αυξανόμενο ερευνητικό ενδιαφέρον. Μια αξιοσημείωτη εφαρμογή των LLM είναι το ChatGPT που προσαρμόζει τα LLM από τη σειρά GPT για διάλογο, επιδεικνύοντας εντυπωσιακές ικανότητες επικοινωνίας με τους ανθρώπους.

Η παρούσα εργασία εστιάζει στην ανάλυση των Pretrained Language Models (PLM) και Large Language Models (LLM), καθώς αυτά αποτελούν τις κυρίαρχες κατηγορίες γλωσσικών μοντέλων που αξιοποιούνται σε εφαρμογές μεγάλης κλίμακας και σχετίζονται άμεσα με ζητήματα αξιοπιστίας, διαφάνειας και προκατάληψης. Αντίθετα, τα Small Language Models (SLM) και τα

Neural Language Models (NLM), αν και έχουν τη δική τους σημασία στη γλωσσική επεξεργασία, δεν παρουσιάζουν τις ίδιες προκλήσεις και επιπτώσεις όσον αφορά τη χρήση τους σε ευρύτερα κοινωνικά και τεχνολογικά πλαίσια. Για τον λόγο αυτό, η μελέτη επικεντρώνεται στα PLM και LLM, τα οποία έχουν τη μεγαλύτερη επίδραση στη σύγχρονη τεχνητή νοημοσύνη

Η έννοια της κλιμάκωσης εφαρμόστηκε ειδικά για τα μεγάλα γλωσσικά μοντέλα (LLM), αλλά έχει εξελιχθεί με την πρόοδο της τεχνητής νοημοσύνης τις τελευταίες δεκαετίες. Τα πρώιμα γλωσσικά μοντέλα στοχεύουν κυρίως στη μοντελοποίηση και παραγωγή δεδομένων κειμένου, ενώ τα πιο σύγχρονα γλωσσικά μοντέλα (π.χ., GPT-4) εστιάζουν στην επίλυση σύνθετων εργασιών. Από τη μοντελοποίηση της γλώσσας στην επίλυση εργασιών, αυτό αποτελεί ένα σημαντικό άλμα στην επιστημονική σκέψη, το οποίο είναι το κλειδί για την κατανόηση της εξέλιξης των γλωσσικών μοντέλων στην ερευνητική ιστορία. Από την άποψη της επίλυσης εργασιών, οι τέσσερις γενιές γλωσσικών μοντέλων έχουν επιδείξει διαφορετικά επίπεδα δυνατοτήτων. Αρχικά, τα στατιστικά γλωσσικά μοντέλα βοήθησαν κυρίως σε ορισμένες συγκεκριμένες εργασίες (π.χ., ανάκτηση ή αναγνώριση ομιλίας), όπου οι προβλεπόμενες ή εκτιμώμενες πιθανότητες μπορούσαν να ενισχύσουν την απόδοση των προσεγγίσεων για συγκεκριμένες εργασίες. Στη συνέχεια, τα νευρωνικά γλωσσικά μοντέλα επικεντρώθηκαν στην εκμάθηση ανεξάρτητων χαρακτηριστικών από τις εργασίες (π.χ., χαρακτηριστικά), στοχεύοντας στη μείωση της ανάγκης για χειροκίνητη ανάλυση χαρακτηριστικών από ανθρώπους. Επιπλέον, τα προεκπαιδευμένα γλωσσικά μοντέλα έμαθαν αναπαραστάσεις που εξαρτώνται από τα συμφραζόμενα και μπορούν να βελτιστοποιηθούν ανάλογα με τις καθοδικές εργασίες. Στην τελευταία γενιά γλωσσικών μοντέλων, τα LLM ενισχύονται εξερευνώντας την επίδραση της κλιμάκωσης στη δυναμικότητα του μοντέλου, και μπορούν να θεωρηθούν ως καθολικοί επιλύτες εργασιών. Συνοψίζοντας, κατά τη διαδικασία εξέλιξης, το εύρος των εργασιών που μπορούν να επιλύσουν τα γλωσσικά μοντέλα έχει επεκταθεί σημαντικά και η απόδοση τους σε εργασίες έχει βελτιωθεί σημαντικά.

Στην υπάρχουσα βιβλιογραφία, τα προεκπαιδευμένα γλωσσικά μοντέλα (PLMs) έχουν συζητηθεί και ανασκοπηθεί εκτενώς, ενώ τα μεγάλα γλωσσικά μοντέλα (LLMs) σπάνια εξετάζονται συστηματικά. Για να εισαγάγουμε την ανασκόπησή μας, αρχικά τονίζουμε τρεις κύριες διαφορές μεταξύ LLM και PLM. Πρώτον, τα LLM παρουσιάζουν ορισμένες εκπληκτικές αναδυόμενες ικανότητες που μπορεί να μην παρατηρούνται σε μικρότερα PLM. Αυτές οι ικανότητες είναι κρίσιμες για την απόδοση των γλωσσικών μοντέλων σε σύνθετες εργασίες, καθιστώντας τους αλγορίθμους τεχνητής νοημοσύνης απροσδόκητα ισχυρούς και αποτελεσματικούς. Δεύτερον, τα LLM θα επαναπροσδιορίσουν τον τρόπο με τον οποίο οι άνθρωποι αναπτύσσουν και χρησιμοποιούν αλγορίθμους τεχνητής νοημοσύνης. Σε αντίθεση με τα μικρά PLM, η κύρια πρόσβαση στα LLM γίνεται μέσω διεπαφών εντολών (π.χ., API GPT-4). Οι άνθρωποι πρέπει να κατανοήσουν πώς λειτουργούν τα LLM και να διαμορφώσουν τις εργασίες τους με τρόπο που τα LLM να μπορούν να ακολουθήσουν. Τρίτον, η ανάπτυξη των LLM πλέον δεν ξεχωρίζει σαφώς την έρευνα από την μηχανική. Η εκπαίδευση των LLM απαιτεί εκτενή πρακτική εμπειρία στην επεξεργασία μεγάλων δεδομένων και στην κατανεμημένη παράλληλη εκπαίδευση. Για να αναπτύξουν ικανά LLM, οι ερευνητές πρέπει να επιλύουν πολύπλοκα μηχανικά ζητήματα, συνεργαζόμενοι με μηχανικούς ή ενεργώντας ως μηχανικοί.

Σήμερα, τα LLM ασκούν σημαντική επιρροή στην κοινότητα της τεχνητής νοημοσύνης, και η εμφάνιση του ChatGPT και του GPT-4 οδηγεί σε επανεξέταση των δυνατοτήτων της τεχνητής γενικής νοημοσύνης (AGI) [6]. Η OpenAI έχει δημοσιεύσει ένα τεχνικό άρθρο με τίτλο

"Σχεδιασμός για AGI και πέρα", το οποίο συζητά τα βραχυπρόθεσμα και μακροπρόθεσμα σχέδια για την προσέγγιση της AGI, ενώ μια πιο πρόσφατη μελέτη έχει υποστηρίξει ότι το GPT-4 μπορεί να θεωρηθεί ως πρώιμη έκδοση ενός συστήματος AGI. Οι τομείς της τεχνητής νοημοσύνης υφίστανται επαναστατικές αλλαγές λόγω της ταχείας προόδου των LLM. Στον τομέα της επεξεργασίας φυσικής γλώσσας, τα LLM μπορούν να χρησιμεύσουν ως καθολικοί επιλύτες γλωσσικών εργασιών (σε κάποιο βαθμό), και το ερευνητικό παράδειγμα έχει μετατοπιστεί προς τη χρήση των LLM. Στον τομέα της ανάκτησης πληροφορίας, οι παραδοσιακές μηχανές αναζήτησης αμφισβητούνται από τον νέο τρόπο αναζήτησης μέσω των chatbot τεχνητής νοημοσύνης (π.χ., ChatGPT), και το New Bing αποτελεί μια αρχική προσπάθεια ενίσχυσης των αποτελεσμάτων αναζήτησης βασισμένων σε LLM. Στον τομέα της όρασης υπολογιστών, οι ερευνητές προσπαθούν να αναπτύξουν μοντέλα όρασης-γλώσσας τύπου ChatGPT που μπορούν να εξυπηρετούν καλύτερα τους διαλόγους πολλαπλών μορφών, και το GPT-4 υποστηρίζει την πολυτροπική είσοδο ενσωματώνοντας τις οπτικές πληροφορίες. Αυτό το νέο κύμα τεχνολογίας θα μπορούσε ενδεχομένως να οδηγήσει σε μια ευημερούσα οικοσύστημα εφαρμογών βασισμένων σε LLM. Για παράδειγμα, το Microsoft 365 ενισχύεται από τα LLM (δηλαδή, Copilot) για την αυτοματοποίηση της γραφειακής εργασίας, και η OpenAI υποστηρίζει τη χρήση προσθηκών στο ChatGPT για την εφαρμογή ειδικών λειτουργιών.

Παρά την πρόοδο και την επιρροή, οι υποκείμενες αρχές των LLM εξακολουθούν να μην έχουν πλήρως εξερευνηθεί. Πρώτον, παραμένει μυστήριο γιατί εμφανίζονται αναδυόμενες ικανότητες στα LLM και όχι στα μικρότερα PLM. Ως γενικότερο ζήτημα, λείπει μια βαθιά, λεπτομερής διερεύνηση των βασικών παραγόντων που συμβάλλουν στις ανώτερες ικανότητες των LLM. Είναι σημαντικό να μελετηθεί πότε και πώς τα LLM αποκτούν τέτοιες ικανότητες. Αν και υπάρχουν κάποιες ουσιαστικές συζητήσεις για αυτό το πρόβλημα, χρειάζονται πιο αρχές έρευνες για να αποκαλυφθούν τα "μυστικά" των LLM. Δεύτερον, είναι δύσκολο για την ερευνητική κοινότητα να εκπαιδεύσει ικανά LLM. Λόγω της τεράστιας ζήτησης πόρων υπολογισμού, είναι πολύ δαπανηρό να πραγματοποιηθούν επαναληπτικές, αποδομητικές μελέτες για την εξέταση της επίδρασης διαφόρων στρατηγικών εκπαίδευσης των LLM. Πράγματι, τα LLM εκπαιδεύονται κυρίως από τη βιομηχανία, όπου πολλές σημαντικές λεπτομέρειες εκπαίδευσης (π.χ., συλλογή και καθαρισμός δεδομένων) δεν αποκαλύπτονται στο κοινό. Τρίτον, είναι προκλητικό να ευθυγραμμιστούν τα LLM με τις ανθρώπινες αξίες ή προτιμήσεις. Παρά τις δυνατότητες, τα LLM είναι επίσης πιθανό να παράγουν τοξικά, ψευδή ή επιβλαβή περιεχόμενα. Απαιτούνται αποτελεσματικές και αποδοτικές μέθοδοι ελέγχου για την εξάλειψη του δυνητικού κινδύνου χρήσης των LLM.

2.1.2 Αναδυόμενες ικανότητες

Στη βιβλιογραφία, οι αναδυόμενες ικανότητες [7] των LLMs ορίζονται επίσημα ως «ικανότητες που δεν εμφανίζονται σε μικρά μοντέλα, αλλά προκύπτουν σε μεγάλα μοντέλα», ένα από τα πιο εξέχοντα χαρακτηριστικά που διακρίνουν τα LLMs από τα προηγούμενα PLMs. Παρουσιάζεται επίσης μια αξιοσημείωτη ιδιότητα όταν εμφανίζονται οι αναδυόμενες ικανότητες: η απόδοση αυξάνεται σημαντικά πάνω από το τυχαίο όταν η κλίμακα φτάνει σε ένα ορισμένο επίπεδο. Κατά αναλογία, ένα τέτοιο αναδυόμενο μοτίβο συνδέεται στενά με το φαινόμενο της αλλαγής φάσης στη φυσική. Καταρχήν, οι αναδυόμενες ικανότητες μπορούν να οριστούν σε σχέση με ορισμένες σύνθετες εργασίες, ενώ μας ενδιαφέρουν περισσότερο οι γενικές ικανότητες που μπορούν να εφαρμοστούν για την επίλυση ποικιλίας εργασιών. Εδώ παρουσιάζουμε συνοπτικά τρεις τυπικές

αναδυόμενες ικανότητες για τα LLMs και αντιπροσωπευτικά μοντέλα που διαθέτουν αυτές τις ικανότητες.

- **Μάθηση εντός συμφραζομένων:** Η ικανότητα μάθησης εντός συμφραζομένων (ICL) [8] εισήχθη επίσημα με το GPT-3: υποθέτοντας ότι το γλωσσικό μοντέλο έχει λάβει μια φυσική γλωσσική οδηγία ή/και αρκετές εργασίες, μπορεί να παράγει την αναμενόμενη έξοδο για τις περιπτώσεις δοκιμής ολοκληρώνοντας τη σειρά λέξεων του εισαγόμενου κειμένου, χωρίς να απαιτείται πρόσθετη εκπαίδευση ή ενημέρωση βαθμίδων. Από τα μοντέλα της σειράς GPT, το GPT-3 175B παρουσίασε ισχυρή ικανότητα ICL γενικά, αλλά όχι τα μοντέλα GPT-1 και GPT-2. Αυτή η ικανότητα εξαρτάται επίσης από τη συγκεκριμένη καθοδική εργασία. Για παράδειγμα, η ικανότητα ICL μπορεί να εμφανιστεί στις αριθμητικές εργασίες (π.χ., πρόσθεση και αφαίρεση τριψήφιων αριθμών) για το GPT-3 13B, αλλά το GPT-3 175B δεν λειτουργεί το ίδιο καλά στην εργασία ερωταποκρίσεων στα περσικά.
- **Ακολούθηση οδηγιών**: Με τη ρύθμιση μέσω συνδυασμού πολλαπλών εργασιών διαμορφωμένων σε περιγραφές φυσικής γλώσσας [1] (ονομάζεται tuning οδηγιών), τα LLMs φαίνεται ότι επιτελούν καλά εργασίες που δεν έχουν δει, όταν αυτές περιγράφονται με τη μορφή οδηγιών. Με το tuning οδηγιών, τα LLMs μπορούν να ακολουθήσουν τις οδηγίες εργασίας για νέες εργασίες χωρίς να χρησιμοποιούν ρητά παραδείγματα, ενισχύοντας έτσι την ικανότητα γενίκευσης. Σύμφωνα με πειράματα, το LaMDA-PT που ρυθμίστηκε μέσω οδηγιών άρχισε να υπερέχει σημαντικά του μη ρυθμισμένου σε εργασίες που δεν είχε δει όταν το μέγεθος του μοντέλου έφτασε τα 68B, αλλά όχι για μοντέλα μεγέθους 8B ή μικρότερα. Πρόσφατη μελέτη διαπίστωσε ότι ένα μέγεθος μοντέλου τουλάχιστον 62B είναι απαραίτητο για το PaLM ώστε να αποδίδει καλά σε διάφορες εργασίες σε τέσσερα αξιολογικά benchmarks, αν και για ορισμένες εργασίες μπορεί να αρκεί μικρότερο μέγεθος.
- **Βήμα προς βήμα συλλογιστική**: Για τα μικρά γλωσσικά μοντέλα, είναι συνήθως δύσκολο να λύσουν σύνθετες εργασίες που απαιτούν πολλαπλά βήματα συλλογισμού, π.χ., προβλήματα μαθηματικών λέξεων. Αντίθετα, με τη στρατηγική του "Chain-of-Thought" (CoT) [9], τα LLMs μπορούν να επιλύουν τέτοιες εργασίες χρησιμοποιώντας τον μηχανισμό εντολών που περιλαμβάνει ενδιάμεσα βήματα συλλογισμού για την εξαγωγή της τελικής απάντησης. Αυτή η ικανότητα πιθανώς αποκτάται μέσω της εκπαίδευσης με κώδικα. Μια εμπειρική μελέτη έδειξε ότι η εντολή CoT μπορεί να προσφέρει οφέλη απόδοσης (σε αριθμητικά benchmarks) όταν εφαρμόζεται σε παραλλαγές του PaLM και LaMDA με μέγεθος μοντέλου μεγαλύτερο από 60B, ενώ το πλεονέκτημά της έναντι της τυπικής εντολής γίνεται πιο εμφανές όταν το μέγεθος του μοντέλου υπερβαίνει τα 100B.

Στη βιβλιογραφία, οι νόμοι κλιμάκωσης και οι αναδυόμενες ικανότητες παρέχουν δύο προοπτικές για την κατανόηση του πλεονεκτήματος των μεγάλων μοντέλων έναντι των μικρών μοντέλων. Γενικά, ο νόμος κλιμάκωσης (συχνά μετριέται από την απώλεια γλωσσικής μοντελοποίησης) περιγράφει μια προβλέψιμη σχέση απόδοσης με πιθανό φαινόμενο μειούμενης απόδοσης, ενώ οι αναδυόμενες ικανότητες (συχνά μετριούνται από την απόδοση σε εργασίες) είναι απρόβλεπτες αλλά πολύ επωφελείς όταν εμφανίζονται. Επειδή οι δύο αυτές προοπτικές αντανακλούν διαφορετικές τάσεις απόδοσης (συνεχής βελτίωση έναντι ξαφνικού άλματος), μπορεί να οδηγήσουν σε μη ευθυγραμμισμένα ευρήματα ή παρατηρήσεις. Υπάρχουν επίσης εκτεταμένες

συζητήσεις για τη λογικότητα των αναδυόμενων ικανοτήτων. Μια δημοφιλής υπόθεση είναι ότι οι αναδυόμενες ικανότητες μπορεί να αποδίδονται εν μέρει στις ρυθμίσεις αξιολόγησης για ειδικές εργασίες (π.χ., οι διακοπτόμενες μετρικές αξιολόγησης): όταν οι μετρικές αξιολόγησης τροποποιούνται ανάλογα, η απότομη καμπύλη της αναδυόμενης ικανότητας θα εξαφανιζόταν. Ωστόσο, η απόδοση των LLMs στις περισσότερες εργασίες γίνεται αντιληπτή από τους χρήστες φυσικά με έναν μη συνεχόμενο τρόπο. Για παράδειγμα, οι τελικοί χρήστες προτιμούν έναν αξιόπιστο κώδικα που δημιουργείται από LLMs και μπορεί να περάσει την περίπτωση δοκιμής, αλλά δεν ενδιαφέρονται για την επιλογή ενός καλύτερου κώδικα με λιγότερα λάθη μεταξύ δύο αποτυχημένων.

Πρόσφατα, μια μελέτη πρότεινε μια νέα ρύθμιση αξιολόγησης που μπορεί να αυξήσει την ανάλυση των μετρικών για τις εργασίες, καθιστώντας την απόδοση των εργασιών πιο προβλέψιμη. Παρά τις προσπάθειες αυτές, απαιτείται πιο θεμελιώδης έρευνα για τη λειτουργία των LLMs ώστε να κατανοηθεί η εμφάνιση συγκεκριμένων ικανοτήτων. Η λεπτή σχέση μεταξύ του νόμου κλιμάκωσης και των αναδυόμενων ικανοτήτων μπορεί να εξηγηθεί κατά αναλογία με την απόκτηση ικανοτήτων στους ανθρώπους. Για παράδειγμα, στην ανάπτυξη γλωσσικής ικανότητας στα παιδιά, η διαδικασία ανάπτυξης μπορεί να θεωρηθεί ως μια πολυεπίπεδη διαδικασία όπου εμφανίζονται «αναδυόμενες ικανότητες».

2.1.3 Κύριες Τεχνικές

Η εξέλιξη των LLMs μέχρι την τρέχουσα κατάστασή τους ως γενικών και ικανών μοντέλων μάθησης έχει περάσει από μακρά διαδικασία. Κατά την πορεία της ανάπτυξής τους, προτάθηκαν αρκετές σημαντικές τεχνικές, οι οποίες βελτίωσαν σημαντικά τις δυνατότητες των LLMs. Εδώ αναφέρουμε εν συντομία μερικές από τις βασικές τεχνικές που (ενδεχομένως) οδήγησαν στην επιτυχία των LLMs:

- **Κλιμάκωση**: Υπάρχει ένα σαφές φαινόμενο κλιμάκωσης [10] στα γλωσσικά μοντέλα Transformer: μεγαλύτερα μεγέθη μοντέλων/δεδομένων και περισσότερη υπολογιστική ισχύς συνήθως οδηγούν σε βελτιωμένη δυναμικότητα του μοντέλου. Ως δύο αντιπροσωπευτικά μοντέλα, τα GPT-3 και PaLM διερεύνησαν τα όρια της κλιμάκωσης, αυξάνοντας το μέγεθος του μοντέλου στα 175B και 540B αντίστοιχα. Δεδομένου ότι ο υπολογιστικός προϋπολογισμός είναι συνήθως περιορισμένος, οι νόμοι κλιμάκωσης μπορούν περαιτέρω να αξιοποιηθούν για πιο αποδοτική κατανομή των πόρων υπολογισμού. Για παράδειγμα, το μοντέλο Chinchilla (με περισσότερα δεδομένα εκπαίδευσης) υπερτερεί του μοντέλου Gopher (με μεγαλύτερο μέγεθος μοντέλου) αυξάνοντας την κλίμακα δεδομένων με τον ίδιο υπολογιστικό προϋπολογισμό. Επιπλέον, η κλιμάκωση δεδομένων πρέπει να συνοδεύεται από μια προσεκτική διαδικασία καθαρισμού, καθώς η ποιότητα των δεδομένων προεκπαίδευσης παίζει κρίσιμο ρόλο στη δυναμικότητα του μοντέλου.
- **Εκπαίδευση**: Λόγω του τεράστιου μεγέθους του μοντέλου, είναι πολύ δύσκολο να εκπαιδευτεί [11] με επιτυχία ένα ικανό LLM. Απαιτούνται αλγόριθμοι κατανεμημένης εκπαίδευσης για την εκμάθηση των παραμέτρων δικτύου των LLMs, όπου διάφορες στρατηγικές παράλληλης επεξεργασίας χρησιμοποιούνται συχνά από κοινού. Για να υποστηριχθεί η κατανεμημένη εκπαίδευση, έχουν κυκλοφορήσει αρκετά πλαίσια βελτιστοποίησης για να διευκολύνουν την εφαρμογή και ανάπτυξη παράλληλων

αλγορίθμων, όπως τα DeepSpeed και Megatron-LM. Επίσης, τρικς βελτιστοποίησης είναι σημαντικά για τη σταθερότητα της εκπαίδευσης και την απόδοση του μοντέλου, π.χ., επανεκκίνηση για να ξεπεραστεί η απότομη αύξηση της απώλειας και εκπαίδευση μεικτής ακρίβειας. Πρόσφατα, το GPT-4 πρότεινε την ανάπτυξη ειδικών υποδομών και μεθόδων βελτιστοποίησης που προβλέπουν αξιόπιστα την απόδοση των μεγάλων μοντέλων με πολύ μικρότερα μοντέλα.

- **Ανάκληση Ικανοτήτων**: Αφού έχουν προεκπαιδευτεί σε μεγάλης κλίμακας σύνολα δεδομένων, τα LLMs διαθέτουν δυνητικές ικανότητες ως καθολικοί επιλύτες εργασιών [1]. Αυτές οι ικανότητες μπορεί να μην εκδηλώνονται ρητά όταν τα LLMs εκτελούν ορισμένες συγκεκριμένες εργασίες. Ως τεχνική προσέγγιση, είναι χρήσιμο να σχεδιάζονται κατάλληλες οδηγίες εργασίας ή ειδικές στρατηγικές μάθησης εντός συμφραζομένων για να ανακληθούν αυτές οι ικανότητες. Για παράδειγμα, το "Chain-of-Thought prompting" έχει αποδειχθεί χρήσιμο για την επίλυση σύνθετων εργασιών συλλογιστικής, περιλαμβάνοντας ενδιάμεσα βήματα συλλογισμού. Επιπλέον, μπορούμε να πραγματοποιήσουμε tuning οδηγιών στα LLMs με περιγραφές εργασιών σε φυσική γλώσσα, για τη βελτίωση της ικανότητας γενίκευσης των LLMs σε μη ορατές εργασίες. Αυτές οι τεχνικές ανάκλησης σχετίζονται κυρίως με τις αναδυόμενες ικανότητες των LLMs, οι οποίες ενδέχεται να μην έχουν το ίδιο αποτέλεσμα σε μικρότερα γλωσσικά μοντέλα.
- **Ευθυγράμμιση**: Επειδή τα LLMs εκπαιδεύονται για να συλλαμβάνουν τα χαρακτηριστικά των δεδομένων προεκπαίδευσης [1] (περιλαμβάνοντας τόσο δεδομένα υψηλής όσο και χαμηλής ποιότητας), είναι πιθανό να παράγουν τοξικό, προκατειλημμένο ή ακόμη και επιβλαβές περιεχόμενο για τους ανθρώπους. Είναι απαραίτητο να ευθυγραμμιστούν τα LLMs με ανθρώπινες αξίες, π.χ., χρήσιμα, ειλικρινή και μη επιβλαβή. Για τον σκοπό αυτό, το InstructGPT σχεδίασε μια αποτελεσματική προσέγγιση ευθυγράμμισης που επιτρέπει στα LLMs να ακολουθούν τις αναμενόμενες οδηγίες, χρησιμοποιώντας την τεχνική της ενισχυτικής μάθησης με ανατροφοδότηση από ανθρώπους. Ενσωματώνει τον άνθρωπο στον κύκλο εκπαίδευσης με προσεκτικά σχεδιασμένες στρατηγικές. Το ChatGPT βασίζεται σε παρόμοια τεχνική με το InstructGPT, παρουσιάζοντας ισχυρή ικανότητα ευθυγράμμισης στην παραγωγή ποιοτικών και μη επιβλαβών απαντήσεων, π.χ., αρνούμενο να απαντήσει σε προσβλητικές ερωτήσεις.
- **Χειρισμός Εργαλείων**: Στην ουσία, τα LLMs εκπαιδεύονται ως γεννήτριες κειμένου πάνω σε τεράστια σώματα απλού κειμένου, συνεπώς αποδίδουν λιγότερο καλά σε εργασίες που δεν εκφράζονται ιδανικά με τη μορφή κειμένου (π.χ., αριθμητικός υπολογισμός). Επιπλέον, οι δυνατότητές τους περιορίζονται επίσης στα δεδομένα προεκπαίδευσης, π.χ., η αδυναμία τους να συλλάβουν ενημερωμένες πληροφορίες. Για την αντιμετώπιση αυτών των ζητημάτων, μια πρόσφατα προταθείσα τεχνική είναι να χρησιμοποιούνται εξωτερικά εργαλεία για να αντισταθμιστούν οι αδυναμίες των LLMs. Για παράδειγμα, τα LLMs μπορούν να χρησιμοποιούν την αριθμομηχανή για ακριβείς υπολογισμούς και να απευθύνονται σε μηχανές αναζήτησης για ανάκτηση άγνωστων πληροφοριών. Πρόσφατα, το ChatGPT έχει ενεργοποιήσει τον μηχανισμό χρήσης εξωτερικών προσθηκών (υπαρχόντων ή νεοδημιουργημένων εφαρμογών). Αυτός ο μηχανισμός μπορεί να επεκτείνει ευρέως τις δυνατότητες των LLMs.

2.2 Έμπιστή Τεχνητή Νοημοσύνη

2.2.1 Εισαγωγή

Ο προσδιορισμός του πόση εμπιστοσύνη πρέπει να αποδοθεί σε ένα αποτέλεσμα που παράγεται από αλγορίθμους Τεχνητής Νοημοσύνης (AI) / Μηχανικής Μάθησης (ML) μπορεί να είναι μια εξαιρετικά δύσκολη διαδικασία, ειδικά για τη λήψη στρατηγικών αποφάσεων. Υπάρχουν κάποιοι σημαντικοί λόγοι, οι οποίοι παρατίθενται στη συνέχεια [12]:

- Η φύση της λογικής μάθησης είναι αντίστροφη, όπου η μηχανή μαθαίνει κανόνες ή μοτίβα που προέρχονται από δεδομένα, αντί να προγραμματίζεται ρητά.
- Ενώ ο David Hume υποστηρίζει ότι δεν υπάρχει λόγος να αναμένεται ότι το μέλλον θα μοιάζει με το παρελθόν, εξακολουθούμε να χρησιμοποιούμε ιστορικά δεδομένα για να κάνουμε προβλέψεις σχετικά με το μέλλον.
- Πολλά μοντέλα ML, όπως τα βαθιά νευρωνικά δίκτυα, οι μηχανές υποστήριξης διανυσμάτων και οι μέθοδοι συνόλων, θεωρούνται μαύρα κουτιά. Λόγω της έλλειψης διαφάνειας, πολλά μοντέλα μαύρων κουτιών θα μπορούσαν να αποτελέσουν σημαντικό κίνδυνο για τη στρατηγική μιας επιχείρησης ή για αποφάσεις υψηλού ρίσκου.
- Η θεμελιώδης αρχή της ML είναι να εντοπίζει συσχετισμούς αντί για αιτιότητα. Ο Pearl υποστηρίζει ότι η παραδοσιακή στατιστική προσέγγιση, που επικεντρώνεται στην ταυτοποίηση συσχετισμών μεταξύ μεταβλητών, δεν επαρκεί για την κατανόηση της αιτιότητας.
- Η ασφάλεια και η προστασία της ιδιωτικότητας των συστημάτων AI είναι κρίσιμες, καθώς οι κυβερνοεπιθέσεις μπορούν να επιφέρουν σοβαρές συνέπειες, όπως κλοπή δεδομένων, οικονομικές απώλειες και παραβιάσεις ασφαλείας. Αυτά τα ζητήματα μπορούν να επηρεάσουν σημαντικά την εμπιστοσύνη των ανθρώπων στα συστήματα AI.
- Η έλλειψη ουσιαστικού ανθρώπινου ελέγχου πάνω στα συστήματα AI αποτελεί σημαντικό εμπόδιο για την οικοδόμηση αξιόπιστης AI.
- Η λεπτότητα των ηθικών ζητημάτων εγείρει μια άλλη σημαντική πρόκληση για την αξιόπιστη AI, καθώς οι ηθικές αξίες είναι πολιτιστικοί κανόνες και εξαρτώνται από πεποιθήσεις.
- Η στρατηγική λήψη αποφάσεων σημαίνει μακροπρόθεσμες συνέπειες για έναν οργανισμό. Αυτές οι αποφάσεις συνήθως βασίζονται σε υποθέσεις για το μέλλον. Ως αποτέλεσμα, μπορεί να είναι δύσκολο να επαληθευτούν οι συνέπειες αμέσως.

Για την συστηματική επίλυση αυτών των προκλήσεων [12] [13], η ποιότητα των δεδομένων είναι ένα καλό σημείο εκκίνησης, καθώς τα δεδομένα αποτελούν κρίσιμο όσο αφορά την εξασφάλιση εμπιστοσύνης. Παραδοσιακά, έχει χρησιμοποιηθεί μια προσέγγιση βασισμένη σε κανόνες, γνωστή ως καλή GOFAI. Με την GOFAI, ορίζεται ένα σύνολο κανόνων με ένα συγκεκριμένο σύνολο δεδομένων και η μηχανή εξάγει ένα επιθυμητό αποτέλεσμα. Με τη ML ή τα βαθιά νευρωνικά δίκτυα, αντιστρέφεται η λογική της μάθησης, όπου καθορίζεται τι ζητείται με ένα συγκεκριμένο σύνολο δεδομένων και η μηχανή ανακαλύπτει κανόνες ή μοτίβα στον χώρο της αναπαράστασης για λογαριασμό των χρηστών. Επιπλέον, οι εφαρμογές ML για στρατηγική λήψη

αποφάσεων είναι διαδεδομένες. Μπορούν να βρεθούν σε πολλές βιομηχανίες και τομείς, όπως η επιχειρηματική ανάλυση, η διαχειριστική οικονομία, ο οργανωτικός σχεδιασμός, η οργανωτική συμπεριφορά, οι οικονομικές επενδύσεις, η διαχείριση περιουσιακών στοιχείων, η διαχείριση εφοδιαστικής αλυσίδας, ο κύκλος ζωής προϊόντων, η ανάλυση αγοράς και η ανταγωνιστική ανάλυση.

Όταν κάποιος εργάζεται με ένα σύνολο δεδομένων που είναι ελεύθερο από διαφθορά, προκαταλήψεις, ελλείποντα δεδομένα και ασυνέπειες, μπορεί να έχει σχετική εμπιστοσύνη στην ακρίβεια των αποτελεσμάτων της μάθησης. Δυστυχώς, τα δεδομένα που υπάρχουν συνήθως είναι χαοτικά ή ελλιπή, κάτι που εμποδίζει την ικανότητα να γίνονται ακριβείς προβλέψεις ή αξιολογήσεις. Το Data Warehousing Institute (TDWI) ανέφερε ότι η κακή ποιότητα δεδομένων κόστισε στις επιχειρήσεις των ΗΠΑ 600 δισεκατομμύρια δολάρια ετησίως το 2012. Πρόσφατα, η IBM ανέφερε ότι το μέσο κόστος της κακής ποιότητας δεδομένων θα έφτανε κατά μέσο όρο τα 15 εκατομμύρια δολάρια ανά επιχείρηση ή περισσότερα από 3,1 τρισεκατομμύρια δολάρια ετησίως το 2022. Οι Batini και Scannapieco υπογράμμισαν οκτώ κοινές ανησυχίες σχετικά με την ποιότητα των δεδομένων: 1) ακρίβεια, 2) πληρότητα, 3) πλεονασμός, 4) αναγνωσιμότητα, 5) προσβασιμότητα και διαθεσιμότητα, 6) συνέπεια, 7) χρησιμότητα, και 8) εμπιστοσύνη.

Επιπλέον, μπορεί να συναντηθούν ζητήματα όπως ελλείποντα χαρακτηριστικά, απίθανες τιμές, ασαφείς έννοιες και μη ανιχνεύσιμα δεδομένα. Συνολικά, αυτά τα ζητήματα μπορούν να κατηγοριοποιηθούν σε δύο μεγάλες κατηγορίες: μη διαθέσιμα και μη χρησιμοποιήσιμα δεδομένα. Όταν τα δεδομένα είναι ακαθάριστα και προκατειλημμένα, τα αποτελέσματα των αλγορίθμων ML γίνονται λιγότερο αξιόπιστα, καθιστώντας δύσκολο να εμπιστευτεί κανείς τα αποτελέσματα.

2.2.2 Βασικές Έννοιες Έμπιστης Τεχνητής Νοημοσύνης

2.2.2.1 Εμπιστοσύνη

Το πλαίσιο της της Έμπιστής Τεχνητής Νοημοσύνης [14] (TAI – Trustworthy AI) έχει σχεδιαστεί για να καλύπτει συγκεκριμένες απαιτήσεις και να εκπροσωπεί εξειδικευμένες γνώσεις σε ορισμένους τομείς. Είναι βασισμένο σε θεμελιώδεις έννοιες, όπως η εμπιστοσύνη, η αξιοπιστία, η AI, η απόφαση, η ηθική, οι αξίες και η πολιτική. Οι Flores και Solomon [14] υποστήριξαν ότι η εμπιστοσύνη από μόνη της είναι μια απόφαση. Προσδιορίζει τους δεσμούς σχέσεων προς κάποιον ή κάτι. Λαμβάνοντας υπόψη την κοινωνική διάσταση, η εμπιστοσύνη δεν είναι μόνο σχέση, αλλά φέρει και σημαντική συναισθηματική βαρύτητα. Όπως όλα τα συναισθήματα, η εμπιστοσύνη είναι δυναμική. Ποτέ δεν σημαίνει μια μόνο περίπτωση εμπιστοσύνης. Συνεπώς, η εμπιστοσύνη μπορεί να δημιουργηθεί, να αποκατασταθεί ή να καταστραφεί. Ομοίως, η αξιοπιστία φέρει την ίδια σημασία.

Συχνά χρησιμοποιούμε αυτούς τους δύο όρους εναλλακτικά. Μερικές φορές αυτοί οι δύο όροι προκαλούν σύγχυση επειδή είναι στενά συνδεδεμένες έννοιες. Λεξικολογικά, η εμπιστοσύνη σημαίνει πίστη στην αξιοπιστία. Οι Flores και Solomon υποστήριξαν ότι η εμπιστοσύνη είναι συχνά το αποτέλεσμα του ότι κάτι γίνεται αντιληπτό ως αξιόπιστο. Αυτοί οι δύο όροι σχηματίζουν ένα προφανές συμπληρωματικό ζεύγος. Η εμπιστοσύνη ανοίγει τη δυνατότητα της αξιοπιστίας. Χωρίς εμπιστοσύνη, η αξιοπιστία είναι αδύνατη.

2.2.2.2 Διαβάθμιση Εμπιστοσύνης

Εκτός από αυτήν τη συμπληρωματική σχέση, οι Flores και Solomon [14] ταξινόμησαν περαιτέρω την εμπιστοσύνη σε πέντε επίπεδα [15] ή βαθμίδες: απλή εμπιστοσύνη, βασική εμπιστοσύνη, τυφλή εμπιστοσύνη, αυθεντική εμπιστοσύνη και διατυπωμένη εμπιστοσύνη. Η απλή εμπιστοσύνη είναι μια μορφή αφελούς, μη αμφισβητούμενης, αναμφισβήτητης και αδιατύπωτης πίστης, όπως τα παιδιά εμπιστεύονται τους γονείς τους. Αυτό το είδος εμπιστοσύνης είναι χωρίς δυσπιστία. Η βασική εμπιστοσύνη είναι ένα σύνολο βασικών κανόνων που θεωρούνται δεδομένοι, όπως η εγγύηση προσωπικής ασφάλειας και ιδιωτικότητας σε μια ειρηνική κοινωνία. Η τυφλή εμπιστοσύνη είναι ένα είδος πεισματικής νοοτροπίας ή αυταπάτης. Η αυθεντική εμπιστοσύνη κρατά την εμπιστοσύνη και τη δυσπιστία σε ισορροπία. Οι άνθρωποι που έχουν αυθεντική εμπιστοσύνη κατανοούν τους κινδύνους και τις ευπάθειές της. Η διατυπωμένη εμπιστοσύνη εκφράζεται ως συμβόλαιο ή συμφωνία με μηχανισμούς επιβολής.

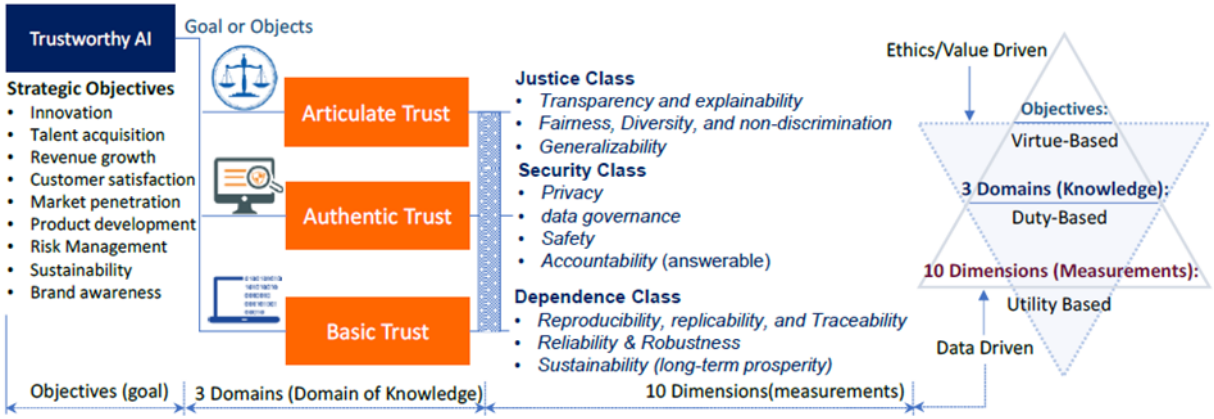
Από τη σκοπιά της μέτρησης, η διατυπωμένη εμπιστοσύνη προσφέρει δικαιοσύνη, διαφάνεια και γενικευσιμότητα, τα οποία μπορούν να συνδεθούν με την κατηγορία της δικαιοσύνης. Στο επόμενο επίπεδο, η αυθεντική εμπιστοσύνη αφορά τους κινδύνους και τη μακροπρόθεσμη λογοδοσία, που μπορούμε να την θεωρήσουμε ως κατηγορία ασφάλειας. Με άλλα λόγια, απαιτούμε ένα σύστημα ΤΑΙ να ποσοτικοποιεί τις απειλές και τα πιθανά οφέλη και να προσφέρει μια επιλογή ισορροπίας. Η βασική εμπιστοσύνη συχνά αγνοείται και παραμένει ανείπωτη, γνωστή ως αδιατύπωτη εμπιστοσύνη. Για παράδειγμα, το αποτέλεσμα μιας στρατηγικής απόφασης που παράγεται από την AI/ML μπορεί τουλάχιστον να είναι αναπαραγώγιμο, χωρίς να αναφέρεται ρητά. Σε κάποιο βαθμό, η βασική εμπιστοσύνη μπορεί να συνδεθεί χαλαρά με την κατηγορία της εξάρτησης. Αν και υπάρχει μια χαλαρή σχέση ανάμεσα σε έναν τομέα και σε μια κατηγορία διάστασης, κάθε διάσταση μπορεί να ευθυγραμμιστεί με έναν διαφορετικό τομέα εμπιστοσύνης. Αυτό εξαρτάται από τους στόχους. Για παράδειγμα, εάν θέλουμε να λάβουμε μια στρατηγική απόφαση για ένα δημόσιο επενδυτικό έργο (π.χ., δημόσιες μεταφορές), η δικαιοσύνη (κατηγορία δικαιοσύνης), η λογοδοσία (κατηγορία ασφάλειας) και η βιωσιμότητα (κατηγορία εξάρτησης) πρέπει να ληφθούν υπόψη. Αυτό το παράδειγμα δίνει αφορμή για τη δημιουργία εννοιών διάστασης.

2.2.2.3 Διαστάσεις Εμπιστοσύνης

Η Εικόνα 2[1] απεικονίζει δέκα εννοιολογικές διαστάσεις [16]. Μπορούμε να τις κατατάξουμε περίπου σε τρεις κατηγορίες: δικαιοσύνη, ασφάλεια και εξάρτηση. Στη δικαιοσύνη, υπάρχει η διάσταση της διαφάνειας και της εξηγησιμότητας. Η λέξη διαφάνεια σημαίνει ότι μια ουσία έχει την ιδιότητα να επιτρέπει στο φως να περνά μέσα από αυτήν και τα αντικείμενα πίσω της μπορούν να φανούν καθαρά. Είναι το αντίθετο της αδιαφάνειας ή του μαύρου κουτιού. Από την οπτική της AI, η διαφάνεια έχει γίνει ένας από τους υποχρεωτικούς κανόνες που απαιτούνται από τον Γενικό Κανονισμό για την Προστασία Δεδομένων (GDPR), διότι η διαδικασία λήψης αποφάσεων με τη βοήθεια της AI υπερβαίνει τις ικανότητες του ανθρώπου-ειδικού. Η διαφάνεια μπορεί να υλοποιηθεί μέσω της αλγοριθμικής διαφάνειας. Ο Wortham υποστήριξε ότι η διαφάνεια συνδέεται συχνά με τις έννοιες της λογοδοσίας και της ευθύνης σε αντικειμενικό επίπεδο. Ωστόσο, ο πιο

[1] *https://www.semanticscholar.org/paper/Survey-of-Trustworthy-AI%3A-A-Meta-Decision-of-AI-Wu-Li/5bfc22ae81d47e70fc71b02c84d907c3381975ea/figure/2*

κοντινός ορισμός της διαφάνειας είναι η εξηγησιμότητα. Ο ορισμός του εξηγήσιμου είναι το να είναι κάτι κατανοητό και ερμηνεύσιμο και να έχει λόγους ύπαρξης στην TAI. Με άλλα λόγια, κάτι ή μια ιστορία είναι δικαιολογημένο και αποδεκτό. Είναι η κατάσταση του να είναι δίκαιο.



Εικόνα 2: Γενικό Πλαίσιο της Έμπιστης Τεχνητής Νοημοσύνης

Η αμεροληψία αποτελεί ουσιαστικό μέρος της κατηγορίας της δικαιοσύνης. Η ποικιλομορφία και η μη διάκριση είναι συμπληρωματικές της αμεροληψίας. Η ποικιλομορφία είναι η κατάσταση του να υπάρχουν ή να αποτελούνται από διαφορετικά στοιχεία. Ένα από τα πλεονεκτήματα της ποικιλομορφίας είναι η σοφία του πλήθους. Με αρκετή ποικιλομορφία δεδομένων, η μηχανή μπορεί να μάθει ή να διατυπώσει κανόνες χωρίς προκαταλήψεις. Η ποικιλομορφία στοχεύει να διατυπώσει σοφία και να μετριάσει την άγνοια ή τη διάκριση. Η μη διάκριση σημαίνει την ίση μεταχείριση. Οι Veale και Binns [15] υποστήριξαν ότι τα αλγοριθμικά μοντέλα μάθησης μπορούν να παράγουν προκαταλήψεις εάν η μηχανή μαθαίνει από ιστορικά δεδομένα, όπως το φύλο, η εθνικότητα, η σεξουαλικότητα ή η αναπηρία. Η μελέτη τους προσφέρει τρία πιθανά μοντέλα για την μείωση της πιθανής διάκρισης.

Η έννοια της μη διάκρισης φέρνει μια άλλη στενά συνδεδεμένη έννοια, τη γενικευσιμότητα. Σημαίνει τον βαθμό στον οποίο μπορούμε να εφαρμόσουμε τα αποτελέσματα της εκπαίδευσης σε ένα ευρύτερο πλαίσιο. Ωστόσο, η γενίκευση συχνά αντιμετωπίζει το ζήτημα της διατήρησης της ιδιωτικότητας.

Κατατάσσουμε την ιδιωτικότητα ως μέρος της ασφάλειας. Είναι η κατάσταση του να μην μοιράζονται οι πληροφορίες με άλλους – η ιδιωτικότητα έχει σημασία για πολλούς λόγους, όπως η διασφάλιση της αυτοδιάθεσης, της ασφάλειας και της εμπιστοσύνης. Η ιδιωτικότητα είναι ζωτικής σημασίας όταν χρησιμοποιούμε AI/ML για ιατρικές εφαρμογές.

Η διακυβέρνηση δεδομένων εγγυάται ότι τα συλλεγόμενα δεδομένα είναι ασφαλή, ποιοτικά, ακεραιότητας και χρήσιμα για εξουσιοδοτημένους χρήστες και ενδιαφερόμενους κατά τη διάρκεια του κύκλου ζωής των δεδομένων, μέχρι τα δεδομένα να καταστραφούν ή να αρχειοθετηθούν. Όταν αντιμετωπίζουμε την αυτοματοποιημένη λήψη αποφάσεων, το ζήτημα της ασφάλειας γίνεται κρίσιμο. Η IEEE διενεργεί αρκετές διαδικασίες σχετικά με την ηθική και την ασφάλεια της AI. Η ασφάλεια της AI είναι ένα συστημικό ζήτημα, διότι πολλές νέες εφαρμογές σε διάφορους τομείς της κοινωνίας και των δημόσιων υποδομών φέρνουν νέους κινδύνους, οι οποίοι οδηγούν σε νέες μορφές παθογενειών.

Επιπλέον, η λογοδοσία αναφέρεται στην προσδοκία ότι οι οργανισμοί ή τα άτομα θα διασφαλίσουν την ορθή λειτουργία, καθ' όλη τη διάρκεια του κύκλου ζωής τους, των συστημάτων

AI που σχεδιάζουν, λειτουργούν ή αναπτύσσουν, σύμφωνα με τους ρόλους τους και τα ισχύοντα ρυθμιστικά πλαίσια, και να το αποδεικνύουν μέσω των πράξεων και της διαδικασίας λήψης αποφάσεων. Ο Loi και ο Spielkamp [15] ποσοτικοποίησαν περαιτέρω τη λογοδοσία ως μια συνθήκη σχέσης που περιλαμβάνει τουλάχιστον τρία στοιχεία: ευθύνη, απαντησιμότητα και δυνατότητα κυρώσεων. Αυτά τα στοιχεία δημιουργούν την κατηγορία της εξάρτησης.

Ένα μέρος της εξάρτησης είναι το να είναι κάτι αναπαραγώγιμο, ανακτήσιμο και ανιχνεύσιμο. Αυτό σημαίνει ότι το πειραματικό αποτέλεσμα πρέπει να είναι αναπαραγώγιμο. Επιπλέον, η αξιοπιστία σημαίνει ότι τα συστήματα AI πληρούν τα αναμενόμενα πρότυπα απόδοσης. Με άλλα λόγια, η ποιότητα των συστημάτων AI είναι αξιόπιστη, εμπιστεύσιμη ή αποδίδει σταθερά καλά. Η ανθεκτικότητα, από την άλλη πλευρά, αναφέρεται στην ικανότητα των συστημάτων AI να λειτουργούν χωρίς αποτυχία υπό ένα ευρύ φάσμα πιθανών συνθηκών, συμπεριλαμβανομένων απροσδόκητων ή αντίξοων καταστάσεων. Στην επιστήμη των υπολογιστών, η ανθεκτικότητα σημαίνει ότι τα συστήματα AI μπορούν να αντιμετωπίσουν σφάλματα κατά την εκτέλεση κωδίκων AI. Εν συντομία, η αξιοπιστία αφορά τη συνεπή απόδοση υπό γνωστές, τυπικές συνθήκες, ενώ η ανθεκτικότητα περιλαμβάνει τη διατήρηση της κανονικής λειτουργίας υπό ένα ευρύ φάσμα πιθανών σεναρίων, συμπεριλαμβανομένων των απρόσμενων συνθηκών.

2.2.3 TAI και Διατυπωμένη Εμπιστοσύνη

Η διατυπωμένη εμπιστοσύνη είναι μια μορφή εμπιστοσύνης που έχει σαφώς οριστεί και γνωρίζει τους κινδύνους και τις αδυναμίες. Απαιτεί μέτρα εξηγησιμότητας, διαφάνειας, δικαιοσύνης και μη διάκρισης, ενώ μπορεί να προστεθούν και άλλες διαστάσεις ανάλογα με το πλαίσιο του συστήματος AI. Οι άνθρωποι δεν θα εμπιστευτούν τα συστήματα AI αν δεν μπορούν να εξηγηθούν οι αποφάσεις που λαμβάνονται με διαφάνεια, δικαιοσύνη και χωρίς διακρίσεις. Το έργο των Bennet και Keyes αναδεικνύει τη μεροληψία των αλγορίθμων και υποστηρίζει ότι η AI ηθική πρέπει να επικεντρωθεί στη δικαιοσύνη.

2.2.3.1 Εξηγησιμότητα και Διαφάνεια

Η μέθοδος των αντιθετικών εξηγήσεων (counterfactuals) [13] [16] χρησιμοποιείται για τη δικαιοσύνη και τη διαφάνεια. Το μοντέλο CERTFAI (Counterfactual Explanations for the Robustness, Transparency, Interpretability and Fairness of AI) αναπτύχθηκε για να βελτιώσει τη διαφάνεια και την ανθεκτικότητα των AI συστημάτων. Παρόλο που οι αντιθετικές εξηγήσεις είναι δημοφιλείς, δεν είναι πάντα βοηθητικές, διότι τα AI μοντέλα είναι πολύπλοκα.

Διάφορες προσεγγίσεις για την εξηγησιμότητα (XAI – Explainable AI) έχουν αναπτυχθεί, οι οποίες περιλαμβάνουν κοινωνικές επιστήμες, υποθέσεις και ηθικά ζητήματα. Ο Gilpin [16] πρότεινε ότι η XAI πρέπει να επικεντρωθεί στο "γιατί" και "γιατί πρέπει", κάτι που εγείρει νέες ερωτήσεις. Υπάρχουν ηθικά διλήμματα μεταξύ εξηγησιμότητας και ερμηνείας, διαφάνειας και χειραγώγησης. Προτείνεται η χρήση των ερωτήσεων "τι, τι και πώς" για να γίνει σαφής η διαφορά μεταξύ ερμηνείας και χειραγώγησης, προσφέροντας μια πλήρη εξήγηση των.

2.2.3.2 Δικαιοσύνη και Ποικιλομορφία

Η δικαιοσύνη και η ποικιλομορφία [16] σχετίζονται στενά με τη δικαιοσύνη στις διαδικασίες και τη μάθηση. Η δικαιοσύνη μπορεί να κατηγοριοποιηθεί σε τρεις βασικές κατηγορίες: α) δικαιοσύνη

στη διαδικασία, β) απώλεια μεροληψίας και γ) δίκαιη μάθηση. Οι αλγόριθμοι δικαιοσύνης, όπως το AI Fairness 360 (AIF360), προσφέρουν εργαλεία για τη μείωση της μεροληψίας. Το AIF360 παρέχει αλγόριθμους πριν, κατά και μετά την επεξεργασία για την εξισορρόπηση της δικαιοσύνης. Ωστόσο, η δικαιοσύνη δεν καλύπτει πάντα κάθε περίπτωση, και η υπερβολική εξισορρόπηση μπορεί να οδηγήσει σε απώλεια πληροφοριών και μείωση της ακρίβειας.

Ο Zhou πρότεινε μια λύση που διατηρεί τη δικαιοσύνη χωρίς απώλεια πληροφοριών, αλλά θυσιάζει την ακρίβεια προβλέψεων με την υπερεκπροσώπηση των μειονοτικών ομάδων. Μια άλλη πρόταση είναι η διαδικαστική δικαιοσύνη, που βασίζεται στην ακρίβεια των εισροών και την αξιοπιστία, αποφεύγοντας τις διακρίσεις. Ωστόσο, η γενικευσιμότητα παραμένει ένα ζήτημα.

2.2.3.3 Γενικευσιμότητα

Η γενικευσιμότητα [16] σχετίζεται με την αναπαραγωγιμότητα των αποτελεσμάτων. Είναι σημαντικό να διασφαλιστεί ότι τα μοντέλα είναι γενικεύσιμα πριν εφαρμοστούν σε μεγάλη κλίμακα, ειδικά σε ιατρικές εφαρμογές. Ο Arjovsky πρότεινε τον αλγόριθμο IRM για την ενίσχυση της γενικευσιμότητας, ο οποίος επιτρέπει τη γενίκευση εκτός του αρχικού συνόλου δεδομένων. Ωστόσο, υπάρχουν περιορισμοί, όπως η αποτυχία του IRM να συλλάβει "φυσικές" αμεταβλητότητες, κάτι που μπορεί να οδηγήσει σε μοντέλα με φτωχή γενικευσιμότητα σε νέα περιβάλλοντα. Συνοπτικά, η γενικευσιμότητα επιτρέπει τη χρήση νέων δεδομένων σε διάφορες τεχνικές μάθησης, αλλά μπορεί να αποκαλύψει ευαίσθητες πληροφορίες, κάτι που συνδέεται με θέματα ασφάλειας και αυθεντικής εμπιστοσύνης.

2.2.4 TAI και Αυθεντική Εμπιστοσύνη

Η αυθεντική εμπιστοσύνη αποδέχεται τη δυσπιστία και επιδιώκει να επιλύσει θέματα ασφάλειας, όπως επιθέσεις μέσω hacking και ransomware. Η ασφάλεια καλύπτει τέσσερις διαστάσεις: ιδιωτικότητα, διακυβέρνηση δεδομένων, ασφάλεια και λογοδοσία. Η ιδιωτικότητα και η διακυβέρνηση δεδομένων είναι αλληλένδετες αλλά διακριτές λόγω διαφορετικών μετρήσεων. Μοντέλα όπως τα Privacy-preserving GANs και PP-SVM επιδιώκουν να προστατεύσουν την ιδιωτικότητα κατά την ανάλυση δεδομένων, ενώ η ομοσπονδιακή μάθηση (FL) εστιάζει στη διατήρηση της ασφάλειας και ιδιωτικότητας τοπικά.

2.2.4.1 Ιδιωτικότητα

Προτάσεις όπως οι Privacy-preserving Support Vector Machines (PP-SVM) και το Gradient Boosting Decision Tree (GBDT) [16] προσφέρουν λύσεις για τη διαφύλαξη της ιδιωτικότητας κατά τη μάθηση μοντέλων. Ωστόσο, αυτές οι μέθοδοι μπορεί να μειώσουν την ακρίβεια ή την απόδοση. Η "k-anonymity" είναι μια παραδοσιακή προσέγγιση για την προστασία της ιδιωτικότητας, αλλά είναι ακατάλληλη για πολυδιάστατα δεδομένα. Η ομοσπονδιακή μάθηση (FL) προσφέρει μια λύση διατηρώντας την τοπική ασφάλεια και ιδιωτικότητα, μειώνοντας τον κίνδυνο διαρροής πληροφοριών.

Η FL χρησιμοποιεί πολλά τοπικά σύνολα δεδομένων αντί να τα μοιράζεται με κεντρικούς διακομιστές. Μοντέλα όπως το Secure Multi-part Communication (SMC) και το Differential Privacy (DP) βελτιώνουν την ασφάλεια και ιδιωτικότητα. Ωστόσο, προστίθεται θόρυβος στα δεδομένα, κάτι που μπορεί να μειώσει την απόδοση του μοντέλου. Το VerifyNet προτείνει ένα

πρωτόκολλο διπλής κάλυψης για την επαλήθευση των αποτελεσμάτων, ενώ τα υβριδικά μοντέλα συνδυάζουν DP και SMC για καλύτερη ισορροπία μεταξύ ιδιωτικότητας και ακρίβειας.

Η Αντιφατική Εκπαίδευση (AT) βελτιώνει την ασφάλεια εισάγοντας θόρυβο στα δεδομένα για να αυξήσει την ανθεκτικότητα των μοντέλων. Η Ενισχυμένη AT (EAT) μειώνει τη τρωτότητα, αλλά δεν αυξάνει την ακρίβεια των μοντέλων κατά τις επιθέσεις black-box.

2.2.4.2 Διακυβέρνηση Δεδομένων

Η διακυβέρνηση δεδομένων είναι η διαδικασία διαχείρισης δεδομένων για τη διασφάλιση της ποιότητας, ακεραιότητας, ασφάλειας και χρηστικότητας των δεδομένων καθ' όλη τη διάρκεια του κύκλου ζωής τους. Περιλαμβάνει την πρόσβαση σε επιτρεπόμενους χρήστες, τη δυνατότητα ελέγχου και τη συμμόρφωση με τους κανονισμούς. Υπάρχουν τρεις προσεγγίσεις στη διακυβέρνηση: αυτοδιακυβέρνηση (προσαρμογή σε κανονιστικά κενά), συν-διακυβέρνηση (συνεργασία για κοινό σκοπό), και ιεραρχική διακυβέρνηση (από πάνω προς τα κάτω προσέγγιση). Η συν-διακυβέρνηση προσφέρει ευελιξία και κλιμάκωση, ενώ η ιεραρχική διακυβέρνηση δίνει έμφαση στην αποτελεσματικότητα. Οι μέθοδοι διακυβέρνησης δεδομένων, όπως η τεκμηριωμένη διακυβέρνηση, εφαρμόζονται για την αντιμετώπιση προκλήσεων κυβερνοασφάλειας και τη διαφάνεια της AI.

2.2.4.3 Ασφάλεια

Η ασφάλεια στην AI είναι κρίσιμο ζήτημα, καθώς κακός σχεδιασμός μπορεί να οδηγήσει σε ανεπιθύμητες συνέπειες. Τα προβλήματα ασφάλειας χωρίζονται σε τέσσερις κατηγορίες: διακοπή λειτουργίας, αποφυγή παρενεργειών, απουσία επιβλέποντος, και εκμετάλλευση συστημάτων επιβράβευσης. Η διακοπή λειτουργίας αφορά τη δυνατότητα ανθρώπινης παρέμβασης σε ενέργειες AI, ενώ η αποφυγή παρενεργειών προσπαθεί να ελαχιστοποιήσει τις αρνητικές επιπτώσεις των AI συστημάτων. Η απουσία επιβλέποντος σχετίζεται με το πώς λειτουργεί η AI όταν δεν υπάρχει ανθρώπινη παρακολούθηση, και το ζήτημα της εκμετάλλευσης αφορά τη χειραγώγηση του συστήματος επιβράβευσης από την AI.

2.2.4.4 Λογοδοσία

Η λογοδοσία είναι κρίσιμη για συστήματα AI, αλλά συναντά εμπόδια όπως η ανάθεση ευθυνών σε πολλούς (πολλά χέρια), τα σφάλματα (bugs), η απόδοση ευθυνών στον υπολογιστή και η ιδιοκτησία χωρίς ευθύνη. Τα προβλήματα bugs περιλαμβάνουν σφάλματα σχεδιασμού και κώδικα, ενώ η απόδοση ευθυνών στον υπολογιστή μειώνει την ευθύνη των δημιουργών και χειριστών. Το ζήτημα της ιδιοκτησίας χωρίς ευθύνη εμφανίζεται όταν η βιομηχανία υπολογιστών αποφεύγει την ευθύνη για ζητήματα που προκύπτουν από τα προϊόντα της.

2.2.5 TAI και Βασική Εμπιστοσύνη

Η βασική εμπιστοσύνη είναι μια θεμελιώδης, αυτονόητη μορφή εμπιστοσύνης. Περιλαμβάνει τρεις διαστάσεις: 1) αναπαραγωγιμότητα και επαναληψιμότητα, 2) αξιοπιστία και ανθεκτικότητα, και 3) βιωσιμότητα. Η αναπαραγωγιμότητα αφορά τη δυνατότητα πλήρους αναπαραγωγής πειραμάτων και δεδομένων. Η αξιοπιστία συνδέεται με τη μάθηση κατά τη διάρκεια επιθέσεων, ενώ η βιωσιμότητα περιλαμβάνει περιβαλλοντικά, οικονομικά και κοινωνικά μοντέλα.

2.2.5.1 Αναπαραγωγιμότητα και Επαναληψιμότητα

Ο McDermott αναγνώρισε ζητήματα αναπαραγωγιμότητας στην υγειονομική περίθαλψη: 1) Τεχνική Επαναληψιμότητα (TR), που αφορά την πλήρη αναπαραγωγή ενός αποτελέσματος με τον ίδιο κώδικα και δεδομένα, 2) Στατιστική Επαναληψιμότητα (ST), που εξαρτάται από τις συνθήκες και τις τυχαίες παραμέτρους, και 3) Εννοιολογική Επαναληψιμότητα (CR), που σημαίνει ότι το αποτέλεσμα μπορεί να επεκταθεί σε άλλες καταστάσεις και μετρήσεις. Χωρίς αυτές τις τρεις διαστάσεις, τα αποτελέσματα δεν μπορούν να αναπαραχθούν πλήρως ούτε να εφαρμοστούν στην πράξη.

Επιπλέον, ο Gundersen πρότεινε έξι μετρικές αναπαραγωγιμότητας βασισμένες σε συνδυασμούς τριών παραγόντων: πείραμα, δεδομένα και μέθοδος. Τα αποτελέσματα έδειξαν ότι κανένα επιστημονικό άρθρο δεν είναι πλήρως αναπαραγώγιμο, με στατιστική βελτίωση σε ορισμένους τομείς, αλλά συνολική έλλειψη προόδου σε άλλες κατηγορίες.

2.2.5.2 Αξιοπιστία και Ανθεκτικότητα

Η αξιοπιστία και η ανθεκτικότητα είναι κρίσιμα για τα συστήματα λήψης αποφάσεων που βασίζονται στη Μηχανική Μάθηση (ML), ειδικά σε εφαρμογές όπως η αυτοκινητοβιομηχανία και η ιατρική διάγνωση. Οι Saria και Subbaswamy επεσήμαναν την απουσία πλαισίου για την αντιμετώπιση πιθανών αποτυχιών και καταστροφικών συνεπειών.

Επιπλέον, ο Brendel ασχολήθηκε με την αξιοπιστία των αλγορίθμων ML απέναντι σε "επιθετικές διαταραχές", που είναι μικρές αλλαγές στα δεδομένα εισόδου που οδηγούν σε λανθασμένα αποτελέσματα. Οι "επιθετικές επιθέσεις" χωρίζονται σε τρεις κατηγορίες: επιθέσεις με βάση το gradient, με βάση το score και μεταφερόμενες επιθέσεις. Οι επιθέσεις με βάση το gradient αντιμετωπίζονται με απόκρυψη των gradients, ενώ οι επιθέσεις με βάση το score αντιμετωπίζονται με την εισαγωγή στοχαστικών στοιχείων στο μοντέλο. Οι μεταφερόμενες επιθέσεις εκμεταλλεύονται τη μεταφορά δεδομένων μεταξύ μοντέλων, με αμυντική στρατηγική την αύξηση των επιθετικών παραδειγμάτων από μοντέλα υποκατάστασης.

Επιπλέον, οι επιθέσεις με βάση την απόφαση επικεντρώνονται στην τελική απόφαση του μοντέλου και είναι ιδιαίτερα ανθεκτικές. Για να ενισχυθεί η ανθεκτικότητα, η AI πρέπει να μπορεί να τροποποιεί τον εαυτό της, να προσαρμόζεται σε αλλαγές περιβάλλοντος (distributional shift), να επιβιώνει από επιθέσεις και να εξερευνά με ασφάλεια νέες συνθήκες.

Ο Moustapha πρότεινε ένα γενικευμένο πλαίσιο για την αξιοπιστία, το οποίο περιλαμβάνει τέσσερα στοιχεία: 1) μοντέλο υποκατάστασης, 2) αλγόριθμο εκτίμησης αξιοπιστίας, 3) συνάρτηση μάθησης και 4) κριτήριο διακοπής. Η ενεργή μάθηση στοχεύει στη μείωση του κόστους προσομοίωσης και της κατανάλωσης ενέργειας, συνδέοντας έτσι την αξιοπιστία με τη βιωσιμότητα.

2.2.5.3 Βιωσιμότητα

Η βιωσιμότητα είναι κρίσιμο ζήτημα για τη λήψη στρατηγικών αποφάσεων με χρήση τεχνολογιών AI, ιδιαίτερα στις επιχειρηματικές λειτουργίες. Ο Vinuesa εξετάζει τη βιωσιμότητα μέσα από τους στόχους βιώσιμης ανάπτυξης (SDGs), χωρίζοντας τους σε τρεις κατηγορίες: περιβάλλον, οικονομία και κοινωνία. Η γρήγορη ανάπτυξη της AI απαιτεί ρύθμιση για να αποφευχθούν κενά στη διαφάνεια, την ασφάλεια και τα ηθικά ζητήματα.

Περιβαλλοντική Βιωσιμότητα: Η AI χρησιμοποιείται για τη δημιουργία μοντέλων προσομοίωσης για την κλιματική δράση, αλλά και για τη μέτρηση του αποτυπώματος άνθρακα μέσω μεθόδων όπως το CO2-eq, το οποίο υπολογίζει τις εκπομπές αερίων θερμοκηπίου. Ωστόσο, η μέτρηση αυτή εξαρτάται από το υλικό και δεν είναι πάντα δίκαιη μεταξύ διαφορετικών μοντέλων. Ο Schwartz προτείνει τη χρήση των πράξεων κινητής υποδιαστολής (FPO) ως πιο ακριβή μέτρηση της εργασίας που γίνεται από ένα μοντέλο.

Οικονομική Βιωσιμότητα: Η οικονομική βιωσιμότητα αφορά τη μακροπρόθεσμη οικονομική ανάπτυξη με ελάχιστες αρνητικές επιπτώσεις στο κοινωνικό και περιβαλλοντικό επίπεδο. Η λήψη στρατηγικών αποφάσεων απαιτεί πόρους, ανάλυση πρόβλεψης και μοντελοποίηση. Ο Grover προτείνει τα Bayesian Belief Networks (BBN) ως μέθοδο για τη μετάβαση από την αντικειμενική στην υποκειμενική ανάλυση, συνδυάζοντας προηγούμενα αποτελέσματα με παρατηρήσεις για πιο ακριβείς αποφάσεις.

Κοινωνική Βιωσιμότητα: Η κοινωνική βιωσιμότητα εστιάζει στην κοινωνική δικαιοσύνη, τα ανθρώπινα δικαιώματα, την ισότητα και την ευημερία. Οι αποφάσεις για την κοινωνική βιωσιμότητα είναι συχνά δύσκολες λόγω των παραδόξων μεταξύ δικαιοσύνης και ηθικής, όπως η ανοχή και η ισότητα. Οι τεχνολογίες ML έχουν υιοθετηθεί για διάφορες κοινωνικές επιστήμες, όπως δίκτυα, τυχαία δάση και βαθιά μάθηση, ανάλογα με το κοινωνικό πρόβλημα και τα δεδομένα.

Συνολικά, η επιδίωξη ενός βιώσιμου μέλλοντος απαιτεί ολιστική προσέγγιση που συνδυάζει οικονομικά, κοινωνικά και περιβαλλοντικά συστήματα. Οι στρατηγικές αποφάσεις πρέπει να εξισορροπούν στόχους και να λαμβάνουν υπόψη ανεπιθύμητες συνέπειες, εμπλέκοντας τους ενδιαφερόμενους σε συνεργατικές διαδικασίες λήψης αποφάσεων.

2.3 Τομείς Υψηλής Κρισιμότητας σύμφωνα με την Οδηγία NIS2

Η οδηγία NIS2 Directive της Ευρωπαϊκής Ένωσης αποτελεί το βασικό κανονιστικό πλαίσιο για την ενίσχυση της κυβερνοασφάλειας σε κρίσιμες υποδομές και υπηρεσίες. Στο πλαίσιο αυτό, προσδιορίζονται συγκεκριμένοι τομείς υψηλής κρισιμότητας, οι οποίοι θεωρούνται θεμελιώδεις για τη διασφάλιση της οικονομικής σταθερότητας, της κοινωνικής συνοχής και της δημόσιας ασφάλειας. Οι τομείς αυτοί χαρακτηρίζονται από αυξημένη εξάρτηση από πληροφοριακά και επικοινωνιακά συστήματα, γεγονός που τους καθιστά ιδιαίτερα ευάλωτους σε κυβερνοαπειλές.

Σύμφωνα με το Παράρτημα Ι της οδηγίας, οι τομείς υψηλής κρισιμότητας περιλαμβάνουν ένα ευρύ φάσμα δραστηριοτήτων που επηρεάζουν άμεσα τη λειτουργία των σύγχρονων κοινωνιών. Η προστασία τους είναι ζωτικής σημασίας, καθώς πιθανές κυβερνοεπιθέσεις μπορούν να προκαλέσουν εκτεταμένες διαταραχές σε κρίσιμες υπηρεσίες, οικονομικές απώλειες και κινδύνους για την ανθρώπινη ζωή.

Ειδικότερα, ο τομέας της ενέργειας περιλαμβάνει την παραγωγή, μεταφορά και διανομή ηλεκτρικής ενέργειας, φυσικού αερίου και άλλων ενεργειακών πόρων, αποτελώντας βασικό πυλώνα της οικονομικής δραστηριότητας. Ο τομέας των μεταφορών καλύπτει τις αεροπορικές, σιδηροδρομικές, θαλάσσιες και οδικές μεταφορές, διασφαλίζοντας τη μετακίνηση ανθρώπων και αγαθών. Ο τραπεζικός τομέας και οι χρηματοπιστωτικές υποδομές αποτελούν κρίσιμα στοιχεία της οικονομικής λειτουργίας, υποστηρίζοντας τη ροή κεφαλαίων και τις χρηματοοικονομικές συναλλαγές.

Παράλληλα ο τομέας της υγείας διαχειρίζεται ευαίσθητα δεδομένα και κρίσιμες υπηρεσίες περίθαλψης, ενώ οι τομείς του πόσιμου νερού και των λυμάτων σχετίζονται άμεσα με τη δημόσια υγεία και την περιβαλλοντική προστασία. Οι ψηφιακές υποδομές και οι υπηρεσίες ΤΠΕ αποτελούν τη βάση της ψηφιακής οικονομίας, υποστηρίζοντας την πλειονότητα των σύγχρονων υπηρεσιών.

Επιπλέον, η δημόσια διοίκηση εξασφαλίζει τη λειτουργία του κράτους και των δημόσιων υπηρεσιών, ενώ ο τομέας του διαστήματος αναδεικνύεται ως ιδιαίτερα σημαντικός λόγω της εξάρτησης κρίσιμων εφαρμογών από δορυφορικά συστήματα, όπως η πλοήγηση και οι τηλεπικοινωνίες.

Συνολικά, οι τομείς αυτοί συνιστούν το θεμέλιο της σύγχρονης κοινωνίας, και η ενίσχυση της κυβερνοασφάλειάς τους αποτελεί βασική προτεραιότητα της Ευρωπαϊκής Ένωσης. Στο πλαίσιο της παρούσας εργασίας, οι τομείς υψηλής κρισιμότητας αξιοποιούνται ως βάση για τη διαμόρφωση σεναρίων αξιολόγησης των γλωσσικών μοντέλων, με στόχο τη διερεύνηση της ικανότητάς τους να παρέχουν αξιόπιστες και τεχνικά ορθές απαντήσεις σε περιβάλλοντα αυξημένης σημασίας. Ο Πίνακας 1 παρουσιάζει τους έντεκα τομείς υψηλής κρισιμότητας που ορίζονται από την οδηγία NIS2, οι οποίοι αποτελούν βασικούς πυλώνες της ευρωπαϊκής οικονομίας και κοινωνίας και χρήζουν αυξημένων μέτρων κυβερνοασφάλειας.

Α/Α	**Τομέας (Αγγλικά)**	**Τομέας (Ελληνικά)**	**Περιγραφή**
1	Energy	Ενέργεια	Παραγωγή, μεταφορά και διανομή ηλεκτρικής ενέργειας, φυσικού αερίου και καυσίμων
2	Transport	Μεταφορές	Αεροπορικές, σιδηροδρομικές, θαλάσσιες και οδικές μεταφορές
3	Banking	Τραπεζικός Τομέας	Πιστωτικά ιδρύματα και χρηματοοικονομικές υπηρεσίες
4	Financial Market Infrastructure	Χρηματοπιστωτικές Υποδομές	Συστήματα εκκαθάρισης συναλλαγών και χρηματιστήρια
5	Health	Υγεία	Νοσοκομεία, κλινικές και υπηρεσίες υγειονομικής περίθαλψης
6	Drinking Water	Πόσιμο Νερό	Παροχή και διανομή νερού για ανθρώπινη κατανάλωση
7	Wastewater	Λύματα	Διαχείριση και επεξεργασία αποβλήτων
8	Digital Infrastructure	Ψηφιακές Υποδομές	Δίκτυα, data centers, cloud και υπηρεσίες internet
9	ICT Service Management	Υπηρεσίες ΤΠΕ	Πάροχοι υπηρεσιών πληροφορικής και επικοινωνιών
10	Public Administration	Δημόσια Διοίκηση	Κυβερνητικοί οργανισμοί και δημόσιες υπηρεσίες
11	Space	Διάστημα	Δορυφορικά συστήματα και διαστημικές υποδομές

Πίνακας 1: Τομείς Υψηλής Κρισιμότητας της Οδηγίας NIS2

Η ανάλυση των τομέων υψηλής κρισιμότητας που ορίζονται από την NIS2 Directive αναδεικνύει το αυξανόμενο επίπεδο εξάρτησης της σύγχρονης κοινωνίας από ψηφιακά και πληροφοριακά συστήματα, καθώς και τη σημασία της διασφάλισης της κυβερνοασφάλειας σε κρίσιμες υποδομές. Οι τομείς αυτοί δεν αποτελούν απλώς τεχνολογικά περιβάλλοντα, αλλά δυναμικά οικοσυστήματα όπου η ακρίβεια, η αξιοπιστία και η έγκαιρη παροχή πληροφορίας είναι καθοριστικής σημασίας για τη λήψη αποφάσεων.

Στο πλαίσιο αυτό, τα γλωσσικά μοντέλα τεχνητής νοημοσύνης καλούνται να διαδραματίσουν έναν ολοένα και σημαντικότερο ρόλο, παρέχοντας πληροφορίες, τεχνικές εξηγήσεις και υποστήριξη σε χρήστες που δραστηριοποιούνται σε περιβάλλοντα αυξημένης κρισιμότητας. Ωστόσο, η αξιοπιστία των απαντήσεών τους σε τέτοια περιβάλλοντα παραμένει ένα ανοιχτό ερευνητικό ζήτημα, ιδίως όταν οι απαιτήσεις ακρίβειας και ασφάλειας είναι ιδιαίτερα αυστηρές.

Με βάση τα παραπάνω, κρίνεται απαραίτητη η αξιολόγηση των γλωσσικών μοντέλων σε σενάρια που αντικατοπτρίζουν πραγματικές συνθήκες λειτουργίας κρίσιμων υποδομών. Για τον σκοπό αυτό, στην παρούσα εργασία σχεδιάζεται ένα σύνολο στοχευμένων ερωτημάτων (prompts), τα οποία αντιστοιχούν στους έντεκα τομείς υψηλής κρισιμότητας της οδηγίας NIS2. Τα ερωτήματα αυτά διαμορφώνονται με τρόπο ώστε να εξετάζουν την ικανότητα των μοντέλων να παρέχουν έγκυρες, τεχνικά ορθές και αξιόπιστες απαντήσεις σε ζητήματα κυβερνοασφάλειας και λειτουργίας κρίσιμων συστημάτων.

Η μεθοδολογία που ακολουθεί βασίζεται στην υποβολή των ίδιων prompts σε δύο διαφορετικά γλωσσικά μοντέλα, συγκεκριμένα στα GPT-4 και Gemini 1.5 Flash, με στόχο τη συγκριτική αξιολόγηση της απόδοσής τους. Μέσα από τη διαδικασία αυτή, επιδιώκεται η διερεύνηση του βαθμού στον οποίο τα μοντέλα μπορούν να ανταποκριθούν στις απαιτήσεις των σύγχρονων κρίσιμων υποδομών, όπως αυτές προσδιορίζονται από το κανονιστικό πλαίσιο της Ευρωπαϊκής Ένωσης.

Η αναλυτική παρουσίαση της διαδικασίας σχεδιασμού των prompts, καθώς και η εφαρμογή τους στους επιμέρους τομείς υψηλής κρισιμότητας, θα παρουσιαστούν εκτενώς στο Κεφάλαιο 5.4, όπου αναπτύσσεται η πειραματική διαδικασία αξιολόγησης των γλωσσικών μοντέλων και καταγράφονται τα αντίστοιχα αποτελέσματα.

3

Σχετική Έρευνα και Εργαλεία Αξιολόγησης

Η εμπιστοσύνη στην Τεχνητή Νοημοσύνη (AI) αποτελεί κρίσιμο παράγοντα για την αποδοχή και την ευρεία εφαρμογή αυτής της τεχνολογίας σε διάφορους τομείς της κοινωνίας. Η έννοια της Trustworthy AI περιλαμβάνει ηθικές, τεχνικές και κοινωνικές διαστάσεις, οι οποίες πρέπει να συνυπολογίζονται κατά την ανάπτυξη και την αξιολόγηση των συστημάτων AI. Το κεφάλαιο αυτό εξετάζει τη σχετική βιβλιογραφία, τα εργαλεία αξιολόγησης και τη σύνδεσή τους με πρότυπα όπως το NIST AI Risk Management Framework (RMF).

3.1 Παρουσίαση και Ανάλυση της Υπάρχουσας Βιβλιογραφίας και των Σχετικών Surveys για την Trustworthy AI

Η σχετική βιβλιογραφία αποτυπώνει την ανάγκη για αξιόπιστα συστήματα AI που να συμμορφώνονται με βασικές αρχές, όπως η διαφάνεια, η δικαιοσύνη, η ιδιωτικότητα, η ασφάλεια, και η λογοδοσία. Αυτές οι αρχές βρίσκονται στο επίκεντρο των διεθνών κατευθυντήριων γραμμών και των επιστημονικών ερευνών.

3.1.1 Θεωρητικά Πλαίσια και Αρχές

Μια από τις πιο σημαντικές συνεισφορές στη βιβλιογραφία είναι οι κατευθυντήριες γραμμές της Ευρωπαϊκής Επιτροπής για την Trustworthy AI (2019). Το έγγραφο αυτό καθορίζει έξι βασικές αρχές που αποτελούν το θεμέλιο για την ανάπτυξη αξιόπιστων συστημάτων AI:

1. Διαφάνεια: Η λειτουργία της AI πρέπει να είναι επεξηγήσιμη και κατανοητή από τους χρήστες.
2. Δικαιοσύνη: Τα συστήματα AI πρέπει να διασφαλίζουν ίση μεταχείριση και να αποφεύγουν τη μεροληψία.
3. Ιδιωτικότητα: Η προστασία προσωπικών δεδομένων πρέπει να είναι εξασφαλισμένη.
4. Ασφάλεια και Ανθεκτικότητα: Τα συστήματα πρέπει να λειτουργούν αξιόπιστα υπό οποιεσδήποτε συνθήκες.
5. Λογοδοσία: Πρέπει να υπάρχουν μηχανισμοί για την ανάθεση ευθυνών.

6. Ανθρωποκεντρική Προσέγγιση: Η AI πρέπει να σχεδιάζεται για την ευημερία των ανθρώπων [18].

Αντίστοιχα, η μελέτη των Jobin, Ienca και Vayena καταγράφει 84 κατευθυντήριες γραμμές από διάφορους οργανισμούς, αποδεικνύοντας ότι, παρά την ομοφωνία στις αρχές, παραμένουν κενά στις πρακτικές εφαρμογές [19]. Οι Floridi et al. (2019) υποστηρίζουν ότι η επιτυχής ενσωμάτωση της ηθικής στην AI απαιτεί όχι μόνο θεωρητική προσέγγιση αλλά και πρακτικά εργαλεία που υποστηρίζουν την εφαρμογή των αρχών [20].

3.1.2 Εργαλεία Επεξηγησιμότητας

Η επεξηγησιμότητα αποτελεί μία από τις κύριες απαιτήσεις για τη διαφάνεια των συστημάτων AI. Εργαλεία όπως το SHAP (SHapley Additive exPlanations) [21] και το LIME (Local Interpretable Model-agnostic Explanations) [22] έχουν αναπτυχθεί για να εξηγούν τις αποφάσεις που λαμβάνονται από μαύρα κουτιά (black box models). Παρότι αυτά τα εργαλεία παρέχουν κρίσιμες πληροφορίες για τη λειτουργία των συστημάτων, η εφαρμογή τους παραμένει τεχνικά απαιτητική και συχνά περιορίζεται σε εξειδικευμένα περιβάλλοντα.

3.1.3 Σχολιασμός των Κριτηρίων και Μεθόδων Αξιολόγησης από τις Έρευνες

Η αξιοπιστία της AI αξιολογείται μέσω συγκεκριμένων κριτηρίων, όπως η διαφάνεια, η δικαιοσύνη, η ιδιωτικότητα, και η ασφάλεια. Αυτά τα κριτήρια στηρίζονται σε εργαλεία και μεθόδους που αναπτύσσονται συνεχώς για να ανταποκρίνονται στις αυξημένες απαιτήσεις των εφαρμογών AI.

Διαφάνεια

Η διαφάνεια αποτελεί κεντρικό στοιχείο για την οικοδόμηση εμπιστοσύνης. Εργαλεία όπως το SHAP και το LIME παρέχουν τη δυνατότητα ανάλυσης των αποφάσεων των αλγορίθμων AI, επιτρέποντας στους χρήστες να κατανοήσουν πώς και γιατί λαμβάνονται συγκεκριμένες αποφάσεις. Ωστόσο, η αποτελεσματικότητα αυτών των εργαλείων περιορίζεται από την πολυπλοκότητα των σύγχρονων μοντέλων AI και τη δυσκολία ερμηνείας των αποτελεσμάτων από μη τεχνικούς χρήστες [21][22].

Δικαιοσύνη

Η δικαιοσύνη στα συστήματα AI αφορά τη διασφάλιση της ίσης μεταχείρισης όλων των χρηστών, ανεξάρτητα από χαρακτηριστικά όπως η φυλή, το φύλο ή η κοινωνική τάξη. Metrics όπως το Disparate Impact Ratio και το Equalized Odds [23] παρέχουν μεθόδους μέτρησης της μεροληψίας, ενώ αλγόριθμοι fairness-aware, όπως οι Adversarial Debiasing Methods, βοηθούν στη μείωση αυτής [24]. Παρόλα αυτά, η εφαρμογή αυτών των μεθόδων παρουσιάζει δυσκολίες, κυρίως λόγω της ανάγκης για συμβιβασμό μεταξύ δικαιοσύνης και ακρίβειας.

Ιδιωτικότητα

Η προστασία της ιδιωτικότητας αποτελεί θεμελιώδες κριτήριο αξιολόγησης. Τεχνικές όπως η Differential Privacy [25] και το Federated Learning [26] διασφαλίζουν ότι τα προσωπικά δεδομένα των χρηστών προστατεύονται κατά την εκπαίδευση των μοντέλων AI. Εντούτοις, η χρήση αυτών των τεχνικών συχνά οδηγεί σε μείωση της απόδοσης των μοντέλων, κάτι που αποτελεί πρόκληση για τους ερευνητές.

Ασφάλεια

Η ασφάλεια και η ανθεκτικότητα των συστημάτων AI διασφαλίζουν την αξιόπιστη λειτουργία τους, ακόμη και υπό κακόβουλες επιθέσεις ή συνθήκες πίεσης. Εργαλεία όπως το Adversarial Testing [27] και το Stress Testing [28] χρησιμοποιούνται για την ανάλυση της ανθεκτικότητας των συστημάτων, αν και η συνεχής εξέλιξη των απειλών απαιτεί συνεχή βελτίωση των εργαλείων.

3.2 AI Risk Management Framework

Η Εθνική Υπηρεσία Προτύπων και Τεχνολογίας των ΗΠΑ (NIST) έκανε μια σημαντική κίνηση με την εισαγωγή του πρωτοποριακού Πλαισίου Διαχείρισης Κινδύνων για την Τεχνητή Νοημοσύνη (AI RMF) [29] [30]. Σχεδιασμένο να ενισχύσει οργανισμούς και άτομα με ολοκληρωμένες οδηγίες διαχείρισης κινδύνων, το AI RMF στοχεύει στη δημιουργία ενός κόσμου όπου η τεχνητή νοημοσύνη μπορεί να αναπτυχθεί υπεύθυνα.

Αυτή η εξέλιξη δείχνει ότι η NIST αναγνωρίζει την επείγουσα ανάγκη για αποτελεσματική διαχείριση κινδύνων από την τεχνητή νοημοσύνη, καθώς αυτή η τεχνολογία επαναστατεί στη βιομηχανία και αναδιαμορφώνει τις διαδικασίες—από τις εξατομικευμένες συστάσεις και την προγνωστική ανάλυση έως τα αυτόνομα οχήματα και τις προηγμένες ιατρικές διαγνώσεις, το δυναμικό της τεχνητής νοημοσύνης φαίνεται απεριόριστο.

Ωστόσο, με τη μεγάλη δύναμη έρχεται και μεγάλη ευθύνη. Καθώς οι δυνατότητες της τεχνητής νοημοσύνης επεκτείνονται, επεκτείνονται επίσης και οι κίνδυνοι που σχετίζονται με την εφαρμογή της, και με την εξάπλωσή της σε διάφορους τομείς, η διαχείριση των εγγενών κινδύνων έχει γίνει ουσιώδης για να εξασφαλιστεί ένα ασφαλές, υπεύθυνο και αξιόπιστο περιβάλλον για την τεχνητή νοημοσύνη.

Για να αντιμετωπίσει αυτές τις προκλήσεις, η NIST δημιούργησε το AI RMF, το οποίο προσφέρει ένα ευέλικτο σύνολο οδηγιών για να βοηθήσει όσους χρησιμοποιούν την τεχνητή νοημοσύνη—είτε πρόκειται για οργανισμούς του δημόσιου ή ιδιωτικού τομέα είτε για μεμονωμένα άτομα—να κατανοήσουν και να μετριάσουν τους μοναδικούς κινδύνους που θέτουν τα συστήματα τεχνητής νοημοσύνης. Στη συνέχεια, θα αναλύσουμε το πλαίσιο αυτό—συμπεριλαμβανομένων των βασικών πληροφοριών και κύριων λειτουργιών του—ώστε καθώς η χρήση της τεχνητής νοημοσύνης συνεχίζει να ανοίγει νέους ορίζοντες, η οργάνωσή σας να είναι καλύτερα ενημερωμένη και να μπορεί να βρίσκεται στη σωστή πλευρά της προόδου.

3.2.1 Τι είναι το AI RMF της NIST

Καθώς η τεχνητή νοημοσύνη συνεχίζει να ενσωματώνεται σε περισσότερες και περισσότερες πτυχές της κοινωνίας, παραβιάσεις της ιδιωτικότητας, ευπάθειες στην ασφάλεια, προκατειλημμένες αποφάσεις και κοινωνικές επιπτώσεις είναι μεταξύ των βασικών ανησυχιών που απαιτούν άμεση προσοχή.

Για να βοηθήσει, η NIST εισήγαγε το AI RMF, ένα σύνολο οδηγιών και βέλτιστων πρακτικών που μπορούν να βοηθήσουν τους οργανισμούς να διαχειριστούν τους κινδύνους που σχετίζονται με τα συστήματα τεχνητής νοημοσύνης. Το πλαίσιο αυτό παρέχει μια δομημένη προσέγγιση για τον εντοπισμό, την αξιολόγηση και τη μείωση των κινδύνων καθ' όλη τη διάρκεια ζωής ενός συστήματος τεχνητής νοημοσύνης.

3.2.2 Κατηγορίες κινδύνων

Το AI RMF ξεκινά με την καθιέρωση τριών βασικών κατηγοριών κινδύνων που πρέπει να λαμβάνουν υπόψη οι χρήστες τεχνητής νοημοσύνης:

- Κίνδυνοι για τους Ανθρώπους: Εξασφάλιση της προστασίας των ατομικών ελευθεριών, της φυσικής και ψυχολογικής ασφάλειας, και της ίσης πρόσβασης, διατηρώντας τη δημοκρατία και την εκπαίδευση.
- Κίνδυνοι για έναν Οργανισμό: Προστασία από διαταραχές στις επιχειρηματικές λειτουργίες, πιθανούς κινδύνους ασφαλείας λόγω παραβιάσεων και ζημιά στη φήμη.
- Κίνδυνοι για το Οικοσύστημα: Πρόληψη διαταραχών σε παγκόσμια οικονομικά ή εφοδιαστικά συστήματα και μείωση της ζημιάς στους φυσικούς πόρους και το περιβάλλον.

3.2.3 Χαρακτηριστικά Αξιόπιστων Συστημάτων Τεχνητής Νοημοσύνης

Για να αποφεύγονται αυτές οι μορφές κινδύνου, το AI RMF στοχεύει να βοηθήσει τους προγραμματιστές να ενσωματώσουν τις παραμέτρους αξιοπιστίας στον σχεδιασμό, την ανάπτυξη, τη χρήση και την αξιολόγηση προϊόντων, υπηρεσιών και συστημάτων τεχνητής νοημοσύνης. Στο πλαίσιο αυτό, το AI RMF καθορίζει επτά βασικά χαρακτηριστικά για αξιόπιστα συστήματα τεχνητής νοημοσύνης:

- Έγκυρα και Αξιόπιστα: Τα συστήματα τεχνητής νοημοσύνης πρέπει να παρέχουν ακριβή και αξιόπιστα αποτελέσματα.
- Ασφαλή: Πρέπει να δίνουν προτεραιότητα στην ασφάλεια των χρηστών και να αποτρέπουν την πρόκληση ζημιάς.
- Ανθεκτικά: Πρέπει να προστατεύονται από κακόβουλες επιθέσεις και να παραμένουν ανθεκτικά στις προκλήσεις.
- Υπεύθυνα και Διαφανή: Πρέπει να είναι διαφανή και υπεύθυνα για τις αποφάσεις τους.
- Επεξηγήσιμα και Ερμηνεύσιμα: Η εσωτερική λειτουργία τους πρέπει να είναι κατανοητή και ερμηνεύσιμη.
- Ενισχυμένη Ιδιωτικότητα: Πρέπει να σέβονται την ιδιωτικότητα των χρηστών και να προστατεύουν τα προσωπικά δεδομένα.
- Δίκαια: Πρέπει να διασφαλίζουν δικαιοσύνη στα αποτελέσματα και να αντιμετωπίζουν τις επιβλαβείς προκαταλήψεις.

3.2.4 Κύριες Λειτουργίες

Για να επιτευχθεί αυτή η ποιότητα στα συστήματα τεχνητής νοημοσύνης, συνιστάται η τακτική αξιολόγηση της αποτελεσματικότητας της διαχείρισης κινδύνων τεχνητής νοημοσύνης για τη συνεχή βελτίωση της στρατηγικής διαχείρισης και μετριασμού των κινδύνων. Ο πυρήνας του AI RMF, ωστόσο, διαιρεί αυτές τις δραστηριότητες σε τέσσερις αλληλένδετες λειτουργίες: Διακυβέρνηση, Χαρτογράφηση, Μέτρηση και Διαχείριση.

Κάθε λειτουργία περιλαμβάνει διάφορες κατηγορίες και υποκατηγορίες, με συγκεκριμένες προτεινόμενες ενέργειες καθ' όλη τη διάρκεια ζωής του συστήματος τεχνητής νοημοσύνης:

- Χαρτογράφηση (Map): Ως η πρώτη λειτουργία του κύκλου ζωής, ζητά τη συγκέντρωση διαφορετικών απόψεων—συμπεριλαμβανομένων των εσωτερικών ομάδων, εξωτερικών συνεργατών, τελικών χρηστών και οποιουδήποτε άλλου ενδεχομένως επηρεάζεται—ώστε να δομηθεί πιο ολοκληρωμένα το πλαίσιο κινδύνων.
- Μέτρηση (Measure): Μετά τη Χαρτογράφηση, η Μέτρηση απαιτεί από τα συστήματα τεχνητής νοημοσύνης να δοκιμάζονται τόσο πριν από την ανάπτυξή τους όσο και τακτικά κατά τη διάρκεια της λειτουργίας τους, για να διατηρείται μια επίκαιρη κατανόηση της λειτουργικότητας και αξιοπιστίας τους.
- Διαχείριση (Manage): Με την πλήρη κατανόηση και τις τακτικές επανεκτιμήσεις των συστημάτων τεχνητής νοημοσύνης, η Διαχείριση επιτρέπει τη σωστή αντιμετώπιση των κινδύνων, με την κατάλληλη κατανομή πόρων για τη μεγιστοποίηση των πλεονεκτημάτων και την ελαχιστοποίηση των αρνητικών επιπτώσεων.
- Διακυβέρνηση (Govern): Αυτή η λειτουργία υποστηρίζει όλες τις άλλες, παρέχοντας οδηγίες για την υλοποίηση δομών, συστημάτων, διαδικασιών και ομάδων για την ανάπτυξη μιας ισχυρής κουλτούρας διαχείρισης κινδύνων, τη δημιουργία πολιτικών και τη σύνδεση μηχανισμών ανατροφοδότησης για την αποτελεσματική αναγνώριση και μείωση κινδύνων.

4

Γλωσσικά Μοντέλα που θα Αξιολογηθούν

4.1 GPT -4

4.1.1 Εισαγωγή

Η OpenAI, ένας ερευνητικός οργανισμός αφιερωμένος στην προώθηση της τεχνητής νοημοσύνης προς όφελος της ανθρωπότητας, μόλις κυκλοφόρησε το GPT-4 [31] [32] (Generative Pre-trained Transformer 4), την πιο πρόσφατη προσθήκη στη σειρά GPT. Αυτό το επαναστατικό μοντέλο τεχνητής νοημοσύνης είναι ικανό να εκτελεί πολυτροπικές εργασίες που υπερβαίνουν την απλή δημιουργία φυσικού γλωσσικού κειμένου. Το GPT-4 έχει εκπαιδευτεί σε τεράστιο όγκο δεδομένων, γεγονός που του επιτρέπει να παράγει ουσιαστικά και συνεκτικά κείμενα χρησιμοποιώντας την αρχιτεκτονική transformer, μια τυπική προσέγγιση στην επεξεργασία φυσικής γλώσσας. Ωστόσο, αυτό που διαφοροποιεί το GPT-4 είναι η ικανότητά του να αναλύει και να δημιουργεί εικόνες και ήχο, καθιστώντας το ένα ευέλικτο μοντέλο που μπορεί να εκτελεί εργασίες σε πολλαπλές μορφές.

Ένα από τα μεγαλύτερα νέα χαρακτηριστικά του GPT-4 είναι η ικανότητά του να κατανοεί πιο σύνθετες και λεπτομερείς προτροπές. Σύμφωνα με την OpenAI, το GPT-4 "εμφανίζει επιδόσεις επιπέδου ανθρώπου σε διάφορα επαγγελματικά και ακαδημαϊκά πρότυπα." Το GPT-4 είναι επίσης πολύγλωσσο και μπορεί να απαντά σε ερωτήσεις πολλαπλής επιλογής με υψηλή ακρίβεια σε 26 γλώσσες, από τα Ιταλικά έως τα Κορεατικά. Επιπλέον, σύμφωνα με την OpenAI, το GPT-4 είναι ασφαλέστερο και καλύτερα ευθυγραμμισμένο. Διαθέτει ευρύτερες γενικές γνώσεις και βαθύτερη κατανόηση διαφόρων πεδίων. Το GPT-4 είναι επίσης ικανό να επεξεργάζεται εικόνες και βίντεο για να βρίσκει σχετικές και ακριβείς πληροφορίες.

4.1.2 Δυνατότητες

Βαθύτερη κατανόηση: Το GPT-4 είναι ικανό να κατανοεί πιο σύνθετες και λεπτομερείς προτροπές, παρουσιάζοντας επιδόσεις επιπέδου ανθρώπου σε διάφορα επαγγελματικά και ακαδημαϊκά πρότυπα, ξεπερνώντας τον προκάτοχό του GPT-3.5 σε τεστ όπως το SAT, το BAR και το GRE,

χωρίς ειδική εκπαίδευση. Διαθέτει επίσης πολύ μεγαλύτερο όριο λέξεων, καθώς μπορεί να διαχειριστεί εισαγωγές μέχρι 25.000 λέξεις, επιτρέποντας έτσι λεπτομερέστερες προτροπές και παράγοντας εκτενέστερες εξόδους. Επιπλέον, το GPT-4 υποστηρίζει περισσότερες από 26 γλώσσες, συμπεριλαμβανομένων γλωσσών με περιορισμένους πόρους, όπως τα Λετονικά, τα Ουαλικά και τα Σουαχίλι.

Πολυτροπικές δυνατότητες: Μία από τις νεότερες λειτουργίες του GPT-4 είναι οι πολυτροπικές δυνατότητές του, οι οποίες του επιτρέπουν να δέχεται τόσο κείμενα όσο και εικόνες ως εισαγωγές. Αυτό σημαίνει ότι η τεχνητή νοημοσύνη μπορεί να δεχτεί μια εικόνα ως είσοδο, να την ερμηνεύσει και να την κατανοήσει, όπως ακριβώς και μια προτροπή κειμένου. Αυτή η δυνατότητα καλύπτει κάθε μέγεθος και τύπο εικόνας και κειμένου, συμπεριλαμβανομένων εγγράφων που τα συνδυάζουν, χειροποίητων σκίτσων και ακόμη και στιγμιότυπων οθόνης. Αυτή η λειτουργία είναι ιδιαίτερα χρήσιμη για προγραμματιστές που γράφουν κώδικα χρησιμοποιώντας μια διάταξη που καθορίζεται σε ένα προσχέδιο.

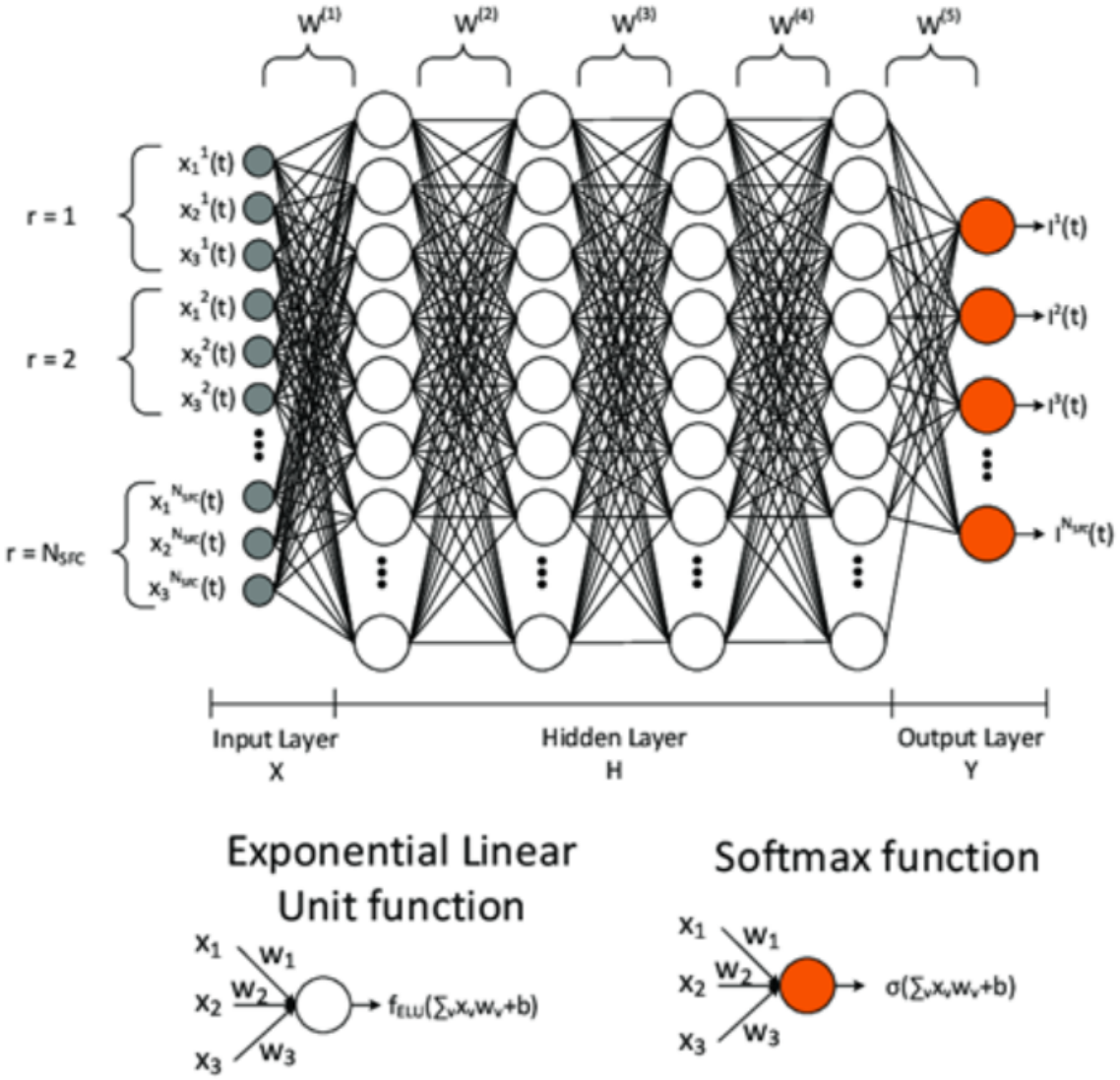
Μεγαλύτερη κατευθυνσιμότητα: Το GPT-4 έχει υψηλό βαθμό κατευθυνσιμότητας, που σημαίνει ότι οι προγραμματιστές μπορούν να καθορίσουν το στυλ και την εργασία της τεχνητής νοημοσύνης περιγράφοντας την κατεύθυνση στο μήνυμα του "συστήματος". Αυτά τα μηνύματα επιτρέπουν στους χρήστες του API να προσαρμόζουν σε μεγάλο βαθμό την εμπειρία χρήστη εντός ορισμένων ορίων. Επιπλέον, είναι λιγότερο πιθανό να αποτύχει όταν εφαρμόζεται σε εφαρμογές που παίζουν έναν συγκεκριμένο ρόλο ή χαρακτήρα.

Ασφάλεια: Η OpenAI έχει κάνει το GPT-4 πιο ασφαλές και πιο ευθυγραμμισμένο, ισχυριζόμενη ότι είναι κατά 82% λιγότερο πιθανό να ανταποκριθεί σε αιτήματα για ακατάλληλο ή μη επιτρεπτό περιεχόμενο, κατά 29% πιο πιθανό να ανταποκριθεί σύμφωνα με τις πολιτικές της OpenAI σε ευαίσθητα αιτήματα, και κατά 40% πιο πιθανό να παράγει ακριβείς απαντήσεις σε σύγκριση με το GPT-3.5. Αυτή είναι μια σημαντική βελτίωση στην ασφάλεια του μοντέλου, καθιστώντας το λιγότερο πιθανό να παράγει επιβλαβείς ή μεροληπτικές απαντήσεις.

Βελτιώσεις Απόδοσης: Η OpenAI αξιολόγησε το GPT-4 σε παραδοσιακά πρότυπα σχεδιασμένα για μοντέλα μηχανικής μάθησης, ισχυριζόμενη ότι "ξεπερνά σημαντικά" τα υπάρχοντα μοντέλα LLM και "τα περισσότερα από τα πιο σύγχρονα μοντέλα." Αυτά τα πρότυπα περιλαμβάνουν τα MMLU, AI2 Reasoning Challenge (ARC), WinoGrande, HumanEval και Drop, όλα τα οποία δοκιμάζουν ατομικές δυνατότητες. Οι βελτιώσεις απόδοσης του GPT-4 το καθιστούν ένα ισχυρό εργαλείο για διάφορες εφαρμογές, από chatbots μέχρι μεταφράσεις γλώσσας και δημιουργία περιεχομένου.

4.1.3 Τρόπος Λειτουργίας

Χρησιμοποιώντας μια βαθιά αρχιτεκτονική νευρωνικού δικτύου, το GPT-4 είναι ικανό να μαθαίνει από τεράστιες ποσότητες δεδομένων κειμένου, επιτρέποντάς του να δημιουργεί κείμενα και να ολοκληρώνει προτροπές με συνεκτικές και φυσικές προτάσεις. Αυτό το μοντέλο έχει εκπαιδευτεί σε έναν τεράστιο όγκο δεδομένων από μια μεγάλη ποικιλία πηγών, όπως μέσα κοινωνικής δικτύωσης, ιστότοπους και βιβλία. Το GPT-4 διαθέτει σημαντικά μεγαλύτερο μέγεθος και ισχύ σε σύγκριση με τον προκάτοχό του, GPT-3, καθώς περιλαμβάνει 170 τρισεκατομμύρια παραμέτρους, σε αντίθεση με τα 175 δισεκατομμύρια παραμέτρους του GPT-3.



Εικόνα 3: Αρχιτεκτονική GPT-4 [33]

Η εκπαίδευση του GPT-4 απαιτεί σημαντική ποσότητα υπολογιστικών πόρων και χρόνου. Εξειδικευμένο υλικό, συμπεριλαμβανομένων GPUs υψηλής απόδοσης, χρησιμοποιείται συνήθως για την επεξεργασία τεράστιων συνόλων δεδομένων και την εκτέλεση περίπλοκων μαθηματικών υπολογισμών. Ανάλογα με το μέγεθος του μοντέλου και τους διαθέσιμους πόρους, η διαδικασία εκπαίδευσης μπορεί να διαρκέσει αρκετές εβδομάδες ή ακόμη και μήνες για να ολοκληρωθεί.

4.1.4 Πεδίο εφαρμογής

Το GPT-4 της OpenAI είναι ένα ισχυρό εργαλείο επεξεργασίας φυσικής γλώσσας με τη δυνατότητα να φέρει επανάσταση σε διάφορους κλάδους. Ακολουθούν μερικές πιθανές εφαρμογές του GPT-4 που θα μπορούσαν να αλλάξουν τον τρόπο που ζούμε και εργαζόμαστε:

- Μετάφραση γλώσσας: Με την ικανότητά του να κατανοεί και να παράγει φυσικό γλωσσικό κείμενο, το GPT-4 θα μπορούσε να βελτιώσει σημαντικά τις εφαρμογές μηχανικής μετάφρασης, παρέχοντας πιο ακριβείς και φυσικές μεταφράσεις.

- Περίληψη κειμένου: Η ικανότητα του GPT-4 να παράγει κείμενα που μοιάζουν με ανθρώπινα το καθιστά ιδανικό για εργασίες όπως η περίληψη κειμένων, όπου το παραγόμενο κείμενο πρέπει να είναι εύκολα κατανοητό.
- Απάντηση σε ερωτήσεις: Το GPT-4 είναι ικανό να απαντά σε ερωτήσεις και να παρέχει λεπτομερείς εξηγήσεις, καθιστώντας το χρήσιμο για εφαρμογές όπως η εξυπηρέτηση πελατών ή η τεχνική υποστήριξη.
- Παραγωγή εικόνων και βίντεο: Η αρχιτεκτονική Transformer του GPT-4 έχει αποδειχθεί αποτελεσματική σε εργασίες μηχανικής μάθησης, συμπεριλαμβανομένης της υπολογιστικής όρασης, γεγονός που το καθιστά δυνητικά χρήσιμο εργαλείο για την παραγωγή εικόνων και βίντεο.
- Άλλες εφαρμογές: Η ευελιξία και η προσαρμοστικότητα του GPT-4 το καθιστούν υποσχόμενο για ένα ευρύ φάσμα εργασιών επεξεργασίας φυσικής γλώσσας, όπως chatbots, αυτοματοποιημένη συγγραφή ειδήσεων και δημιουργική συγγραφή.

4.2 Gemini 1.5 Flash

4.2.1 Εισαγωγή

Το Gemini 1.5 Flash [34] [35] είναι η πιο πρόσφατη προσθήκη στην οικογένεια των μεγάλων γλωσσικών μοντέλων Gemini και έχει σχεδιαστεί ειδικά για να είναι γρήγορο, αποδοτικό και οικονομικό για εργασίες υψηλού όγκου. Αυτό επιτυγχάνεται επειδή είναι ένα πιο ελαφρύ μοντέλο σε σύγκριση με το Gemini 1.5 Pro. Σύμφωνα με την εργασία από την Google DeepMind, το Gemini 1.5 Flash είναι «μια πιο ελαφριά παραλλαγή σχεδιασμένη για αποδοτικότητα με ελάχιστη υποχώρηση στην ποιότητα» και χρησιμοποιεί την αρχιτεκτονική μοντέλου transformer decoder και τις πολυτροπικές δυνατότητες του Gemini 1.5 Pro, «σχεδιασμένη για αποδοτική χρήση των TPUs (Tensor Processing Units) με χαμηλότερη καθυστέρηση για την εξυπηρέτηση του μοντέλου».

Επιπλέον, εκτελεί παράλληλη υπολογιστική επεξεργασία για τα στοιχεία attention και feedforward, και επίσης υφίσταται online distillation από το πολύ μεγαλύτερο μοντέλο Gemini 1.5 Pro. Εκπαιδεύεται με μεθόδους προετοιμασίας ανώτερης τάξης (higher-order preconditioned methods) για βελτιωμένη ποιότητα.

Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο προσαρμόζει επαναληπτικά τις εσωτερικές παραμέτρους του για να ελαχιστοποιήσει μια συνάρτηση απώλειας (Gradient Descent). Αυτή η προσαρμογή καθοδηγείται από τη βαθμίδα της συνάρτησης απώλειας (gradient). Η μέθοδος higher-order preconditioning αναλύει τη συνάρτηση απώλειας και τις παραγώγους της (πληροφορίες ανώτερης τάξης). Βάσει αυτής της ανάλυσης, μπορεί να προσαρμόσει τη βαθμίδα (gradient) καθιστώντας την πιο ακριβή εκπροσώπηση του πώς οι αλλαγές στις παραμέτρους θα επηρεάσουν την απώλεια. Αυτή η στρατηγική βελτιώνει τη σύγκλιση και μειώνει την πιθανότητα το μοντέλο να κολλήσει σε τοπικά ελάχιστα (υποβέλτιστες λύσεις).

4.2.2 Χαρακτηριστικά

Ταχύτητα και Απόδοση: Το Gemini 1.5 Flash είναι το ταχύτερο μοντέλο της σειράς Gemini, καθιστώντας το ιδανικό για εργασίες που απαιτούν αποκρίσεις σε πραγματικό χρόνο, όπως εφαρμογές ιστού και chatbots. Σκεφτείτε μια εφαρμογή ιστού που εκτελείται σε έναν μόνο container

με δυνατότητα αυτόματης κλιμάκωσης (autoscaling). Αν χρησιμοποιήσετε ένα μεγάλο μοντέλο, η εφαρμογή μπορεί να εξυπηρετήσει 5.000 χρήστες μέσα στην ημέρα, καθώς ο χρόνος επεξεργασίας είναι 30 tokens/δευτερόλεπτο. Όταν η ζήτηση αυξηθεί, το autoscaling θα ενεργοποιηθεί, αυξάνοντας το κόστος. Τώρα, σκεφτείτε τον ίδιο container, αλλά με χρόνο επεξεργασίας 60 tokens/δευτερόλεπτο. Με το ίδιο container, μπορείτε να εξυπηρετήσετε πολύ περισσότερους από 5.000 χρήστες ημερησίως, μειώνοντας το κόστος και καθυστερώντας την ανάγκη για autoscaling.

Ακολουθεί ένας πίνακας που συγκρίνει τον μέσο χρόνο ανά χαρακτήρα εξόδου για τα καλύτερα υπάρχοντα μοντέλα (Ιούνιος 2024):

Language	Gemini 1.5 Flash	Gemini 1.5 Pro	GPT-3.5 Turbo	GPT-4 Turbo	Claude 3 Haiku	Claude 3 Sonnet	Claude 3 Opus
English	1.5	4.3	2.6	6.8	2.2	6.2	10.5
Japanese	4.3	10.9	12.9	35.4	10.7	23.9	46.6
Chinese	5.2	14.1	18.4	53.4	12.8	30	55.3
French	1.9	4.7	4.1	10.7	3.0	6.8	13.4

Εικόνα 4: Συγκριτικά στοιχεία μέσων χρόνων [36]

Οικονομική Απόδοση: Επειδή είναι μικρότερο μοντέλο, το 1.5 Flash είναι πιο οικονομικό στη χρήση σε σύγκριση με άλλα μοντέλα Gemini. Είναι, μάλιστα, 1/10 του κόστους του Gemini 1.5 Pro και φθηνότερο από το GPT-3.5.

Εκτεταμένο Παράθυρο Συμφραζομένων: Το Flash 1.5 μπορεί να επεξεργάζεται πληροφορίες από μεγαλύτερο ιστορικό κειμένου, παρέχοντας πιο ολοκληρωμένες και ενημερωτικές απαντήσεις χωρίς να μειώνεται η ακρίβεια. Διαθέτει ένα παράθυρο συμφραζομένων ένα εκατομμύριο tokens από προεπιλογή, το οποίο σημαίνει ότι μπορεί να επεξεργαστεί:

- Μία ώρα βίντεο,
- 11 ώρες ήχου,
- Κώδικες με περισσότερες από 30.000 γραμμές,
- ή πάνω από 700.000 λέξεις.

Αυτή η δυνατότητα εκτεταμένων συμφραζομένων δεν επηρεάζει τις βασικές πολυτροπικές δυνατότητες του μοντέλου. Εκτεταμένες αξιολογήσεις δείχνουν ότι το Gemini 1.5 Flash, αν και μικρότερο και πιο αποδοτικό, παρουσιάζει εντυπωσιακές βελτιώσεις απόδοσης σε σχέση με το Gemini 1.0 Pro.

Πολυτροπική Συλλογιστική: Μπορεί να συλλογίζεται χρησιμοποιώντας διαφορετικούς τύπους δεδομένων, όπως κείμενο, εικόνες, ήχο, βίντεο, PDF και πίνακες μέσα σε εικόνες. Αυτή η δυνατότητα είναι ιδιαίτερα χρήσιμη για σύνθετες εργασίες. Επίσης, υποστηρίζει λειτουργίες κλήσεων API (π.χ. οικονομικά, καιρικές συνθήκες, χάρτες) και δυνατότητα ενημέρωσης με πρόσβαση σε δεδομένα πραγματικού κόσμου και σύγχρονες πληροφορίες.

Υψηλή Απόδοση: Το Gemini 1.5 Flash, αν και μικρότερο, πιο αποδοτικό και ταχύτερο, διατηρεί υψηλά επίπεδα απόδοσης ακόμα και όταν αυξάνεται το παράθυρο συμφραζομένων του.

4.2.3 Πεδίο εφαρμογής

Ακολουθούν μερικά παραδείγματα για το πού μπορεί να χρησιμοποιηθεί το Gemini 1.5 Flash:

Περίληψη Κειμένων: Μπορεί να δημιουργήσει περιλήψεις από μεγάλα έγγραφα, άρθρα, ηχητικά, βίντεο ή τμήματα κειμένου.

Εφαρμογές Συνομιλίας: Μπορεί να χρησιμοποιηθεί για την υποστήριξη chatbots ή εικονικών βοηθών, προσφέροντας φυσικές και αποτελεσματικές αλληλεπιδράσεις.

Υποτιτλισμός και Περίληψη Εικόνων και Βίντεο: Μπορεί να δημιουργήσει υπότιτλους για εικόνες και βίντεο, να εξάγει συγκεκριμένα δεδομένα που περιέχονται στα μέσα και να παρέχει πολύτιμες πληροφορίες.

Εξαγωγή Δεδομένων από Μεγάλα Έγγραφα και Πίνακες: Μπορεί να εξάγει συγκεκριμένες πληροφορίες από μεγάλα έγγραφα και πίνακες, συμπεριλαμβανομένων μη δομημένων PDF και φορμών.

Ενοποίηση Γνώσης από Διάφορες Πηγές: Το Gemini 1.5 Flash μπορεί να αξιοποιήσει την πολυτροπικότητά του για να συλλέξει δεδομένα από πηγές σε διαφορετικές μορφές και στη συνέχεια να εκτελέσει σύνθετες αναλύσεις για αυτά τα δεδομένα.

Ανάκτηση Εμπλουτισμένης Παραγωγής (Retrieval Augmented Generation): Η υψηλής ποιότητας πολυτροπική έξοδός του το καθιστά ιδανικό για RAG (Retrieval Augmented Generation) και για την αναγνώριση οντοτήτων σε Γραφήματα Γνώσης (Knowledge Graphs).

5

Αξιολόγηση Μοντέλων

Στα πλαίσια του συγκεκριμένου κεφαλαίου, παρουσιάζονται 3 κατηγορίες ερωτήσεων με βάση τις οποίες αξιολογήθηκαν τα 2 παραπάνω γλωσσικά μοντέλα, καθώς και οι απαντήσεις που ελήφθησαν από αυτά:

- Εξειδικευμένη Συμβουλευτική
- Δημιουργικότητα
- Επεξήγηση και Διάγνωση Προβλημάτων

5.1 Κατηγορία 1: Εξειδικευμένη Συμβουλευτική

5.1.1 Ερώτηση 1: Τι είναι το μοντέλο "Gradient Boosting" στη μηχανική μάθηση;

Απάντηση GPT-4:

Το "Gradient Boosting" είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιείται για την κατασκευή ισχυρών μοντέλων πρόβλεψης. Αποτελείται από μια σειρά αδύναμων μοντέλων, συνήθως δέντρα απόφασης, που εκπαιδεύονται διαδοχικά. Σε κάθε βήμα, το νέο μοντέλο διορθώνει τα σφάλματα του προηγούμενου, ελαχιστοποιώντας μια συνάρτηση απώλειας με βάση τη μέθοδο gradient descent.

Απάντηση Gemini 1.5 Flash:

Gradient Boosting: Μια Ενισχυμένη Μέθοδος για Ακριβείς Προβλέψεις Το Gradient Boosting είναι μια ισχυρή τεχνική μηχανικής μάθησης που συνδυάζει πολλούς αδύναμους μαθητές (συνήθως δέντρα αποφάσεων) για να δημιουργήσει ένα ισχυρό μοντέλο πρόβλεψης. **Πώς λειτουργεί;** 1. **Αρχική Πρόβλεψη:** Ξεκινάμε με ένα απλό μοντέλο (π.χ., τον μέσο όρο της τιμής στόχου).

2. **Υπολογισμός Σφαλμάτων:** Υπολογίζουμε τα σφάλματα που κάνει αυτό το αρχικό μοντέλο στις προβλέψεις του.
3. **Δημιουργία Νέων Μοντέλων:** Κατασκευάζουμε ένα νέο μοντέλο που προσπαθεί να μειώσει τα σφάλματα του προηγούμενου.
4. **Συνδυασμός Μοντέλων:** Συνδυάζουμε το νέο μοντέλο με τα προηγούμενα, δίνοντας μεγαλύτερο βάρος στα μοντέλα που έχουν καλύτερη απόδοση.
5. **Επανάληψη:** Επαναλαμβάνουμε τα βήματα 3 και 4 για έναν καθορισμένο αριθμό επαναλήψεων ή μέχρι να επιτευχθεί ένα συγκεκριμένο επίπεδο ακρίβειας.

Η διαδικασία αυτή είναι παρόμοια με την εκμάθηση από τα λάθη. Κάθε νέο μοντέλο "μαθαίνει" από τα λάθη των προηγούμενων και προσπαθεί να τα διορθώσει. Με αυτόν τον τρόπο, το τελικό μοντέλο είναι σε θέση να κάνει πολύ πιο ακριβείς προβλέψεις.

Γιατί να χρησιμοποιήσουμε Gradient Boosting;

- **Υψηλή ακρίβεια:** Τα μοντέλα Gradient Boosting συχνά επιτυγχάνουν εξαιρετικά αποτελέσματα σε διάφορα προβλήματα, όπως ταξινόμηση, παλινδρόμηση και ανίχνευση ανωμαλιών.
- **Ευελιξία:** Μπορούν να χειριστούν διαφορετικούς τύπους δεδομένων και να προσαρμοστούν σε διάφορες λειτουργίες απώλειας.
- **Ερμηνευσιμότητα:** Αν και πιο σύνθετα από τα μεμονωμένα δέντρα αποφάσεων, τα μοντέλα Gradient Boosting μπορούν να γίνουν κατανοητά μέσω τεχνικών όπως η ανάλυση σημαντικότητας μεταβλητών.

Δημοφιλείς Αλγόριθμοι Gradient Boosting

- **XGBoost:** Ένας πολύ αποδοτικός και ευέλικτος αλγόριθμος που χρησιμοποιείται ευρέως στην πράξη.
- **LightGBM:** Ένας ταχύτερος και πιο ελαφρύς αλγόριθμος σε σύγκριση με τον XGBoost.
- **CatBoost:** Ένας αλγόριθμος που χειρίζεται κατηγορικές μεταβλητές χωρίς την ανάγκη μετατροπής τους σε αριθμητικές.

Εφαρμογές

Το Gradient Boosting χρησιμοποιείται σε πολλούς τομείς, όπως:

- **Οικονομικά:** Πρόβλεψη τιμών μετοχών, ανίχνευση απάτης.
- **Μάρκετινγκ:** Στόχευση πελατών, ανάλυση συμπεριφοράς πελατών.
- **Υγεία:** Πρόβλεψη ασθενειών, ανάλυση ιατρικών εικόνων.

5.1.2 Ερώτηση 2: Πώς λειτουργούν τα ηλιακά πάνελ;

Απάντηση GPT -4:

Τα ηλιακά πάνελ αποτελούνται από φωτοβολταϊκά κύτταρα που μετατρέπουν την ηλιακή ακτινοβολία σε ηλεκτρική ενέργεια. Όταν το φως του ήλιου πέσει πάνω στα κύτταρα, απελευθερώνει ηλεκτρόνια από τα άτομα τους μέσω του φωτοηλεκτρικού φαινομένου. Αυτά τα ηλεκτρόνια δημιουργούν ηλεκτρικό ρεύμα, το οποίο συλλέγεται και μπορεί να χρησιμοποιηθεί για την τροφοδοσία ηλεκτρικών συσκευών.

Απάντηση Gemini 1.5 Flash

Τα ηλιακά πάνελ είναι συσκευές που μετατρέπουν την ενέργεια του ήλιου σε ηλεκτρική ενέργεια. Αυτό το φαινόμενο ονομάζεται φωτοβολταϊκό φαινόμενο.

Η Επιστήμη Πίσω από τα Πάνελ

- **Φωτόνια και Ηλεκτρόνια:** Όταν το φως του ήλιου, που αποτελείται από μικροσκοπικά πακέτα ενέργειας που ονομάζονται φωτόνια, προσπίπτει σε ένα ηλιακό κύτταρο, τα φωτόνια χτυπούν τα άτομα του υλικού του κυττάρου (συνήθως πυρίτιο). Αυτή η σύγκρουση απελευθερώνει ηλεκτρόνια από τα άτομα, δημιουργώντας ένα ηλεκτρικό ρεύμα.
- **Ηλεκτρικό Πεδίο:** Το ηλιακό κύτταρο είναι κατασκευασμένο με τέτοιο τρόπο ώστε να δημιουργεί ένα ηλεκτρικό πεδίο. Αυτό το πεδίο αναγκάζει τα ελεύθερα ηλεκτρόνια να κινούνται σε μια συγκεκριμένη κατεύθυνση, δημιουργώντας έτσι συνεχές ρεύμα.
- **Σύνδεση των Κυττάρων:** Πολλά ηλιακά κύτταρα συνδέονται σε σειρά για να σχηματίσουν ένα ηλιακό πάνελ. Τα πάνελ μπορούν να συνδεθούν παράλληλα ή σε σειρά για να αυξηθεί η ισχύς που παράγεται.

Η Διαδικασία με Απλά Λόγια

1. **Απορρόφηση Φωτός:** Τα ηλιακά πάνελ απορροφούν το φως του ήλιου.
2. **Δημιουργία Ηλεκτρικού Ρεύματος:** Το φως μετατρέπεται σε ηλεκτρικό ρεύμα μέσα στα ηλιακά κύτταρα.
3. **Συλλογή Ενέργειας:** Το ηλεκτρικό ρεύμα συλλέγεται και μπορεί να χρησιμοποιηθεί για την τροφοδότηση ηλεκτρικών συσκευών ή να αποθηκευτεί σε μπαταρίες.

Πλεονεκτήματα των Ηλιακών Πάνελ

- **Ανανεώσιμη Ενέργεια:** Ο ήλιος είναι μια ανανεώσιμη πηγή ενέργειας, σε αντίθεση με τα ορυκτά καύσιμα.
- **Φιλικά προς το Περιβάλλον:** Τα ηλιακά πάνελ δεν εκπέμπουν ρύπους κατά τη λειτουργία τους.
- **Μειωμένο Ενεργειακό Κόστος:** Με την πάροδο του χρόνου, τα ηλιακά πάνελ μπορούν να μειώσουν σημαντικά το κόστος της ηλεκτρικής ενέργειας.
- **Αυτονομία:** Τα ηλιακά συστήματα μπορούν να παρέχουν ενέργεια σε απομακρυσμένες περιοχές χωρίς πρόσβαση στο ηλεκτρικό δίκτυο.

5.1.3 Ερώτηση 3: Τι σημαίνει "πληθωρισμός" και ποια είναι τα αίτιά του;

Απάντηση GPT -4:

Ο πληθωρισμός είναι η αύξηση του γενικού επιπέδου τιμών αγαθών και υπηρεσιών σε μια οικονομία σε μια χρονική περίοδο. Τα κύρια αίτια περιλαμβάνουν:

1. **Αύξηση ζήτησης** (Demand-pull inflation): Όταν η ζήτηση υπερβαίνει την προσφορά.
2. **Αύξηση κόστους** (Cost-push inflation): Όταν το κόστος παραγωγής αυξάνεται.
3. **Νομισματική πολιτική**: Όταν υπάρχει υπερβολική προσφορά χρήματος.

Απάντηση Gemini 1.5 Flash:

Ο πληθωρισμός είναι η συνεχής αύξηση του γενικού επιπέδου τιμών των αγαθών και των υπηρεσιών σε μια οικονομία σε μια συγκεκριμένη χρονική περίοδο. Αυτό σημαίνει ότι με την πάροδο του χρόνου, τα χρήματα αγοράζουν λιγότερα προϊόντα και υπηρεσίες, μειώνοντας έτσι την αγοραστική δύναμη των καταναλωτών.

Για παράδειγμα: Αν σήμερα ένα κιλό μήλα κοστίζει 2 ευρώ, σε ένα χρόνο μπορεί να κοστίζει 2,20 ευρώ. Όταν αυτό συμβαίνει σε ένα ευρύ φάσμα προϊόντων και υπηρεσιών, τότε μιλάμε για πληθωρισμό.

Ποια είναι τα αίτια του πληθωρισμού;

Τα αίτια του πληθωρισμού είναι πολυεπίπεδα και μπορούν να διαφέρουν από περίπτωση σε περίπτωση. Ωστόσο, οι κύριοι παράγοντες που συνήθως συνδέονται με τον πληθωρισμό είναι:

- **Αύξηση της ζήτησης:** Όταν η ζήτηση για αγαθά και υπηρεσίες υπερβαίνει την προσφορά, οι επιχειρήσεις μπορούν να αυξήσουν τις τιμές τους, καθώς οι καταναλωτές είναι διατεθειμένοι να πληρώσουν περισσότερα για να τα αποκτήσουν.
- **Μείωση της προσφοράς:** Αν η παραγωγή αγαθών και υπηρεσιών μειωθεί λόγω έλλειψης πρώτων υλών, εργατικού δυναμικού ή άλλων παραγόντων, οι τιμές τείνουν να αυξηθούν.
- **Αύξηση του κόστους παραγωγής:** Όταν αυξάνεται το κόστος παραγωγής (π.χ., ενέργεια, εργατικό κόστος), οι επιχειρήσεις μετακυλίουν συνήθως αυτό το αυξημένο κόστος στον καταναλωτή μέσω υψηλότερων τιμών.
- **Αύξηση της προσφοράς χρήματος:** Όταν υπάρχει υπερβολική ποσότητα χρήματος στην αγορά, αυτό μπορεί να οδηγήσει σε αύξηση της ζήτησης και, κατά συνέπεια, σε πληθωρισμό.
- **Προσδοκίες για πληθωρισμό:** Αν οι καταναλωτές και οι επιχειρήσεις αναμένουν ότι οι τιμές θα αυξηθούν στο μέλλον, ενδέχεται να αυξήσουν τις δαπάνες τους ή τις τιμές τους αντίστοιχα, δημιουργώντας έτσι μια αυτοεκπληρούμενη προφητεία.

Είδη πληθωρισμού

Υπάρχουν διάφοροι τύποι πληθωρισμού, όπως:

- **Υπερπληθωρισμός:** Είναι μια εξαιρετικά υψηλή και ανεξέλεγκτη αύξηση των τιμών, που μπορεί να καταστρέψει μια οικονομία.
- **Στάσιμος πληθωρισμός:** Συνδυάζει χαμηλή οικονομική ανάπτυξη με υψηλό πληθωρισμό.
- **Εμφοβιστικός πληθωρισμός:** Είναι ο πληθωρισμός που προκαλείται από τις προσδοκίες των καταναλωτών και των επιχειρήσεων για μελλοντική αύξηση των τιμών.

Η διαχείριση του πληθωρισμού είναι ένας από τους κύριους στόχους της οικονομικής πολιτικής των κυβερνήσεων. Οι κεντρικές τράπεζες χρησιμοποιούν διάφορα εργαλεία, όπως η νομισματική πολιτική, για να διατηρήσουν τον πληθωρισμό σε ένα σταθερό και χαμηλό επίπεδο.

5.1.4 Ερώτηση 4: Ποια είναι τα συμπτώματα της έλλειψης βιταμίνης D;

Απάντηση GPT-4:

Τα κοινά συμπτώματα της έλλειψης βιταμίνης D περιλαμβάνουν:

- Κούραση και αδυναμία.
- Πόνο στα οστά και μυϊκή αδυναμία.
- Κατάθλιψη και αλλαγές στη διάθεση.
- Συχνές λοιμώξεις λόγω εξασθενημένου ανοσοποιητικού.

Απάντηση Gemini 1.5 Flash:

Συμπτώματα Έλλειψης Βιταμίνης D: Όταν ο Ήλιος Λείπει από τον Οργανισμό μας

Η βιταμίνη D, συχνά αποκαλούμενη και "βιταμίνη του ήλιου", παίζει σημαντικό ρόλο σε πολλές λειτουργίες του οργανισμού μας. Από την υγεία των οστών μέχρι την ενίσχυση του ανοσοποιητικού μας συστήματος, η επαρκής πρόσληψή της είναι ζωτικής σημασίας. Όταν όμως υπάρχει έλλειψη, εμφανίζονται μια σειρά από συμπτώματα που μπορεί να επηρεάσουν την καθημερινότητά μας.

Ποια είναι τα πιο συχνά συμπτώματα της έλλειψης βιταμίνης D;

- **Κόπωση και αδυναμία:** Ένα από τα πιο κοινά σημάδια έλλειψης βιταμίνης D είναι η συνεχής αίσθηση κόπωσης και έλλειψης ενέργειας. Ακόμη και μετά από επαρκή ύπνο, μπορεί να νιώθετε εξαντλημένοι.
- **Πόνοι στα οστά και στις αρθρώσεις:** Η βιταμίνη D είναι απαραίτητη για την απορρόφηση του ασβεστίου, το οποίο είναι το κύριο συστατικό των οστών. Η έλλειψή της μπορεί να οδηγήσει σε πόνο, ευαισθησία και αδυναμία στα οστά, καθώς και σε αυξημένο κίνδυνο καταγμάτων.
- **Μυϊκή αδυναμία:** Η βιταμίνη D συμβάλλει στην υγεία των μυών. Η έλλειψή της μπορεί να προκαλέσει μυϊκή αδυναμία και πόνο, καθιστώντας δύσκολες τις καθημερινές δραστηριότητες.
- **Κατάθλιψη και διαταραχές της διάθεσης:** Υπάρχουν ενδείξεις ότι η έλλειψη βιταμίνης D μπορεί να συνδέεται με την εμφάνιση κατάθλιψης και άλλων διαταραχών της διάθεσης.
- **Αδύναμο ανοσοποιητικό σύστημα:** Η βιταμίνη D παίζει σημαντικό ρόλο στην ενίσχυση του ανοσοποιητικού συστήματος. Η έλλειψή της μπορεί να αυξήσει τον κίνδυνο λοιμώξεων.
- **Πρόβλημα στα δόντια:** Η βιταμίνη D είναι απαραίτητη για την υγεία των δοντιών. Η έλλειψή της μπορεί να συμβάλλει στην εμφάνιση τερηδόνας και άλλων οδοντικών προβλημάτων.
- **Χρόνιες παθήσεις:** Η ανεπάρκεια βιταμίνης D έχει συνδεθεί με μια σειρά από χρόνιες παθήσεις, όπως ο διαβήτης, οι καρδιακές παθήσεις και ορισμένοι τύποι καρκίνου.

5.2 Κατηγορία 2: Δημιουργικότητα

5.2.1 Ερώτηση 1: Δημιούργησε μια σύντομη ιστορία για έναν ταξιδιώτη που ανακαλύπτει έναν χαμένο πολιτισμό.

Απάντηση GPT - 4

Ο Λέανδρος, εξερευνητής και γεωγράφος, περπατούσε στις βαθιές ζούγκλες της Νότιας Αμερικής. Κάποια στιγμή, αντίκρισε μια πελώρια πέτρινη πύλη καλυμμένη με αρχαία σύμβολα. Μέσα από την πύλη βρέθηκε σε μια πόλη γεμάτη μαρμάρινους ναούς και κρυστάλλινες λίμνες. Οι κάτοικοι, ντυμένοι με χρυσοκέντητους μανδύες, τον καλωσόρισαν, αναφέροντας ότι ανήκει στον πολιτισμό των Αστέριων, που είχαν εξαφανιστεί πριν από χίλια χρόνια. Η γνώση τους για την αστρονομία και τη θεραπευτική τέχνη άφησε τον Λέανδρο άφωνο. Όταν ήρθε η ώρα να φύγει, υποσχέθηκε να προστατεύσει το μυστικό τους, αφήνοντας πίσω μια νέα κατανόηση του κόσμου.

Απάντηση Gemini 1.5 Flash:

Ο Αλεξάντερ, ένας περιπετειώδης ταξιδιώτης, είχε πάντα ένα πάθος για την εξερεύνηση άγνωστων εδαφών και την ανακάλυψη χαμένων πολιτισμών. Ένα πρωί, ενώ έπινε τον καφέ του σε ένα μικρό καφέ στην Αμαζονοχώρα, άκουσε μια ιστορία για έναν χαμένο πολιτισμό που ζούσε βαθιά μέσα στο τροπικό δάσος.

Η ιστορία τον ενθουσίασε και αποφάσισε να ξεκινήσει μια αποστολή για να βρει αυτόν τον χαμένο πολιτισμό. Εξοπλίστηκε με προμήθειες, χάρτες και ένα GPS, και ξεκίνησε το ταξίδι του.

Οι πρώτες μέρες ήταν δύσκολες. Ο Αλεξάντερ έπρεπε να αντιμετωπίσει την πυκνή βλάστηση, τα επικίνδυνα ζώα και τις δυσμενείς καιρικές συνθήκες. Αλλά η περιέργειά του ήταν πιο δυνατή από την εξάντληση και την απογοήτευση.

Μετά από εβδομάδες ταξιδιού, ο Αλεξάντερ έφτασε σε ένα μικρό χωριό που κατοικούνταν από έναν αρχαίο λαό. Οι άνθρωποι ήταν φιλικοί και φιλόξενοι, και ο Αλεξάντερ ένιωσε σαν να είχε βρεθεί σε άλλο κόσμο.

Ο πολιτισμός τους ήταν μοναδικός και γεμάτος παράδοση. Είχαν τη δική τους γλώσσα, τη δική τους θρησκεία και τη δική τους τέχνη. Ο Αλεξάντερ πέρασε μήνες με τους ανθρώπους, μαθαίνοντας για τη ζωή τους και τη κουλτούρα τους.

Μια μέρα, ο αρχηγός του χωριού αποκάλυψε στον Αλεξάντερ ότι ο πολιτισμός τους ήταν κρυμμένος από τον κόσμο για αιώνες, για να προστατευτεί από την καταστροφή. Ο Αλεξάντερ κατάλαβε ότι είχε την ευκαιρία να μοιραστεί την ιστορία τους με τον κόσμο.

Όταν επέστρεψε στο σπίτι του, ο Αλεξάντερ έγραψε ένα βιβλίο για τον χαμένο πολιτισμό που ανακάλυψε. Το βιβλίο έγινε best seller και ο κόσμος έμαθε για την ύπαρξη αυτού του μοναδικού λαού.

Ο Αλεξάντερ συνέχισε να ταξιδεύει και να εξερευνά, αλλά η ανακάλυψη του χαμένου πολιτισμού παρέμεινε η πιο συναρπαστική εμπειρία της ζωής το

5.2.2 Ερώτηση 2: Δημιούργησε στίχους για ένα τραγούδι εμπνευσμένο από τον χειμώνα.

Απάντηση GPT-4:

Ρεφρέν:
Στον χειμώνα τον λευκό, οι σκιές χορεύουν, μες στο κρύο οι ψυχές τους ήλιους γυρεύουν.

Το χιόνι πέφτει, μα η καρδιά ζεσταίνει κάθε ανάσα, μια φλόγα που καίει.

Στίχοι:
Στα δάση τα παγωμένα, τ' αστέρια σιγομιλούν, οι άνεμοι σαν φύλακες, τα μυστικά φυλούν.

Η νύχτα αγκαλιάζει, τον κόσμο σκεπάζει, μα η αγάπη μας πάντα το σκοτάδι διαπερνά.

Απάντηση Gemini 1.5 Flash:

Στίχοι:

Χιόνια πέφτουν, σιγανά, απαλά σ' ένα χάρτινο κόσμο, μαγικά κρύβονται τα δέντρα, ντυμένα στα λευκά και τα πουλιά σιωπούν, μακριά

Φωτιά τρεμοπαίζει, ζεσταίνει την καρδιά μια ιστορία λέγεται, μια νύχτα μαγεία κάτω από τον ουρανό, γεμάτο αστέρια Ονειρευόμαστε μαζί, μια ζωή καλύτερη

5.3 Κατηγορία 3: Επεξήγηση και Διάγνωση Προβλημάτων

5.3.1 Ερώτηση 1: Ποια είναι τα πλεονεκτήματα και μειονεκτήματα της χρήσης μεγάλων συνόλων δεδομένων για εκπαίδευση AI μοντέλων;

Απάντηση GPT-4:

Πλεονεκτήματα:

1. Μεγαλύτερη ακρίβεια και γενίκευση στα αποτελέσματα.
2. Ανίχνευση λεπτομερών μοτίβων.
3. Καλύτερη απόδοση σε πραγματικές εφαρμογές.

Μειονεκτήματα:

1. Αυξημένες απαιτήσεις σε υπολογιστική ισχύ.
2. Κίνδυνος υπερπροσαρμογής (overfitting) αν τα δεδομένα δεν είναι κατάλληλα ισορροπημένα.
3. Ζητήματα ιδιωτικότητας και δεοντολογίας.

Απάντηση Gemini 1.5 Flash:

Η εκπαίδευση μοντέλων τεχνητής νοημοσύνης (AI) βασίζεται σε μεγάλο βαθμό στην ποιότητα και την ποσότητα των δεδομένων που χρησιμοποιούνται. Τα μεγάλα σύνολα δεδομένων, που περιέχουν τεράστιους όγκους πληροφοριών, προσφέρουν σημαντικά πλεονεκτήματα αλλά και ενέχουν ορισμένους κινδύνους.

Πλεονεκτήματα:

- **Αύξηση της ακρίβειας:** Όσο μεγαλύτερο και ποικιλότερο είναι το σύνολο δεδομένων, τόσο πιο ακριβείς προβλέψεις μπορεί να κάνει ένα μοντέλο AI. Τα μεγάλα σύνολα δεδομένων επιτρέπουν στα μοντέλα να μαθαίνουν από μια ευρύτερη γκάμα παραδειγμάτων, οδηγώντας σε καλύτερη κατανόηση των υποκείμενων σχέσεων.
- **Βελτίωση της γενίκευσης:** Μοντέλα εκπαιδευμένα σε μεγάλα σύνολα δεδομένων είναι πιο πιθανό να γενικεύσουν σε νέα, αόρατα δεδομένα. Αυτό σημαίνει ότι μπορούν να εφαρμοστούν σε πραγματικές καταστάσεις με μεγαλύτερη επιτυχία.
- **Ανάπτυξη πιο πολύπλοκων μοντέλων:** Μεγάλα σύνολα δεδομένων επιτρέπουν την εκπαίδευση πιο πολύπλοκων μοντέλων AI, όπως τα βαθιά νευρωνικά δίκτυα, τα οποία μπορούν να εκτελέσουν πιο σύνθετες εργασίες, όπως η αναγνώριση εικόνων και η επεξεργασία φυσικής γλώσσας.
- **Ανακάλυψη νέων πληροφοριών:** Η ανάλυση μεγάλων συνόλων δεδομένων μπορεί να οδηγήσει στην ανακάλυψη νέων πληροφοριών και τάσεων που δεν ήταν προφανείς από την αρχική παρατήρηση.

Μειονεκτήματα:

- **Κόστος και υπολογιστική ισχύς:** Η συλλογή, αποθήκευση και επεξεργασία μεγάλων συνόλων δεδομένων απαιτεί σημαντικούς πόρους, τόσο σε όρους κόστους όσο και υπολογιστικής ισχύος.
- **Ποιότητα δεδομένων:** Η ποιότητα των δεδομένων είναι εξίσου σημαντική με την ποσότητα. Αν τα δεδομένα είναι θορυβώδη, ελλιπή ή μεροληπτικά, μπορεί να οδηγήσουν σε λανθασμένα αποτελέσματα.
- **Απορρητικότητα:** Η συλλογή και η χρήση προσωπικών δεδομένων εγείρει σοβαρά ζητήματα απορρήτου και ηθικής. Είναι σημαντικό να διασφαλίζεται ότι τα δεδομένα συλλέγονται και χρησιμοποιούνται με ηθικό και νόμιμο τρόπο.
- **Ερμηνευσιμότητα:** Τα πολύπλοκα μοντέλα AI που εκπαιδεύονται σε μεγάλα σύνολα δεδομένων μπορεί να είναι δύσκολο να ερμηνευθούν. Αυτό μπορεί να δυσκολέψει την κατανόηση των αποφάσεων που λαμβάνονται από το μοντέλο και να αυξήσει τον κίνδυνο λανθασμένων συμπερασμάτων.
- **Προκαταλήψεις:** Τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων AI μπορεί να περιέχουν προκαταλήψεις που αντικατοπτρίζουν τις προκαταλήψεις του πραγματικού κόσμου. Αυτές οι προκαταλήψεις μπορούν να ενσωματωθούν στο μοντέλο και να οδηγήσουν σε δυσμενείς επιπτώσεις.

5.3.2 Ερώτηση 2: Γιατί ένας υπολογιστής μπορεί να λειτουργεί αργά;

Απάντηση GPT-4:

1. **Υπερφόρτωση RAM**: Πάρα πολλές εφαρμογές ανοιχτές ταυτόχρονα.
2. **Κακόβουλο λογισμικό**: Ιοί ή malware μπορεί να χρησιμοποιούν πόρους του συστήματος.
3. **Παλιό hardware**: Ο επεξεργαστής ή ο δίσκος μπορεί να μην ανταποκρίνεται στις σύγχρονες απαιτήσεις.

4. **Αποθήκευση**: Έλλειψη χώρου στον σκληρό δίσκο.

Απάντηση Gemini 1.5 Flash:

Υπάρχουν πολλοί λόγοι για τους οποίους ένας υπολογιστής μπορεί να λειτουργεί αργά. Αυτοί οι λόγοι μπορούν να ποικίλλουν από απλά προβλήματα λογισμικού μέχρι πιο σοβαρά προβλήματα υλικού. Ακολουθούν μερικοί από τους πιο συχνούς λόγους:

Λόγοι λογισμικού:

- **Πολλά προγράμματα ανοιχτά ταυτόχρονα:** Όταν έχετε πολλά προγράμματα ανοιχτά, ο υπολογιστής πρέπει να διαχειριστεί περισσότερες διαδικασίες, κάτι που μπορεί να επιβραδύνει τη λειτουργία του.
- **Ιοί και κακόβουλο λογισμικό:** Οι ιοί και το κακόβουλο λογισμικό μπορούν να καταναλώσουν πόρους του συστήματος και να επιβραδύνουν τον υπολογιστή σας.
- **Προγράμματα εκκίνησης:** Πολλά προγράμματα ξεκινούν αυτόματα όταν ανάβετε τον υπολογιστή σας. Όσο περισσότερα προγράμματα εκκίνησης έχετε, τόσο περισσότερος χρόνος χρειάζεται για να ξεκινήσει ο υπολογιστής σας και να λειτουργήσει ομαλά.
- **Λειτουργικό σύστημα:** Ένα παλιό ή κατεστραμμένο λειτουργικό σύστημα μπορεί να προκαλέσει προβλήματα απόδοσης.
- **Περιορισμένος χώρος στον σκληρό δίσκο:** Όταν ο σκληρός δίσκος σας είναι γεμάτος, ο υπολογιστής δυσκολεύεται να αποθηκεύσει και να ανακτήσει αρχεία, κάτι που μπορεί να επιβραδύνει τη λειτουργία του.
- **Προβλήματα με τους οδηγούς:** Εάν οι οδηγοί των συσκευών σας είναι παλιοί ή κατεστραμμένοι, μπορεί να προκαλέσουν προβλήματα συμβατότητας και να επιβραδύνουν τον υπολογιστή σας.

Λόγοι υλικού:

- **Παλαιός επεξεργαστής:** Ένας παλιός επεξεργαστής μπορεί να μην είναι αρκετά ισχυρός για να εκτελέσει τις σύγχρονες εφαρμογές.
- **Λίγη μνήμη RAM:** Η μνήμη RAM είναι απαραίτητη για την εκτέλεση των προγραμμάτων. Αν δεν έχετε αρκετή RAM, ο υπολογιστής σας μπορεί να αρχίσει να χρησιμοποιεί τον σκληρό δίσκο ως εικονική μνήμη, κάτι που είναι πολύ πιο αργό.
- **Προβλήματα με τον σκληρό δίσκο:** Ένας κατεστραμμένος ή κατακερματισμένος σκληρός δίσκος μπορεί να προκαλέσει προβλήματα στην ανάγνωση και τη γραφή των δεδομένων, επιβραδύνοντας έτσι τον υπολογιστή σας.
- **Υπερθέρμανση:** Αν ο υπολογιστής σας υπερθερμαίνεται, μπορεί να μειώσει αυτόματα την ταχύτητα του επεξεργαστή για να προστατευτεί.
- **Βλάβη σε άλλα εξαρτήματα:** Άλλα προβλήματα υλικού, όπως μια χαλασμένη κάρτα γραφικών ή τροφοδοτικό, μπορούν επίσης να προκαλέσουν αργή λειτουργία.

5.3.3 Ερώτηση 3: Ποια είναι η διαφορά μεταξύ κβαντικών υπολογιστών και κλασικών υπολογιστών;

Απάντηση GPT-4:

Οι κλασικοί υπολογιστές χρησιμοποιούν bits, που μπορούν να είναι είτε 0 είτε 1. Αντίθετα, οι κβαντικοί υπολογιστές χρησιμοποιούν qubits, τα οποία μπορούν να βρίσκονται σε κατάσταση 0, 1 ή και στις δύο ταυτόχρονα (υπερθέση). Αυτό επιτρέπει στους κβαντικούς υπολογιστές να λύνουν συγκεκριμένα προβλήματα πολύ πιο γρήγορα από τους κλασικούς.

Απάντηση Gemini 1.5 Flash:

Η διαφορά μεταξύ κβαντικών υπολογιστών και κλασικών υπολογιστών είναι θεμελιώδης και αφορά τον τρόπο με τον οποίο επεξεργάζονται την πληροφορία.

Κλασικοί Υπολογιστές:

- **Βασίζονται σε bits:** Κάθε bit μπορεί να έχει μία από δύο τιμές: 0 ή 1.
- **Λειτουργούν γραμμικά:** Εκτελούν υπολογισμούς βήμα προς βήμα.
- **Κατάλληλοι για τις περισσότερες καθημερινές εργασίες:** Είναι εξαιρετικοί σε εργασίες όπως η επεξεργασία κειμένου, η αναζήτηση στο διαδίκτυο και οι υπολογισμοί.

Κβαντικοί Υπολογιστές:

- **Βασίζονται σε qubits:** Τα qubits μπορούν να βρίσκονται σε πολλαπλές καταστάσεις ταυτόχρονα (υπέρθεση), σε αντίθεση με τα bits.
- **Εκμεταλλεύονται την κβαντική διεμπλοκή:** Τα qubits μπορούν να συνδεθούν μεταξύ τους, ανεξάρτητα από την απόσταση που τα χωρίζει.
- **Παράλληλη επεξεργασία:** Μπορούν να εκτελούν πολλούς υπολογισμούς ταυτόχρονα, κάτι που τους κάνει ιδιαίτερα αποτελεσματικούς σε ορισμένα προβλήματα.
- **Κατάλληλοι για πολύπλοκα προβλήματα:** Είναι ιδανικοί για την επίλυση προβλημάτων που είναι πολύ δύσκολα ή αδύνατα για τους κλασικούς υπολογιστές, όπως η ανάπτυξη νέων φαρμάκων, η δημιουργία κρυπτογραφικών συστημάτων και η προσομοίωση κβαντικών συστημάτων.

1. Εξειδικευμένη Συμβουλευτική

Κριτήρια Επιλογής:

Τεχνική Εξειδίκευση: Ερωτήσεις που απαιτούν βαθιά γνώση σε συγκεκριμένα επιστημονικά ή τεχνικά θέματα.

Ακρίβεια και Λεπτομέρεια: Αξιολόγηση της ικανότητας των μοντέλων να παρέχουν ακριβείς και λεπτομερείς απαντήσεις.

Εφαρμοσιμότητα: Ερωτήσεις που έχουν πρακτική εφαρμογή και μπορούν να χρησιμοποιηθούν σε πραγματικές καταστάσεις.

2. Δημιουργικότητα

Κριτήρια Επιλογής:

Φαντασία και Καινοτομία: Ερωτήσεις που απαιτούν δημιουργική σκέψη και φαντασία.

Ποικιλία και Πρωτοτυπία: Αξιολόγηση της ικανότητας των μοντέλων να δημιουργούν πρωτότυπο περιεχόμενο.

Συναισθηματική Έκφραση: Ερωτήσεις που επιτρέπουν στα μοντέλα να εκφράσουν συναισθήματα και να δημιουργήσουν συναισθηματικά φορτισμένο περιεχόμενο.

3. Επεξήγηση και Διάγνωση Προβλημάτων

Κριτήρια Επιλογής:

Αναλυτική Σκέψη: Ερωτήσεις που απαιτούν ανάλυση και διάγνωση προβλημάτων.

Ερμηνευτική Ικανότητα: Αξιολόγηση της ικανότητας των μοντέλων να εξηγούν σύνθετα θέματα με σαφήνεια.

Πρακτική Επίλυση: Ερωτήσεις που έχουν πρακτική εφαρμογή στην επίλυση προβλημάτων.

Συνολική Στρατηγική

Η επιλογή αυτών των κατηγοριών ερωτήσεων εξασφαλίζει ότι τα γλωσσικά μοντέλα θα αξιολογηθούν σε ένα ευρύ φάσμα δεξιοτήτων, από την τεχνική εξειδίκευση και την αναλυτική σκέψη μέχρι τη δημιουργικότητα και τη συναισθηματική έκφραση. Αυτό επιτρέπει μια ολοκληρωμένη αξιολόγηση των δυνατοτήτων τους και της αποτελεσματικότητάς τους σε διαφορετικά σενάρια.

5.4 Αξιολόγηση Γλωσσικών Μοντέλων βάσει Τομέων Υψηλής Κρισιμότητας (NIS2)

Στο πλαίσιο της παρούσας διπλωματικής, η αξιολόγηση των γλωσσικών μοντέλων επεκτείνεται ώστε να συμπεριλάβει σενάρια χρήσης που προέρχονται από πραγματικά περιβάλλοντα υ υψηλής κρισιμότητας, όπως αυτά ορίζονται από την NIS2 Directive. Η προσέγγιση αυτή επιτρέπει τη διερεύνηση της απόδοσης των μοντέλων σε τομείς όπου η ακρίβεια, η αξιοπιστία και η ασφάλεια της πληροφορίας είναι ιδιαίτερα κρίσιμες.

Για τον σκοπό αυτό, σχεδιάστηκε ένα σύνολο ερωτημάτων (prompts), τα οποία αντιστοιχούν στους έντεκα τομείς υψηλής κρισιμότητας της οδηγίας NIS2. Για κάθε τομέα δημιουργήθηκαν τουλάχιστον πέντε εξειδικευμένα ερωτήματα, τα οποία στοχεύουν στην αξιολόγηση της ικανότητας των γλωσσικών μοντέλων να παρέχουν τεχνικά ορθές, πλήρεις και αξιόπιστες απαντήσεις.

Η διαδικασία αξιολόγησης πραγματοποιήθηκε μέσω της υποβολής των ίδιων ερωτημάτων στα δύο επιλεγμένα μοντέλα, συγκεκριμένα στο GPT-4 και στο Gemini 1.5 Flash. Οι απαντήσεις των μοντέλων συγκρίθηκαν ως προς την πληρότητα, την ακρίβεια, τη σαφήνεια και τη συνάφεια με το εκάστοτε πεδίο εφαρμογής. Για λόγους διαφάνειας και τεκμηρίωσης της πειραματικής διαδικασίας, στο Παράρτημα της εργασίας παρατίθενται ενδεικτικά στιγμιότυπα οθόνης (screenshots) από τις απαντήσεις των δύο γλωσσικών μοντέλων για τα εξεταζόμενα ερωτήματα. Τα παραρτήματα αυτά επιτρέπουν την επαλήθευση των αποτελεσμάτων και ενισχύουν την αξιοπιστία της μελέτης, παρέχοντας άμεση πρόσβαση στα πρωτογενή δεδομένα της αξιολόγησης. Στις επόμενες υποενότητες παρουσιάζονται τα ερωτήματα που χρησιμοποιήθηκαν για κάθε τομέα, καθώς και ενδεικτική ανάλυση της απόδοσης των μοντέλων.

5.4.1 Τομέας Ενέργειας (Energy)

Ο τομέας της ενέργειας αποτελεί έναν από τους πλέον κρίσιμους τομείς υποδομών, καθώς υποστηρίζει το σύνολο της οικονομικής δραστηριότητας και της κοινωνικής λειτουργίας. Η αυξανόμενη ψηφιοποίηση των ενεργειακών συστημάτων, όπως τα έξυπνα δίκτυα (smart grids) και τα συστήματα βιομηχανικού ελέγχου (SCADA/ICS), έχει αυξήσει σημαντικά την επιφάνεια επίθεσης, καθιστώντας τον τομέα ιδιαίτερα ευάλωτο σε κυβερνοαπειλές.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- Explain the main cybersecurity threats for smart grid infrastructures.
- What types of cyberattacks can affect power generation systems?
- How can ransomware impact energy distribution networks?
- What security measures are required for energy operators under NIS2?
- How can artificial intelligence help detect cyber threats in energy systems?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 εντοπίζει βασικές απειλές για τα smart grids, όπως προηγμένες επίμονες απειλές (APTs), κακόβουλο λογισμικό, ransomware και επιθέσεις άρνησης υπηρεσίας (DoS). Επιπλέον, αναφέρει κινδύνους από επιθέσεις man-in-the-middle, εκμετάλλευση IoT συσκευών και επιθέσεις phishing ή insider threats, οι οποίες μπορούν να οδηγήσουν σε διαταραχές λειτουργίας ή ακόμη και εκτεταμένες διακοπές ρεύματος.

Σχετικά με τα συστήματα παραγωγής ενέργειας, το μοντέλο αναφέρεται σε επιθέσεις σε συστήματα SCADA/ICS, οι οποίες μπορούν να οδηγήσουν σε διακοπή λειτουργίας εξοπλισμού ή αλλοίωση δεδομένων λειτουργίας. Όσον αφορά το ransomware, επισημαίνεται ότι μπορεί να καταστήσει μη προσβάσιμα κρίσιμα συστήματα, προκαλώντας καθυστερήσεις στη διανομή ενέργειας και σημαντικές οικονομικές απώλειες.

Σε επίπεδο κανονιστικής συμμόρφωσης, το GPT-4 αναφέρει ότι η οδηγία NIS2 απαιτεί την εφαρμογή μέτρων διαχείρισης κινδύνου, ελέγχου πρόσβασης, κρυπτογράφησης και συνεχούς παρακολούθησης, καθώς και την ανάπτυξη σχεδίων επιχειρησιακής συνέχειας. Τέλος, τονίζεται ο ρόλος της τεχνητής νοημοσύνης στην ανίχνευση ανωμαλιών και την έγκαιρη αναγνώριση απειλών μέσω ανάλυσης δεδομένων σε πραγματικό χρόνο.

Απάντηση Gemini

Το Gemini εστιάζει σε πιο εξειδικευμένες απειλές, όπως οι επιθέσεις από κρατικούς φορείς (APTs) και οι επιθέσεις στην εφοδιαστική αλυσίδα (supply chain attacks). Στον τομέα της παραγωγής ενέργειας, αναφέρεται σε malware ειδικά σχεδιασμένο για βιομηχανικά συστήματα, όπως το Triton, καθώς και σε επιθέσεις ψευδών δεδομένων (False Data Injection), που μπορούν να παραπλανήσουν τους διαχειριστές των συστημάτων.

Σχετικά με το ransomware, το Gemini εισάγει την έννοια της «τετραπλής εκβίασης» (quadruple extortion), όπου οι επιτιθέμενοι όχι μόνο κρυπτογραφούν τα δεδομένα, αλλά και διαρρέουν ευαίσθητες πληροφορίες και εκτελούν επιθέσεις DDoS για να ενισχύσουν την πίεση προς τους οργανισμούς.

Σε ό,τι αφορά την οδηγία NIS2, το μοντέλο αναφέρει συγκεκριμένες απαιτήσεις, όπως η υποχρέωση έγκαιρης ειδοποίησης περιστατικών (εντός 24 ωρών), η ευθύνη διοικητικών στελεχών και ο έλεγχος της εφοδιαστικής αλυσίδας. Επιπλέον, τονίζει τη χρήση προηγμένων τεχνολογιών τεχνητής νοημοσύνης, όπως τα digital twins, για την προσομοίωση και πρόληψη επιθέσεων.

Ανάλυση-Σύγκριση

Η σύγκριση των δύο μοντέλων αναδεικνύει διαφοροποίηση ως προς το επίπεδο τεχνικής εξειδίκευσης και τον τρόπο παρουσίασης της πληροφορίας. Το GPT-4 παρέχει πιο δομημένες και

συνολικές απαντήσεις, καλύπτοντας ένα ευρύ φάσμα απειλών και μέτρων ασφάλειας, με έμφαση στη σαφήνεια και την κατανόηση των βασικών εννοιών. Αντίθετα, το Gemini εστιάζει περισσότερο σε εξειδικευμένα και σύγχρονα σενάρια επιθέσεων, εισάγοντας τεχνικούς όρους και παραδείγματα που προσδίδουν μεγαλύτερο βάθος στην ανάλυση.

Όσον αφορά τη συμμόρφωση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, ωστόσο το Gemini παρουσιάζει πιο συγκεκριμένες κανονιστικές απαιτήσεις, ενώ το GPT-4 προσφέρει πιο γενικευμένη αλλά ολοκληρωμένη εικόνα των μέτρων ασφάλειας. Στον τομέα της τεχνητής νοημοσύνης, και τα δύο μοντέλα αναγνωρίζουν τη σημασία της, με το Gemini να δίνει μεγαλύτερη έμφαση σε προηγμένες εφαρμογές.

Συνολικά, και τα δύο μοντέλα κρίνονται αξιόπιστα, με το GPT-4 να υπερέχει σε σαφήνεια και πληρότητα και το Gemini σε τεχνική εξειδίκευση και καινοτόμες προσεγγίσεις.

5.4.2 Τομέας Μεταφορών (Transport)

Ο τομέας των μεταφορών αποτελεί έναν από τους πιο σύνθετους και κρίσιμους τομείς υποδομών, καθώς περιλαμβάνει ένα ευρύ φάσμα συστημάτων που υποστηρίζουν τη μετακίνηση ανθρώπων και αγαθών. Η ενσωμάτωση ψηφιακών τεχνολογιών, όπως τα έξυπνα συστήματα διαχείρισης κυκλοφορίας, τα αυτόνομα οχήματα και τα αεροδρομιακά πληροφοριακά συστήματα, έχει αυξήσει σημαντικά την εξάρτηση του τομέα από την κυβερνοασφάλεια.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity risks in transportation systems?
- How can cyberattacks affect railway control systems?
- What are the risks of autonomous vehicles in terms of cybersecurity?
- How can airports protect their digital infrastructure?
- What measures does NIS2 require for transport sector security?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 αναγνωρίζει βασικούς κινδύνους για τα συστήματα μεταφορών, όπως επιθέσεις σε συστήματα ελέγχου, κακόβουλο λογισμικό, ransomware και παραβιάσεις δεδομένων. Επισημαίνει ότι η αυξημένη συνδεσιμότητα μέσω IoT, GPS και δικτύων επικοινωνίας ενισχύει την ευπάθεια των υποδομών, γεγονός που μπορεί να οδηγήσει σε διαταραχές λειτουργίας και κινδύνους ασφάλειας.

Στον τομέα των σιδηροδρομικών συστημάτων, αναφέρεται ότι επιθέσεις σε συστήματα σηματοδότησης και SCADA μπορούν να επηρεάσουν τις διαδρομές των τρένων ή να διακόψουν τη λειτουργία τους, προκαλώντας καθυστερήσεις ή ακόμη και ατυχήματα. Επιπλέον, για τα αυτόνομα οχήματα τονίζεται η ευπάθεια λόγω εξάρτησης από αισθητήρες, λογισμικό και συνδεσιμότητα, με πιθανά σενάρια επιθέσεων που περιλαμβάνουν χειραγώγηση πλοήγησης ή μη εξουσιοδοτημένο έλεγχο.

Όσον αφορά τα αεροδρόμια, το GPT-4 προτείνει μέτρα όπως ισχυρός έλεγχος πρόσβασης, παρακολούθηση δικτύων, κρυπτογράφηση, τακτικές ενημερώσεις και εκπαίδευση προσωπικού, καθώς και διαχωρισμό κρίσιμων συστημάτων. Τέλος, αναφέρει ότι η οδηγία NIS2 απαιτεί

εφαρμογή πολιτικών ασφάλειας, διαχείριση κινδύνου, αναφορά περιστατικών και ενίσχυση της ανθεκτικότητας των συστημάτων.

Απάντηση Gemini

Το Gemini παρουσιάζει πιο εξειδικευμένη προσέγγιση, αναδεικνύοντας κινδύνους όπως η συστημική διαταραχή των συστημάτων διαχείρισης κυκλοφορίας και οι επιθέσεις στην εφοδιαστική αλυσίδα μέσω κακόβουλων ενημερώσεων λογισμικού. Επίσης, δίνει έμφαση στην κλοπή ευαίσθητων δεδομένων, όπως βιομετρικά και δεδομένα τοποθεσίας επιβατών.

Στα σιδηροδρομικά συστήματα, αναφέρεται σε επιθέσεις παρεμβολής σημάτων και ψευδείς ενδείξεις παρουσίας συρμών, που μπορούν να παραλύσουν τη λειτουργία των δικτύων. Για τα αυτόνομα οχήματα, εισάγει έννοιες όπως sensor spoofing και έλεγχο μέσω παραβίασης τηλεματικών συστημάτων ή API.

Στον τομέα των αεροδρομίων, το Gemini προτείνει την υιοθέτηση αρχιτεκτονικής Zero Trust, αυστηρό διαχωρισμό δικτύων και χρήση τεχνητής νοημοσύνης για ανίχνευση ανωμαλιών. Σε ό,τι αφορά την οδηγία NIS2, δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η ευθύνη διοίκησης, η άμεση αναφορά περιστατικών εντός 24 ωρών και ο έλεγχος της εφοδιαστικής αλυσίδας.

Ανάλυση-Σύγκριση

Η συγκριτική ανάλυση των δύο μοντέλων δείχνει ότι αμφότερα παρέχουν επαρκή κατανόηση των βασικών κινδύνων κυβερνοασφάλειας στον τομέα των μεταφορών, ωστόσο διαφοροποιούνται ως προς το επίπεδο τεχνικής λεπτομέρειας και εξειδίκευσης. Το GPT-4 παρουσιάζει μια πιο γενικευμένη αλλά ολοκληρωμένη προσέγγιση, καλύπτοντας ένα ευρύ φάσμα θεμάτων με σαφή και δομημένο τρόπο, γεγονός που το καθιστά ιδιαίτερα χρήσιμο για συνολική κατανόηση του πεδίου.

Αντίθετα, το Gemini εστιάζει περισσότερο σε εξειδικευμένες απειλές και τεχνικές έννοιες, όπως sensor spoofing και Zero Trust αρχιτεκτονική, προσδίδοντας μεγαλύτερο βάθος στην ανάλυση. Επιπλέον, παρουσιάζει πιο συγκεκριμένες και πρακτικές προσεγγίσεις όσον αφορά την εφαρμογή μέτρων ασφάλειας.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, ωστόσο το Gemini εμφανίζεται πιο ακριβές ως προς συγκεκριμένες κανονιστικές απαιτήσεις, ενώ το GPT-4 προσφέρει μια πιο συνολική εικόνα της διαχείρισης κινδύνου και της ανθεκτικότητας.

Συνολικά, και τα δύο μοντέλα θεωρούνται κατάλληλα για την ανάλυση του τομέα, με το GPT-4 να υπερέχει σε σαφήνεια και συνοχή και το Gemini σε τεχνική εξειδίκευση και πρακτική προσέγγιση.

5.4.3 Τομέας Τραπεζικών Υπηρεσιών (Banking)

Ο τραπεζικός τομέας αποτελεί έναν από τους πλέον κρίσιμους πυλώνες της σύγχρονης οικονομίας, καθώς διαχειρίζεται ευαίσθητα οικονομικά δεδομένα και υποστηρίζει ένα ευρύ φάσμα χρηματοοικονομικών συναλλαγών. Η αυξανόμενη ψηφιοποίηση των τραπεζικών υπηρεσιών, σε συνδυασμό με την εξάπλωση των online banking πλατφορμών και των fintech εφαρμογών, έχει

ενισχύσει σημαντικά την επιφάνεια επίθεσης, καθιστώντας τον τομέα ιδιαίτερα ευάλωτο σε κυβερνοαπειλές.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity threats faced by banking institutions in modern digital environments?
- How can cyberattacks such as phishing and ransomware impact banking systems and customer data?
- What security measures are required for banks under the NIS2 directive to ensure the protection of financial data?
- How can artificial intelligence be used to detect fraudulent transactions in banking systems?
- What are the risks associated with online banking platforms and how can they be mitigated?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 εντοπίζει βασικές απειλές για τα τραπεζικά ιδρύματα, όπως phishing, κακόβουλο λογισμικό, ransomware, επιθέσεις DDoS και παραβιάσεις δεδομένων. Επισημαίνει ότι η υψηλή ψηφιοποίηση και η ευρεία χρήση online υπηρεσιών καθιστούν τις τράπεζες ιδιαίτερα ελκυστικούς στόχους, με σημαντικούς κινδύνους οικονομικών απωλειών και διαρροής ευαίσθητων δεδομένων πελατών.

Σχετικά με τις επιθέσεις phishing και ransomware, το μοντέλο αναφέρει ότι οι επιθέσεις phishing μπορούν να οδηγήσουν σε κλοπή διαπιστευτηρίων και μη εξουσιοδοτημένη πρόσβαση σε λογαριασμούς, ενώ το ransomware μπορεί να μπλοκάρει κρίσιμα συστήματα, να διακόψει τις υπηρεσίες και να προκαλέσει σημαντική ζημία στη φήμη των οργανισμών.

Σε επίπεδο συμμόρφωσης με την οδηγία NIS2, το GPT-4 επισημαίνει την ανάγκη για εφαρμογή μέτρων διαχείρισης κινδύνου, ισχυρής ταυτοποίησης, κρυπτογράφησης, αναφοράς περιστατικών και συνεχούς παρακολούθησης. Επιπλέον, αναφέρεται στη σημασία της ασφάλειας της εφοδιαστικής αλυσίδας και της επιχειρησιακής συνέχειας.

Όσον αφορά την τεχνητή νοημοσύνη, το GPT-4 υπογραμμίζει τη χρήση της για την ανάλυση συναλλακτικών προτύπων σε πραγματικό χρόνο, επιτρέποντας την ταχεία ανίχνευση ύποπτων δραστηριοτήτων και τη μείωση των ψευδώς θετικών αποτελεσμάτων. Τέλος, επισημαίνει ότι οι κίνδυνοι του online banking μπορούν να μετριαστούν μέσω πολυπαραγοντικής ταυτοποίησης, κρυπτογράφησης και συνεχούς ενημέρωσης των συστημάτων.

Απάντηση Gemini

Το Gemini προσεγγίζει το θέμα με μεγαλύτερη τεχνική εξειδίκευση, αναδεικνύοντας σύγχρονες απειλές όπως τα deepfakes που αξιοποιούνται για επιθέσεις κοινωνικής μηχανικής, καθώς και το μοντέλο Ransomware-as-a-Service (RaaS), το οποίο στοχεύει κρίσιμα τραπεζικά συστήματα. Επιπλέον, τονίζει τη σημασία των επιθέσεων στην εφοδιαστική αλυσίδα μέσω τρίτων παρόχων, όπως fintech εταιρείες και υπηρεσίες cloud.

Σε σχέση με το phishing και το ransomware, το Gemini αναφέρει ότι τα phishing attacks εξελίσσονται μέσω χρήσης τεχνητής νοημοσύνης για τη δημιουργία εξατομικευμένων επιθέσεων,

ενώ το ransomware υιοθετεί τεχνικές «τετραπλής εκβίασης», συνδυάζοντας κρυπτογράφηση δεδομένων, διαρροή πληροφοριών και επιθέσεις DDoS.

Όσον αφορά την οδηγία NIS2, το μοντέλο δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η υποχρέωση έγκαιρης ειδοποίησης εντός 24 ωρών, η ευθύνη της διοίκησης και ο αυστηρός έλεγχος της εφοδιαστικής αλυσίδας. Στον τομέα της τεχνητής νοημοσύνης, το Gemini παρουσιάζει πιο προηγμένες εφαρμογές, όπως η χρήση behavioral biometrics και ανίχνευση μικρο-ανωμαλιών στη συμπεριφορά των χρηστών.

Τέλος, αναδεικνύει κινδύνους όπως το account takeover και οι ευπάθειες API στο open banking, προτείνοντας λύσεις όπως Zero Trust αρχιτεκτονική και βιομετρική επαλήθευση με έλεγχο «ζωντάνιας» (liveness detection).

Ανάλυση-Σύγκριση

Η συγκριτική αξιολόγηση των δύο μοντέλων καταδεικνύει ότι και τα δύο διαθέτουν υψηλό επίπεδο κατανόησης των κινδύνων κυβερνοασφάλειας στον τραπεζικό τομέα, ωστόσο παρουσιάζουν διαφοροποιήσεις ως προς την εστίαση και το επίπεδο τεχνικής λεπτομέρειας. Το GPT-4 παρέχει πιο ισορροπημένες και δομημένες απαντήσεις, καλύπτοντας με σαφήνεια τα βασικά θέματα και προσφέροντας μια ολοκληρωμένη εικόνα των απειλών και των μέτρων αντιμετώπισης.

Αντίθετα, το Gemini εστιάζει περισσότερο σε σύγχρονες και εξελιγμένες μορφές επιθέσεων, όπως deepfakes και RaaS, καθώς και σε τεχνικές αιχμής όπως behavioral biometrics και Zero Trust αρχιτεκτονική. Η προσέγγισή του χαρακτηρίζεται από μεγαλύτερη τεχνική εξειδίκευση και έμφαση σε καινοτόμες μεθόδους.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρουσιάζουν σχετικές πληροφορίες, με το Gemini να προσφέρει πιο συγκεκριμένες κανονιστικές λεπτομέρειες, ενώ το GPT-4 παρέχει μια πιο γενικευμένη αλλά ολοκληρωμένη προσέγγιση της συμμόρφωσης.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια, συνοχή και πληρότητα, ενώ το Gemini σε τεχνική εμβάθυνση και εξειδίκευση, γεγονός που καθιστά τη συνδυαστική αξιοποίηση των δύο μοντέλων ιδιαίτερα χρήσιμη για την κατανόηση του τομέα.

5.4.4 Χρηματοπιστωτικές Υποδομές

Οι χρηματοπιστωτικές υποδομές, όπως τα χρηματιστήρια και τα συστήματα εκκαθάρισης συναλλαγών, αποτελούν κρίσιμα στοιχεία της παγκόσμιας οικονομίας, καθώς διασφαλίζουν την ομαλή λειτουργία των αγορών και τη ροή κεφαλαίων. Η εξάρτησή τους από πληροφοριακά συστήματα υψηλής ταχύτητας και διαθεσιμότητας τις καθιστά ιδιαίτερα ευάλωτες σε κυβερνοεπιθέσεις, με δυνητικά σοβαρές επιπτώσεις στη χρηματοπιστωτική σταθερότητα.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity risks affecting financial market infrastructures such as stock exchanges and clearing systems?
- How can cyberattacks disrupt trading platforms and financial transaction systems?
- What are the potential consequences of a cyberattack on a stock exchange?
- What security requirements does the NIS2 directive impose on financial market infrastructures?

- How can advanced technologies such as artificial intelligence improve the resilience of financial transaction systems?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 αναγνωρίζει βασικούς κινδύνους για τις χρηματοπιστωτικές υποδομές, όπως επιθέσεις DDoS, κακόβουλο λογισμικό, εσωτερικές απειλές και αλλοίωση δεδομένων. Επισημαίνει ότι η υψηλή συνδεσιμότητα και η ανάγκη για λειτουργία σε πραγματικό χρόνο καθιστούν τα συστήματα αυτά ιδιαίτερα ευάλωτα σε διαταραχές και οικονομική απάτη.

Σχετικά με τη διατάραξη των trading platforms, το μοντέλο αναφέρει ότι οι κυβερνοεπιθέσεις μπορούν να επιβραδύνουν ή να διακόψουν τις συναλλαγές, να επηρεάσουν τις τιμές και να προκαλέσουν καθυστερήσεις στις χρηματοοικονομικές πράξεις. Αυτό μπορεί να οδηγήσει σε αστάθεια της αγοράς και απώλεια εμπιστοσύνης από τους επενδυτές.

Σε επίπεδο επιπτώσεων, το GPT-4 επισημαίνει ότι μια επίθεση σε χρηματιστήριο μπορεί να προκαλέσει διακοπή συναλλαγών, εσφαλμένες πράξεις, διαρροή δεδομένων και πανικό στην αγορά, με σημαντικές οικονομικές συνέπειες. Όσον αφορά την οδηγία NIS2, αναφέρει την ανάγκη για ισχυρά μέτρα κυβερνοασφάλειας, όπως διαχείριση κινδύνου, έλεγχο πρόσβασης, κρυπτογράφηση, συνεχή παρακολούθηση και διασφάλιση ανθεκτικότητας.

Τέλος, τονίζεται ο ρόλος της τεχνητής νοημοσύνης στην ανίχνευση ανωμαλιών σε συναλλακτικά πρότυπα, την πρόληψη απάτης και την αυτοματοποίηση των μηχανισμών απόκρισης σε απειλές.

Απάντηση Gemini

Το Gemini προσεγγίζει το θέμα με μεγαλύτερη τεχνική εμβάθυνση, αναδεικνύοντας κινδύνους όπως επιθέσεις αλλοίωσης δεδομένων (data integrity attacks), οι οποίες μπορούν να επηρεάσουν κρίσιμα αρχεία συναλλαγών, καθώς και επιθέσεις από κρατικούς φορείς με στόχο τη μακροχρόνια αποσταθεροποίηση των αγορών.

Σχετικά με τη διατάραξη των συναλλαγών, το μοντέλο αναφέρει τεχνικές όπως η χειραγώγηση αλγορίθμων (algorithm manipulation), η οποία μπορεί να προκαλέσει τεχνητές καταρρεύσεις της αγοράς (flash crashes), καθώς και επιθέσεις DDoS που επηρεάζουν τη ρευστότητα των αγορών.

Όσον αφορά τις επιπτώσεις, το Gemini εισάγει έννοιες όπως το «πάγωμα ρευστότητας» (liquidity freeze) και ο συστημικός κίνδυνος (systemic risk), που μπορούν να οδηγήσουν σε

ευρύτερη οικονομική κρίση. Σε επίπεδο συμμόρφωσης με την οδηγία NIS2, το μοντέλο δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η έγκαιρη αναφορά περιστατικών, ο έλεγχος της εφοδιαστικής αλυσίδας και η ευθύνη της διοίκησης.

Τέλος, το Gemini αναδεικνύει προηγμένες εφαρμογές τεχνητής νοημοσύνης, όπως predictive threat hunting και self-healing architectures, που ενισχύουν την ανθεκτικότητα των συστημάτων μέσω αυτόματης απομόνωσης απειλών.

Ανάλυση-Σύγκριση

Η συγκριτική ανάλυση των δύο μοντέλων καταδεικνύει ότι και τα δύο παρέχουν αξιόπιστες και σχετικές πληροφορίες για τον τομέα των χρηματοπιστωτικών υποδομών, ωστόσο διαφοροποιούνται ως προς το επίπεδο εξειδίκευσης και τον τρόπο προσέγγισης. Το GPT-4

παρουσιάζει μια πιο γενικευμένη και δομημένη ανάλυση, καλύπτοντας τις βασικές έννοιες και τους κύριους κινδύνους με σαφήνεια και συνοχή.

Αντίθετα, το Gemini εστιάζει περισσότερο σε εξειδικευμένες και σύνθετες απειλές, όπως data integrity attacks και algorithm manipulation, προσδίδοντας μεγαλύτερο βάθος στην ανάλυση. Επιπλέον, εισάγει έννοιες συστημικού κινδύνου που αναδεικνύουν τις ευρύτερες επιπτώσεις των κυβερνοεπιθέσεων στις αγορές.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, με το Gemini να παρουσιάζει πιο συγκεκριμένες και αυστηρές κανονιστικές απαιτήσεις, ενώ το GPT-4 προσφέρει μια πιο συνολική και ισορροπημένη προσέγγιση.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια και πληρότητα, ενώ το Gemini σε τεχνική εμβάθυνση και ανάλυση σύνθετων σεναρίων, γεγονός που αναδεικνύει τη συμπληρωματική αξία των δύο μοντέλων.

5.4.5 Υγεία

Ο τομέας της υγείας αποτελεί έναν από τους πιο ευαίσθητους και κρίσιμους τομείς, καθώς συνδέεται άμεσα με την ανθρώπινη ζωή και την παροχή ιατρικών υπηρεσιών. Η αυξανόμενη ψηφιοποίηση των νοσοκομείων και των συστημάτων υγείας, σε συνδυασμό με τη χρήση διασυνδεδεμένων ιατρικών συσκευών (IoMT), έχει αυξήσει σημαντικά την έκθεση του τομέα σε κυβερνοαπειλές. Η προστασία των πληροφοριακών συστημάτων υγείας είναι ζωτικής σημασίας, τόσο για τη διασφάλιση της λειτουργίας των υπηρεσιών όσο και για την προστασία των ευαίσθητων δεδομένων των ασθενών.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity risks affecting healthcare information systems?
- How can cyberattacks impact hospital operations and patient safety?
- What types of sensitive data are targeted in healthcare cyberattacks?
- What security measures are required for healthcare organizations under the NIS2 directive?
- How can artificial intelligence be used to enhance cybersecurity in healthcare systems?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 εντοπίζει βασικούς κινδύνους για τα συστήματα υγείας, όπως ransomware, κακόβουλο λογισμικό, phishing και παραβιάσεις δεδομένων. Επισημαίνει ότι η χρήση διασυνδεδεμένων ιατρικών συσκευών και παλαιών πληροφοριακών συστημάτων αυξάνει σημαντικά την επιφάνεια επίθεσης, καθιστώντας τα νοσοκομεία ιδιαίτερα ευάλωτα.

Σχετικά με τις επιπτώσεις των κυβερνοεπιθέσεων, το μοντέλο αναφέρει ότι μπορούν να διακόψουν τη λειτουργία των νοσοκομείων, να καθυστερήσουν τη διάγνωση και τη θεραπεία και να εμποδίσουν την πρόσβαση σε κρίσιμα δεδομένα ασθενών. Αυτό μπορεί να οδηγήσει σε σοβαρούς κινδύνους για την ασφάλεια των ασθενών.

Όσον αφορά τα δεδομένα που στοχοποιούνται, το GPT-4 αναφέρει προσωπικά στοιχεία, ιατρικούς φακέλους, ασφαλιστικές πληροφορίες και δεδομένα τιμολόγησης, τα οποία έχουν υψηλή αξία για επιθέσεις απάτης και κλοπής ταυτότητας.

Σε επίπεδο συμμόρφωσης με την οδηγία NIS2, το μοντέλο επισημαίνει την ανάγκη για διαχείριση κινδύνου, ισχυρή ταυτοποίηση, κρυπτογράφηση, αναφορά περιστατικών και συνεχή παρακολούθηση, καθώς και εκπαίδευση προσωπικού και διασφάλιση ανθεκτικότητας των συστημάτων. Τέλος, αναδεικνύει τον ρόλο της τεχνητής νοημοσύνης στην ανίχνευση ανωμαλιών και την έγκαιρη αντιμετώπιση απειλών.

Απάντηση Gemini

Το Gemini παρουσιάζει πιο εξειδικευμένη προσέγγιση, αναδεικνύοντας κινδύνους όπως ευπάθειες στην εφοδιαστική αλυσίδα μέσω τρίτων παρόχων λογισμικού και επιθέσεις phishing που αξιοποιούν τεχνητή νοημοσύνη για την εξαπάτηση του προσωπικού.

Σε ό,τι αφορά τις επιπτώσεις, το μοντέλο εισάγει την έννοια του «clinical downtime», όπου οι κυβερνοεπιθέσεις, όπως το ransomware, οδηγούν σε καθυστερήσεις χειρουργικών επεμβάσεων και εκτροπή ασθενών. Επιπλέον, αναφέρεται σε επιθέσεις αλλοίωσης δεδομένων (data integrity attacks), όπου τροποποιούνται κρίσιμες ιατρικές πληροφορίες, όπως ομάδες αίματος ή αλλεργίες, με δυνητικά σοβαρές συνέπειες.

Σχετικά με τα δεδομένα, το Gemini τονίζει τη σημασία των Protected Health Information (PHI) και των γονιδιωματικών δεδομένων, τα οποία έχουν υψηλή αξία λόγω της μόνιμης φύσης τους και της δυνατότητας εκμετάλλευσης για εκβιασμό.

Όσον αφορά την οδηγία NIS2, το μοντέλο παρουσιάζει συγκεκριμένες απαιτήσεις, όπως η έγκαιρη ειδοποίηση περιστατικών, η ευθύνη της διοίκησης και ο διαχωρισμός δικτύων για την προστασία κρίσιμων ιατρικών συσκευών. Τέλος, αναδεικνύει προηγμένες εφαρμογές τεχνητής νοημοσύνης, όπως behavioral biometrics και αυτόματη ανίχνευση απειλών σε συσκευές IoMT.

Ανάλυση-Σύγκριση

Η σύγκριση των δύο μοντέλων αναδεικνύει υψηλό επίπεδο κατανόησης των κινδύνων κυβερνοασφάλειας στον τομέα της υγείας, με σαφείς διαφοροποιήσεις ως προς την εστίαση και την τεχνική εμβάθυνση. Το GPT-4 παρέχει πιο ισορροπημένες και κατανοητές απαντήσεις, καλύπτοντας τις βασικές κατηγορίες απειλών και τα μέτρα προστασίας με σαφή και δομημένο τρόπο.

Αντίθετα, το Gemini εστιάζει σε πιο εξειδικευμένες και σύγχρονες απειλές, όπως data integrity attacks και επιθέσεις σε γονιδιωματικά δεδομένα, προσδίδοντας μεγαλύτερο βάθος και τεχνική ακρίβεια. Επιπλέον, εισάγει έννοιες που σχετίζονται άμεσα με την ασφάλεια των ασθενών, ενισχύοντας τη σημασία της κυβερνοασφάλειας πέρα από την απλή προστασία δεδομένων.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, με το Gemini να προσφέρει πιο συγκεκριμένες κανονιστικές λεπτομέρειες, ενώ το GPT-4 παρουσιάζει μια πιο συνολική και γενικευμένη προσέγγιση.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια και πληρότητα, ενώ το Gemini σε τεχνική εξειδίκευση και ανάλυση σύνθετων σεναρίων, γεγονός που αναδεικνύει τη συμπληρωματική αξία των δύο μοντέλων στον τομέα της υγείας.

5.4.6 Πόσιμο νερό

Ο τομέας του πόσιμου νερού αποτελεί κρίσιμη υποδομή για τη δημόσια υγεία και την κοινωνική ευημερία, καθώς εξασφαλίζει την παροχή ασφαλούς και καθαρού νερού στους πολίτες.

Η λειτουργία των συστημάτων ύδρευσης βασίζεται σε μεγάλο βαθμό σε πληροφοριακά και βιομηχανικά συστήματα ελέγχου (SCADA/ICS), γεγονός που αυξάνει την έκθεσή τους σε κυβερνοαπειλές. Η προστασία των υποδομών αυτών είναι ζωτικής σημασίας, καθώς τυχόν διαταραχές μπορεί να έχουν άμεσες και σοβαρές επιπτώσεις στη δημόσια υγεία.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity threats to drinking water supply systems?
- How can cyberattacks affect water treatment and distribution processes?
- What are the potential consequences of a cyberattack on water supply infrastructure?
- What security measures are required under the NIS2 directive for drinking water providers?
- How can monitoring systems help detect cyber threats in water supply networks?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 αναγνωρίζει βασικές απειλές για τα συστήματα ύδρευσης, όπως ransomware, κακόβουλο λογισμικό, εσωτερικές απειλές και μη εξουσιοδοτημένη πρόσβαση σε συστήματα SCADA/ICS. Επισημαίνει ότι η χρήση παλαιών υποδομών και η ανεπαρκής ασφάλεια απομακρυσμένης πρόσβασης αυξάνουν σημαντικά τους κινδύνους.

Σχετικά με τις επιπτώσεις, το μοντέλο αναφέρει ότι κυβερνοεπιθέσεις μπορούν να επηρεάσουν κρίσιμες λειτουργίες, όπως η δοσολογία χημικών ουσιών, η λειτουργία αντλιών και η ροή του νερού. Αυτό μπορεί να οδηγήσει σε υποβάθμιση της ποιότητας του νερού ή σε διακοπή της παροχής.

Σε επίπεδο συνεπειών, το GPT-4 επισημαίνει κινδύνους όπως μόλυνση του νερού, διακοπές υπηρεσιών, απειλές για τη δημόσια υγεία και απώλεια εμπιστοσύνης των πολιτών. Όσον αφορά την οδηγία NIS2, αναφέρει την ανάγκη για διαχείριση κινδύνου, έλεγχο πρόσβασης, κρυπτογράφηση, συνεχή παρακολούθηση και σχέδια αντιμετώπισης έκτακτων περιστατικών.

Τέλος, αναδεικνύεται ο ρόλος των συστημάτων παρακολούθησης στην ανίχνευση ανωμαλιών, όπως μη φυσιολογικές ροές ή συμπεριφορές συστημάτων, επιτρέποντας την έγκαιρη αντιμετώπιση πιθανών επιθέσεων.

Απάντηση Gemini

Το Gemini προσεγγίζει το θέμα με μεγαλύτερη τεχνική εμβάθυνση, αναδεικνύοντας κινδύνους όπως επιθέσεις από κρατικούς φορείς (nation-state sabotage) και ευπάθειες σε παλαιά συστήματα SCADA με ανεπαρκή κρυπτογράφηση και μη ασφαλή απομακρυσμένη πρόσβαση.

Σχετικά με τη διατάραξη των διαδικασιών, το μοντέλο αναφέρει ότι κυβερνοεπιθέσεις μπορούν να επηρεάσουν τα συστήματα δοσολογίας χημικών ουσιών, οδηγώντας σε επικίνδυνα επίπεδα, ή να στοχεύσουν προγραμματιζόμενους λογικούς ελεγκτές (PLCs), προκαλώντας διακοπές λειτουργίας και βλάβες στο δίκτυο.

Όσον αφορά τις συνέπειες, το Gemini εισάγει πιο ακραία σενάρια, όπως η πιθανότητα δηλητηρίασης του νερού, η οικονομική παράλυση και η πλήρης απώλεια εμπιστοσύνης των πολιτών στις δημόσιες υποδομές.

Σε σχέση με την οδηγία NIS2, το μοντέλο δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η έγκαιρη ειδοποίηση περιστατικών, η ευθύνη της διοίκησης και ο αυστηρός διαχωρισμός των

δικτύων επιχειρησιακής τεχνολογίας (OT) από τα εταιρικά δίκτυα. Επιπλέον, αναδεικνύει τη χρήση τεχνητής νοημοσύνης και digital twins για την ανίχνευση ανωμαλιών και την άμεση αναγνώριση επιθέσεων.

Ανάλυση-Σύγκριση

Η συγκριτική ανάλυση των δύο μοντέλων δείχνει ότι και τα δύο παρέχουν αξιόπιστες και σχετικές πληροφορίες για τον τομέα του πόσιμου νερού, με διαφοροποιήσεις ως προς την προσέγγιση και το επίπεδο τεχνικής εμβάθυνσης. Το GPT-4 παρουσιάζει μια πιο ισορροπημένη και κατανοητή ανάλυση, καλύπτοντας τις βασικές απειλές και τις συνέπειες με σαφήνεια και συνοχή.

Αντίθετα, το Gemini εστιάζει περισσότερο σε εξειδικευμένα και υψηλού κινδύνου σενάρια, όπως επιθέσεις σε συστήματα δοσολογίας και PLCs, καθώς και σε στρατηγικές επιθέσεις από κρατικούς φορείς, προσδίδοντας μεγαλύτερο βάθος και ρεαλισμό στην ανάλυση.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, με το Gemini να παρουσιάζει πιο συγκεκριμένες κανονιστικές απαιτήσεις και τεχνικά μέτρα, ενώ το GPT-4 προσφέρει μια πιο γενικευμένη αλλά ολοκληρωμένη προσέγγιση.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια και πληρότητα, ενώ το Gemini σε τεχνική εξειδίκευση και ανάλυση κρίσιμων σεναρίων, γεγονός που αναδεικνύει τη συμπληρωματική αξία των δύο μοντέλων στον συγκεκριμένο τομέα.

5.4.7 Λύματα

Ο τομέας διαχείρισης λυμάτων αποτελεί κρίσιμη υποδομή για την προστασία της δημόσιας υγείας και του περιβάλλοντος, καθώς εξασφαλίζει την ασφαλή επεξεργασία και διάθεση των αποβλήτων. Η λειτουργία των εγκαταστάσεων βασίζεται σε αυτοματοποιημένα συστήματα ελέγχου, όπως SCADA/ICS, γεγονός που τις καθιστά ευάλωτες σε κυβερνοεπιθέσεις. Η διατάραξη της λειτουργίας των συστημάτων αυτών μπορεί να έχει σοβαρές περιβαλλοντικές και κοινωνικές επιπτώσεις.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity risks affecting wastewater treatment facilities?
- How can cyberattacks disrupt wastewater management and processing systems?
- What are the environmental impacts of a cyberattack on wastewater infrastructure?
- What security measures are required under the NIS2 directive for wastewater service providers?
- How can real-time monitoring systems help detect and prevent cyber threats in wastewater systems?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 αναγνωρίζει βασικούς κινδύνους για τις εγκαταστάσεις επεξεργασίας λυμάτων, όπως ransomware, κακόβουλο λογισμικό, μη εξουσιοδοτημένη πρόσβαση και επιθέσεις σε συστήματα SCADA/ICS. Επισημαίνει ότι η χρήση παλαιών συστημάτων και η ύπαρξη απομακρυσμένης πρόσβασης αυξάνουν την ευπάθεια των υποδομών.

Σχετικά με τη διατάραξη των διαδικασιών, το μοντέλο αναφέρει ότι κυβερνοεπιθέσεις μπορούν να επηρεάσουν τη λειτουργία αντλιών, να αλλοιώσουν τις διαδικασίες επεξεργασίας ή να απενεργοποιήσουν τα συστήματα παρακολούθησης, οδηγώντας σε δυσλειτουργίες ή ανεπαρκή επεξεργασία των λυμάτων.

Σε επίπεδο περιβαλλοντικών επιπτώσεων, το GPT-4 επισημαίνει ότι πιθανές συνέπειες περιλαμβάνουν την απελευθέρωση μη επεξεργασμένων λυμάτων, τη ρύπανση υδάτινων πόρων και τη δημιουργία κινδύνων για τη δημόσια υγεία, καθώς και παραβιάσεις κανονιστικών απαιτήσεων.

Όσον αφορά την οδηγία NIS2, το μοντέλο αναφέρει την ανάγκη για διαχείριση κινδύνου, ισχυρούς ελέγχους πρόσβασης, κρυπτογράφηση, συνεχή παρακολούθηση και σχεδιασμό ανθεκτικότητας, συμπεριλαμβανομένης της ασφάλειας της εφοδιαστικής αλυσίδας. Τέλος, αναδεικνύεται ο ρόλος των συστημάτων παρακολούθησης σε πραγματικό χρόνο για την ανίχνευση ανωμαλιών και την άμεση απόκριση σε απειλές.

Απάντηση Gemini

Το Gemini προσεγγίζει το θέμα με μεγαλύτερη τεχνική εμβάθυνση, αναδεικνύοντας κινδύνους όπως επιθέσεις από κρατικούς φορείς (nation-state sabotage) και ransomware που εκμεταλλεύεται τη σημασία της συνεχούς λειτουργίας των εγκαταστάσεων. Επιπλέον, επισημαίνει ότι οι ευπάθειες προκύπτουν συχνά από μη ασφαλή απομακρυσμένη πρόσβαση σε παλαιά συστήματα SCADA.

Σχετικά με τη διατάραξη των διαδικασιών, το μοντέλο αναφέρει ότι κυβερνοεπιθέσεις μπορούν να επηρεάσουν προγραμματιζόμενους λογικούς ελεγκτές (PLCs), να σταματήσουν αντλίες και να προκαλέσουν υπερχειλίσεις λυμάτων. Επιπλέον, μπορεί να αλλοιώσουν τη δοσολογία χημικών ουσιών, οδηγώντας σε καταστροφή των βιολογικών φίλτρων ή στην απελευθέρωση επικίνδυνων αερίων.

Σε επίπεδο περιβαλλοντικών επιπτώσεων, το Gemini αναδεικνύει σοβαρά σενάρια, όπως η απελευθέρωση ακατέργαστων λυμάτων σε φυσικά οικοσυστήματα, η καταστροφή της υδρόβιας ζωής και η ρύπανση του εδάφους, με σημαντικές νομικές και οικονομικές συνέπειες.

Όσον αφορά την οδηγία NIS2, το μοντέλο δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η έγκαιρη αναφορά περιστατικών, η ευθύνη της διοίκησης και ο διαχωρισμός των δικτύων επιχειρησιακής τεχνολογίας (OT) από τα εταιρικά συστήματα. Επιπλέον, αναδεικνύει τη χρήση προηγμένων τεχνολογιών, όπως ανίχνευση ανωμαλιών μέσω τεχνητής νοημοσύνης και digital twins για την έγκαιρη αναγνώριση απειλών.

Ανάλυση-Σύγκριση

Η συγκριτική ανάλυση των δύο μοντέλων καταδεικνύει ότι και τα δύο παρέχουν αξιόπιστες πληροφορίες για τον τομέα των λυμάτων, με διαφοροποιήσεις ως προς το επίπεδο λεπτομέρειας και τεχνικής εμβάθυνσης. Το GPT-4 προσφέρει μια πιο ισορροπημένη και σαφή προσέγγιση, καλύπτοντας τις βασικές απειλές και επιπτώσεις με κατανοητό τρόπο.

Αντίθετα, το Gemini εστιάζει περισσότερο σε εξειδικευμένα και κρίσιμα σενάρια, όπως επιθέσεις σε PLCs και καταστροφή βιολογικών φίλτρων, προσδίδοντας μεγαλύτερο βάθος και τεχνική ακρίβεια. Επιπλέον, αναδεικνύει πιο έντονα τις περιβαλλοντικές συνέπειες των κυβερνοεπιθέσεων.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, με το Gemini να εμφανίζεται πιο συγκεκριμένο ως προς τις κανονιστικές απαιτήσεις, ενώ το GPT-4 παρέχει μια πιο συνολική και γενικευμένη εικόνα της διαχείρισης κινδύνου.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια και συνοχή, ενώ το Gemini σε τεχνική εξειδίκευση και ανάλυση κρίσιμων σεναρίων, γεγονός που αναδεικνύει τη συμπληρωματική αξία των δύο μοντέλων στον συγκεκριμένο τομέα.

5.4.8 Ψηφιακές Υποδομές

Ο τομέας των ψηφιακών υποδομών αποτελεί τη βάση της σύγχρονης ψηφιακής οικονομίας, καθώς περιλαμβάνει κρίσιμα συστήματα όπως data centers, υπηρεσίες cloud, δίκτυα επικοινωνιών και βασικές διαδικτυακές υπηρεσίες. Η αυξημένη διασυνδεσιμότητα, η εξάρτηση από κατανεμημένα υπολογιστικά περιβάλλοντα και η χρήση κοινόχρηστων πόρων ενισχύουν σημαντικά την επιφάνεια επίθεσης, καθιστώντας τον τομέα ιδιαίτερα εκτεθειμένο σε κυβερνοαπειλές. Η ασφάλεια και η ανθεκτικότητα των ψηφιακών υποδομών είναι κρίσιμης σημασίας, καθώς η διατάραξή τους μπορεί να επηρεάσει άμεσα πλήθος άλλων τομέων.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity threats targeting digital infrastructure such as data centers and cloud services?
- How can cyberattacks affect the availability and integrity of internet services?
- What are the risks associated with Distributed Denial of Service (DDoS) attacks on digital infrastructure?
- What security requirements does the NIS2 directive impose on digital infrastructure providers?
- How can redundancy and resilience strategies improve the security of digital infrastructure systems?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 αναγνωρίζει βασικές απειλές για τις ψηφιακές υποδομές, όπως επιθέσεις DDoS, ransomware, παραβιάσεις δεδομένων, εσωτερικές απειλές και λανθασμένες ρυθμίσεις συστημάτων. Επισημαίνει ότι η υψηλή συνδεσιμότητα και η χρήση κοινόχρηστων υποδομών, όπως τα περιβάλλοντα cloud, αυξάνουν σημαντικά την έκθεση των οργανισμών σε κυβερνοεπιθέσεις.

Σε σχέση με τη διαθεσιμότητα και την ακεραιότητα των υπηρεσιών, το μοντέλο αναφέρει ότι οι κυβερνοεπιθέσεις μπορούν να προκαλέσουν διακοπές υπηρεσιών, επιβράδυνση της απόδοσης και αλλοίωση δεδομένων, επηρεάζοντας ιστοσελίδες, cloud πλατφόρμες και κρίσιμες online υπηρεσίες. Όσον αφορά τις επιθέσεις DDoS, το GPT-4 επισημαίνει ότι η μαζική υπερφόρτωση των εξυπηρετητών μπορεί να καταστήσει τις υπηρεσίες μη διαθέσιμες και να επηρεάσει ταυτόχρονα μεγάλο αριθμό χρηστών.

Σε επίπεδο συμμόρφωσης με την οδηγία NIS2, το GPT-4 αναφέρει την ανάγκη για διαχείριση κινδύνου, ισχυρό έλεγχο πρόσβασης, κρυπτογράφηση, αναφορά περιστατικών, συνεχή παρακολούθηση και ασφάλεια της εφοδιαστικής αλυσίδας. Τέλος, τονίζει τη σημασία της

πλεοναστικότητας και των στρατηγικών ανθεκτικότητας, όπως τα εφεδρικά συστήματα και οι πολλαπλοί εξυπηρετητές, για τη διατήρηση της διαθεσιμότητας και τη γρήγορη αποκατάσταση των υπηρεσιών.

Απάντηση Gemini

Το Gemini προσεγγίζει το θέμα με μεγαλύτερη τεχνική εξειδίκευση, αναδεικνύοντας ως κύριες απειλές τις λανθασμένες ρυθμίσεις στο cloud, οι οποίες αποτελούν βασική αιτία παραβιάσεων δεδομένων, καθώς και τις επιθέσεις στην εφοδιαστική αλυσίδα μέσω διαχειριστικών APIs. Επιπλέον, αναφέρεται σε προηγμένο κακόβουλο λογισμικό που αξιοποιεί τεχνικές απόκρυψης και παραμένει αδρανές για μεγάλα χρονικά διαστήματα.

Σε ό,τι αφορά τη διαθεσιμότητα και την ακεραιότητα, το μοντέλο αναδεικνύει την πιθανότητα μικροδιαταραχών που μπορούν να επηρεάσουν συνδεδεμένα βιομηχανικά αισθητήρια, καθώς και επιθέσεις αλλοίωσης δεδομένων (data poisoning), που ενδέχεται να επηρεάσουν τη λήψη αποφάσεων από αυτοματοποιημένα συστήματα.

Για τις επιθέσεις DDoS, το Gemini παρουσιάζει ένα πιο σύγχρονο και τεχνικό σενάριο, αναφερόμενο σε επιθέσεις terabit-scale που συνδυάζουν μεγάλου όγκου κυκλοφορία με επιθέσεις application-layer, με στόχο όχι μόνο τη διακοπή της υπηρεσίας αλλά και την εξάντληση των πόρων των ομάδων πληροφορικής.

Σε σχέση με την οδηγία NIS2, το Gemini δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η έγκαιρη αναφορά περιστατικών, η ευθύνη της διοίκησης και η εφαρμογή risk management που καλύπτει τόσο κυβερνοαπειλές όσο και φυσικούς κινδύνους. Τέλος, αναδεικνύει προηγμένες στρατηγικές ανθεκτικότητας, όπως immutable backups, self-healing networks και γεωγραφική πλεοναστικότητα.

Ανάλυση-Σύγκριση

Η συγκριτική αξιολόγηση των δύο μοντέλων δείχνει ότι και τα δύο παρέχουν ουσιαστικές και αξιόπιστες πληροφορίες για τον τομέα των ψηφιακών υποδομών, με διακριτές διαφορές ως προς την έμφαση και το επίπεδο τεχνικής ανάλυσης. Το GPT-4 προσφέρει μια πιο ισορροπημένη και εύληπτη προσέγγιση, καλύπτοντας τους βασικούς κινδύνους και τις πρακτικές άμυνας με σαφή και συνεκτικό τρόπο.

Αντίθετα, το Gemini εστιάζει περισσότερο σε σύγχρονες και εξειδικευμένες απειλές, όπως data poisoning, cloud misconfigurations και προηγμένα σενάρια DDoS, προσδίδοντας μεγαλύτερο βάθος και τεχνική εμβάθυνση στην ανάλυση. Επιπλέον, παρουσιάζει πιο προηγμένες έννοιες ανθεκτικότητας, όπως self-healing networks, που αναδεικνύουν μια πιο τεχνολογικά εξελιγμένη προσέγγιση.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, με το Gemini να εμφανίζεται πιο συγκεκριμένο και αυστηρό ως προς τις κανονιστικές απαιτήσεις, ενώ το GPT-4 προσφέρει μια πιο γενικευμένη αλλά ολοκληρωμένη εικόνα.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια, συνοχή και πληρότητα, ενώ το Gemini σε τεχνική εξειδίκευση και ανάλυση σύγχρονων απειλών, γεγονός που καθιστά τη συνδυαστική αξιοποίηση των δύο μοντέλων ιδιαίτερα χρήσιμη για τον συγκεκριμένο τομέα.

5.4.9 Υπηρεσίες ΤΠΕ

Ο τομέας των υπηρεσιών τεχνολογιών πληροφορικής και επικοινωνιών (ΤΠΕ) αποτελεί βασικό πυλώνα της ψηφιακής οικονομίας, καθώς υποστηρίζει τη λειτουργία πολλών άλλων κρίσιμων τομέων, όπως η ενέργεια, οι μεταφορές, η υγεία και το χρηματοπιστωτικό σύστημα. Οι πάροχοι υπηρεσιών ΤΠΕ διαχειρίζονται υποδομές, δεδομένα και υπηρεσίες υψηλής σημασίας, γεγονός που τους καθιστά ελκυστικούς στόχους για κυβερνοεπιθέσεις. Η ασφάλεια των υπηρεσιών αυτών είναι καθοριστική, καθώς τυχόν διαταραχές μπορούν να έχουν αλυσιδωτές επιπτώσεις σε πολλαπλά επίπεδα.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity risks faced by ICT service providers?
- How can cyberattacks on ICT services affect other critical sectors?
- What are the risks associated with third-party ICT service providers and supply chain attacks?
- What security measures are required under the NIS2 directive for ICT service management providers?
- How can incident response strategies improve the resilience of ICT service providers?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 εντοπίζει βασικούς κινδύνους για τους παρόχους υπηρεσιών ΤΠΕ, όπως ransomware, παραβιάσεις δεδομένων, επιθέσεις DDoS, εσωτερικές απειλές και λανθασμένες ρυθμίσεις συστημάτων. Επισημαίνει ότι ο κεντρικός ρόλος των παρόχων αυτών τους καθιστά στόχους υψηλής αξίας για επιτιθέμενους.

Σε ό,τι αφορά τις επιπτώσεις σε άλλους τομείς, το μοντέλο αναφέρει ότι κυβερνοεπιθέσεις σε υπηρεσίες ΤΠΕ μπορούν να προκαλέσουν εκτεταμένες διακοπές σε κρίσιμους τομείς, όπως η ενέργεια, η τραπεζική και η υγεία, οδηγώντας σε λειτουργικές δυσλειτουργίες και απώλεια υπηρεσιών.

Σχετικά με τους κινδύνους της εφοδιαστικής αλυσίδας, το GPT-4 επισημαίνει ότι τρίτοι πάροχοι μπορούν να εισάγουν ευπάθειες μέσω μη ασφαλούς λογισμικού ή πρόσβασης, ενώ οι επιθέσεις supply chain μπορούν να διαδοθούν σε μεγάλο αριθμό οργανισμών.

Σε επίπεδο συμμόρφωσης με την οδηγία NIS2, το μοντέλο αναφέρει την ανάγκη για διαχείριση κινδύνου, έλεγχο πρόσβασης, κρυπτογράφηση, αναφορά περιστατικών, συνεχή παρακολούθηση και ασφάλεια της εφοδιαστικής αλυσίδας. Τέλος, τονίζεται η σημασία των στρατηγικών απόκρισης σε περιστατικά για τον εντοπισμό, την απομόνωση και την αποκατάσταση από επιθέσεις.

Απάντηση Gemini

Το Gemini προσεγγίζει το θέμα με μεγαλύτερη τεχνική εξειδίκευση, αναδεικνύοντας σύγχρονες απειλές όπως η κλοπή διοικητικών διαπιστευτηρίων μέσω deepfakes και η χρήση αυτοματοποιημένου κακόβουλου λογισμικού που κινείται γρήγορα εντός των δικτύων για αποφυγή ανίχνευσης.

Σε ό,τι αφορά τις επιπτώσεις, το μοντέλο εισάγει την έννοια του «skeleton key», όπου μια επιτυχής επίθεση σε πάροχο ΤΠΕ μπορεί να επιτρέψει την ταυτόχρονη πρόσβαση σε πολλαπλούς κρίσιμους τομείς μέσω αξιόπιστων διαύλων επικοινωνίας.

Σχετικά με την εφοδιαστική αλυσίδα, το Gemini αναδεικνύει τους κινδύνους από ευπάθειες σε δευτερογενείς και τριτογενείς προμηθευτές (N-th party vulnerabilities), καθώς και τη χρήση κακόβουλου κώδικα σε βιβλιοθήκες ανοικτού λογισμικού, που μπορεί να δημιουργήσει αόρατες «πίσω πόρτες» στα συστήματα.

Όσον αφορά την οδηγία NIS2, το μοντέλο δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η έγκαιρη ειδοποίηση περιστατικών, η εκπαίδευση της διοίκησης και ο αυστηρός έλεγχος προμηθευτών. Επιπλέον, παρουσιάζει προηγμένες στρατηγικές απόκρισης, όπως τα αυτοματοποιημένα κέντρα επιχειρήσεων ασφάλειας (Agentic SOCs) και τα immutable backups.

Ανάλυση-Σύγκριση

Η συγκριτική αξιολόγηση των δύο μοντέλων αναδεικνύει ότι και τα δύο παρέχουν ουσιαστικές και αξιόπιστες πληροφορίες για τον τομέα των υπηρεσιών ΤΠΕ, με διαφοροποιήσεις ως προς την εστίαση και το επίπεδο τεχνικής εμβάθυνσης. Το GPT-4 προσφέρει μια πιο δομημένη και κατανοητή προσέγγιση, καλύπτοντας τους βασικούς κινδύνους και τις επιπτώσεις με σαφήνεια και συνοχή.

Αντίθετα, το Gemini εστιάζει σε πιο σύνθετα και σύγχρονα σενάρια απειλών, όπως deepfake-based επιθέσεις και N-th party vulnerabilities, προσδίδοντας μεγαλύτερο βάθος και τεχνική ακρίβεια. Επιπλέον, αναδεικνύει πιο έντονα τις συστημικές επιπτώσεις των επιθέσεων σε πολλαπλούς τομείς.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, με το Gemini να παρουσιάζει πιο συγκεκριμένες και εξειδικευμένες απαιτήσεις, ενώ το GPT-4 προσφέρει μια πιο γενικευμένη αλλά ολοκληρωμένη προσέγγιση.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια και συνοχή, ενώ το Gemini σε τεχνική εξειδίκευση και ανάλυση πολύπλοκων απειλών, γεγονός που αναδεικνύει τη συμπληρωματική αξία των δύο μοντέλων στον συγκεκριμένο τομέα.

5.4.10 Δημόσια Διοίκηση

Ο τομέας της δημόσιας διοίκησης αποτελεί βασικό πυλώνα της λειτουργίας του κράτους, καθώς περιλαμβάνει κρίσιμες υπηρεσίες που σχετίζονται με τη διακυβέρνηση, τη δημόσια ασφάλεια και την εξυπηρέτηση των πολιτών. Η αυξανόμενη ψηφιοποίηση των δημόσιων υπηρεσιών και η ανάπτυξη πλατφορμών ηλεκτρονικής διακυβέρνησης (e-governance) έχουν ενισχύσει την αποτελεσματικότητα των διαδικασιών, αλλά παράλληλα έχουν αυξήσει την έκθεση των συστημάτων σε κυβερνοαπειλές. Η προστασία των πληροφοριακών συστημάτων του δημόσιου τομέα είναι ιδιαίτερα σημαντική, καθώς ενδεχόμενες επιθέσεις μπορούν να επηρεάσουν άμεσα την εμπιστοσύνη των πολιτών και τη λειτουργία του κράτους.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity threats affecting public administration systems?
- How can cyberattacks impact government services and citizen data?

- What are the risks associated with digital government platforms and e-governance systems?
- What security measures are required under the NIS2 directive for public sector organizations?
- How can cybersecurity awareness and training improve the protection of public administration systems?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 εντοπίζει βασικές απειλές για τα συστήματα δημόσιας διοίκησης, όπως ransomware, phishing, παραβιάσεις δεδομένων, επιθέσεις DDoS και εσωτερικές απειλές. Επισημαίνει ότι οι κυβερνητικές πλατφόρμες αποτελούν συχνό στόχο λόγω της διαχείρισης ευαίσθητων δεδομένων και της σημασίας των παρεχόμενων υπηρεσιών.

Σχετικά με τις επιπτώσεις, το μοντέλο αναφέρει ότι κυβερνοεπιθέσεις μπορούν να διαταράξουν τη λειτουργία δημόσιων υπηρεσιών, να καθυστερήσουν διοικητικές διαδικασίες και να εκθέσουν προσωπικά δεδομένα πολιτών, οδηγώντας σε απώλεια εμπιστοσύνης και πιθανές περιπτώσεις κλοπής ταυτότητας.

Όσον αφορά τις πλατφόρμες ηλεκτρονικής διακυβέρνησης, το GPT-4 επισημαίνει κινδύνους όπως επιθέσεις hacking, διαρροές δεδομένων, τεχνικές αστοχίες και αδύναμοι μηχανισμοί ταυτοποίησης. Σε επίπεδο συμμόρφωσης με την οδηγία NIS2, αναφέρει την ανάγκη για διαχείριση κινδύνου, ισχυρό έλεγχο πρόσβασης, κρυπτογράφηση, συνεχή παρακολούθηση και ασφάλεια της εφοδιαστικής αλυσίδας.

Τέλος, υπογραμμίζεται η σημασία της εκπαίδευσης του προσωπικού, καθώς η ευαισθητοποίηση των εργαζομένων συμβάλλει στην αναγνώριση απειλών, όπως phishing επιθέσεις, και στη μείωση των ανθρώπινων λαθών.

Απάντηση Gemini

Το Gemini προσεγγίζει το θέμα με μεγαλύτερη τεχνική και στρατηγική εμβάθυνση, αναδεικνύοντας απειλές όπως επιθέσεις κοινωνικής μηχανικής με χρήση τεχνητής νοημοσύνης (deepfakes) και προηγμένες επίμονες απειλές (APTs) που στοχεύουν κυβερνητικά δίκτυα για κατασκοπεία ή μελλοντική δολιοφθορά.

Σε ό,τι αφορά τις επιπτώσεις, το μοντέλο επισημαίνει ότι η παραβίαση κυβερνητικών συστημάτων μπορεί να οδηγήσει σε έκθεση μόνιμων προσωπικών δεδομένων (όπως αριθμοί ταυτότητας και φορολογικά στοιχεία), τα οποία δεν μπορούν να αλλάξουν, καθώς και σε παράλυση βασικών υπηρεσιών, όπως πληρωμές κοινωνικών παροχών ή υπηρεσίες ταυτοποίησης.

Σχετικά με τις πλατφόρμες ηλεκτρονικής διακυβέρνησης, το Gemini αναδεικνύει τον κίνδυνο συγκέντρωσης πολλαπλών υπηρεσιών σε ενιαία συστήματα, δημιουργώντας «σημεία μοναδικής αποτυχίας» (single points of failure). Επιπλέον, επισημαίνει τον κίνδυνο ακούσιας διαρροής δεδομένων μέσω χρήσης εργαλείων τεχνητής νοημοσύνης από το προσωπικό.

Όσον αφορά την οδηγία NIS2, το μοντέλο δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η έγκαιρη ειδοποίηση περιστατικών, η ευθύνη της διοίκησης και ο έλεγχος της εφοδιαστικής αλυσίδας. Τέλος, αναδεικνύει σύγχρονες πρακτικές εκπαίδευσης, όπως προσομοιώσεις επιθέσεων και ανάπτυξη «identity literacy» για την αναγνώριση deepfake απειλών.

Ανάλυση-Σύγκριση

Η συγκριτική ανάλυση των δύο μοντέλων δείχνει ότι και τα δύο παρέχουν επαρκή και αξιόπιστη κατανόηση των κινδύνων κυβερνοασφάλειας στον τομέα της δημόσιας διοίκησης, με σαφείς διαφοροποιήσεις ως προς την προσέγγιση. Το GPT-4 παρουσιάζει μια πιο δομημένη και κατανοητή ανάλυση, καλύπτοντας τις βασικές απειλές και επιπτώσεις με σαφήνεια και συνοχή.

Αντίθετα, το Gemini εστιάζει περισσότερο σε σύγχρονες και εξελιγμένες μορφές επιθέσεων, όπως deepfake-based social engineering και APTs, προσδίδοντας μεγαλύτερο βάθος και στρατηγική διάσταση στην ανάλυση. Επιπλέον, αναδεικνύει σημαντικές προκλήσεις που σχετίζονται με τη συγκέντρωση δεδομένων και την πολυπλοκότητα των ψηφιακών πλατφορμών.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, με το Gemini να παρουσιάζει πιο συγκεκριμένες κανονιστικές απαιτήσεις, ενώ το GPT-4 προσφέρει μια πιο γενικευμένη αλλά ολοκληρωμένη εικόνα.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια και πληρότητα, ενώ το Gemini σε τεχνική εξειδίκευση και ανάλυση σύγχρονων απειλών, γεγονός που αναδεικνύει τη συμπληρωματική αξία των δύο μοντέλων στον τομέα της δημόσιας διοίκησης.

5.4.11 Διάστημα

Ο τομέας του διαστήματος αποτελεί έναν αναδυόμενο αλλά εξαιρετικά κρίσιμο τομέα, καθώς υποστηρίζει βασικές υπηρεσίες όπως οι τηλεπικοινωνίες, η πλοήγηση (GPS/GNSS), η παρακολούθηση του περιβάλλοντος και η άμυνα. Η εξάρτηση από δορυφορικά συστήματα έχει αυξηθεί σημαντικά, καθιστώντας τις διαστημικές υποδομές βασικό στόχο κυβερνοεπιθέσεων. Η πολυπλοκότητα των συστημάτων και η εξάρτηση από επίγειους σταθμούς ελέγχου ενισχύουν τις προκλήσεις στον τομέα της κυβερνοασφάλειας.

Για την αξιολόγηση των γλωσσικών μοντέλων στον συγκεκριμένο τομέα, χρησιμοποιήθηκαν τα ακόλουθα ερωτήματα:

- What are the main cybersecurity risks affecting satellite systems and space infrastructure?
- How can cyberattacks disrupt satellite communication and navigation systems?
- What are the potential consequences of a cyberattack on space-based services such as GPS?
- What security measures are required under the NIS2 directive for space infrastructure?
- How can advanced monitoring and detection systems improve the cybersecurity of space assets?

Απαντήσεις Μοντέλων

Απάντηση GPT-4

Το GPT-4 εντοπίζει βασικούς κινδύνους για τα δορυφορικά και διαστημικά συστήματα, όπως παρεμβολές σημάτων (jamming), spoofing, κακόβουλο λογισμικό και μη εξουσιοδοτημένη πρόσβαση σε επίγειους σταθμούς ελέγχου. Επισημαίνει ότι η απομακρυσμένη και πολύπλοκη φύση των συστημάτων αυτών αυξάνει την ευπάθειά τους.

Σχετικά με τη διατάραξη των υπηρεσιών, το μοντέλο αναφέρει ότι κυβερνοεπιθέσεις μπορούν να επηρεάσουν τη μετάδοση σημάτων, να διακόψουν τις επικοινωνίες ή να αλλοιώσουν δεδομένα πλοήγησης, επηρεάζοντας κρίσιμες υπηρεσίες όπως το GPS και τα δίκτυα τηλεπικοινωνιών.

Όσον αφορά τις συνέπειες, το GPT-4 επισημαίνει ότι επιθέσεις σε συστήματα GPS μπορούν να προκαλέσουν σφάλματα πλοήγησης, προβλήματα στις μεταφορές και κινδύνους για την ασφάλεια, καθώς και επιπτώσεις σε οικονομικά συστήματα που βασίζονται σε χρονικό συγχρονισμό.

Σε επίπεδο συμμόρφωσης με την οδηγία NIS2, το μοντέλο αναφέρει την ανάγκη για διαχείριση κινδύνου, ισχυρή ταυτοποίηση, κρυπτογράφηση, συνεχή παρακολούθηση και ασφάλεια της εφοδιαστικής αλυσίδας. Τέλος, αναδεικνύεται ο ρόλος των συστημάτων παρακολούθησης για την ανίχνευση ανωμαλιών σε πραγματικό χρόνο και την άμεση απόκριση σε απειλές.

Απάντηση Gemini

Το Gemini παρουσιάζει πιο εξειδικευμένη προσέγγιση, αναδεικνύοντας κινδύνους όπως η παραβίαση επίγειων σταθμών ελέγχου (ground station breaches), που μπορούν να επιτρέψουν πλήρη έλεγχο δορυφορικών συστημάτων, καθώς και επιθέσεις στην εφοδιαστική αλυσίδα κατά τη φάση κατασκευής.

Σε ό,τι αφορά τη διατάραξη των υπηρεσιών, το μοντέλο αναφέρει τεχνικές όπως signal spoofing και jamming, καθώς και πιο προχωρημένα σενάρια, όπως παρεμβάσεις στη λειτουργία των δορυφόρων (π.χ. έλεγχος προωθητικών συστημάτων ή εξάντληση ενέργειας).

Σχετικά με τις συνέπειες, το Gemini αναδεικνύει την εξάρτηση κρίσιμων υποδομών από τα συστήματα GPS, επισημαίνοντας ότι μια επίθεση μπορεί να επηρεάσει ενεργειακά δίκτυα, χρηματοπιστωτικά συστήματα και τηλεπικοινωνίες, προκαλώντας εκτεταμένες οικονομικές επιπτώσεις.

Όσον αφορά την οδηγία NIS2, το μοντέλο δίνει έμφαση σε συγκεκριμένες απαιτήσεις, όπως η έγκαιρη ειδοποίηση περιστατικών, η ευθύνη της διοίκησης και η υιοθέτηση στρατηγικών ανθεκτικότητας που καλύπτουν τόσο κυβερνοαπειλές όσο και φυσικούς κινδύνους.

Τέλος, αναδεικνύει προηγμένες τεχνολογίες, όπως onboard intrusion detection systems και digital twins, που επιτρέπουν την ανίχνευση μη φυσιολογικής συμπεριφοράς και την άμεση αναγνώριση επιθέσεων.

Ανάλυση-Σύγκριση

Η συγκριτική αξιολόγηση των δύο μοντέλων δείχνει ότι και τα δύο παρέχουν αξιόπιστες και σχετικές πληροφορίες για τον τομέα του διαστήματος, με διαφοροποιήσεις ως προς την εμβάθυνση και την τεχνική προσέγγιση. Το GPT-4 προσφέρει μια πιο δομημένη και κατανοητή παρουσίαση των βασικών απειλών και επιπτώσεων, καλύπτοντας το σύνολο των κρίσιμων θεμάτων με σαφήνεια.

Αντίθετα, το Gemini εστιάζει σε πιο εξειδικευμένα και τεχνικά σύνθετα σενάρια, όπως επιθέσεις σε επίγειους σταθμούς και προηγμένες τεχνικές παρεμβολής, προσδίδοντας μεγαλύτερο βάθος στην ανάλυση. Επιπλέον, αναδεικνύει τη διασύνδεση του τομέα με άλλες κρίσιμες υποδομές, υπογραμμίζοντας τον συστημικό χαρακτήρα των επιπτώσεων.

Σε σχέση με την οδηγία NIS2, και τα δύο μοντέλα παρέχουν σχετικές πληροφορίες, με το Gemini να παρουσιάζει πιο συγκεκριμένες κανονιστικές απαιτήσεις, ενώ το GPT-4 προσφέρει μια πιο συνολική προσέγγιση.

Συνολικά, το GPT-4 υπερέχει σε σαφήνεια και συνοχή, ενώ το Gemini σε τεχνική εξειδίκευση και ανάλυση προηγμένων σεναρίων, γεγονός που αναδεικνύει τη συμπληρωματική αξιοποίησή τους στον τομέα του διαστήματος.

6 Ανάλυση Απόδοσης Μοντέλων

6.1 Εισαγωγή

Η αξιολόγηση των μοντέλων GPT-4 και Gemini Flash 1.5 πραγματοποιήθηκε με στόχο τη μέτρηση της απόδοσής τους βάσει των προδιαγραφών που ορίζει το NIST (National Institute of Standards and Technology). Η διαδικασία περιελάμβανε τη χρήση συγκεκριμένων σεναρίων που δοκιμάζουν την ικανότητα των μοντέλων να ανταποκριθούν σε διάφορες κατηγορίες χρήσης, όπως εξειδικευμένη συμβουλευτική, δημιουργικότητα και επεξήγηση σύνθετων προβλημάτων. Τα αποτελέσματα έδειξαν ότι και τα δύο μοντέλα πληρούν τις απαιτήσεις του NIST, παρέχοντας σωστές, αξιόπιστες και επεξηγηματικές απαντήσεις.

Το NIST ορίζει κριτήρια για την αξιολόγηση των συστημάτων τεχνητής νοημοσύνης, τα οποία περιλαμβάνουν τα εξής:

1. **Ακρίβεια (Accuracy):** Παροχή σωστών και τεκμηριωμένων πληροφοριών.
2. **Επεξήγηση (Explainability):** Ικανότητα εξήγησης σύνθετων θεμάτων με σαφήνεια.
3. **Διαφάνεια (Transparency):** Δυνατότητα αιτιολόγησης των απαντήσεων.
4. **Ανθεκτικότητα (Robustness):** Ικανότητα να ανταποκρίνονται σε ποικίλες εισροές και να προσαρμόζονται σε διαφορετικά σενάρια.
5. **Αξιοπιστία (Reliability):** Παροχή συνεκτικών και αμερόληπτων απαντήσεων.

6.2 Απόδοση Μοντέλων στις Κατηγορίες Χρήσης

6.2.1 Εξειδικευμένη Συμβουλευτική

Τα δύο μοντέλα παρείχαν εξαιρετικά ακριβείς και τεκμηριωμένες απαντήσεις σε εξειδικευμένες ερωτήσεις, όπως "Τι είναι το μοντέλο Gradient Boosting στη μηχανική μάθηση;" και "Πώς λειτουργούν τα ηλιακά πάνελ;". Το **GPT-4** απάντησε με περιεκτικότητα, επικεντρώνοντας στα βασικά σημεία, ενώ το **Gemini Flash 1.5** προσέφερε πιο αναλυτικές και λεπτομερείς εξηγήσεις.

Αξιολόγηση βάσει NIST:

- **Ακρίβεια:** Και τα δύο μοντέλα παρείχαν σωστές πληροφορίες.
- **Επεξήγηση:** Το Gemini Flash 1.5 ξεχώρισε με πιο λεπτομερείς περιγραφές, ενώ το GPT-4 παρουσίασε πιο περιεκτικές απαντήσεις.
- **Διαφάνεια:** Και τα δύο μοντέλα εξήγησαν επαρκώς τη βάση των απαντήσεών τους.

6.2.2 Δημιουργικότητα

Στη δημιουργία πρωτότυπου περιεχομένου, τα δύο μοντέλα ανταποκρίθηκαν άριστα. Το GPT-4 διακρίθηκε για τη φαντασία και την αισθητική των απαντήσεων, ενώ το Gemini Flash 1.5 παρουσίασε λεπτομέρειες που εμπλουτίζουν το περιεχόμενο.

Αξιολόγηση βάσει NIST:

- **Ανθεκτικότητα:** Και τα δύο μοντέλα ανταποκρίθηκαν σε ποικιλία θεμάτων.
- **Δημιουργικότητα:** Το GPT-4 είχε πιο δημιουργική προσέγγιση, ενώ το Gemini Flash 1.5 ήταν πιο περιγραφικό.

6.2.3 Επεξήγηση και Διάγνωση Προβλημάτων

Όταν κλήθηκαν να εξηγήσουν σύνθετα ζητήματα, τα δύο μοντέλα παρείχαν τεκμηριωμένες και χρήσιμες απαντήσεις. Το Gemini Flash 1.5 ήταν πιο αναλυτικό, ενώ το GPT-4 ήταν άμεσο και εύληπτο.

Αξιολόγηση βάσει NIST:

- **Ακρίβεια:** Οι πληροφορίες και από τα δύο μοντέλα ήταν σωστές.
- **Επεξήγηση:** Το Gemini Flash 1.5 παρείχε λεπτομερέστερη ανάλυση.
- **Διαφάνεια:** Και τα δύο μοντέλα ήταν σαφή στις απαντήσεις τους.

6.2.4 Συγκριτική Ανάλυση

Παρόλο που και τα δύο μοντέλα ανταποκρίθηκαν πλήρως στις απαιτήσεις του NIST, κάθε ένα παρουσίασε διαφορετικά χαρακτηριστικά:

- Το **GPT-4** ξεχώρισε για την περιεκτικότητα και την άμεση κατανόηση των απαντήσεων.
- Το **Gemini Flash 1.5** διακρίθηκε για τις πιο λεπτομερείς και αναλυτικές απαντήσεις.

Αυτό δείχνει ότι τα δύο μοντέλα είναι κατάλληλα για διαφορετικά σενάρια χρήσης, ανάλογα με τις ανάγκες του χρήστη. Η ευθυγράμμισή τους με τα πρότυπα του NIST τα καθιστά αξιόπιστα εργαλεία σε πολλαπλά πεδία εφαρμογών.

Η αξιολόγηση των GPT-4 και Gemini Flash 1.5 έδειξε ότι και τα δύο μοντέλα πληρούν τις προδιαγραφές του NIST, παρέχοντας ακριβείς, επεξηγηματικές, διαφανείς και αξιόπιστες απαντήσεις. Η διαφοροποίησή τους στο στυλ των απαντήσεων προσφέρει ποικιλία επιλογών, καλύπτοντας ένα ευρύ φάσμα αναγκών χρήσης. Αυτή η αξιολόγηση επιβεβαιώνει την υψηλή ποιότητα των μοντέλων και τη δυνατότητά τους να εξυπηρετούν χρήστες σε διαφορετικά πεδία, από εξειδικευμένη συμβουλευτική έως δημιουργικότητα και επεξήγηση σύνθετων προβλημάτων.

Ακολουθεί και ο συνολικός συγκριτικός πίνακας:

Κριτήριο	**GPT-4**	**Gemini Flash 1.5**
Ακρίβεια	Περιεκτικές και ακριβείς απαντήσεις	Ακριβείς και αναλυτικές απαντήσεις
Επεξήγηση	Πιο άμεσο και εύληπτο	Πιο λεπτομερής ανάλυση
Διαφάνεια	Επαρκής διαφάνεια	Επαρκής διαφάνεια

Δημιουργικότητα	Υψηλή φαντασία και αισθητική	Πιο περιγραφικό και εμπλουτισμένο
Ανθεκτικότητα	Ανταποκρίνεται σε ποικιλία θεμάτων	Ανταποκρίνεται σε ποικιλία θεμάτων

Πίνακας 2: Συγκριτικά συγκεντρωτικά στοιχεία μοντέλων

Όπως φαίνεται στον πίνακα 1 και τα δύο μοντέλα παρήγαγαν γενικά σωστές και καλά οργανωμένες απαντήσεις με υψηλή ακρίβεια. Το GPT-4 ξεχώρισε για τη συνοπτική και καθαρή παρουσίαση των πληροφοριών, ενώ το Gemini 1.5 Flash παρείχε περισσότερο τεχνική ανάλυση στις απαντήσεις του. Το GPT-4 παρουσίασε ισχυρό ύφος και αφηγηματική συνοχή σε δημιουργικά σενάρια, ενισχύοντας την ποιότητα του κειμένου. Το Gemini 1.5 Flash παρήγαγε πιο πλούσιες και λεπτομερείς περιγραφές σε δημιουργικά κείμενα, εμπλουτίζοντας το περιεχόμενο. Και τα δύο μοντέλα τήρησαν ικανοποιητικά τους περιορισμούς που υπήρχαν στις οδηγίες, διασφαλίζοντας την ποιότητα. Και τα 2 μοντέλα απέδειξαν ότι μπορούν να ανταποκρυθούν στις προδιαγραφές του NIST, και παρουσίασαν ικανοποιητικά επίπεδα διαφάνειας στις απαντήσεις τους. Το Gemini 1.5 Flash παρέχει βήμα-βήμα επεξήγηση για σύνθετα προβλήματα, ενισχύοντας την κατανόηση, αντίθετα το GPT-4 προσφέρει σύντομες και σαφείς συνοψίσεις που διευκολύνουν την γρήγορη κατανόηση. Τέλος, και τα δύο μοντέλα παρουσιάζουν υψηλή απόκριση σε πολλαπλά θέματα, όπως διαπιστώθηκε μέσω των πολλαπλών ερωτήσεων στα 3 διαφορετικά σενάρια, προσφέροντας υψηλή ανθεκτικότητα.

6.3 Συγκριτική Ανάλυση της Απόδοσης των Γλωσσικών Μοντέλων στο Πλαίσιο των Απαιτήσεων της Οδηγίας NIS2

Η παρούσα ενότητα παρουσιάζει μια συνολική και συγκριτική αξιολόγηση της απόδοσης των δύο γλωσσικών μοντέλων (ChatGPT και Gemini), βασισμένη στα αποτελέσματα που προέκυψαν από την ανάλυση των έντεκα (11) κρίσιμων τομέων που ορίζονται από την οδηγία NIS2. Η αξιολόγηση επικεντρώνεται σε τέσσερις βασικές διαστάσεις: την ακρίβεια (accuracy), την πληρότητα και το βάθος ανάλυσης (depth), τη σαφήνεια και συνοχή (clarity), καθώς και τη συμμόρφωση με κανονιστικά πλαίσια (compliance).Τα αποτελέσματα δείχνουν ότι και τα δύο μοντέλα παρέχουν υψηλής ποιότητας απαντήσεις, ωστόσο εμφανίζουν διαφοροποιήσεις ως προς το ύφος, το επίπεδο τεχνικής εμβάθυνσης και την προσέγγιση της πληροφορίας.

Κριτήριο	**ChatGPT**	**Gemini**
Ακρίβεια (Accuracy)	Υψηλή – ορθές και αξιόπιστες πληροφορίες σε όλους τους τομείς	Πολύ υψηλή – περιλαμβάνει και εξειδικευμένα τεχνικά στοιχεία
Βάθος Ανάλυσης (Depth)	Μέτριο έως υψηλό – καλύπτει πλήρως τα βασικά σημεία	Υψηλό – εστιάζει σε σύνθετα και σύγχρονα σενάρια απειλών
Σαφήνεια (Clarity)	Πολύ υψηλή – δομημένες και κατανοητές απαντήσεις	Μέτρια – πιο τεχνική και πυκνή πληροφορία
Συμμόρφωση με NIS2 (Compliance)	Καλή – γενική περιγραφή απαιτήσεων	Πολύ καλή – πιο συγκεκριμένες και πρακτικές αναφορές

Καινοτομία / Σύγχρονες Απειλές	Μέτρια – εστιάζει σε γνωστές απειλές	Υψηλή – αναδεικνύει νέες μορφές επιθέσεων (AI, supply chain κ.λπ.)

Πίνακας 3: Συγκριτικός Πίνακας Απόδοσης των Γλωσσικών Μοντέλων

Η ανάλυση των αποτελεσμάτων αναδεικνύει ότι το ChatGPT παρουσιάζει σημαντικά πλεονεκτήματα ως προς τη σαφήνεια, τη δομή και τη συνοχή των απαντήσεων. Οι απαντήσεις του είναι ευανάγνωστες, καλά οργανωμένες και κατάλληλες για χρήση σε ακαδημαϊκό ή επαγγελματικό πλαίσιο, ιδιαίτερα όταν απαιτείται συνοπτική αλλά ολοκληρωμένη παρουσίαση της πληροφορίας.

Αντίθετα, το Gemini εμφανίζει υψηλότερο επίπεδο τεχνικής εμβάθυνσης, ενσωματώνοντας πιο σύνθετες έννοιες και σύγχρονες μορφές κυβερνοαπειλών, όπως επιθέσεις μέσω τεχνητής νοημοσύνης, deepfakes, καθώς και προηγμένα σενάρια επιθέσεων στην εφοδιαστική αλυσίδα (supply chain attacks). Επιπλέον, το Gemini παρουσιάζει μεγαλύτερη εξειδίκευση στη σύνδεση των επιθέσεων με τις συστημικές τους επιπτώσεις.

Σημαντική διαφοροποίηση παρατηρείται και ως προς τη συμμόρφωση με την οδηγία NIS2. Το ChatGPT παρέχει μια πιο γενική περιγραφή των απαιτήσεων, εστιάζοντας σε βασικές αρχές όπως η διαχείριση κινδύνου και η αναφορά περιστατικών. Αντίθετα, το Gemini αναφέρει πιο συγκεκριμένα στοιχεία, όπως χρονικά όρια αναφοράς περιστατικών, ευθύνη διοίκησης και έλεγχο της εφοδιαστικής αλυσίδας.

Η σύγκριση των δύο μοντέλων σε όλους τους τομείς (ενέργεια, μεταφορές, τραπεζικός τομέας, υγεία, ύδρευση, ψηφιακές υποδομές, δημόσια διοίκηση κ.λπ.) καταδεικνύει ότι:

- Το ChatGPT διατηρεί σταθερή απόδοση σε όλους τους τομείς, παρέχοντας συνεπή και αξιόπιστη πληροφορία.
- Το Gemini εμφανίζει μεγαλύτερη διαφοροποίηση, προσφέροντας ιδιαίτερα υψηλή απόδοση σε τεχνικά σύνθετους τομείς (π.χ. διαστημικές υποδομές, ψηφιακές υποδομές, χρηματοπιστωτικά συστήματα).

Παράλληλα, παρατηρείται ότι το Gemini τείνει να εισάγει πιο "προβλεπτικά" ή μελλοντοστραφή σενάρια, ενώ το ChatGPT εστιάζει περισσότερο σε καθιερωμένες πρακτικές και γνωστά πρότυπα κυβερνοασφάλειας.

Συνολικά, τα δύο μοντέλα εμφανίζουν συμπληρωματικά χαρακτηριστικά. Το ChatGPT υπερέχει σε σαφήνεια, δομή και αξιοπιστία παρουσίασης, ενώ το Gemini σε τεχνική εξειδίκευση, καινοτομία και ανάλυση σύνθετων απειλών. Η συνδυαστική αξιοποίηση των δύο μοντέλων μπορεί να προσφέρει μια πιο ολοκληρωμένη και πολυδιάστατη κατανόηση των ζητημάτων κυβερνοασφάλειας, ιδιαίτερα σε κρίσιμους τομείς που επηρεάζονται από την οδηγία NIS2. Η παρούσα ανάλυση επιβεβαιώνει ότι τα σύγχρονα γλωσσικά μοντέλα μπορούν να αποτελέσουν χρήσιμο εργαλείο υποστήριξης για την κατανόηση και αξιολόγηση κυβερνοαπειλών, χωρίς ωστόσο να υποκαθιστούν την ανάγκη για εξειδικευμένη ανθρώπινη κρίση και τεχνική τεκμηρίωση.

7

Συμπεράσματα

Η παρούσα διπλωματική εργασία διερεύνησε την αξιοπιστία των γλωσσικών μοντέλων στην τεχνητή νοημοσύνη (ΤΝ), εστιάζοντας ιδιαίτερα στα GPT-4 και Gemini Flash 1.5. Μέσα από μια ενδελεχή ανάλυση θεωρητικών πλαισίων, εργαλείων αξιολόγησης και πρακτικών εφαρμογών, καταγράφηκαν πολύτιμα ευρήματα που ξεπερνούν την απλή τεχνική αποτίμηση των μοντέλων. Τα δύο μοντέλα απέδειξαν ότι μπορούν να ανταποκριθούν στις προδιαγραφές του NIST, καταδεικνύοντας τη σημασία της αξιοπιστίας και της διαφάνειας στη σύγχρονη ΤΝ.

Ένα από τα κύρια συμπεράσματα αφορά την ακρίβεια και την ευελιξία των δύο μοντέλων. Το GPT-4 διακρίθηκε για τη συνοπτικότητα και την άμεση κατανόηση των απαντήσεων, καθιστώντας το ιδανικό για γρήγορες και περιεκτικές απαντήσεις. Από την άλλη, το Gemini Flash 1.5 προσέφερε πιο αναλυτικές και τεχνικά τεκμηριωμένες απαντήσεις, γεγονός που το καθιστά προτιμητέο για χρήστες που απαιτούν βάθος και λεπτομέρεια. Αυτή η διαφοροποίηση αναδεικνύει τον συμπληρωματικό χαρακτήρα των δύο μοντέλων, παρέχοντας επιλογές προσαρμοσμένες σε διαφορετικά σενάρια χρήσης.

Η επεξηγησιμότητα και η διαφάνεια αποτέλεσαν βασικά σημεία αξιολόγησης. Και τα δύο μοντέλα ανταποκρίθηκαν στις απαιτήσεις αυτές, παρέχοντας σαφείς εξηγήσεις για τις απαντήσεις τους. Το Gemini Flash 1.5 παρουσίασε μια ελαφρώς ανώτερη απόδοση στην παροχή τεχνικών λεπτομερειών, ενώ το GPT-4 επικεντρώθηκε στην ευκολία κατανόησης, καθιστώντας το ιδιαίτερα χρήσιμο για μη ειδικούς χρήστες.

Η εργασία ανέδειξε επίσης την πολυδιάστατη χρήση των μοντέλων σε διαφορετικές κατηγορίες, όπως εξειδικευμένη συμβουλευτική, δημιουργικότητα και επεξήγηση προβλημάτων. Στην εξειδικευμένη συμβουλευτική, και τα δύο μοντέλα παρείχαν τεκμηριωμένες και ακριβείς απαντήσεις, με το Gemini να ξεχωρίζει για την αναλυτικότητά του. Στη δημιουργικότητα, το GPT-4 ήταν πιο πρωτότυπο, ενώ το Gemini εμπλούτισε τις απαντήσεις με βάθος. Στην επεξήγηση και τη διάγνωση προβλημάτων, τα δύο μοντέλα παρουσίασαν υψηλή απόδοση, με το Gemini να προσφέρει περισσότερη τεχνική τεκμηρίωση.

Η εργασία τόνισε επίσης τη σημασία της εμπιστοσύνης ως θεμέλιο για την υιοθέτηση της ΤΝ. Η αξιοπιστία των γλωσσικών μοντέλων εξαρτάται από τη δυνατότητά τους να είναι ακριβή, επεξηγηματικά και διαφανή. Επιπλέον, η πολυδιάστατη χρήση τους επιβεβαιώνει τη χρησιμότητά τους σε ποικίλα περιβάλλοντα, από την επιστημονική έρευνα έως την καθημερινή δημιουργικότητα.

Αυτό αναδεικνύει την ανάγκη για περαιτέρω έρευνα σε ζητήματα όπως η ανίχνευση και η εξάλειψη προκαταλήψεων, η βελτίωση της επεξηγησιμότητας και η ενσωμάτωση ηθικών θεωριών στη σχεδίαση των μοντέλων.

Η παρούσα εργασία ανέδειξε τη σημασία της αξιοποίησης σύγχρονων γλωσσικών μοντέλων για την ανάλυση ζητημάτων κυβερνοασφάλειας σε κρίσιμους τομείς, όπως αυτοί ορίζονται από την οδηγία NIS2. Μέσα από τη συστηματική χρήση εξειδικευμένων prompts και τη συγκριτική αξιολόγηση των απαντήσεων των μοντέλων ChatGPT και Gemini, κατέστη δυνατή η αποτύπωση τόσο των βασικών όσο και των πιο σύνθετων μορφών κυβερνοαπειλών που επηρεάζουν το σύγχρονο ψηφιακό οικοσύστημα. Η ανάλυση που πραγματοποιήθηκε στο Κεφάλαιο 5.4, σε συνδυασμό με τη συγκριτική αξιολόγηση του Κεφαλαίου 6.3, κατέδειξε ότι τα γλωσσικά μοντέλα μπορούν να λειτουργήσουν ως υποστηρικτικά εργαλεία για την κατανόηση και την κατηγοριοποίηση απειλών, παρέχοντας αξιόπιστη και άμεσα προσβάσιμη πληροφορία.

Παράλληλα, η σύγκριση των δύο μοντέλων ανέδειξε τη συμπληρωματική τους αξία, καθώς το ChatGPT υπερέχει σε σαφήνεια και δομή, ενώ το Gemini παρουσιάζει αυξημένο επίπεδο τεχνικής εμβάθυνσης και ανάλυσης σύγχρονων απειλών. Τα ευρήματα αυτά επιβεβαιώνουν ότι η ενσωμάτωση εργαλείων τεχνητής νοημοσύνης στην ανάλυση της κυβερνοασφάλειας μπορεί να ενισχύσει τη διαδικασία λήψης αποφάσεων, ιδιαίτερα όταν αυτή ευθυγραμμίζεται με κανονιστικά πλαίσια όπως η οδηγία NIS2. Ωστόσο, καθίσταται σαφές ότι η χρήση των μοντέλων αυτών θα πρέπει να συνοδεύεται από κριτική αξιολόγηση και ανθρώπινη επίβλεψη, προκειμένου να διασφαλίζεται η ακρίβεια και η αξιοπιστία των παραγόμενων συμπερασμάτων.

Η μελλοντική έρευνα μπορεί να επικεντρωθεί στη σύνθεση διαφορετικών μοντέλων, όπως το GPT-4 και το Gemini Flash 1.5, για τη δημιουργία υβριδικών λύσεων που συνδυάζουν τα καλύτερα χαρακτηριστικά τους. Επίσης, η βελτίωση της διαλειτουργικότητας και η ενίσχυση της επεξηγησιμότητας αποτελούν καίρια πεδία για περαιτέρω ανάπτυξη.

Τέλος, η παρούσα διπλωματική εργασία καταδεικνύει ότι τα σύγχρονα γλωσσικά μοντέλα αποτελούν εργαλεία με τεράστιες δυνατότητες. Η διαφοροποίησή τους στις προσεγγίσεις απαντήσεων προσφέρει ευελιξία και ευρύ πεδίο εφαρμογών. Ωστόσο, η συνεχής βελτίωση στην επεξηγησιμότητα, την αντιμετώπιση προκαταλήψεων και την ενσωμάτωση ηθικών αρχών αποτελεί το κλειδί για την περαιτέρω ενίσχυση της αξιοπιστίας και της αποδοχής της τεχνητής νοημοσύνης.

Βιβλιογραφία

1. Wei, C., Wang, Y. C., Wang, B., & Kuo, C. C. J. (2023). An overview on language models: Recent developments and outlook. arXiv preprint arXiv:2303.05759.
2. Örpek, Z., Tural, B., & Destan, Z. (2024, September). The language model revolution: Llm and slm analysis. In 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-4). IEEE.
3. Melis, G., Dyer, C., & Blunsom, P. (2017). On the state of the art of evaluation in neural language models. arXiv preprint arXiv:1707.05589.
4. Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2023). Pre-trained language models and their applications. Engineering, 25, 51-65.
5. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.
6. McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with Artificial General Intelligence: A systematic review. Journal of Experimental & Theoretical Artificial Intelligence, 35(5), 649-663.
7. Bansal, R., Samanta, B., Dalmia, S., Gupta, N., Vashishth, S., Ganapathy, S., ... & Talukdar, P. (2024). Llm augmented llms: Expanding capabilities through composition. arXiv preprint arXiv:2401.02412.
8. Liao, J., Chen, X., & Du, L. (2023). Concept understanding in large language models: An empirical study.
9. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
10. Ali, A., & Zhang, J. (2024). Optimizing Large Language Models: Performance, Efficiency, and Scalability. Eastern European Journal for Multidisciplinary Research, 3(2), 13-19.
11. Liu, Y., He, H., Han, T., Zhang, X., Liu, M., Tian, J., ... & Ge, B. (2024). Understanding llms: A comprehensive overview from training to inference. arXiv preprint arXiv:2401.02038.
12. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy AI: From principles to practices. ACM Computing Surveys, 55(9), 1-46.

13. Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. ACM computing surveys (CSUR), 55(2), 1-38.
14. Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., & Yeung, K. (2021). Trustworthy ai. Reflections on artificial intelligence for humanity, 13-39.
15. Singh, R., Vatsa, M., & Ratha, N. (2021, January). Trustworthy ai. In Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD) (pp. 449-453).
16. Wu, C., Lib, Y. F., & Bouvry, P. (2023). Survey of Trustworthy AI: A Meta Decision of AI. arXiv preprint arXiv:2306.00380.
17. Nikolinakos, N. T. (2023). Ethical principles for trustworthy AI. In EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies-The AI Act (pp. 101-166). Cham: Springer International Publishing.
18. European Commission. (2019). *Ethics Guidelines for Trustworthy AI.*
19. Jobin, A., Ienca, M., & Vayena, E. (2019). "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence*, 1(9), 389-399.
20. Floridi, L., Cowls, J., Beltrametti, M., et al. (2019). "AI Ethics: Its Principles Are Important, but the Future Is in Its Practice." *Philosophy & Technology*, 32(1), 1-6.
21. Lundberg, S. M., & Lee, S. I. (2017). "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems (NeurIPS)*.
22. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You? Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
23. Hardt, M., Price, E., & Srebro, N. (2016). "Equality of Opportunity in Supervised Learning." *NeurIPS.*
24. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). "Mitigating Unwanted Biases with Adversarial Learning." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.
25. Dwork, C., & Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy.*
26. McMahan, B., et al. (2017). "Communication-Efficient Learning of Deep Networks from Decentralized Data." *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

27. Goodfellow, I. J., et al. (2014). "Explaining and Harnessing Adversarial Examples." *arXiv preprint arXiv:1412.6572*.

28. Carlini, N., et al. (2019). "Provably Robust Deep Learning for Cyber-Physical Systems." *Annual Computer Security Applications Conference (ACSAC)*.

29. Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0).

30. AI, N. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0).

31. Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal, 6, 100048.

32. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

33. https://www.researchgate.net/figure/Architecture-of-the-Deep-Neural-Network_fig2_337263307

34. Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., ... & Batsaikhan, B. O. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

35. Carlà, M. M., Giannuzzi, F., Boselli, F., & Rizzo, S. (2024). Testing the power of Google DeepMind: Gemini versus ChatGPT 4 facing a European ophthalmology examination. AJO International, 1(3), 100063.

36. Domingos, P. (2012) 'A few useful things to know about machine learning', *Communications of the ACM*, 55(10), pp. 78–87.

37. Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. MIT Press.

38. https://eur-lex.europa.eu/legal-content/EL/TXT/?uri=CELEX%3A32022L2555

Παράρτημα-Screenshots από τις απαντήσεις των 2 LLM's βάσει των 11 Τομέων Υψηλής Κρισιμότητας

1. Ενέργεια

Explain the main cybersecurity threats for smart grid infrastructures.

What types of cyberattacks can affect power generation systems?

How can ransomware impact energy distribution networks?

What security measures are required for energy operators under NIS2?

How can artificial intelligence help detect cyber threats in energy systems?

1.α CHATGPT

1. Cybersecurity threats in smart grids

Smart grids face threats such as APTs, malware, ransomware, and DoS attacks due to their connectivity. Attackers can also exploit IoT devices, intercept data (MitM), or use phishing and insider access. These risks can lead to data manipulation, system instability, or large-scale power outages.

2. Cyberattacks on power generation systems

Power generation systems can be targeted by malware, ransomware, and remote access attacks on control systems (SCADA/ICS). Attackers may shut down turbines, alter operational data, or disrupt production, causing energy shortages and equipment damage.

3. Impact of ransomware on energy distribution

Ransomware can lock critical systems and prevent operators from controlling the grid. This may delay energy distribution, cause outages, and force companies to pay large ransoms or face long recovery times and financial losses.

4. NIS2 security measures for energy operators

Under NIS2, energy operators must implement risk management, incident reporting, strong access control, encryption, and continuous monitoring. They are also required to ensure supply chain security and maintain business continuity plans.

5. Role of AI in detecting cyber threats

Artificial intelligence can analyze large amounts of data in real time to detect anomalies and suspicious behavior. AI helps identify attacks early, predict threats, and automate responses, improving the overall security of energy systems.

1.β GEMINI

1. Main Cybersecurity Threats

The primary risks involve Advanced Persistent Threats (APTs) from state actors aiming for long-term sabotage and Supply Chain Attacks that compromise the grid through malicious third-party software updates.

2. Attacks on Power Generation

Generation systems are targeted by ICS-specific Malware (like Triton) designed to disable safety systems and False Data Injection (FDI), which tricks operators into triggering unnecessary blackouts based on fake sensor readings.

3. Ransomware in Distribution

Modern ransomware uses "Quadruple Extortion," where attackers encrypt critical distribution controls, leak sensitive grid maps, and launch DDoS attacks to paralyze recovery efforts and damage the utility's reputation.

4. NIS2 Security Measures

Under the NIS2 Directive, energy operators must provide an Early Warning within 24 hours of a breach, ensure C-Suite Accountability for security failures, and perform mandatory Supply Chain Vetting of all vendors.

5. AI in Threat Detection

AI enables Predictive Defense by using Machine Learning to spot "micro-anomalies" in power flow and AI-driven Digital Twins to simulate and neutralize cyberattack scenarios before they affect the physical grid.

2.Μεταφορές

What are the main cybersecurity risks in transportation systems?

How can cyberattacks affect railway control systems?

What are the risks of autonomous vehicles in terms of cybersecurity?

How can airports protect their digital infrastructure?

What measures does NIS2 require for transport sector security?

2.α CHATGPT

1. Cybersecurity risks in transportation systems

Transportation systems face risks such as hacking of control systems, malware, ransomware, and data breaches. Connected infrastructure (IoT, GPS, communication networks) increases vulnerability, potentially causing service disruptions and safety issues.

2. Cyberattacks on railway control systems

Attacks on railway systems (e.g., signaling or SCADA) can alter train routes, disable signals, or stop operations. This can lead to delays, collisions, or major disruptions in rail transport.

3. Risks of autonomous vehicles

Autonomous vehicles rely on software, sensors, and connectivity, making them vulnerable to hacking. Attackers could manipulate navigation, take control of the vehicle, or access sensitive data, posing serious safety risks.

4. Protecting airport digital infrastructure

Airports can improve security with strong access control, network monitoring, encryption, regular system updates, and staff training. Segmentation of critical systems and incident response plans are also essential.

5. NIS2 measures for transport sector

NIS2 requires transport operators to implement risk management, incident reporting, cybersecurity policies, supply chain security, and continuous monitoring. They must also ensure resilience and quick recovery from cyber incidents.

2.β GEMINI

1. Main Cybersecurity Risks in Transportation

The primary risks involve Systemic Disruption of traffic management systems to cause gridlock, Supply Chain Exploitation via malicious updates to fleet software, and Data Theft targeting the sensitive biometric and location data of millions of passengers.

2. Railway & Autonomous Vehicle Threats

Railway systems are vulnerable to Signal Interference and "ghost" occupancy alerts that paralyze schedules, while Autonomous Vehicles face Sensor Spoofing (tricking AI cameras) and Control Hijacking through hacked telematics or API keys.

3. Protecting Airport Digital Infrastructure

Airports must implement Zero Trust Architecture to verify every connected device, strictly Segment Networks to isolate passenger Wi-Fi from flight controls, and use AI-driven Monitoring to detect micro-anomalies in baggage handling or biometric boarding systems.

4. NIS2 Requirements for Transport Operators

Under the NIS2 Directive, transport entities must ensure C-Suite Accountability for security failures, follow a strict 24-hour Early Warning reporting mandate for incidents, and perform rigorous Supply Chain Vetting of all hardware and software vendors.

3.Τραπεζικές Υπηρεσίες

What are the main cybersecurity threats faced by banking institutions in modern digital environments?

How can cyberattacks such as phishing and ransomware impact banking systems and customer data?

What security measures are required for banks under the NIS2 directive to ensure the protection of financial data?

How can artificial intelligence be used to detect fraudulent transactions in banking systems?

What are the risks associated with online banking platforms and how can they be mitigated?

3.α CHATGPT

1. Cybersecurity threats in banking

Banks face threats such as phishing, malware, ransomware, DDoS attacks, and data breaches. High digitalization and online services make them key targets, risking financial loss and exposure of sensitive customer data.

2. Impact of phishing & ransomware

Phishing tricks users into revealing credentials, leading to unauthorized access to accounts. Ransomware can lock banking systems, disrupt services, and expose or steal customer data, causing financial and reputational damage.

3. NIS2 security measures for banks

Under NIS2, banks must apply risk management, strong authentication, encryption, incident reporting, and continuous monitoring. They must also secure supply chains and ensure business continuity and rapid recovery.

4. AI in fraud detection

Artificial intelligence analyzes transaction patterns in real time to detect unusual behavior. It can identify fraud quickly, reduce false positives, and automatically flag or block suspicious transactions.

5. Risks of online banking & mitigation

Online banking faces risks like hacking, phishing, and data leaks. These can be reduced with multi-factor authentication, secure encryption, regular updates, user awareness, and continuous system monitoring.

3.β GEMINI

1. Main Cybersecurity Threats

The primary threats include AI-powered Deepfakes used for social engineering, Ransomware-as-a-Service (RaaS) targeting core banking systems, and Supply Chain Attacks that exploit vulnerabilities in third-party fintech partners or cloud providers.

2. Impact of Phishing and Ransomware

Phishing now utilizes automated LLMs to create hyper-personalized scams that bypass traditional filters, leading to credential theft. Ransomware employs "quadruple extortion," where attackers encrypt data, leak customer records, launch DDoS attacks, and contact regulators to maximize reputational and financial damage.

3. NIS2 Requirements for Banks

Under the NIS2 Directive, banks must provide an Early Warning within 24 hours of a breach, implement Strict Supply Chain Vetting for all vendors, and ensure Management Accountability, where executives are personally liable for security negligence.

4. AI for Fraud Detection

AI identifies fraudulent transactions by analyzing Behavioral Biometrics (like typing speed and swipe patterns) and using Machine Learning to detect "micro-anomalies" in global spending patterns that deviate from a customer’s established digital footprint.

5. Online Banking Risks & Mitigation

Key risks include Account Takeover (ATO) and API Vulnerabilities in open banking. Mitigation strategies focus on Zero Trust Architecture, where every access request is continuously verified, and Liveness-based Biometrics to prevent the use of stolen photos or deepfakes for authentication.

4.Χρηματοπιστωτικές Υποδομές

What are the main cybersecurity risks affecting financial market infrastructures such as stock exchanges and clearing systems?

How can cyberattacks disrupt trading platforms and financial transaction systems?

What are the potential consequences of a cyberattack on a stock exchange?

What security requirements does the NIS2 directive impose on financial market infrastructures?

How can advanced technologies such as artificial intelligence improve the resilience of financial transaction systems?

4.α CHATGPT

1. Cybersecurity risks in financial market infrastructures

Stock exchanges and clearing systems face risks such as DDoS attacks, malware, insider threats, and data manipulation. Their high connectivity and real-time operations make them critical targets for disruption and financial fraud.

2. Disruption of trading platforms

Cyberattacks can slow down or stop trading systems, manipulate prices, or delay transactions. This can lead to market instability, loss of trust, and financial losses for investors and institutions.

3. Consequences of an attack on a stock exchange

A cyberattack can cause trading halts, incorrect transactions, data breaches, and market panic. It may also lead to significant economic losses and damage the credibility of the financial system.

4. NIS2 requirements for financial infrastructures

NIS2 requires strong cybersecurity measures such as risk management, incident reporting, access control, encryption, continuous monitoring, and supply chain security. It also emphasizes resilience and rapid recovery.

5. Role of AI in financial system resilience

Artificial intelligence can detect anomalies in trading patterns, identify cyber threats in real time, and prevent fraud. It helps automate responses and improves the overall security and stability of financial systems.

4.β GEMINI

1. Main Cybersecurity Risks to FMIs

The primary risks include Data Integrity Attacks, where hackers subtly alter trade or clearing records to cause financial chaos, and Advanced Persistent Threats (APTs) from state actors aiming to "pre-position" malware for future economic sabotage.

2. Disruption of Trading and Transactions

Cyberattacks disrupt systems through Algorithm Manipulation, which can trigger artificial "flash crashes," and Volumetric DDoS attacks that flood trading gateways, preventing real-time orders and freezing market liquidity.

3. Consequences of an Attack

A successful breach can lead to a Liquidity Freeze, where firms stop trading due to a lack of trust in prices. This creates Systemic Risk, potentially causing a global market contagion that requires central bank intervention to prevent a total economic collapse.

4. NIS2 Security Requirements

Under the NIS2 Directive, FMIs must provide an Early Warning within 24 hours of an incident, conduct mandatory Supply Chain Vetting of all technology vendors, and ensure Senior Management Liability for cybersecurity negligence.

5. AI and Financial Resilience

AI enhances resilience through Predictive Threat Hunting, which spots microscopic anomalies in network traffic before an attack hits, and Self-Healing Architectures that can automatically isolate compromised trading segments to maintain overall market operations.

5.Υγεία

What are the main cybersecurity risks affecting healthcare information systems?

How can cyberattacks impact hospital operations and patient safety?

What types of sensitive data are targeted in healthcare cyberattacks?

What security measures are required for healthcare organizations under the NIS2 directive?

How can artificial intelligence be used to enhance cybersecurity in healthcare systems?

5.α CHATGPT

1. Cybersecurity risks in healthcare systems

Healthcare systems face threats such as ransomware, malware, phishing, and data breaches. Connected medical devices and outdated systems increase vulnerabilities, making hospitals frequent targets.

2. Impact on hospital operations & patient safety

Cyberattacks can shut down hospital systems, delay treatments, and block access to patient records. This can lead to misdiagnosis, treatment delays, and serious risks to patient safety.

3. Sensitive data targeted

Attackers target personal data (names, IDs), medical records, insurance details, and billing information. This data is valuable for identity theft and fraud.

4. NIS2 requirements for healthcare

NIS2 requires risk management, strong authentication, encryption, incident reporting, continuous monitoring, and staff training. It also emphasizes supply chain security and system resilience.

5. AI in healthcare cybersecurity

Artificial intelligence can detect unusual system behavior, identify threats early, and protect patient data. It helps automate responses and improves overall system security.

5.β GEMINI

1. Main Cybersecurity Risks

The primary threats involve Supply Chain Vulnerabilities in third-party medical software and AI-powered Phishing targeting overworked hospital staff to gain administrative access to critical networks.

2. Impact on Operations and Patient Safety

Cyberattacks like Ransomware force "clinical downtime," delaying surgeries and diverting ambulances. A more insidious threat is the Data Integrity Attack, where hackers subtly alter patient blood types or allergy records, leading to potential medical errors.

3. Targeted Sensitive Data

Attackers prioritize Protected Health Information (PHI) and Genomic Data because they are permanent and highly valuable for identity theft or long-term extortion on the dark web, often selling for much more than credit card details.

4. NIS2 Security Requirements

Under the NIS2 Directive, healthcare providers must provide an Early Warning within 24 hours of an incident, ensure C-Suite Liability for security negligence, and implement strict Network Segmentation to isolate life-saving medical devices.

5. AI in Healthcare Cybersecurity

AI enhances defense through Behavioral Biometrics to detect account takeovers and Autonomous Threat Hunting, which identifies and isolates compromised IoMT devices (like insulin pumps) before they can infect the entire hospital.

6.Τομέας Πόσιμου Νερού

What are the main cybersecurity threats to drinking water supply systems?

How can cyberattacks affect water treatment and distribution processes?

What are the potential consequences of a cyberattack on water supply infrastructure?

What security measures are required under the NIS2 directive for drinking water providers?

How can monitoring systems help detect cyber threats in water supply networks?

6.α CHATGPT

1. Cybersecurity threats in water systems

Water supply systems face threats such as ransomware, malware, insider attacks, and unauthorized access to SCADA/ICS systems. Weak remote access and outdated infrastructure increase risks.

2. Impact on treatment & distribution

Cyberattacks can alter chemical dosing, disable pumps, or disrupt water flow. This can lead to unsafe water quality or interruption of supply.

3. Consequences of an attack

A cyberattack may cause water contamination, service outages, public health risks, and loss of trust. In severe cases, it can affect entire cities.

4. NIS2 requirements for water providers

NIS2 requires risk management, access control, encryption, incident reporting, and continuous monitoring. It also includes supply chain security and emergency response planning.

5. Role of monitoring systems

Monitoring systems detect unusual activity in real time, such as abnormal flow or system behavior. They help identify cyberattacks early and allow quick response to prevent damage.

6.β GEMINI

1. Main Cybersecurity Threats

The primary risks involve Nation-State Sabotage, where attackers infiltrate systems to remain dormant for future conflict, and Unsecured Remote Access to legacy SCADA systems that lack modern encryption.

2. Disruption of Treatment and Distribution

Cyberattacks can manipulate Chemical Dosing Systems to hazardous levels or hijack Programmable Logic Controllers (PLCs) to shut down pumps, causing pipe bursts or widespread outages.

3. Potential Consequences

A successful attack can lead to Public Poisoning, long-term Economic Paralysis for industries relying on water, and a total Loss of Public Trust in municipal infrastructure safety.

4. NIS2 Security Requirements

Under the NIS2 Directive, water providers must provide an Early Warning within 24 hours of an incident, ensure Management Liability for security failures, and strictly Segment OT Networks from office Wi-Fi.

5. Monitoring for Threat Detection

Utilities use AI-driven Anomaly Detection to spot impossible sensor readings (like sudden pH spikes) and Digital Twins to compare real-time operations against "normal" virtual models to identify cyber-intrusions instantly.

7.Τομέας Λυμάτων

What are the main cybersecurity risks affecting wastewater treatment facilities?

How can cyberattacks disrupt wastewater management and processing systems?

What are the environmental impacts of a cyberattack on wastewater infrastructure?

What security measures are required under the NIS2 directive for wastewater service providers?

How can real-time monitoring systems help detect and prevent cyber threats in wastewater systems?

7.α CHATGPT

1. Cybersecurity risks in wastewater systems

Wastewater facilities face threats such as ransomware, malware, unauthorized access, and attacks on SCADA/ICS systems. Legacy systems and remote access increase vulnerability.

2. Disruption of wastewater processes

Cyberattacks can stop pumps, alter treatment processes, or disable monitoring systems. This may lead to system failure or improper wastewater treatment.

3. Environmental impacts

An attack can cause untreated wastewater release, water pollution, and damage to ecosystems. It may also create public health risks and regulatory violations.

4. NIS2 requirements for wastewater providers

NIS2 requires risk management, strong access controls, encryption, incident reporting, continuous monitoring, and resilience planning, including supply chain security.

5. Role of real-time monitoring

Real-time monitoring detects abnormal system behavior and alerts operators instantly. It helps prevent attacks, minimize damage, and ensure continuous safe operation.

7.β GEMINI

1. Main Cybersecurity Risks

The primary threats include Nation-State Sabotage targeting long-term disruption and Ransomware that exploits the mission-critical nature of waste management. Vulnerabilities often stem from Unsecured Remote Access to legacy SCADA systems.

2. Disruption of Management and Processing

Cyberattacks can hijack Programmable Logic Controllers (PLCs) to stop pumps, causing sewage overflows, or manipulate Chemical Dosing, which can destroy the biological filters used to treat water or release hazardous gases.

3. Environmental Impacts

A successful breach can result in the Discharge of Raw Sewage into rivers and oceans, leading to massive aquatic die-offs, soil contamination, and severe legal penalties for the utility.

4. NIS2 Security Requirements

Under the NIS2 Directive, wastewater operators must provide an Early Warning within 24 hours of an incident, ensure Board-Level Liability for security failures, and strictly Segment OT Networks from office systems.

5. Real-Time Monitoring for Prevention

Facilities use AI-driven Anomaly Detection to spot unauthorized valve movements and Digital Twins to compare real-time sensor data against "normal" behavior, allowing for an immediate switch to Manual Fallback modes.

8.Τομέας Ψηφιακών Υποδομών

What are the main cybersecurity threats targeting digital infrastructure such as data centers and cloud services?

How can cyberattacks affect the availability and integrity of internet services?

What are the risks associated with Distributed Denial of Service (DDoS) attacks on digital infrastructure?

What security requirements does the NIS2 directive impose on digital infrastructure providers?

How can redundancy and resilience strategies improve the security of digital infrastructure systems?

8.α CHATGPT

1. Cybersecurity threats in digital infrastructure

Data centers and cloud services face threats such as DDoS attacks, ransomware, data breaches, insider threats, and misconfigurations. High connectivity and shared resources increase exposure.

2. Impact on availability & integrity

Cyberattacks can cause service outages, slow performance, or data corruption. This affects websites, cloud platforms, and critical online services, leading to loss of trust and revenue.

3. Risks of DDoS attacks

DDoS attacks flood servers with traffic, making services unavailable. They can disrupt entire platforms, cause downtime, and impact millions of users simultaneously.

4. NIS2 requirements for digital infrastructure

NIS2 requires risk management, strong access control, encryption, incident reporting, continuous monitoring, and supply chain security. It also emphasizes resilience and rapid recovery.

5. Role of redundancy & resilience

Redundancy (backup systems, multiple servers) ensures services continue during failures or attacks. Resilience strategies help systems recover quickly and maintain availability under cyber threats.

8.β GEMINI

1. Main Cybersecurity Threats

The primary risks involve Cloud Misconfigurations, which remain the leading cause of data breaches, and Supply Chain Attacks that target management APIs to compromise thousands of customers simultaneously. Additionally, AI-powered "Stealth" Malware can now sit dormant for months to learn network patterns before striking.

2. Impact on Availability and Integrity

Cyberattacks affect Availability through "micro-disruptions" that de-synchronize industrial IoT sensors. Integrity is threatened by Data Poisoning, where attackers subtly alter cloud-hosted AI training data to cause incorrect automated decisions without triggering traditional security alarms.

3. DDoS Attacks (Distributed Denial of Service)

In 2026, DDoS attacks have reached "terabit-scale," combining massive volumetric floods with low-intensity application-layer tactics that mimic legitimate users. These attacks aim to create "strategic fatigue," exhausting IT resources and degrading hardware over time.

4. NIS2 Requirements for Providers

Under the NIS2 Directive, infrastructure providers must provide an Early Warning within 24 hours of an incident, prove Management Liability for security negligence, and implement "all-hazards" risk management covering both cyber and physical threats.

5. Redundancy and Resilience Strategies

Security is improved through Immutable Backups (WORM storage) that cannot be encrypted by ransomware, Self-Healing Networks that automatically isolate compromised servers, and Geographic Redundancy to ensure services stay online even during regional power grid failures.

9.Υπηρεσίες ΤΠΕ

What are the main cybersecurity risks faced by ICT service providers?

How can cyberattacks on ICT services affect other critical sectors?

What are the risks associated with third-party ICT service providers and supply chain attacks?

What security measures are required under the NIS2 directive for ICT service management providers?

How can incident response strategies improve the resilience of ICT service providers?

9.α CHATGPT

1. Cybersecurity risks in ICT service providers

ICT providers face threats such as ransomware, data breaches, DDoS attacks, insider threats, and system misconfigurations. Their central role makes them high-value targets.

2. Impact on other critical sectors

Attacks on ICT services can disrupt multiple sectors like energy, banking, and healthcare. A single breach can cause widespread outages and operational failures.

3. Risks from third-party & supply chain

Third-party providers may introduce vulnerabilities through insecure software or access. Supply chain attacks can spread malware across many organizations at once.

4. NIS2 requirements for ICT providers

NIS2 requires risk management, access control, encryption, incident reporting, continuous monitoring, and supply chain security. Providers must ensure resilience and secure service delivery.

5. Role of incident response strategies

Effective incident response helps detect, contain, and recover from attacks. It reduces damage, ensures quick restoration of services, and improves overall resilience.

9.β GEMINI

1. Main Cybersecurity Risks

The top threats are Administrative Identity Theft via AI-deepfakes and Autonomous Malware that moves through provider networks at machine speed to evade detection.

2. Cascading Effects on Critical Sectors

A breach in an ICT provider acts as a "skeleton key," allowing attackers to simultaneously disable Hospital EHRs, Bank payment gateways, and Power grid controls through trusted maintenance tunnels.

3. Third-Party & Supply Chain Risks

The greatest danger lies in N-th party vulnerabilities—malicious code hidden in sub-components or open-source libraries that the primary ICT provider uses, creating an invisible "backdoor" into the entire digital ecosystem.

4. NIS2 Requirements for ICT Providers

Under NIS2, ICT management providers must provide an Early Warning within 24 hours, undergo mandatory C-Suite Cybersecurity Training, and perform rigorous High-Risk Vendor Vetting for all hardware and software components.

5. Incident Response & Resilience

Resilience is maintained through Agentic SOCs that automate 90% of threat triage and Immutable Backups (WORM storage) that ensure data cannot be deleted or encrypted by ransomware during an attack.

10.Τομέας Δημόσιας Διοίκησης

What are the main cybersecurity threats affecting public administration systems?

How can cyberattacks impact government services and citizen data?

What are the risks associated with digital government platforms and e-governance systems?

What security measures are required under the NIS2 directive for public sector organizations?

How can cybersecurity awareness and training improve the protection of public administration systems?

10.α CHATGPT

1. Cybersecurity threats in public administration

Public systems face threats such as ransomware, phishing, data breaches, DDoS attacks, and insider threats. Government platforms are frequent targets due to sensitive data and critical services.

2. Impact on government services & data

Cyberattacks can disrupt public services, delay operations, and expose citizen data. This may lead to identity theft, loss of trust, and administrative failures.

3. Risks in e-governance platforms

Digital platforms face risks like hacking, data leaks, system failures, and weak authentication. High interconnectivity increases the attack surface.

4. NIS2 requirements for public sector

NIS2 requires risk management, strong access control, encryption, incident reporting, continuous monitoring, and supply chain security. It also focuses on resilience and fast recovery.

5. Role of awareness & training

Cybersecurity training helps employees recognize threats like phishing and avoid mistakes. It reduces human error and strengthens overall system protection.

10.β GEMINI

1. Main Cybersecurity Threats

The primary risks include AI-Powered Social Engineering (using deepfakes to impersonate officials) and State-Aligned APTs that infiltrate government networks to monitor communications or prepare for future sabotage.

2. Impact on Services and Citizen Data

Breaches compromise "Permanent Data" (Social Security and tax IDs) that cannot be reset. Service Paralysis from ransomware can freeze welfare payments, digital IDs, and legal services, leaving citizens without essential support for weeks.

3. Risks of E-Governance Platforms

Centralized Digital ID platforms create a "single point of failure," where one breach provides a key to all connected services. There is also the risk of Shadow AI, where employees inadvertently leak sensitive data into public AI models.

4. NIS2 Security Requirements

Under the NIS2 Directive, public bodies must provide an Early Warning within 24 hours, ensure Management Liability for security failures, and perform mandatory Supply Chain Vetting of all software vendors.

5. Training and Awareness

Modern defense relies on Continuous Simulation (live phishing drills) and Identity Literacy, teaching staff to treat every digital voice or video as a potential deepfake and to verify requests via a "Zero Trust" protocol.

11.Τομέας Διαστήματος

What are the main cybersecurity risks affecting satellite systems and space infrastructure?

How can cyberattacks disrupt satellite communication and navigation systems?

What are the potential consequences of a cyberattack on space-based services such as GPS?

What security measures are required under the NIS2 directive for space infrastructure?

How can advanced monitoring and detection systems improve the cybersecurity of space assets?

11.α CHATGPT

1. Cybersecurity risks in satellite systems

Satellite and space systems face threats such as signal jamming, spoofing, malware, and unauthorized access to ground control stations. Their remote and complex nature increases vulnerability.

2. Disruption of communication & navigation

Cyberattacks can interfere with satellite signals, block communications, or alter navigation data. This affects GPS, telecommunications, and military or aviation systems.

3. Consequences of attacks on GPS

Attacks on GPS can cause navigation errors, transport disruptions, and safety risks in aviation and shipping. They may also impact financial systems and critical infrastructure relying on timing signals.

4. NIS2 requirements for space infrastructure

NIS2 requires risk management, strong authentication, encryption, incident reporting, continuous monitoring, and supply chain security. It also emphasizes resilience and rapid recovery.

5. Role of monitoring & detection systems

Advanced monitoring systems detect anomalies in signals and system behavior in real time. They help identify cyberattacks early and enable quick response to protect space assets.

11.β GEMINI

1. Main Cybersecurity Risks

The primary threats involve Ground Station Breaches, which provide a "backdoor" to command entire satellite fleets, and Supply Chain Contamination, where malicious code is embedded into hardware during the long manufacturing cycle.

2. Disruption of Space Infrastructure

Attackers utilize Signal Spoofing to send fake navigation data to receivers and Uplink/Downlink Jamming to block critical communications. Cyber-Kinetic attacks can even manipulate thrusters to cause orbital collisions or exhaust satellite batteries.

3. Consequences of GPS/GNSS Attacks

A disruption of GPS signals triggers Infrastructure Failure, as power grids and financial markets lose the "timing signals" needed to synchronize transactions. Experts estimate a total outage would cost the global economy over $1 billion per day.

4. NIS2 Requirements for Space Providers

Under NIS2, space operators must provide an Early Warning within 24 hours, ensure Management Liability for security negligence, and implement "All-Hazards" Resilience plans covering both cyberattacks and physical space weather.

5. Advanced Monitoring and Detection

Defense relies on Onboard Intrusion Detection (OIDS)—AI agents that monitor telemetry for "impossible" movements—and Digital Twins that compare a satellite's real-time behavior against a virtual model to instantly detect hijacking attempts.