

Anàlisi Arrel IA: Ciberseguretat Pràctica

Arnau Mata Melià

Nom del Programa
Seguretat Informàtica

Tutor/a de TF

Gerard Farràs Ballabriga

**Professor/a responsable de
l'assignatura**

Andreu Pere Isern Deyà

Universitat Oberta
de Catalunya

9 de Gener de 2024



Aquesta obra està subjecta a una llicència de [Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

C) Copyright

© Arnau Mata Melià

Reservats tots els drets. Està prohibit la reproducció total o parcial d'aquesta obra per qualsevol mitjà o procediment, compresos la impressió, la reprografia, el microfilm, el tractament informàtic o qualsevol altre sistema, així com la distribució d'exemplars mitjançant lloguer i préstec, sense l'autorització escrita de l'autor o dels límits que autoritzi la Llei de Propietat Intel·lectual.

FITXA DEL TREBALL FINAL

Títol del treball:	<i>Anàlisi Arrel IA: Ciberseguretat Pràctica</i>
Nom de l'autor:	<i>Arnau Mata Melià</i>
Nom del consultor/a:	<i>Gerard Farràs Ballabriga</i>
Nom del PRA:	<i>Andreu Pere Isern Deyà</i>
Data de lliurament (mm/aaaa):	<i>01/2024</i>
Titulació o programa:	<i>Grau d'Enginyeria Informàtica</i>
Àrea del Treball Final:	<i>Seguretat informàtica</i>
Idioma del treball:	<i>Català</i>
Paraules clau	<i>Ciberseguretat, Intel·ligència Artificial, Resposta a Incidents</i>

Resum del Treball

Aquest treball presenta una exploració i aplicació pràctica dels Grans Models Lingüístics (LLM) en l'àmbit de la ciberseguretat, amb un enfocament dirigit a l'anàlisi de la causa arrel dels incidents de ciberseguretat. Els LLM, coneguts per la seva habilitat per processar i generar text a partir de grans conjunts de dades, són utilitzats en aquest context per millorar la comprensió del context en el que succeeixen els incidents de ciberseguretat. En el projecte, destaca la utilització de tècniques de transferència d'aprenentatge per adaptar models preentrenats a conjunts de dades específics. El treball també incorpora el desenvolupament d'una solució tecnològica completa, incloent una aplicació Python amb APIs FastAPI i una interfície de presentació web Dash. A més, es descriu la implementació de pràctiques d'MLOps per mitjà del desplegament de les aplicacions i les seves dependències a través d'un entorn virtual en python, facilitant així el desplegament i la portabilitat de la solució. Aquest estudi representa una contribució significativa a la pràctica de la ciberseguretat, demostrant l'aplicabilitat dels LLM en la detecció i anàlisi d'incidents, així com també s'evidencia la millora dels LLM a través de la integració eficaç de models d'aprenentatge automàtic.

Abstract

Using a methodological framework that integrates Large Language Models (LLMs) with Reinforcement Learning (RL) supervision (Question-Answer), a solution has been developed that includes APIs for interacting with the models and for training using Python language data. Despite technical challenges and time constraints in the design and execution of the project, which limited the acquisition of real datasets for intensive training of the models and their subsequent processing, promising results have been achieved. It has been observed that supervising the LLM model with an RL model enhances the accuracy of LLM responses, largely due to the RL model training based on rewards for correct answers. The containerization of the solution in a Docker image, which includes a small server that launches the API endpoints, facilitates the interaction and continuous training of the models in a local environment, allowing for

future improvements such as automated validation of training data and programmatic generation of training files. The findings suggest that establishing a learning cycle for both models in a real network environment, training with real incidents, could significantly improve the system's response capability. This research paves the way for future studies where the dynamic integration of LLMs and RL in real cybersecurity environments could be key for a more accurate and effective analysis of the root causes of security incidents.

Índex

1.	Introducció	1
1.1.	Context i justificació del Treball.....	2
1.2.	Objectius del Treball	4
1.3.	Impacte en sostenibilitat, ètic-social i de diversitat	7
1.4.	Enfocament i mètode seguit	8
1.5.	Planificació del Treball.....	9
1.6.	Breu sumari de productes obtinguts	9
1.7.	Breu descripció dels altres capítols de la memòria	10
2.	Materials i mètodes	12
2.1.	Planificació i gestió del projecte	12
2.2.	Entorn de treball i configuració	12
2.3.	Recerca de fonts de dades	17
2.3.1.	Estratègies de recerca	18
2.3.2.	Construcció dels jocs de dades.....	19
2.3.3.	Construcció dels jocs de dades per el model d'aprenentatge per reforç....	22
2.4.	Elecció del model.....	23
2.5.	Desenvolupament de codi	26
2.6.	Execució del model (interacció)	48
2.7.	Aplicació FastAPI (backend).....	48
2.8.	Aplicació Dash (capa de presentació).....	52
2.9.	Funcionament de la Interfície, Logs i Monitorització.....	53
3.	Resultats.....	56
4.	Conclusions i treballs futurs	58
5.	Glossari.....	60
6.	Bibliografia	64
7.	Annexos.....	66
7.1	Annex 1: Manual d'administració	66
7.2	Annex 2: Manual d'instruccions per generar una imatge en Docker.....	73
7.3	Annex 3: Recursos de suport al treball	78

Llista de figures

Figura 1 - Exercicis CFT de cyber defensa.	3
Figura 2 - Grans fabricants de models LLM generalistes [12].	3
Figura 3 - Enfocament del treball a realitzar [3].	4
Figura 4 - Taulell Kanban del projecte (fet amb l'eina Trello).	8
Figura 5 - Planificació del projecte (Gantt).	9
Figura 6 - Repositori públic del projecte GitHub (codi font).	16
Figura 7 - Repositori GitHub del projecte NLP Gym (utilitzat en el projecte).	16
Figura 8 - Informe de l'anàlisi d'un incident originat per IcedID info stealer (https://malware-traffic-analysis.net [11]).	21
Figura 9 – Inicialització del model LaMini-LM, en el que es basa el model LLM seleccionat [15].	26
Figura 10 – Principals fases del projecte.	27
Figura 11 - Codi comentat de com s'ha entrenat el model LLM (CyberVigilant).	28
Figura 12 - Codi modificat amb adaptacions.	29
Figura 13 - Primera interacció amb el model (pregunta i resposta obtinguda).	30
Figura 14 - Interacció amb el model LLM.	30
Figura 15 - Interacció amb el model LLM després d'una ronda de <i>fine-tune</i>	31
Figura 16 - Segona interacció amb el model LLM després d'una ronda de <i>fine-tune</i>	31
Figura 17 - Interacció amb el model LLM després d'un nou procés de <i>fine-tune</i>	32
Figura 18 - Interacció amb el model LLM després de fer el segon procés de <i>fine-tune</i>	32
Figura 19 - Cuadern Jupyter que mostra el procés de càrrega i aprenentatge del model LLM.	34
Figura 20 - Cuadern Jupyter amb interacció amb el model LLM.	35
Figura 21 - Cuadern Jupyter amb el procés d'entrenament del model RL.	39
Figura 22 - Procés complet d'entrenament i emmagatzematge en local del model RL.	44
Figura 23 - Procés d'entrenament del model RL.	47
Figura 24 - Interacció amb la combinació de models RL i LLM (CyberVigilant).	47
Figura 25 - Segona interacció amb la combinació de models RL i LLM (CyberVigilant).	47
Figura 26 - Interacció inicial amb el model LLM.	48
Figura 27 - Estructura del codi font del projecte (1).	50
Figura 28 - Estructura del codi font del projecte (2).	51
Figura 29 - Interfície web de l'aplicació.	53
Figura 30 - Interfície web de l'aplicació amb una resposta del model.	54

Figura 31 - Interfície web de l'aplicació amb dues respostes.	54
Figura 32 - Codi font mostrant la construcció de les 2 pistes per el model RL [https://github.com/rajcscw/NLPGym].	55
Figura 33 - Repositori public del projecte a GitHub.	67
Figura 34 - Terminal d'execució de comandes python.	67
Figura 35 - Terminal d'execució de comandes python (aixecant servidor uvicorn).	68
Figura 36 - Interfície web de l'aplicació.	68
Figura 37 - Estructura de fitxers a l'espai Google Drive compartit amb membres de la UOC.	69
Figura 38 - Interfície web de l'aplicació.	70
Figura 39 - Contingut carpeta de logs de l'aplicació.	71
Figura 40 - Espai a Hugging Face amb el model LLM entrenat en aquest projecte.	71
Figura 41 - Imatge Docker carregada.	75
Figura 42 - Contenidor Docker en execució.	76
Figura 43 - Monitorització dels recursos consumits per el contenidor Docker.	76
Figura 44 - Logs d'execució de l'aplicació en el contenidor Docker.	77

Llista de taules

Taula 1 - Joc de dades per entrenar el model.....	79
Taula 2 - Joc de dades per entrenar el model en termes de ciberseguretat i d'anàlisi de xarxes.....	86
Taula 3 - Joc de dades per entrenar a fer Anàlisi de Causa Arrel.	87

1. Introducció

Aquest treball explora i aplica Grans Models Lingüístics (LLM en endavant per les seves sigles en anglès) en l'àmbit de la ciberseguretat, específicament en l'anàlisi de causes arrel d'incidents de ciberseguretat [2][3].

Els LLM són construccions neurals de gran escala, dissenyades per capturar intricades dependències lingüístiques i semàntiques a partir de vastes quantitats de text. Gràcies a la seva capacitat de generalització, els LLM poden generar, completar i comprendre textos en diversos contextos, des de tasques simples fins a aplicacions més complexes.

Els LLM s'han establert com a eines poderoses per a la comprensió i generació de text basat en grans conjunts de dades. S'han dut a terme investigacions que han evidenciat la seva eficàcia en tasques relacionades en l'àmbit de la ciberseguretat, com per exemple en processos de detecció de correus electrònics maliciosos o en la identificació de comportaments anòmals en textos [2] [3].

Un dels reptes més rellevants que representa la gran mida d'aquests models es mesura en termes de tenir una gran càrrega computacional. Tot i que molts LLM estan dissenyats i entrenats en clústers de GPU (Unitats de Processament Gràfic per les seves sigles en anglès) o TPU (Unitats de Processament Tensor per les seves sigles en anglès, tecnologia de Google utilitzada per accelerar la IA) de gran escala, la seva execució per inferència, és a dir, generació o comprensió de text, pot ser realitzada en ordinadors personals sota certes condicions en quant a recursos disponibles de maquinari, gestió de memòria i fent ús de models reduïts de codi obert.

Els LLM de gran format, com és el cas de ChatGPT, han estat prèviament entrenats, inicialment era pràcticament impossible adaptar-los per a casos d'ús específics com es busca amb l'execució d'aquest treball, que no té com a objectiu identificar la causa arrel en incidents de ciberseguretat. En l'actualitat, OpenAI ha obert noves eines per a poder entrenar models GPT-4 en temes específics, tot i que segueixen sense ser de codi obert ni es poden utilitzar en entorns aïllats. En el cas del present treball, es pot fer servir la tècnica de "*transfer learning*" (transferència d'aprenentatge en Català), que consisteix a seleccionar un model preentrenat en un gran conjunt de dades i afinar-lo o adaptar-lo a un conjunt de dades més petit i específic, aquest model tindrà un nombre molt menor de paràmetres i serà de codi obert. Els passos involucrats en aquest procés són els següents:

1. Preprocessament de dades: Consisteix a convertir les dades, en aquest cas, logs de sistemes afectats, en un format que el model pugui interpretar, això són generalment vectors.
2. Càrrega del model preentrenat: Importar un LLM de codi obert i que compleixi les condicions per a ser executat en un ordinador personal.
3. Afina ment: Entrenar el model amb el conjunt de dades específic, ajustant només les últimes capes del model o l'arquitectura necessària, mantenint la majoria de pesos originals.

4. Validació i prova: Avaluar el model especialitzat en un conjunt de prova per garantir la seva precisió i eficàcia.

Amb aquest enfocament, es pot aprofitar la capacitat dels LLMs per capturar complexitats lingüístiques i adaptar-les a tasques específiques, com ara l'anàlisi de ciberseguretat, tot optimitzant la seva execució en recursos d'ordinadors personals. Addicionalment, un dels punts clau del projecte ha estat la identificació d'un model d'aprenentatge reforçat (RL per les seves sigles en anglès) que s'ha desplegat entre l'usuari i el model LLM seleccionat, aquest segon model disposa de mètodes per millorar la qualitat de les respostes a través de la selecció, d'entre un conjunt de diferents opcions proporcionades, de la que més s'ajusta als paràmetres amb els que s'ha entrenat aquest segon model. Aquest enfocament és, sense dubte, un punt innovador i ha comportat un alt grau de dedicació per a poder ser integrat en una única solució, l'assoliment d'aquesta fita ha estat molt rellevant i permet obrir noves vies d'investigació cap a la optimització d'eines interactives que permeten un alt grau d'especialització.

Aquest treball, a més de la seva contribució teòrica en l'ús dels LLM en ciberseguretat, ha integrat pràcticament aquestes capacitats en una solució tecnològica completa. Això s'assoleix amb el desenvolupament d'una aplicació en el llenguatge *Python* que aixeca dos servidors en un mateix port de l'equip en el que es desplega, un amb les *API RESTful* (Application Programming Interface), que permeten la interacció amb els models LLM i RL, així com un altre servei que ofereix la capa de presentació web, que facilita als usuaris la interacció amb els models d'una forma eficaç i senzilla.

Per a reforçar la viabilitat i la facilitat de desplegament d'aquesta solució, es va implementar una estratègia de *MLOps* [9], específicament en la construcció d'una imatge de *Docker*. les operacions d'aprenentatge automàtic (*MLOps* com són conegudes en anglès) fa referència a pràctiques d'enginyeria per la producció i desplegament automatitzat de models d'aprenentatge automàtic. Aquestes pràctiques representen una extensió de les operacions de desenvolupament [4] (*DevOps* com es coneixen en anglès), que busca augmentar l'agilitat en l'entrega d'aplicacions i serveis, millorant la col·laboració entre equips de desenvolupament i operacions [4]. L'adopció d'*MLOps* [9] es veu com a essencial per a la implementació efectiva i escalable dels models d'aprenentatge automàtic en diversos entorns [5].

1.1. Context i justificació del Treball

L'avanç tecnològic experimentat en les últimes dècades ha portat a un augment significatiu de la connectivitat i la digitalització, fet que ha resultat en que la ciberseguretat sigui una prioritat per a organitzacions i individus. En aquest context, la necessitat d'identificar i respondre ràpidament als incidents de ciberseguretat és crucial, la detecció eficaç de causes arrel pot prevenir futurs incidents i minimitzar l'impacte de certs cyber atacs. En aquest marc, existeixen eines potents i modernes com per exemple els exercicis de Capture-The-Flag (CTF) que permeten formar professionals en la identificació i mitigació d'atacs. No obstant això, amb l'aparició de tecnologies com els Large Language Models (LLM), s'obren noves vies de capacitació de professionals en matèria de ciberseguretat, un sector en el que els professionals amb més coneixement estan altament demandats i disposen de molt poc temps per transmetre el seu coneixement a les persones que es van incorporant al mercat laboral, en aquest direcció, es pot disposar d'eines com la que s'ha construït en el present treball per a que puguin entrenar models que, al seu temps, podran transmetre

el coneixement a noves generacions de professionals, així com també serviran com a eines que els permeti interactuar amb l'elevat nombre de solucions tecnològiques que existeixen en l'actualitat i que segueix creixent a través d'un nexa comú, la integració dels registres que generen i la seva interpretació en el context de resposta a incidents de seguretat.



Figura 1 - Exercicis CFT de cyber defensa.

Els grans LLM actuals, com ara OpenAI ChatGPT, Google Bard i Microsoft Bing, són models neurals entrenats amb enormes conjunts de dades amb l'objectiu de comprendre i generar llenguatge humà. La seva capacitat per processar, interpretar i generar text basat en aquesta comprensió els fa candidats ideals per ser utilitzats en una àmplia varietat de tasques, incloent la ciberseguretat. De fet, hi ha evidències que suggereixen que els LLM poden ser eficaços en tasques específiques relacionades amb la ciberseguretat, com ara la identificació de correus electrònics maliciosos o la detecció de comportaments anòmals en textos.

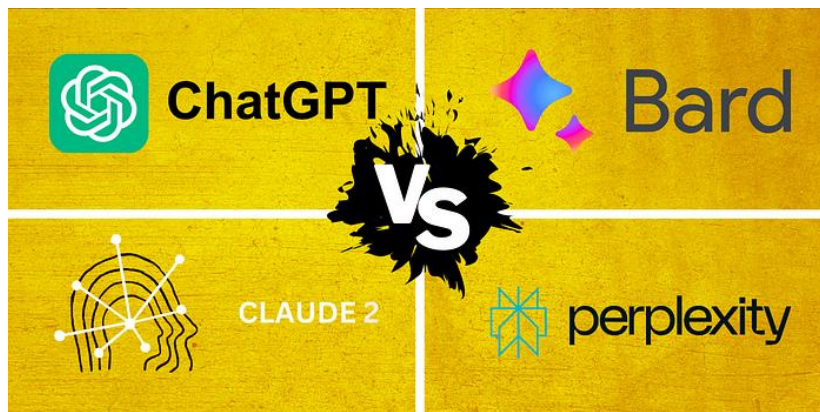


Figura 2 - Grans fabricants de models LLM generalistes [12].

Amb aquesta eficàcia, també sorgeixen reptes, especialment quan es tracta d'entrenar i adaptar aquests models a tasques específiques. Per exemple, si bé models com els d'OpenAI estan tancats i no poden ser re-entrenats, hi ha altres models, com els que són de codi lliure, que permeten l'execució i entrenament en local. Aquests models poden ser adaptats a dades específiques mitjançant la tècnica de *transfer learning*, on el coneixement prèviament adquirit per el LLM es pot transferir a una nova tasca.

En aquest marc, el desenvolupament d'aplicacions en Python emergeix com una peça clau degut a l'elevat nombre de llibreries disponibles per interactuar amb els models prèviament esmentats, juntament amb la possibilitat de crear aplicacions seguint paradigmes de programació avançats i separar l'arquitectura en models de tres capes (presentació, lògica i accés a dades). Aquestes aplicacions, en definitiva, aporten una major flexibilitat i eficiència en la gestió de les sol·licituds i en la comunicació entre els usuaris o altres aplicacions amb els models de LLM i de RL. Aquest enfocament s'integra perfectament en el paradigma d'MLOps, on la dockerització de les aplicacions juga un rol fonamental. La creació d'imatges Docker per a aquestes aplicacions facilita el seu desplegament i escalabilitat, assegurant una implementació consistent i eficient en diferents entorns. Això permet una millor col·laboració entre els equips de desenvolupament i operacions, així com una major agilitat en el llançament i actualització de les aplicacions.

1.2. Objectius del Treball

Amb la creixent rellevància dels Large Language Models (LLM) en l'àmbit de la ciberseguretat, aquest treball es proposa aprofundir en la seva aplicabilitat específica per a l'anàlisi de causes arrel d'incidents de seguretat.

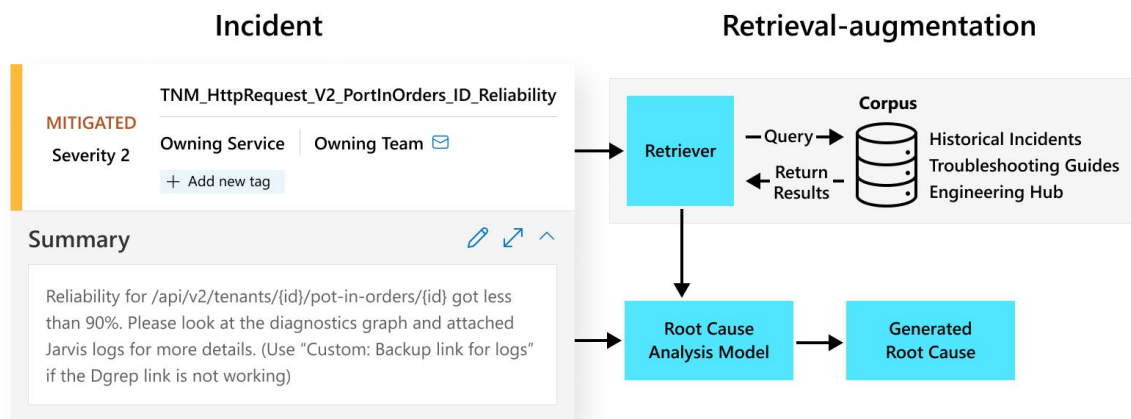


Figura 3 - Enfocament del treball a realitzar [3].

A continuació, es detallen els objectius específics:

- Entrenament d'un LLM amb dades reals d'incidents de ciberseguretat a través del transfer learning:
 - Recollida i preprocessament de dades: S'han recollit dades d'incidents de ciberseguretat (proporcionades per l'equip Unit42 de Paloalto i analitzades a

malware, que posteriorment seran preprocessades per assegurar-se que estiguin en un format que el model LLM pugui interpretar.

- Selecció del model LLM de codi obert: a diferència d'altres models com els d'OpenAI, ens permet fer fine-tune i execució en local. Es descarrega des de Hugging Face el model *MBZUAI/LaLaMini-T5-738M* [15].
- Aplicació de transfer learning: El transfer learning permet aprofitar el coneixement adquirit pel model en tasques prèvies per adaptar-se a una nova tasca amb menys dades. En aquest cas, "LaMini-T5-783M" serà re-entrenat amb les dades específiques d'incidents de ciberseguretat.
 - Carregar el model pre-entrenat "LaMini".
 - Definir els nous paràmetres i outputs específics de la tasca de detecció de causes arrel.
 - Re-entrenar el model amb les dades preprocessades, ajustant-lo a les necessitats concretes del projecte.
- Desenvolupament d'un *MLOps* per la contenidorització i automatització del model adaptat:
 - Definició del pipeline *MLOps* [9]: Permet fer les tasques d'entrenament, validació, desplegament i monitorització del model de forma automatitzada i escalable. Això s'ha assolit a través de les API desplegades en el servidor local d'aplicacions.
 - Contenedorització del model: Aquesta etapa assegura que el model pugui ser desplegat en qualsevol entorn sense dependències externes, facilitant la seva portabilitat i integració en diferents sistemes.
- Avaluació exhaustiva del model en contextos reals:
 - Definició d'escenaris de prova: Es seleccionen o defineixen situacions reals en què es vulgui detectar la causa arrel d'incidents de ciberseguretat.
 - Proves i validació: Es posarà a prova el model en aquests escenaris, comparant la seva eficàcia i precisió amb altres mètodes tradicionals.
 - Anàlisi de resultats: S'analitzaran els resultats obtinguts, identificant àrees de millora i validant la utilitat pràctica del model adaptat.
- Desenvolupament d'una aplicació en Python utilitzant FastAPI per presentar quatre *endpoints* (consumibles des d'altres aplicacions):
 - Creació d'un *endpoint* per fer preguntes directes al model LLM.
 - Desenvolupament d'un *endpoint* que permeti enviar preguntes amb dues pistes al model LLM (format esperat per part del model de RL).

- Implementació d'un *endpoint* per a l'entrenament del model LLM a partir d'un fitxer en format *csv* carregat prèviament.
- Establiment d'un *endpoint* per a l'entrenament del model RL a través d'un fitxer amb format *jsonl* carregat prèviament.
- Desenvolupament d'una aplicació frontal amb Dash, accessible des d'un navegador web:
 - Creació d'una interfície web que permeti als usuaris interactuar amb els quatre API endpoints desenvolupats.
- Contenedorització i Dockerització del projecte dins un entorn MLOps [9]:
 - Construcció d'una imatge Docker que inclogui tot el projecte, inclòs els models, per assegurar la interoperabilitat amb altres sistemes.

L'objectiu final és combinar la potència dels LLM i els RL amb les avantatges del *MLOps* [9], creant una solució robusta i eficient per a l'anàlisi de causes arrel en incidents de ciberseguretat [10].

1.3. Impacte en sostenibilitat, ètic-social i de diversitat

L'aplicació de LLM en l'àmbit de la ciberseguretat, tot i que pot tenir un impacte positiu en termes de sostenibilitat i eficiència, també planteja reptes ètics i socials que preocupen a la comunitat acadèmica. En aquest context, és de vital importància alinear el desenvolupament i l'aplicació d'aquestes tecnologies amb els valors i principis ètics establerts per institucions com la UOC.

Sostenibilitat: Aquest treball contribueix a l'Objectiu de Desenvolupament Sostenible (ODS) de les Nacions Unides número 9, "Indústria, innovació i infraestructura", promovent la construcció d'infraestructures resistents i fomentant la innovació. També es relaciona amb l'ODS 4, "Educació de qualitat", ja que busca proporcionar eines i recursos per millorar la formació en ciberseguretat.

Ètic-social: La capacitat d'aquests models per interpretar, generar i potencialment manipular informació pot ser utilitzada de manera no ètica. És crucial que la comunitat acadèmica i professional sigui conscient d'aquests reptes i treballi per establir pautes i normatives que assegurin un ús responsable i ètic d'aquestes eines. Aquesta preocupació està alineada amb la missió de la UOC, que busca compartir coneixement en favor del desenvolupament de les persones i la millora de la societat.

Diversitat: És essencial assegurar-se que els LLM no perpetuen biaixos ni discriminacions presents en les dades amb les quals s'entrenen. La diversitat i la inclusió han de ser considerades en tots els aspectes del desenvolupament i aplicació d'aquests models, des de la recopilació de dades fins a la interpretació dels resultats.

1.4. Enfocament i mètode seguit

Per a la realització d'aquest Treball Final de Grau, s'utilitzarà la metodologia *Kanban*, metodologia d'origen japonès, va ser inicialment desenvolupat per l'empresa del sector automobilístic Toyota durant la dècada dels anys 40 com a sistema per optimitzar la producció en la seva cadena de muntatge. El terme "Kanban" es tradueix com "targeta visual" o "taulell" i es basa en la gestió de la informació visual per millorar l'eficiència dels processos de treball.

La metodologia *Kanban* es basa en un taulell dividit en diferents columnes que representen les fases del procés de treball. Les tasques es representen amb targetes que es mouen de columna a columna a mesura que avancen en el procés. Tot això resulta en una visualització clara del treball, adaptabilitat a canvis o imprevistos, optimització de fluxos de treball ja que es poden identificar ràpidament colls d'ampolla i l'autogestió ja que permet i facilita la organització de les tasques.

Com a eina que permeti la implementació d'aquesta metodologia de seguiment i gestió de les tasques del projecte, s'ha escollit *Trello*, una plataforma en línia que facilita la creació de taulells *Kanban* de manera intuïtiva. Disposa d'una interfície amigable i de funcionalitats, com ara etiquetes, llistes de comprovació i dates límit, que el fan idoni per a la gestió individual d'un projecte com aquest. La principal raó per utilitzar *Trello* en aquest context és la seva capacitat per implementar taulells *Kanban*, permetent una organització visual de les tasques i una gestió eficient del projecte.

A continuació es mostra el taulell per el TFG completat.

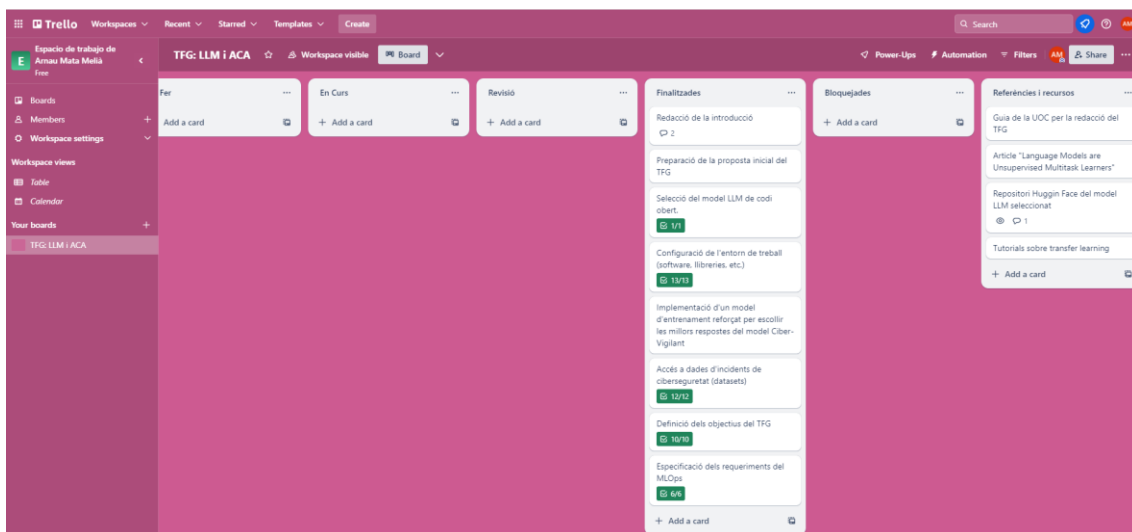


Figura 4 - Taulell Kanban del projecte (fet amb l'eina Trello).

1.5. Planificació del Treball

La realització d'aquest projecte s'ha dividit en diverses fases al llarg de sis PAC, marcades com fites en la planificació. A continuació, es detalla la planificació temporal per a cada PAC:

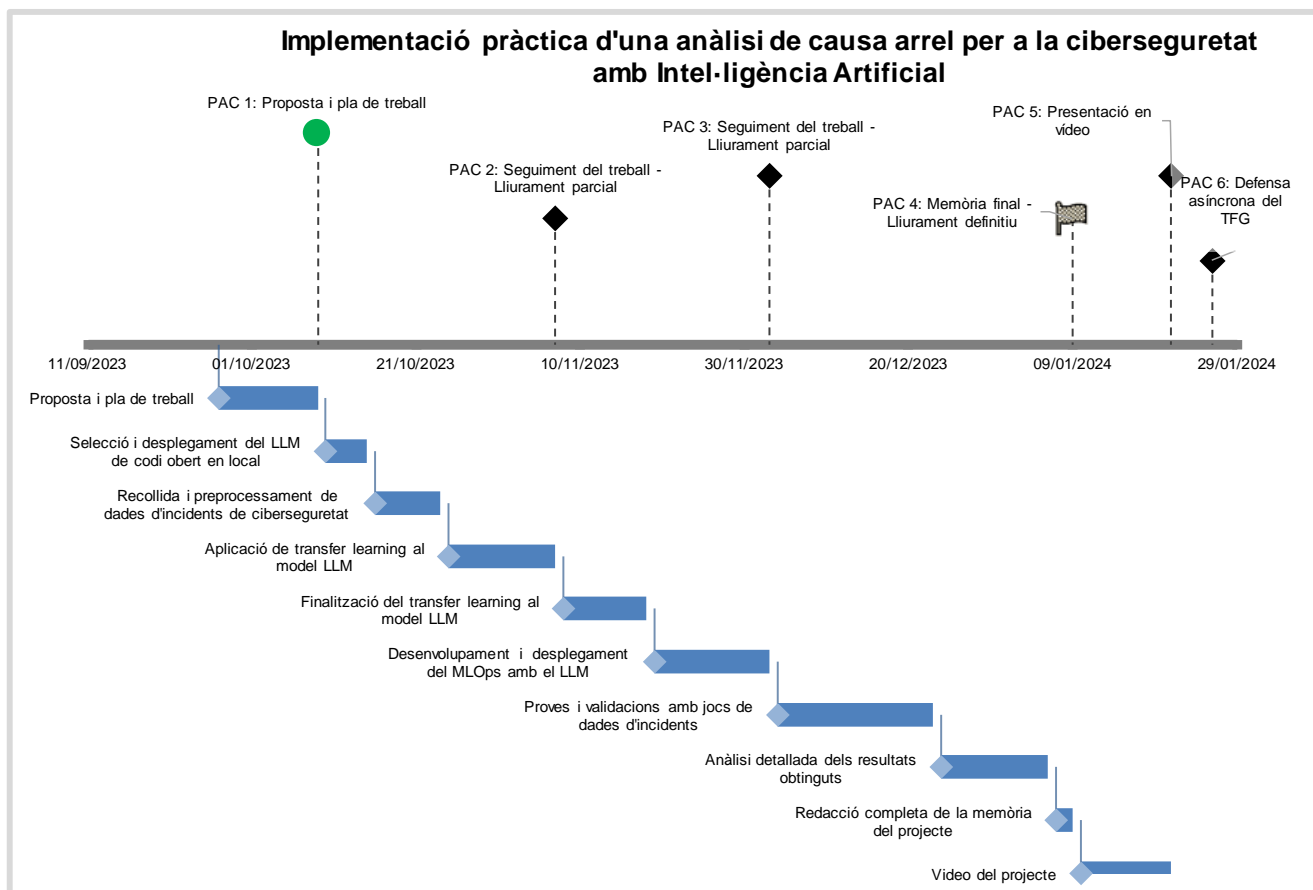


Figura 5 - Planificació del projecte (Gantt).

1.6. Breu sumari de productes obtinguts

En el marc d'aquest treball s'han desenvolupat diversos productes clau, que es detallen a continuació, complint amb els objectius establerts i seguint les directrius de les millors pràctiques en ciberseguretat, desenvolupament de software i MLOps [9]:

1. Model LLM adaptat per a l'anàlisi d'incidents de ciberseguretat: s'ha aconseguit l'entrenament d'un model LLM de codi obert, específicament adaptat per a l'anàlisi de dades d'incidents de ciberseguretat. Aquest model, anomenat "LaMini" [15], ha estat re-entrenat amb dades preprocessades d'incidents reals, ajustant-se específicament a les necessitats d'anàlisi de causes arrel.
2. Aplicació en python fent ús de FastAPI: s'ha creat una aplicació en Python que utilitza FastAPI per oferir quatre *endpoints* diferents que es poden consumir des d'altres aplicacions. Aquests *endpoints* permeten la interacció amb el model

LLM per fer preguntes directes, preguntes amb pistes, entrenar el model LLM i entrenar el model RL, facilitant així una interacció més dinàmica i versàtil amb els models.

3. Aplicació per a la interfície d'usuari desenvolupada amb python i la llibreria Dash, que permet avaluar els productes desenvolupats: s'ha desenvolupat una interfície web d'usuari, utilitzant Dash, amb la que poden interactuar fàcilment amb els *endpoints* de l'API. Aquest frontal proporciona una experiència d'usuari intuïtiva i simplifica l'accés a les funcionalitats del model, així com també estableix pipelines d'automatització de certs processos com l'entrenament dels models.
4. Imatge Docker del projecte: com a part de l'estratègia MLOps [9], s'ha creat una imatge Docker que conté tot el projecte, incloent els models LLM i RL. Això assegura una fàcil distribució, portabilitat i interoperabilitat del projecte en diferents entorns, fent-lo més accessible i fàcil de desplegar.

Aquests productes representen una contribució significativa al camp de la ciberseguretat, demostrant com la integració dels LLM i les pràctiques d'MLOps [9] poden millorar l'anàlisi de causes arrel en incidents de seguretat. La combinació d'aquests elements ofereix una solució potent i flexible, preparada per afrontar els desafiaments actuals i futurs en la detecció i gestió d'incidents de ciberseguretat.

1.7. Breu descripció dels altres capítols de la memòria

Materials i mètodes: Aquest capítol es centra en la descripció detallada dels materials i mètodes utilitzats en el projecte. Inclou la planificació i gestió del projecte, la configuració de l'entorn de treball amb Anaconda, i l'ús de llibreries específiques com HuggingFace, PyTorch, Gymnasium [1], NLP_Gym [1] i stable_baselines3. També s'aborda la recerca de fonts de dades, la construcció de jocs de dades, l'elecció del model LLM, i el desenvolupament de codi per al *scrapping* i transformació de dades, així com l'entrenament i la interacció amb el model.

Resultats: Aquesta secció presenta els resultats obtinguts de l'aplicació del model LLM en l'anàlisi de causes arrel en incidents de ciberseguretat. Es detallen els resultats de l'entrenament del model, així com la seva eficàcia en contextos reals.

Conclusions i treballs futurs: S'ofereix una anàlisi crítica dels resultats obtinguts i es discuteixen les conclusions principals del projecte. A més, es proposen línies de treball futur que podrien ampliar i millorar la recerca realitzada.

Glossari: Aquest capítol proporciona definicions clares i concises dels termes tècnics utilitzats al llarg de la memòria, facilitant la comprensió del contingut.

Bibliografia: Inclou totes les referències bibliogràfiques utilitzades en la redacció de la memòria, oferint una base sòlida i fiable per a la investigació realitzada.

Annexos: Conté informació addicional com el manual d'usuari per desplegar i utilitzar l'aplicació, el seu codi font i altres materials que complementen la memòria i proporcionen més detalls sobre els processos i resultats del projecte.

En els capítols de "Materials i Mètodes" i "Resultats", es detalla el desenvolupament d'una aplicació integral en Python fent ús de FastAPI, que inclou quatre endpoints per interactuar amb els models LLM i RL. A més, s'explica el desenvolupament d'un frontend amb Dash que permet a l'usuari accedir a una web per interactuar amb aquests endpoints. També es descriu l'arrencada del servidor Uvicorn per a l'execució de l'aplicació FastAPI i la dockerització de tot el projecte, assegurant la seva interoperabilitat i facilitant el seu desplegament dins del marc del MLOps [9].

2. Materials i mètodes

2.1. Planificació i gestió del projecte

Com es pot veure en la [Figura 4](#) i la [Figura 5](#), totes les tasques del projecte s'han finalitzat.

2.2. Entorn de treball i configuració

Un dels principals objectius del treball és aconseguir individualitzar mitjançant tècniques d'entrenament, un model de llenguatge LLM en un ordinador convencional, amb l'objectiu d'apropar les grans capacitats de la Intel·ligència Artificial a persones, institucions o fins i tot empreses amb recursos molt limitats. En concret, aquest projecte s'està desenvolupant i executant íntegrament en un ordinador adquirit l'any 2014 amb les següents especificacions:

- Processador Intel i7-4770K
- RAM 16GB
- Disc dur mecànic de 2TB
- GPU Nvidia 1050

Com a sistema operatiu, s'ha utilitzat Windows 10. El projecte s'ha desenvolupat íntegrament sota l'entorn de treball Anaconda, des d'on es pot gestionar tots els recursos necessaris.

▪ **Anaconda**

Sobre l'equip descrit al punt anterior, inicialment s'ha instal·lat Anaconda, una plataforma de gestió de paquets i entorns de treball per a ciència de dades en Python. Això proporciona un marc de treball robust per gestionar llibreries i dependències, i facilita l'execució de diversos projectes en entorns aïllats.

Des d'Anaconda, s'ha utilitzat tant la pròpia interfície gràfica disponibles en la instal·lació per defecte sobre Windows, així com també l'eina CMD.exe que incorpora l'eina per a gestionar els paquets necessaris per desenvolupar el codi necessari en python i que s'explica tot seguit, així com els quaderns Jupyter per a gestionar i executar el codi i els scripts en python.

▪ **Visual Studio Code i entorn virtual**

Passada la fase d'interacció i fine-tuning del model LLM, s'ha optat per utilitzar Visual Studio Code (VSC) com a entorn de desenvolupament principal, substituint l'ús anterior d'Anaconda i els quaderns Jupyter. Visual Studio Code és un editor de codi font potent i versàtil que suporta una àmplia varietat de llenguatges de programació i ofereix un gran conjunt d'eines i extensions per a desenvolupadors.

La gestió del projecte s'ha realitzat a través d'un repositori a GitHub, el qual s'ha clonat en local per treballar directament des de VSC. Això ha permès una millor integració amb el control de versions i ha facilitat la col·laboració i el seguiment de canvis en el codi.

Per a la gestió de dependències i llibreries, s'ha establert un entorn virtual dins de VSC. Els entorns virtuals en Python són una pràctica estàndard per crear espais aïllats on es poden instal·lar llibreries específiques per a cada projecte, evitant així conflictes entre dependències i garantint la consistència dels entorns de desenvolupament. Aquesta aproximació facilita també el desplegament i la possibilitat de replicar els projectes en diferents màquines o entorns, tot i que tot plegat s'ha acabat desplegant en una imatge Docker.

Dins de VSC, s'ha fet ús del terminal PowerShell integrat per executar comandes per a la gestió dels entorns virtuals, la instal·lació de llibreries i l'execució dels scripts de Python. Aquesta integració del terminal dins de l'entorn de desenvolupament proporciona una major eficiència i fluïdesa en el treball, permetent executar i provar el codi de manera ràpida i còmoda.

En resum, la transició a VSC i l'ús d'entorns virtuals, combinats amb la gestió de codi a través de GitHub, han suposat una millora significativa en el desenvolupament del projecte, aportant major flexibilitat, control i eficiència en el procés de desenvolupament de software.

▪ **Docker**

Una altra peça clau en l'estructura de treball d'aquest projecte ha estat l'ús de Docker, una plataforma de contenidorització que permet empaquetar aplicacions i les seves dependències en un format estàndard anomenat contenidor. Aquest enfocament garanteix que el programari funcioni de manera fiable i consistent en qualsevol entorn.

Per a aquest projecte, s'ha creat un *Dockerfile*, que és un script de configuració utilitzat per construir imatges Docker personalitzades. Aquest Dockerfile inclou tots els components necessaris del projecte: les aplicacions en Python, els models LLM i RL, així com totes les llibreries i dependències requerides. Aquesta imatge es pot utilitzar per llançar contenidors Docker que executaran el projecte de manera aïllada i consistent en qualsevol entorn, sigui en un servidor local, en el núvol, o en un entorn de producció.

L'ús de Docker aporta diversos beneficis significatius al projecte. Primerament, simplifica el procés de desplegament, ja que tots els requisits i dependències del projecte s'empaqueten en una sola imatge. Això elimina els problemes comuns relacionats amb "aquesta aplicació funciona en la meua màquina, però no en l'entorn de producció". A més, Docker facilita l'escalabilitat i la distribució del projecte, permetent que es desplegui ràpidament en diversos entorns o escali segons les necessitats.

Per executar el projecte, l'ús de Docker és opcional, tot i que aquest garanteix que l'entorn de treball romangui consistent i que l'aplicació operi segons l'esperat, independentment de les variacions locals en la configuració del maquinari o del programari, també és possible carregar el projecte en un entorn virtual de python i aixecar el servidor des de la pròpia línia de comandaments. Docker s'integra dins del

marc de MLOps [9], facilitant la gestió, el desplegament i l'operació dels models de manera eficaç i fiable.

▪ **Llibreries HuggingFace i PyTorch**

Per tal de disposar de les eines necessàries a l'entorn de treball anteriorment descrit i d'utilitzar el potencial de les llibreries desenvolupades sobre python per a fer projectes relacionats amb les ciències de dades, com és el nostre cas, s'ha fet ús de dos grans paquets que posteriorment ens han permès invocar i utilitzar llibreries concretes per cada cas d'ús que s'han especificat en el projecte.

HuggingFace: plataforma de codi obert que ofereix eines per a la intel·ligència artificial i el processament del llenguatge natural. Hugging Face Inc., una empresa amb seu a Nova York (Estats Units d'Amèrica) promou i manté la plataforma. Els usuaris poden accedir a una gran varietat de models pre-entrenats, així com a eines per a la creació de models personalitzats. Addicionalment, es poden compartir els models afinats i col·laborar amb altres membres de la comunitat. En aquest treball en concret, es fan servir les següents llibreries:

- Per a l'accés i gestió de models i datasets en la plataforma Hugging Face:
 - Huggingface_hub.
- Per a l'entrenament del model:
 - Transformers [6][13]: es tracta d'una API que permet accedir a models pre-entrenats per a la intel·ligència artificial i el processament del llenguatge natural com és el cas de LaLaMini-T5-738M [15]. En el treball s'utilitzen per descarregar el mode, ajustar-lo als conjunts de dades específics d'anàlisi de causa arrel d'incidents de seguretat i utilitzar-los per la inferència. Les classes que s'utilitzen són:
 - Trainer: proporciona eines per entrenar models de llenguatge natural, s'utilitza per entrenar el model personalitzat i ajustar-lo al nostre cas d'ús concret.
 - TrainingArguments: proporciona eines per personalitzar els paràmetres d'entrenament dels models de llenguatge natural, amb ells s'ajusten els híper paràmetres del model per personalitzar el procés d'entrenament.
 - AutoTokenizer: proporciona eines per a la tokenització automàtica de textos, el que ens facilita la conversió de textos en seqüències de tokens que el nostre model de llenguatge pot processar.
 - T5ForConditionalGeneration: proporciona eines per a la generació de textos condicionals amb el model T5, s'utilitza per processar dades i preparar-les per a l'entrenament del nostre model.

- Per a l'execució del model:
 - Transformers [6] [13]. En aquest cas, s'utilitzen les següents classes:
 - Pipeline: es tracta d'una API dissenyada per a utilitzar models pre-entrenats per a la inferència, en el nostre cas sobre un model de generació de text basat en T5.
- Per a la recollida de dades (scrappers):
 - OS: facilita la interacció amb el sistema operatiu sobre el que es treballa, facilita les operacions de lectura de fitxers i accés a directoris, així com gestionar excepcions en l'ús de fitxers i el seu emmagatzematge.
 - Requests: permet enviar sol·licituds http/https a través de python.
 - BeautifulSoup: proporciona eines per analitzar i manipular documents html i xml, permet extreure dades de pàgines web, manipular fitxers xml.
 - Urljoin: funció que permet crear URL absolutes a partir de relatives i de les base.
 - Urlparse: permet analitzar i manipular URLs.

PyTorch – es tracta d'un marc de treball enfocat a treballar amb xarxes neuronals, els models com el que utilitzo en aquest treball el necessiten per poder executar el codi que porten per darrere.

- **Llibreria Gymnasium, NLPGym [1] i stable_baselines3**

Gymnasium: Una biblioteca per a la creació d'entorns d'aprenentatge per reforç, basada en l'original 'gym' d'OpenAI [1]. Derivat de l'original 'Gym' d'OpenAI, és una biblioteca que ha portat avenços significatius en l'àmbit de l'aprenentatge per reforç. Aquesta biblioteca ofereix entorns estandarditzats per a l'entrenament d'algoritmes d'aprenentatge per reforç, el que facilita el desenvolupament i la comparació de diferents estratègies d'aprenentatge.

NLPGym: Una extensió de la biblioteca Gym [1] específicament dissenyada per a l'aprenentatge automàtic en el camp del processament del llenguatge natural (NLP). Aquesta biblioteca facilita la implementació i l'avaluació d'algoritmes en tasques com la comprensió de textos, el *question answering* (QA), i altres aplicacions del processament de llenguatge natural (NLP). Aquesta biblioteca s'ha personalitzat ja que la seva versió oficial té una dependència amb una llibreria obsoleta, amb el que s'ha procedit a habilitar un *fork* a github amb el canvi corresponent per a poder treballar amb ella, l'error ha estat notificat als desenvolupadors, que no han contestat a la demanda a data de redacció del treball. A continuació es pot veure la modificació, s'ha afegit la llibreria "scikit-learn" per solucionar el problema.

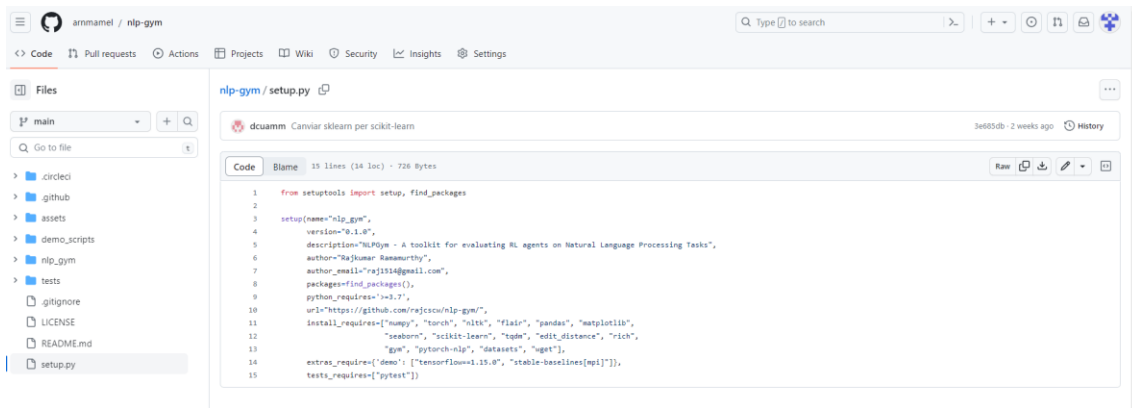


Figura 6 - Repositori públic del projecte GitHub (codi font).

En la següent imatge es pot veure el cas reportat a l'equip desenvolupador:

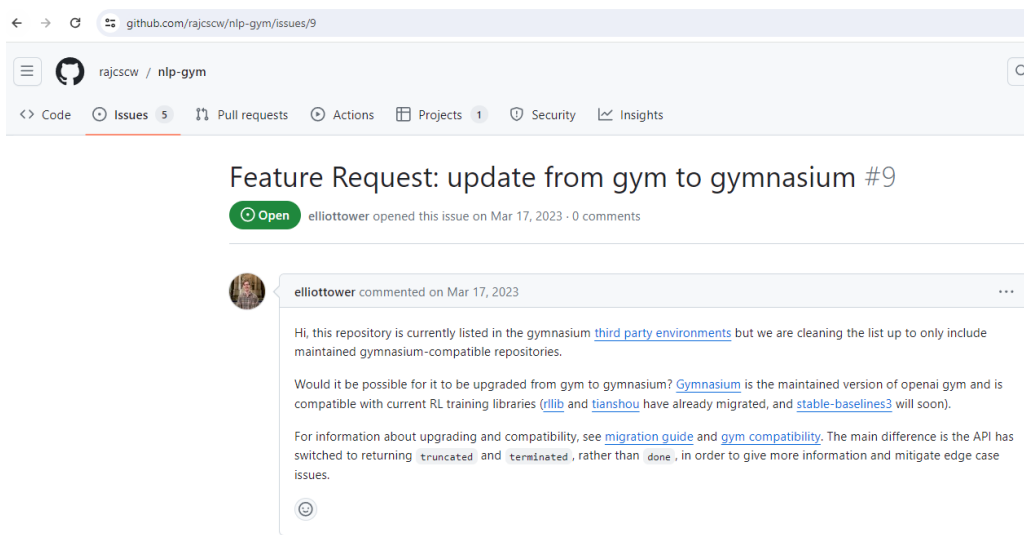


Figura 7 - Repositori GitHub del projecte NLP Gym (utilitzat en el projecte).

Shimmy: És la llibreria que ofereix la solució per a que NLP Gym [1] pugui interactuar eficaçment amb Gymnasium d'OpenAI, en lloc del Gym [1] inclòs per defecte. "Shimmy" actua com a capa pont, permetent que dues biblioteques o components de software que originalment no estaven dissenyades per treballar juntes, ho puguin fer de manera eficient. Aquesta capacitat d'interoperabilitat és crucial en projectes d'aprenentatge automàtic i NLP, ja que permet combinar diferents eines i marcs de treball per aconseguir els millors resultats, especialment en l'àmbit de la seguretat cibernètica on la flexibilitat i la integració de múltiples eines són clau.

Stable_baselines3: Potent biblioteca per a l'aprenentatge per reforç que ofereix implementacions de primera qualitat d'alguns dels més avançats algorismes en aquest camp, com ara *Q-Learning*, *Soft Actor-Critic (SAC)* i *Proximal Policy Optimization (PPO)*. La veritable "màgia" d'aquesta biblioteca rau en la seva capacitat per implementar polítiques sofisticades que determinen com un agent ha d'actuar en un determinat entorn. Això es tradueix en una major eficàcia i precisió en l'entrenament dels models.

Aquestes llibreries són fonamentals per a la implementació de tècniques avançades d'aprenentatge automàtic i enforteixen la capacitat del projecte per abordar problemes complexos en el camp de la ciberseguretat.

- **Llibreria FastAPI**

Biblioteca que permet fer desenvolupament ràpid en el procés de creació d'APIs amb versions de Python a partir de la 3.7. Es basa en estàndards com OpenAPI i JSON schema. És essencial en aquest projecte per a la construcció d'unes API robustes, eficients i fàcilment escalables. S'ha utilitzat per desenvolupar quatre endpoints API específics: (1) pregunta al model LLM, (2) pregunta amb dues pistes al model LLM, (3) entrenament del model LLM, i (4) entrenament del model RL. Aquests endpoints permeten la interacció dinàmica amb els models LLM i RL, facilitant l'anàlisi de dades i la implementació d'algoritmes d'aprenentatge per reforç en el camp de la ciberseguretat.

- **Llibreria Dash**

Biblioteca dedicada a la creació d'aplicacions web analítiques en Python, s'utilitza en aquest projecte per desenvolupar un frontend web interactiu que permet als usuaris accedir i utilitzar els *endpoints* API creats amb FastAPI. Aquesta interfície proporciona una manera intuïtiva i accessible per als usuaris d'interactuar amb el sistema, realitzar consultes i veure els resultats de les anàlisis efectuades pels models LLM i RL.

- **Llibreries Uvicorn**

Uvicorn és un servidor ASGI (*Asynchronous Server Gateway Interface*) lleuger i de gran rendiment per a Python 3, s'utilitza com a servidor per executar l'aplicació FastAPI. Aquesta elecció es deu a la seva rapidesa i eficiència en el maneig de sol·licituds asincròniques, essencial per a operacions ràpides d'API.

Ademés, Uvicorn facilita la creació de rutes i *handlers* d'aplicacions asincròniques, que són fonamentals per a la presentació dels servidors d'API i del frontal web en un únic port a través del servidor local. Aquesta integració ofereix una solució simplificada i eficient per a la gestió de sol·licituds web i la interacció amb els models LLM i RL.

2.3. Recerca de fonts de dades

En el camp de la seguretat informàtica, la recerca de fonts de dades fiables i actualitzades és de vital importància per a poder realitzar les anàlisis i comprendre els incidents de seguretat. Aquest procés implica una cerca exhaustiva i meticulosa per identificar bases de dades que continguin informació detallada i rellevant sobre amenaces recents, tàctiques, tècniques i procediments (TTPs) d'actors malintencionats [2], així com informes d'incidentes i anàlisis de malware, el terme de TTP prové del marc de treball Mitre Att&ck.

- **El marc de treball ATT&CK i Mitre**

ATT&CK, desenvolupat i mantingut per Mitre, és un marc de treball àmpliament reconegut que categoritza i descriu els diversos TTPs utilitzats per actors malintencionats en l'àmbit de la ciberseguretat. Mitre, una organització sense ànim de lucre que opera centres de recerca i desenvolupament finançats pel govern dels Estats

Units d'Amèrica, ha estat un actor clau en el desenvolupament d'eines i marcs per a la millora de la seguretat informàtica. El seu marc ATT&CK serveix com una base de coneixement que detalla les múltiples fases d'un cyber atac, dotant als analistes d'eines per obtenir una comprensió profunda dels mètodes emprats pels atacants.

▪ **La Importància del context en els incidents de seguretat**

Comprendre el context en el qual es produeix un incident de seguretat és vital, la informació aïllada sobre un atac o una vulnerabilitat, sense un enteniment del context més ampli, pot ser enganyosa o insuficient per a una identificació precisa del incident. Aquesta comprensió contextual inclou no només els detalls tècnics de l'atac, sinó també la comprensió dels objectius, estratègies i patrons de comportament dels atacants. Per tant, la recerca de fonts de dades ha de ser orientada no només a recopilar informació detallada, sinó també a entendre el panorama general i la manera en què cada peça d'informació s'integra en el mateix.

Tenint en compte aquest enfocament, es detallen a continuació les estratègies de recerca utilitzades, així com a descriure les fonts de dades seleccionades i la metodologia d'accés i anàlisi de les mateixes.

2.3.1. Estratègies de recerca

Aquest apartat es centra en la recerca i la identificació de conjunts de dades en ciberseguretat i, sobretot, dades generades a partir d'investigacions dutes a terme en resposta a incidents de seguretat, aquest punt ha començat amb una revisió de literatura acadèmica, on destaquen les següents publicacions de rellevància:

- Ferrag, M. A. (2023, 25 junio). *Revolutionizing cyber threat detection with large language models*. arXiv.org. <https://arxiv.org/abs/2306.14263> [2]
- Hughes, A. (2023, 11 mayo). *Large-language models for automatic cloud incident Management - Microsoft Research*. Microsoft Research. <https://www.microsoft.com/en-us/research/blog/large-language-models-for-automatic-cloud-incident-management/> [3]
- *Scrapers with code - language models are unsupervised multitask learners*. (2019, 14 febrero). <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask> [6]
- Quentin, L., et al. (2021). *Datasets: A Community Library for Natural Language Processing*. <https://arxiv.org/abs/2109.02846> [7]
- Raffin, A. (2021). *Stable-Baselines3: Reliable Reinforcement learning Implementations*. <https://jmlr.org/papers/v22/20-1364.html> [8]
- Rodriguez, A. (2019). *MLOps: Toward Effective Implementation of Machine Learning Models*. International Conference on Software Engineering Proceedings. [9]
- Sessions, N. (2023, December 14). *Bolstering cybersecurity: How large language models and Generative AI are Transforming Digital Security*. NVIDIA Technical

Blog. <https://developer.nvidia.com/blog/bolstering-cybersecurity-how-large-language-models-and-generative-ai-are-transforming-digital-security/> [10]

- Web: YourHub4Tech. (2023, 3 novembre). *Comparing the top generative AI tools! - illumination's mirror - Medium*. <https://medium.com/illuminations-mirror/comparing-the-top-generative-ai-tools-5fea222b603e> [12]
- Wolf et al., EMNLP (2020). *Transformers: State-of-the-Art Natural Language Processing*. <https://aclanthology.org/2020.emnlp-demos.6> [13]
- Wu, M., Waheed, A., Zhang, C., Abdul-Mageed, M., & Aji, A. F. (2023). LAMini-LM: a diverse herd of distilled models from Large-Scale instructions. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2304.14402> [14]

A diferència de la majoria dels treballs que s'han trobat, en el cas del present projecte, les tècniques utilitzades han estat diferents, ja que el busquen és, per un costat, minar dades rellevants que ens permetin entrenar el nostre model LLM i per un altre, que aquesta informació, juntament amb tots els jocs d'instruccions amb què s'ha pre-entrenat el model i les característiques que vagi assolint ajudin a poder interactuar amb aquest model en format de xat per obtenir informació rellevant d'incidents de seguretat.

La investigació de l'equip de Microsoft [3], d'altra banda, està dirigida exactament en la mateixa línia cap a la que ho està el present projecte, però en el seu cas enfocant-se a l'anàlisi d'incidents de seguretat que afectin sistemes al núvol, estratègia que segueixen per integrar capacitats d'intel·ligència artificial generativa en la seva solució d'automatització de la resposta a incidents (copilot). De totes formes, l'objectiu del present projecte pretén una eina més que permeti als analistes de seguretat descarregar una de les funcions que més feina manual suposa, la de l'anàlisi i unificació de causes arrel d'un incident de seguretat. En certs casos serà molt útil disposar d'aquesta eina, els més obvis són aquells escenaris que no puguin ser coberts per eines tradicionals que es basen en conèixer exactament el comportament de les amenaces, aquí és on la intel·ligència artificial pot aportar més valor, construint formes d'identificar amenaces abans no hagin estat analitzades íntegrament, això sí, sempre i quan se li hagi fet un procés d'entrenament significatiu en situacions reals, amb informació de tot el context en el que es produeixen els incidents de seguretat.

2.3.2. Construcció dels jocs de dades

Per a comprendre la construcció dels jocs de dades per entrenar el model LLM local, és important abans tenir clars els trets característics d'aquests tipus de models i del seu funcionament intern [7]:

- Eficiència i *fine-tune*: els models LLM pre-entrenats, com LaLaMini-T5-738M [15], requereixen menys interaccions per al *fine-tuning* comparat amb els models d'aprenentatge automàtic tradicionals. Això es deu a que ja han après una gran quantitat de característiques generals del llenguatge i només necessiten ajustaments específics per a tasques concretes.
- Comparació amb l'aprenentatge automàtic tradicional: En l'aprenentatge automàtic convencional, sovint es requereix un entrenament més extens des de l'inici, el que pot ser més intensiu en termes de temps i recursos

computacionals. Aquest no és el cas dels models LLM com el que s'utilitza en el treball.

D'altra banda, hi ha un factor que té impacte sobre la selecció i construcció dels jocs de dades per a l'entrenament del model LLM, les limitacions de l'entrenament en un ordinador personal amb pocs recursos computacionals. Aquest fet es tradueix en un entrenament molt lent, en el cas d'aquest treball la mitja de temps que tarda l'entrenament és d'entre 20 minuts i unes hores per línia dins el fitxer que s'envia a l'entrenament (en format csv), aquest punt suposa una limitació a l'hora d'entrenar el model.

Després d'una avaluació inicial, es van seleccionar diferents fonts principals per a aquest estudi: NIST, Recorded Future (sandbox d'anàlisi de malware) i malware-traffic-analysis.net [11].

- Conceptes bàsics de ciberseguretat i anàlisi de tràfic de xarxa: glossari de termes de la NIST proporciona una base sòlida i estandarditzada de conceptes i terminologia en ciberseguretat (<https://csrc.nist.gov/glossary>).
- Anàlisi d'Incidents de Seguretat: Els resums d'incidents de seguretat, com els disponibles a la web *malware traffic analysis*, ofereixen dades reals i pràctiques. Convertir fitxers pcap en csv utilitzant l'eina tshark que forma part de la instal·lació de Wireshark, juntament amb logs d'endpoints i eines de seguretat, proporciona un ric conjunt de dades per a l'entrenament (<https://malware-traffic-analysis.net> [11]). A continuació es mostra un exemple d'infecció analitzada en aquesta web:

```
malware-traffic-analysis.net/2023/07/25/index.html
2023-07-25 (TUESDAY): ICEDID (BOKBOT) INFECTION FROM WAVE OF MALSPAM ON MONDAY 2023-07-24

REFERENCES:
- https://twitter.com/k3dg3/status/1683544196341219341
- https://twitter.com/JAMESWT_MHT/status/1683747418641252360
- https://twitter.com/Tac_Mangusta/status/1683741137784147969
- https://bazaar.abuse.ch/browse/tag/1561373935/
- https://tria.ge/s/campaign:1561373935

INFECTION CHAIN:
- email --> PDF attachment --> link from PDF --> TDS redirect --> zip download --> run extracted EXE --> IcedID infection

ASSOCIATED FILES:
- SHA256 hash: 8715a27a8a263f59335601258c584963cc50822d2c24ba81a9d1b33b48753a70
- File size: 86,558 bytes
- File name: INV-Details-JUL23.pdf
- File type: PDF document, version 1.4, 1 pages
- File description: example of PDF file attachment

- SHA256 hash: fed2651d0e6ef1a2bb6283dba0c52a24a90b3b69ae2b8f08b883dcc6d8ecf4a1
- File size: 185,994 bytes
- File name: INV-Details-JUL2023(14).zip
- File type: Zip archive data, at least v2.0 to extract, compression method=deflate
- File description: example of password-protected zip archive
- Password: 724

- SHA256 hash: 2cf9bef62939a9b645d43d05959e181f3962e7f10c9e4d73288c7f36af7d5ec2
- File size: 458,416 bytes
- File name: INV-Details-JUL2023(14).exe
- File type: PE32+ executable (GUI) x86-64 (stripped to external PDB), for MS Windows
- File description: extracted from the above zip archive, installer EXE for IcedID

- SHA256 hash: 729a9743319f548134a1535291bea0bfdac6cd9a8c4c9ba19ab1b8524d5e8ea7
- File size: 1,036,133 bytes
- File location: hxxp://filitaferamoza[.]com/
- File type: gzip compressed data, was "Pledge.txt", from FAT filesystem (MS-DOS, OS/2, NT), original size modulo 2^32 2501970
- File description: gzip binary used to create persistent IcedID DLL and license.dat data binary

- SHA256 hash: 332afc80371187881ef9a6f80e5c244b44af746b20342b8722f7b56b61604953
- File size: 354,474 bytes
- File location: C:\Users\user\AppData\Roaming\DisplayBronze\license.dat
- File type: data
- File description: data binary needed to run persistent IcedID DLL

- SHA256 hash: 110ca2ce6308f2731162fa5acf0dd8a89ff22d4c545702c0118c1bccf43e9bee
- File size: 680,904 bytes
- File location: C:\Users\user\AppData\Roaming\user\{45613AD5-6C95-8231-C86F-BF8027DF7176}\Afebwc.dll
- File type: PE32+ executable (DLL) (GUI) x86-64, for MS Windows
- File description: Persistent DLL for IcedID infection
- Run method: rundll32 [filename],init --abim="[path to license.dat]"

LINK FROM PDF:
- 173.232.146[.]10 port 80 - 173.232.146[.]10 - GET /

TDS REDIRECT FOR ZIP DOWNLOAD:
- 109.234.165[.]110 port 443 - hxxps://www.sophrologie-yoga[.]fr/palatial/

INSTALLER EXE CALLS FOR GZIP BINARY:
- 128.199.3[.]164 port 80 - filitaferamoza[.]com - GET /

ICEDID HTTPS C2:
- 2.56.177[.]122 port 443 - magizanoqomo[.]com - HTTPS traffic
- 2.56.177[.]122 port 443 - magiketchinn[.]com - HTTPS traffic
- 140.99.221[.]138 port 443 - lohmotarufos[.]com - HTTPS traffic
- 140.99.221[.]138 port 443 - flarkonafaero[.]com - HTTPS traffic

BACKCONNECT TRAFFIC/KEYHOLE VNC IP AND PORT:
- 165.232.175[.]216 port 443 - encoded/encrypted traffic (no Keyhole VNC during initial pcap)

NOTE:
- I started up the Windows host the next day and saw also saw Keyhole VNC traffic from the continued IcedID infection.
- I've added a pcap of this infection traffic with Keyhole VNC: 2023-07-26-IcedID-with-Keyhole-VNC-traffic.pcap.zip
```

Figura 8 - Informe de l'anàlisi d'un incident originat per IcedID infostealer (<https://malware-traffic-analysis.net> [11]).

Per accedir a les dades d'aquestes fonts, es va procedir a crear comptes d'usuari, quan va ser necessari, i a familiaritzar-se amb les seves interfícies i formats de dades. En el cas de tria.ge, es van utilitzar les seves eines d'exportació de dades per obtenir informació detallada sobre mostres específiques de malware a través de l'endpoint a la seva API, al que es va obtenir accés mitjançant sol·licitud. Per a malware-traffic-analysis.net [11], es van descarregar les captures de tràfic i se'n va fer una anàlisi forense utilitzant eines com *tshark* seguint els informes detallats disponibles en la pròpia web.

Aquestes tasques s'han automatitzat en gran mesura utilitzant codi desenvolupat en python i les llibreries que ja s'han referenciat al punt [2.2](#) del present document. De totes formes, tractant-se d'un punt més dins el marc del projecte, no ha estat possible utilitzar totes les fonts de dades descarregades per entrenar els models.

2.3.3. Construcció dels jocs de dades per el model d'aprenentatge per reforç

El format de fitxer *jsonl* (*JSON Lines*) és essencial per a la manipulació eficient de grans volums de dades estructurades en l'aprenentatge automàtic [5]. Aquest format emmagatzema objectes JSON en línies separades, facilitant la seva lectura i processament per part dels models. Per a la construcció del joc de dades del nostre model de RL CyberVigilant, s'ha utilitzat aquest format per representar preguntes, respostes i fets relacionats amb més de 10 incidents de seguretat diferents. Aquesta estructura permet no només una càrrega eficient de les dades sinó també un entrenament més efectiu del model d'aprenentatge per reforç.

Les dades utilitzades per a aquests jocs de dades provenen d'incidents reals de seguretat analitzats i documentats en fonts de confiança, que ja es detallen en la secció [2.3.3](#) del present document. Cada incident és desglossat en preguntes específiques amb diverses opcions de resposta, permetent al model comprendre millor els patrons i les dinàmiques dins del camp de la ciberseguretat [10]. Aquest enfocament no només millora la capacitat predictiva del model sinó que també aporta una base sòlida per a futures investigacions en l'àrea.

Aquesta expansió proporciona una visió més detallada del procés d'entrenament del model i la seva aplicació pràctica en l'àmbit de la seguretat informàtica, a continuació es mostra una línia d'exemple del fitxer generat.

```
{
  "id": "M1",
  "question": {
    "stem": "What was the initial attack vector in the DEV-0569 activity on 2023-02-03?",
    "choices": [
      {"text": "Email phishing", "label": "A"},
      {"text": "Malicious Google Ad", "label": "B"},
      {"text": "USB device", "label": "C"},
      {"text": "Network intrusion", "label": "D"},
      {"text": "Website vulnerability", "label": "E"}
    ]
  }
}
```

```
]
},
"answerKey": "B",
"fact1": "The DEV-0569 activity began with a malicious Google Ad.",
"fact2": "Google Ads are a common vector for cyber attacks.",
"combinedfact": "The DEV-0569 malware infection initiated through a malicious Google Ad.",
"formatted_question": "What was the initial attack vector in the DEV-0569 activity on 2023-02-03? (A) Email phishing (B) Malicious Google Ad (C) USB device (D) Network intrusion (E) Website vulnerability"
}
```

Figura 9 - Format objecte JSON entrenament RL.

2.4. Elecció del model

▪ Revisió de l'article publicat respecte el model LaMini-T5-783M

En les referències, al final del document, es pot trobar l'enllaç a l'article *LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions* [14] que descriu el treball dut a terme per crear el model de llenguatge natural que s'ha seleccionat per fer el present projecte, a continuació es detalla el procés i l'argumentari que s'ha seguit.

▪ Contextualització de la necessitat de model LLM

La tria d'un model de processament de llenguatge natural adequat en el camp de l'anàlisi d'incidents de seguretat informàtica ha estat un aspecte fonamental. Per un costat, el model seleccionat ha de tenir la capacitat de gestionar grans volums de dades que tinguin en compte diferents àmbits, assegurant l'eficiència tant en la detecció de patrons com en l'extracció d'informació rellevant.

La selecció d'un model de llenguatge de gran magnitud (LLM) de codi obert per a l'afinament i el seu ús en l'entorn operatiu per analitzar la causa arrel dels incidents de ciberseguretat, basat en les dades contextuais obtingudes de fonts fiables, ofereix diversos avantatges. A continuació, es presenten algunes raons clau per les quals aquests models poden ser particularment útils en aquest context. En aquest sentit, l'ús de models LLM de codi obert en el camp de la ciberseguretat pot oferir avantatges significatius en termes de detecció i reparació de vulnerabilitats, anàlisi de dades d'intel·ligència, la gestió de la privacitat i la millora de processos d'anàlisi d'incidents de seguretat, porció que s'ha explorat en més detall en el present projecte.

▪ Criteris de selecció: execució en local i limitacions de maquinari

Un dels principals criteris de selecció per al model LaMini-T5-738M han estat les limitacions de maquinari en l'entorn on s'implementarà. Considerant que l'entrenament i l'execució del model es realitzen en un ordinador local sense targetes gràfiques de gran potència, la LaMini-T5-738M emergeix com una opció ideal.

Els models com LaMini-T5-738M estan dissenyats per ser més lleugers i eficients que les versions completes dels LLMs, permetent la seva execució en ordinadors amb menys recursos de processament i memòria.

Aquests models poden seguir el procés d'afinament localment, adaptant-se a necessitats específiques, incloent l'anàlisi de dades de ciberseguretat, tot i que poden requerir alguna optimització per minimitzar l'ús de recursos.

La capacitat d'operar sense connexió a Internet és crucial en entorns de baixos recursos o amb limitacions de connectivitat. Això permet processar dades sensibles o crítiques sense necessitat de connectar-se a servidors externs.

▪ **Capacitats i aplicacions del model**

L'article descriu detalladament les capacitats de la LaMini-T5-738M, incloent la seva eficiència en el processament i la seva adaptabilitat a diferents tasques de processament de llenguatge natural. Es fa especial èmfasi en la seva idoneïtat per a l'anàlisi de dades de seguretat, incloent la identificació de TTPs (Tàctiques, Tècniques i Protocols) i l'avaluació d'informes d'incidents o de tràfic de xarxa en general.

- Processament de Llenguatge Natural (*NLP*, que és com es coneix per les sigles en anglès): els LLMs com LaMini-T5-738M poden analitzar i interpretar textos relacionats amb ciberseguretat, incloent informes d'incidents, logs i comunicacions.
- Identificació de tendències i patrons: poden detectar patrons en dades de ciberseguretat, ajudant a identificar les causes arrels d'incidents de seguretat.
- Suport en la presa de decisions: proporcionen anàlisi de dades que poden ser útils per a la presa de decisions en resposta a incidents de ciberseguretat.
- Capacitació i sensibilització en seguretat: aquest model pot ser utilitzat per a entrenar personal en temes de ciberseguretat, a través de la generació automàtica de materials formatius o simulacions d'atacs.

▪ **Comparació amb altres models**

El desenvolupament i la selecció de models d'intel·ligència artificial en el camp de la seguretat informàtica requereixen tenir en compte diverses variables, incloent la capacitat de processament, la mida del model, la velocitat d'execució i la precisió dels resultats. En aquest context, l'article realitza una comparació exhaustiva entre la LaMini-T5-738M i altres models que prevalen en el camp.

La LaMini-T5-738M destaca sobre altres models per la seva capacitat d'oferir un bon equilibri entre mida i eficiència. A diferència de models més grans i complexos, com el BERT o GPT-3, la LaMini-T5-738M està específicament optimitzada per funcionar en entorns amb recursos limitats. Això és particularment rellevant per a institucions amb

infraestructures de maquinari menys potents, on l'ús de models més grans podria no ser viable.

Un dels avantatges clau de la LaMini-T5-738M és la seva velocitat d'execució en comparació amb models de major escala. Aquesta velocitat es tradueix en una resposta més ràpida en l'anàlisi de dades de seguretat, un factor crític en situacions on el temps de resposta pot ser molt important. Tot i la seva mida reduïda, el model manté una alta precisió en l'anàlisi de textos, essent capaç d'identificar patrons i anomalies amb una eficàcia comparable a la dels seus homòlegs més grans.

Un altre punt destacat en la comparativa és la capacitat de la LaMini-T5-738M per a ser executada localment, sense la necessitat d'accés a serveis basats en el núvol. Això proporciona una major flexibilitat i seguretat per a les organitzacions que prefereixen o requereixen mantenir les seves dades i processos dins del seu propi entorn de maquinari.

En resum, la LaMini-T5-738M representa una solució ideal per a entorns amb limitacions de maquinari, oferint un equilibri entre mida, velocitat i precisió que la fa especialment atractiva per a la implementació en l'anàlisi d'incidents de seguretat. La seva capacitat per operar eficaçment en entorns locals, juntament amb la seva competència en l'entrega de resultats precisos, la posiciona com una alternativa viable i eficient davant d'altres models de major mida i requisits de recursos.

2.5. Desenvolupament de codi

En aquest apartat es descriu de forma exhaustiva el codi que s'ha utilitzat en el treball, que ha estat desenvolupat íntegrament en python a través de Jupyter Notebooks que es troben a la plataforma local Anaconda.

▪ Descàrrega del model LaMini-T5-738M de Hugging Face

Fent ús de la llibreria *transformers* de Hugging Face en Python, s'ha pogut descarregar els fitxers del model en l'ordinador local, això s'ha aconseguit a través de codi concret implementat per descarregar el model específic LaMini-T5-738M.

▪ *Scrapping* i transformació de dades

En aquesta primera fase del treball, centrada en la recerca del millor model per poder-lo executar en un equip amb recursos limitats, la obtenció de dades ha estat executada principalment de forma manual, tot i que s'han preparat scripts en python per obtenir, de forma automatitzada (fent ús de les llibreries *requests* i *BeautifulSoup*), continguts específics respecte d'anàlisis d'infeccions de malware des de la web de *malware-traffic-analysis.net* [11].

▪ Entrenament del model (fine-tuning)

La següent imatge il·lustra com l'equip investigador que ha desenvolupat el model LLM sobre el que es durà a terme l'entrenament ha realitzat precisament la tasca d'entrenament inicial, que requereix molta més intensitat ja que el model ha de ser capaç de processar, entendre i respondre al llenguatge natural.

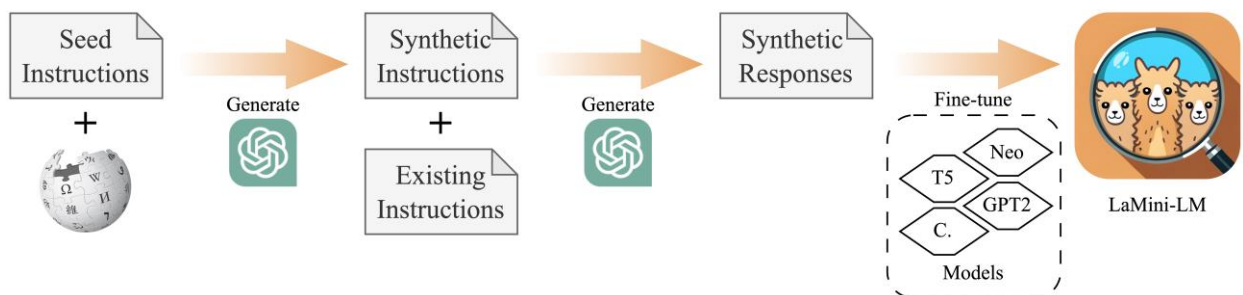


Figura 10 – Inicialització del model LaMini-LM, en el que es basa el model LLM seleccionat [15].

En tot el transcurs del projecte, el model ha estat entrenat en 5 iteracions diferents, es mostren tres dels jocs de dades més significatius ja que és una representació fidedigna del que s'ha seguit executant.

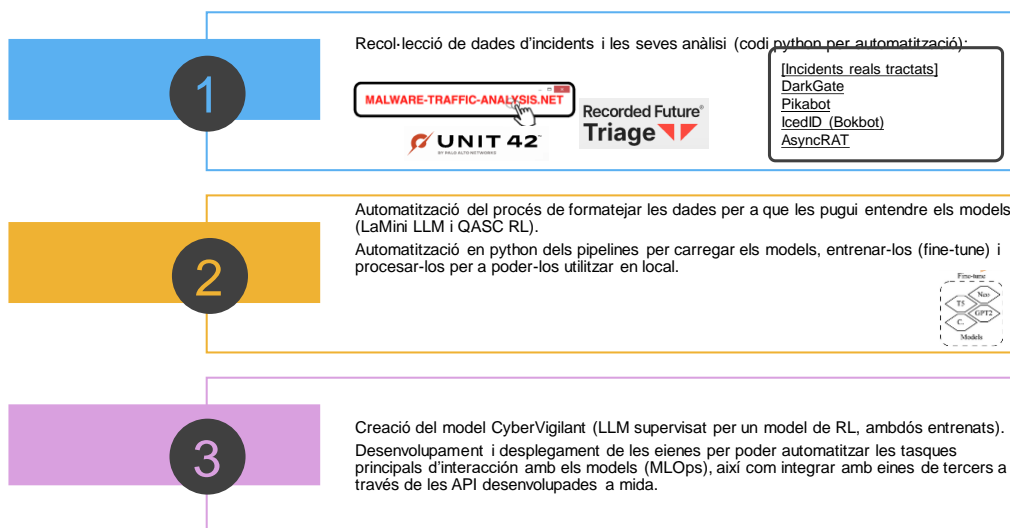


Figura 11 – Principals fases del projecte.

JD1. Incident Redline Stealer, en concret, s'han inclòs les següents dades indicant els loC com a "input" i que es tracta del "RedLine Stealer C2" com a "target", de forma que s'ensenyen al model que aquests objectes corresponen a un tipus d'atac informàtic. Consultar la Taula 1 disponible a [l'Annex 3 del present document](#).

En la primera interacció, s'ha assignat l'anterior text a l'entrada al model i com a sortida s'ha provat de posar un text estàtic als efectes de validar que els mètodes funcionen correctament, no s'ha preparat un joc de dades per a fer-ne la validació, doncs no és part dels objectius del present treball, es fa de forma manual un cop hagi finalitzat el procés d'entrenament, que és llarg i costós. A continuació es mostra el codi en python utilitzat per fer l'entrenament del model, que carrega un fitxer CSV, pre-processa les dades, utilitza un model de Hugging Face per a la generació condicional de text (T5), i entrena el model. Els comentaris expliquen cada pas i la seva funció.

```

import pandas as pd
from datasets import Dataset
from transformers import T5Tokenizer, T5ForConditionalGeneration
from transformers import Trainer, TrainingArguments

# Funció per carregar i processar els fitxers CSV
def carregar_csv(ruta_arxiu_csv, tokenizer):
    # Carrega el fitxer CSV en un DataFrame
    df = pd.read_csv(ruta_arxiu_csv).head(1)

    # Combina les columnes en 'input_text'
    df['input_text'] = df.apply(lambda x: [str(x.values)], axis=1).apply(lambda y: " ".join(y))

    # Defineix el text objectiu com a text estàtic
    df['target_text'] = "Això és un registre de xarxa, hauria de ser més informatiu, però no puc"

    # Preprocessament de les dades
    # En aquest punt es realitzarien els passos de preprocessament necessaris,
    # com ara la tokenització, l'afegiment de tokens especials, la truncació, etc.
    # Per exemple, suposem que el teu CSV té dues columnes: 'input_text' i 'target_text'
    df['input_ids'] = df['input_text'].apply(lambda x: tokenizer.encode(x, truncation=True, padding='max_length', max_length=512))
    df['labels'] = df['target_text'].apply(lambda x: tokenizer.encode(x, truncation=True, padding='max_length', max_length=512))

    # Converteix el DataFrame en un objecte 'Dataset' de Hugging Face
    dataset = Dataset.from_pandas(df)

    # Potser caldrà establir el format per a PyTorch o TensorFlow si s'utilitzen aquests frameworks
    # dataset.set_format(type='torch', columns=['input_ids', 'attention_mask', 'labels'])

    return dataset

# Carrega el tokenizer i el model
tokenizer = T5Tokenizer.from_pretrained('MBZUAI/LaMini-T5-738M')
model = T5ForConditionalGeneration.from_pretrained('MBZUAI/LaMini-T5-738M')

# Prepara les dades (això és una simplificació)
train_dataset = carregar_csv('./train_redlinestealer.csv', tokenizer)

# Defineix els arguments d'entrenament
training_args = TrainingArguments(
    output_dir='/resultats/',
    num_train_epochs=3,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=64,
    warmup_steps=500,
    weight_decay=0.01,
    logging_dir='/registres/',
    logging_steps=10,
)

# Inicialitza l'entrenador
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=None
)

# Entrena el model
trainer.train()

# Desa el model
model.save_pretrained('/model/')

```

Figura 12 - Codi comentat de com s'ha entrenat el model LLM (CyberVigilant).

En la primera execució del codi d'entrenament del model, es va comprovar que els paràmetres no eren els òptims per l'equip en el que s'estava executant, aquest equip va quedar malmès degut a les altes temperatures dels components després de varies hores d'execució del mateix. Per aquest motiu, es van retocar els paràmetres de l'entrenament i es va acabar amb el següent codi.

```

import pandas as pd
from datasets import Dataset
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM, TrainingArguments, Trainer

# Definim el camí on hem guardat el model preentrenat
model_path = './model/'
print("Carregant el tokenizer i el model...")

# Carreguem el tokenizer i el model
tokenizer = AutoTokenizer.from_pretrained('MBZUAI/LaMini-T5-738M')
model = AutoModelForSeq2SeqLM.from_pretrained(model_path)

# Funció per carregar i processar el conjunt de dades
def carregar_i_processar_conjunt_de_dades(ruta_arxiu_csv, tokenizer, max_length=512):
    print(f"Carregant i processant el conjunt de dades des de {ruta_arxiu_csv}...")
    # Carreguem el conjunt de dades
    df = pd.read_csv(ruta_arxiu_csv, delimiter=';')

    # Extraïem el text 'input' i 'target' del dataframe
    text_input = df['input'].tolist()
    text_objectiu = df['target'].tolist()

    # Tokenitzem el text 'input' i 'target'
    encoding = tokenizer(text_input, truncation=True, padding='max_length', max_length=max_length, return_tensors="pt")
    etiquetes = tokenizer(text_objectiu, truncation=True, padding='max_length', max_length=max_length, return_tensors="pt")['input_ids']

    # Creem un conjunt de dades a partir del text tokenitzat
    conjunt_de_dades = Dataset.from_dict({
        'input_ids': encoding['input_ids'],
        'attention_mask': encoding['attention_mask'],
        'labels': etiquetes
    })

    return conjunt_de_dades

# Carreguem els conjunts de dades d'entrenament i avaluació
print("Carregant conjunts de dades d'entrenament i avaluació...")
conjunt_entrenament = carregar_i_processar_conjunt_de_dades('C:\\Users\\arnma\\OneDrive - Universitat Oberta de Catalunya\\TFG\\tr
conjunt_avaluacio = carregar_i_processar_conjunt_de_dades('C:\\Users\\arnma\\OneDrive - Universitat Oberta de Catalunya\\TFG\\tr

# Definim els arguments d'entrenament
print("Definint els arguments d'entrenament...")
arguments_entrenament = TrainingArguments(
    output_dir='./resultats/',
    num_train_epochs=3, # Ajusteu les èpoques segons la vostra observació
    per_device_train_batch_size=4, # Augmenteu la mida del lot segons la capacitat del vostre maquinari
    logging_dir='./registres/',
    logging_steps=10,
    evaluation_strategy="epoch" # Avaluem al final de cada època
)

# Inicialitzem el Trainer
print("Inicialitzant el Trainer...")
entrenador = Trainer(
    model=model,
    args=arguments_entrenament,
    train_dataset=conjunt_entrenament,
    eval_dataset=conjunt_avaluacio
)

# Entrenem i avaluem el model
print("Començant l'entrenament i l'avaluació...")
entrenador.train()

# Desem el model després de l'entrenament
print("Desant el model entrenat...")
model.save_pretrained(model_path)

# Avaluem el model i desem els resultats
print("Avaluant el model...")
resultats_metrics = entrenador.evaluate()
with open('resultats_avaluacio.txt', 'w') as fitxer:
    for clau, valor in resultats_metrics.items():
        fitxer.write(f"{clau}: {valor}\n")

print("L'entrenament i l'avaluació han estat completats.")

```

Figura 13 - Codi modificat amb adaptacions.

JD2. Per afegir nous coneixements al model, es va decidir fer un nou joc de dades amb termes bàsics de ciberseguretat, així com d'anàlisi de tràfic de xarxes i entrenar al model amb aquests per tal que reconegués i disposés d'eines de llenguatge natural específiques de la ciberseguretat. A continuació es mostra el nou joc de dades amb el què s'ha entrenat el model (consultar la Taula 2, que es troba a [l'Annex 3 del present document](#)).

▪ Detalls de l'entrenament del model LLM local

En la primera interacció amb el model original LaLaMini-T5-738M [15], vam fer les següents avaluacions amb els resultats que es detallen.

Pregunta 1	Resposta 1
Analyze the following network traffic: No.,Time,Source,Destination,Protocol,Length,Info 1,0.000000,10.10.31.101,10.10.31.1,DNS,75,Standard query 0x7ec7 A grafielucho.com 2,0.113504,10.10.31.1,10.10.31.101,DNS,107,Standard query response 0x7ec7 A grafielucho.com A 104.21.32.6 A 172.67.182.27 3,0.117203,10.10.31.101,grafielucho.com,TCP,66,56108 > 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM 4,0.161532,grafielucho.com,10.10.31.101,TCP,66,80 > 56108 [SYN, ACK] Seq=0 Ack=1 Win=64240 Len=0 MSS=1380 SACK_PERM WS=8192 5,0.161684,10.10.31.101,grafielucho.com,TCP,60,56108 > 80 [ACK] Seq=1 Ack=1 Win=131072 Len=0 6,0.161896,10.10.31.101,grafielucho.com,HTTP,361,GET / HTTP/1.1 Conduct a Root Cause Analysis based on this log	The given network traffic is a traffic log which includes information about the user's network server configurations and IP addresses. Specifically, it shows the IP addresses, server addresses, DNS records, and message information. The log also provides a number of error messages along with trace back to IP addresses. It also shows the error messages and trace back to a specific session on the server.

Figura 14 - Primera interacció amb el model (pregunta i resposta obtinguda).

Aquest resultat és d'esperar, el model no té coneixements específics sobre incidents de seguretat i no pot trobar encara els problemes que es poden trobar en les comunicacions que s'han capturat, en aquest cas es tracta de tràfic de xarxa sintètic i no seria fàcil tampoc per un analista de seguretat identificar la interacció com un incident de seguretat.

Pregunta 2	Resposta 2
Analyze the following network traffic log: SIEM System: "CyberGuard SIEM Pro" Network Devices Reporting: "Cisco NGFW, Cisco Catalyst 9300 Switch, Cisco ISR 4000 Router" Network Segment: "Corporate VLAN 10 - Finance Department" Associated User: "bob.finance@company.com - Last login: 2023-10-15 14:55:00 from 192.168.1.15" Recent Network Changes: "New VPN policy applied at 2023-10-14 18:30:00" Suspicious Events: "5 failed login attempts on 192.168.1.15 at 2023-10-15 15:00:00" Traffic Details: Timestamp Source IP Destination IP Protocol Port Payload ----- 2023-10-15 15:10:01 192.168.1.15 195.123.45.67 TCP 4444 "Encoded Payload Detected" 2023-10-15 15:10:02 195.123.45.67 192.168.1.15 TCP 4444 "Response with encoded data" Network Behavior: "Unusual outbound traffic to an external IP on a non-standard port" Other Observations: "The external IP 195.123.45.67 was previously flagged in threat intelligence feeds for C2 activity." Conduct a Root Cause Analysis based on this traffic.	The root cause of the traffic log is a failure of the VPN policy.

Figura 15 - Interacció amb el model LLM.

En aquest cas, el model ha emès un veredict que no segueix un mètode consistent ni justifica el motiu per el què considera que hi ha un problema i que aquest tingui el seu origen en la política de la VPN. Aquesta avaluació també s'ha basat en un incident sintètic i en informació parcial, és cert que un dels punts que s'ha inclòs és un canvi en la política de la VPN i que pot tenir relació amb el comportament de la xarxa en el que s'acaba indicant que hi ha interaccions amb una IP externa que ha estat identificada

com que està essent utilitzada per activitat maliciosa, però no acaba de justificar la decisió.

Després d'aquesta primera fase, havent passat el primer joc de dades (JD1) al model per entrenar-lo, s'ha tornat a avaluar les dues primeres preguntes, a continuació es veuen els canvis que ha tingut el model respecte de l'anterior ronda de preguntes.

Pregunta 1 (JD1)	Resposta 1 (JD1)
<p>Analyze the following network traffic:</p> <p>No.,Time,Source,Destination,Protocol,Length,Info</p> <p>1,0.000000,10.10.31.101,10.10.31.1,DNS,75,Standard query 0x7ec7 A grafielucho.com</p> <p>2,0.113504,10.10.31.1,10.10.31.101,DNS,107,Standard query response 0x7ec7 A grafielucho.com A 104.21.32.6 A 172.67.182.27</p> <p>3,0.117203,10.10.31.101,grafielucho.com,TCP,66,56108 > 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM</p> <p>4,0.161532,grafielucho.com,10.10.31.101,TCP,66,80 > 56108 [SYN, ACK] Seq=0 Ack=1 Win=64240 Len=0 MSS=1380 SACK_PERM WS=8192</p> <p>5,0.161684,10.10.31.101,grafielucho.com,TCP,60,56108 > 80 [ACK] Seq=1 Ack=1 Win=131072 Len=0</p> <p>6,0.161896,10.10.31.101,grafielucho.com,HTTP,361,GET / HTTP/1.1</p> <p>Conduct a Root Cause Analysis based on this log</p>	<p>This is an example of a packet sniffer protocol connection, where the packet travels through multiple server domains with a maximum length of 228 bytes, as well as a 107-character message body. The probe sends a request to a reverse DNS server to check the port for the packet. However, the response contains a broken header and a sequence of characters that match a pattern of file paths for all the requests.</p>

Figura 16 - Interacció amb el model LLM després d'una ronda de *fine-tune*.

Comprovem, de nou, que la resposta proporcionada per part del model no satisfà el que se li demana, proporciona informació inexacta sobre el registre de trànsit de xarxa registrada a "Pregunta 1". Tampoc realitza una anàlisi precisa de la causa arrel basada en el registre.

Pregunta 2 (JD1)	Resposta 2 (JD1)																		
<p>Analyze the following network traffic log:</p> <p>SIEM System: "CyberGuard SIEM Pro"</p> <p>Network Devices Reporting: "Cisco NGFW, Cisco Catalyst 9300 Switch, Cisco ISR 4000 Router"</p> <p>Network Segment: "Corporate VLAN 10 - Finance Department"</p> <p>Associated User: "bob.finance@company.com - Last login: 2023-10-15 14:55:00 from 192.168.1.15"</p> <p>Recent Network Changes: "New VPN policy applied at 2023-10-14 18:30:00"</p> <p>Suspicious Events: "5 failed login attempts on 192.168.1.15 at 2023-10-15 15:00:00"</p> <p>Traffic Details:</p> <table border="1"> <thead> <tr> <th>Timestamp</th> <th>Source IP</th> <th>Destination IP</th> <th>Protocol</th> <th>Port</th> <th>Payload</th> </tr> </thead> <tbody> <tr> <td>2023-10-15 15:10:01</td> <td>192.168.1.15</td> <td>195.123.45.67</td> <td>TCP</td> <td>4444</td> <td>"Encoded Payload Detected"</td> </tr> <tr> <td>2023-10-15 15:10:02</td> <td>195.123.45.67</td> <td>192.168.1.15</td> <td>TCP</td> <td>4444</td> <td>"Response with encoded data"</td> </tr> </tbody> </table> <p>Network Behavior: "Unusual outbound traffic to an external IP on a non-standard port"</p> <p>Other Observations: "The external IP 195.123.45.67 was previously flagged in threat intelligence feeds for C2 activity."</p> <p>Conduct a Root Cause Analysis based on this traffic.</p>	Timestamp	Source IP	Destination IP	Protocol	Port	Payload	2023-10-15 15:10:01	192.168.1.15	195.123.45.67	TCP	4444	"Encoded Payload Detected"	2023-10-15 15:10:02	195.123.45.67	192.168.1.15	TCP	4444	"Response with encoded data"	<p>Network Traffic Log: Network Access Request (NTAP/DNS) - Finance Department - 5 failed login attempts on 192.168.1.15</p>
Timestamp	Source IP	Destination IP	Protocol	Port	Payload														
2023-10-15 15:10:01	192.168.1.15	195.123.45.67	TCP	4444	"Encoded Payload Detected"														
2023-10-15 15:10:02	195.123.45.67	192.168.1.15	TCP	4444	"Response with encoded data"														

Figura 17 - Segona interacció amb el model LLM després d'una ronda de *fine-tune*.

En aquest cas, la resposta del model sembla estar simplificant la informació registrada en la "Pregunta 2" en comptes de fer-ne una anàlisi de causa arrel.

Després d'aquesta segona fase, havent passat el segon joc de dades (JD2) al model per entrenar-lo, s'han tornat a avaluar les dues primeres preguntes, a continuació es veuen els canvis que ha tingut el model respecte de l'anterior ronda de preguntes.

Pregunta 1 (JD2)	Resposta 1 (JD2)
<p>a) Analyze the following network traffic:</p> <p>No.,Time,Source,Destination,Protocol,Length,Info</p> <p>1.0.000000,10.10.31.101,10.10.31.1,DNS,75,Standard query 0x7ec7 A grafielucho.com</p> <p>2.0.113504,10.10.31.1,10.10.31.101,DNS,107,Standard query response 0x7ec7 A grafielucho.com A 104.21.32.6 A 172.67.182.27</p> <p>3.0.117203,10.10.31.101,grafielucho.com,TCP,66,56108 > 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM</p> <p>4.0.161532,grafielucho.com,10.10.31.101,TCP,66,80 > 56108 [SYN, ACK] Seq=0 Ack=1 Win=64240 Len=0 MSS=1380 SACK_PERM WS=8192</p> <p>5.0.161684,10.10.31.101,grafielucho.com,TCP,60,56108 > 80 [ACK] Seq=1 Ack=1 Win=131072 Len=0</p> <p>6.0.161896,10.10.31.101,grafielucho.com,HTTP,361,GET / HTTP/1.1</p> <p>Conduct a Root Cause Analysis based on this log</p>	<p>The network traffic could potentially represent a network connection between a computer and a client, causing the computer to communicate with the client over HTTP protocol. However, since the client is on a regular basis and has a connection to the client's mail server, it may be insecure as the client is unable to determine its current IP address. Moreover, due to improper data compression, the client may not be able to locate the correct file or port on their server.</p>

Figura 18 - Interacció amb el model LLM després d'un nou procés de *fine-tune*.

En aquesta ocasió, el model entrenat ofereix una visió general del registre de tràfic de xarxa i planteja certa informació que pot ser potencial preocupant, tot i que a aquestes alçades segueix sense proporcionar una anàlisi de la causa arrel detallada.

Pregunta 2 (JD1)	Resposta 2 (JD1)																		
<p>Analyze the following network traffic log:</p> <p>SIEM System: "CyberGuard SIEM Pro"</p> <p>Network Devices Reporting: "Cisco NGFW, Cisco Catalyst 9300 Switch, Cisco ISR 4000 Router"</p> <p>Network Segment: "Corporate VLAN 10 - Finance Department"</p> <p>Associated User: "bob.finance@company.com - Last login: 2023-10-15 14:55:00 from 192.168.1.15"</p> <p>Recent Network Changes: "New VPN policy applied at 2023-10-14 18:30:00"</p> <p>Suspicious Events: "5 failed login attempts on 192.168.1.15 at 2023-10-15 15:00:00"</p> <p>Traffic Details:</p> <table border="1"> <thead> <tr> <th>Timestamp</th> <th>Source IP</th> <th>Destination IP</th> <th>Protocol</th> <th>Port</th> <th>Payload</th> </tr> </thead> <tbody> <tr> <td>2023-10-15 15:10:01</td> <td>192.168.1.15</td> <td>195.123.45.67</td> <td>TCP</td> <td>4444</td> <td>"Encoded Payload Detected"</td> </tr> <tr> <td>2023-10-15 15:10:02</td> <td>195.123.45.67</td> <td>192.168.1.15</td> <td>TCP</td> <td>4444</td> <td>"Response with encoded data"</td> </tr> </tbody> </table> <p>Network Behavior: "Unusual outbound traffic to an external IP on a non-standard port"</p> <p>Other Observations: "The external IP 195.123.45.67 was previously flagged in threat intelligence feeds for C2 activity."</p> <p>Conduct a Root Cause Analysis based on this traffic.</p>	Timestamp	Source IP	Destination IP	Protocol	Port	Payload	2023-10-15 15:10:01	192.168.1.15	195.123.45.67	TCP	4444	"Encoded Payload Detected"	2023-10-15 15:10:02	195.123.45.67	192.168.1.15	TCP	4444	"Response with encoded data"	<p>A total of 23 traffic failures between two different devices on the network were intercepted, two of which were taken to a web server and one of which was debugged.</p>
Timestamp	Source IP	Destination IP	Protocol	Port	Payload														
2023-10-15 15:10:01	192.168.1.15	195.123.45.67	TCP	4444	"Encoded Payload Detected"														
2023-10-15 15:10:02	195.123.45.67	192.168.1.15	TCP	4444	"Response with encoded data"														

Figura 19 - Interacció amb el model LLM després de fer el segon procés de *fine-tune*.

La resposta del model proporciona una descripció de l'activitat de tràfic de xarxa capturat i registrat en la pregunta, posa importància en el nombre de connexions fallides de tràfic interceptades entre diferents dispositius de xarxa i mencionant l'existència de comunicació amb un servidor web i una operació de depuració.

En aquest punt, s'identifiquen els següents reptes per assolir l'objectiu d'aconseguir fer el *fine-tuning* del model que s'ha marcat a nivell de projecte, que no és un altre que acabi sent un assistent que pugui fer anàlisi de causa arrel d'incidents de seguretat:

- Detalls de l'Anàlisi de Causa Arrel (ACA): les respostes actuals proporcionen una visió general dels fets que se li presenten, però falta una anàlisi més aprofundida. Una ACA completa hauria d'incloure un desglossament de les fases del incident, identificant la causa arrel de cada esdeveniment sospitós.
- Context històric: és de vital importància conèixer l'historial de canvis de xarxa, com ara la nova política VPN aplicada plantejada en una de les preguntes. Això podria estar relacionat amb els esdeveniments actuals i s'hauria d'avaluar com a part de la ACA. Aquest punt no és de fàcil resolució, ja que en el món real, els registres de canvis a la xarxa no es processen ni s'emmagatzemen juntament amb captures de tràfic de la xarxa, caldria aprofundir amb eines com poden ser un SIEM i llegir-hi tant fluxos de xarxa com canvis en configuracions.
- Connexions anòmales: l'activitat de xarxa inusual cap a una IP externa i un port no estàndard és una alerta crítica. L'ACA hauria de determinar per què això va ocórrer i si representa una amenaça.
- Fonts d'intel·ligència: s'observa en l'exemple anterior que una de les adreces IP externa a la que s'estableix una connexió estava prèviament marcada per activitat C2 (Command and Control) és significativa. S'ha de considerar com a part de l'ACA i es podria investigar més a fons l'activitat relacionada amb aquesta adreça IP.

Tot seguit, s'ha procedit a entrenar el model amb un incident real de seguretat (informació aportada per l'equip Unit42 de l'empresa PaloAlto), en concret he realitzat l'ACA de dos incidents i els he documentat, a la Taula 3, que es troba a [l'Annex 3 del present document](#), es pot veure el joc de dades (un cas d'infecció del programari maliciós IcedID amb dues opcions d'anàlisi), el punt important és ensenyar el procés d'anàlisi d'ACA (1) identificar la causa arrel, (2) entendre i minimitzar l'impacte del problema, (3) prevenir incidents futurs, lliçons apreses.

```

In [1]: # Codi adaptat a L'entrenament (sense avaluació) de 2 RCAs
import pandas as pd
from datasets import Dataset
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM, TrainingArguments, Trainer

# Definim el camí on hem guardat el model preentrenat
model_path = './model/'
print("Carregant el tokenizer i el model...")

# Carreguem el tokenizer i el model
tokenizer = AutoTokenizer.from_pretrained('MBZUAI/LaMini-T5-738M')
model = AutoModelForSeq2SeqLM.from_pretrained(model_path)

# Funció per carregar i processar el conjunt de dades
def carregar_i_processar_conjunt_de_dades(ruta_arxiu_csv, tokenizer, max_length=512):
    print(f"Carregant i processant el conjunt de dades des de {ruta_arxiu_csv}...")
    df = pd.read_csv(ruta_arxiu_csv, delimiter=';')
    text_input = df['input'].tolist()
    text_objectiu = df['target'].tolist()

    encoding = tokenizer(text_input, truncation=True, padding='max_length', max_length=max_length, return_tensors="pt")
    etiquetes = tokenizer(text_objectiu, truncation=True, padding='max_length', max_length=max_length, return_tensors="pt")['input_ids']

    conjunt_de_dades = Dataset.from_dict({
        'input_ids': encoding['input_ids'],
        'attention_mask': encoding['attention_mask'],
        'labels': etiquetes
    })

    return conjunt_de_dades

print("Carregant conjunts de dades d'entrenament...")
conjunt_entrenament = carregar_i_processar_conjunt_de_dades('C:\\Users\\aromma\\OneDrive - Universitat Oberta de Catalunya\\TFG\\train\\rca_t\\rca_train.csv')

print("Definint els arguments d'entrenament...")
arguments_entrenament = TrainingArguments(
    output_dir='./resultats/',
    num_train_epochs=3,
    per_device_train_batch_size=2, # Adjusted for Longer phrases
    logging_dir='./registres/',
    logging_steps=10
)

print("Començant l'entrenament...")
entrenador.train()

print("Desant el model entrenat...")
model.save_pretrained(model_path)

print("L'entrenament ha estat completat.")

```

D:\aromma\anaconda3\envs\TFGRCAAI\lib\site-packages\tqdm\auto.py:21: TqdmWarning: IPProgress not found. Please update jupyter and ipywidgets. See https://ipywidgets.readthedocs.io/en/stable/user_install.html
from .autonotebook import tqdm as notebook_tqdm

Carregant el tokenizer i el model...
Carregant conjunts de dades d'entrenament...
Carregant i processant el conjunt de dades des de C:\Users\aromma\OneDrive - Universitat Oberta de Catalunya\TFG\train\rca_t\rca_train.csv...
Definint els arguments d'entrenament...
Iniciant el Trainer...
Començant l'entrenament...

[3/3 55:29, Epoch 3/3]

Step	Training Loss
Desant el model entrenat...	
L'entrenament ha estat completat.	

Figura 20 - Quadern Jupyter que mostra el procés de càrrega i aprenentatge del model LLM.

Tot seguit, es torna a interactuar amb el model, repetint les preguntes que ja s'havia fet i així poder comparar els resultats:

```

In [6]: # Interactuem amb el model entrenat - anàlisi d'una captura de tràfic de xarxa
input_prompt_10 = "Analyze the following network traffic:\n\n" \
"No.,Time,Source,Destination,Protocol,Length,Info\n" \
"1,0.000000,10.10.31.101,10.10.31.1,DNS,75,Standard query 0x7ec7 A grafielucho.com\n" \
"2,0.113504,10.10.31.1,10.10.31.101,DNS,107,Standard query response 0x7ec7 A grafielucho.com A 104.21.32.6 A 172.67.182.27\n" \
"3,0.117203,10.10.31.101,grafielucho.com,TCP,66,56108 > 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM\n" \
"4,0.161532,grafielucho.com,10.10.31.101,TCP,66,80 > 56108 [SYN, ACK] Seq=0 Ack=1 Win=64240 Len=0 MSS=1380 SACK_PERM WS=8192\n" \
"5,0.161684,10.10.31.101,grafielucho.com,TCP,60,56108 > 80 [ACK] Seq=1 Ack=1 Win=131072 Len=0\n" \
"6,0.161896,10.10.31.101,grafielucho.com,HTTP,361,GET / HTTP/1.1\n" \
"Conduct a Root Cause Analysis based on this log."
generated_text_10 = pipe(input_prompt_10, max_length=512, do_sample=True)[0]['generated_text']
print("Response 10:", generated_text_10)
# Response 21/11/2023: The given network traffic is a traffic log which includes information about the user's network server cony

```

Response 10: The provided network traffic consists of packets of HTTP requests for server location and referring to an internal network for location

```
In [7]: # Interactuem amb el model carregat a huggingface - anàlisi d'una captura de tràfic de xarxa
input_prompt_t1 = "Analyze the following network traffic:\n\n" \
"No.,Time,Source,Destination,Protocol,Length,Info\n" \
"1,0.000000,10.10.31.101,10.10.31.1,DNS,75,Standard query 0x7ec7 A grafielucho.com\n" \
"2,0.113504,10.10.31.1,10.10.31.101,DNS,107,Standard query response 0x7ec7 A grafielucho.com A 104.21.32.6 A 172.67.182.27\n" \
"3,0.117203,10.10.31.101,grafielucho.com,TCP,66,56108 > 80 [SVN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM\n" \
"4,0.161532,grafielucho.com,10.10.31.101,TCP,66,80 > 56108 [SVN, ACK] Seq=0 Ack=1 Win=64240 Len=0 MSS=1380 SACK_PERM WS=8192\n" \
"5,0.161684,10.10.31.101,grafielucho.com,TCP,60,56108 > 80 [ACK] Seq=1 Ack=1 Win=131072 Len=0\n" \
"6,0.161896,10.10.31.101,grafielucho.com,HTTP,361,GET / HTTP/1.1\n" \
"Conduct a Root Cause Analysis based on this log."
generated_text_t1 = pipe(input_prompt_t1, max_length=4096, do_sample=True)[0]['generated_text']
print("Response:", generated_text_t1)
## Response 10: The given network traffic is a traffic log which includes information about the user's network server
# configurations and IP addresses. Specifically, it shows the IP addresses, server addresses, DNS records, and message
# information. The log also provides a number of error messages along with trace back to IP addresses. It also shows the error
# messages and trace back to a specific session on the server.
# ##### After training 1 (41 Lines) #####
## Response: This is an example of a packet sniffer protocol connection, where the packet travels through multiple server
# domains with a maximum length of 228 bytes, as well as a 107-character message body. The probe sends a request to a reverse
# DNS server to check the port for the packet. However, the response contains a broken header and a sequence of characters that
# match a pattern of file paths for all the requests.
# ##### After further training 1 (41 Lines) #####
## Response 06/11/2023: The network traffic is related to a race or a football match involving a professional football team.
# Specifically, the log provides information about the network topology, routing tables, and protocols used for the game.
# It also includes data about the server and its configuration, as well as the IP address of the player involved. The traffic
# seems to be related to the sport of football and is likely caused by high-speed connectivity between the player and the server.
# ##### After further training 1 (41 Lines) #####
## Response 07/11/2023: The network traffic is likely a connection between an external server and a web server. The DNS server
# resolves the connection by sending a request to a DNS resource, and after validating the request, a 500 page response is sent
# from it, which is not supported by the web server. However, it seems that there is some correspondence there. The IP address
# is 45000 with the domain name "resolved" was replaced with the user's full name.
# Response 09/11/2023: This is a network traffic log that involves a connection test between two computers. Two different
# computers (one from different networks) are also being communicated with, based on the information provided, appear to be
# trying to establish a connection to the src server to exchange network traffic.
## Response 10/11/2023: This network traffic sent an attempt to connect to an Failed.
## Response 10/11/23 b: Network traffic analysis of network traffic between two HTTP request for adnswlan: 56304/417285
## Response 11/11/23 a: The network traffic could potentially represent a network connection between a computer and a client, cau
## Response 11/11/23 b: The network traffic seems to be a computer system communicating with an internet service provider (ISP) c
## Response 13/11/23: This network traffic appears to be related to a hacking attempt perpetrated by a hacker using the Kerberos
## Response 20/11/23: The network traffic consists of packets sent between two computers that are connected to the same virtual p
## Response 22/11/2023: The network traffic is a packet loss attack (SDN) on an IPv4 network. The packet comes from the destinati
```

Response: The provided network traffic seems to involve a denial of service (DoS) attack on a server operating from several servers located in multiple locations. The denial of service was detected through a time-based analysis on received packets, displaying "Hawking" as the IPv4 address.

Figura 21 - Quadern Jupyter amb interacció amb el model LLM.

Es comprova que es repeteixen patrons respecte amb respostes anteriors, en aquest punt, procedeix a avaluar si el model, que ha estat pre-entrenat per part de l'equip investigador que el va definir i jo he enfocat a aprendre glossari i termes de ciberseguretat, així com a analitzar incidents de seguretat, és capaç de generar diferents respostes per una mateixa pregunta en una única interacció.

Per fer-ho, es selecciona un cas real d'infecció del *Rhadamanthys Stealer*, que es va observar en un anunci de Google que conduïa a una pàgina que pretenia ser de Notepad++, però que en realitat contenia el malware prèviament esmentat.

A continuació, es mostra una anàlisi completa de la infecció d'aquest *stealer* (tipus de malware dirigit a extreure dades dels equips infectats):

NOTES:

- The Google ad for this infection chain did not hide the fake Notepad++ site (malicious site was visible in the ad).
- Hasankahrimanoglu[.]com[.]tr was used before in December 2022 for the same type of fake Notepad++ site.

- The downloaded zip contains a Rhadamanthys Stealer EXE inflated to approx 802 MB and a folder with unrelated files.
- This infection process used an image with data hidden through steganography, but I don't know what this data is.
- After the steganography image, the infection switched to websocket traffic to encrypt the post-infection traffic.
- Special thanks to @500mk500, @da_667, and @ex_raritas for identifying this malware!

GOOGLE AD URL:

- https://www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwiDiu-13Kv8AhWkE9QBHa7UADwYABACGgJvYQ&ae=2&ohost=www.google.com&cid=CAASJORopA-3gIku5H1e8Y7FuoHCKJjSFjgbPRpqqj2ZKXrbPcnfRQ&sig=AOD64_2JnosseZ0C9qLEszOg47HtRfUYig&q&adurl&ved=2ahUKEw-jup-i13Kv8AhVckmoFHeyWDP8Q0Qx6BAgKEAE

FAKE NOTEPAD++ SITE:

- <https://noteepad.hasankahrimanoglu.com.tr/>

ZIP DOWNLOAD URL:

- <https://noteepad.hasankahrimanoglu.com.tr/ing.php>

DOWNLOADED ZIP ARCHIVE:

- SHA256 hash:

56840aba173e384469ea4505158eead4e7612c41caa59738fcf5efe9b2e10864

- File size: 69,728,905 bytes
- File name: Nottepaad_lastNeWx32x64.zip

EXE FOR RHADAMANTHYS STEALER EXTRACTED FROM ABOVE ZIP ARCHIVE:

- SHA256 hash:
8d0e8baffed28f5c709a99392f7ab42430635839f7aba92a01c956c10702c8f

- File size: 802,160,640 bytes
- File name: Noteppad_SetupX32iX64.exe
- Note: This file has more than 801 MB of extra bytes to make the EXE too big for services like VirusTotal to analyze.

RHADAMANTHYS STEALER EXE CARVED TO REMOVE PADDING:

- SHA256 hash:
af67a6bd0baf78191617c97aad2d21b7d6133e879c92c97b1b1345d629f79661

- File size: 333,344 bytes
- File name: Noteppad_SetupX32iX64-carved.exe
- Analysis: <https://app.any.run/tasks/96a0206a-5683-47c1-9804-04aff3c55228>
- Analysis: <https://tria.ge/230103-tr9agsfb8w>

POST INFECTION TRAFFIC:

- 162.33.178[.]106 port 80 - 162.33.178[.]106 - GET /gjntrrm/zzn2o.hgfq
- 162.33.178[.]106 port 80 - 162.33.178[.]106 - GET /gjntrrm/zzn2o.hgfq

NOTES ON THE POST-INFECTION TRAFFIC:

- The first HTTP GET request returns a 929,566 byte .jpg image that's 95x120 pixels and has obfuscated or encrypted data,

so it seems there's steganography involved here.

- The second HTTP GET request upgrades the traffic, switching to encrypted websocket activity.

STEGANOGRAPHY IMAGE:

- SHA256 hash:
c4b7e2de87630bde08e367c75d9a2b9ae79b1d4f03ee8014531239c9597efc2e

- File size: 929,566 bytes

- Location: hxxp://162.33.178[.]106/gjntrrm/zznb2o.hgfq

- File description: JPEG image 95x120 bytes

- Note: Same size, but different file hash seen from infections on at least 2 different Win10 hosts.

Figura 22 - Anàlisi integral d'un incident de seguretat.

Es procedeix a efectuar una nova interacció amb el model per valorar com es comporta quan en cas que se li demanin 10 respostes de forma iterativa amb el mateix text d'entrada:

```
In [9]: # Definim la prompt d'entrada
input_prompt = "Make a root cause analysis of the following:\n\n" \
"The Google ad for this infection chain did not hide the fake Notepad++ site (malicious site was visible in the ad).\n" \
"Hasankahrmanoglu[.]com[.]tr was used before in December 2022 for the same type of fake Notepad++ site\n" \
"The downloaded zip contains a Rhadamanthys Stealer EXE inflated to approx 802 MB and a folder with unrelated files.\n" \
"This infection process used an image with data hidden through steganography, but I don't know what this data is.\n" \
"After the steganography image, the infection switched to websocket traffic to encrypt the post-infection traffic.\n" \
"hxxps://www.googleadservices[.]com/pagead/acLk?sa=L&ai=DChcSEwiDiu-13Kv8AhWkE9QBHa7UADwYABACGg3vYQ&ae=2&ohost=www.google.com&ic" \
"hxxps://notepad.hasankahrmanoglu[.]com[.]tr/\n" \

# Generem 10 respostes utilitzant la funció pipe
for i in range(10):
    generated_text = pipe(input_prompt, max_length=4096, do_sample=True, num_return_sequences=1)[0]['generated_text']
    print(f"Resposta {i+1}: {generated_text}\n")
```

Resposta 1: To conduct a root cause analysis of the Google ad for this infection chain, it's important to first analyze the URL of the infected files using the msv directory in the search results and remove the malicious gist from the browser. This will likely determine the cause of the infection.

Resposta 2: The root cause of the malware attack in the Google ad chain is unclear, as it only describes a specific attack on a specific machine that downloaded an infected file onto a system and required manual scanning for virus activity.

Resposta 3: A possible root cause for the Ransomware infection for this website is the ingestion of malicious code and the use of a template for the root directory analysis.

Resposta 4: Root Cause: Misleading images in Google ads

Resposta 5: Root cause analysis: The malware was able to take control of a significant portion of the root server by using a random password. The password was a random password (ransom) issued by the software developer that was obtained through a ransomware distribution.

Resposta 6: Root Cause Analysis: Undiscovered Trojan software using a fake notepad++ site

Resposta 7: The root cause of the Notepad++ malware infection is the use of the root kit of the ransomware distributor to hide the malware

Resposta 8: The root cause for the A03 ransomware attacks is that a portion of the downloaded file on one of the infected systems had a file named "jamllan" modified.

Resposta 9: Root cause analysis: The downloaded file contains an embedded exploit that exploits some vulnerability in the Notepad++ program using Rhadamanthys Stealer EXE extension

Resposta 10: The root cause of the recent malware attack on the Google ad network was not necessarily related to the thief mentioned in the previous message: "Disaster: Malware actor"

Figura 23 - Quadern Jupyter amb el procés d'entrenament del model RL.

En el marc del desenvolupament de models de llenguatge de gran escala (LLM) generalistes i en concret en el cas que s'està duent a terme per a l'anàlisi d'incidents de seguretat, s'ha observat que el model pot generar respostes amb variacions en la seva precisió. Aquest fenomen es deu, en gran part, a les anomenades al·lucinacions de models LLM que es produeixen durant la fase de preentrenament, un aspecte ja destacat en l'article del model de referència LaMini-Flan-T5-738M. Per tal d'optimitzar la qualitat de les respostes generades, especialment en el context de l'Anàlisi de Causa Arrel (ACA) en ciberseguretat, es proposa la incorporació d'un analista de seguretat experimentat, anomenat aquí RL, combinació que resulta amb el que s'anomena dins el context del present projecte 'CyberVigilant', que pugui avaluar objectivament les respostes del model LLM afinat de forma personalitzada.

Per millorar aquesta interacció i augmentar la precisió en la resposta del model, es recomana l'ús de tècniques avançades d'aprenentatge per reforç (RL). En aquest context, es fa ús de la llibreria OpenAI Gymnasium, una eina de codi obert que proporciona un entorn estàndard per a l'experimentació en RL. Gymnasium permet l'entrenament i l'avaluació de models RL en diversos entorns simulats, facilitant l'aplicació de tècniques d'aprenentatge automatitzat en problemes complexos.

Aquesta llibreria s'integra l'ús de NPLGym, una llibreria específica per a l'entrenament de models RL en el context de processament del llenguatge natural. NPLGym ofereix models RL preentrenats que es poden afinar per a tasques específiques com l'anàlisi d'incidents de seguretat, proporcionant una base sòlida per a la personalització i l'entrenament específic.

Finalment, per implementar polítiques d'aprenentatge efectives, s'incorpora l'ús de Stable-Baselines3 [8], una llibreria que permet l'aplicació de polítiques Deep Q-Network (DQN) i Proximal Policy Optimization (PPO), entre d'altres. Aquestes polítiques són fonamentals per a l'entrenament de models RL complexos, com ho demostra el seu ús en projectes com ChatGPT d'OpenAI. La combinació d'aquestes eines permetrà al CyberVigilant interactuar eficientment amb el model, sol·licitant un mínim de 10 respostes per cada consulta i seleccionant la més adequada, millorant així la qualitat i la fiabilitat de l' Anàlisi de Causa Arrel (ACA) en ciberseguretat.

En el present projecte, es fa ús de la llibreria NLPGym com ja s'ha descrit, que ens ajuda a establir un entorn virtual, en concret s'escull el Question-Answering (QA), que funciona seguint un conjunt d'etapes que són fonamentals per a l'aprenentatge basat en recompenses. Aquí tens una explicació detallada del seu funcionament:

1. Inicialització:

- S'inicialitza l'entorn QA de NLPGym. Aquest entorn és un marc que simula situacions de QA on l'agent (model RL) interactua.
- Els paràmetres com la configuració de l'entorn, el nombre de possibles accions, i l'espai d'estats són definits.

2. Estats:

- En l'entorn de QA, un estat pot ser una pregunta que necessita una resposta.
- L'estat canvia quan es rep una nova pregunta o quan l'entorn evoluciona basat en l'acció de l'agent.

3. Accions:

- Les accions en aquest context podrien ser les diferents respostes possibles que l'agent pot donar a una pregunta.
- L'agent selecciona una acció basada en la seva política d'aprenentatge, que podria ser determinista o estocàstica.

4. Política d'aprendre:

- La política determina com l'agent selecciona accions basades en l'estat actual. Pot ser tan simple com una taula de cerca o tan complexa com una xarxa neuronal.

5. Recompenses:

- Una vegada l'agent realitza una acció (resposta a una pregunta), l'entorn avalua aquesta acció i proporciona una recompensa.
- Les recompenses són senyals per a l'agent per entendre quin de ben fet ha estat el seu comportament. Per exemple, respostes correctes poden rebre

recompenses positives mentre que les incorrectes reben recompenses negatives o cap recompensa.

6. Aprenentatge:

- L'agent utilitza les recompenses per ajustar la seva política d'aprenentatge, intentant millorar les seves decisions futures.
- Això es realitza a través de mètodes com Q-learning, Deep Q-Networks (DQN), o altres algorismes d'RL.

7. Iteració:

- El procés es repeteix amb l'agent rebent noves preguntes (estats), seleccionant accions, rebent recompenses, i ajustant la seva política.

A continuació, es mostra la primera fase ja executada, que consisteix a configurar l'entorn sobre el qual es carrega, entrena i guarda el model RL que forma part del CyberVigilant:

```

import gymnasium
from nlp_gym.data_pools.custom_question_answering_pools import QASC
from nlp_gym.envs.question_answering.env import QAEEnv
from nlp_gym.data_pools.question_answering_pool import QADDataPool, Sample
import pandas as pd
from datasets import load_dataset

# Entorn operatiu amb l'exemple de NPL Gym - RL
from nlp_gym.envs.question_answering.featurizer import InformedFeaturizer, SimpleFeaturizer
from stable_baselines3 import DQN
from stable_baselines3.dqn.policies import MlpPolicy as DQNPoly
from stable_baselines3.common.env_util import make_vec_env
from stable_baselines3.common.evaluation import evaluate_policy
import tqdm
from nlp_gym.envs.question_answering.env import QAEEnv as OriginalQAEEnv

D:\arnma\anaconda3\envs\TFGRCAAI\lib\site-packages\tqdm\auto.py:21: TqdmWarning: IPProgress not found. Please update jupyter and ipywidgets. See https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm

```

```

# data pool
pool = QASC.prepare("train")

# creem un entorn per question answering
env = QAEEnv()
for sample, weight in pool:
    env.add_sample(sample)

# executem un episodi a mode d'exemple i per provar el correcte funcionament de l'entorn
done = False
state = env.reset()
total_reward = 0
while not done:
    action = env.action_space.sample()
    state, reward, done, info = env.step(action)
    total_reward += reward
    env.render()
    print(f"Action: {env.action_space.ix_to_action(action)}")

Found cached dataset qasc (C:/Users/arnma/.cache/huggingface/datasets/qasc/default/0.1.0/a8c2ff717429f8f9041f665234865cc42c93d4b1b3c4f16a1e119a85366714ad)
100%|██████████| 3/3 [00:00<00:00, 150.00it/s]

```

Step 0

Question: What do sea lions learn how to behave from?

Fact: Playing is one way mammals and other animals learn how to behave.

Fact: Sea lions are tactile mammals.

Choice A: exercise

Action: ANSWER

```

file_path = "../../datasets/RL/QASC_Cyber_Dataset/sample1.jsonl"
jsonObj = pd.read_json(path_or_buf=file_path, lines=True)
class QASC_json(QADDataPool):
    """
    Source: Custom QASC Cyber
    """
    @classmethod
    def prepare(cls):
        ds = load_dataset('json', data_files=file_path)['train']
        samples = []
        for datapoint in ds:
            sample_id = datapoint["id"]
            facts = [str(datapoint["fact1"]), str(datapoint["fact2"])]
            question = datapoint["question"]['stem']

            choices = {qi['label']: qi['text']
                       for qi in datapoint['question']['choices']}

            answer = datapoint["answerKey"]
            sample = Sample(sample_id, question, facts, choices, answer)

            samples.append(sample)

        return QASC(samples)

```

```

pool = QASC_json.prepare()

```

Downloading and preparing dataset json/default to C:/Users/arnma/.cache/huggingface/datasets/json/default-6be6581a1fe7093e/0.0.0/e347ab1c932092252e717ff3f949105a4dd28b27e842dd53157d2f72e276c2e4...

Downloading data files: 100%|██████████| 1/1 [00:00<?, ?it/s]
Extracting data files: 100%|██████████| 1/1 [00:00<00:00, 333.23it/s]

Dataset json downloaded and prepared to C:/Users/arnma/.cache/huggingface/datasets/json/default-6be6581a1fe7093e/0.0.0/e347ab1c932092252e717ff3f949105a4dd28b27e842dd53157d2f72e276c2e4. Subsequent calls will reuse this data.

100%|██████████| 1/1 [00:00<00:00, 17.54it/s]

```
env = QAEVn()
# question answering env

for sample, weight in pool:
    env.add_sample(sample)
```

```
# play an episode
done = False
state = env.reset()
total_reward = 0
while not done:
    action = env.action_space.sample()
    state, reward, done, info = env.step(action)
    total_reward += reward
    env.render()
    print(f"Action: {env.action_space.ix_to_action(action)}")
```

Step 1

Question: How should the solution to prevent IcedID infections be implemented?

Fact: Enhancing email security involves implementing advanced filtering and scanning protocols.

Fact: This approach targets and mitigates the risk of email-based malware delivery.

Choice A: By training employees on physical security

Fact: Enhancing email security involves implementing advanced filtering and scanning protocols.

Fact: This approach targets and mitigates the risk of email-based malware delivery.

Choice B: Upgrading all software systems

Action: CONTINUE

Step 1

Question: How should the solution to prevent IcedID infections be implemented?

Fact: Enhancing email security involves implementing advanced filtering and scanning protocols.

Fact: This approach targets and mitigates the risk of email-based malware delivery.

Choice A: By training employees on physical security

Fact: Enhancing email security involves implementing advanced filtering and scanning protocols.

Fact: This approach targets and mitigates the risk of email-based malware delivery.

Choice B: Upgrading all software systems

Action: ANSWER

```
# Procedim a entrenar el model
model = DQN(DQNPolicy, env, gamma=0.99, batch_size=32, learning_rate=1e-4,
            buffer_size=10000, exploration_fraction=0.1, exploration_final_eps=0.1,
            target_update_interval=100, policy_kwargs={"net_arch": [64, 64]},
            verbose=1)
model.learn(total_timesteps=int(1e+5))
# Save the trained model
model.save("../datasets/RL/trained/model_v2")
```

learning_rate	0.0001
loss	0.0428
n_updates	11017

rollout/	
ep_len_mean	2.64
ep_rev_mean	0.67
exploration_rate	0.1
time/	
episodes	43916
fps	167
time_elapsed	560
total_timesteps	94078
train/	

Figura 24 - Procés complet d'entrenament i emmagatzematge en local del model RL.

En aquest punt es disposa de la part més interessant del projecte, es carrega el model que s'ha entrenat i se'n fa l'avaluació, aquest model RL té una pregunta i múltiples respostes entre les que escollir-ne una, un cop entrenat, és capaç de decidir la resposta que li sembla correcte, mentre no la trobi seguirà demanant respostes:

```
state = env.reset()
for i in range(10):
    ## aquí és on deixem correr la màgia -
    ## el model és la policy, indica el millor curs d'acció entre les diferents opcions (Action Space)
    action, _states = model.predict(state, deterministic=False)

    action = action.item()

    state, reward, done, info = env.step(action)
    env.render()
    # VecEnv resets automatically
    # if done:
    # obs = env.reset()
env.close()
```

Step 1

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice A: Inadequate monitoring of the network leading to unnoticed intrusion

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice B: Failure to update security patches causing system vulnerability

Step 2

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice A: Inadequate monitoring of the network leading to unnoticed intrusion

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice B: Failure to update security patches causing system vulnerability

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice C: Lack of employee awareness training on phishing attacks

Step 2

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice A: Inadequate monitoring of the network leading to unnoticed intrusion

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice B: Failure to update security patches causing system vulnerability

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice C: Lack of employee awareness training on phishing attacks

Step 3

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice A: Inadequate monitoring of the network leading to unnoticed intrusion

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice B: Failure to update security patches causing system vulnerability

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice C: Lack of employee awareness training on phishing attacks

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice D: Email with a malicious PDF initiating a complex infection chain

Step 3

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice A: Inadequate monitoring of the network leading to unnoticed intrusion

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice B: Failure to update security patches causing system vulnerability

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice C: Lack of employee awareness training on phishing attacks

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice D: Email with a malicious PDF initiating a complex infection chain

Step 3

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice A: Inadequate monitoring of the network leading to unnoticed intrusion

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice B: Failure to update security patches causing system vulnerability

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice C: Lack of employee awareness training on phishing attacks

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice D: Email with a malicious PDF initiating a complex infection chain

Step 3

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.

Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice A: Inadequate monitoring of the network leading to unnoticed intrusion

Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice B: Failure to update security patches causing system vulnerability
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice C: Lack of employee awareness training on phishing attacks
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice D: Email with a malicious PDF initiating a complex infection chain

Step 3

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice A: Inadequate monitoring of the network leading to unnoticed intrusion
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice B: Failure to update security patches causing system vulnerability
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice C: Lack of employee awareness training on phishing attacks
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice D: Email with a malicious PDF initiating a complex infection chain

Step 3

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice A: Inadequate monitoring of the network leading to unnoticed intrusion
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice B: Failure to update security patches causing system vulnerability
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice C: Lack of employee awareness training on phishing attacks
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice D: Email with a malicious PDF initiating a complex infection chain

Step 3

Question: What is the root cause analysis of the IcedID (Bokbot) infection's effect on 2023-01-16?

Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice A: Inadequate monitoring of the network leading to unnoticed intrusion
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice B: Failure to update security patches causing system vulnerability
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.
Choice C: Lack of employee awareness training on phishing attacks
Fact: The IcedID infection began with an email containing a malicious PDF attachment.
Fact: This initiated a complex infection chain leading to backconnect traffic, VNC, and Cobalt Strike.

Choice D: Email with a malicious PDF initiating a complex infection chain

Figura 25 - Procés d'entrenament del model RL.

Es comprova com l'entrenament ha estat efectiu i el CyberVigilant ha estat capaç d'encertar les respostes a totes les preguntes que se li han plantejat (seleccionades a l'atzar). Per fer aquest exercici, s'han preparat nous jocs de dades que queden recollits en l'apartat corresponent.

Per acabar de fer les verificacions, s'han fet les mateixes preguntes al model final de CyberVigilant amb la selecció de resposta òptima per part del model RL, es veuen a continuació els resultats:

Pregunta 1 (JD2)	Resposta 1 (JD2)
Analyze the following network traffic: No.,Time,Source,Destination,Protocol,Length,Info 1,0.000000,10.10.31.101,10.10.31.1,DNS,75,Standard query 0x7ec7 A grafielucho.com 2,0.113504,10.10.31.1,10.10.31.101,DNS,107,Standard query response 0x7ec7 A grafielucho.com A 104.21.32.6 A 172.67.182.27 3,0.117203,10.10.31.101,grafielucho.com,TCP,66,56108 > 80 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM 4,0.161532,grafielucho.com,10.10.31.101,TCP,66,80 > 56108 [SYN, ACK] Seq=0 Ack=1 Win=64240 Len=0 MSS=1380 SACK_PERM WS=8192 5,0.161684,10.10.31.101,grafielucho.com,TCP,60,56108 > 80 [ACK] Seq=1 Ack=1 Win=131072 Len=0 6,0.161896,10.10.31.101,grafielucho.com,HTTP,361,GET / HTTP/1.1 Conduct a Root Cause Analysis based on this log	The network traffic being analyzed may indicate that an e-commerce site has been attacked by an e-commerce website on October 4th through 10th

Figura 26 - Interacció amb la combinació de models RL i LLM (CyberVigilant).

Pregunta 2 (CyberVigilant)	Resposta 2 (CyberVigilant)
Analyze the following network traffic log: SIEM System: "CyberGuard SIEM Pro" Network Devices Reporting: "Cisco NGFW, Cisco Catalyst 9300 Switch, Cisco ISR 4000 Router" Network Segment: "Corporate VLAN 10 - Finance Department" Associated User: "bob.finance@company.com - Last login: 2023-10-15 14:55:00 from 192.168.1.15" Recent Network Changes: "New VPN policy applied at 2023-10-14 18:30:00" Suspicious Events: "5 failed login attempts on 192.168.1.15 at 2023-10-15 15:00:00" Traffic Details: Timestamp Source IP Destination IP Protocol Port Payload ----- 2023-10-15 15:10:01 192.168.1.15 195.123.45.67 TCP 4444 "Encoded Payload Detected" 2023-10-15 15:10:02 195.123.45.67 192.168.1.15 TCP 4444 "Response with encoded data" Network Behavior: "Unusual outbound traffic to an external IP on a non-standard port" Other Observations: "The external IP 195.123.45.67 was previously flagged in threat intelligence feeds for C2 activity." Conduct a Root Cause Analysis based on this traffic.	The network log for network event 608899-01-01 is available.

Figura 27 - Segona interacció amb la combinació de models RL i LLM (CyberVigilant).

Es pot comprovar, amb les dues últimes respostes, que el model de RL ajuda a fer una selecció de respostes més acurades i amb un índex d'allucinacions menor al que es veu anteriorment, tot i així, la conclusió final és que falta seguir entrenant els models per tal que adquireixin més coneixements i habilitats per interpretar les dades d'incidents de seguretat.

2.6. Execució del model (interacció)

La primera interacció amb el model, que està disponible com ja s'ha descrit prèviament a la plataforma Hugging Face, s'executa el codi python que comença per descarregar-lo en local en un fitxer que s'anomena "model.safetensors" i que pesa 2,74 GB en disc dur. Un cop descarregat, se'n pot fer us en local, ja s'ha visualitzat a l'entrenament com després de fer el *fine-tune* del model, es desa en el disc dur local de l'ordinador en el que s'executa. A continuació es pot veure el codi que he utilitzat per interactuar amb el model LaLaMini-T5-738M i com s'hi interactua de forma programàtica amb python.

```
# Carreguem la llibreria transformers de hugging face per tal de descarregar i d'interactuar amb el model LLM en local
from transformers import pipeline, AutoModelForSeq2SeqLM, AutoTokenizer

# Defineix el camí on has guardat el teu model entrenat
model_path = './model/'

# Load the tokenizer and model
tokenizer = AutoTokenizer.from_pretrained('MBZUAI/LaMini-T5-738M')
model = AutoModelForSeq2SeqLM.from_pretrained(model_path)

# Inicialitza la pipeline amb el teu model i tokenizer
pipe = pipeline('text2text-generation', model=model, tokenizer=tokenizer)

# Defineix el teu prompt
input_prompt = 'Please let me know your thoughts on the given place and why you think it deserves to be visited: \n"Barcelona, Sp

# Genera el text utilitzant el teu model
generated_text = pipe(input_prompt, max_length=512, do_sample=True)[0]['generated_text']

print("Response:", generated_text)
# Response 06/11/2023: Barcelona is a city known for its rich history, breathtaking architecture, charming neighborhoods, and
# famous nightlife. It deserves to be visited for its historic significance, its gastronomic experiences, the rich culture, and
# the charming atmosphere. Barcelona is also the birthplace of the Spanish bullfighting community.
```

Figura 28 - Interacció inicial amb el model LLM.

2.7. Aplicació FastAPI (backend)

L'aplicació FastAPI es presenta estructurada al voltant de quatre rutes principals, definides en el fitxer main.py. Aquestes rutes proporcionen endpoints específics per a la interacció amb els models de Machine Learning:

1. `analyze_question`: Permet fer preguntes directes al model RL, que a continuació consulta el model LLM per a generar respostes i seleccionar la millor opció.
2. `qa_with_facts`: Similar a `analyze_question`, però incorpora dues pistes addicionals. Això s'ajusta al format d'entrenament del model RL, optimitzant així la selecció de respostes.
3. `train_llm`: Ofereix la funcionalitat per carregar i utilitzar un fitxer CSV per entrenar el model LLM CyberVigilant, actualitzant les versions del model i mantenint les anteriors accessibles.
4. `train_rl`: Permet carregar un fitxer JSONL per l'entrenament del model RL, que avalua les respostes del model LLM CyberVigilant.

Cada ruta interactua amb els models corresponents (RL i LLM) a través dels fitxers situats en la carpeta `./app/routes/`, totes les funcions de backend en les que es treballa amb els models es troben dins el directori `./app/models/`, fins i tot aquí és on s'emmagatzemen els propis fitxers dels models. Els desafiaments tècnics d'aquesta

integració han inclòs la gestió de models complexos i la realització d'operacions numèriques en entorns sense connexió a Internet. Aquesta complexitat s'ha superat amb una planificació acurada i implementacions tècniques adaptades als recursos disponibles.

Per a la gestió d'errors, s'ha implementat un sistema unificat de logs a través de tota l'aplicació. Cada execució del servidor d'aplicacions genera un nou fitxer de log amb una nomenclatura única basada en la data i l'hora, emmagatzemat en `./app/logs/`. Aquests logs registren informació detallada, incloent missatges informatius, de depuració i errors, facilitant així el seguiment, l'anàlisi de problemes i la solució d'errors de manera eficaç. Aquesta aproximació no només millora la traçabilitat i la transparència de l'aplicació, sinó que també contribueix a la seva seguretat i fiabilitat operativa.

A continuació es pot veure l'estructura completa del projecte:

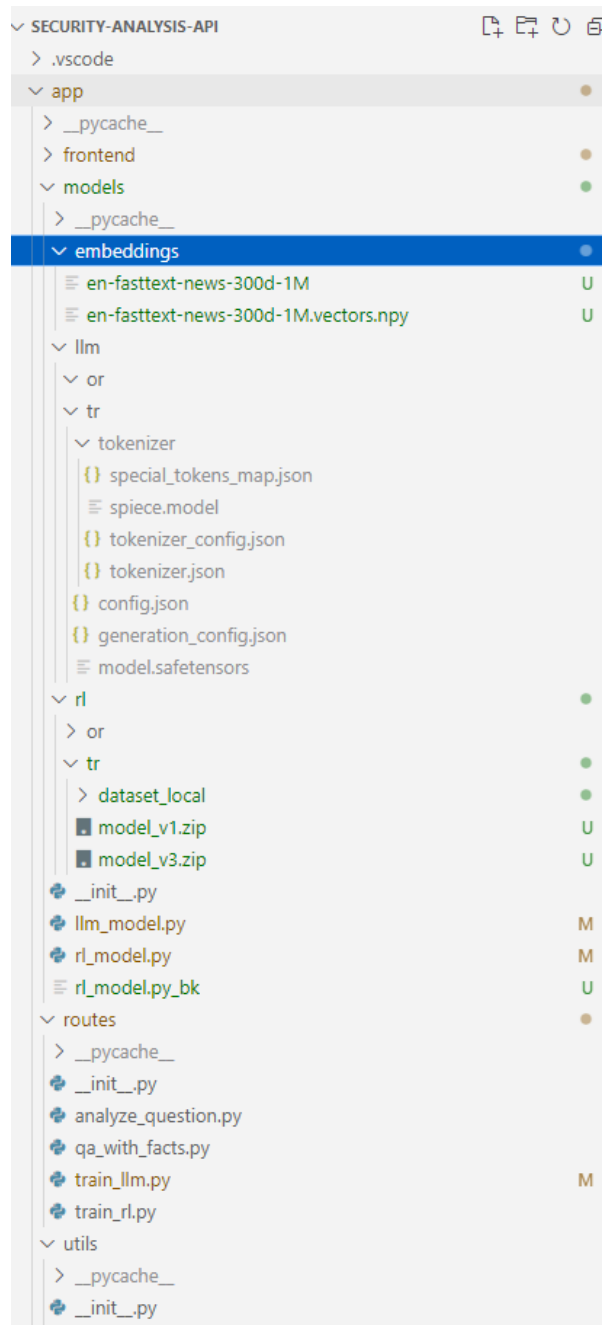


Figura 29 - Estructura del codi font del projecte (1).

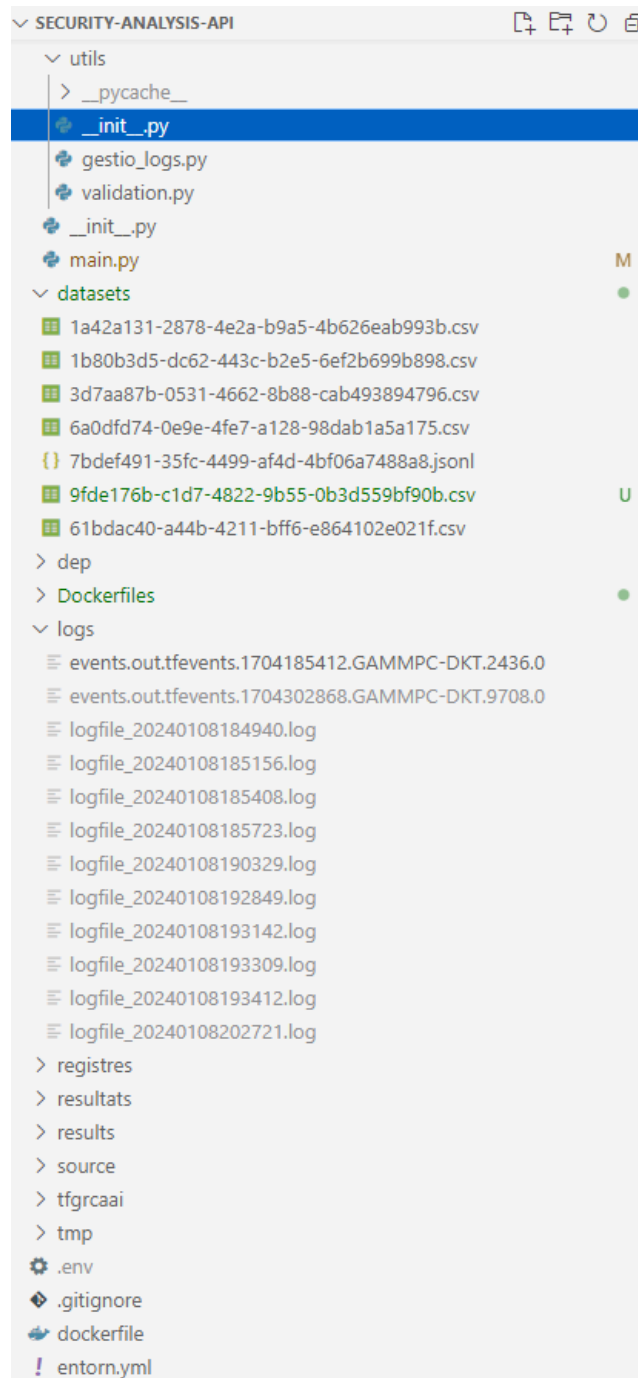


Figura 30 - Estructura del codi font del projecte (2).

2.8. Aplicació Dash (capa de presentació)

L'aplicació Dash s'ha dissenyat per oferir una interfície d'usuari clara i fàcil d'utilitzar, que permet als usuaris interactuar directament amb els models RL i LLM. L'estructura de la pàgina inclou seccions separades per a l'enviament de preguntes al model LLM, l'enviament de preguntes amb pistes, així com per a l'entrenament dels models LLM i RL. S'ha posat especial atenció en assegurar que els camps d'entrada siguin intuïtius i els botons clarament etiquetats, amb l'objectiu de facilitar l'experiència d'usuari i minimitzar possibles confusions.

La integració de Dash amb FastAPI s'ha realitzat mitjançant peticions HTTP POST a l'API des dels callbacks de Dash, permetent així processar les sol·licituds dels usuaris i obtenir respostes dels models. Aquest enfocament assegura una comunicació eficaç entre la interfície d'usuari i la lògica dels models en el servidor. Els desafiaments en aquesta integració han inclòs la gestió de peticions asincròniques i la correcta interpretació de les respostes de l'API, superats mitjançant una curosa gestió d'estats i errors, així com amb la implementació unificada dels callbacks de totes les funcionalitats implementades.

L'aplicació permet als usuaris carregar fitxers (CSV per a LLM i JSONL per a RL) per a l'entrenament dels models. S'ha implementat una lògica per a la càrrega segura i el processament d'aquests fitxers, incloent la generació de noms de fitxers únics per evitar col·lisions i la correcta ubicació dels fitxers en el sistema de fitxers del servidor. A més, s'ha gestionat el procés d'entrenament de manera que es realitzi eficientment i sense afectar la disponibilitat de la interfície d'usuari.

Per a la traçabilitat i el seguiment, s'ha establert un sistema de logs unificat. Cada acció realitzada en l'aplicació, incloent peticions de l'usuari, respostes de l'API i errors, es registra en fitxers de log amb nomenclatura basada en data i hora. Aquesta pràctica no només permet un seguiment detallat de l'activitat de l'aplicació, sinó també facilita la detecció i solució de problemes.

En resum, l'aplicació Dash proporciona una interfície d'usuari potent i intuïtiva per a la interacció amb els models de Machine Learning, alhora que integra eficaçment la lògica de l'API FastAPI i gestiona de forma segura i eficient la càrrega i el processament de fitxers per l'entrenament dels models.

Com a punts de millora, no s'han implementat mètodes per a que l'usuari sigui conscient de l'estat d'execució dels processos de backend en la capa de presentació, els processos de generació de text dels models d'IA són costosos ja que s'executen en ordinadors amb pocs recursos, fet que pot provocar que l'usuari hagi d'esperar un temps sense tenir informació actualitzada de l'estat d'aquestes execucions. Per posar un exemple extrem, com s'ha comentat en altres punts del present treball, el procés d'entrenament del LLM pot tardar més de 4 hores per línia processada, sí que s'han implementat mètodes per a poder mantenir activa la sessió en local i així rebre el missatge quan s'hagi finalitzat el procés.

A continuació es mostra la interfície a través del navegador web:

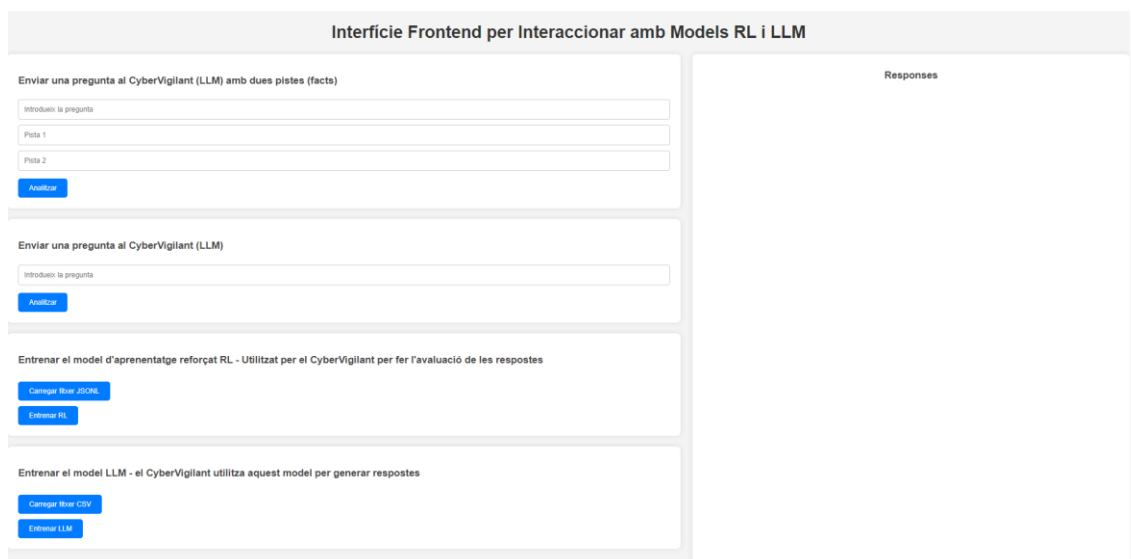


Figura 31 - Interfície web de l'aplicació.

2.9. Funcionament de la Interfície, Logs i Monitorització

Les aplicacions *FastAPI* i *Dash* s'aixequen utilitzant un servidor *Uvicorn*, el qual permet l'accés tant local com des d'altres equips en la mateixa xarxa, amb les configuracions de xarxa i tallafocs pertinents. El desplegament de la solució requereix Python (s'ha provat la versió 3.10.13 publicada a l'agost), Git per gestionar el repositori de GitHub (<https://github.com/arnmamel/security-analysis-api>), i un entorn virtual Python per aïllar les dependències. Els usuaris poden clonar el codi de l'aplicació des de GitHub i instal·lar les dependències necessàries. Després d'executar *Uvicorn*, l'aplicació és accessible a través d'un navegador a les adreces "http://127.0.0.1:8000/dash" o "http://localhost:8000/dash". Els usuaris poden interactuar amb l'aplicació enviant sol·licituds, carregant fitxers i rebent respostes dels models.

S'ha implementat un sistema de logs unificat per a tota l'aplicació. Cada sessió del servidor genera un nou fitxer de log amb una nomenclatura única, que inclou informació detallada sobre les accions realitzades, les sol·licituds dels usuaris, les respostes de l'API i qualsevol error. Aquests logs estan emmagatzemats en una carpeta específica i són crucials per a la solució de problemes i el seguiment de l'activitat de l'aplicació.

Per a la monitorització del rendiment de l'aplicació i dels models, s'utilitzen els logs generats i eines de seguiment integrades en l'entorn de desenvolupament com ara Docker. Aquestes eines permeten visualitzar l'ús dels recursos, el comportament dels models i l'estat de les sol·licituds dels usuaris. Aquesta monitorització ajuda a detectar i solucionar ràpidament problemes, assegurant que l'aplicació funcioni de manera òptima.

A mode il·lustratiu, es mostra una imatge en la que es pot veure com les respostes generades per els nostres models s'integren en la interfície d'usuari:

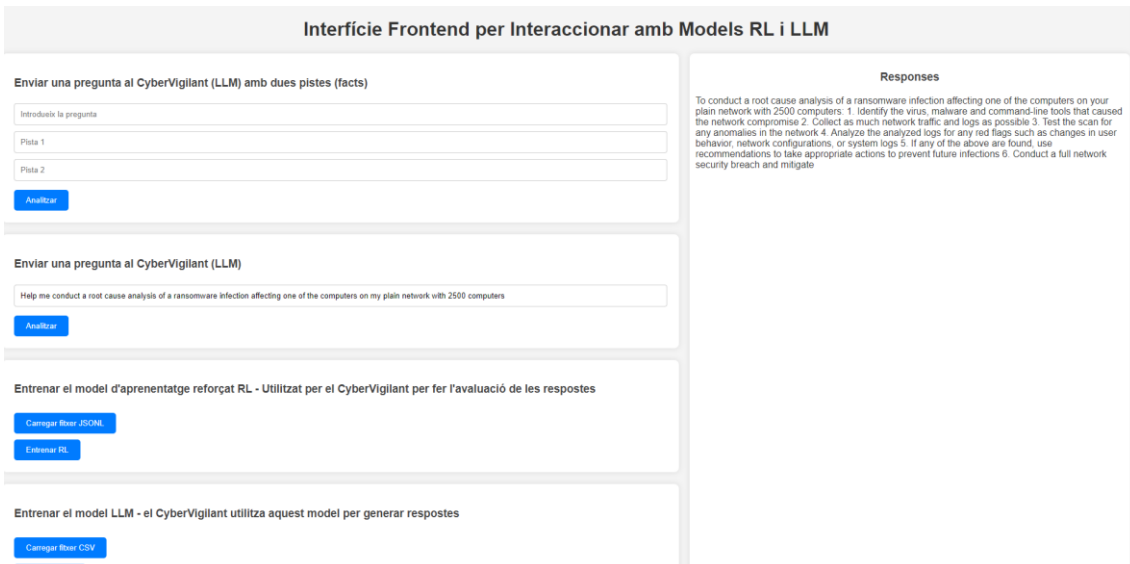


Figura 32 - Interfície web de l'aplicació amb una resposta del model.

En aquest cas, es pot comprovar que les respostes han millorat significativament respecte de les primeres versions del model gràcies, en gran mesura, a la intervenció del model de RL que avalua diverses respostes del nostre CyberVigilant i n'escull la que considera més adient en base a l'entrenament rebut.

Cada cop que es fa una pregunta, la resposta s'afegeix a continuació de l'anterior, amb el que l'usuari manté l'historial de respostes, es veu en la següent imatge.

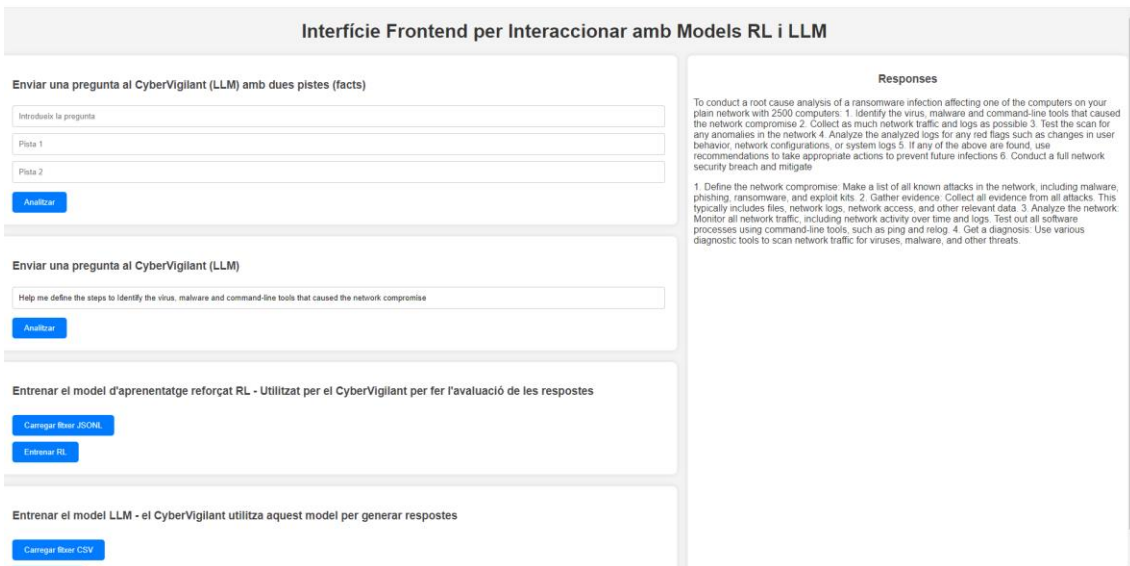


Figura 33 - Interfície web de l'aplicació amb dues respostes.

En aquestes imatges es pot visualitzar com s'interactua amb el model proporcionant només la pregunta, en aquest cas, la implementació fa que el CyberVigilant generi les 2 pistes o que es posi un text fix en cas que cap de les dues opcions anteriors sigui possible, que són dades necessàries per a que el model RL faci l'avaluació de les respostes, a continuació es mostra el flux d'informació:

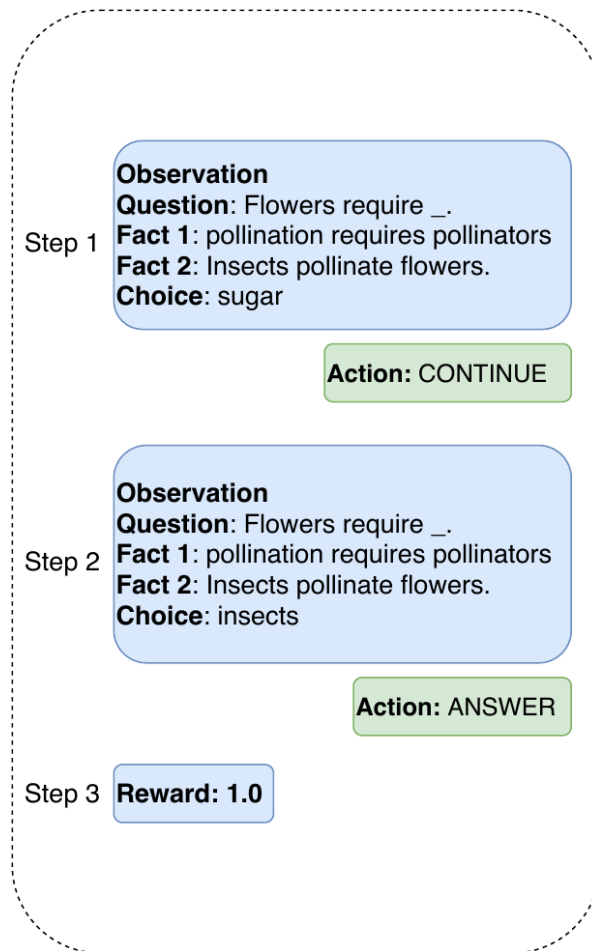


Figura 34 - Codi font mostrant la construcció de les 2 pistes per el model RL [https://github.com/rajcscw/NLPGym].

Finalment, s'ha desenvolupat una funcionalitat que permet consultar el fitxer de logs generat en la interacció amb l'aplicació, el sistema està preparat per desar els últims 10 fitxers de log generats, de forma que comprova si s'ha arribat a aquest número abans de crear un fitxer nou, cas en el que procediria a esborrar el fitxer més antic.

Una de les funcionalitats implementades és la de demanar quatre respostes úniques al model LLM (CyberVigilant), de forma que quan es detecta que s'ha generat una resposta que ja existia, se'n demana una de nova abans d'afegir-la a la llista.

3. Resultats

Eficiència dels models LLM i RL:

El model LLM inicialment presentava una baixa precisió en la identificació de causes arrel en incidents de ciberseguretat. No obstant això, a través d'un entrenament progressiu amb dades reals i incidents específics, la precisió ha millorat notablement, encara que encara existeix marge per a més avanços. La incorporació del model RL ha reforçat significativament la precisió del model LLM. A través de l'entrenament amb més de 150 preguntes basades en casos reals i la selecció de respostes correctes, el model RL ha aconseguit filtrar les respostes menys precises del model LLM, reduint considerablement les respostes invàlides.

Interacció amb els endpoints API:

L'ús dels endpoints API ha millorat notablement l'automatització dels processos, que anteriorment requerien una intervenció manual significativa. A més, aquests endpoints han facilitat la integració amb altres eines i la capacitat d'escalar la solució de manera eficient.

Experiència d'usuari amb el frontend Dash:

La interfície d'usuari Dash ha estat rebuda favorablement pels usuaris, proporcionant una plataforma fàcil i intuïtiva per a la interacció amb els models. Aquesta interfície elimina la necessitat de coneixements tècnics avançats per part dels usuaris finals.

Rendiment i escalabilitat:

El rendiment general del sistema ha estat satisfactori, destacant-se especialment en termes de processament de dades d'entrenament per al model LLM. La dockerització i l'ús d'Uvicorn i Starlette han contribuït positivament al rendiment, encara que el processament dels jocs de dades requereix una quantitat de temps considerable.

Comparació amb altres mètodes o models:

No s'ha realitzat una comparativa directa amb altres models, com ara els basats en BERT de Google, a causa de diferències en els mètodes d'accés i els costos associats.

Desafiaments i limitacions:

El principal desafiament ha estat la limitació dels recursos de computació, que ha restringit l'entrenament del model LLM a dades més extenses i representatives. A més, el temps disponible per a l'entrenament i la preparació dels jocs de dades ha estat un factor limitant.

Implicacions pràctiques:

Els resultats obtinguts ofereixen una eina valuosa per als professionals en els camps de l'anàlisi forense i operacions de ciberseguretat. Encara que actualment està en una fase inicial, amb un entrenament adequat i més temps, aquesta eina té el potencial de ser integrada amb solucions com SIEM, actuant com un assistent en la gestió d'incidents de seguretat. Es preveu que, amb el desenvolupament i entrenament adequats, aquesta eina podria estar llesta per a la seva integració en entorns professionals en un termini de 6-8 mesos.

4. Conclusions i treballs futurs

Després d'analitzar els resultats, es pot concloure que el projecte ha aconseguit integrar amb èxit els models LLM i RL en un entorn de ciberseguretat, tot i que amb certes limitacions. La precisió del model LLM ha millorat amb l'entrenament, però encara hi ha espai per a l'optimització a través de la generació de jocs de dades d'entrenament extretes d'entorns reals. Una opció que queda oberta i podria ajudar a que la eina desenvolupada pogués competir amb models desplegats per grans empreses tecnològiques com el que disposa Google (BERT), és deixar obert l'espai de GitHub i el de HuggingFace per tal que la comunitat de ciberseguretat contribueixi en la millora dels models, tant sigui contribuint amb codi per millorar o optimitzar les funcionalitats existents, com per construir-ne de noves, com sobretot amb dades d'incidentes amb conclusions extretes després de les anàlisi amb la informació contextual, fet que ajudaria molt a millorar els resultats.

La combinació del model LLM amb el model RL ha resultat ser una estratègia eficaç, millorant la selecció de respostes i reduint l'aleatorietat en les respostes que proporcionava en certes ocasions el model LLM. Els resultats obtinguts han estat en línia amb les expectatives, encara que la necessitat de grans quantitats de dades per a l'entrenament ha estat un desafiament major del previst. Per construir els jocs de dades necessaris, s'han destinat un gran nombre d'hores, primer en la recerca de fonts de dades fiables, reproduint i analitzant incidents i programari maliciós i el seu comportament, així com formatant la informació extreta en fitxers que poguessin ser ingerits per part dels mètodes implementats per entrenar els models.

Reflexió crítica sobre l'assoliment dels objectius

Tot i que s'han assolit la gran majoria dels objectius plantejats, alguns aspectes, com l'entrenament més profund dels models amb dades de major qualitat i quantitat, han quedat pendents per restriccions de temps i recursos. Això subratlla la importància de tenir recursos suficients i una planificació realista en projectes d'aquesta naturalesa. Tot i això, s'han superat en altres àmbits els objectius, principalment amb el desenvolupament i el desplegament d'una aplicació completa per interactuar amb els models, el que deixa la opció de connectar sistemes externs amb els mateixos, per exemple per entrenar els models a través de les API.

Anàlisi crítica del seguiment de la planificació i metodologia

La planificació inicial s'ha ajustat al desenvolupament del projecte, tot i que ocasionalment ha estat necessari fer algunes adaptacions, sobretot en relació amb la gestió dels recursos, en l'inici del projecte es va espatllar l'únic ordinador del que disposava, i el temps d'entrenament dels models, que ha comportat més temps del previst precisament per la necessitat de treballar amb un ordinador més antic i menys adaptat als requeriments d'aplicacions intenses com les que s'han gestionat en aquest treball. La metodologia adoptada ha estat adequada.

Impactes ètic-socials, de sostenibilitat i de diversitat

Aquest projecte ha tingut en compte la necessitat de gestionar els recursos computacionals de manera eficient per minimitzar l'impacte ambiental, especialment en termes de consum energètic associat amb l'execució i l'entrenament dels models LLM i RL que habitualment consumeixen grans quantitats de recursos, el projecte desenvolupat es pot executar en un ordinador personal de capacitats mitjanes (s'ha desenvolupat en un ordinador amb més de 10 anys d'antiguitat).

El projecte ha plantejat reptes ètics i socials, sobretot en relació amb la potencial manipulació de la informació per part dels models LLM. S'ha fet un esforç conscient per disposar d'una arquitectura auto continguda, de forma que es pugui tenir major control sobre l'enfocament dels models cap a la gestió d'incidents de ciberseguretat.

En l'àmbit de la diversitat, s'ha posat un èmfasi especial en evitar que els LLM perpetuïn biaixos o discriminacions i, sobretot, s'ha posat èmfasi en desenvolupar un projecte que pugui utilitzar-se en institucions amb pocs recursos econòmics disponibles.

En conclusió, mentre que el projecte ha contribuït significativament en l'àmbit de la ciberseguretat, també ha reconegut i abordat els desafiaments ètics, socials i de sostenibilitat associats amb l'ús d'aquests models avançats. El compromís amb la sostenibilitat, l'ètica i la diversitat ha estat un pilar fonamental en totes les etapes del desenvolupament del projecte.

Impactes no previstos

A l'inici del projecte, es va seleccionar un model LLM de codi obert que, un cop carregat en local en l'ordinador, va fer un excés de consum de recursos per un temps massa llarg durant la seva inicialització i va provocar que es produís una incidència a nivell de maquinari. Inicialment es va procedir a seleccionar un altre model LLM de codi obert i executar-lo en un ordinador antic que estava en desús. Recentment s'ha solucionat amb la substitució de la placa base de l'ordinador portàtil en qüestió.

A part de l'anterior, no s'han identificat impactes no previstos d'importància durant el desenvolupament del projecte.

Línies de treball futur

Pel que fa a les línies de treball futur en el context del projecte de l'ús de models LLM i RL en l'anàlisi de causes arrel en incidents de ciberseguretat, s'identifiquen quatre objectius principals.

En primer lloc, es posa a disposició del públic el codi font desenvolupat i els models afinats en el transcurs del projecte. Aquesta acció permet no només un major escrutini i validació per part de la comunitat científica i professional de la ciberseguretat, sinó també la possibilitat d'enriquir l'entrenament dels models amb un ventall més ampli i divers de dades, en aquest punt és molt rellevant el fet que s'hagin escollit eines i models de codi obert, que permeten dur a terme les accions esmentades.

En segon lloc, es considera clau l'exploració de la integració del sistema amb eines avançades de ciberseguretat, com per exemple amb els sistemes de gestió de la informació i esdeveniments de seguretat (SIEM). Això permetria potenciar l'aplicació

pràctica del projecte, de forma que es podria també oferir una solució més robusta i integrada dins de l'ecosistema de la seguretat informàtica.

El tercer objectiu es centra en el desenvolupament de mètodes automatitzats per a la generació i processament de dades d'entrenament per els models de codi obert que s'han posat a disposició del públic. Tot i que ja s'han desenvolupat funcions per cercar contingut en línia, descarregar-lo i processar-lo per a l'entrenament dels models, les fonts d'informació d'incidents reals són encara molt poques i amb informació, en certes situacions, desfasada, pel que encara persisteixen desafiaments relacionats tant amb l'adaptació dels formats de les dades originals com en la utilització de dades generades en entorns reals.

Finalment i en línia amb els punts anteriors, es planteja la necessitat d'avaluar l'aplicació del sistema desenvolupat en entorns de xarxa reals, proporcionant així una retroalimentació valuosa per a una millora contínua del sistema.

Aquestes línies de treball no només són essencials per a l'avanç del projecte, sinó que també obren noves vies de recerca i aplicació en el camp de la ciberseguretat, marcant el camí cap a una protecció més efectiva i eficient en el món digital.

5. Glossari

- **AI-lucinacions en models LLM:** Es refereixen a quan aquests models generen respostes o textos que no estan basats en dades reals o que són incorrectes, sense sentit o fora de context. Aquest fenomen sol ocórrer quan el model interpreta erròniament la informació d'entrada o quan generalitza de manera inadequada a partir del seu entrenament.
- **Anàlisi de causes arrel:** Mètode d'investigació utilitzat per identificar la causa fonamental d'un problema o incident.
- **API RESTful (Application Programming Interface):** Una interfície de programació d'aplicacions que utilitza HTTP per comunicar-se amb aplicacions o serveis web.
- **Aplicació frontal:** Interfície d'usuari que permet la interacció amb una aplicació, sovint accessible a través d'un navegador web.
- **Aprenentatge profund (*Deep Learning* en anglès):** Branca de l'aprenentatge automàtic que utilitza xarxes neuronals amb moltes capes per aprendre representacions de dades amb nivells creixents d'abstracció. Aquest mètode permet a models computacionals aprendre directament de grans quantitats de dades no estructurades, millorant la seva capacitat de reconeixement, classificació i predicció.
- **Automatització:** Procés d'utilització de sistemes o programari per realitzar tasques repetitives sense intervenció humana.

- **Ciberseguretat:** Pràctica de protegir els sistemes informàtics, xarxes i dades dels atacs, danys o accessos no autoritzats.
- **Command and Control (C2):** En ciberseguretat, es refereix als servidors utilitzats per controlar ordinadors infectats, com en el cas de xarxes de bots o *malware*. Aquests servidors C2 envien ordres als dispositius compromesos i reben dades d'ells, actuant com a centres de comandament per a activitats malicioses com l'enviament de *spam*, el llançament d'atacs *DDoS*, el furt de dades, entre altres.
- **Contenidorització:** Tècnica d'empaquetament de programari que inclou tot el necessari per executar una aplicació, com el codi, les biblioteques i les dependències, en un "contenidor" estàndard.
- **Dades preprocessades:** Dades que han estat preparades o transformades en un format específic per a ser utilitzades eficaçment en processos d'anàlisi o models d'aprenentatge automàtic.
- **Dash:** Framework per a crear aplicacions web interactives en Python.
- **Deep Q-Network (DQN):** Un DQN és un tipus d'algoritme d'aprenentatge profund per reforç que combina xarxes neuronals amb un mètode de control basat en Q-learning, utilitzat principalment en problemes de decisió seqüencial.
- **Dependències lingüístiques i semàntiques:** Relacions entre paraules i frases en un idioma que determinen el seu significat i estructura.
- **Desenvolupament d'una aplicació en Python:** Creació d'un programari utilitzant el llenguatge de programació Python.
- **Desplegament:** Procés de posar una aplicació o servei a disposició dels usuaris, típicament en un entorn de producció.
- **DevOps (operacions de desenvolupament):** Conjunt de pràctiques destinades a millorar la col·laboració entre els equips de desenvolupament de programari i operacions.
- **Dockerització:** Procés de creació d'imatges Docker per a aplicacions, facilitant la seva distribució i desplegament.
- **Endpoints:** Punts finals d'una API que proporcionen accés a funcions específiques del programari.
- **FastAPI:** Framework modern i ràpid per a construir API amb Python, basat en estàndards com OpenAPI.
- **Fitxer csv:** Format de fitxer que utilitza valors separats per comes per emmagatzemar dades tabulars.
- **Fitxer jsonl:** Format de fitxer que conté dades en format JSON, amb cada objecte JSON en una línia separada.

- **Generalització:** Capacitat d'un model d'aprenentatge automàtic de realitzar prediccions o conclusions precises en dades noves que no s'han utilitzat durant l'entrenament.
- **GPU (Unitats de Processament Gràfic):** Dispositius especialitzats en processar gràfics i imatges, però també utilitzats en l'aprenentatge automàtic per la seva capacitat de realitzar càlculs paral·lels ràpidament.
- **Grans Models Lingüístics (LLM):** Models d'intel·ligència artificial de gran escala dissenyats per comprendre, generar i manipular llenguatge humà. Són models neurals que aprenen a partir de grans conjunts de dades.
- **Hugging Face:** Plataforma coneguda per allotjar models d'aprenentatge automàtic, especialment models de processament de llenguatge natural.
- **Inferència:** En el context de l'aprenentatge automàtic, es refereix a l'ús d'un model entrenat per fer prediccions o prendre decisions basades en noves dades.
- **Interoperabilitat:** Capacitat de diferents sistemes per interactuar i funcionar conjuntament.
- **Kanban:** Metodologia d'organització i gestió de projectes que utilitza un sistema visual per representar les tasques i el seu estat dins del procés de treball.
- **MBZUAI/LaLaMini-T5-738M:** Model específic disponible a la plataforma Hugging Face, utilitzat per processar llenguatge, ha estat la base del model que s'ha afinat durant el projecte.
- **MLOps (operacions d'aprenentatge automàtic):** Pràctiques d'enginyeria destinades a agilitzar el cicle de vida dels models d'aprenentatge automàtic, des de la creació fins a la producció.
- **Polítiques d'acció en entorns de decisió:** Les polítiques d'acció en entorns de decisió en l'aprenentatge profund són estratègies utilitzades per agents intel·ligents per decidir quines accions realitzar en un determinat entorn. Aquestes polítiques es basen en models d'aprenentatge profund que permeten a l'agent aprendre i adaptar-se a l'entorn a través de l'experiència, amb l'objectiu d'optimitzar alguna mesura de rendiment o recompensa.
- **Proximal Policy Optimization (PPO):** PPO és un algoritme d'aprenentatge per reforç que busca optimitzar les polítiques d'acció en entorns de decisió, equilibrant l'eficàcia dels passos d'aprenentatge amb la estabilitat de l'entrenament.
- **Q-learning:** Q-learning és un algoritme d'aprenentatge automàtic basat en el principi de reforç, on un agent aprèn a determinar l'acció òptima a realitzar en un determinat estat, amb l'objectiu de maximitzar la recompensa total obtinguda a llarg termini.
- **Question-Answering:** És un enfocament on un agent d'aprenentatge per reforç aprèn a respondre preguntes basades en text. En aquests entorns, l'agent

interactua amb un conjunt de dades de text, rebent recompenses basades en l'exactitud i la pertinència de les seves respostes, amb l'objectiu de millorar la seva capacitat de comprensió i generació de respostes correctes a preguntes formulades en llenguatge natural.

- **Reinforcement Learning / RL (aprenentatge per reforç):** Tipus d'aprenentatge automàtic on un agent aprèn a prendre decisions optimitzant recompenses en un entorn determinat.
- **Re-entrenament:** Procés d'entrenar un model d'aprenentatge automàtic ja existent amb noves dades per millorar-ne el rendiment o adaptar-lo a noves tasques.
- **Scrapping (web):** *Scrapping* és el procés d'extracció automàtica d'informació des de llocs web utilitzant programes o scripts, convertint dades no estructurades en dades estructurades per al seu posterior anàlisi.
- **TPU (Unitats de Processament Tensor):** Processadors desenvolupats per Google específicament per accelerar tasques d'aprenentatge automàtic.
- **Transfer learning (transferència d'aprenentatge):** Tècnica en l'aprenentatge automàtic on un model desenvolupat per a una tasca es reutilitza com a punt de partida per a una altra tasca.
- **Trello:** Eina en línia que permet la creació i gestió de taulells Kanban per a la planificació i seguiment de projectes.
- **TTP (Tàctiques, Tècniques i Procediments):** TTP són els mètodes i pràctiques utilitzades per actors de ciberseguretat o adversaris per dur a terme les seves missions o atacs, incloent els passos específics i els mètodes operatius.
- **Unit42 de Paloalto:** Un equip d'investigació de seguretat cibernètica conegut per proporcionar informació i dades sobre amenaces de seguretat.

6. Bibliografía

- [1] Brockman, G. (2016, 5 junio). *OpenAI Gym*. arXiv.org. <https://arxiv.org/abs/1606.01540>
- [2] Ferrag, M. A. (2023, 25 junio). *Revolutionizing cyber threat detection with large language models*. arXiv.org. <https://arxiv.org/abs/2306.14263>
- [3] Hughes, A. (2023, 11 mayo). *Large-language models for automatic cloud incident Management - Microsoft Research*. Microsoft Research. <https://www.microsoft.com/en-us/research/blog/large-language-models-for-automatic-cloud-incident-management/>
- [4] Katal, A., Bajoria, V., & Dahiya, S. (2019). DevOps: Bridging the gap between Development and Operations. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1-7, doi: 10.1109/ICCMC.2019.8819631. <https://ieeexplore.ieee.org/document/8819631>
- [5] McKinney, W. (2010). *Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference*. https://www.researchgate.net/publication/340177686_Data_Structures_for_Statistical_Computing_in_Python
- [6] *Papers with code - language models are unsupervised multitask learners*. (2019, 14 febrero). <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask>
- [7] Quentin, L., et al. (2021). *Datasets: A Community Library for Natural Language Processing*. <https://arxiv.org/abs/2109.02846>
- [8] Raffin, A. (2021). *Stable-Baselines3: Reliable Reinforcement learning Implementations*. <https://jmlr.org/papers/v22/20-1364.html>
- [9] Rodriguez, A. (2019). MLOps: Toward Effective Implementation of Machine Learning Models. International Conference on Software Engineering Proceedings.
- [10] Sessions, N. (2023, December 14). *Bolstering cybersecurity: How large language models and Generative AI are Transforming Digital Security*. NVIDIA Technical Blog. <https://developer.nvidia.com/blog/bolstering-cybersecurity-how-large-language-models-and-generative-ai-are-transforming-digital-security/>
- [11] Web: *Malware-traffic-analysis.net*. (s. f.). malware-traffic-analysis.net. <https://malware-traffic-analysis.net/>
- [12] Web: YourHub4Tech. (2023, 3 noviembre). *Comparing the top generative AI tools! - illumination's mirror - Medium*. <https://medium.com/illuminations-mirror/comparing-the-top-generative-ai-tools-5fea222b603e>

- [13] Wolf et al., EMNLP (2020). *Transformers: State-of-the-Art Natural Language Processing*. <https://aclanthology.org/2020.emnlp-demos.6>
- [14] Wu, M., Waheed, A., Zhang, C., Abdul-Mageed, M., & Aji, A. F. (2023). LAMini-LM: a diverse herd of distilled models from Large-Scale instructions. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2304.14402>
- [15] Web: Wu, M., Waheed, A., Zhang, C., Abdul-Mageed, M., & Aji, A. F. (2023). *MBZUAI/LALAMINI-T5-738M*. Hugging face. (s. f.). <https://huggingface.co/MBZUAI/LaLaMini-T5-738M>

7. Annexos

7.1 Annex 1: Manual d'administració

1. Introducció a l'aplicació

En aquest apartat es descriuen les passes a seguir, des de la perspectiva d'un usuari avançat (donat que l'eina per a l'aplicació de ciberseguretat que integra els models LLM i RL - CyberVigilant), per instal·lar i desplegar l'aplicació en un ordinador. Aquesta aplicació, desenvolupada amb FastAPI i Dash, ofereix una interfície robusta i intuïtiva per interactuar amb aquests models avançats d'intel·ligència artificial.

Per fer el desplegament de l'aplicació, és necessari disposar de connectivitat a Internet, ja que caldrà accedir a repositoris en línia de GitHub per descarregar codi font i dependències, així com també de HuggingFaces per descarregar els models que s'han entrenat.

A nivell de programari, l'aplicació i els models són compatibles amb arquitectures Intel, des de la seva gamma de i5, i AMD , a partir dels Ryzen 5 actuals. Durant l'entrenament dels models, moment en què més recursos es consumeixen, es poden arribar a requerir 16GB de memòria RAM, es recomana executar l'aplicació sobre un disc d'estat sòlid.

Es recomana utilitzar un sistema operatiu basat en Windows al que s'hi ha de d'instal·lar python 3.10.13 (a través de la web oficial de python), Visual Studio Code (VSC) i git, que es pot desplegar amb l'extensió disponible dins el propi VSC. També es fa us de l'explorador de fitxers o de l'eina cmd.exe o powershell per treballar amb fitxers al sistema operatiu.

2. Configuració i desplegament

Per desplegar l'aplicació, cal seguir aquests passos:

Clonar, fent ús de Git, el repositori principal del projecte:

Fer ús de:

```
git clone https://github.com/arnmamel/security-analysis-api
```

per clonar el repositori en la màquina.

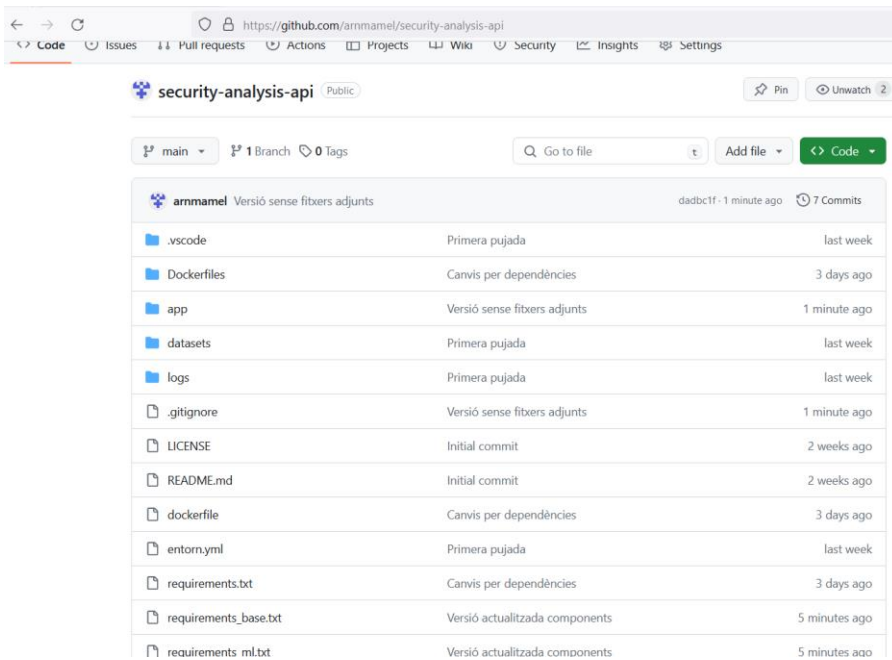


Figura 35 - Repositori públic del projecte a GitHub.

Configurar entorn virtual:

En una línia de comandaments o terminal, per exemple dins de `tfgrcaai`, crea un entorn virtual amb VSC:

```
python -m venv tfgrcaai
```

activa-ho amb

```
tfgrcaai\Scripts\activate.
```

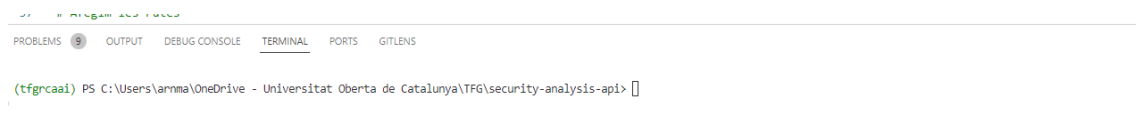


Figura 36 - Terminal d'execució de comandes python.

Instal·lar Dependències:

Executa

```
pip install --no-cache-dir -r requirements_base.txt && pip install --no-cache-dir -r requirements_ml.txt && pip install --no-cache-dir -r requirements_ml2.txt
```

per instal·lar les dependències necessàries (aproximadament 5GB).

Instal·lar Uvicorn: Assegura't que Uvicorn estigui instal·lat amb

```
pip install uvicorn
```

Instal·lar NLPGYM: Navega fins a la carpeta `./dep/` i executa

`git clone https://github.com/arnmamel/NLPGym`

Executar servidor d'aplicacions: des del terminal, utilitza

`uvicorn app.main:app --reload`

per executar el servidor.

```
(tfgcaai) PS C:\Users\arnma\OneDrive - Universitat Oberta de Catalunya\TFG\security-analysis-api>
* History restored

PS C:\Users\arnma\OneDrive - Universitat Oberta de Catalunya\TFG\security-analys
INFO: Will watch for changes in these directories: ['C:\Users\arnma\OneDrive - Universitat Oberta de Catalunya\TFG\security-analysis-api']
INFO: Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)
INFO: Started reloader process [9964] using statreload
2024-01-09 20:50:49.306747: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'cudart64_110.dll'; dlderror: cudart64_110.dll not found
2024-01-09 20:50:49.307334: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if you do not have a GPU set up on your machine.
INFO: Started server process [2580]
INFO: Waiting for application startup.
INFO: Application startup complete.
WARNING: StatReload detected file change in 'app/main.py'. Reloading...
```

Figura 37 - Terminal d'execució de comandes python (aixecant servidor uvicorn).

Accedir a l'Aplicació: Obre un navegador i visita <http://127.0.0.1:8000/dash> o <http://localhost:8000/dash>.

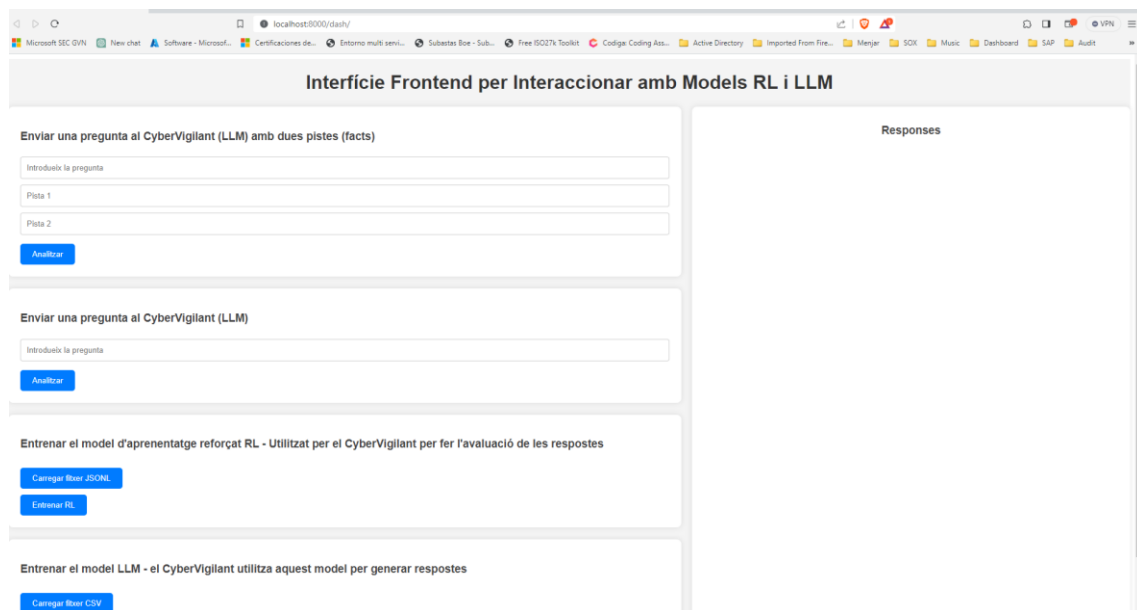


Figura 38 - Interfície web de l'aplicació.

Descarrega i desplaçament dels models en local

Per tal que el projecte funcioni correctament, és necessari descarregar els models, això es faria habitualment des de huggingface, però com que no es disposa dels comptes d'usuari necessaris, s'ha optat per utilitzar l'eina Google Drive amb el compte de la UOC per a posar-los a disposició del professorat, companyes i companys, a part també he carregat les models a huggingface per a que la comunitat en pugui fer ús.

Els fitxers es poden obtenir en [aquest enllaç](#).

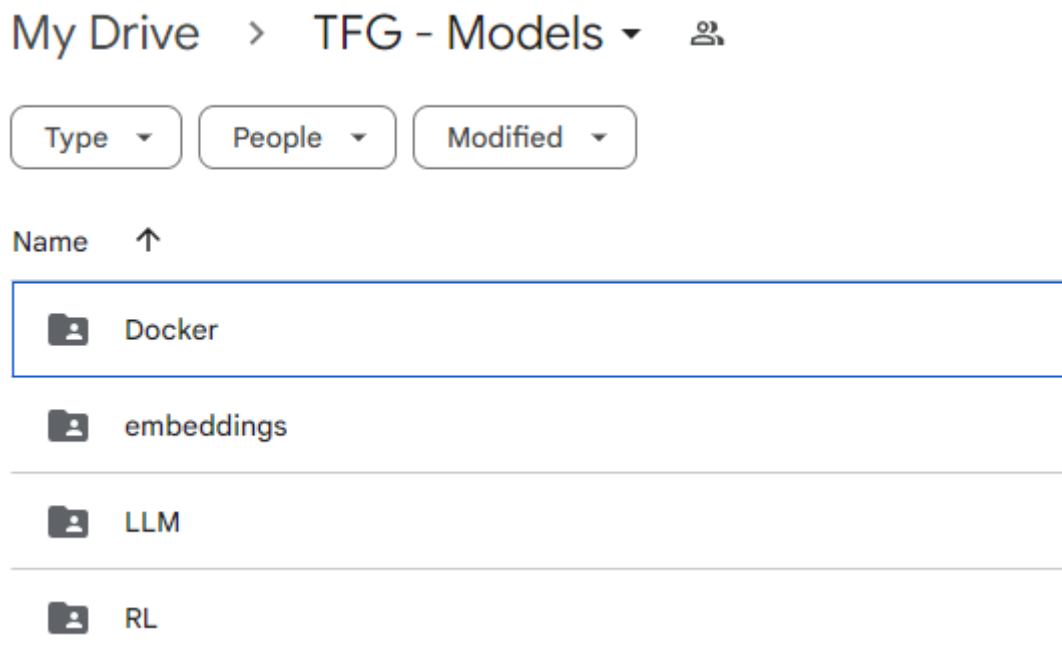


Figura 39 - Estructura de fitxers a l'espai Google Drive compartit amb membres de la UOC.

El contingut de la carpeta del drive “*embeddings*” (sense la carpeta en sí) s’ha de posar dins de “*./app/models/embeddings/*”.

El contingut de la carpeta del drive “*LLM*” (sense la carpeta en sí) s’ha de posar dins de “*./app/models/llm/tr/*”.

El contingut de la carpeta del drive “*RL*” (sense la carpeta en sí) s’ha de posar dins de “*./app/models/rl/tr/*”.

3. Utilització de l'aplicació FastAPI

L'aplicació FastAPI inclou quatre endpoints principals:

- Analyze question: Envia preguntes al model RL que consulta el model LLM per generar i seleccionar respostes.
- QA with facts: Similar a l'anterior, però afegeix dues pistes per optimitzar la selecció de respostes.
- Train LLM: Permet carregar un fitxer CSV i entrenar el model LLM.
- Train RL: Permet carregar un fitxer JSONL i entrenar el model RL.

4. Utilització de l'Aplicació Dash (Capa de Presentació)

La interfície d'usuari Dash permet interactuar amb els models de manera senzilla:

- Enviar Preguntes: Fes servir els camps d'entrada per enviar preguntes als models.
- Carregar Fitxers per a l'Entrenament: Utilitza la funcionalitat de càrrega per aportar fitxers CSV o JSONL per entrenar els models.
- Visualització de Respostes: Les respostes dels models es mostren de forma clara i accessible.

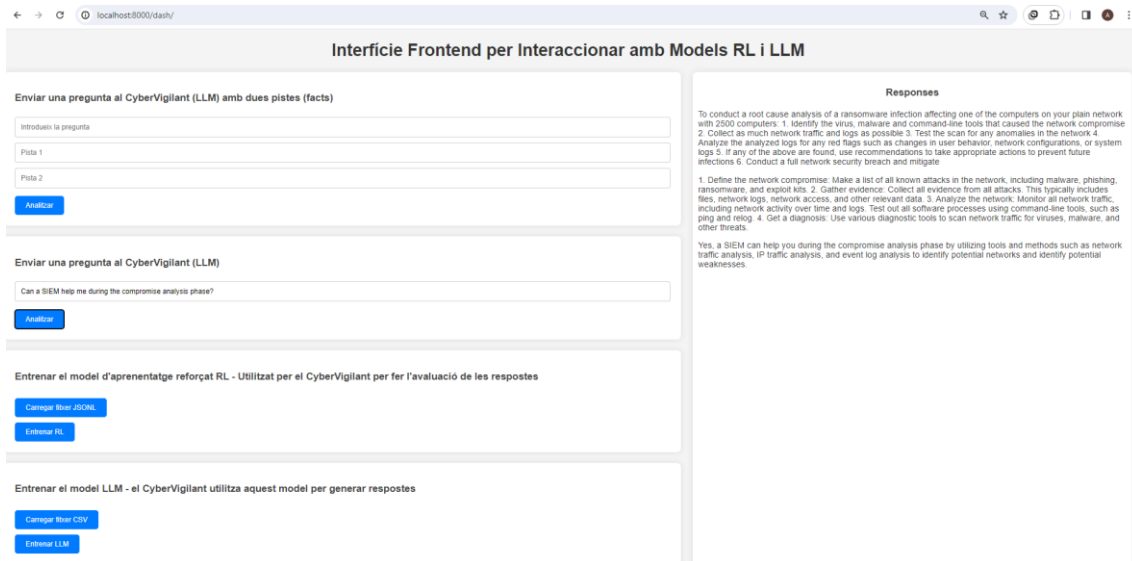


Figura 40 - Interfície web de l'aplicació.

5. Monitorització i Logs

L'aplicació genera registres d'accions detallats en cada sessió de l'aplicació (en el moment d'iniciar-la), que inclouen informació sobre accions, sol·licituds, respostes i errors. Aquests logs es guarden a la carpeta `./app/logs/` i faciliten la investigació i resolució de problemes, així com també el seguiment de l'activitat dins el context de l'aplicació.

security-analysis-api > logs

Nombre	Fecha de modificación	Tipo	Tamaño
logfile_20240103182246.log	03/01/2024 18:31	Documento de te...	11 KB
events.out.tfevents.1704302868.GAMMP...	03/01/2024 18:27	Archivo 0	6 KB
logfile_20240103175215.log	03/01/2024 17:53	Documento de te...	9 KB
logfile_20240103175141.log	03/01/2024 17:51	Documento de te...	1 KB
logfile_20240103175106.log	03/01/2024 17:51	Documento de te...	1 KB
logfile_20240103175035.log	03/01/2024 17:50	Documento de te...	1 KB
logfile_20240103174601.log	03/01/2024 17:47	Documento de te...	9 KB
logfile_20240103174208.log	03/01/2024 17:43	Documento de te...	9 KB
logfile_20240103173941.log	03/01/2024 17:39	Documento de te...	1 KB
logfile_20240103173906.log	03/01/2024 17:39	Documento de te...	1 KB
logfile_20240103173518.log	03/01/2024 17:35	Documento de te...	1 KB
events.out.tfevents.1704185412.GAMMP...	02/01/2024 10:11	Archivo 0	7 KB

Figura 41 - Contingut carpeta de logs de l'aplicació.

Per veure el contingut dels registres, es pot fer ús de qualsevol editor de text.

Cal tenir en compte que s'emmagatzemen els deu últims fitxers `.log` generats, els anteriors s'esborren de forma automàtica, aquesta comprovació es fa cada cop que es genera un nou fitxer.

Per un altre costat, en aquest directori també s'hi emmagatzemen registres d'utilització dels models d'intel·ligència artificial, les dades generades no són rellevants a nivell del projecte, per el que es poden esborrar si causen problemes en algun moment del cicle de vida de l'aplicació.

6. Descarrega dels models

Com he comentat anteriorment, el model LLM també s'ha carregat a [huggingface](https://huggingface.co):

The screenshot shows the Hugging Face interface for the model 'tfg-rca/ciber-responder'. The page includes a search bar, navigation tabs for Models, Datasets, Spaces, Docs, Solutions, and Pricing. The model card displays 'Text2Text Generation', 'Transformers', 'Safetensors', 't5', 'Inference Endpoints', and 'text-generation-inference' with a 'License: gpl-3.0'. Below the card, the 'Files and versions' tab is active, showing a list of files in the 'main' branch:

File Name	Size	Upload Method	Time
tfg-rca	Upload 3 files	#f8c267	about 1 month ago
.gitattributes	1.52 KB	initial commit	about 1 month ago
README.md	25 Bytes	initial commit	about 1 month ago
config.json	1.51 KB	Upload 3 files	about 1 month ago
generation_config.json	142 Bytes	Upload 3 files	about 1 month ago
model.safetensors	2.95 GB	Upload 3 files	about 1 month ago

Figura 42 - Espai a Hugging Face amb el model LLM entrenat en aquest projecte.

7. Problemes Comuns i Solucions

- El model LLM es pot entrenar amb un document csv seguint el format que es mostra en aquest treball: {"input";"target"}, el que equival a pregunta, resposta. L'entrenament d'un model d'aquest tipus no és trivial, pel que no és necessari entrenar-lo amb totes les possibles combinacions.

input;target

"infection chain of TA577 ICEDID (BOKBOT);"thread-hijacked email, which contains a link for zip download, which gets into the downloaded zip, which is extracted into a .js file and then retrieves IcedID installer, this gets to IcedID infection with BackConnect traffic: <https://www.malware-traffic-analysis.net/2023/10/16/index.html>"

- El model RL es pot entrenar amb un document jsonl, seguint el format que es pot veure en aquest treball com en el cas anterior. El fitxer per entrenar aquest model és més complex de confeccionar i requereix coneixements més específics:

```
{"id": "1", "question": {"stem": "What was the initial attack vector in the DEV-0569 activity on 2023-02-03?", "choices": [{"text": "Email phishing", "label": "A"}, {"text": "Malicious Google Ad", "label": "B"}, {"text": "USB device", "label": "C"}, {"text": "Network intrusion", "label": "D"}, {"text": "Website vulnerability", "label": "E"}]}, "answerKey": "B", "fact1": "The DEV-0569 activity began with a malicious Google Ad.", "fact2": "Google Ads are a common vector for cyber attacks.", "combinedfact": "The DEV-0569 malware infection initiated through a malicious Google Ad.", "formatted_question": "What was the initial attack vector in the DEV-0569 activity on 2023-02-03? (A) Email phishing (B) Malicious Google Ad (C) USB device (D) Network intrusion (E) Website vulnerability"}
```

- Lentitud en l'Entrenament: Degut a limitacions de recursos, l'entrenament pot ser lent. Es recomana paciència i planificació.
- Errors en la Càrrega de Fitxers: Assegura't que els formats i els continguts dels fitxers siguin correctes abans de carregar-los.
- Problemes d'Accés: Verifica les configuracions de xarxa i tallafocs si no pots accedir a l'aplicació des d'altres equips.

8. Suport i Contacte

Per a qualsevol dubte o problema, es poden posar en contacte a través de l'espai de suport a GitHub.

Nota: aquest manual està subjecte a actualitzacions per reflectir millores i canvis en l'aplicació.

7.2 Annex 2: Manual d'instruccions per generar una imatge en Docker

Una part important del projecte ha estat paquetitzar la solució desplegada fent ús de *Docker*, en aquest sentit, es descriuen a continuació les passes seguides per crear aquesta imatge, des de la que es poden aixecar contenidors que continguin tot el necessari per només haver d'accedir al navegador i interactuar amb l'aplicació.

Aquesta solució no és òptima per a ser utilitzada en una aplicació que interactua amb models LLM, per el que l'experiència de l'usuari es veurà afectada si se'n fa ús (es recomana seguir el mètode descrit a l'annex 1 del present document), ja que requereixen consumir un nombre elevat de recursos, fet per el que Docker no està dissenyat. En qualsevol cas, s'ha pogut dur a terme aquesta tasca com es mostra a continuació.

En primer lloc, s'ha generat un fitxer "*dockerfile*" que conté les instruccions que seguirà Docker per crear la imatge, a continuació es descriuen els continguts del fitxer generat, que està disponible també al repositori de github del projecte:

1. FROM python:3.10.13: Aquesta línia especifica la imatge base que s'utilitzarà. En aquest cas, és una imatge oficial de Python 3.10.13.
2. WORKDIR /usr/src/sec-ai: Estableix el directori de treball dins del contenidor com a /usr/src/sec-ai.
3. RUN apt-get update && apt-get install -y git: Actualitza els paquets i instal·la Git dins del contenidor.
4. COPY requirements_*.txt /usr/src/sec-ai/: Copia els arxius de requisits de Python al directori de treball del contenidor.
5. RUN pip install --no-cache-dir -r requirements_*.txt: Instal·la les dependències de Python especificades en els arxius de requisits.
6. COPY dep /usr/src/sec-ai/dep: Copia el directori dep al contenidor.
7. RUN pip install /usr/src/sec-ai/dep/NLPGym: Instal·la un paquet local de Python situat dins del directori dep.
8. ENV PYTHONPATH="\${PYTHONPATH}:/usr/src/sec-ai/dep": Configura la variable d'entorn PYTHONPATH per incloure el directori dep, que és on es copia el *fork* que s'ha generat de NLPGym.
9. COPY app logs datasets tmp /usr/src/sec-ai/: Copia els directoris app, logs, datasets i tmp al contenidor.
10. RUN pip install uvicorn: Instal·la Uvicorn, un servidor ASGI per a aplicacions web.
11. EXPOSE 8000: Exposa el port 8000 del contenidor.
12. CMD ["uvicorn", "app.main:app", "--host", "0.0.0.0", "--port", "8000", "--reload"]: Defineix el comandament per defecte que s'executarà quan s'iniciï el contenidor.

13. LABEL author="Arnau Mata Melià" ...: Afegeix metadades a la imatge Docker.

Un cop descrit els continguts del fitxer dockerfile, es procedeix a revisar les passes a seguir per desplegar la imatge Docker:

1. Instal·lar docker desktop: Docker desktop ha d'estar instal·lat en l'equip Windows. Es pot descarregar des de Docker Hub.
2. Obrir powershell: Navegar al directori on es troba el Dockerfile i obrir PowerShell en aquesta ubicació.
3. Construir la imatge: Executar el següent comandament per construir la imatge Docker:

```
docker build -t sec-ana-ai .
```

4. Executar el contenidor amb requisits específics: Utilitzar el següent comandament per executar un contenidor amb un mínim de 4GB de RAM i 4 CPUs:

```
docker run -d --name sec-macgiver -p 8000:8000 --memory 4g --cpus 4 nom-sec-ana-ai
```

Explicacions addicionals de les comandes:

- *docker build -t sec-ana-ai .*: Aquest comandament construeix una imatge Docker a partir del Dockerfile en el directori actual (indicat per *.*) i l'etiqueta amb el nom especificat.
- *docker run*: Aquest comandament crea i inicia un contenidor basat en la imatge especificada.
- *--name*: Assigna un nom al contenidor per a una identificació fàcil.
- *-p 8000:8000*: Mapeja el port 8000 del contenidor al port 8000 de l'amfitrió.
- *--memory 4g*: Assigna 4GB de memòria al contenidor.
- *--cpus 4*: Assigna 4 CPUs al contenidor.
- *-d*: Executa el contenidor en mode desacoplat (en segon pla).

Per facilitar aquestes passes, s'ha generat una imatge docker amb tots els requeriments complerts i s'ha emmagatzemant també a l'espai de google drive del projecte. S'ha pogut fer seguint les següents passes:

1. Obrir PowerShell: S'ha obert PowerShell a l'equip Windows amb la instal·lació de Docker i on està present la imatge Docker a exportar
2. S'ha fet ús de la comanda "docker save" per guardar la imatge generada en un fitxer amb extensió ".tar":

```
docker save -o tfg-rcaai.tar sec-ana-ai
```

Per carregar aquesta imatge en una instal·lació Docker, només cal seguir les instruccions que es mostren a continuació:

1. Descarregar la imatge del google drive del projecte: [enllaç](#).
2. Obrir powershell o un terminal.
3. Carregar la imatge

```
docker load -i tfg-rcaai.tar
```

4. Verificar la imatge mitjançant l'execució de la següent comanda:

```
docker images
```

A continuació, es mostra el resultat d'executar les comandes anteriors, que també es poden dur a terme mitjançant la interfície gràfica que ofereix Docker Desktop al sistema operatiu Window:

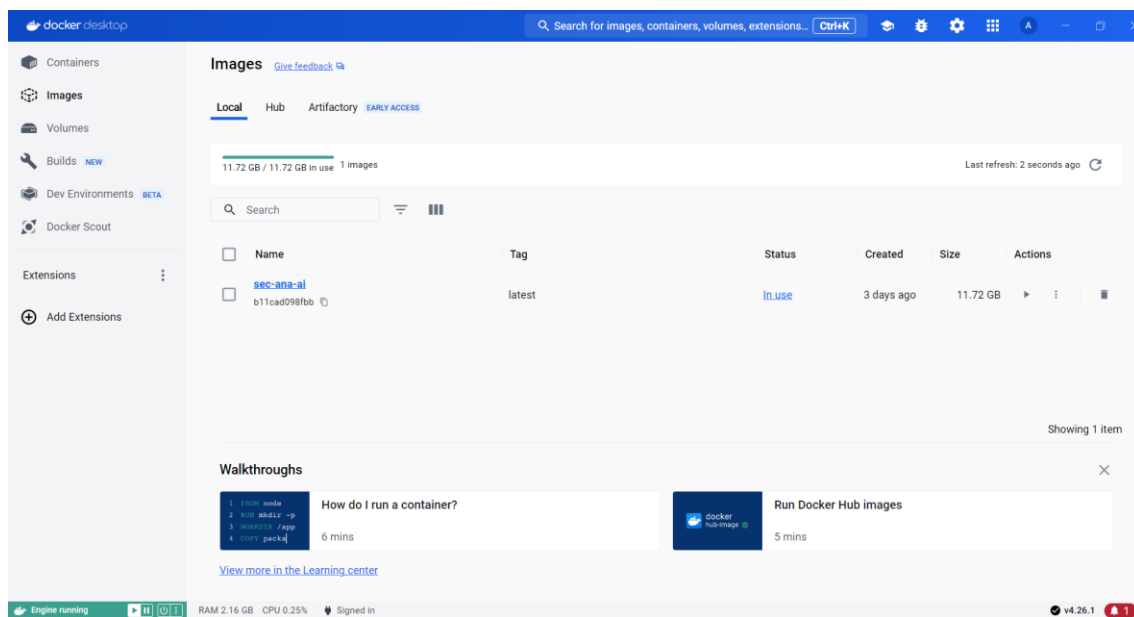


Figura 43 - Imatge Docker carregada.

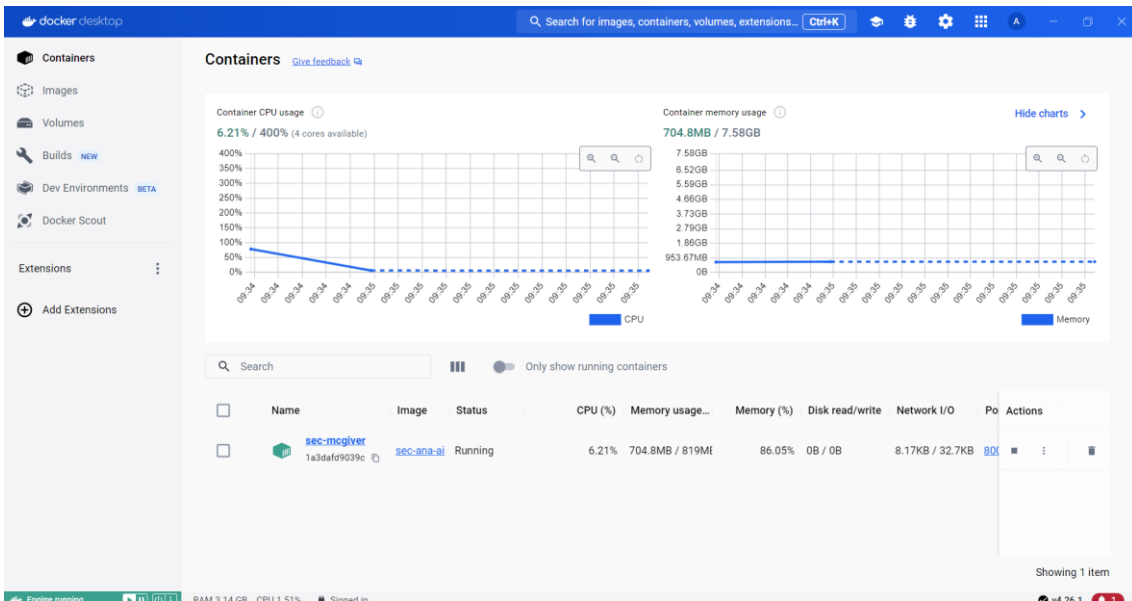


Figura 44 - Contenedor Docker en ejecución.

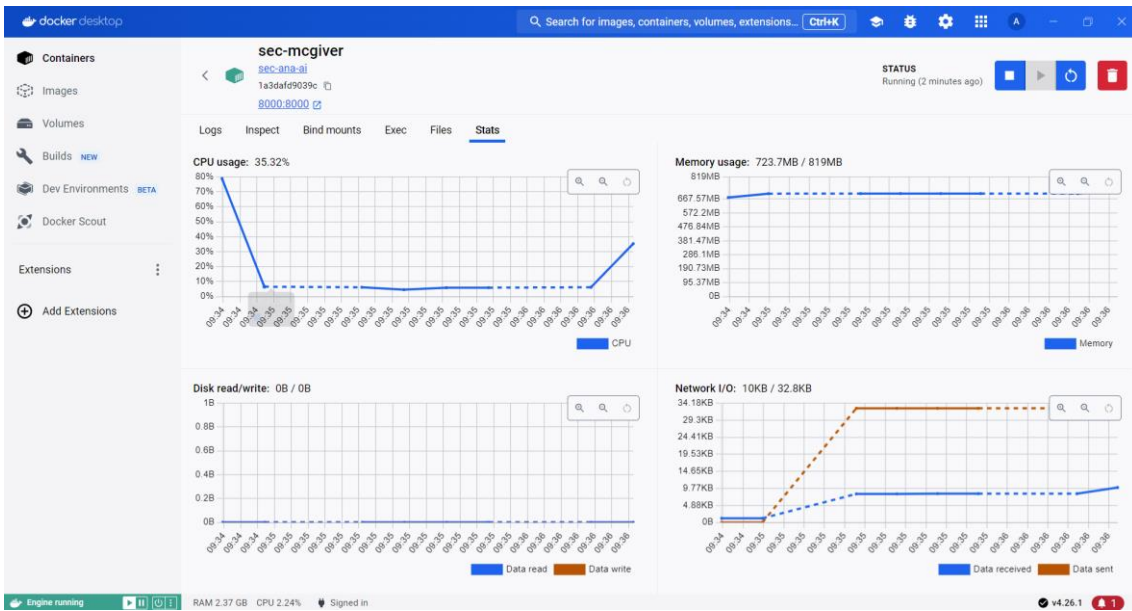


Figura 45 - Monitorització dels recursos consumits per el contenidor Docker.

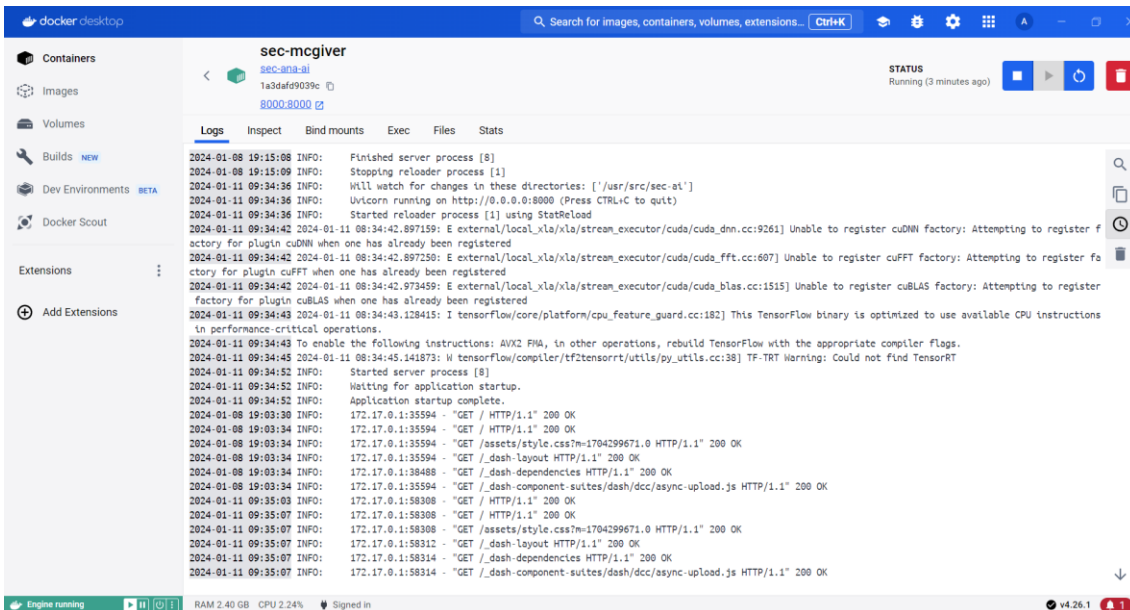


Figura 46 - Logs d'execució de l'aplicació en el contenidor Docker.

Explicacions addicionals de les comandes:

- *docker save*: Aquesta comanda crea un arxiu tar que conté tota la informació necessària de la imatge Docker, incloent les capes i etiquetes.
- *-o*: Especifica el nom del fitxer de sortida on es guardarà la imatge.
- *docker load*: Aquesta comanda carrega una imatge des d'un arxiu tar, permetent que sigui utilitzada en qualsevol entorn Docker.
- *-i*: Especifica el fitxer d'entrada des del qual es carregarà la imatge.

Nota important: la imatge *docker* ha estat definida per executar-se en un entorn de Linux, els contenidor que s'aixequen tenen aquest sistema per sota. Les proves que s'han realitzat han estat en entorns Windows, amb el què només s'ha pogut comprovar el correcte funcionament fent ús del subsistema Linux, en la seva versió 2 (WSL2), no es garanteix el correcte funcionament en altres entorns diferents als esmentats, tot i que l'aplicació s'ha desenvolupat per ser independent de la capa de sistema operatiu sobre la que s'executa.

7.3 Annex 3: Recursos de suport al treball

loc,tag
http://www.hhgenice.top/,"RedLine Stealer C2"
www.hhgenice.top,"RedLine Stealer C2"
hhgenice.top,"RedLine Stealer C2"
hhgenice,"RedLine Stealer C2"
185.215.113.45/g4MbvE/index.php,"RedLine Stealer C2"
185.215.113.45,"RedLine Stealer C2"
g4MbvE/index.php,"RedLine Stealer C2"
http://membro.at/upload/,"RedLine Stealer C2"
membro.at/upload/,"RedLine Stealer C2"
membro.at,"RedLine Stealer C2"
http://jeevanpunetha.com/upload/,"RedLine Stealer C2"
jeevanpunetha.com/upload/,"RedLine Stealer C2"
jeevanpunetha.com,"RedLine Stealer C2"
http://misipu.cn/upload/,"RedLine Stealer C2"
misipu.cn/upload/,"RedLine Stealer C2"
misipu.cn,"RedLine Stealer C2"
http://zavodooo.ru/upload/,"RedLine Stealer C2"
zavodooo.ru/upload/,"RedLine Stealer C2"
zavodooo.ru,"RedLine Stealer C2"
http://targiko.ru/upload/,"RedLine Stealer C2"
targiko.ru/upload/,"RedLine Stealer C2"
targiko.ru,"RedLine Stealer C2"
http://vues3d.com/upload/,"RedLine Stealer C2"
vues3d.com/upload/,"RedLine Stealer C2"

vues3d.com,"RedLine Stealer C2"
91.121.67.60:51630,"RedLine Stealer C2"
91.121.67.60,"RedLine Stealer C2"
51630,"RedLine Stealer C2"
185.215.113.38/f8dfksdj3/index.php,"RedLine Stealer C2"
185.215.113.38,"RedLine Stealer C2"
f8dfksdj3/index.php,"RedLine Stealer C2"
185.215.113.85:10018,"RedLine Stealer C2"
185.215.113.85,"RedLine Stealer C2"
10018,"RedLine Stealer C2"
http://ugll.org/test3/get.php,"RedLine Stealer C2"
ugll.org/test3/get.php,"RedLine Stealer C2"
ugll.org,"RedLine Stealer C2"
http://zerit.top/dl/build2.exe,"RedLine Stealer Payload"
http://zerit.top,"RedLine Stealer Payload_domain"
zerit.top,"RedLine Stealer Payload_domain"
http://ugll.org/files/1/build3.exe,"RedLine Stealer Payload"
http://ugll.org,"RedLine Stealer Payload_domain"
ugll.org,"RedLine Stealer Payload_domain"
185.215.113.24:15994,"RedLine Stealer C2"
185.215.113.24,"RedLine Stealer C2"
15994,"RedLine Stealer C2"

Taula 1 - Joc de dades per entrenar el model.

1	Cybersecurity	Threat Intelligence	The collected and analyzed information about current or potential attacks that threaten an organization
---	---------------	---------------------	---------------------------------------------------------------------------------------------------------

2	Cybersecurity	Intrusion Detection System (IDS)	A device or software application that monitors network or system activities for malicious activities or policy violations
3	Cybersecurity	Security Information and Event Management (SIEM)	A solution that aggregates and analyzes activity from many different resources across your IT infrastructure
4	Cybersecurity	Malware Analysis	The process of determining the functionality, origin, and potential impact of a given malware sample
5	Cybersecurity	Network Forensics	The process of capturing, recording, and analyzing network events to discover the source of security attacks or other problem incidents
6	Cybersecurity	Encryption	The process of converting data into a coded form to prevent unauthorized access
7	Cybersecurity	Firewall	A network security device that monitors and filters incoming and outgoing network traffic based on an organization's previously established security policies
8	Cybersecurity	Botnet	A network of private computers infected with malicious software and controlled as a group without the owners' knowledge, e.g., to send spam messages
9	Cybersecurity	Phishing	A form of fraud where an attacker masquerades as a reputable entity or person in email or other communication channels
10	Cybersecurity	Ransomware	Malware that encrypts the victim's data and demands payment in exchange for the decryption key
11	Cybersecurity	Zero-Day Exploit	A cyber attack that occurs on the same day a weakness is discovered in software, before the software developers have an opportunity to create a patch to fix the vulnerability
12	Cybersecurity	Denial of Service (DoS)	An interruption in an authorized user's access to a computer network, typically one caused with malicious intent
13	Cybersecurity	Distributed Denial of Service (DDoS)	A type of attack where multiple compromised systems are used to target a single system causing a Denial of Service (DoS) attack
14	Cybersecurity	Penetration Testing	An authorized simulated cyberattack on a computer system, performed to evaluate the security of the system
15	Cybersecurity	Vulnerability Assessment	The process of identifying, quantifying, and prioritizing (or ranking) the vulnerabilities in a system

16	Cybersecurity	Incident Response	The methodology an organization uses to respond to and manage a cyberattack
17	Cybersecurity	Rootkit	A set of software tools that enable an unauthorized user to gain control of a computer system without being detected
18	Cybersecurity	Spyware	Software that enables a user to obtain covert information about another's computer activities by transmitting data covertly from their hard drive
19	Cybersecurity	Adware	Any software application in which advertising is displayed while the program is running
20	Cybersecurity	Trojan Horse	Any malware which misleads users of its true intent
21	Cybersecurity	Worm	A standalone malware computer program that replicates itself in order to spread to other computers
22	Cybersecurity	Logic Bomb	A piece of code intentionally inserted into a software system that will set off a malicious function when specified conditions are met
23	Cybersecurity	Backdoor	A technique in which a system security mechanism is bypassed undetectably to access a computer or its data
24	Cybersecurity	Social Engineering	The use of deception to manipulate individuals into divulging confidential or personal information that may be used for fraudulent purposes
25	Cybersecurity	Packet Sniffing	The practice of monitoring and capturing all data packets passing through a given network
26	Cybersecurity	Man-in-the-Middle Attack	A cyberattack where the attacker secretly relays and possibly alters the communication between two parties who believe they are directly communicating with each other
27	Cybersecurity	Advanced Persistent Threat (APT)	A prolonged and targeted cyberattack in which an intruder gains access to a network and remains undetected for an extended period of time
28	Cybersecurity	Patch Management	The process of distributing and applying updates to software
29	Cybersecurity	Security Audit	A systematic evaluation of the security of a company's information system by measuring how well it conforms to a set of established criteria
30	Cybersecurity	Honeypot	A decoy system designed to lure cyberattackers away from legitimate targets
31	Cybersecurity	Network Packet	A formatted unit of data carried by a packet-switched network

32	Cybersecurity	IP Address	A numerical label assigned to each device connected to a computer network that uses the Internet Protocol for communication
33	Cybersecurity	Port Scanning	The act of systematically scanning a computer's ports to detect listening ports and infer the presence of vulnerable services
34	Cybersecurity	Protocol Analysis	The process of examining the protocols used over a network in real-time, or through examining a capture of network traffic
35	Cybersecurity	Signature-Based Detection	The use of predefined signatures of known cyber threats to identify malicious activity
36	Cybersecurity	Anomaly-Based Detection	The process of comparing definitions of what is considered normal behavior with observed events to identify significant deviations
37	Cybersecurity	Behavioral Analysis	The monitoring of network traffic to identify patterns that may indicate a security breach
38	Cybersecurity	Heuristic Analysis	A technique designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution
39	Cybersecurity	SSL/TLS Inspection	The interception of encrypted SSL/TLS traffic to monitor and filter out malicious content
40	Cybersecurity	Deep Packet Inspection (DPI)	A form of computer network packet filtering that examines the data part (and possibly also the header) of a packet as it passes an inspection point
41	Cybersecurity	Network Layer	The third layer in the OSI model of computer networking responsible for packet forwarding including routing through intermediate routers
42	Cybersecurity	Application Layer	The top layer in the OSI model, which provides methods for software applications to communicate over a network
43	Cybersecurity	Transport Layer Security (TLS)	A cryptographic protocol designed to provide communications security over a computer network
44	Cybersecurity	Secure Sockets Layer (SSL)	The standard security technology for establishing an encrypted link between a web server and a browser
45	Cybersecurity	Virtual Private Network (VPN)	A service that allows you to connect to the Internet via an encrypted tunnel to ensure your online privacy and protect your sensitive data
46	Cybersecurity	Domain Name	The hierarchical and decentralized naming system used to

		System (DNS)	identify computers reachable through the Internet
47	Cybersecurity	DNS Security Extensions (DNSSEC)	A suite of Internet Engineering Task Force (IETF) specifications for securing certain kinds of information provided by the Domain Name System (DNS) as used on Internet Protocol (IP) networks
48	Cybersecurity	Network Address Translation (NAT)	A method of remapping one IP address space into another by modifying network address information in the IP header of packets while they are in transit across a traffic routing device
49	Cybersecurity	Subnet	A subdivision of an IP network that can be independently routed in a larger network
50	Cybersecurity	Access Control List (ACL)	A list of permissions attached to an object that specifies which users or system processes are granted access to objects, as well as what operations are allowed on given objects
51	Cybersecurity	Network Topology	The arrangement of the elements (links, nodes, etc.) of a communication network
52	Cybersecurity	Intrusion Prevention System (IPS)	A network security/threat prevention technology that examines network traffic flows to detect and prevent vulnerability exploits
53	Cybersecurity	Network Segmentation	The practice of splitting a computer network into subnetworks, each being a network segment or network layer to enhance performance and security
54	Cybersecurity	Endpoint Detection and Response (EDR)	A cybersecurity technology that addresses the need for continuous monitoring and response to advanced threats
55	Cybersecurity	Network Access Control (NAC)	A security approach that unifies endpoint security technology, user or system authentication and network security enforcement
56	Cybersecurity	Session Hijacking	The exploitation of a valid computer session to gain unauthorized access to information or services in a computer system
57	Cybersecurity	Secure Shell (SSH)	A cryptographic network protocol for operating network services securely over an unsecured network
58	Cybersecurity	Hypertext Transfer Protocol Secure (HTTPS)	An extension of the Hypertext Transfer Protocol (HTTP) for secure communication over a computer network

59	Cybersecurity	Simple Network Management Protocol (SNMP)	An Internet Standard protocol for collecting and organizing information about managed devices on IP networks and for modifying that information to change device behavior
60	Cybersecurity	Network Probe	A test or action taken to investigate the state or health of a computer network
61	MITRE ATT&CK	Tactic	The highest level of the framework's hierarchy that describes the "why" of an ATT&CK technique or the adversary's goal
62	MITRE ATT&CK	Technique	The "how" in the framework, describing the behavior an adversary uses to achieve a tactical goal
63	MITRE ATT&CK	Sub-technique	A more specific description of the technique detailing the adversary's actions
64	MITRE ATT&CK	Lateral Movement	A tactic where an adversary moves through an environment from one system to another to gain control of additional systems within a network
65	MITRE ATT&CK	Command and Control (C2)	The means by which an adversary communicates with systems under their control within a victim network
66	MITRE ATT&CK	Exploit Public-Facing Application	The use of vulnerabilities in internet-facing computer programs to gain initial access to a network
67	MITRE ATT&CK	Drive-by Compromise	The compromise of a system through a user visiting a website over the normal course of browsing
68	MITRE ATT&CK	Spearphishing Attachment	Sending targeted emails with an attachment to gain access to a network
69	MITRE ATT&CK	Spearphishing Link	Sending targeted emails with a link to a malicious website to gain access to a network
70	MITRE ATT&CK	Valid Accounts	The use of legitimate credentials to gain access and move laterally in a network
71	MITRE ATT&CK	Remote Services	The use of remote services such as VPNs, Citrix, and other access mechanisms to gain entry to a network
72	MITRE ATT&CK	Replication Through Removable Media	The introduction of malware through removable media to move laterally through a network
73	MITRE ATT&CK	External Remote Services	Accessing external remote services as a means of maintaining access to a network

74	MITRE ATT&CK	Internal Spearphishing	The act of sending emails from a compromised account to internal users to gain access to additional systems
75	MITRE ATT&CK	Taint Shared Content	The manipulation of shared content to spread an attack laterally through a network
76	MITRE ATT&CK	Third-party Software	Exploiting third-party software to compromise a network and move laterally
77	MITRE ATT&CK	Multi-hop Proxy	Using multiple proxies to help obfuscate their C2 communications
78	MITRE ATT&CK	Multilayer Encryption	Encrypting C2 communications multiple times to avoid detection
79	MITRE ATT&CK	Custom Cryptographic Protocol	Using custom encryption schemes for C2 communications to avoid standard decryption techniques
80	MITRE ATT&CK	Domain Fronting	Using different domain names in a single HTTPS connection to discreetly direct C2 communications to a hidden server
81	MITRE ATT&CK	Data Encoding	Encoding data in a way that is difficult for defenders to detect or decode to hide C2 communications
82	MITRE ATT&CK	Data Obfuscation	Obfuscating data to make C2 communications less obvious and harder to detect
83	MITRE ATT&CK	Fallback Channels	Establishing additional C2 channels that can be used if the primary channels are discovered and cut off
84	MITRE ATT&CK	Standard Cryptographic Protocol	Using standard encryption protocols to disguise C2 traffic as normal encrypted web traffic
85	MITRE ATT&CK	Standard Application Layer Protocol	Using standard application layer protocols to blend in C2 communications with normal traffic
86	MITRE ATT&CK	Uncommonly Used Port	Using non-standard ports for C2 communications to bypass security devices that monitor commonly used ports
87	MITRE ATT&CK	Web Service	Using a web service like Twitter or GitHub as a means of controlling compromised systems within a network
88	MITRE ATT&CK	Application Access Token	Using tokens from applications (like OAuth tokens) to access resources and move laterally in a network
89	MITRE ATT&CK	Pass the Ticket	Using Kerberos tickets to authenticate to systems and move laterally in a network
90	MITRE	Remote File	Copying files from one system to another to stage tools or

	ATT&CK	Copy	data for exfiltration
--	--------	------	-----------------------

Taula 2 - Joc de dades per entrenar el model en termes de ciberseguretat i d'anàlisi de xarxes.

Input	Target
<p>Create a Root Cause Analysis of: A machine with IP address 192.168.1.10 accessed various suspicious domains and downloaded a zip file which extracted to a malicious .js file. This script initiated a download of an installer DLL for the IcedID malware</p>	<p>Root Cause Analysis (RCA) Report: Event Description: A machine with IP address 192.168.1.10 accessed various suspicious domains and downloaded a zip file which extracted to a malicious .js file. This script initiated a download of an installer DLL for the IcedID malware. Objective 1: Discovering the Cause Root Cause: The initial compromise occurred via a phishing email containing a link to a zip file. The extracted JavaScript file ran a command script which downloaded and installed the IcedID malware. Objective 2: Understanding and Addressing the Issue Impact: The machine is infected with IcedID, a banking trojan that can lead to financial fraud, data exfiltration, and potential access for further malicious activities within the network. Recommendations: Isolate the infected machine from the network to prevent further spread of the malware. Conduct a forensic analysis of the infected machine to determine the initial entry point and any lateral movements or data exfiltration. Review security policies and training to prevent future phishing attacks. Update antivirus and antimalware programs across the network to detect and prevent IcedID infections. Objective 3: Preventing Future Incidents Lessons Learned: Users can be deceived into downloading malware through seemingly innocuous email links. Robust endpoint protection and user training are essential to prevent malware infections. Preventative Measures: Implement an email filtering solution that scans and blocks suspicious attachments and links. Conduct regular security awareness training to help users recognize and report phishing attempts. Keep systems and applications updated with the latest security patches. Employ a network monitoring solution to detect and alert on unusual network traffic patterns. Conclusion: Understanding the root cause of the IcedID infection and implementing strong preventative measures can significantly reduce the risk of future incidents and strengthen the organization's cybersecurity</p>

	defenses
<p>Make a root cause analysis of the following events: A machine with IP address 192.168.1.10 accessed various suspicious domains and downloaded a zip file which extracted to a malicious .js file. This script initiated a download of an installer DLL for the IcedID malware</p>	<p>Root Cause Analysis (RCA) Report: Event Description: A machine with IP address 192.168.1.10 accessed various suspicious domains and downloaded a zip file which extracted to a malicious .js file. This script initiated a download of an installer DLL for the IcedID malware. Objective 1: Discovering the Cause Root Cause: The initial compromise occurred via a phishing email containing a link to a zip file. The extracted JavaScript file ran a command script which downloaded and installed the IcedID malware. Objective 2: Understanding and Addressing the Issue Impact: The machine is infected with IcedID, a banking trojan that can lead to financial fraud, data exfiltration, and potential access for further malicious activities within the network. Recommendations: Isolate the infected machine from the network to prevent further spread of the malware. Conduct a forensic analysis of the infected machine to determine the initial entry point and any lateral movements or data exfiltration. Review security policies and training to prevent future phishing attacks. Update antivirus and antimalware programs across the network to detect and prevent IcedID infections. Objective 3: Preventing Future Incidents Lessons Learned: Users can be deceived into downloading malware through seemingly innocuous email links. Robust endpoint protection and user training are essential to prevent malware infections. Preventative Measures: Implement an email filtering solution that scans and blocks suspicious attachments and links. Conduct regular security awareness training to help users recognize and report phishing attempts. Keep systems and applications updated with the latest security patches. Employ a network monitoring solution to detect and alert on unusual network traffic patterns. Conclusion: Understanding the root cause of the IcedID infection and implementing strong preventative measures can significantly reduce the risk of future incidents and strengthen the organization's cybersecurity defenses</p>

Taula 3 - Joc de dades per entrenar a fer Anàlisi de Causa Arrel.