
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ


# Enhancing Structured Cyber Threat Intelligence Extraction with Local Large Language Models and Retrieval-Augmented Generation

DIPLOMA THESIS

by


**Konstantina Psarrou**



**Επιβλέπων:** Παναγιώτης Τσανάκας

Καθηγητής Ε.Μ.Π.


Αθήνα, Οκτώβριος 2024


ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ


# Enhancing Structured Cyber Threat Intelligence Extraction with Local Large Language Models and Retrieval-Augmented Generation

DIPLOMA THESIS

by


**Konstantina Psarrou**

**Επιβλέπων:** Παναγιώτης Τσανάκας

Καθηγητής Ε.Μ.Π.


Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Οκτωβρίου, 2024.

| | | |
|---|---|---|
| ……..……. | ……..……. | ……..……. |
| Παναγιώτης Τσανάκας | Αθανάσιος Βουλόδημος | Ανδρέας-Γεώργιος Σταφυλοπάτης |
| Καθηγητής Ε.Μ.Π. | Επίκουρος Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. |

Αθήνα, Οκτώβριος 2024

……..

**ΚΩΝΣΤΑΝΤΙΝΑ ΨΑΡΡΟΥ**

Διπλωματούχος Ηλεκτρολόγος Μηχανικός

και Μηχανικός Υπολογιστών Ε.Μ.Π.

# Περίληψη

Οι Κυβερνοαπειλές εξελίσσονται συνεχώς. Η εξαγωγή αξιοποιήσιμων δεδομένων από μη δομημένα δεδομένα που αφορούν Πληροφορίες Κυβερνοαπειλών (Cyber Threat Intelligence, CTI) είναι απαραίτητη για τη λήψη τεκμηριωμένων αποφάσεων για την διαφύλαξη των ψηφιακών υποδομών και την άμυνα έναντι εξελιγμένων κυβερνοαπειλών. Οι πρόσφατες εξελίξεις στα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models, LLMs) έχουν προκαλέσει ένα κύμα ερευνών πάνω στη χρήση τους για αυτοματοποίηση της εξαγωγής CTI πληροφορίας. Ωστόσο, τα LLMs είναι επιρρεπή σε ψευδαισθήσεις, οι οποίες μπορεί να οδηγήσουν σε ανακρίβειες ή αναξιόπιστες πληροφορίες, ένας κίνδυνος που δεν είναι αποδεκτός σε κρίσιμα σενάρια Κυβερνοασφάλειας.

Για να αντιμετωπιστούν αυτές οι προκλήσεις, η παρούσα διατριβή προτείνει μια νέα προσέγγιση που χρησιμοποιεί μια διάταξη Επαυξημένης με Ανάκτηση Παραγωγής (Retrieval-Augmented Generation, RAG) σε πλήρως τοπικό περιβάλλον, η οποία ενισχύει την εξαγωγή CTI παρέχοντας σχετικές με το περιεχόμενο πληροφορίες στο LLM. Η γνώση εξάγεται δομημένη σε STIX 2.1 bundles, επιτρέποντας την απερίσπαστη συμμόρφωση με τα πρότυπα της Κυβερνοασφάλειας. Το τοπικό περιβάλλον χρησιμοποιείται για την διασφάλιση του απορρήτου των δεδομένων και τη μείωση της εξάρτησης από υπηρεσίες υπολογιστικού νέφους, ένας παράγοντας σημαντικός για τους οργανισμούς που χειρίζονται ευαίσθητα δεδομένα.

Επιπλέον, η παρούσα μελέτη αξιολογεί τις επιδόσεις διάφορων κορυφαίων τοπικών LLMs πάνω στη συγκεκριμένη εργασία, ενώ παράλληλα διερευνά τεχνικές για την βελτίωση της ακρίβειας και τον μετριασμό των ψευδαισθήσεων. Τα πειραματικά αποτελέσματα καταδεικνύουν ότι οι τεχνικές αυτές βελτιώνουν σημαντικά την συνάφεια των παραγόμενων αποτελεσμάτων και αναδεικνύουν τις δυνατότητες της εφαρμογής του RAG στα LLMs στην αντιμετώπιση των περιορισμών στην αυτοματοποιημένη εξαγωγή CTI.

Λέξεις Κλειδιά— Γενεσιουργός Τεχνητή Νοημοσύνη, Μεγάλα Γλωσσικά Μοντέλα, Επαυξημένη με Ανάκτηση Παραγωγή, Few-shot Learning, Prompt Engineering, Πληροφορίες Κυβερνοαπειλών, Κυβερνοασφάλεια, STIX.

# Abstract

Cyber threats are constantly evolving. Extracting actionable intelligence from unstructured Cyber Threat Intelligence (CTI) data is essential for making informed decisions to safeguard digital infrastructures and defend against sophisticated cyber threats. Recent advancements in Large Language Models (LLMs) have sparked a surge of research into their use for automating CTI extraction. However, LLMs are prone to hallucinations, which can result in inaccurate or unreliable information, an unacceptable risk in critical cybersecurity scenarios.

To address these challenges, this thesis proposes a novel approach utilizing a fully local Retrieval-Augmented Generation (RAG) pipeline, which enhances CTI extraction by providing relevant contextual information to decoder-only LLMs. The extracted intelligence is structured into STIX 2.1 bundles, enabling seamless integration with industry-standard cybersecurity frameworks. The local infrastructure is employed to ensure data privacy and reduce dependence on third-party cloud services, an important factor for organizations handling sensitive threat intelligence data.

Additionally, this study evaluates the performance of several leading local LLMs in this task, while exploring optimization techniques to improve accuracy and mitigate hallucinations. The experimental results demonstrate that these techniques significantly enhance the reliability and relevance of the generated outputs and highlight the potential of RAG-enhanced LLMs in a local environment to overcome key limitations in the automated processing of CTI.

Keywords — Generative AI, Large Language Models, Retrieval-Augmented Generation, Prompt Engineering, Few-shot Learning, Cyber Threat Intelligence, Cybersecurity, STIX

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τους καθηγητές Γιώργο Σπανουδάκη, ιδιοκτήτη της Sphynx, και Παναγιώτη Τσανάκα, κοσμήτορα της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, για την ευκαιρία που μου έδωσαν να εκπονήσω την διπλωματική μου εργασία στο πλαίσιο πρακτικής στη Sphynx, καθώς επίσης και τους συναδέλφους μου Κωνσταντίνο Φυσαράκη, Δημήτρη Ελευθερίου και Ραφαήλ Ελληνιτάκη για την πολύτιμη βοήθεια και καθοδήγηση τους. Η συνεργασία και η υποστήριξη τους ήταν καθοριστική στην ολοκλήρωση της εργασίας μου.

Επιπλέον, θα ήθελα να εκφράσω την ευγνωμοσύνη μου προς την οικογένεια μου, που με στήριξε και με ενθάρρυνε καθ' όλη τη διάρκεια των σπουδών μου. Ευχαριστώ θερμά τους φίλους μου και τους συμφοιτητές μου για τα υπέροχα φοιτητικά χρόνια, είμαι ευγνώμων για αυτό το ταξίδι γνώσης που μοιραστήκαμε. Στα επόμενα ταξίδια!

# Περιεχόμενα

2. Η τεχνολογία των Large Language Models ..... 13
2.1 Η επεξεργασία φυσικής γλώσσας και τα γλωσσικά μοντέλα ..... 13
2.2 Τα πρώτα βήματα της έρευνας των γλωσσικών μοντέλων ..... 13
2.2.1 Στατιστικά γλωσσικά μοντέλα ..... 13
2.2.2 Νευρωνικά γλωσσικά μοντέλα ..... 13
2.3 Βαθιά νευρωνικά δίκτυα και προεκπαιδευμένα γλωσσικά μοντέλα ..... 15
2.4.1 Tokenization ..... 17
2.4.2 Embeddings ..... 17
2.4.3 Encoder ..... 17
2.4.4 Decoder ..... 18
2.4.5 Output Generation ..... 18
2.5 Large Language Models (LLM) ..... 19
2.5.1 Ορισμός ..... 19
2.5.2 Σχεδιασμός και Δυνατότητες ..... 19
2.5.3 Encoder-only και decoder-only μοντέλα ..... 20
2.5.4 Περιορισμοί ..... 22
2.6 Retrieval-Augmented Generation (RAG) ..... 22
3. Το Cybersecurity και το Cyber Threat Intelligence ..... 24
3.1 Το Cybersecurity ..... 24
3.1.1 Ορισμός και Σκοπός ..... 24
3.1.2 Είδη επιθέσεων στον Κυβερνοχώρο ..... 24
3.1.3 Σύγχρονες Προκλήσεις στην Ασφάλεια Πληροφοριών ..... 25
3.2 Το Cyber Threat Intelligence ..... 26
3.2.1 Ορισμός και Σκοπός ..... 26
3.2.2 Είδη Cyber Threat Intelligence ..... 27
3.2.3 Πηγές ..... 28
3.2.4 Εργαλεία ..... 29
3.2.5 Προκλήσεις ..... 30
3.3 Το Πλαίσιο STIX 2.1 ..... 31
3.3.1 Ορισμός και Σκοπός ..... 31
3.3.2 Δομή ..... 32
3.3.3 Χρήση και Προκλήσεις ..... 33
4. Χρήση Τεχνητής Νοημοσύνης για την Εξαγωγή Πληροφοριών Κυβερνοαπειλών ..... 34

4.1 Κίνητρο εργασίας ..... 34
4.2 Σχετικές εργασίες και Προηγούμενες Μελέτες ..... 35
4.3 Συνεισφορά της παρούσας εργασίας ..... 37
4.4 Περιορισμοί ..... 37
5. Πειραματική Διάταξη ..... 38
5.1 Περιγραφή ..... 38
5.2 Εργαλεία και Frameworks ..... 38
5.2.1 Ollama ..... 38
5.2.2 Chroma DB ..... 39
5.2.3 OpenAI Client ..... 40
5.3 Το pipeline του RAG Agent ..... 40
5.3.1 Περιγραφή ..... 40
5.3.2 Λειτουργία ..... 40
5.3.3 Οφέλη ..... 41
6. Αποτελέσματα ..... 43
6.1 Μέθοδος αξιολόγησης της απόδοσης του μοντέλου ..... 43
6.2.1 Prompt Engineering ..... 44
6.2.2 Few-shot Learning ..... 44
6.3 Σύγκριση αποτελεσμάτων των Large Language Models και Prompt Engineering ..... 45
6.4 Αποτελέσματα της τεχνικής few-shot learning ..... 51
6.5 Αποτελέσματα της αρχιτεκτονικής του RAG ..... 52
6.5.1 Περίληψη πειράματος ..... 52
6.5.2 Εκτέλεση πειράματος ..... 53
7. Συμπέρασμα ..... 58
8. Μελλοντικές κατευθύνσεις ..... 59
9. Βιβλιογραφία ..... 60

# 2. Η τεχνολογία των Large Language Models

## 2.1 Η επεξεργασία φυσικής γλώσσας και τα γλωσσικά μοντέλα

Η επεξεργασία φυσικής γλώσσας (natural language processing, NLP) είναι ο κλάδος της τεχνητής νοημοσύνης που ασχολείται με την κατανόηση, επεξεργασία και δημιουργία ανθρώπινης γλώσσας από τις μηχανές. [9] Η NLP έχει πολλές εφαρμογές, όπως την κατηγοριοποίηση κειμένου, ανάλυση συναισθήματος, αναγνώριση οντοτήτων κοκ. Αυτές οι εφαρμογές εμβαθύνουν στην ανάλυση δεδομένων, ένα ζήτημα που ερευνητικά παρουσιάζει ταχεία εξάπλωση τα τελευταία χρόνια, με την συνεχή αύξηση του όγκου των δεδομένων που παράγονται καθημερινά.[10]

Τα γλωσσικά μοντέλα (language models, LMs) είναι υπολογιστικά μοντέλα που μπορούν να κατανοούν και να παράγουν ανθρώπινη γλώσσα. [17] Βασικός τους στόχος είναι να προβλέπουν την πιθανότητα ακολουθιών λέξεων και να δημιουργούν νέο κείμενο, με βάση κάποια δεδομένη είσοδο. Τα παραδοσιακά LMs, στα οποία θα αναφερθούμε πιο αναλυτικά και παρακάτω, έχουν μια πληθώρα πλεονεκτημάτων, ωστόσο παρουσιάζουν πολλές δυσκολίες, ειδικά όταν πρέπει να αντιμετωπίσουν άγνωστες σε αυτά λέξεις και σύνθετα γλωσσικά φαινόμενα. [18] Αυτές οι προκλήσεις οδήγησαν πολλές έρευνες στην αναζήτηση νέων αρχιτεκτονικών και μεθόδων εκπαίδευσης για την βελτίωση των μοντέλων γλώσσας.

## 2.2 Τα πρώτα βήματα της έρευνας των γλωσσικών μοντέλων

### 2.2.1 Στατιστικά γλωσσικά μοντέλα

Τα στατιστικά γλωσσικά μοντέλα (statistical language models, SLMs) είναι οι πρώτες μέθοδοι που αναπτύχθηκαν για την πρόβλεψη ακολουθιών λέξεων. Βασίζονται σε στατιστικές μεθόδους μάθησης και υπολογίζουν την κατανομή διαφόρων φαινόμενων φυσικής γλώσσας, με στόχο την αναγνώριση λόγου. [2] Η βασική ιδέα πίσω από αυτά είναι η πρόβλεψη της επόμενης λέξης με βάση το πιο πρόσφατο περιεχόμενο χρησιμοποιώντας τη υπόθεση Markov. [1] Τα SLMs για συγκεκριμένο περιεχόμενο μήκους n είναι γνωστά και ως n-gram γλωσσικά μοντέλα και έχουν ευρεία εφαρμογή στην ανάκτηση πληροφορίας (information retrieval, IR) και στην NLP. Ωστόσο, παρουσιάζουν σημαντικούς περιορισμούς, καθώς απαιτούν τον υπολογισμό εκθετικού αριθμού πιθανοτήτων. Κάποιες από τις τεχνικές εξομάλυνσης αυτού του ζητήματος είναι η backoff εκτίμηση και η Good-Turing εκτίμηση.[1] Τα n-gram γλωσσικά μοντέλα ανέδειξαν την σημασία των συμφραζόμενων στη γλώσσα και έθεσαν τις βάσεις για πιο προηγμένες τεχνικές.

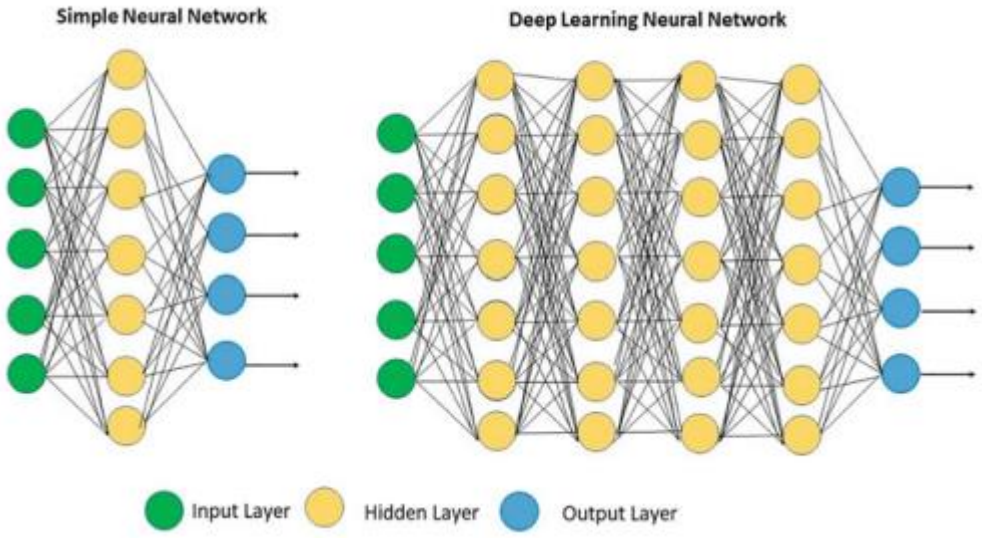
### 2.2.2 Νευρωνικά γλωσσικά μοντέλα

Τα νευρωνικά γλωσσικά μοντέλα (neural language models, NLMs) υπολογίζουν την πιθανότητα ακολουθιών λέξεων χρησιμοποιώντας νευρωνικά δίκτυα (neural networks, NNs). Σε αντίθεση με τα παραδοσιακά στατιστικά μοντέλα, τα NLMs χάρη στην χρήση νευρωνικών δικτύων μπορούν να μάθουν πιο σύνθετες σχέσεις μεταξύ των λέξεων. Νευρωνικά δίκτυα όπως τα πολυεπίπεδα perceptron (multilayer perceptron, MLP), τα τροφοδοτούμενα νευρωνικά δίκτυα (feedforward neural networks, FNNs), τα επαναλαμβανόμενα νευρωνικά δίκτυα (recurrent neural networks, RNNs), και οι εξελιγμένες μορφές τους όπως τα δίκτυα μακράς βραχύχρονης μνήμης (Long Short-Term Memory, LSTM) και τα συνελεκτικά

νευρωνικά δίκτυα (convolutional neural networks, CNNs) καθόρισαν το τοπίο της επεξεργασίας φυσικής γλώσσας και βαθιάς μάθησης, ανοίγοντας νέους ορίζοντες στην κατανόηση της ανθρώπινης γλώσσας.

Τα τροφοδοτούμενα νευρωνικά δίκτυα (feedforward neural networks, FNNs) έχουν μονοκατευθυντική ροή πληροφορίας από την είσοδο προς την έξοδο, χωρίς επαναλαμβανόμενες διαδρομές ή βρόγχους. Επεξεργάζονται τα δεδομένα σειριακά, μια λέξη τη φορά και δεν μπορούν να κρατήσουν συμφραζόμενα. Αντίθετα, τα επαναλαμβανόμενα νευρωνικά δίκτυα (recurrent neural networks, RNNs), έχουν αμφίδρομη ροή, και μπορούν να προβλέπουν την επόμενη λέξη λαμβάνοντας υπόψη την τρέχουσα λέξη και την προηγούμενη κρυφή κατάσταση. Δημιουργούν δηλαδή μια ακολουθία κρυφών καταστάσεων, όπου κάθε κρυφή κατάσταση εξαρτάται από την προηγούμενη. Συνεπώς, τα RNNs ξεπέρασαν τους περιορισμούς των FNNs, ωστόσο αντιμετώπισαν δυσκολίες στην μοντελοποίηση μακροπρόθεσμων εξαρτήσεων, λόγω του προβλήματος εξαφανιζόμενης κλίσης (vanishing gradient problem), πρόβλημα που έλυσαν τα πιο εξελιγμένα, τύπου RNN, LSTM νευρωνικά δίκτυα.[19]

Παρά την εξέλιξη των RNNs όμως στην επεξεργασία συμφραζόμενων, εξακολουθούν να αντιμετωπίζουν δυσκολίες στην μοντελοποίηση σύνθετων γλωσσικών σχέσεων. Σημαντική βελτίωση σε αυτό το ζήτημα ήταν η εισαγωγή της έννοιας της κατανεμημένης αναπαράστασης λέξεων (distributed word vectors)[3], η οποία εισήγαγε την δημιουργία συναρτήσεων πρόβλεψης λέξεων που εξαρτώνται από τα συμφραζόμενα. Ένας επίσης πολύ σημαντικός αλγόριθμος είναι ο αλγόριθμος word2vec, που πρότεινε ένα απλοποιημένο νευρωνικό δίκτυο για την εκμάθηση κατανεμημένων αναπαραστάσεων λέξεων, δείχνοντας την αποτελεσματικότητα του σε NLP εργασίες. [11] Αυτές κι άλλες αντίστοιχες μελέτες είχαν σημαντικό αντίκτυπο στον τομέα της επεξεργασίας φυσικής γλώσσας, και χάραξαν τον δρόμο για την αναπαράσταση της μάθησης πέρα από την μοντελοποίηση ακολουθίας λέξεων.[1]




*Εικόνα 1 Σύγκριση μεταξύ απλού νευρωνικού δικτύου και βαθύ νευρωνικού δικτύου [32]*

## 2.3 Βαθιά νευρωνικά δίκτυα και προεκπαιδευμένα γλωσσικά μοντέλα

Τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks - DNNs) [32] είναι μια εξελιγμένη μορφή νευρωνικών δικτύων, η οποία περιλαμβάνει πολλαπλά κρυφά επίπεδα (layers) μεταξύ της εισόδου και της εξόδου. Η δυνατότητα των DNNs να διαχειρίζονται πολυεπίπεδες και σύνθετες πληροφορίες τα καθιστά εξαιρετικά αποτελεσματικά στη μοντελοποίηση πολύπλοκων σχέσεων και μοτίβων στην επεξεργασία φυσικής γλώσσας (NLP). Σε αντίθεση με τα παραδοσιακά νευρωνικά δίκτυα, που συνήθως έχουν μόνο ένα ή δύο επίπεδα, τα DNNs μπορούν να περιέχουν δεκάδες ή ακόμα και εκατοντάδες επίπεδα. Αυτό επιτρέπει την ανάπτυξη ιεραρχικών αναπαραστάσεων, όπου τα πρώτα επίπεδα μαθαίνουν πιο βασικά χαρακτηριστικά, όπως λέξεις ή φράσεις, ενώ τα επόμενα μαθαίνουν πιο σύνθετα χαρακτηριστικά, όπως οι σχέσεις και τα συμφραζόμενα αυτών των λέξεων.

Τα DNNs έθεσαν τις βάσεις για την ανάπτυξη των σύγχρονων γλωσσικών μοντέλων. Χάρη στη δυνατότητά τους να μαθαίνουν και να διαχειρίζονται πολύπλοκα και μεγάλα σύνολα δεδομένων, τα DNNs χρησιμοποιούνται για την εκπαίδευση προεκπαιδευμένων γλωσσικών μοντέλων (pre-trained language models.

Η προεκπαίδευση (pre-training) είναι μια θεμελιώδης διαδικασία της εκπαίδευσης των νευρωνικών δικτύων, ιδίως στον τομέα της επεξεργασίας φυσικής γλώσσας. Κατά την προεκπαίδευση, το μοντέλο εκπαιδεύεται σε ένα μεγάλο σύνολο δεδομένων, για την εκμάθηση γενικών χαρακτηριστικών, πριν ακολουθήσει η κατάλληλη προσαρμογή του σε κάποια συγκεκριμένη εργασία (fine-tuning). Αυτή η τεχνική βοηθάει το μοντέλο να αποκτήσει ευρεία γνώση των δομών δεδομένων και των μοτίβων, χωρίς να χρειάζεται δεδομένα με ετικέτες (labeled data). [12]

Μια πρώιμη προσπάθεια για προεκπαίδευση που παρουσιάστηκε είναι το ELMo (embeddings from language models), το οποίο αρχικά προεκπαιδεύει ένα δίκτυο αμφίδρομου LSTM (bidirectional Long Short-Term Memory, biLSTM) κι έπειτα εκτελεί κατάλληλη προσαρμογή (fine-tuning) του δικτύου σε συγκεκριμένες εργασίες με σκοπό να καταγράφει αναπαραστάσεις λέξεων με βάση τα συμφραζόμενα.[5]

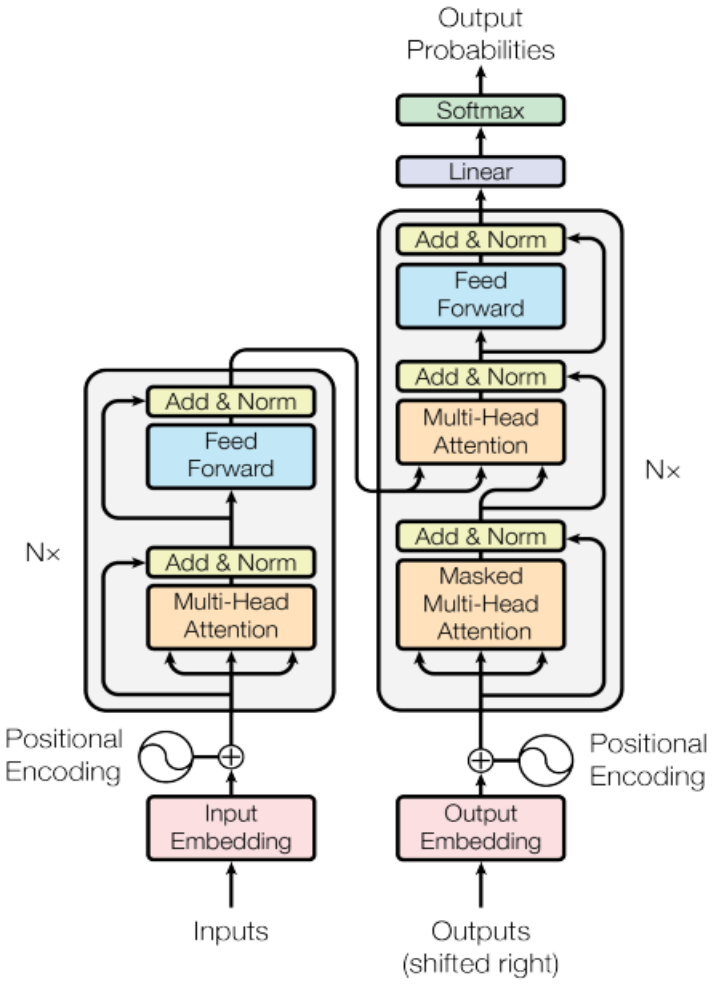
Παρόλο που αυτές οι εξελιγμένες παραλλαγές των νευρωνικών δικτύων για τις οποίες μιλήσαμε έχουν δείξει αποτελεσματικότητα στη μοντελοποίηση ακολουθιών, είναι αρκετά περιοριστικές καθώς εξακολουθούν να παρουσιάζουν δυσκολίες στην παράλληλη επεξεργασία και στην διατήρηση μακροπρόθεσμων εξαρτήσεων, λόγω της σειριακής τους φύσης. Τα DNNs επέτρεψαν τη δημιουργία γλωσσικών μοντέλων που δεν στηρίζονται μόνο στη σειριακή επεξεργασία λέξεων, αλλά εκμεταλλεύονται μηχανισμούς προσοχής (attention mechanisms) και τη δυνατότητα παράλληλης επεξεργασίας που προσφέρουν οι αρχιτεκτονικές των Μετασχηματιστών (Transformers). Αυτό επέτρεψε τη δραστική βελτίωση της απόδοσης και της ακρίβειας των γλωσσικών μοντέλων σε διάφορες εργασίες NLP, όπως είναι η κατανόηση κειμένου και η παραγωγή απαντήσεων σε φυσική γλώσσα.[15][16]

## 2.4 Transformer

Οι Μετασχηματιστές (Transformers)[7] αποτελούν μια από τις πιο σημαντικές και καινοτόμες εξελίξεις στην επεξεργασία φυσικής γλώσσας. Η αρχιτεκτονική τους ήρθε να αντικαταστήσει αυτή της σειριακής επεξεργασίας δεδομένων, χρησιμοποιώντας τον

μηχανισμό αυτό-προσοχής (self-attention mechanism), ο οποίος επιτρέπει στο μοντέλο να δίνει προσοχή σε όλες τις λέξεις μιας πρότασης ταυτόχρονα, χωρίς να περιορίζεται από τη θέση της λέξης στο κείμενο. Συγκεκριμένα, κάθε λεκτική μονάδα (token) επεξεργάζεται λαμβάνοντας υπόψη το ευρύτερο πλαίσιο μέσω του μηχανισμού παράλληλης προσοχής πολλαπλών κεφαλών (multi-head attention mechanism), κατά τον οποίο ενισχύεται η σημασία των βασικών λέξεων και μειώνεται των λιγότερο σημαντικών. Ουσιαστικά, μιμείται την ανθρώπινη προσοχή, αποδίδοντας διαφορετικά επίπεδα σημασίας στις λέξεις σε μια πρόταση.

Η τεχνολογία αυτή επιτρέπει στους Μετασχηματιστές να καταγράφουν πιο αποδοτικά μεγάλες εξαρτήσεις συγκριτικά με τις προηγούμενες αρχιτεκτονικές, καθώς μείωσε σημαντικά τον αναγκαίο χρόνο για εκπαίδευση λόγω απουσίας επαναλαμβανόμενων δικτύων, καθιστώντας τους ιδανικούς για εκπαίδευση πάνω σε μεγάλα σύνολα δεδομένων (datasets). Επίσης, επιτρέπει σημαντικά μεγαλύτερο παραλληλισμό συγκριτικά με προηγούμενες αρχιτεκτονικές, φτάνοντας πολύ μεγαλύτερα επίπεδα ποιότητας μετάφρασης μετά από εκπαίδευση πολύ λιγότερου χρόνου. Η αρχιτεκτονική του Μετασχηματιστή χωρίζεται σε δυο κύρια μέρη, τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder). Ακολουθεί περιγραφή της αρχιτεκτονικής του Μετασχηματιστή.




*Εικόνα 2 Το Transformer - αρχιτεκτονική μοντέλου [7]*

### 2.4.1 Tokenization

Αρχικά, το κείμενο εισόδου πρέπει να αναπαρασταθεί αριθμητικά, ώστε να μπορούν να γίνουν οι απαραίτητοι υπολογισμοί πάνω σε αυτό. Η διαδικασία παραγωγής λεκτικών μονάδων (tokenization) αναλύει ένα κείμενο και το μετατρέπει σε μια συμπιεσμένη ακολουθία συμβόλων. Η διαδικασία αυτή δημιουργεί ένα διάνυσμα από ακεραίους, όπου κάθε ακέραιος αντιστοιχεί σε ένα κομμάτι κειμένου. Η πρώτη δημοσίευση για τον Μετασχηματιστή [13] χρησιμοποιεί byte-pair encoding (BPE) ως μέθοδο tokenization. Το BPE είναι μια μορφή συμπίεσης, όπου οι πιο συνηθισμένοι διαδοχικοί χαρακτήρες (bytes) συμπιέζονται σε έναν μόνο ακέραιο (byte).

Πρόσφατες έρευνες δείχνουν ότι το BPE δεν είναι η βέλτιστη μέθοδος. Πιο πρόσφατα γλωσσικά μοντέλα, όπως το BERT, χρησιμοποιούν WordPiece παραγωγούς λεκτικών μονάδων (tokenizers), οι οποίοι σε αντίθεση με το BPE που κωδικοποιεί κομμάτια λέξεων, μπορεί να συγκεντρώνει ολόκληρες λέξεις σε ένα token, το οποίο μπορεί να είναι και διαφορετικό σύμβολο από ακέραιο όπως πχ. Γράμμα της αλφαβήτου. Αυτό τον καθιστά πιο εύκολο στην αποκωδικοποίηση και πιο διαισθητικό.

### 2.4.2 Embeddings

Τα embeddings [20] είναι διανυσματικές αναπαραστάσεις των λεκτικών μονάδων (tokens), οι οποίες αποδίδουν τη σημασιολογική ομοιότητα μεταξύ των λέξεων. Οι αναπαραστάσεις λέξεων (word embeddings) είναι διανύσματα πραγματικών τιμών που κωδικοποιούν την σημασία μιας λέξης με τέτοιο τρόπο ώστε οι λέξεις που βρίσκονται πιο κοντά στον διανυσματικό χώρο να βρίσκονται πιο κοντά σημασιολογικά.

Αφού κάθε λεκτική μονάδα (token) της πρότασης έχει μετατραπεί σε διάνυσμα μέσω των embeddings, ακολουθεί η μετατροπή αυτού του συνόλου διανυσμάτων σε ένα διάνυσμα. Ο πιο συνήθης τρόπος να γίνει αυτό είναι να προστεθούν μεταξύ τους κατά συνιστώσες. Στην περίπτωση των Μετασχηματιστών όμως αυτό δεν μπορεί να λειτουργήσει, καθώς η πρόσθεση είναι αντιμεταθετική, οπότε αν προστεθούν οι ίδιοι αριθμοί με διαφορετική σειρά, θα πάρουμε το ίδιο αποτέλεσμα. Για παράδειγμα η πρόταση «I’m not , I’m happy» και η πρόταση «I’m not happy, I’m sad» θα καταλήξουν να έχουν το ίδιο διάνυσμα, καθώς έχουν τις ίδιες λέξεις με αντίθετη σειρά, ενώ έχουν το ακριβώς αντίθετο νόημα.

Οι κωδικοποιήσεις θέσης (positional encodings) είναι διανυσματικές αναπαραστάσεις σταθερού μεγέθους που εισάγουν πληροφορία για την θέση των λέξεων σε μια ακολουθία δεδομένων, μια και οι Μετασχηματιστές δεν γνωρίζουν την σειρά των λέξεων καθώς δεν έχουν επαναληπτική δομή. Χρησιμοποιούν ημιτονοειδείς συναρτήσεις οι οποίες επιτρέπουν στο μοντέλο να κατανοεί τη σειρά των λέξεων, διατηρώντας έτσι τη σημασία της θέσης τους στο κείμενο.

### 2.4.3 Encoder

Ο κωδικοποιητής (encoder) στους Μετασχηματιστές αποτελείται από πολλαπλά επαναλαμβανόμενα επίπεδα (layers), συνήθως έξι. Κάθε επίπεδο περιλαμβάνει δύο κύρια

υποεπίπεδα: τον μηχανισμό προσοχής πολλαπλών κεφαλών (multi-head self-attention) και ένα πλήρως συνδεδεμένο τροφοδοτούμενο νευρωνικό δίκτυο (feedforward neural network, FNN). Ο μηχανισμός προσοχής πολλαπλών κεφαλών επιτρέπει στο μοντέλο να εξετάζει ταυτόχρονα πληροφορίες από διαφορετικές θέσεις του κειμένου. Μπορεί δηλαδή να παρατηρήσει διαφορετικές σχέσεις ή αλληλεπιδράσεις ανάμεσα στις λέξεις του κειμένου, βελτιώνοντας έτσι την κατανόηση. Μετά τον μηχανισμό προσοχής, η αναπαράσταση κάθε λεκτικής μονάδας (token) περνάει από το τροφοδοτούμενο νευρωνικό δίκτυο με βάση τη θέση (poition-wise feedforward neural network, FNN). Αυτό εισάγει τη μη γραμμικότητα και βελτιώνει την αναπαράσταση των λεκτικών μονάδων ακόμη περισσότερο. [7] Τέλος, χρησιμοποιείται μια υπολειμματική σύνδεση (residual connection) γύρω από καθένα από τα δύο υποεπίπεδα, ακολουθούμενη κανονικοποίηση επιπέδου (layer normalization).

### 2.4.4 Decoder

Ο αποκωδικοποιητής (decoder) στους Μετασχηματιστές αποτελείται κι αυτός από πολλαπλά επαναλαμβανόμενα επίπεδα (layers), όπως ο κωδικοποιητής. Εκτός από τα δύο υποεπίπεδα σε κάθε επίπεδο κωδικοποιητή, ο αποκωδικοποιητής έχει και ένα τρίτο υποεπίπεδο, το οποίο εκτελεί προσοχή πολλαπλών κεφαλών (multi-head encoder-decoder attention) στην έξοδο του κωδικοποιητή. Ομοίως με τον κωδικοποιητή, χρησιμοποιούνται υπολειμματικές συνδέσεις γύρω από κάθε υποεπίπεδο, ακολουθούμενη από κανονικοποίηση επιπέδου. Επιπλέον, τροποποιείται το επίπεδο αυτό-προσοχής (masked multi-head self-attention) στον αποκωδικοποιητή, ώστε να αποτρέπεται η προσοχή σε μελλοντικές θέσεις. Αυτή η απόκρυψη (masking), σε συνδυασμό με το γεγονός ότι οι ενσωματώσεις της εξόδου μετατοπίζονται κατά μία θέση, διασφαλίζει ότι οι προβλέψεις για μια θέση i εξαρτώνται μόνο από τις ήδη γνωστές εξόδους σε θέσεις μικρότερες από το i. [7]

### 2.4.5 Output Generation

Η διαδικασία παραγωγής εξόδου (output generation) στους Μετασχηματιστές ξεκινά μετά το τέλος της επεξεργασίας από τον αποκωδικοποιητή. Στο τελικό επίπεδο του αποκωδικοποιητή, οι αναπαραστάσεις που έχουν παραχθεί μετατρέπονται σε κατανομές πιθανότητας για το λεξιλόγιο της εξόδου. Αυτό γίνεται μέσω μιας γραμμικής μετατροπής (linear transformation) που μετατρέπει τα διανύσματα εξόδου σε βαθμολογίες (logits) για κάθε πιθανή λέξη του λεξιλογίου. Στη συνέχεια, η συνάρτηση softmax εφαρμόζεται για να μετατρέψει αυτές τις βαθμολογίες σε πιθανότητες.

Κατά την εκπαίδευση, η διαδικασία είναι αυτοπαλινδρομική (auto-regressive), δηλαδή το μοντέλο χρησιμοποιεί τις λέξεις που έχει ήδη παράξει για να προβλέψει την επόμενη λέξη. Αυτό επιτρέπει στο μοντέλο να δημιουργήσει προτάσεις λέξη προς λέξη, με τη σειρά που χρειάζεται, με βάση στην προηγούμενη έξοδο. Η συνάρτηση softmax διασφαλίζει ότι η τελική έξοδος αντιστοιχεί στην πιο πιθανή λέξη, καθιστώντας τη διαδικασία πρόβλεψης αρκετά αποτελεσματική και ακριβή. [20]

## 2.5 Large Language Models (LLM)

### 2.5.1 Ορισμός

Τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models, LLMs) είναι η εξέλιξη των μοντέλων γλώσσας, η οποία διαθέτει τεράστιο αριθμό παραμέτρων και εξαιρετικές δυνατότητες μάθησης. Ο πυρήνας των πιο προηγμένων LLMs είναι ο μηχανισμός αυτοπροσοχής (self-attention mechanism) του Μετασχηματιστή (Transformer), ο οποίος αποτελεί τη θεμελιώδη δομική μονάδα για εργασίες μοντελοποίησης γλώσσας. Τα μοντέλα αυτά εκπαιδεύονται σε τεράστιες ποσότητες δεδομένων κειμένου και έχουν τη δυνατότητα να παράγουν υψηλής ποιότητας κείμενο.[18] Σε αντίθεση με τα πρώτα γλωσσικά μοντέλα που προαναφέραμε, τα οποία περιορίζονται στο να μοντελοποιούν και παράγουν κείμενο, τα LLMs νέας γενιάς μπορούν να λύνουν σύνθετα προβλήματα και να κάνουν πολύπλοκες εργασίες.

### 2.5.2 Σχεδιασμός και Δυνατότητες

Τα LLMs υλοποιούνται κυρίως με αρχιτεκτονική Μετασχηματιστών, η οποία αποτελείται από πολλά επίπεδα προσοχής πολλαπλών κεφαλών (multi-head attention), δημιουργώντας ένα ισχυρό νευρωνικό δίκτυο που μπορεί να μάθει πολύπλοκες σχέσεις στο κείμενο. Στα LLMs, αυτή η αρχιτεκτονική είναι σε πολύ μεγαλύτερη κλίμακα συγκριτικά με μικρότερα μοντέλα, καθώς χρησιμοποιούνται εκατοντάδες ή και χιλιάδες επίπεδα προσοχής. Πολλές έρευνες έχουν δείξει ότι η κλιμάκωση του μοντέλου μπορεί να το βελτιώσει σε μεγάλο βαθμό την απόδοσή του, καθώς έτσι το μοντέλο μαθαίνει πιο πολύπλοκες συσχετίσεις στα δεδομένα του. Συγκεκριμένα, μια μελέτη της OpenAI συνέδεσε την απόδοση του μοντέλου με τρεις βασικούς παράγοντες: το μέγεθος του, το μέγεθος των δεδομένων και την υπολογιστική ισχύ. Σύμφωνα με αυτή τη μελέτη, η απόδοση του μοντέλου αυξάνεται, όσο αυξάνονται αυτοί οι τρεις παράγοντες. [27]

Όσο το μέγεθος του μοντέλου αυξάνεται, νέες ικανότητες «ξεκλειδώνονται» σε αυτό, φανερώνοντας πολύ ενδιαφέροντα χαρακτηριστικά των LLMs που μπορούν να αξιοποιηθούν για διάφορους σκοπούς. Ουσιαστικά, παρουσιάζονται έντονες βελτιώσεις σε συγκεκριμένες λειτουργίες όταν το μοντέλο φτάσει σε μια συγκεκριμένη κλίμακα. Αυτές οι ικανότητες είναι οι εξής:

- Μάθηση με βάση τα συμφραζόμενα (In-Context Learning, ICL): Αυτή η ικανότητα επιτρέπει στο μοντέλο να ολοκληρώνει εργασίες με βάση οδηγίες ή παραδείγματα που του δίνονται κατά τη χρήση του, χωρίς επιπλέον εκπαίδευση.
- Ακολούθηση οδηγιών (Instruction Following): Μέσω της εκπαίδευσης σε σύνολα δεδομένων που αναφέρουν πολλές διαφορετικές εργασίες που περιγράφονται σε φυσική γλώσσα (instruction tuning), τα LLMs μπορούν να ακολουθούν οδηγίες για νέες εργασίες, βελτιώνοντας την ικανότητά τους να γενικεύουν σε αυτές τις εργασίες.
- Σκέψη βήμα-βήμα (Step-by-Step Reasoning): Αυτή η ικανότητα επιτρέπει στα LLMs να λύνουν πολύπλοκες εργασίες λογικής και αριθμητικών υπολογισμών, ακολουθώντας ενδιάμεσα βήματα συλλογισμού. Αυτή η προσέγγιση είναι γνωστή και ως Chain-of-Thought (CoT) Prompting.

Αξίζει να σημειωθεί ότι η εκπαίδευση μοντέλων τέτοιου μεγέθους είναι μεγάλη πρόκληση, καθώς για αυτή απαιτούνται πολύ μεγάλου μεγέθους υπολογιστικοί πόροι, εξειδικευμένο

hardware όπως TPUs και GPU σε υπερυπολογιστές και αρκετός χρόνος. Παρόλο που είναι εξαιρετικά δαπανηρή διαδικασία, τα αποτελέσματα των μοντέλων έχουν εντυπωσιακές δυνατότητες και μεγάλα ποσοστά ακρίβειας.

Οι εξελίξεις στα LLMs οφείλονται σε μεγάλο βαθμό στην υιοθέτηση της αρχιτεκτονικής του Μετασχηματιστή. Όπως είδαμε παραπάνω, κεντρικό ρόλο σε αυτή την αρχιτεκτονική παίζουν ο κωδικοποιητής (encoder) και ο αποκωδικοποιητής (decoder), τα οποία έχουν χρησιμοποιηθεί διακριτά σε διάφορα LLMs. [21]

### 2.5.3 Encoder-only και decoder-only μοντέλα

#### 2.5.3.1 Encoder-only models

Τα encoder-only μοντέλα αποτελούν τα πρώτα βήματα προς την εξέλιξη των LLMs, με βασικό εκπρόσωπο τους το BERT (Bidirectional Encoder Representations from Transformers). Το BERT είναι ένα προεκπαιδευμένο μοντέλο βασισμένο αποκλειστικά στον κωδικοποιητή του Μετασχηματιστή, που του επιτρέπει να κατανοεί σε βάθος το πλαίσιο και τις λεπτομέρειες της γλώσσας. Ένα από τα κύρια χαρακτηριστικά του είναι η αμφίδρομη εκπαίδευση (bidirectional training), η οποία επιτρέπει στο μοντέλο να κατανοεί τη γλώσσα λαμβάνοντας υπόψη τόσο το προηγούμενο όσο και το επόμενο κείμενο, σε αντίθεση με άλλα μοντέλα που χρησιμοποιούν μόνο μονοκατευθυντική προσοχή. Αυτή του η δυνατότητα του επιτρέπει να έχει μια συνολική κατανόηση της δομής και των σχέσεων στο κείμενο, καθιστώντας το εξαιρετικά αποδοτικό σε εργασίες όπως η απάντηση ερωτήσεων, η ταξινόμηση κειμένου και η ανάλυση συναισθημάτων.[23]

Μια βελτιωμένη εκδοχή του BERT είναι το RoBERTa (Robustly Optimized BERT Pre-training Approach), το οποίο έκανε ενδελεχή αξιολόγηση των παραμέτρων εκπαίδευσης και του μεγέθους του συνόλου δεδομένων που χρησιμοποιήθηκαν αρχικά από το BERT, και επέκτεινε τις δυνατότητες του. Η μελέτη αυτή ανέδειξε πως το BERT δεν είχε εκπαιδευτεί επαρκώς σε σχέση με τις δυνατότητές του, και πρότειναν το RoBERTa ως βέλτιστη μέθοδο προεκπαίδευσης. Οι βελτιώσεις αυτές περιλάμβαναν μεγαλύτερη διάρκεια εκπαίδευσης του μοντέλου σε μεγαλύτερες παρτίδες δεδομένων (batches) και μεγαλύτερα σύνολα δεδομένων (datasets), αλλά και σε μεγαλύτερες ακολουθίες, καθώς την κατάργηση της πρόβλεψης επόμενης ακολουθίας (next sequence prediction) που υπήρχε στο BERT. Με αυτές τις βελτιώσεις το RoBERTa κατέστη ένα από τα πιο ισχυρά και αποτελεσματικά μοντέλα κατανόησης γλώσσας.

#### 2.5.3.2 Decoder-only models

Τα Decoder-Only μοντέλα έχουν διαμορφώσει ριζικά το τοπίο των Μεγάλων Γλωσσικών Μοντέλων (LLMs). Η εκπαίδευσή τους γίνεται σε τεράστια σύνολα δεδομένων, περιλαμβάνοντας δισεκατομμύρια παραμέτρους, και απαιτεί σημαντικούς υπολογιστικούς πόρους και υποδομή. [24] Χαρακτηριστικό τους είναι ότι χρησιμοποιούν μόνο τον αποκωδικοποιητή (decoder) από την αρχιτεκτονική του Μετασχηματιστή, ο οποίος τους επιτρέπει να παράγουν κείμενο με αυτοπαλίνδρομο τρόπο (autoregressive text generation). Αυτή η ιδιότητα τους επιτρέπει να παράγουν το τελικό κείμενο λέξη προς λέξη, δημιουργώντας φυσικές προτάσεις πολύ κοντά στον ανθρώπινο λόγο. [1]

Σε αντίθεση με τα encoder-only μοντέλα, τα οποία εστιάζουν στην κατανόηση και την ερμηνεία του κειμένου, τα decoder-only μοντέλα εξειδικεύονται στην παραγωγή συνεκτικού και λογικά διαρθρωμένου κειμένου [25]. Ως αυτοπαλινδρομικά, μπορούν να προβλέπουν την επόμενη λέξη με βάση τις προηγούμενες στην ακολουθία, καθώς υποθέτουν ότι οι μελλοντικές λέξεις εξαρτώνται από τις προγενέστερες τους. Αυτά τα μοντέλα μαθαίνουν τα στατιστικά μοτίβα και εξαρτήσεις στα δεδομένα εκπαίδευσης, και στη συνέχεια χρησιμοποιούν αυτή τη γνώση για να δημιουργήσουν κείμενο που θα είναι συνεκτικό και σχετικό με το περιεχόμενο.

Ακολουθεί περιγραφή των πιο δημοφιλών decoder-only μοντέλων:

- GPT-4 (OpenAI): Το GPT-4 είναι η πιο πρόσφατη εξέλιξη της σειράς μοντέλων από την OpenAI και θεωρείται ένα από τα πιο προηγμένα γλωσσικά μοντέλα σήμερα, καθώς μπορεί να επεξεργάζεται και κείμενο και εικόνα. Αποτελείται από πολλά δισεκατομμύρια παραμέτρους, περισσότερες από τον προκάτοχο του GPT-3.5, οι οποίες βελτίωσαν τις δυνατότητες του κατά 40% σε θέματα ακρίβειας και ασφάλειας. [26] Είναι κλειστού κώδικα (closed-source) μοντέλο, δηλαδή δεν είναι δημόσια διαθέσιμο για τροποποίηση ή περεταίρω εκπαίδευση από την κοινότητα, μπορεί όμως να χρησιμοποιηθεί συνδρομητικά.
- LLaMa 3.1 (Meta): Το LLaMa 3.1 είναι η τελευταία έκδοση στη LLaMa από την Meta και προσφέρει διάφορα μεγέθη μοντέλων, από μικρότερα με λιγότερες παραμέτρους έως πολύ μεγάλα που ξεπερνούν τα 100 δισεκατομμύρια παραμέτρους. Η Meta τονίζει πως το μοντέλο αυτό βελτιώνει τη δυνατότητα της δημιουργίας κειμένου και την ακρίβεια των απαντήσεων, ειδικά σε πολύπλοκες λεκτικές προκλήσεις. Είναι ανοιχτού κώδικα (open-source), γεγονός που επιτρέπει στους προγραμματιστές να πειραματιστούν με αυτό, να το προσαρμόσουν στην ανάγκες τους με επιπλέον εκπαίδευση, σε εξειδικευμένα δεδομένα και να το βελτιώσουν για τις εργασίες που το χρειάζονται.
- Mistral (Mistral AI): Το Mistral είναι ένα μοντέλο ανοιχτού κώδικα, σχεδιασμένο με στόχο την υψηλή απόδοση και αποτελεσματικότητα. Διατίθεται σε διάφορα μεγέθη, με την πιο προηγμένη έκδοση να έχει 12 δισεκατομμύρια παραμέτρους. Το μοντέλο έχει βελτιωθεί για να ανταποκρίνεται με ακρίβεια σε γλωσσικές δοκιμασίες και να επεξεργάζεται μεγάλες σειρές δεδομένων χωρίς προβλήματα. Η Mistral έχει κατασκευάσει το μοντέλο αυτό ώστε να είναι πλήρως διαθέσιμο στην κοινότητα, προωθώντας τη διαφάνεια και τη συνεργασία.
- Gemma-2 (Google): Το Gemma-2 είναι ένα άλλο σημαντικό LLM που εστιάζει στην ακρίβεια και τη γλωσσική πολυμορφία. Προσφέρει πολλαπλές εκδόσεις με διάφορους αριθμούς παραμέτρων, επιτρέποντας την κλιμάκωση ανάλογα με τις ανάγκες των χρηστών. Οι δημιουργοί του τονίζουν ότι το Gemma-2 είναι βελτιστοποιημένο για πολυγλωσσική επεξεργασία και έχει σχεδιαστεί για να υποστηρίζει ειδικές εφαρμογές, όπως η ανάλυση συναισθημάτων και η εξαγωγή πληροφοριών από δομημένα δεδομένα.

### 2.5.4 Περιορισμοί

Τα Μεγάλα Γλωσσικά Μοντέλα, παρά τη σημαντική πρόοδο που έχουν σημειώσει στην κατανόηση και παραγωγή πληροφορίας, παρουσιάζουν σημαντικούς περιορισμούς που χρήζουν προσοχής.

- Ψευδαισθήσεις (Hallucinations): Τα LLMs συχνά παράγουν πληροφορίες που έρχονται σε αντίθεση με την πραγματικότητα ή που δεν μπορούν να επαληθευτούν. Αυτό το ζήτημα είναι υπαρκτό ακόμη και στα πιο εξελιγμένα όπως το GPT-4, καθώς και αυτά δυσκολεύονται να αναγνωρίσουν και να διορθώσουν τέτοια αποτελέσματα. Για την αντιμετώπιση αυτού του προβλήματος, υπάρχουν λύσεις όπως ο συντονισμός ευθυγράμμισης (alignment tuning), όπου τα μοντέλα εκπαιδεύονται σε υψηλής ποιότητας δεδομένα και παίρνουν παρατηρήσεις από ανθρώπους.[28]
- Επικαιρότητα Γνώσης (Knowledge Recency): Τα LLMs αδυνατούν να ανταποκριθούν σε εργασίες που χρειάζονται πολύ ενημερωμένη γνώση, επειδή η ενημέρωση τω μοντέλων με νέα δεδομένα δεν είναι συχνή καθώς είναι μια χρονοβόρα και δαπανηρή διαδικασία. Επιπλέον, η συχνή ενημέρωση των μοντέλων με νέα δεδομένα μπορεί να προκαλέσει το φαινόμενο καταστροφικής λήθης (catastrophic forgetting), κατά το οποίο το μοντέλο στην εκμάθηση της νέας πληροφορίας, «ξεχνάει» πληροφορίες στις οποίες έχει εκπαιδευτεί στο παρελθόν. Για να αντιμετωπιστεί αυτό το ζήτημα, αναπτύσσονται μέθοδοι που χρησιμοποιούν εξωτερικές πηγές γνώσης, όπως μηχανές αναζήτησης, για την εύρεση επίκαιρων πληροφοριών.

## 2.6 Retrieval-Augmented Generation (RAG)

Παρά την αποτελεσματικότητα των LLMs στο να ακολουθούν οδηγίες (ερωτήσεις-απαντήσεις), είναι επιρρεπή σε ψευδαισθήσεις στις παραγόμενες απαντήσεις τους.[28] Αυτό οφείλεται στο ότι βασίζονται σε μεγάλο βαθμό σε γνώση στην οποία έχουν προεκπαιδευτεί, ένας περιορισμός που επιφέρει μεγάλο ρίσκο όταν τα LLMs χρησιμοποιούνται σε εργασίες που χρειάζονται μεγάλη ακρίβεια. Για να προσαρμοστεί η τεχνολογία τους σε νέα πληροφορία και να μπορέσει να τη χρησιμοποιήσει, προτάθηκε η επαυξημένη με ανάκτηση παραγωγή (Retrieval-Augmented Generation, RAG).[29][30] Το RAG είναι μια μέθοδος που ενισχύει τις δυνατότητες των LLMs, ιδίως σε εργασίες που απαιτούν πιο συγκεκριμένη και επίκαιρη γνώση, αξιοποιώντας εξωτερικές πηγές γνώσης.

Η τεχνολογία του RAG μπορεί να εντοπίζει, μέσα από μια βάση γνώσης (knowledge base), που του έχει διαθέσει ο προγραμματιστής, πληροφορία η οποία σχετίζεται με την εργασία που έχει αναθέσει ο χρήστης στο LLM. Έπειτα, αυτή η πληροφορία συμπεριλαμβάνεται στην είσοδο του χρήστη, ώστε να παρέχει στο LLM επιπλέον περιεχόμενο σχετικό με το αντικείμενο του αιτήματος του χρήστη.

Σε αρχικό στάδιο, η προσέγγιση αυτή απαιτούσε και αυτή εκπαίδευση του μοντέλου για κάθε εργασία. Ωστόσο, έρευνες [31] έδειξαν ότι χρησιμοποιώντας ένα προεκπαιδευμένο embedding μοντέλο μπορεί να φέρει καλή απόδοση χωρίς τη χρήση περεταίρω προεκπαίδευσης. Για την υλοποίηση της αρχιτεκτονικής του RAG χρειάζεται ένα προεκπαιδευμένο μοντέλο αναπαράστασης embedding (pre-trained embedding model) και μια διανυσματική βάση δεδομένων (vector database). Συγκεκριμένα, το RAG δημιουργεί μια

αναπαράσταση (embedding) του ερωτήματος του χρήστη στο LLM, η οποία συγκρίνεται με τις αναπαραστάσεις (embeddings) των πληροφοριών που έχουν δοθεί από τον προγραμματιστή στο RAG και βρίσκονται αποθηκευμένες σε μια διανυσματική βάση δεδομένων, ώστε να βρεθούν τα κορυφαία K κομμάτια πληροφορίας που είναι πιο σχετικά. Στην συνέχεια η πληροφορία αυτή ενώνεται με το αίτημα του χρήστη και δίνεται στο LLM για να εκτελέσει την εργασία, πλέον εξοπλισμένο με επικαιροποιημένη πληροφορία που θα το οδηγήσει σε σωστό αποτέλεσμα.[31]

Η προσέγγιση του RAG που του επιτρέπει να ενσωματώνει πληροφορίες σε πραγματικό χρόνο το καθιστά μια καινοτόμα λύση που ξεπερνάει τους περιορισμούς των LLMs και του fine-tuning. Παρουσιάζει ευελιξία στην ενσωμάτωση εξωτερικών πηγών γνώσης, καθιστώντας το ιδιαίτερα αποτελεσματικό σε τομείς που χρήζουν συνεχή ενημέρωση. Ωστόσο, το RAG παρουσιάζει επίσης κάποιους περιορισμούς. Η διαδικασία ανάκτησης της σωστής πληροφορίας από το περιεχόμενο μπορεί να δυσκολέψει, όσο οι πληροφορία που ανεβαίνει στο RAG μεγαλώνει. Τότε, επειδή είναι μεγάλη η ποσότητα της πληροφορίας, η κατανομή της σχετικότητας μεταξύ των πληροφοριών διασκορπίζεται, και το σύστημα πολλές φορές αποτυγχάνει να επιστρέψει την πιο κρίσιμη πληροφορία. Αυτή η σύγχυση μπορεί να οδηγήσει σε ανακρίβειες, για αυτό πρέπει η φόρτωση πληροφορίας στο σύστημα του RAG να είναι πολύ προσεκτική.[13]

# 3. Το Cybersecurity και το Cyber Threat Intelligence

## 3.1 Το Cybersecurity

### 3.1.1 Ορισμός και Σκοπός

Η Κυβερνοασφάλεια (Cybersecurity) [32] είναι η διαδικασία προστασίας των δικτύων, των συσκευών και των δεδομένων από μη εξουσιοδοτημένη πρόσβαση ή εγκληματική χρήση. και η πρακτική της διασφάλισης της εμπιστευτικότητας (confidentiality), της ακεραιότητας (integrity) και της διαθεσιμότητας των πληροφοριών (availability), που αποτελούν τα τρία θεμέλια της ασφάλειας πληροφοριών.

Στην σύγχρονη εποχή, η εξάρτηση από την τεχνολογία και το διαδίκτυο είναι πιο μεγάλη από ποτέ. Η πλειονότητα των ανθρώπινων δραστηριοτήτων βασίζεται σε ψηφιακά συστήματα και υποδομές. Από την επικοινωνία (π.χ. ηλεκτρονικό ταχυδρομείο, smartphones), την ψυχαγωγία (π.χ. διαδραστικά βιντεοπαιχνίδια, μέσα κοινωνικής δικτύωσης, εφαρμογές ), τις μεταφορές (π.χ. συστήματα πλοήγησης), τις ηλεκτρονικές αγορές  (πιστωτικές κάρτες) και την ιατρική (π.χ. ιατρικός εξοπλισμός, ιατρικά αρχεία), και άλλους πολλούς εξίσου κρίσιμους τομείς, τα ψηφιακά συστήματα είναι αναπόσπαστο μέρος τις ομαλής τους λειτουργίας.

Μαζί με την αυξανόμενη εξάρτηση, προκύπτουν και νέοι σοβαροί κίνδυνοι. Μεγάλος όγκος από προσωπικά και ευαίσθητα δεδομένα αποθηκεύεται είτε σε τοπικές συσκευές ή σε πλατφόρμες υπολογιστικού νέφους. Όπου υπάρχουν δεδομένα, υπάρχει και στόχος για κακόβουλες επιθέσεις. Οι παραβιάσεις δεδομένων, οι επιθέσεις με κακόβουλο λογισμικό και οι απόπειρες μη εξουσιοδοτημένης πρόσβασης είναι πλέον καθημερινές απειλές που αντιμετωπίζουν οι χρήστες και οι οργανισμοί.

Με τόσες πολλές κρίσιμες δραστηριότητες να εξαρτώνται από ψηφιακές υποδομές και με τους κινδύνους να αυξάνονται διαρκώς, γίνεται εμφανής η επιτακτική ανάγκη για ολοκληρωμένες λύσεις κυβερνοασφάλειας. Οι οργανισμοί είναι αναγκαίο να επενδύουν σε προληπτικά μέτρα, όπως είναι η κρυπτογράφηση, η πολυπαραγοντική ταυτοποίηση (multi-factor authentication) και τα συστήματα ανίχνευσης εισβολών (Intrusion Detection Systems), ώστε να προστατεύσουν την ιδιωτικότητα και την ακεραιότητα των δεδομένων τους και να διασφαλίσουν την αδιάλειπτη λειτουργία των συστημάτων τους. Η εκπαίδευση των χρηστών και η συνεχής ενημέρωση των συστημάτων ασφαλείας, έτσι ώστε να είναι προετοιμασμένα για νέες φιλοσοφημένες απειλές, είναι απολύτως απαραίτητες για την αντιμετώπιση των σύγχρονων προκλήσεων στον ψηφιακό κόσμο.

### 3.1.2 Είδη επιθέσεων στον Κυβερνοχώρο

Μαζί με την ψηφιοποίηση των διαδικασιών και την εξέλιξη της τεχνολογίας, οι κυβερνοαπειλές εξελίσσονται και γίνονται όλο και πιο πολύπλοκες. Οι κακόβουλοι παράγοντες (threat actors) αναπτύσσουν συνεχώς νέες μεθόδους για να αποκτούν μη εξουσιοδοτημένη πρόσβαση και να εκμεταλλεύονται αδυναμίες στις δικλείδες ασφαλείας. Οι οργανισμοί οφείλουν να προστατευθούν από τέτοιες απειλές, καθώς οι συνέπειες από επιτυχείς επιθέσεις μπορεί να είναι καταστροφικές, όπως μεγάλες οικονομικές απώλειες, διαρροή προσωπικών δεδομένων και μεγάλο πλήγμα στη φήμη του οργανισμού. Πριν παρθούν τα μέτρα προστασίας των πληροφοριακών αγαθών ενός οργανισμού, είναι

σημαντικό να είναι γνωστό αρχικά ποιες είναι οι κύριες κατηγορίες επιθέσεων από τις οποίες κινδυνεύει. Μερικές από τις πιο δημοφιλείς κατηγορίες [33] είναι οι παρακάτω:

- Malware: Κακόβουλο λογισμικό που έχει σκοπό να βλάψει το σύστημα ή να αποκτήσει μη εξουσιοδοτημένη πρόσβαση σε αυτό
- Ransomware: Μια μορφή κακόβουλου λογισμικού ή ενέργειας, που συνήθως κρυπτογραφεί δεδομένα και ο κυβερνοεγκληματίας ζητάει πληρωμή για να τα αποδεσμεύσει
- Phishing and social engineering: Εξαπάτηση των χρηστών μέσω ψεύτικων email, μηνυμάτων ή κλήσεων
- Man-in-the-middle attack: Όταν ο κυβερνοεγκληματίας «κρυφακούει» σε μια σύνδεση μη σωστά προστατευμένου δικτύου με σκοπό την υποκλοπή δεδομένων
- Denial-of-service: Μια επίθεση που κατακλύζει ένα σύστημα από ψεύτικη κίνηση, καθιστώντας το αργό στη χρήση ή μη διαθέσιμο

Η ασφάλιση των οργανισμών από τέτοιες απειλές απαιτεί έναν συνδυασμό από μέτρα προστασίας και κυρίως εκπαιδευτική πολιτική για τους νόμιμους χρήστες των συστημάτων του οργανισμού, καθώς πολλές φορές για μεγάλες επιθέσεις ευθύνεται η μη συμμορφωμένη συμπεριφορά με τους κανονισμούς Κυβερνοασφάλειας.

### 3.1.3 Σύγχρονες Προκλήσεις στην Ασφάλεια Πληροφοριών

Οι κυβερνοεπιθέσεις εξελίσσονται συνεχώς και γίνονται πιο περίπλοκες, καθώς κάθε νέα εξέλιξη της τεχνολογίας των ψηφιακών υποδομών μπορεί να δημιουργήσει νέες, σύνθετες προκλήσεις στην ασφάλεια πληροφοριών. Κάποιες από τις πιο συνήθεις προκλήσεις αναφέρονται παρακάτω.[33]

- Επιθέσεις Μηδενικής Ημέρας (Zero-day vulnerabilities): Οι επιθέσεις μηδενικής ημέρας εκμεταλλεύονται άγνωστες ευπάθειες λογισμικού πριν οι προγραμματιστές έχουν τη δυνατότητα να κυκλοφορήσουν ενημερώσεις ασφαλείας. Αυτές οι ευπάθειες είναι ιδιαίτερα επικίνδυνες, καθώς αφήνουν τα συστήματα εκτεθειμένα χωρίς δυνατότητα άμεσης προστασίας. Οι επιτιθέμενοι μπορούν να αποκτήσουν πρόσβαση σε δεδομένα ή να παραβιάσουν υποδομές προτού εντοπιστεί το πρόβλημα
.
- Ασφάλεια σε Υποδομές Υπολογιστικού Νέφους (Cloud Security): Με τη μετάβαση των δεδομένων σε πλατφόρμες υπολογιστικού νέφους, οι οργανισμοί αντιμετωπίζουν προκλήσεις που σχετίζονται με την ασφάλεια τους. Οι κίνδυνοι περιλαμβάνουν την λανθασμένη πρόσβασης, τις παραβιάσεις δεδομένων, καθώς και την πιθανή διαρροή ευαίσθητων πληροφοριών λόγω εσφαλμένων διαμορφώσεων ή ευπαθειών στις υπηρεσίες νέφους.

- Ανάπτυξη της Τεχνητής Νοημοσύνης (AI) και των Μεγάλων Γλωσσικών Μοντέλων (LLMs): Η τεχνητή νοημοσύνη μπορεί να χρησιμοποιηθεί τόσο για την προστασία όσο και για την ενίσχυση των κυβερνοεπιθέσεων. Από τη μια, τα εργαλεία τεχνητής νοημοσύνης μπορούν να ενισχύσουν τα συστήματα ανίχνευσης απειλών και να αυτοματοποιήσουν τις απαντήσεις σε επιθέσεις. Από την άλλη, οι επιτιθέμενοι μπορούν να χρησιμοποιήσουν τεχνητή νοημοσύνη και LLMs πχ. για τη δημιουργία

πιο πειστικών phishing μηνυμάτων ή για την ανάλυση συστημάτων με σκοπό την εύρεση αδυναμιών.

- Κανονισμοί και Συμμόρφωση: Η αυξανόμενη αυστηρότητα των ρυθμιστικών πλαισίων για την προστασία των προσωπικών δεδομένων, όπως το GDPR, επιβάλλει πρόσθετες προκλήσεις στους οργανισμούς. Η μη συμμόρφωση με αυτούς τους κανονισμούς μπορεί να οδηγήσει σε βαριά πρόστιμα και ζημία στη φήμη των οργανισμών.

Οι σύγχρονες προκλήσεις στην ασφάλεια πληροφοριών απαιτούν από τους οργανισμούς προσεκτική στρατηγική, συνεχή εκπαίδευση και επένδυση σε τεχνολογίες αιχμής. Ο συνδυασμός προληπτικών μέτρων, όπως η τακτική ενημέρωση των συστημάτων και η χρήση κρυπτογράφησης, με τον έλεγχο της ανθρώπινης συμπεριφοράς και τη συμμόρφωση με τα διεθνή πρότυπα, είναι απαραίτητος για την προστασία των πληροφοριακών αγαθών σε έναν διαρκώς μεταβαλλόμενο ψηφιακό κόσμο.

## 3.2 To Cyber Threat Intelligence

### 3.2.1 Ορισμός και Σκοπός

Η Πληροφορία Κυβερνοαπειλών (Cyber Threat Intelligence, CTI) [34] αναφέρεται στη διαδικασία συλλογής, ανάλυσης και οργάνωσης πληροφοριών που αφορούν πιθανές και υπαρκτές απειλές στον κυβερνοχώρο. Ο σκοπός του CTI είναι να βοηθήσει τους οργανισμούς να προβλέπουν, να αναγνωρίζουν και να αποκρίνονται αποτελεσματικά σε κυβερνοεπιθέσεις, δίνοντας τους μια πιο συνειδητή και προληπτική προσέγγιση στον τρόπο με τον οποίο ασφαλίζονται. Η γνώση των απειλών επιτρέπει στους οργανισμούς να προσαρμόζουν την άμυνα τους απέναντι σε αυτές, να κατανοούν καλύτερα τους κινδύνους που αντιμετωπίζουν και να εφαρμόζουν πιο αποτελεσματικές στρατηγικές άμυνας.

Σε μια εποχή όπου το τοπίο των κυβερνοαπειλών εξελίσσεται ραγδαία, η σημασία των πληροφοριών κυβερνοαπειλών (Cyber Threat Intelligence, CTI) είναι πιο κρίσιμη από ποτέ. Για τους οργανισμούς που επιδιώκουν να προστατεύσουν τα πληροφοριακά αγαθά τους, ο ρόλος του CTI στο «κυνήγι απειλών» (threat hunting) είναι καθοριστικός.

Ο κύκλος ζωής του CTI [35] πρόκειται για έναν συνεχή και επαναλαμβανόμενο κύκλο που δίνει τη δυνατότητα στις ομάδες κυβερνοασφάλειας να προβλέπουν, να ανιχνεύουν και να ανταποκρίνονται σε απειλές με μεγαλύτερη αποτελεσματικότητα. Αποτελείται από έξι φάσεις:

1. Κατεύθυνση (Direction): Ξεκινά με τον καθορισμό στόχων, αντικειμενικών σκοπών, εύρους και μεθοδολογίας για τη συλλογή πληροφοριών απειλών, βάσει των απαιτήσεων των εμπλεκομένων μερών. Η αναγνώριση των αναγκών είναι κρίσιμη για να διασφαλιστεί ότι η διαδικασία ευθυγραμμίζεται με τους επιχειρηματικούς στόχους και τους κινδύνους.
2. Συλλογή (Collection): Αφορά τη συλλογή δεδομένων για την αντιμετώπιση των σημαντικότερων αναγκών. Η συλλογή μπορεί να γίνει μέσω εσωτερικών δικτύων,

logs, εξειδικευμένων πηγών όπως το dark web, καθώς και από δημόσιες πηγές όπως ειδήσεις, blogs, και φόρουμ.
3. Επεξεργασία (Processing): Τα δεδομένα που συλλέγονται μετατρέπονται σε μορφή που να είναι εύχρηστη για τον οργανισμό. Αυτή η φάση περιλαμβάνει το φιλτράρισμα άχρηστων δεδομένων, τη δόμηση των πληροφοριών και τον εμπλουτισμό τους με συμφραζόμενα για την επόμενη φάση ανάλυσης.
4. Ανάλυση (Analysis): Σε αυτή τη φάση, οι αναλυτές ασφαλείας μετατρέπουν τις επεξεργασμένες πληροφορίες σε αξιοποιήσιμες πληροφορίες που μπορούν να επηρεάσουν τη λήψη αποφάσεων.
5. Διάδοση (Dissemination): Τα αποτελέσματα της ανάλυσης διανέμονται με τη μορφή αναφορών, εξασφαλίζοντας την ασφαλή διανομή και συνεργασία με εμπιστευτικότητα.
6. Ανατροφοδότηση (Feedback): Αφορά τη λήψη ανατροφοδότησης για την αποτελεσματικότητα των πληροφοριών από απειλές (threat intelligence) και την αναγνώριση ευκαιριών για συνεχή βελτίωση.

### 3.2.2 Είδη Cyber Threat Intelligence

Το Cyber Threat Intelligence διακρίνεται σε διαφορετικούς τύπους ανάλογα με τη φύση των πληροφοριών που παρέχει και τον σκοπό τους. Κάθε τύπος CTI εξυπηρετεί διαφορετικές ανάγκες μέσα σε έναν οργανισμό, όπως τη στρατηγική cybersecurity που ακολουθεί, την τεχνική ανάλυση και την ανταπόκριση σε πραγματικό χρόνο σε απειλές. Παρακάτω αναλύονται οι τέσσερις κύριοι τύποι CTI [36]:

1. Στρατηγικό CTI (Strategic CTI): Παρέχει μια υψηλού επιπέδου επισκόπηση του τοπίου των απειλών και συνοψίζει τις πιθανές κυβερνοεπιθέσεις και τις συνέπειές τους. Απευθύνεται κυρίως σε μη τεχνικά στελέχη και υπεύθυνους λήψης αποφάσεων, βοηθώντας τους να κατανοήσουν τις μακροπρόθεσμες απειλές για τον οργανισμό τους. Αν και δεν εμβαθύνει σε τεχνικές λεπτομέρειες, βοηθάει στη διαμόρφωση στρατηγικών αποφάσεων για την προστασία από μελλοντικές απειλές.

2. Τακτικό CTI (Tactical CTI): Επικεντρώνεται σε άμεσες και συγκεκριμένες πληροφορίες σχετικά με τρέχουσες και αναδυόμενες απειλές. Αυτός ο τύπος CTI παρέχει πληροφορίες για νέους τύπους κακόβουλου λογισμικού, μεθόδους επιθέσεων και συγκεκριμένους threat actors. Οι τακτικές πληροφορίες είναι απαραίτητες για τις καθημερινές επιχειρησιακές δραστηριότητες, καθώς βοηθούν τους οργανισμούς να ανταποκριθούν γρήγορα σε απειλές που βρίσκονται σε εξέλιξη ή αναμένονται στο εγγύς μέλλον. Οι αναλυτές ασφαλείας και οι ομάδες διαχείρισης απειλών αξιοποιούν αυτές τις πληροφορίες για την άμεση προστασία του οργανισμού.

3. Τεχνικό CTI (Technical CTI): Παρέχουν μια λεπτομερή τεχνική ανάλυση των απειλών. Περιλαμβάνουν πληροφορίες σχετικά με τα τεχνικά χαρακτηριστικά του κακόβουλου λογισμικού, τις ευπάθειες και τις μεθόδους επίθεσης. Οι τεχνικές πληροφορίες χρησιμοποιούνται από τεχνικές ομάδες ασφαλείας ενός οργανισμού για την ανίχνευση και την απόκριση σε απειλές. Βοηθούν στην κατανόηση των

επιτιθέμενων συστημάτων και παρέχουν λεπτομέρειες για το πώς να τα αντιμετωπιστούν τεχνικά. Οι δείκτες παραβίασης (Indicators of Compromise, IoCs)) όπως διευθύνσεις IP, hashes αρχείων και ονόματα domain, είναι συχνά μέρος αυτών των πληροφοριών.

4. Επιχειρησιακό CTI (Operational CTI): Παρέχουν πληροφορίες σε πραγματικό χρόνο για τις κυβερνοεπιθέσεις που βρίσκονται σε εξέλιξη και τα περιστατικά που αφορούν τον οργανισμό. Αυτές οι πληροφορίες συλλέγονται από διάφορες πηγές, όπως τα κοινωνικά μέσα, καταγραφές από λογισμικό antivirus και δεδομένα από προηγούμενα περιστατικά. Αυτός ο τύπος CTI βοηθά τους οργανισμούς να αντιδρούν άμεσα σε απειλές και να λαμβάνουν τα απαραίτητα μέτρα για να τις μετριάσουν.

Κάθε τύπος CTI παίζει καθοριστικό ρόλο στην προστασία ενός οργανισμού. Από τις στρατηγικές αποφάσεις μέχρι την άμεση αντίδραση σε απειλές, το CTI βοηθάει τους οργανισμούς να διατηρήσουν ένα ολοκληρωμένο και προσαρμοσμένο σχέδιο ασφαλείας, ώστε να μπορούν να αντιμετωπίζουν τις απειλές πιο αποτελεσματικά την ασφάλεια των δεδομένων και των υποδομών τους.

### 3.2.3 Πηγές

Το CTI στηρίζεται στη συλλογή δεδομένων από διάφορες πηγές, οι οποίες παρέχουν πολύτιμες πληροφορίες σχετικά με υπάρχουσες και αναδυόμενες κυβερνοαπειλές. Η διαδικασία συλλογής δεδομένων από αξιόπιστες πηγές είναι κρίσιμη, καθώς επηρεάζει άμεσα την αποτελεσματικότητα της ανάλυσης απειλών και την προσαρμογή των οργανισμών απέναντι σε πιθανές επιθέσεις.

- Open Source Intelligence (OSINT): Οι ανοιχτές πηγές πληροφοριών περιλαμβάνουν δεδομένα που είναι δημόσια διαθέσιμα και προέρχονται από διάφορες πηγές στο διαδίκτυο. Αυτές οι πληροφορίες είναι ελεύθερα προσβάσιμες και μπορούν να συλλεχθούν χωρίς την ανάγκη ειδικών δικαιωμάτων ή άδειας. Παραδείγματα:
    - Ενημερωτικές ιστοσελίδες: Αναφορές για κυβερνοεπιθέσεις, κυβερνοασφάλεια, και διαρροές δεδομένων.
    - Μέσα κοινωνικής δικτύωσης: Αναρτήσεις και συζητήσεις που μπορεί να υποδεικνύουν κυβερνοεπιθέσεις ή την ύπαρξη απειλών. (πχ. X (ex. Twitter) feeds)
    - Δημόσια φόρουμ: Συζητήσεις γύρω από ευπάθειες συστημάτων, κακόβουλο λογισμικό και νέες επιθέσεις.
    - Blogs και white papers: Δημοσιεύσεις από ειδικούς του κλάδου της κυβερνοασφάλειας που αναλύουν πρόσφατες τάσεις και τεχνικές επιθέσεων.

  Αν και οι ανοιχτές πηγές είναι εύκολα προσβάσιμες και συχνά παρέχουν χρήσιμες πληροφορίες, υπάρχει πάντα η ανάγκη για αξιολόγηση της ακρίβειας και της αξιοπιστίας τους, καθώς μπορεί να περιλαμβάνουν εσφαλμένες ή παραπλανητικές πληροφορίες.

- Closed Source Intelligence: Οι κλειστές πηγές αποτελούνται από ιδιωτικά δεδομένα που δεν είναι διαθέσιμα στο ευρύ κοινό και παρέχονται μόνο μέσω συνδρομητικών υπηρεσιών ή εξειδικευμένων προμηθευτών πληροφοριών ασφαλείας. Αυτές οι πηγές θεωρούνται πιο αξιόπιστες και ακριβείς, καθώς προέρχονται από επαγγελματίες της κυβερνοασφάλειας και εξειδικευμένες ομάδες. Παραδείγματα:
    - Ειδικές πλατφόρμες συλλογής πληροφοριών απειλών: Εταιρείες που ειδικεύονται στη συλλογή πληροφοριών από κυβερνοεγκληματικά δίκτυα και παρέχουν πληροφορίες για αναδυόμενες απειλές σε πραγματικό χρόνο. (πχ. OpenCTI)
    - Ιδιωτικά reports: Εκθέσεις από ιδιωτικές εταιρείες ή κυβερνητικούς οργανισμούς σχετικά με συγκεκριμένες επιθέσεις, απειλές ή ευπάθειες.
    - Ανταλλαγή πληροφοριών μεταξύ οργανισμών: Οι οργανισμοί ενδέχεται να ανταλλάσσουν μεταξύ τους κλειστές πληροφορίες για κυβερνοεπιθέσεις στο πλαίσιο συνεργασίας για την αντιμετώπιση απειλών.

  Η χρήση κλειστών πηγών προσφέρει σημαντικά πλεονεκτήματα στην ανάλυση CTI, καθώς οι πληροφορίες αυτές είναι πιο εξειδικευμένες και αξιόπιστες. Παρόλα αυτά, μπορεί να είναι δαπανηρές και να απαιτούν συνδρομές ή συνεργασίες με εξωτερικούς φορείς.

- Dark Web: Το Dark Web αποτελεί επίσης σημαντική πηγή πληροφοριών για την ανάλυση κυβερνοαπειλών. Αν και το περιεχόμενό του δεν είναι δημόσια προσβάσιμο, τα κυβερνοεγκληματικά δίκτυα συχνά χρησιμοποιούν το dark web για να προωθούν ή να ανταλλάσσουν πληροφορίες, κακόβουλο λογισμικό, ή ευαίσθητα δεδομένα που έχουν αποκτηθεί από παραβιάσεις. Παραδείγματα:
    - Φόρουμ και αγορές στο Dark Web: Συζητήσεις και συναλλαγές που σχετίζονται με παράνομες δραστηριότητες, όπως πώληση προσωπικών δεδομένων, κακόβουλου λογισμικού, ή εργαλεία hacking.
    - Αναφορές για επερχόμενες επιθέσεις: Ανακοινώσεις για μελλοντικές επιθέσεις και δραστηριότητες που μπορεί να θέσουν σε κίνδυνο οργανισμούς.

  Η συλλογή πληροφοριών από το dark web απαιτεί εξειδικευμένα εργαλεία και τεχνικές, καθώς και ιδιαίτερη προσοχή λόγω της φύσης του περιεχομένου. Παρά τον κίνδυνο, οι πληροφορίες αυτές μπορεί να αποδειχθούν πολύτιμες για την πρόβλεψη και αντιμετώπιση κυβερνοεπιθέσεων.

### 3.2.4 Εργαλεία

Τα εργαλεία Cyber Threat Intelligence (CTI) είναι απαραίτητα για τη συλλογή, ανάλυση και διαχείριση πληροφοριών απειλών, επιτρέποντας στους οργανισμούς να ανιχνεύουν, να παρακολουθούν και να ανταποκρίνονται σε απειλές κυβερνοασφάλειας σε πραγματικό χρόνο. Ακολουθεί μια πιο αναλυτική περιγραφή των κύριων εργαλείων CTI που χρησιμοποιούνται από τις ομάδες ασφαλείας:

1. Πλατφόρμες Συλλογής και Διαχείρισης CTI: Οι πλατφόρμες αυτές συνδυάζουν πληροφορίες από πολλαπλές πηγές, ανοιχτές και κλειστές, προσφέροντας μια ενοποιημένη άποψη του τοπίου απειλών. Τα δεδομένα που συλλέγονται ενσωματώνονται σε μια κεντρική πλατφόρμα, διευκολύνοντας την ανάλυση και την

εφαρμογή μέτρων προστασίας. Οι πληροφορίες που αντλούνται περιλαμβάνουν αναφορές για νέες ευπάθειες, επιθέσεις και μοτίβα απειλών. (π.χ. Recorded Future, ThreatConnect. [37] [38]

2. SIEM (Security Information and Event Management): Τα συστήματα SIEM είναι σχεδιασμένα για τη συλλογή, ανάλυση και συσχέτιση δεδομένων από διαφορετικά σημεία του δικτύου ενός οργανισμού. Χρησιμοποιούν τις πληροφορίες αυτές για να εντοπίσουν απειλές και παραβιάσεις σε πραγματικό χρόνο. Τα εργαλεία SIEM ανιχνεύουν ανωμαλίες και παρέχουν προειδοποιήσεις ασφαλείας, επιτρέποντας στους αναλυτές να απαντήσουν άμεσα σε ύποπτα περιστατικά. (π.χ. Splunk, IBM QRadar).[39],[40]

3. Εργαλεία Ανάλυσης Κακόβουλου Λογισμικού (Malware Analysis Tools): Τα εργαλεία αυτά βοηθούν στην ανάλυση κακόβουλου λογισμικού (malware) και παρέχουν λεπτομέρειες για τον τρόπο λειτουργίας του, επιτρέποντας στις ομάδες ασφαλείας να κατανοήσουν πώς λειτουργεί ένα malware και πώς να το αντιμετωπίσουν. (π.χ. Cuckoo Sandbox, VirusTotal).[41], [42]

4. TI Feed Aggregators (Αθροιστές Πηγών Πληροφοριών Απειλών): Αυτά τα εργαλεία παρέχουν πληροφορίες από πολλαπλές πηγές, επιτρέποντας στις ομάδες ασφαλείας να συλλέξουν πληροφορίες σε πραγματικό χρόνο για αναδυόμενες απειλές. Τα δεδομένα προέρχονται από διαφορετικές πηγές CTI, συγκεντρώνονται και παρουσιάζονται σε μια ενιαία μορφή για ευκολότερη κατανόηση και χρήση. (π.χ. AlienVault OTX, MISP). [43], [44]

5. Automated Threat Intelligence Platforms (Αυτοματοποιημένες Πλατφόρμες Πληροφοριών Απειλών): Αυτά τα εργαλεία χρησιμοποιούν αυτοματισμούς για την ανίχνευση απειλών και την ανταπόκριση σε πραγματικό χρόνο, εξοικονομώντας χρόνο στις ομάδες ασφαλείας και μειώνοντας τα ψευδώς θετικά αποτελέσματα. Χρησιμοποιούν τεχνικές τεχνητής νοημοσύνης και μηχανικής μάθησης για να αναλύσουν τα δεδομένα και να προσφέρουν προτάσεις σχετικά με τις απαραίτητες ενέργειες. (π.χ. Anomali ThreatStream).[45]

Η χρήση των κατάλληλων εργαλείων CTI επιτρέπει στους οργανισμούς να ενισχύσουν την άμυνά τους απέναντι σε κυβερνοαπειλές, βελτιώνοντας την ανίχνευση και την απόκριση σε επιθέσεις. Η επιλογή του σωστού εργαλείου εξαρτάται από τις ανάγκες του οργανισμού, το διαθέσιμο προσωπικό και την υποδομή ασφάλειας. Συνδυάζοντας πλατφόρμες CTI, συστήματα SIEM, και αυτοματοποιημένα εργαλεία ανάλυσης, οι οργανισμοί μπορούν να βελτιώσουν σημαντικά την ανθεκτικότητά τους απέναντι σε κυβερνοαπειλές.

### 3.2.5 Προκλήσεις

Το Cyber Threat Intelligence (CTI) προσφέρει απαραίτητη γνώση στους οργανισμούς για την αντιμετώπιση κυβερνοαπειλών, αλλά έχει σημαντικές προκλήσεις. Αυτές οι προκλήσεις

έχουν να κάνουν με τη διαχείριση του όγκου των δεδομένων, με την ακρίβεια των πληροφοριών, αλλά και την ευθυγράμμιση με τους κανονισμούς.

Ένα από τα σημαντικότερα ζητήματα που αντιμετωπίζει το CTI είναι η υπερφόρτωση δεδομένων. [46] Με τον τεράστιο όγκο πληροφοριών που παράγουν τα εργαλεία συλλογής, οι ομάδες ασφαλείας μπορεί να βρεθούν αντιμέτωπες με δυσκολίες στη διαχείριση των δεδομένων. Πολλές φορές, τα δεδομένα αυτά είναι αδόμητα και περιέχουν μη σχετικές πληροφορίες, καθιστώντας δύσκολο τον εντοπισμό των σημαντικών πληροφοριών. Ένα άλλο μεγάλο ζήτημα είναι τα ψευδώς θετικά (false positives) και ψευδώς αρνητικά (false negatives) αποτελέσματα. Τα ψευδώς θετικά μπορούν να οδηγήσουν σε σπατάλη πόρων και άσκοπες αντιδράσεις, ενώ τα ψευδώς αρνητικά μπορεί να αφήσουν έναν οργανισμό εκτεθειμένο σε απειλές.

Τέλος, η συμμόρφωση σε κανονισμούς αποτελεί μια από τις πιο σημαντικές προκλήσεις για τους οργανισμούς, ιδιαίτερα τα τελευταία χρόνια που είναι αυξημένη η αυστηρότητα των κανονιστικών πλαισίων, όπως είναι το GDPR (Γενικός Κανονισμός για την Προστασία Δεδομένων). Οι οργανισμοί που συλλέγουν και επεξεργάζονται πληροφορίες απειλών πρέπει να διαχειρίζονται σωστά τα προσωπικά δεδομένα και τις ευαίσθητες πληροφορίες, προκειμένου να προστατεύσουν με τους σχετικούς νόμους, γιατί διαφορετικά μπορεί να οδηγηθούν σε αυστηρές κυρώσεις, μεγάλα πρόστιμα και ζημία στη φήμη τους.[47]

## 3.3 Το Πλαίσιο STIX 2.1

### 3.3.1 Ορισμός και Σκοπός

Το STIX 2.1 (Structured Threat Information eXpression) [48] είναι ένα ανοιχτό πρότυπο για την ανταλλαγή πληροφοριών απειλών στον κυβερνοχώρο, σχεδιασμένο να βοηθάει τους οργανισμούς στην αναγνώριση, ανάλυση, και ανταπόκριση σε κυβερνοαπειλές. Το STIX αναπτύχθηκε από το MITRE και υποστηρίζεται από την OASIS (Organization for the Advancement of Structured Information Standards), προκειμένου να τυποποιήσει τον τρόπο με τον οποίο οι πληροφορίες απειλών κοινοποιούνται και επεξεργάζονται.

Το STIX προσφέρει μια κοινή γλώσσα που επιτρέπει στους οργανισμούς να ανταλλάσσουν CTI με έναν τρόπο που είναι κατανοητός τόσο από τα συστήματα όσο και από τους ανθρώπους της κυβερνοασφάλειας. Ο βασικός στόχος του είναι να διευκολύνει την ανταλλαγή δεδομένων κυβερνοασφάλειας μεταξύ διαφορετικών οργανισμών και εργαλείων, αυξάνοντας την αποτελεσματικότητα της ανίχνευσης και απόκρισης σε απειλές. Το STIX χρησιμοποιείται από οργανισμούς σε όλο τον κόσμο, όπως κρατικές υπηρεσίες, επιχειρήσεις, και ομάδες ασφαλείας, για να κατανοήσουν και να ανταποκριθούν καλύτερα στις κυβερνοεπιθέσεις. Παρέχει ένα κοινό πλαίσιο αναφοράς, διευκολύνοντας την επικοινωνία μεταξύ διαφορετικών φορέων και εξασφαλίζοντας ότι οι απειλές καταγράφονται και αξιοποιούνται με τυποποιημένο τρόπο.

Το STIX χρησιμοποιείται συχνά σε συνδυασμό με το TAXII (Trusted Automated eXchange of Indicator Information), ένα πρωτόκολλο που επιτρέπει την ασφαλή και αξιόπιστη ανταλλαγή CTI σε πραγματικό χρόνο μεταξύ οργανισμών. Το TAXII επιτρέπει στις πλατφόρμες CTI να μοιράζονται γρήγορα πληροφορίες χρησιμοποιώντας τα δομημένα δεδομένα STIX.

### 3.3.2 Δομή

Το STIX 2.1 είναι δομημένο με τρόπο που επιτρέπει την περιγραφή του CTI μέσω τυποποιημένων οντοτήτων και σχέσεων, τοποθετημένες μέσα σε ένα STIX bundle, που αποτελεί έναν σημαντικό μηχανισμό για την συλλογή πολλών αντικειμένων που σχετίζονται μεταξύ τους σε μια ενιαία μονάδα. Τα STIX Bundles χρησιμοποιούνται για να συγκεντρώσουν πολλά STIX Objects σε μια ενιαία μορφή, η οποία μπορεί να αποσταλεί και να επεξεργαστεί από διάφορα συστήματα. Κάθε bundle περιλαμβάνει τα σχετικά δεδομένα απειλών, επιτρέποντας την ανταλλαγή μιας ολοκληρωμένης εικόνας απειλών με ένα μόνο πακέτο.

Η βασική δομή του STIX βασίζεται σε διάφορα αντικείμενα δεδομένων που ορίζουν τις απειλές και τις σχετικές δραστηριότητες. Αυτά τα αντικείμενα διασυνδέονται για να δώσουν μια πλήρη εικόνα των επιθέσεων, των επιτιθέμενων και των στόχων. Τα κύρια αντικείμενα του STIX 2.1, γνωστά και ως STIX Domain Objects (SDOs) είναι τα παρακάτω:

- Indicators (Δείκτες): Σημάδια που υποδεικνύουν μια πιθανή κυβερνοαπειλή, όπως IP διευθύνσεις, domain names ή hashes αρχείων.
- Threat Actors (Επιτιθέμενοι Παράγοντες): Πληροφορίες σχετικά με τις ομάδες ή τους μεμονωμένους επιτιθέμενους που συμμετέχουν σε κακόβουλες δραστηριότητες.
- Campaigns (Εκστρατείες): Οργανωμένες και συντονισμένες επιθέσεις που εκτελούνται από επιτιθέμενους με συγκεκριμένους στόχους.
- Malware (Κακόβουλο Λογισμικό): Αναλυτικά στοιχεία για τα κακόβουλα λογισμικά που χρησιμοποιούνται σε κυβερνοεπιθέσεις.
- Courses of Action (Πλάνα Δράσης): Στρατηγικές και μέτρα που μπορούν να ληφθούν για την αντιμετώπιση των απειλών.

Οι οντότητες που συνδέουν μεταξύ τους τα SDOs και δημιουργούν σχέσεις μεταξύ τους, λέγονται STIX Relationship Objects (SROs) και είναι πολύ σημαντικά για την κατανόηση της συσχέτισης μεταξύ των αντικειμένων που βρίσκονται στο bundle. Βοηθάει να δοθεί μια πλήρης εικόνα της απειλής και των σχετικών με αυτή στοιχείων. Οι τύπου των SROs αναφέρονται παρακάτω:

- Relationship: Χρησιμοποιείται για τη δημιουργία άμεσων σχέσεων μεταξύ δύο αντικειμένων STIX. Για παράδειγμα, μπορεί να υπάρχει σχέση μεταξύ ενός Threat Actor και ενός Campaign που αυτός καθοδηγεί ή ανάμεσα σε έναν Indicator και ένα Malware που εντοπίστηκε.
- Sighting: Τα Sighting Objects χρησιμοποιούνται για να δείξουν ότι μια συγκεκριμένη δραστηριότητα ή απειλή έχει παρατηρηθεί στην πράξη. Ένα Sighting μπορεί να συνδέει έναν Indicator με μια πραγματική εμφάνιση ή δραστηριότητα που εντοπίστηκε σε ένα σύστημα. Αυτό βοηθά τις ομάδες ασφαλείας να δουν πότε και πού παρατηρήθηκε μια απειλή, διευκολύνοντας την ανάλυση και την αντίδραση.

Η δομή των αντικειμένων στο STIX βασίζεται σε JSON, καθιστώντας το εύκολο για τα συστήματα να επεξεργάζονται και να ανταλλάσσουν πληροφορίες. Κάθε αντικείμενο πρέπει να περιλαμβάνει βασικά πεδία όπως type (τύπος), id (μοναδικό αναγνωριστικό), created και modified (ημερομηνίες δημιουργίας και τροποποίησης), τα οποία διασφαλίζουν την ακριβή παρακολούθηση και ενημέρωση των πληροφοριών. Επιπλέον, τα αντικείμενα διαθέτουν

labels για κατηγοριοποίηση και descriptions που παρέχουν επιπλέον πληροφορίες σχετικά με αυτά.

### 3.3.3 Χρήση και Προκλήσεις

Το STIX έχει αναδειχθεί σε ένα από τα πιο σημαντικά πρότυπα στο Cybersecurity για την ανταλλαγή και τυποποίηση πληροφοριών στο πλαίσιο του CTI. Η δυνατότητά του να περιγράφει πλήρως τα χαρακτηριστικά μιας κυβερνοαπειλής το καθιστά ένα απαραίτητο εργαλείο για τους οργανισμούς που επιθυμούν να βελτιώσουν την ανίχνευση, πρόληψη και ανταπόκριση σε απειλές. Το STIX επιτρέπει στις ομάδες ασφαλείας να συλλέγουν, να οργανώνουν και να μοιράζονται δεδομένα απειλών με έναν τυποποιημένο και αυτοματοποιημένο τρόπο, κάτι που βελτιώνει την αξιοποιησιμότητα των πληροφοριών που διαμοιράζονται σε CTI πλατφόρμες και συστήματα SIEM.

Ωστόσο, παρά τα οφέλη του, υπάρχουν διάφορες προκλήσεις και περιορισμοί που αντιμετωπίζουν οι οργανισμοί κατά τη χρήση του STIX στο πλαίσιο του CTI. Μια από αυτές τις προκλήσεις αφορά την συμβατότητα των εργαλείων. Παρόλο που το STIX έχει σχεδιαστεί για να είναι διαλειτουργικό, δεν είναι όλα τα εργαλεία CTI ή τα συστήματα ασφαλείας πλήρως συμβατά με αυτό. Αυτό μπορεί να προκαλέσει προβλήματα όταν ένα περιβάλλον χρειάζεται διαφορετικά εργαλεία για να ασφαλιστεί. Επιπλέον, υπάρχουν ανησυχίες σχετικά με την ευαισθησία των δεδομένων που διαμοιράζονται μέσω STIX. Η ανταλλαγή ευαίσθητων πληροφοριών μεταξύ οργανισμών πρέπει να γίνεται με προσοχή, για να εξασφαλιστεί η συμμόρφωση με κανονισμούς.

Τέλος, η χρήση του STIX απαιτεί εκπαίδευση και εξειδίκευση, καθώς η πλήρης αξιοποίηση των δυνατοτήτων του απαιτεί από τις ομάδες ασφαλείας να είναι εξοικειωμένες με την τεχνική του φύση και τα μοτίβα ανάλυσης απειλών που αυτό προσφέρει. Είναι σημαντικό, για την αυτοματοποίηση και την κλιμάκωση της δημιουργίας των STIX bundles, να υιοθετηθούν νέες τεχνικές που δεν βασίζονται στον ανθρώπινο παράγοντα. Αυτό θα οδηγήσει στην αύξηση της ταχύτητας και της ακρίβειας, επιτρέποντας στις ομάδες ασφαλείας να ανταποκρίνονται πιο γρήγορα στις απειλές και να διασφαλίζουν τη συνεχή προστασία των συστημάτων τους.

# 4. Χρήση Τεχνητής Νοημοσύνης για την Εξαγωγή Πληροφοριών Κυβερνοαπειλών

## 4.1 Κίνητρο εργασίας

Ο κλάδος του Cyber Threat Intelligence βρίσκεται αντιμέτωπος με τη συνεχή ανάγκη για νέες και πιο αποδοτικές τεχνικές, καθώς καλείται να διαχειριστεί έναν τεράστιο όγκο νέας πληροφορίας, που συχνά είναι μη δομημένη, και να εντοπίσει εξελιγμένες απειλές που εμφανίζονται καθημερινά από πολλαπλές και ποικίλες πηγές. Οι πληροφορίες που αφορούν νέες απειλές και ευάλωτα σημεία δημοσιεύονται πρώτα σε πηγές όπως το dark web, το X (Twitter), τα μέσα κοινωνικής δικτύωσης και διάφορα hacking forums, πριν φτάσουν στα επίσημα κανάλια επικοινωνίας, στα οποία μάλιστα γίνονται γνωστές συνήθως αφού πρώτα επηρεάσουν αρνητικά κάποιον οργανισμό.

Η εξαγωγή πληροφοριών από μη δομημένες πηγές είναι καθοριστική για την κυβερνοασφάλεια, καθώς περιέχουν πολύτιμα δεδομένα για την έγκαιρη αναγνώριση και αντιμετώπιση εξελισσόμενων απειλών. Ο έγκαιρος εντοπισμός και αξιολόγηση τέτοιων πληροφοριών μπορεί να σώσει τους οργανισμούς από σοβαρές συνέπειες, όπως είναι η διαρροή δεδομένων, η καταστροφή συστημάτων και οι οικονομικές απώλειες. Η αναγνώριση μιας νέας απειλής ή ευπάθειας σε πρώιμο στάδιο επιτρέπει στους επαγγελματίες της κυβερνοασφάλειας να λάβουν άμεσα προληπτικά μέτρα, να ενισχύσουν τα συστήματά τους και να προστατεύσουν τις υποδομές τους προτού γίνει μια μεγάλης κλίμακας επίθεση.

Η πληροφορίες κυβερνοαπειλών είναι σημαντικό να αναπαρίσταται σε τυποποιημένη μορφή, ώστε να γίνεται κατανοητή από τους επαγγελματίες στον τομέα της κυβερνοασφάλειας, αλλά και για να μπορεί να αξιοποιηθεί από συστήματα ανάλυσης δεδομένων μαζικά. Το STIX επιτρέπει την τυποποίηση και την αυτοματοποίηση της ανταλλαγής πληροφοριών CTI, καθιστώντας την πιο προσβάσιμη και κατανοητή. Έτσι, τόσο οι αναλυτές κυβερνοασφάλειας όσο και τα αυτόματα συστήματα ανάλυσης δεδομένων μπορούν να αξιοποιήσουν άμεσα αυτήν τη γνώση, ενισχύοντας την ικανότητα των οργανισμών να αντιδρούν γρήγορα και αποτελεσματικά στις αναδυόμενες απειλές.

Η παραδοσιακή προσέγγιση για την εξαγωγή πληροφοριών κυβερνοαπειλών περιέχει εργαλεία που βασίζονταν κυρίως σε συνδυασμούς τεχνικών αναγνώρισης μοτίβων και κανόνων, καθώς και σε μη αυτόματη ανάλυση από ειδικούς κυβερνοασφάλειας. Αυτά τα εργαλεία χρησιμοποιούνταν για την ανάλυση logs, αναφορών, και αρχείων καταγραφής δικτύων, ενώ η ταξινόμηση απειλών γινόταν με τη βοήθεια στατικών κανόνων και προδιαγραφών. Παρά το γεγονός ότι αυτές οι μέθοδοι έχουν αποδειχθεί αποτελεσματικές σε κάποιο βαθμό, έχουν περιορισμούς όσον αφορά την κλιμάκωση και την ταχύτητα ανάλυσης, και δεν έχουν την δυνατότητα να εντοπίσουν σύνθετες και νέες απειλές που εξελίσσονται ραγδαία σε πραγματικό χρόνο.

Η εξάρτηση από την ανθρώπινη ανάλυση και στατικά μοντέλα συνεπάγεται ότι χρειάζεται χρόνος για την αναγνώριση και ανάλυση των κυβερνοαπειλών από τους οργανισμούς, με αποτέλεσμα συχνά να μην μπορούν να ανταποκριθούν γρήγορα σε νέες και άγνωστες απειλές. Επιπλέον, τα παραδοσιακά εργαλεία που αναφέραμε είναι πολύ περιορισμένα από την αδυναμία τους να διαχειριστούν τεράστιους όγκους μη δομημένων δεδομένων με ακατέργαστο κείμενο.

Με την εξέλιξη της τεχνητής νοημοσύνης και των τεχνικών μηχανικής μάθησης, η διαδικασία εξαγωγής πληροφοριών CTI έχει αλλάξει ριζικά. Αντί για στατικά και περιορισμένα εργαλεία, πλέον χρησιμοποιούνται αυτόματα συστήματα ανάλυσης που βασίζονται στο AI, τα οποία είναι σε θέση να αναλύουν τεράστιες ποσότητες δεδομένων σε πραγματικό χρόνο και να αναγνωρίζουν περίπλοκα μοτίβα απειλών. Η τεχνητή νοημοσύνη έχει τη δυνατότητα να μαθαίνει δεδομένα που ανανεώνονται συνεχώς, αναγνωρίζοντας νέες απειλές χωρίς να απαιτείται προγραμματισμός εκ των προτέρων πάνω σε αυτά.

Τα μοντέλα τεχνητής νοημοσύνης, όπως τα Μεγάλα Γλωσσικά Μοντέλα (LLMs), μπορούν να επεξεργάζονται μη δομημένα δεδομένα, να εντοπίζουν συσχετισμούς και να παρέχουν στους επαγγελματίες της κυβερνοασφάλειας αξιοποιήσιμες πληροφορίες με πολύ μεγαλύτερη ταχύτητα και ακρίβεια σε σύγκριση με τα παραδοσιακά εργαλεία. Η ικανότητα αυτών των μοντέλων να προσαρμόζονται σε νέα δεδομένα και να αναγνωρίζουν εξελιγμένες τακτικές κυβερνοεπιθέσεων καθιστά την τεχνητή νοημοσύνη ένα αναπόσπαστο κομμάτι της σύγχρονης ανάλυσης CTI.

Λαμβάνοντας υπόψη τις προκλήσεις που αναφέραμε παραπάνω, αναπτύξαμε μια εφαρμογή που θα αξιοποιεί τις δυνατότητες των LLMs με σκοπό την εξαγωγή δομημένης πληροφορίας από ακατέργαστα κείμενα. Η εφαρμογή αυτή θα λειτουργεί σε τοπικά περιβάλλοντα (local), εξασφαλίζοντας την ασφάλεια και την ιδιωτικότητα των δεδομένων, ιδιαίτερα σε περιπτώσεις που οργανισμοί χειρίζονται ευαίσθητα δεδομένα, μειώνοντας παράλληλα την εξάρτηση από υπηρεσίες cloud που μπορεί να εκθέσουν τα δεδομένα σε κινδύνους ή καθυστερήσεις.

Η εφαρμογή αυτή θα λαμβάνει μη δομημένα δεδομένα Cyber Threat Intelligence (CTI) και θα επιστρέφει την πιο σημαντική πληροφορία σε STIX 2.1 bundles, που είναι μια δομή κατανοητή τόσο από επαγγελματίες κυβερνοασφάλειας όσο και από συστήματα ανάλυσης δεδομένων. Με αυτόν τον τρόπο, θα παρέχουμε μια αυτοματοποιημένη λύση για την ανάλυση και επεξεργασία μεγάλου όγκου δεδομένων, διευκολύνοντας την ταχύτερη και πιο ακριβή ανταπόκριση στις αναδυόμενες κυβερνοαπειλές.

Επιπλέον, μελετήθηκαν διεξοδικά οι τελευταίες εξελίξεις στο state-of-the-art της εξαγωγής CTI και χρησιμοποιήθηκαν σύγχρονες τεχνικές, στις οποίες θα εμβαθύνουμε παρακάτω.

## 4.2 Σχετικές εργασίες και Προηγούμενες Μελέτες

Η χρήση προεκπαιδευμένων μοντέλων, όπως το BERT, έχει ήδη δείξει σημαντική επιτυχία στην εξαγωγή CTI από κείμενα. Το BERT, με την αμφίδρομη εκπαίδευσή του, έχει τη δυνατότητα να κατανοεί τη δομή της γλώσσας και να αποδίδει ακριβέστερα νοήματα μέσα από μη δομημένα δεδομένα, κάτι που το καθιστά ιδανικό για εξειδικευμένα κείμενα στον τομέα της κυβερνοασφάλειας [49], [50], [51]. Παραλλαγές του, όπως τα SciBERT, TIM και CyBERT, χρησιμοποιούνται για τη βελτιστοποίηση της απόδοσης του BERT σε συγκεκριμένες εργασίες CTI μέσω της τεχνικής του fine-tuning [51], [52].

Επιπλέον, με την ανάπτυξη του RoBERTa, το οποίο επεκτείνει τις δυνατότητες του BERT, έχει καταστεί δυνατό να αυξηθεί η ακρίβεια στις εργασίες εξαγωγής απειλών, χρησιμοποιώντας μεγαλύτερα datasets και εκτεταμένη εκπαίδευση [53], [54], [57]. Ένα χαρακτηριστικό παράδειγμα είναι το SecureBERT, το οποίο, βασισμένο στο RoBERTa, έχει βελτιστοποιηθεί για την κυβερνοασφάλεια και μπορεί να εντοπίσει απειλές με μεγαλύτερη

ακρίβεια, κάνοντας χρήση μεγαλύτερων δεδομένων εκπαίδευσης και εξαλείφοντας την ανάγκη για το "Next Sentence Prediction", ένα χαρακτηριστικό που περιόριζε το αρχικό BERT [53]. Επιπλέον, πλατφόρμες όπως το LADDER και το TTPHunter έχουν αναδείξει τον συνδυασμό αυτών των μοντέλων για την εξαγωγή πληροφορίας TTPs από μη δομημένα δεδομένα [54], [57].

Παρά τις επιτυχίες των μοντέλων αυτών, υπάρχουν περιορισμοί που σχετίζονται με την ικανότητά τους να ανταποκρίνονται σε νέες απειλές σε πραγματικό χρόνο. Για παράδειγμα, τα encoder-only μοντέλα, όπως το BERT και το RoBERTa, εξακολουθούν να βασίζονται σε στατικές προεκπαιδευμένες γνώσεις και αδυνατούν να παράγουν πιο λεπτομερή και εξειδικευμένη πληροφορία σχετικά με μια επίθεση. Αυτοί οι περιορισμοί οδηγούν την έρευνα στη διερεύνηση πιο εξελιγμένων μοντέλων, όπως τα decoder-only μοντέλα [53], [55], [56].

Τα decoder-only LLMs έχουν αρχίσει να χρησιμοποιούνται ευρέως για την ανάλυση Cyber Threat Intelligence (CTI), προσφέροντας νέες δυνατότητες στην εξαγωγή χρήσιμων πληροφοριών από μη δομημένα δεδομένα. Έχουν την ικανότητά να παράγουν ανθρώπινης ποιότητας κείμενα, αλλάζοντας  ριζικά την προσέγγιση της ανάλυσης CTI. Έχουν δημοσιευτεί διάφορες μελέτες που αξιοποιούν τις δυνατότητες των decoder-only LLMs, όπως είναι το aCTIon, ένα εργαλείο που χρησιμοποιεί το GPT-3.5 για να μετατρέπει αναφορές κυβερνοαπειλών σε δομημένη μορφή STIX, εστιάζοντας μόνο σε συγκεκριμένες οντότητες, όπως Malware, Threat Actor, και Attack Pattern [56]. Επίσης, σε μια πρόσφατη μελέτη προτείνεται η χρήση LLMs σε συνδυασμό με γραφήματα γνώσης (knowledge graphs) για την αυτόματη εξαγωγή πληροφοριών, βελτιστοποιώντας την ανάλυση μέσω τεχνικών όπως το fine-tuning και το guidance framework. [58]

Μια αρκετά ενδιαφέρουσα έρευνα εστιάζει στην σύγκριση των SecureBERT, GPT-3.5 και GPT-3.5 με RAG αρχιτεκτονικών πάνω στο ζήτημα της κατανόησης και σύνοψης των Tactics, Techniques, and Procedures (TTPs) στο πλαίσιο του MITRE ATT&CK framework. Η μελέτη έδειξε ότι το decoder-only LLM χωρίς RAG υπερτερούσε του SecureBERT σε Recall, αλλά εμφάνιζε χαμηλότερη Precision, γεγονός που συνδέεται με την τάση των LLMs να δημιουργούν ψευδαισθήσεις. Ωστόσο, όταν το LLM συνδυάστηκε με RAG, η απόδοσή του βελτιώθηκε σημαντικά, επιτυγχάνοντας ανώτερα αποτελέσματα στο F1 score.

Παρά την σημαντική πρόοδο στη χρήση LLMs για την εξαγωγή και ανάλυση Cyber Threat Intelligence (CTI), εξακολουθεί να υπάρχει ένα σημαντικό κενό στην έρευνα: η έλλειψη μιας πλήρως τοπικής (local) εφαρμογής που να μπορεί να παράγει ολοκληρωμένα STIX bundles από μη δομημένα δεδομένα. Οι περισσότερες υπάρχουσες λύσεις είτε βασίζονται σε online μοντέλα και υποδομές, θέτοντας ζητήματα απορρήτου και ασφαλείας, είτε επικεντρώνονται στην εξαγωγή περιορισμένων οντοτήτων, όπως Malware ή TTPs, χωρίς να παράγουν ολοκληρωμένες δομημένες πληροφορίες για όλες τις οντότητες που καλύπτει το STIX πρότυπο. Επιπλέον, οι υπάρχουσες λύσεις συχνά απαιτούν σημαντική προσαρμογή και fine-tuning των LLMs, γεγονός που αυξάνει το κόστος και την πολυπλοκότητα της εφαρμογής τους. Έτσι, υπάρχει μια σαφής ανάγκη για την ανάπτυξη μιας τοπικής λύσης που όχι μόνο διασφαλίζει την ιδιωτικότητα των δεδομένων, αλλά και παρέχει μια ολοκληρωμένη παραγωγή STIX bundles, διευκολύνοντας την απρόσκοπτη εξαγωγή πληροφοριών και την άμεση ανάλυση τους για έγκαιρη πρόληψη και ασφάλιση των συστημάτων.

## 4.3 Συνεισφορά της παρούσας εργασίας

Σε αυτή τη διατριβή, προτείνεται η χρήση ενός τοπικού (local) LLM σε συνδυασμό με την τεχνολογία Retrieval-Augmented Generation (RAG) για την εξαγωγή και δομή πληροφοριών CTI σε μορφή STIX bundles από ακατέργαστο κείμενο CTI. Η RAG τεχνολογία έχει στόχο την ενίσχυση της απόδοσης των LLMs, καθώς παρέχει στο μοντέλο οδηγίες από το επίσημο STIX 2.1 documentation σχετικές με τις απαιτήσεις του αιτήματος του χρήστη. Η εφαρμογή μας καθιστά εφικτή την ασφαλή και ακριβή παραγωγή πλήρως λειτουργικών και δομημένων STIX bundles.

Οι συνεισφορές της παρούσας διατριβής σημειώνονται συνοπτικά παρακάτω:

- Μια ολοκληρωμένη αυτοματοποιημένη διαδικασία για την εξαγωγή CTI πληροφορίας από ακατέργαστο κείμενο και παραγωγή σωστά δομημένων STIX 2.1 bundles σε JSON, με τη χρήση open-source local Large Language Models.
- Σύγκριση συνολικά έξι διαφορετικών open-source local LLMs στην επίδοση τους πάνω στη συγκεκριμένη εργασία και καταγραφή των αποτελεσμάτων τους.
- Βελτιστοποίηση της επίδοσης των LLMs στη συγκεκριμένη εργασία με χρήση τεχνικών prompting και few-shot learning, και σύγκριση των αποτελεσμάτων τους.
- Ένας ολοκληρωμένος RAG Agent ο οποίος ανακτά σχετική πληροφορία σύμφωνα με τις ανάγκες του εκάστοτε παραδείγματος από το documentation του STIX 2.1, βελτιστοποιώντας τις επιδόσεις των LLMs και ανάδειξη του πιο αποδοτικού LLM για την συγκεκριμένη εργασία.

## 4.4 Περιορισμοί

Για την υλοποίηση του συστήματος είχαμε περιορισμένες δυνατότητες, οι οποίες επηρεάζουν σημαντικά την επίδοση του συστήματος. Αρχικά, υπάρχουν περιορισμένοι υπολογιστικοί πόροι, και λόγω των δυνατοτήτων του υλικού, είχαμε τη δυνατότητα να τρέξουμε γλωσσικά μοντέλα μέχρι και 9 δισεκατομμυρίων παραμέτρων, αποκλείοντας τη χρήση μεγαλύτερων και πιο ισχυρών μοντέλων. Αυτό είχε συνέπειες στην ακρίβεια και ποιότητα των αποτελεσμάτων μας. Αν και τα μικρότερα μοντέλα μπορούν να είναι αρκετά αποδοτικά, τα μεγαλύτερα γλωσσικά μοντέλα παρέχουν συνήθως βελτιωμένη απόδοση σε σύνθετες εργασίες, όπως είναι και η ανάλυση μη δομημένων πληροφοριών απειλών.

Επιπλέον, ένας άλλος περιορισμός αφορούσε την αδυναμία μας να ακολουθήσουμε την τεχνική του fine-tuning, καθώς υπάρχει μεγάλη έλλειψη στα δημόσια δεδομένα που χρειαζόμαστε. Για να επιτύχουμε υψηλής ακρίβειας αποτελέσματα, ήταν απαραίτητα ειδικά σύνολα δεδομένων που να περιλαμβάνουν CTI reports και τα ακριβώς αντίστοιχα STIX bundles τους. Ωστόσο, τέτοια σύνολα δεδομένων δεν είναι δημόσια διαθέσιμα, περιορίζοντας έτσι τις δυνατότητες που μπορεί να επιτύχει το σύστημα μας.

# 5. Πειραματική Διάταξη

## 5.1 Περιγραφή

Για το πειραματισμό μας έχουν υλοποιηθεί δυο παρόμοια συστήματα, το ένα περιλαμβάνει την υλοποίηση της τεχνολογίας του RAG, ενώ το άλλο λειτουργεί χωρίς το RAG με σκοπό να αναδείξει ποια από τα υποψήφια LLMs έχουν καλύτερη επίδοση στο ζητούμενο και την βελτίωση των αποτελεσμάτων τους με τεχνικές όπως το prompt engineering και το few-shot learning.

## 5.2 Εργαλεία και Frameworks

### 5.2.1 Ollama

Το Ollama είναι ένα open-source framework που έχει σχεδιαστεί για να διευκολύνει την εγκατάσταση Large Language Models σε τοπικά περιβάλλοντα. Στόχος του είναι να απλοποιήσει την εκτέλεση και διαχείριση των open-source μοντέλων, προσφέροντας μια εύκολη εμπειρία χρήστη σε διάφορα λειτουργικά συστήματα. Το Ollama παρέχει εργαλεία για τη βελτιστοποίηση της απόδοσης και της κλιμάκωσης των μοντέλων, επιτρέποντας στους χρήστες να πειραματίζονται, χωρίς την ανάγκη για σύνδεση σε συνδρομητικές ή cloud υπηρεσίες. Με αυτόν τον τρόπο, εξασφαλίζει μεγαλύτερη ασφάλεια και έλεγχο των δεδομένων, ενώ ταυτόχρονα μειώνει το κόστος λειτουργίας και τις καθυστερήσεις που συνήθως παρατηρούνται σε απομακρυσμένες υποδομές.

Ακολουθεί μια σύντομη περιγραφή των LLMs που θα μελετηθούν, όλα μεταξύ 7B-9B μεγέθη παραμέτρων, λόγω περιορισμένης υπολογιστικής ισχύος:

1. Llama3.1 (7B): Το μοντέλο αυτό από την Meta διαθέτει ένα πολύ διευρυμένο μήκος πλαισίου (context window) που φτάνει τους 128K tokens, γεγονός που του επιτρέπει να επεξεργάζεται πολύ μεγάλα κείμενα με μεγάλη ακρίβεια και αποδοτικότητα. Είναι σχεδιασμένο για προηγμένες εργασίες κατανόησης και συλλογισμού, ενσωματώνοντας βελτιστοποιήσεις για ταχύτερη επεξεργασία και υψηλή απόδοση.

2. Gemma2 (9B): Το μοντέλο από την Google υποστηρίζει context window 8192 tokens και έχει εκπαιδευτεί για αποδοτική επεξεργασία κειμένου και είναι κατάλληλο για χρήσεις όπως η δημιουργία κειμένου, τα chatbots, και η σύνοψη κειμένου. Επιπλέον, χρησιμοποιεί τεχνικές) για να βελτιώσει την ακρίβεια και τη φυσικότητα των απαντήσεών του.

3. Mistral (7B): Το μοντέλο από την Mistral AI έχει context window 4K tokens. Είναι βελτιστοποιημένο για ταχύτητα και αποδοτικότητα, κάνοντάς το κατάλληλο για περιβάλλοντα με περιορισμένους πόρους και εφαρμογές πραγματικού χρόνου, όπως chatbots και ανακτήσεις πληροφοριών.

4. Llava (7B): Το μοντέλο από την Microsoft διαθέτει context window των 4K tokens και είναι σχεδιασμένο για εφαρμογές που συνδυάζουν πολυμέσα, όπως περιγραφές εικόνων και ανάλυση πολυτροπικών δεδομένων. Οι πολυτροπικές δυνατότητές του το

καθιστούν εξαιρετικά χρήσιμο σε πλατφόρμες που απαιτούν συνδυασμένη ανάλυση οπτικών και γλωσσικών πληροφοριών.

Τα μοντέλα που αναφέραμε παραπάνω, αλλά και αυτά που ακολουθούν ανήκουν στην κατηγορία των Instruct μοντέλων. Τα Instruct μοντέλα είναι γλωσσικά μοντέλα που έχουν εκπαιδευτεί ειδικά για να ακολουθούν οδηγίες και να εκτελούν εργασίες βάσει αυτών. Ωστόσο, τα μοντέλα που ακολουθούν έχουν υποβληθεί σε περαιτέρω fine-tuning ειδικά για να ακολουθούν οδηγίες, με αποτέλεσμα να έχουν βελτιστοποιημένη κατανόηση στην εκτέλεση εντολών, καταφέρνοντας να δίνουν απαντήσεις πιο σχετικές και ακριβείς. Τα (περισσότερο) Instruct μοντέλα που θα μελετήσουμε είναι τα παρακάτω:

5. Dolphin-llama3 (7B): Το μοντέλο αυτό είναι μια ειδική έκδοση του Llama3, προσαρμοσμένο για αυξημένη υπακοή σε οδηγίες, απόδοση και ακρίβεια. Υποστηρίζει context window μέχρι 4K tokens.

6. Dolphin-mistral (7B): Το μοντέλο αυτό συνδυάζει βελτιώσεις του Mistral με δυνατότητες των Instruct μοντέλων. Υποστηρίζει 4K tokens για context window.

Τέλος, για την δημιουργία των embeddings θα χρησιμοποιείται το παρακάτω μοντέλο από το Ollama:

7. mxbai-embed-large: Αυτό το embedding μοντέλο υποστηρίζει έως 512 tokes ανά chunk. Παράγει embeddings υψηλής ποιότητας που αποτυπώνουν τη σημασιολογική σημασία του κειμένου. Το μοντέλο είναι encoder-only, και χρησιμοποιείται ευρέως για αναζήτηση πληροφοριών, σύγκριση κειμένων και ομαδοποίηση δεδομένων, καθώς οι αναπαραστάσεις που δημιουργεί είναι αποδοτικές για εντοπισμό συνάφειας και ανάλυση ομοιοτήτων σε μεγάλες κλίμακες δεδομένων

### 5.2.2 Chroma DB

Το Chroma DB είναι ένα open-source σύστημα βάσης δεδομένων ειδικά σχεδιασμένο για την αποθήκευση και διαχείριση διανυσματικών δεδομένων (vector data), τα οποία χρησιμοποιούνται συχνά σε εφαρμογές μηχανικής μάθησης και τεχνητής νοημοσύνης. Στόχος του είναι να προσφέρει έναν απλό και αποδοτικό τρόπο για την οργάνωση και ανάκτηση δεδομένων, διασφαλίζοντας παράλληλα υψηλή απόδοση και συμβατότητα με διαφορετικά λειτουργικά συστήματα. Το Chroma DB επιτρέπει την εκτέλεση σύνθετων ερωτημάτων και αναζητήσεων σε πραγματικό χρόνο, καθιστώντας το ιδανικό για τοπικές υποδομές όπου είναι κρίσιμη η ασφάλεια των δεδομένων. Υποστηρίζει επίσης ευέλικτη προσαρμογή και επεκτασιμότητα, δίνοντας στους προγραμματιστές τη δυνατότητα να ενσωματώσουν τις λειτουργίες του σε διάφορες εφαρμογές χωρίς την ανάγκη για εξωτερικές ή cloud λύσεις, προσφέροντας έτσι αυτονομία και μειωμένο κόστος λειτουργίας.

### 5.2.3 OpenAI Client

Ο OpenAI client χρησιμοποιείται για τη σύνδεση και την εκτέλεση των μοντέλων μέσω του τοπικού API που έχει εγκατασταθεί στον server. Η χρήση του συγκεκριμένου client επιτρέπει την επικοινωνία με τοπικά εγκατεστημένα μοντέλα, παρέχοντας τη δυνατότητα διαχείρισης αιτημάτων και επεξεργασίας δεδομένων σε πραγματικό χρόνο. Αυτό εξασφαλίζει την αυτονομία του συστήματος, αποφεύγοντας την ανάγκη για εξωτερικές υπηρεσίες και διατηρώντας την ασφάλεια και τον έλεγχο των δεδομένων. Η παραμετροποίηση του client γίνεται μέσω API key, διασφαλίζοντας έτσι την αυθεντικοποίηση και εξουσιοδότηση των αιτημάτων.

## 5.3 Το pipeline του RAG Agent

### 5.3.1 Περιγραφή

Το RAG είναι μια μέθοδος που συνδυάζει την ανάκτηση πληροφοριών με τη δημιουργία κειμένου από τα LLMs. Στόχος του είναι να βελτιώσει την ακρίβεια και τη συνάφεια των απαντήσεων που παράγονται από το μοντέλο, ενσωματώνοντας εξωτερική γνώση σε πραγματικό χρόνο. Ενισχύει τα LLMs, που βασίζονται αποκλειστικά στη γνώση που έχουν μάθει κατά την εκπαίδευση τους, καθώς επιτρέπει στο σύστημα να αντλεί δεδομένα δυναμικά, να τα επεξεργάζεται και να τα χρησιμοποιεί στην παραγωγή των απαντήσεων.

Το σύστημα που σχεδιάστηκε για αυτή τη διατριβή για την αλληλεπίδραση με αυτό χρησιμοποιεί FastAPI. Ο χρήστης που επιθυμεί να εξάγει από ένα κείμενο το αντίστοιχο STIX 2.1 bundle, εκτελεί την εντολή «curl -X POST "http://127.0.0.1:8000/process/" -F “pdf_file=@Report.pdf”» , όπου το Report.pdf είναι το αρχείο με το κείμενο. Τότε, αρχικοποιούνται στο σύστημα δύο σημαντικές τιμές, η ***specific_question***, η οποία περιέχει μια συγκεκριμένη ερώτηση που θα τεθεί στο LLM, και το ***system_message***, το οποίο είναι ένα προκαθορισμένο μήνυμα που παρέχεται στο LLM για να καθορίσει το πλαίσιο, τους κανόνες και τις οδηγίες που θα ακολουθήσει κατά τη διάρκεια επεξεργασίας του αιτήματος.

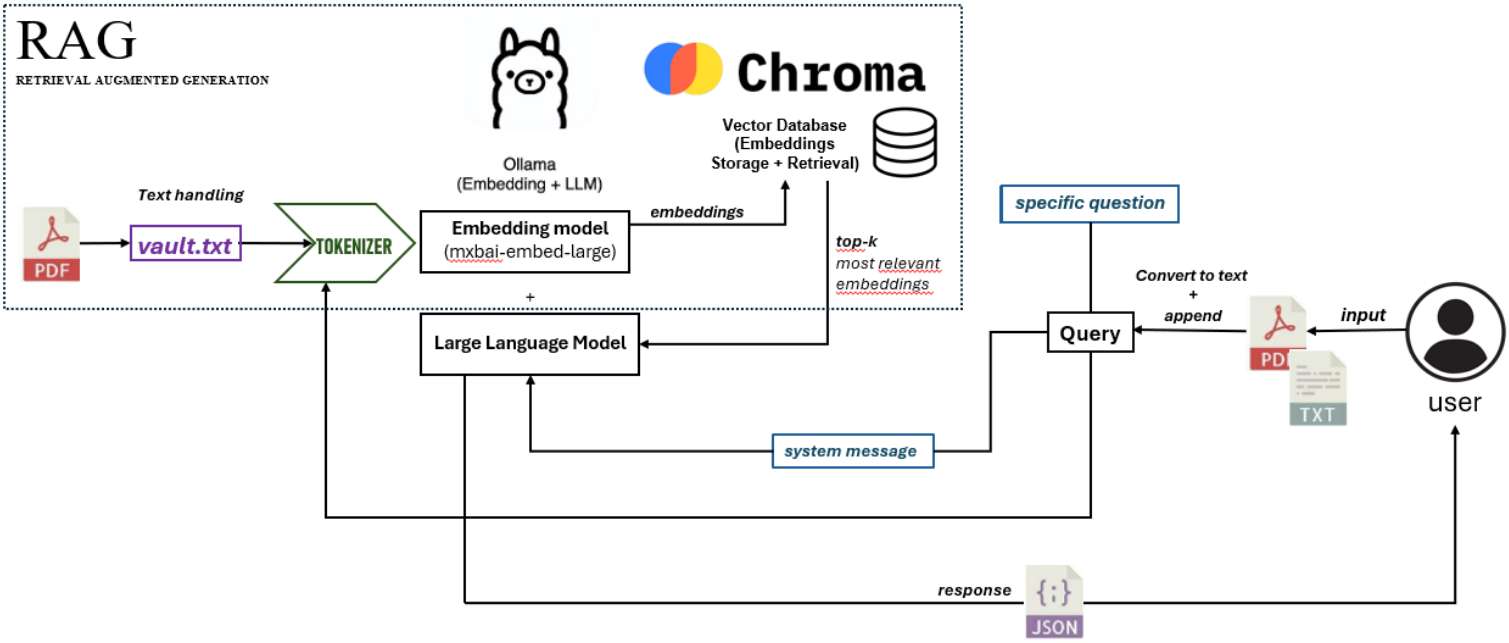
### 5.3.2 Λειτουργία

Η διαδικασία λειτουργίας του συστήματος με το RAG ακολουθεί τα παρακάτω βήματα:

- Ανάκτηση Πληροφορίας: Όταν το σύστημα λαμβάνει το αίτημα από τον χρήστη, το πρώτο βήμα είναι να αναζητήσει σχετικές πληροφορίες από μια εξωτερική βάση δεδομένων. Στην περίπτωση του πειράματός μας, χρησιμοποιείται το Chroma DB για την αποθήκευση και ανάκτηση αυτών των πληροφοριών. Τα δέδομενα αυτά είναι αποθηκευμένα σε μια μόνιμη διανυσματική βάση δεδομένων (persistent vector database) σε μορφή embeddings. Αν αυτά τα embeddings δεν υπάρχουν ήδη στην βάση, το σύστημα τα δημιουργεί και τα αποθηκεύει στην πρώτη του εκτέλεση.
- Δημιουργία Embeddings: Οι διανυσματικές αναπαραστάσεις (embeddings) δημιουργούνται χρησιμοποιώντας ένα embedding μοντέλο από το Ollama, το «mxbai-embed-large» που αναλύθηκε παραπάνω. Φορτώνεται το αρχείο vault.txt, το οποίο περιέχει το κείμενο με την χρήσιμη πληροφορία, και για αρχή περνάει από τον tokenizer του μοντέλου, ο οποίος το διασπά στα αντίστοιχα tokens, τα οποία μετατρέπονται σε αριθμητικές αναπαραστάσεις, προκειμένου να είναι κατανοητό από

το νευρωνικό δίκτυο του μοντέλου. Μόλις το κείμενο μετατραπεί σε tokens, το μοντέλο «mxbai-embed-large» χρησιμοποιεί την αρχιτεκτονική του για να δημιουργήσει embeddings για κάθε token. Στη συνέχεια, αυτά τα embeddings αποθηκεύονται στη vector database του Chroma DB, δημιουργώντας την βάση δεδομένων για την οποία μιλήσαμε στη προηγούμενη παράγρφο, από την οποία το σύστημα μπορεί να αναζητήσει πληροφορίες.

- Αναζήτηση Πληροφορίας: Όταν ο χρήστης υποβάλλει το αίτημα στο σύστημα, αυτό παράγει νέα embeddings για το αίτημα χρησιμοποιώντας τον ίδιο tokenizer και το ίδιο μοντέλο. Αυτά τα νέα embeddings συγκρίνονται με αυτά στην vector database, μέσω similarity check, με σκοπό να βρεθούν τα K πιο συναφή embeddings σε σχέση με αυτά της εισόδου, όπου K ο αριθμός των κορυφαίων πιο σχετικών embeddings που ορίζεται από τον προγραμματιστή.
- Ανάκτηση Πληροφορίας και Δημιουργία Απάντησης: Αφού εντοπιστούν τα πιο σχετικά κομμάτια κειμένου, χρησιμοποιούνται ως πλαίσιο για την παραγωγή της απάντησης του LLM, ώστε η απάντηση που θα παράγει να βασίζεται και σε αυτά τα δεδομένα. Έτσι, η απάντηση δεν βασίζεται αποκλειστικά στην αρχική εκπαίδευση του μοντέλου, αλλά ενσωματώνει και εξωτερικές πληροφορίες.




5.3.3 Οφέλη

Το RAG προσφέρει πολλά πλεονεκτήματα σε σχέση με τις παραδοσιακές προσεγγίσεις:

- Δυναμική Ανάκτηση Πληροφοριών: Το σύστημα μπορεί να αντλήσει πληροφορίες που δεν περιέχονται στο αρχικό μοντέλο, επιτρέποντας την ενημέρωση και την προσαρμογή σε πραγματικό χρόνο.
- Βελτιωμένη Ακρίβεια και Συνάφεια: Η ενσωμάτωση σχετικού περιεχομένου ως πλαίσιο επιτρέπει στο μοντέλο να παράγει απαντήσεις πιο ακριβείς και προσαρμοσμένες στην ερώτηση του χρήστη.

- Αυτονομία και Ασφάλεια Δεδομένων: Χρησιμοποιώντας τοπικές βάσεις δεδομένων, όπως το Chroma DB, εξασφαλίζεται η προστασία και η ασφάλεια των δεδομένων, χωρίς την ανάγκη για απομακρυσμένες υπηρεσίες ή cloud υποδομές.

# 6. Αποτελέσματα

## 6.1 Μέθοδος αξιολόγησης της απόδοσης του μοντέλου

Στην επεξεργασία φυσικής γλώσσας (NLP), η αξιολόγηση της απόδοσης των μοντέλων που παράγουν κείμενο είναι μια πολύπλοκη δουλειά. Τα μοντέλα που παράγουν περιλήψεις, μεταφράσεις ή άλλες μορφές κειμένου πρέπει να αξιολογούνται όχι μόνο ως προς την ευχέρεια και τη συνοχή τους, αλλά και ως προς το πόσο καλά διατηρούν τις κρίσιμες πληροφορίες από το αρχικό περιεχόμενο. Για να επιτευχθεί αυτό, χρησιμοποιούνται μετρικές αυτόματης αξιολόγησης για τη μέτρηση της ποιότητας του παραγόμενου κειμένου. Μια ευρέως χρησιμοποιούμενη μετρική για την αξιολόγηση των αποτελεσμάτων των μοντέλων παραγωγής κειμένου, ιδίως σε εργασίες περίληψης, είναι η Recall-Oriented Understudy for Gisting Evaluation (ROUGE).

Η μετρική ROUGE έχει σχεδιαστεί για να μετρά πόσες από τις σημαντικές πληροφορίες του κειμένου αναφοράς καταγράφονται από την παραγόμενη περίληψη. Ενώ οι ανθρώπινες αξιολογήσεις προσφέρουν μια βαθύτερη κατανόηση της ποιότητας του κειμένου, η ROUGE παρέχει έναν αποτελεσματικό και κλιμακούμενο τρόπο σύγκρισης του κειμένου που παράγεται από μηχανήματα με τις αναφορές που γράφτηκαν από ανθρώπους. Είναι μια πρακτική που χαρακτηρίζεται από την απλότητα και την αποτελεσματικότητα της, χωρίς μεγάλο κόστος.

Υπάρχουν διάφορες παραλλαγές της μετρικής ROUGE, αλλά πιο συχνά χρησιμοποιούνται οι παρακάτω:

- Rouge-1: Μετράει την επικάλυψη των unigrams (μεμονωμένες λέξεις) μεταξύ της προβλεπόμενης εξόδου και της αναφοράς. Δίνει μια βασική αίσθηση της ομοιότητας σε επίπεδο λέξεων.
- Rouge-2: Μετράει την επικάλυψη των bigrams (ακολουθίες δύο λέξεων). Δίνει μεγαλύτερη εικόνα της δομικής ακρίβειας και στα ζεύγη λέξεων στην έξοδο.
- Rouge-L: Μετρά τη μεγαλύτερη κοινή υποακολουθία μεταξύ των προβλεπόμενων κειμένων και των κειμένων αναφοράς. Αυτή η μετρική είναι ευαίσθητη στη δομή της πρότασης και τη σειρά των λέξεων.

Για κάθε μία από αυτές τις μετρικές Rouge, συνήθως υπολογίζονται τρία στοιχεία:

- Precision: Ακρίβεια, το ποσοστό των n-grams στην προβλεπόμενη έξοδο που υπάρχουν και στην αναφορά.
- Recall: Ανάκληση, το ποσοστό των n-grams στην αναφορά που καταγράφονται από την προβλεπόμενη έξοδο.
- F1-Score: Ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης, παρέχοντας ένα ισορροπημένο μέτρο μεταξύ τους. Υπολογίζεται από την παρακάτω σχέση:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Στο πλαίσιο της παρούσας διατριβής, χρησιμοποιείται η μετρική ROUGE για να αξιολογήσουμε τα μοντέλα στο πόσο καλά απέδωσαν το περιεχόμενο των Reports σε STIX 2.1 bundles, συγκρίνοντας το bundle που παράγουν με το σωστό bundle του εκάστοτε Report, το οποίο ήδη γνωρίζουμε. Η αξιολόγηση έγινε υπολογίζοντας τον μέσο όρο των F1-scores από ένα σύνολο διαφορετικών παραδειγμάτων, έτσι ώστε τα αποτελέσματα αντιπροσωπευτικά.

Ωστόσο, η εύρεση CTI πληροφορίας σε ακατέργαστο κείμενο και η μορφοποίηση της σε STIX 2.1 bundle σε JSON, δεν είναι μόνο ένα πρόβλημα summarization. Εκτός από τη σωστή απόδοση της πληροφορίας που εισάγει το μοντέλο στο STIX 2.1 bundle, είναι εξίσου σημαντική η δομή του bundle, καθώς μια ελάχιστα διαφορετική δομή μπορεί να του δώσει άλλο νόημα ή να το καταστήσει μη κατανοητό.

Η μετρική Rouge δεν έχει τη δυνατότητα να αξιολογήσει τη σωστή δομή του bundle και του JSON αρχείου. Για να μπορέσουμε να αξιολογήσουμε εάν το μοντέλο κατάφερε να αποδώσει σωστά τη δομή του STIX 2.1 bundle, χρησιμοποιήσαμε το εργαλειο STIX Visualizer. Ο STIX Visualizer είναι ένα εργαλείο που εμφανίζει τα STIX bundles ως γραφικές αναπαραστάσεις, όπου τα αντικείμενα του STIX αναπαρίστανται ως κόμβοι, και οι σχέσεις μεταξύ αυτών των αντικειμένων ως ακμές. Αυτό διευκολύνει την κατανόηση της δομής και των συσχετίσεων μεταξύ των δεδομένων, είναι εύκολο στη χρήση και μπορεί να αναγνωρίσει αν οι απαντήσεις του μοντέλου έχουν αποδώσει σωστά τη πληροφορία στο bundle και αν το JSON αρχείο είναι σωστά διαμορφωμένο.

## 6.2 Τεχνικές βελτιστοποίησης της επίδοσης του μοντέλου

Σε αυτή τη διατριβή θα ακολουθήσουμε τις παρακάτω τεχνικές προκειμένου να πάρουμε από το μοντέλο που θα επιλέξουμε το βέλτιστο αποτέλεσμα.

### 6.2.1 Prompt Engineering

Το prompt engineering είναι η διαδικασία σχεδιασμού και βελτιστοποίησης των ερωτημάτων ή εντολών (prompt) που δίνουμε σε LLM για να παραγάγει την καλύτερη δυνατή απάντηση. Η τεχνική αυτή είναι ιδιαίτερα σημαντική για τη μεγιστοποίηση της απόδοσης των instruct μοντέλων, καθώς έχουν εκπαιδευτεί να ανταποκρίνονται σε οδηγίες.

Ο στόχος του prompt engineering είναι να διαμορφώσει το prompt με τέτοιο τρόπο, ώστε το μοντέλο να κατανοεί ακριβώς τι ζητά ο χρήστης και να δίνει την πιο ακριβή, χρήσιμη και κατανοητή απάντηση. Ένα καλά σχεδιασμένο prompt μπορεί να βελτιώσει δραματικά την ακρίβεια και τη συνέπεια της απάντησης, ενώ ένα κακώς σχεδιασμένο prompt μπορεί να οδηγήσει σε ασαφείς ή λανθασμένες απαντήσεις.

### 6.2.2 Few-shot Learning

Το Few-shot Learning είναι μια τεχνική μάθησης που χρησιμοποιείται για τη βελτίωση της απόδοσης ενός μοντέλου, ειδικά σε περιπτώσεις όπου διατίθεται περιορισμένο πλήθος παραδειγμάτων για εκπαίδευση. Στα παραδοσιακά συστήματα μηχανικής μάθησης, τα μοντέλα απαιτούν μεγάλες ποσότητες δεδομένων για να επιτύχουν ακρίβεια και συνέπεια.

Ωστόσο, τα σύγχρονα LLMs με τη βοήθεια του few-shot learning μπορούν να παράγουν ακριβείς και αξιόπιστες προβλέψεις, ακόμα και με λίγα παραδείγματα.

Η βασική ιδέα πίσω από το few-shot learning είναι η παροχή στο μοντέλο με μερικά παραδείγματα μιας συγκεκριμένης εργασίας μέσα στο prompt, ώστε να κατανοήσει το πρότυπο ή την επιθυμητή απάντηση πριν προσπαθήσει να απαντήσει σε νέα δεδομένα. Για παράδειγμα, αντί να εκπαιδεύσουμε το μοντέλο με χιλιάδες δείγματα, μπορούμε να δώσουμε μερικά μόνο παραδείγματα από Reports και τα αντίστοιχα STIX bundles τους, τα οποία επιτρέπουν στο μοντέλο να "μάθει" την εργασία και να παράγει σωστές απαντήσεις σε νέες καταστάσεις.

Ωστόσο, έχουμε περιορισμούς στην εφαρμογή αυτής της τεχνικής, καθώς τα μοντέλα που χρησιμοποιούμε έχουν μικρό context window, με αποτέλεσμα να μη μπορούμε να τους δώσουμε πάνω από δύο τέτοια παραδείγματα. Θα δοκιμάσουμε όμως την τεχνική κι έτσι, για να δούμε κατά πόσο βελτιώνει την έξοδο του συστήματος.

## 6.3 Σύγκριση αποτελεσμάτων των Large Language Models και Prompt Engineering

### 1. Πρώτο πείραμα: Απλό prompt

Ξεκινάμε τα πειράματα με τη πιο απλή προσέγγιση, με σκοπό να αξιολογήσουμε τις γνώσεις που έχουν ήδη τα μοντέλα πάνω στο ζητούμενο, με σκοπό να επιλέξουμε τα πιο αποτελεσματικά μοντέλα. Θα συγκρίνουμε έξι (6) open-source Large Language Models, τα οποία έχουμε κατεβάσει τοπικά από το Ollama. Σε όλα τα μοντέλα ρυθμίστηκε η παράμετρος temperature= 0.1, έτσι ώστε να μειωθεί η στοχαστικότητα των απαντήσεων, και να παράγουν πιο συνεπείς και ακριβείς απαντήσεις. Σε αυτό το στάδιο, ακολουθούμε ένα απλό pipeline, παρακάτω ακολουθούν οι παράμετροι που έχουμε δώσει στο μοντέλο:

**Prompt**= "Extract all cyber threat related information from the following text and format it as a STIX 2.1 Bundle in JSON format. The text:" + input file

**System message**= "You are an assistant that transforms raw Cyber Threat Intelligence (CTI) reports text into STIX 2.1 format. Return the STIX2.1 format output as a json with the appropriate keys."

Αυτές οι παράμετροι είναι αρκετά απλές και δεν δίνουν αναλυτικές οδηγίες στο μοντέλο. Έτσι, θα αναδειχθεί αποκλειστικά η γνώση που έχει ήδη το μοντέλο πάνω στον σχεδιασμό STIX 2.1 bundle.

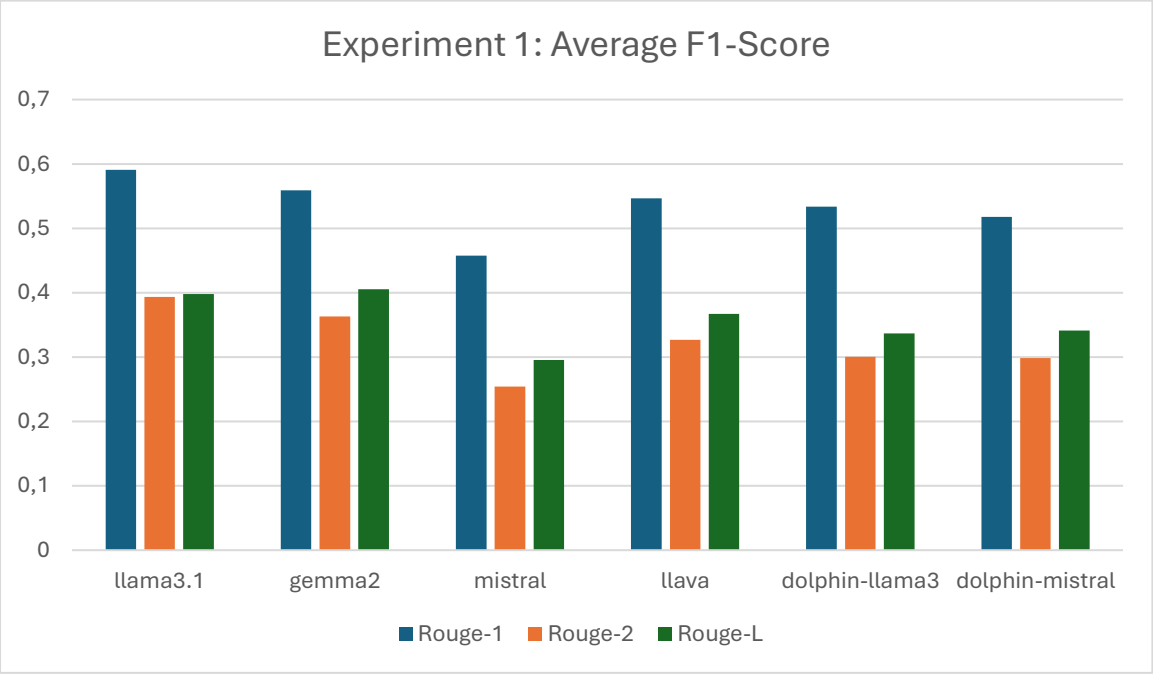
Η διάταξη αυτή έτρεξε για εννέα (9) διαφορετικά reports από την επίσημη σελίδα του STIX, των οποίων γνωρίζουμε τα ακριβή αντίστοιχα bundles, έτσι ώστε να μπορούμε να αξιολογήσουμε τα αποτελέσματα των μοντέλων ως προς αυτά. Για την αξιολόγηση, αφαιρέθηκαν από όλα τα bundles οι αριθμοί των ids και οι ημερομηνίες, καθώς συνήθως τα μοντέλα βάζουν μη έγκυρα. Για την παραγωγή σωστών ids και ημερομηνιών στα bundles, θα προστίθενται με ένα απλό script, χωρίς την χρήση τεχνητής νοημοσύνη.

Στον παρακάτω πίνακα παρουσιάζεται ο **μέσος όρος** των μετρικών Rouge, και συγκεκριμένα

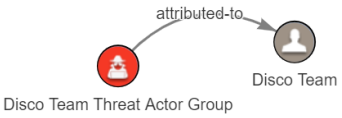
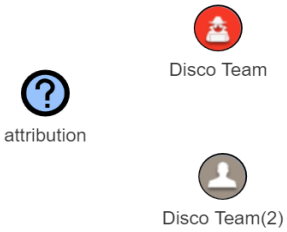
των **F1-Scores**, στο πείραμα:

| Μοντέλο | Rouge-1 (F1-Score) | Rouge-2 (F1-Score) | Rouge-L (F1-Score) |
|---|---|---|---|
| **llama3.1** | **0.5907** | **0.3935** | **0.3981** |
| gemma2 | 0.5592 | 0.3630 | 0.4052 |
| mistral | 0.4575 | 0.2543 | 0.2955 |
| llava | 0.5467 | 0.3270 | 0.3670 |
| dolphin-llama3 | 0.5338 | 0.3006 | 0.3367 |
| dolphin-mistral | 0.5177 | 0.2987 | 0.3413 |

Παρατηρούμε ότι την καλύτερη επίδοση σε όλες της μετρικές είχε το μοντέλο llama3.1. Βλέπουμε ότι είχε το καλύτερο Rouge-1 και Rouge-2, γεγονός που δείχνει πως απέδωσε σωστά μεμονωμένες λέξεις και ακολουθίες δύο λέξεων, ενώ μόνο το Rouge-L του ήταν λίγο χειρότερο από του gemma2, το οποίο αντιπροσωπεύει την επίδοση του στην απόδοση της μεγαλύτερης κοινής υποακολουθίας. Το Rouge-L δεν σημαίνει απαραίτητα πως το μοντέλο απέδωσε σωστότερα την δομή του bundle, όπως και αναφέρθηκε παραπάνω, καθώς σαν μετρική δεν έχει αυτή τη δυνατότητα. Αυτό μπορούμε να το διαπιστώσουμε ανεβάζοντας τα bundles στον STIX Visualizer.




Για να έχουμε μια καλύτερη εικόνα ως προς την απόδοση της δομής του bundle, περάσαμε όλα τα bundles που παράχθηκαν από το llama3.1 από τον STIX Visualizer. Όλα τα bundles διαβάστηκαν επιτυχώς, που σημαίνει ότι το μοντέλο ξέρει να παράγει έγκυρα bundles, ωστόσο πολλές οντότητες χαρακτηρίστηκαν ως άγνωστες, καθώς το μοντέλο δεν τους είχε δώσει τη σωστή ονομασία όπως ορίζεται από το documentation του STIX. Τέλος, το μοντέλο δεν κατάφερε να εντοπίσει τις περισσότερες σχέσεις μεταξύ των οντοτήτων, και σε όσες εντόπισε δεν έδωσε relationship type που τους αντιστοιχεί, με αποτέλεσμα να είναι κενές σχέσεις. Το μοντέλο gemma2, ενώ είχε το υψηλότερο Rouge-L score, κανένα από τα bundles που παρήγαγε δεν διαβάστηκε από τον STIX Visualizer, παρουσίασε syntax error, γεγονός που επιβεβαίωσε ότι το Rouge-L δεν σημαίνει σωστότερη δομή.

| Original Bundle | Llama3.1 | Other Models |
|---|---|---|
|  |  | Error. |

2a. Δεύτερο πείραμα: Εισαγωγή οδηγιών στο prompt/ Prompt engineering

Το προηγούμενο πείραμα ανέδειξε το llama3.1 ως το πιο αποδοτικό LLM για το ζητούμενο. Ωστόσο, η προηγούμενη διάταξη μπορούμε να πούμε πως αδίκησε τα fine-tuned instruct μοντέλα, καθώς αυτά τα μοντέλα έχουν εκπαιδευτεί ειδικά στο να ακολουθούν οδηγίες. Για να εκμεταλλευτούμε, λοιπόν αυτή τους τη δυνατότητα και να ελέγξουμε αν όντως βελτιώνονται τα αποτελέσματα τους στη συγκεκριμένη εργασία με τις κατάλληλες οδηγίες, θα τους δώσουμε μια μικρή καθοδήγηση. Παρακάτω ακολουθούν οι παράμετροι που έχουμε δώσει στο μοντέλο:

**Prompt**= "Extract all cyber threat related information from the following text and format it as a STIX 2.1 Bundle in JSON format, please make sure you format it correctly, the bundle should always start with type:bundle, spec version:2.1, an id and then the objects. Make sure you show the relationships between the objects correctly. The text:" + input file

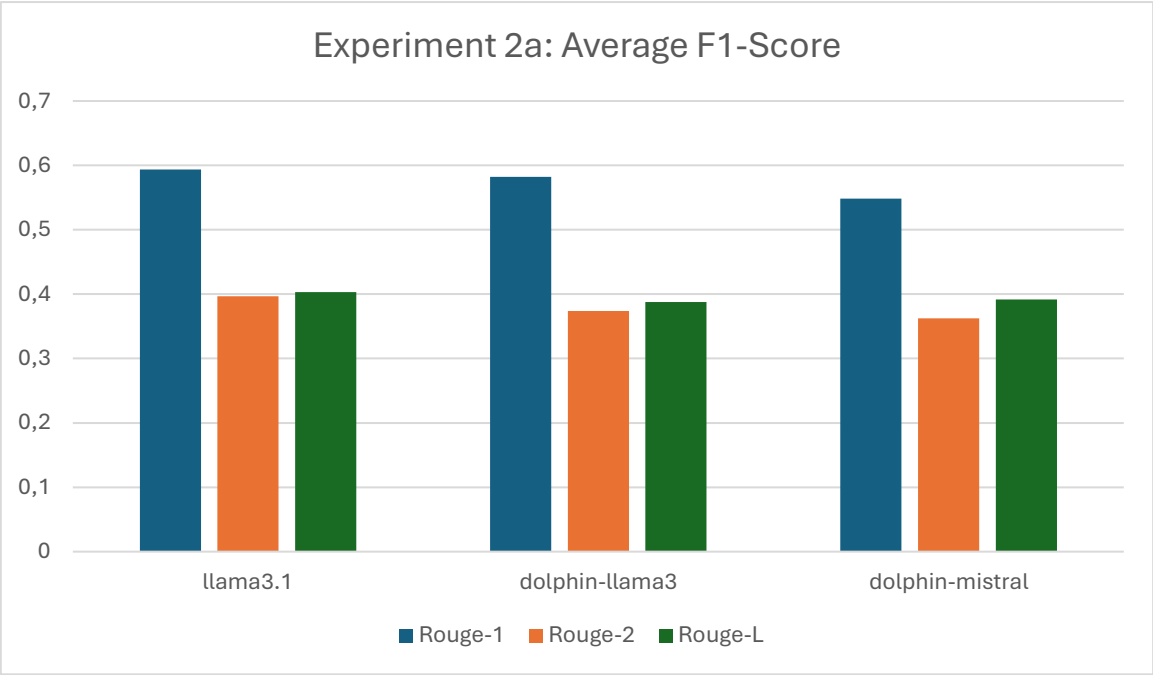
**System message**= "You are an assistant that transforms raw Cyber Threat Intelligence (CTI) reports text into STIX 2.1 format. Return the STIX2.1 format output as a json with the appropriate keys."

Η καθοδήγηση που δίνουμε αφορά κάποια βασικά χαρακτηριστικά των bundles, που απουσίαζαν από τα αποτελέσματα των Instruct στο προηγούμενο πείραμα, και βοηθάνε τόσο στο περιεχόμενο όσο και στην κατασκευή της δομής του bundle.

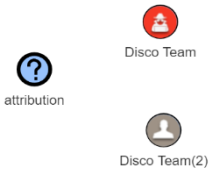
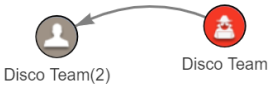
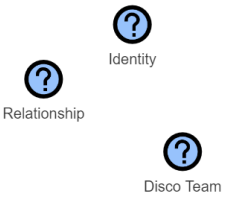
Στον παρακάτω πίνακα παρουσιάζεται ο μέσος όρος των μετρικών Rouge, και συγκεκριμένα των F1-Scores, στο πείραμα:

| Μοντέλο | Rouge-1 (F1-Score) | Rouge-2 (F1-Score) | Rouge-L (F1-Score) |
|---|---|---|---|
| **llama3.1** | **0.5938** | **0.3965** | **0.4033** |
| dolphin-llama3 | 0.5823 | 0.3739 | 0.3877 |
| dolphin-mistral | 0.5486 | 0.3626 | 0.3918 |

Παρατηρούμε ότι αμέσως, με λίγες αλλά σημαντικές οδηγίες περιεχομένου και δομής στο prompt, η επίδοση των fine-tuned instruct μοντέλων βελτιώθηκε σημαντικά, πλησιάζοντας την επίδοση του llama3.1. Από αυτά τα αποτελέσματα μπορούμε να συμπεράνουμε πως τα μοντέλα αυτά μπορούν να φανούν εξίσου αποδοτικά, και ενδεχομένως πιο αποδοτικά, αν τους δοθεί η κατάλληλη καθοδήγηση.




Στη συνέχεια, περάσαμε στον STIX Visualizer τα αποτελέσματα των μοντέλων, με σκοπό να αξιολογήσουμε πως αποδόθηκε η δομή του bundle:

| Original Bundle | Llama3.1 | Dolphin-llama3 | Dolphin-mistral |
|---|---|---|---|
|  |  |  |  |









Από την αναπαράσταση των αποτελεσμάτων στον STIX Visualizer είναι φανερό πως τα instruct μοντέλα με τις κατάλληλες εντολές μπορούν να δημιουργήσουν bundles που διαβάζονται από αυτόν. Μάλιστα, το dolphin-llama3 φαίνεται πως αναγνώρισε πως υπάρχει κάποια σχέση μεταξύ των οντοτήτων, ενώ το llama3.1 δεν είχε σχεδόν καμία βελτίωση. Επομένως, μπορούμε να συμπεράνουμε ότι το dolphin-llama3 με την κατάλληλη καθοδήγηση ενδεχομένως και να περάσει τις επιδόσεις του llama3.1.

b. Τα instruct μοντέλα έχουν εκπαιδευτεί για να ακολουθούν συγκεκριμένες οδηγίες και υποδείξεις μέσω διαλόγων και prompts που διατυπώνονται με συγκεκριμένη μορφή. Με αυτή τη λογική, μπορούμε να αξιοποιήσουμε τον τρόπο με τον οποίο το μοντέλο έχει εκπαιδευτεί να κατανοεί και να επεξεργάζεται τις οδηγίες, δίνοντας στο system message τη μορφή που ξέρει καλύτερα το μοντέλο.

Τα system messages που ακολουθούν προέκυψαν από τις επίσημες σελίδες των μοντέλων, όπου αναφέρεται πως τα δεδομένα στα οποία εκπαιδεύτηκαν πάνω είχαν αυτή τη μορφή.

Για τα llama3.1 και dolphin-llama3 θα θέσουμε:

**Prompt**= "Extract all cyber threat related information from the following text and format it as a STIX 2.1 Bundle in JSON format, please make sure you format it correctly, the bundle should always start with type:bundle, spec version:2.1, an id and then the objects. Make sure you show the relationships between the objects correctly. The text:" + input file

**System message**= "<|begin_of_text|> <|start_header_id|>system<|end_header_id|>

You are an assistant that transforms raw Cyber Threat Intelligence (CTI) reports text into STIX 2.1 format. Return the STIX2.1 format output as a json with the appropriate keys.<|eot_id|>
<|start_header_id|>user<|end_header_id|>

{prompt + input}(Here ends the CTI Report)<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>"

Για το dolphin-mistral θα θέσουμε:

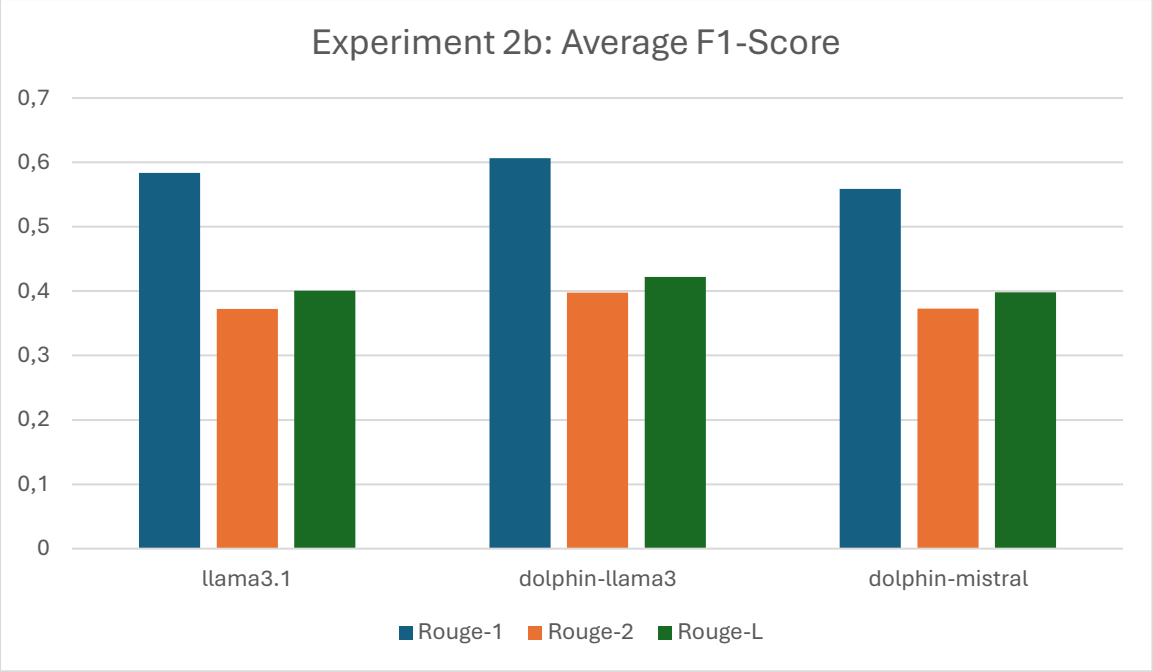
**Prompt**= "Extract all cyber threat related information from the following text and format it as a STIX 2.1 Bundle in JSON format, please make sure you format it correctly, the bundle should always start with type:bundle, spec version:2.1, an id and then the objects. Make sure you show the relationships between the objects correctly. The text:" + input file

**System message**= "[INST] You are an assistant that transforms raw Cyber Threat Intelligence (CTI) reports text into STIX 2.1 format. Return the STIX2.1 format output as a json with the appropriate keys [/INST]"

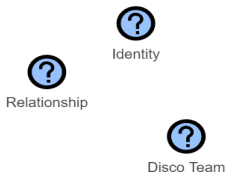
Στον παρακάτω πίνακα παρουσιάζεται ο μέσος όρος των μετρικών Rouge, και συγκεκριμένα των F1-Scores, στο πείραμα:

| Μοντέλο | Rouge-1 (F1-Score) | Rouge-2 (F1-Score) | Rouge-L (F1-Score) |
|---|---|---|---|
| llama3.1 | 0.5838 | 0.3724 | 0.4005 |
| **dolphin-llama3** | **0.6067** | **0.3979** | **0.4223** |
| dolphin-mistral | 0.5590 | 0.3729 | 0.3984 |

Παρατηρούμε ότι όντως, στην περίπτωση του dolphin-llama3, το μοντέλο φαίνεται πιο δεκτικό στις οδηγίες που δίνονται μέσω του system message. Αντίθετα, φαίνεται το llama3.1 να μην είναι τόσο ευαίσθητο στη δομή του system message, καθώς δεν έχει δεχτεί το fine tuning του dolphin μοντέλου. Αυτό μπορεί να αποδοθεί στο γεγονός ότι τα instruct μοντέλα, όπως το dolphin-llama3, εκπαιδεύονται σε μεγάλες ποσότητες δεδομένων με οδηγίες όπου τα system messages είναι θεμελιώδη στοιχεία για τη διαμόρφωση των απαντήσεων.




Στη συνέχεια, περάσαμε στον STIX Visualizer τα αποτελέσματα των μοντέλων, με σκοπό να αξιολογήσουμε πως αποδόθηκε η δομή του bundle:

| Original Bundle | Llama3.1 | Dolphin-llama3 | Dolphin-mistral |
|---|---|---|---|
|   | Error. |   |   |

Από την αναπαράσταση των bundles στον STIX Visualizer παρατηρούμε ότι το συγκεκριμένο prompt «χάλασε» την δομή του llama3.1, με αποτέλεσμα να μη διαβάζεται από το εργαλείο. Αντίθετα, το dolphin-llama3 βελτίωσε τα αποτελέσματα του, προσθέτοντας στο relationship τύπο. Αυτός ο τύπος που απέδωσε όμως δεν είναι σωστός, ούτε υπαρκτός στο STIX documentation.

Επομένως, από αυτά τα πειράματα μπορούμε να βγάλουμε το εξής συμπέρασμα:

Με τη κατάλληλη καθοδήγηση μέσω του system message και του κατάλληλου prompting, το μοντέλο dolphin-llama3 μπορεί να φτάσει, ή ακόμη και να ξεπεράσει τις επιδόσεις του πιο σύγχρονου llama3.1. Φαίνεται ότι το dolphin-llama3 είναι πιο δεκτικό στις οδηγίες, πιθανώς λόγω της πιο εκτενούς εκπαίδευσής του σε datasets με οδηγίες και instruct-based αλληλεπιδράσεις. Αυτό το χαρακτηριστικό του επιτρέπει να επωφελείται περισσότερο από την παροχή σαφών prompts και να προσαρμόζει τις απαντήσεις του με μεγαλύτερη ακρίβεια και συνέπεια. Αντίθετα, το llama3.1 φαίνεται να επηρεάζεται αρνητικά από πιο σύνθετα prompts ή system messages, γεγονός που μπορεί να

οφείλεται σε λιγότερη εκπαίδευση σε τέτοια πλαίσια, οδηγώντας το σε ελαφρύ αποπροσανατολισμό και μείωση της απόδοσής του.

Επίσης, παρά την καλύτερη επίδοση του, φαίνεται πως το dolphin-llama3 δεν γνωρίζει τις σωστές ορολογίες που ορίζονται από το STIX 2.1 documentation, με αποτέλεσμα να πέφτει σε hallucinations, βάζοντας μη υπαρκτούς τύπους.

### 6.4 Αποτελέσματα της τεχνικής few-shot learning

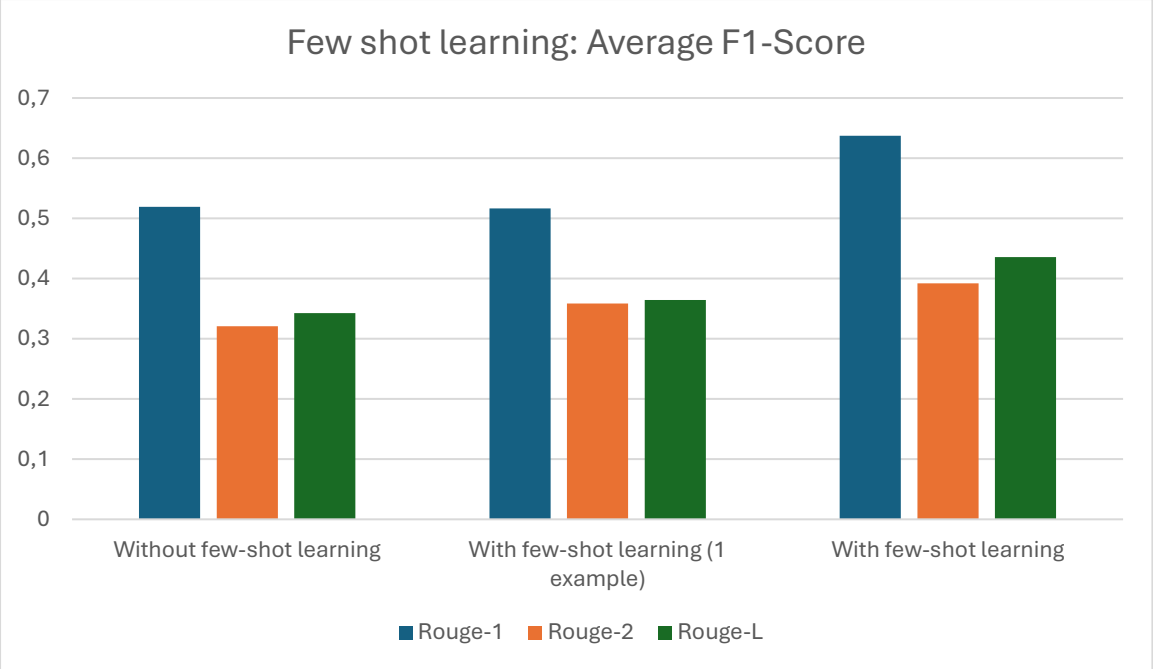
Από το σύνολο Reports-Bundles που διαθέτουμε αφαιρούμε δύο, με σκοπό να τα χρησιμοποιήσουμε στην υλοποίηση της τεχνικής του few-shot learning. Ουσιαστικά, με αυτή τη τεχνική δίνουμε παραδείγματα στο μοντέλο με σκοπό να του δείξουμε ακριβώς τι θέλουμε να κάνει.

Για να μπορέσουμε να συγκρίνουμε τα αποτελέσματα του few-shot learning με αυτά από τα προηγούμενα πειράματα, υπολογίζουμε τα αντίστοιχα average f1-scores τους για τα 7 αντί για τα 9 παραδείγματα, ώστε να είναι δίκαιη σύγκριση. Το πείραμα θα τρέξει για το καλύτερο μέχρι στιγμής μοντέλο, δηλαδή το dolphin-llama3 και για το prompt που μας έδωσε την καλύτερη του επίδοση, συν τα παραδείγματα για το few-shot learning. Πρώτα για ένα παράδειγμα και έπειτα για δύο.

Στον παρακάτω πίνακα παρουσιάζεται ο μέσος όρος των μετρικών Rouge, και συγκεκριμένα των F1-Scores, στο πείραμα:

| Πείραμα | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|
| Without few-shot learning | 0.5194 | 0.3210 | 0.3424 |
| With few-shot learning (1 example) | 0.5165 | 0.3585 | 0.3643 |
| With few-shot learning (2 examples) | 0.6373 | 0.3920 | 0.4355 |

Παρατηρούμε ότι το few-shot learning είχε εξαιρετικά αποτελέσματα στην απόδοση του κειμένου, καθώς βελτίωσε σημαντικά το Rouge-2 που αντιπροσωπεύει τις δυάδες λέξεων. Ωστόσο, δοκιμάζοντας τα αποτελέσματα στον STIX Visualizer, παρατηρούμε ότι δεν διαβάζονται από αυτόν, καθώς πολλές από τις απαντήσεις κόπηκαν πριν τελειώσει η πλήρης δημιουργία του bundle. Αυτό ευθύνεται στο ότι το μοντέλο που χρησιμοποιούμε έχει μικρό context window, και τα μεγάλα κείμενα και bundles που του παρείχαμε κατέλαβαν πολύ χώρο.




## 6.5 Αποτελέσματα της αρχιτεκτονικής του RAG

### 6.5.1 Περίληψη πειράματος

Σε αυτή την ενότητα θα αναλύσουμε και θα σχολιάσουμε τα αποτελέσματα του RAG. Όπως αναλύθηκε και στο προηγούμενο κεφάλαιο, το pipeline που σχεδιάσαμε έχει ως εξής:

Αρχικά, τρέχοντας τον server, το σύστημα ελέγχει αν υπάρχουν ήδη αποθηκευμένα τα embeddings στη vector database του vault.txt, αρχείο το οποίο περιέχει το κείμενο με τις επιπλέον πληροφορίες που θέλουμε να εμπλουτίσουμε τις γνώσεις του μοντέλου. Το vault.txt στο πείραμα μας περιέχει το documentation του STIX 2.1, ελαφρώς τροποποιημένο ώστε να είναι ευανάγνωστοι οι πίνακες και να μην έχει περιττή για το ζητούμενο πληροφορία. Αφού παραχθούν τα embeddings, με τον τρόπο που εξηγήσαμε στο προηγούμενο κεφάλαιο, αποθηκεύονται στην vector database για μελλοντική χρήση.

Ο χρήστης καλείται να δώσει στο σύστημα ένα αρχείο κειμένου που περιέχει το CTI Report. Αυτό το Report ενώνεται με μια συγκεκριμένη ερώτηση, που είναι ουσιαστικά το αίτημα που θέτουμε στο μοντέλο, και αυτά τα δύο μαζί δημιουργούν το prompt. Αυτό το prompt στην συνέχεια σπάει σε embeddings, χρησιμοποιώντας τον ίδιο tokenizer με τα vault embeddings, και το σύστημα ψάχνει να βρει ομοιότητες στις οδηγίες που του έδωσε ο χρήστης με τις μεθοδολογίες που βρίσκει στα embeddings που παράχθηκαν από το STIX 2.1 documentation, συγκεκριμένα τις κορυφαίες K από αυτές. Αφού βρει τις κορυφαίες K επιλογές, της προσθέτει στο system message για να της λάβει υπόψη το LLM στην παραγωγή της απάντησης του.

Τέλος, η απάντηση του μοντέλου περνάει από ένα ειδικό script για να πάρει το bundle την τελική του μορφή, το οποίο αφαιρεί το περιττό κείμενο και βάζει σωστά ids και ημερομηνίες.

### 6.5.2 Εκτέλεση πειράματος

Η πειραματική διάταξη για το σύστημα που χρησιμοποιεί την τεχνολογία του RAG έτρεξε χρησιμοποιώντας τα μοντέλα llama3.1 και dolphin-llama3, για εννέα (9) διαφορετικά reports από την επίσημη σελίδα του STIX, των οποίων γνωρίζουμε τα ακριβή αντίστοιχα bundles, έτσι ώστε να μπορούμε να αξιολογήσουμε τα αποτελέσματα των μοντέλων ως προς αυτά.

Και στα δύο μοντέλα ρυθμίστηκε η παράμετρος temperature= 0.1, έτσι ώστε να μειωθεί η στοχαστικότητα των απαντήσεων του μοντέλου, ώστε να παράγει πιο συνεπείς και ακριβείς απαντήσεις. Επιπλέον, θέσαμε K=10, όπου K τα κορυφαία αποτελέσματα της αναζήτησης του RAG στο STIX 2.1 documentation. Στο πλαίσιο του πειραματισμού, το πείραμα έτρεξε για K=5 και K=12, φέρνοντας χειρότερα αποτελέσματα, λόγω μη επαρκούς πληροφορίας που ανακτήθηκε και λόγω υπέρβασης του χώρου του context window αντίστοιχα.

Για την αξιολόγηση, αφαιρέθηκαν από όλα τα bundles οι αριθμοί των ids και οι ημερομηνίες, καθώς συνήθως τα μοντέλα βάζουν μη έγκυρα. Για την παραγωγή σωστών ids και ημερομηνιών στα bundles, θα προστίθενται με ένα απλό script, χωρίς την χρήση τεχνητής νοημοσύνη.

Για το κάθε μοντέλο χρησιμοποιήθηκε το prompt και το system message με τα οποία έφερε τα καλύτερα αποτελέσματα στη προηγούμενη υποενότητα, συν οι οδηγίες από το STIX 2.1 documentation τις οποίες θα ανακτήσει το RAG.

Για το llama3.1 είναι:

**Prompt**= "Extract all cyber threat related information from the following text and format it as a STIX 2.1 Bundle in JSON format, please make sure you format it correctly, the bundle should always start with type:bundle, spec version:2.1, an id and then the objects. Make sure you show the relationships between the objects correctly. The text:" + input file

**System message**= "You are an assistant that transforms raw Cyber Threat Intelligence (CTI) reports text into STIX 2.1 format. Return the STIX2.1 format output as a json with the appropriate keys.

You are being provided with information from STIX 2.1 documentation, the most relevant information to the user iput. The information: {context}"

Για το dolphin-llama3 είναι:

**Prompt**= "Extract all cyber threat related information from the following text and format it as a STIX 2.1 Bundle in JSON format, please make sure you format it correctly, the bundle should always start with type:bundle, spec version:2.1, an id and then the objects. Make sure you show the relationships between the objects correctly. The text:" + input file

**System message**= <|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are an assistant that transforms raw Cyber Threat Intelligence (CTI) reports text into STIX 2.1 format.

Return the STIX2.1 format output as a json with the appropriate keys.

You are being provided with information from STIX 2.1 documentation, the most relevant information to the user input. The information: {context}

<|eot_id|>

<|start_header_id|>user<|end_header_id|>

{prompt + input}(Here ends the CTI

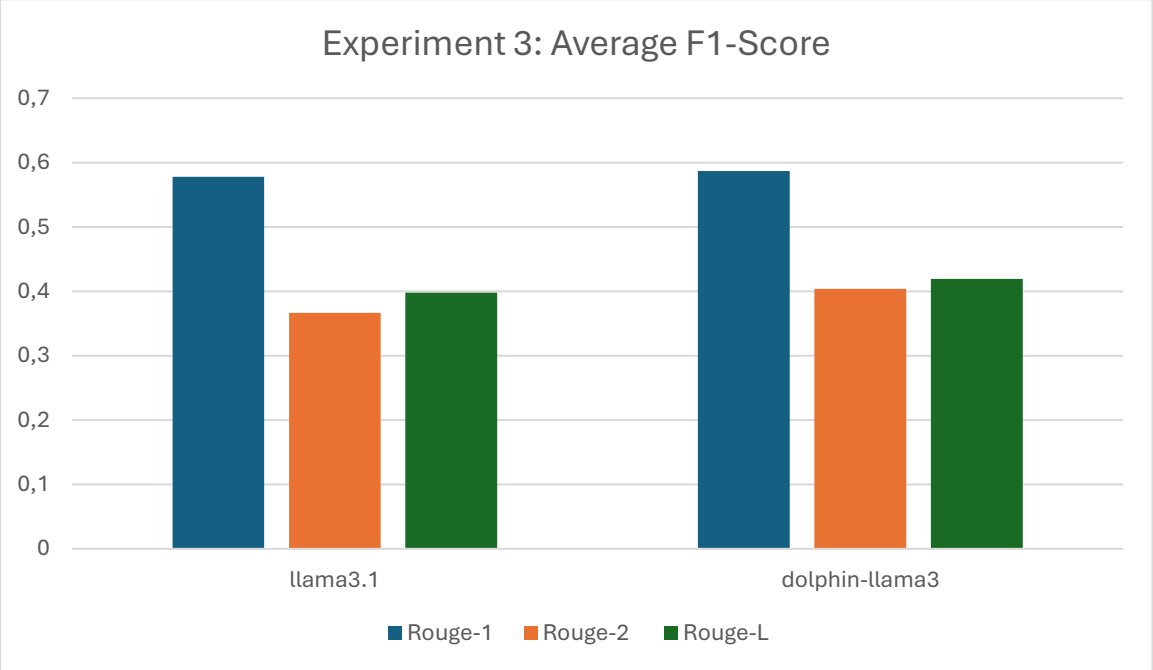
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>"

Στον παρακάτω πίνακα παρουσιάζεται ο μέσος όρος των μετρικών Rouge, και συγκεκριμένα των F1-Scores, στο πείραμα:

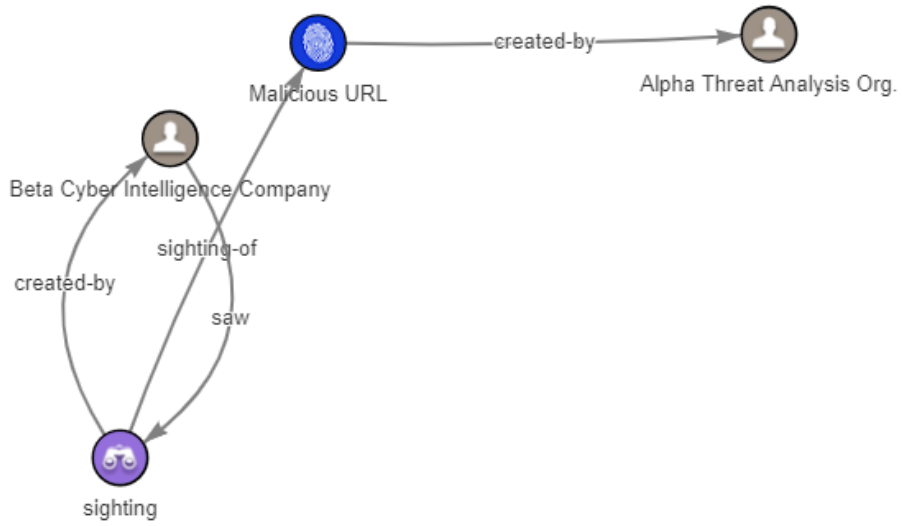
| Μοντέλο | Rouge-1 (F1-Score) | Rouge-2 (F1-Score) | Rouge-L (F1-Score) |
|---|---|---|---|
| llama3.1 | 0.5780 | 0.3665 | 0.3980 |
| **dolphin-llama3** | **0.5869** | **0.4040** | **0.4194** |

Παρατηρούμε ότι το RAG με το STIX 2.1 documentation, για K=10, με μοντέλο το dolphin-llama3 έφερε καλύτερα αποτελέσματα συγκριτικά με το προηγούμενο πείραμα χωρίς το RAG, ως προς τη μετρική Rouge και συγκεκριμένα το Rouge-2. Αυτό μας δείχνει πως αποτυπώθηκαν καλύτερα οι ακολουθίες δύο λέξεων και κατ' επέκταση το νοηματικό περιεχόμενο των bundles.

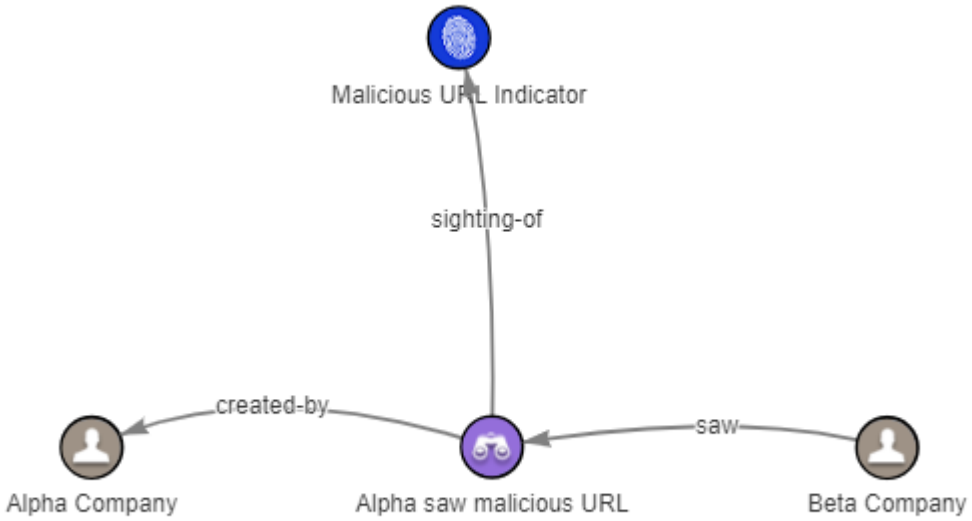


Παρακάτω φαίνονται τα αποτελέσματα του RAG σε αναπαράσταση από τον STIX Visualizer:

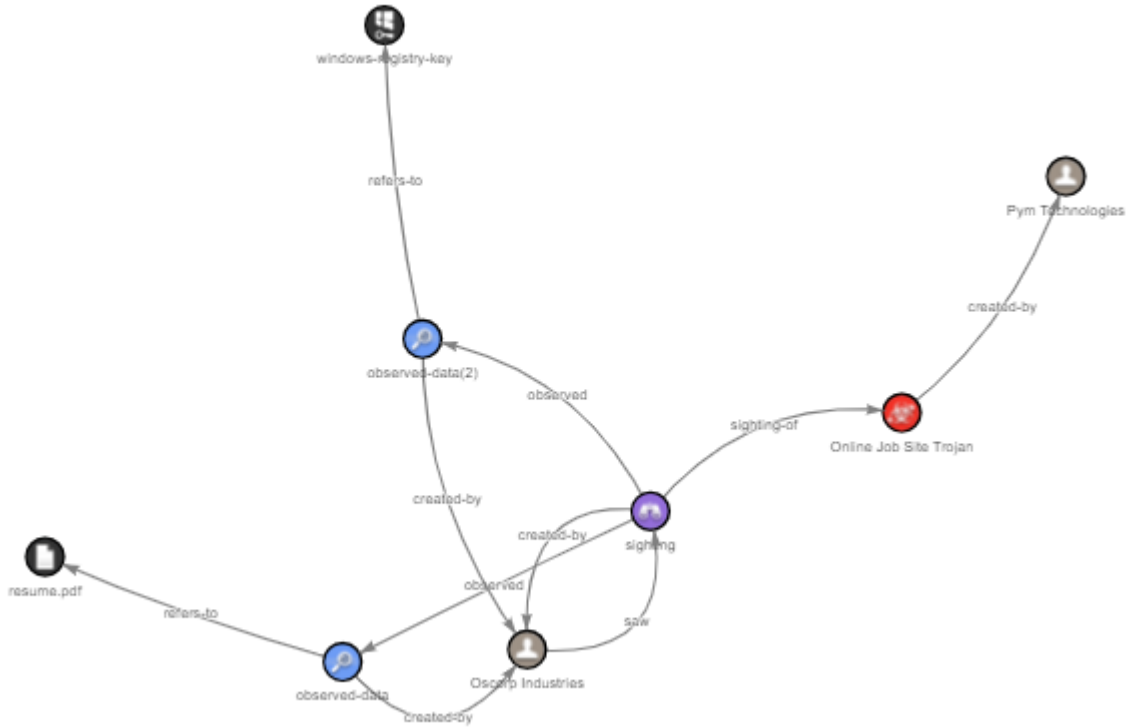
Original Bundle:



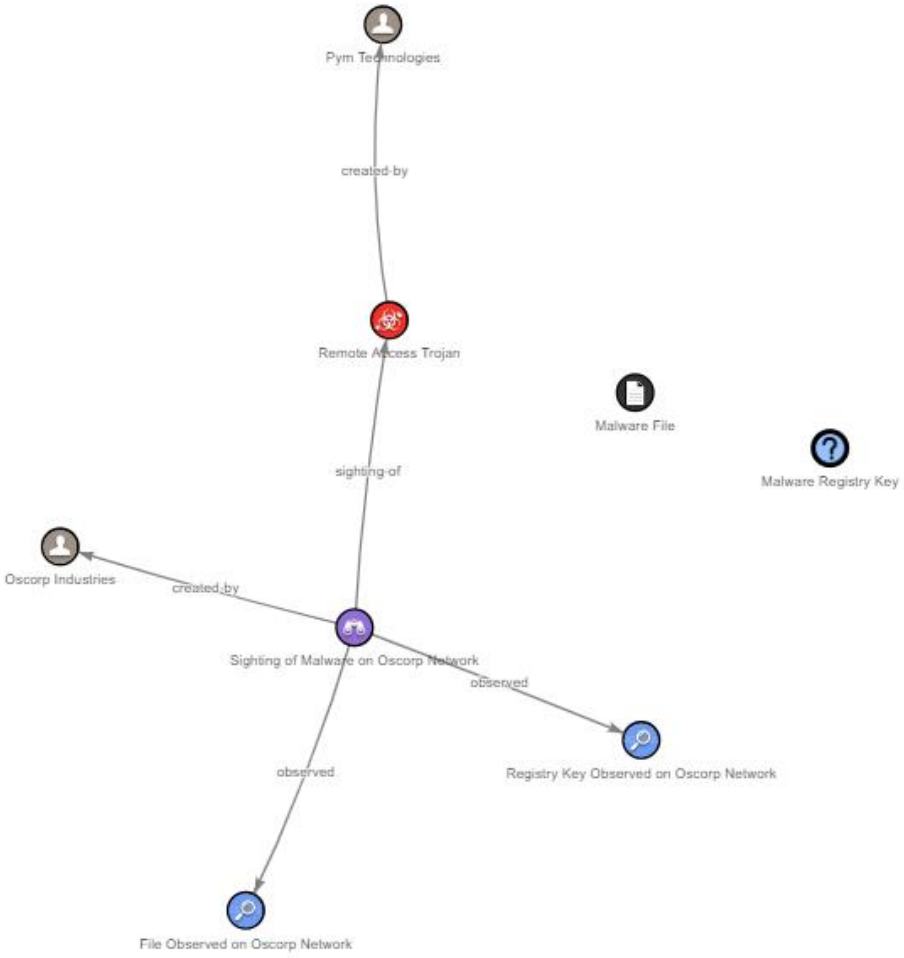

Dolphin-llama3 with RAG:



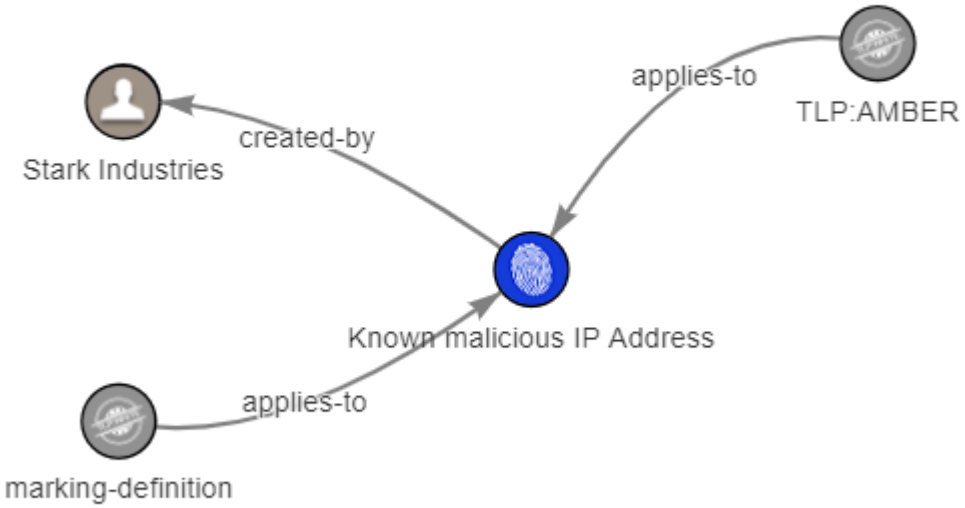
Original Bundle:



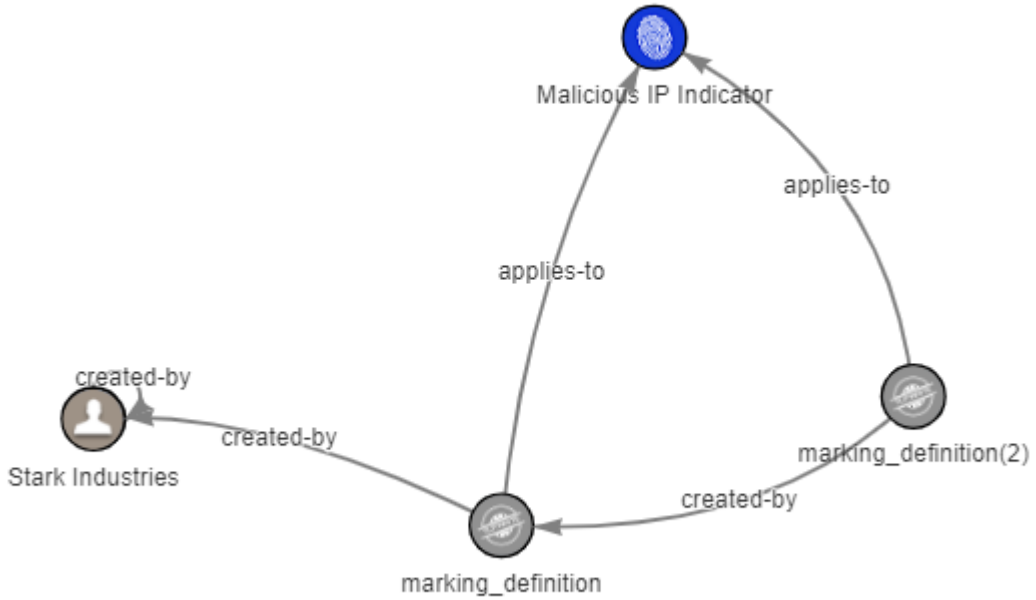

Dolphin-llama3 with RAG:



Original Bundle:




Dolphin-llama3 with RAG:



# 7. Συμπέρασμα

Τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) αποτελούν το μέλλον της ανακάλυψης και ανάλυσης πληροφοριών Cyber Threat Intelligence (CTI), προσφέροντας εξαιρετικές δυνατότητες στην επεξεργασία και ανάλυση τεράστιων όγκων δεδομένων σε πραγματικό χρόνο και την εκτέλεση σύνθετων εργασιών παραγωγής κειμένου. Αυτές οι δυνατότητες αναδείχθηκαν μέσα από τα πειράματα μας, που αφορούν όχι μόνο την παραγωγή κειμένου, αλλά πιο συγκεκριμένα JSON αρχείο και ταυτόχρονα με δομή STIX 2.1.

Τα αποτελέσματα των πειραμάτων ανέδειξαν την επιτυχημένη λειτουργία τοπικών LLMs μικρής κλίμακας σε τέτοιες εργασίες, αποδεικνύοντας ότι είναι ικανά να αποδώσουν ικανοποιητικά στην ανάλυση και εξαγωγή πληροφοριών CTI, ακόμα και όταν λειτουργούν σε περιβάλλοντα με περιορισμένους πόρους. Αυτά τα μοντέλα προσφέρουν μια βιώσιμη λύση για τους οργανισμούς κρατώντας χαμηλά το κόστος, ενώ παράλληλα διασφαλίζουν την ασφάλεια και ακεραιότητα των ευαίσθητων δεδομένων του οργανισμού.

Επιπλέον, τα αποτελέσματα ανέδειξαν τα LLaMa μοντέλα ως τα πιο αποδοτικά για την εξαγωγής δομημένης πληροφορίας στο πρότυπο του STIX, με το Llama3.1 να αποδεικνύεται το πιο εξοικειωμένο με το αντικείμενο, ενώ το dolphin-llama3 διακρίθηκε για την ικανότητά του να παράγει πιο σωστά STIX 2.1 bundles με τις κατάλληλες οδηγίες και βοηθήματα.

Εξίσου σημαντική ήταν και η συμβολή των τεχνικών prompt engineering, ιδιαίτερα για τα Instruct μοντέλα Η κατάλληλη διαμόρφωση του prompt και η παροχή αναλυτικών οδηγιών βελτίωσε σημαντικά την απόδοσή τους, δείχνοντας πως το σωστό prompt παίζει καθοριστικό ρόλο στην επίδοση του μοντέλου. Το πείραμα έδειξε ότι τα Instruct μοντέλα, όταν λαμβάνουν σαφείς και ακριβείς οδηγίες, μπορούν να ανταγωνιστούν πιο σύνθετα μοντέλα, προσφέροντας ευελιξία και ακρίβεια.

Τέλος, η ενσωμάτωση της τεχνολογίας του RAG σε συνδυασμό με ένα Instruct μοντέλο είχε καθοριστική συμβολή στη βελτίωση της απόδοσης. Με το RAG, το σύστημα μπόρεσε να αντλήσει πληροφορίες από το STIX 2.1 documentation, ενσωματώνοντας τες στην διαδικασία παραγωγής της απάντησης του μοντέλου, παρέχοντας ακριβείς και ενημερωμένες απαντήσεις, ευθυγραμμισμένες με το πρότυπο STIX. Αυτή η ενίσχυση βελτίωσε σημαντικά την ακρίβεια και την ορθότητα των παραγόμενων STIX bundles, αποδεικνύοντας ότι οι αποδόσεις αυξάνονται όταν το μοντέλο έχει πρόσβαση σε σχετικές πληροφορίες κατά τη διάρκεια της επεξεργασίας.

# 8. Μελλοντικές κατευθύνσεις

Υπάρχουν αρκετές κατευθύνσεις που μπορούν να ακολουθηθούν ώστε να βελτιωθεί η απόδοση και η ακρίβεια του συστήματος στην παραγωγή σωστότερων STIX 2.1 bundles. Πρώτα, θα μπορούσε να βελτιωθεί η τεχνολογία του RAG με σύγχρονες τεχνικές, όπως τη δημιουργία γράφου γνώσης (knowledge graph) που βοηθάει καλύτερα στην κατανόηση των σχέσεων μεταξύ των embeddings ή η υιοθέτηση της τεχνικής του Graph-RAG, ένα νέο framework όπου η ανάκτηση δεν βασίζεται μόνο στην ομοιότητα αλλά και στις σημασιολογικές σχέσεις μεταξύ των οντοτήτων στον γράφο γνώσης.

Είναι σαφές πως τα LLMs που δοκιμάσαμε χρειάζονται περισσότερη πληροφορία σχετικά με το αντικείμενο προκειμένου να παράγουν τα αποτελέσματα που θέλουμε. Οπότε αυτό που πραγματικά χρειάζεται είναι να βρεθούν τα κατάλληλα δεδομένα και οι υπολογιστικοί πόροι ώστε το μοντέλο να γίνει fine-tuning στην συγκεκριμένη εργασία. Αυτό μπορεί να επιτευχθεί μόνο με δεδομένα κειμένου και τα ακριβώς αντίστοιχα τους bundles, και χρειάζεται set δεδομένων που δεν υπάρχουν διαθέσιμα δημοσίως.

Ο τελικός προορισμός αυτής της εφαρμογής είναι η ένωση της με έναν web scrapper, ο οποίος θα επιτρέπει την άντληση δεδομένων σε πραγματικό χρόνο από online πηγές όπως forums του darkweb ή feeds από το twitter. Αυτό έχει ως σκοπό η εφαρμογή να παράγει και να ενημερώνει αυτόματα τα bundles σε πραγματικό χρόνο, βασιζόμενη στις  πιο πρόσφατες πληροφορίες για κυβερνοαπειλές, μόλις αυτές πρωτοεμφανιστούν.

Συνολικά, τα επόμενα βήματα θα περιλαμβάνουν την ενσωμάτωση νέων τεχνολογιών που αποσκοπούν στην βελτίωση των αποτελεσμάτων, την περαιτέρω εκπαίδευση των μοντέλων σε συγκεκριμένα δεδομένα και την συνολική βελτιστοποίηση της απόδοσης του συστήματος, με στόχο την άμεση και αξιόπιστη αναπαράσταση των πιο καινούργιων απειλών στον κυβερνοχώρο τη στιγμή που αυτές δημοσιευθούν στο διαδίκτυο.

# 9. Βιβλιογραφία

[1] Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., & Wen, J. (2023). A Survey of Large Language Models. *ArXiv*, abs/2303.18223. https://doi.org/10.48550/arXiv.2303.18223.

[2] Zhai, C. (2008). Statistical Language Models for Information Retrieval: A Critical Review. *Found. Trends Inf. Retr.*, 2, 137-213. https://doi.org/10.1561/1500000008.

[3] Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3, 1137-1155.

[4] Airoldi, E. (2016). Rejoinder. *Journal of the American Statistical Association*, 111, 1410 - 1412. https://doi.org/10.1080/01621459.2016.1245070.

[5] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *ArXiv*, abs/1802.05365. https://doi.org/10.18653/v1/N18-1202.

[6] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. , 4171-4186. https://doi.org/10.18653/v1/N19-1423.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv, abs/1706.03762*. https://doi.org/10.48550/arXiv.1706.03762.

[8] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

[9] Chang, K.-H. (2023). Natural Language Processing: Recent Development and Applications. *Applied Sciences*, *13*(20), 11395. https://doi.org/10.3390/app132011395

[10] Schopf, T., Arabi, K., & Matthes, F. (2023). Exploring the Landscape of Natural Language Processing Research. , 1034-1045. https://doi.org/10.26615/978-954-452-092-2_111.

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv, abs/1301.3781*. https://doi.org/10.48550/arXiv.1301.3781

[12] Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010). Why Does Unsupervised Pre-training Help Deep Learning?. , 201-208. https://doi.org/10.5555/1756006.1756025.

[13] Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. *ArXiv preprint*, *arXiv:2401.05856*. https://doi.org/10.48550/arXiv.2401.05856

[14] Mei, S., Montanari, A., & Nguyen, P.-M. (2022). Pretraining. In *Principles of Deep Learning Theory* (pp. 363-402). Cambridge University Press. https://doi.org/10.1017/9781108560793.013

[15] Shiri, F., Perumal, T., Mustapha, N., & Mohamed, R. (2023). A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU. *ArXiv*, abs/2305.17473. https://doi.org/10.48550/arXiv.2305.17473.

[16] Mienye, I. D. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information*, *15*(9), 517. https://doi.org/10.3390/info15090517

[17] Shanahan, M. (2023). Talking About Large Language Models. *Communications of the ACM*, *66*(10), 36-38. https://doi.org/10.1145/3624724

[18] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). A survey on evaluation of large language models. *ArXiv preprint*, *arXiv:2307.03109*. https://doi.org/10.48550/arXiv.2307.03109

[19] ] Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 1045-1048). International Speech Communication Association (ISCA).

[20] Lepelaars, C. (2023). A Deep Dive Into the Transformer Architecture. *Weights & Biases*. https://wandb.ai/carlolepelaars/transformer_deep_dive/reports/A-Deep-Dive-Into-the-Transformer-Architecture--VmlldzozODQ4NDQ

[21] Fayyazi, R., Taghdimi, R., & Yang, S. J. (2024). Advancing TTP analysis: Harnessing the power of large language models with retrieval augmented generation. *ArXiv preprint*, *arXiv:2401.00280*. https://doi.org/10.48550/arXiv.2401.00280

[22] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. , 4171-4186. https://doi.org/10.18653/v1/N19-1423.

[23] Nyandwi, J. (n.d.). The Transformer Blueprint: A Holistic Guide to the Transformer Neural Network Architecture. Carnegie Mellon University. https://deeprevision.github.io/posts/001-transformer/

[24] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv preprint*, *arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

[25] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI*. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

[26] OpenAI. (2023). GPT-4 Technical Report. *ArXiv preprint*, *arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

[27] Kaplan, J., McCandlish, S., Henighan, T., Brown, T., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *ArXiv*, abs/2001.08361.

[28] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ArXiv*, abs/2311.05232. https://doi.org/10.48550/arXiv.2311.05232.

[29] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., De Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J., Elsen, E., & Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. *Proceedings of the 39th International Conference on Machine Learning*, *PMLR 162*, 2206-2240. https://proceedings.mlr.press/v162/borgeaud22a.html

[30] Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 11, 1316-1331. https://doi.org/10.1162/tacl_a_00605.

[31] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv*, abs/2005.11401.

[32] Cybersecurity and Infrastructure Security Agency. (n.d.). What is cybersecurity? *CISA*. Retrieved October 17, 2024, from https://www.cisa.gov/news-events/news/what-cybersecurity

[33] BM. (n.d.). Cyber threats: Types and what to know. *IBM*. Retrieved October 17, 2024, from https://www.ibm.com/think/topics/cyberthreats-types

[34] IBM. (n.d.). What is threat intelligence? *IBM*. Retrieved October 17, 2024, from https://www.ibm.com/topics/threat-intelligence

[35] SISA. (n.d.). The six phases of threat intelligence lifecycle. *SISA Infosec*. Retrieved October 17, 2024, from https://www.sisainfosec.com/blogs/the-six-phases-of-threat-intelligence-lifecycle/

[36] BlueVoyant. (n.d.). Cyber Threat Intelligence (CTI): Definition, Types, Process. *BlueVoyant*. Retrieved October 17, 2024, from https://www.bluevoyant.com/knowledge-center/cyber-threat-intelligence-cti-definition-types-process

[37] Recorded Future. (n.d.). *Recorded Future*. Retrieved October 17, 2024, from https://www.recordedfuture.com/

[38] ThreatConnect. (n.d.). *ThreatConnect*. Retrieved October 17, 2024, from https://threatconnect.com/

[39] Splunk. (n.d.). *Splunk*. Retrieved October 17, 2024, from https://www.splunk.com/

[40] IBM. (n.d.). *IBM QRadar*. Retrieved October 17, 2024, from https://www.ibm.com/qradar

[41] Cuckoo Sandbox. (n.d.). *Cuckoo Sandbox*. Retrieved October 17, 2024, from https://cuckoosandbox.org/

[42] VirusTotal. (n.d.). *VirusTotal*. Retrieved October 17, 2024, from https://www.virustotal.com/gui/home/upload

[43] MISP Project. (n.d.). *MISP Threat Sharing*. Retrieved October 17, 2024, from https://www.misp-project.org/

[44] AlienVault. (n.d.). *OTX (Open Threat Exchange)*. Retrieved October 17, 2024, from https://otx.alienvault.com/

[45] Anomali. (n.d.). *ThreatStream*. Retrieved October 17, 2024, from https://www.anomali.com/products/threatstream

[46] Cyber Defense Magazine. (2023). How to Overcome the Most Common Challenges with Threat Intelligence. *Cyber Defense Magazine*. Retrieved October 17, 2024, from https://www.cyberdefensemagazine.com/how-to-overcome-the-most-common-challenges-with-threat-intelligence/

[47] ISACA. (2021). Cyberthreat Intelligence as a Proactive Extension to Incident Response. *ISACA Journal*, *Volume 6*. Retrieved October 17, 2024, from https://www.isaca.org/resources/isaca-journal/issues/2021/volume-6/cyberthreat-intelligence-as-a-proactive-extension-to-incident-response

[48] OASIS. (2020). *STIX™ Version 2.1. OASIS Committee Specification Public Review Draft 01*. Retrieved October 17, 2024, from https://docs.oasis-open.org/cti/stix/v2.1/csprd01/stix-v2.1-csprd01.html

[49] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. , 3613-3618. https://doi.org/10.18653/v1/D19-1371 .

[50] Bayer, M., Kuehn, P., Shanehsaz, R., & Reuter, C. (2022). CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain. *ArXiv*, abs/2212.02974. https://doi.org/10.48550/arXiv.2212.02974 .

[51] Ameri, K., Hempel, M., Sharif, H., Lopez, J., & Perumalla, K. (2021). CyBERT: Cybersecurity Claim Classification by Fine-Tuning the BERT Language Model. *Journal of Cybersecurity and Privacy*. https://doi.org/10.3390/jcp1040031 .

[52] You, Y., Jiang, J., Jiang, Z., Yang, P., Liu, B., Feng, H., Wang, X., & Li, N. (2021). TIM: Threat context-enhanced TTP intelligence mining on unstructured threat data. *Cybersecurity*, 4(1), 1-19. https://doi.org/10.1186/s42400-021-00106-5

[53] Benayas, A., Sicilia, M. A., & Mora-Cantallops, M. (2023). A Comparative Analysis of Encoder Only and Decoder Only Models in Intent Classification and Sentiment Analysis: Navigating the Trade-Offs in Model Size and Performance. *ResearchGate*. https://www.researchgate.net/publication/377467915

[54] Alam, M., Bhusal, D., Park, Y., & Rastogi, N. (2022). Looking Beyond IoCs: Automatically Extracting Attack Patterns from External CTI. *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*. https://doi.org/10.1145/3607199.3607208 .

[55] Fayyazi, R., Taghdimi, R., & Yang, S. J. (2024). Advancing TTP Analysis: Harnessing the Power of Large Language Models with Retrieval Augmented Generation. *arXiv*. https://arxiv.org/abs/2401.00280

[56] Siracusano, G., Sanvito, D., González, R., Srinivasan, M., Kamatchi, S., Takahashi, W., Kawakita, M., Kakumaru, T., & Bifulco, R. (2023). Time for aCTIon: Automated Analysis of Cyber Threat Intelligence in the Wild. *ArXiv*, abs/2307.10214. https://doi.org/10.48550/arXiv.2307.10214.

[57] Jin, Y., Jang, E., Cui, J., Chung, J., Lee, Y., & Shin, S. (2023). DarkBERT: A Language Model for the Dark Side of the Internet. *ArXiv*, abs/2305.08596. https://doi.org/10.48550/arXiv.2305.08596.

[58] Fieblinger, R., Alam, M. T., & Rastogi, N. (2024). Actionable Cyber Threat Intelligence using Knowledge Graphs and Large Language Models. *arXiv*. https://arxiv.org/pdf/2407.02528