SECRYPT conference in Porto (Portugal), July 2026.

Fine-tuning LLMs for Operational Phishing Email Detection

Armand Florent Tsafack Piugie[1,2], Mathieu Valois[1], Emmanuel Giguet[2], Christophe Rosenberger[2] and Philippe Chauvat[1]

[1]téïcée company, 14760 Bretteville-sur-Odon, France

[2]Université de Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France

{ftsafack,mvalois,pchauvat}@teicee.com,
emmanuel.giguet@cnrs.fr, christophe.rosenberger@ensicaen.fr

téïcée

GREYC
Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

Outline

1. Introduction
2. State of the art
3. Methodology
4. Experimental protocol
5. Results
6. Conclusion

1. Introduction : context

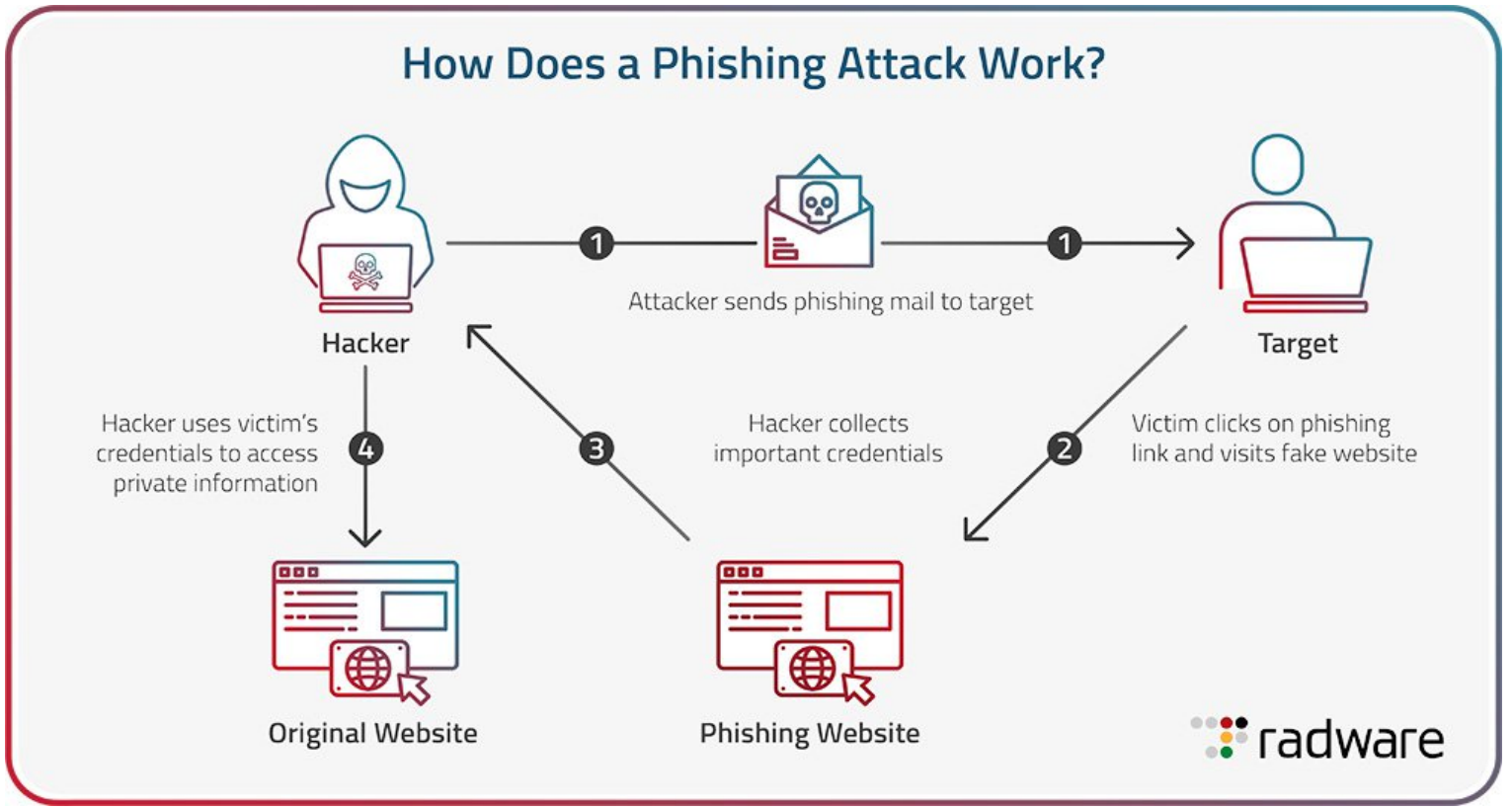
Phishing : a threat for organizations

- **~91% of cyber attacks** *[Baker and Cartier, 2025]* .
- **~50% of organizations** *[Barracuda Networks,2023]* .
- **~4,65M USD** *[Baker and Cartier, 2025]* .

1. Introduction : context

Phishing : a threat for organizations

- **~91% of cyber attacks** *[Baker and Cartier, 2025]* .
- **~50% of organizations** *[Barracuda Networks,2023]* .
- **~4,65M USD** *[Baker and Cartier, 2025]* .



Phishing attack steps - https://www.radware.com/cyberpedia/bot-attacks/phishing-attack/

Definition

A social engineering technique used to deceive users into revealing sensitive information

Social engineering techniques :

- Fake urgency
- Invitation to follow a suspicious link
- Brand impersonation.

1. Introduction : problem and goals

téicée®
la raison de votre confiance

GREYC
Electronics and Computer Science Laboratory

Problem :
Attackers are refining their techniques to increase their profitability (by also using AI).

Main goal :
Design a protection system with automatic detection and human moderation.

Specifically :
- Using AI and NLP techniques for phishing email detection,
- Collaborative approach : Combine automated analysis results and user reports, with IT/CISCO moderators validating the email's status.

2. State of the art

Table 1: Literature review of phishing email detection methods using deep learning and large language model.

Ref.	Method	Dataset	Parts	Features	Metrics
(Altwaijry et al., 2024)	1D-CNN with LSTM, Bi-LSTM, GRU, Bi-GRU	Phishing Corpus + Spam Assassin	Body and Subject	GloVe	Accuracy: 98.9%-99.7%, F1-score: 99.2%-99.6%
(Atawneh and Aljehani, 2023)	CNN, RNN, LSTM, BERT	Enron, SpamAssassin, UCI	Body	Tokenization (Keras and WordPiece)	Accuracy: 98.6%-99.6%
(Fang et al., 2019)	Enhanced RCNN with attention mechanism	IWSPA-AP 2018	Subject and Body	Word2Vec	Accuracy: 99.8%, False Positives: 0.043%
(Ra et al., 2018)	CNN, RNN, LSTM, MLP	IWSPA-AP 2018	Body	N-grams and NLP	Accuracy: 96.2%-99.1%
(Alhogail and Alsabih, 2021)	GCN and NLP	Fraud Dataset	Body	NLP and Graph Features	Accuracy: 98.2%, False Positives: 0.015
(Doshi et al., 2023)	ANN, RNN, CNN	Phishing Corpus + SpamAssassin	Body	TF-IDF	Accuracy: 99.5%, F1-score: 99.5%
(Dewis and Viana, 2022)	LSTM	Phishing Email Collection + UCI Spambase	Body	TF-IDF and NLP	Accuracy: 99.0%
(Muralidharan and Nissim, 2023)	CNN and BERT (ensemble)	Virus Total	Body, header, attachments	Contextual Embeddings	AUC: 99.3%, TPR: 94.7%
(Gogoi and Ahmed, 2022)	BERT, DistilBERT	-	-	Transformer embeddings	Accuracy: 99.0%
(Kuikel et al., 2025)	BERT, Llama 7B, Llama 8B and Wizard7B	Enron and Nazario	Sender, Subject and Body	Transformer embeddings	Accuracy: 98.5%, 90.9%, 93.3%, 83.7%
(Koide et al., 2024)	ChatSpamDetector(using GPT-4, GPT-3.5-Turbo, Llama2-70B and Gemini Pro)	Enron, SpamAssassin, TREC, CSDMC SPAM corpus	Body	Transformer embeddings	Accuracy, precision, Recall (99.7%)
(Dalmiere et al., 2025a)	GPT-4o-mini with In-Context Learning (few-shot)	Real-world French phishing emails (SignalSpam)	Email body (semantic content)	LLM in-context reasoning	Accuracy: 76%
(Uddin and Sarker, 2024)	RoBERTa	Kaggle phishing Email Dataset	Body	Trasnformer embeddings	Accuracy: 97.9%

2. State of the art

Table 1: Literature review of phishing email detection methods using deep learning and large language model.

Ref.	Method	Dataset	Parts	Features	Metrics
(Altwaijry et al., 2024)	1D-CNN with LSTM, Bi-LSTM, GRU, Bi-GRU	Phishing Corpus + Spam Assassin	Body and Subject	GloVe	Accuracy: 98.9%-99.7%, F1-score: 99.2%-99.6%
(Atawneh and Aljehani, 2023)	CNN, RNN, LSTM, BERT	Enron, SpamAssassin, UCI	Body	Tokenization (Keras and WordPiece)	Accuracy: 98.6%-99.6%
(Fang et al., 2019)	Enhanced RCNN with attention mechanism	IWSPA-AP 2018	Subject and Body	Word2Vec	Accuracy: 99.8%, False Positives: 0.043%
(Ra et al., 2018)	CNN, RNN, LSTM, MLP	IWSPA-AP 2018	Body	N-grams and NLP	Accuracy: 96.2%-99.1%
(Alhogail and Alsabih, 2021)	GCN and NLP	Fraud Dataset	Body	NLP and Graph Features	Accuracy: 98.2%, False Positives: 0.015
(Doshi et al., 2023)	ANN, RNN, CNN	Phishing Corpus + SpamAssassin	Body	TF-IDF	Accuracy: 99.5%, F1-score: 99.5%
(Dewis and Viana, 2022)	LSTM	Phishing Email Collection + UCI Spambase	Body	TF-IDF and NLP	Accuracy: 99.0%
(Muralidharan and Nissim, 2023)	CNN and BERT (ensemble)	Virus Total	Body, header, attachments	Contextual Embeddings	AUC: 99.3%, TPR: 94.7%
(Gogoi and Ahmed, 2022)	BERT, DistilBERT	-	-	Transformer embeddings	Accuracy: 99.0%
(Kuikel et al., 2025)	BERT, Llama 7B, Llama 8B and Wizard7B	Enron and Nazario	Sender, Subject and Body	Transformer embeddings	Accuracy: 98.5%, 90.9%, 93.3%, 83.7%
(Koide et al., 2024)	ChatSpamDetector(using GPT-4, GPT-3.5-Turbo, Llama2-70B and Gemini Pro)	Enron, SpamAssassin, TREC, CSDMC SPAM corpus	Body	Transformer embeddings	Accuracy, precision, Recall (99.7%)
(Dalmiere et al., 2025a)	GPT-4o-mini with In-Context Learning (few-shot)	Real-world French phishing emails (SignalSpam)	Email body (semantic content)	LLM in-context reasoning	Accuracy: 76%
(Uddin and Sarker, 2024)	RoBERTa	Kaggle phishing Email Dataset	Body	Trasnformer embeddings	Accuracy: 97.9%

Research works

- A wide range of solution proposals based on DL and LLM models.

2. State of the art

Table 1: Literature review of phishing email detection methods using deep learning and large language model.

Ref.	Method	Dataset	Parts	Features	Metrics
(Altwaijry et al., 2024)	1D-CNN with LSTM, Bi-LSTM, GRU, Bi-GRU	Phishing Corpus + Spam Assassin	Body and Subject	GloVe	Accuracy: 98.9%-99.7%, F1-score: 99.2%-99.6%
(Atawneh and Aljehani, 2023)	CNN, RNN, LSTM, BERT	Enron, SpamAssassin, UCI	Body	Tokenization (Keras and WordPiece)	Accuracy: 98.6%-99.6%
(Fang et al., 2019)	Enhanced RCNN with attention mechanism	IWSPA-AP 2018	Subject and Body	Word2Vec	Accuracy: 99.8%, False Positives: 0.043%
(Ra et al., 2018)	CNN, RNN, LSTM, MLP	IWSPA-AP 2018	Body	N-grams and NLP	Accuracy: 96.2%-99.1%
(Alhogail and Alsabih, 2021)	GCN and NLP	Fraud Dataset	Body	NLP and Graph Features	Accuracy: 98.2%, False Positives: 0.015
(Doshi et al., 2023)	ANN, RNN, CNN	Phishing Corpus + SpamAssassin	Body	TF-IDF	Accuracy: 99.5%, F1-score: 99.5%
(Dewis and Viana, 2022)	LSTM	Phishing Email Collection + UCI Spambase	Body	TF-IDF and NLP	Accuracy: 99.0%
(Muralidharan and Nissim, 2023)	CNN and BERT (ensemble)	Virus Total	Body, header, attachments	Contextual Embeddings	AUC: 99.3%, TPR: 94.7%
(Gogoi and Ahmed, 2022)	BERT, DistilBERT	-	-	Transformer embeddings	Accuracy: 99.0%
(Kuikel et al., 2025)	BERT, Llama 7B, Llama 8B and Wizard7B	Enron and Nazario	Sender, Subject and Body	Transformer embeddings	Accuracy: 98.5%, 90.9%, 93.3%, 83.7%
(Koide et al., 2024)	ChatSpamDetector(using GPT-4, GPT-3.5-Turbo, Llama2-70B and Gemini Pro)	Enron, SpamAssassin, TREC, CSDMC SPAM corpus	Body	Transformer embeddings	Accuracy, precision, Recall (99.7%)
(Dalmiere et al., 2025a)	GPT-4o-mini with In-Context Learning (few-shot)	Real-world French phishing emails (SignalSpam)	Email body (semantic content)	LLM in-context reasoning	Accuracy: 76%
(Uddin and Sarker, 2024)	RoBERTa	Kaggle phishing Email Dataset	Body	Trasnformer embeddings	Accuracy: 97.9%

Research works

- A wide range of solution proposals based on DL and LLM models.
- Content analysis performed on the subject and body of the email message

2. State of the art

Table 1: Literature review of phishing email detection methods using deep learning and large language model.

Ref.	Method	Dataset	Parts	Features	Metrics
(Altwaijry et al., 2024)	1D-CNN with LSTM, Bi-LSTM, GRU, Bi-GRU	Phishing Corpus + Spam Assassin	Body and Subject	GloVe	Accuracy: 98.9%-99.7%, F1-score: 99.2%-99.6%
(Atawneh and Aljehani, 2023)	CNN, RNN, LSTM, BERT	Enron, SpamAssassin, UCI	Body	Tokenization (Keras and WordPiece)	Accuracy: 98.6%-99.6%
(Fang et al., 2019)	Enhanced RCNN with attention mechanism	IWSPA-AP 2018	Subject and Body	Word2Vec	Accuracy: 99.8%, False Positives: 0.043%
(Ra et al., 2018)	CNN, RNN, LSTM, MLP	IWSPA-AP 2018	Body	N-grams and NLP	Accuracy: 96.2%-99.1%
(Alhogail and Alsabih, 2021)	GCN and NLP	Fraud Dataset	Body	NLP and Graph Features	Accuracy: 98.2%, False Positives: 0.015
(Doshi et al., 2023)	ANN, RNN, CNN	Phishing Corpus + SpamAssassin	Body	TF-IDF	Accuracy: 99.5%, F1-score: 99.5%
(Dewis and Viana, 2022)	LSTM	Phishing Email Collection + UCI Spambase	Body	TF-IDF and NLP	Accuracy: 99.0%
(Muralidharan and Nissim, 2023)	CNN and BERT (ensemble)	Virus Total	Body, header, attachments	Contextual Embeddings	AUC: 99.3%, TPR: 94.7%
(Gogoi and Ahmed, 2022)	BERT, DistilBERT	-	-	Transformer embeddings	Accuracy: 99.0%
(Kuikel et al., 2025)	BERT, Llama 7B, Llama 8B and Wizard7B	Enron and Nazario	Sender, Subject and Body	Transformer embeddings	Accuracy: 98.5%, 90.9%, 93.3%, 83.7%
(Koide et al., 2024)	ChatSpamDetector(using GPT-4, GPT-3.5-Turbo, Llama2-70B and Gemini Pro)	Enron, SpamAssassin, TREC, CSDMC SPAM corpus	Body	Transformer embeddings	Accuracy, precision, Recall (99.7%)
(Dalmiere et al., 2025a)	GPT-4o-mini with In-Context Learning (few-shot)	Real-world French phishing emails (SignalSpam)	Email body (semantic content)	LLM in-context reasoning	Accuracy: 76%
(Uddin and Sarker, 2024)	RoBERTa	Kaggle phishing Email Dataset	Body	Trasnformer embeddings	Accuracy: 97.9%

Research works

- A wide range of solution proposals based on DL and LLM models.
- Content analysis performed on the body of the email message
- Very good performance on small unrerepresentative datasets

3. Methodology

Contributions

- An operational pipeline for phishing detection **from raw emails** (parsing, preprocessing, classification, prediction and explanation) ;
- A comparison of three pretrained LLMs under the same protocol, datasets, and metrics, including published LLM-based baselines ;
- A trade-off analysis between accuracy, training cost, and inference throughput on corpora up to 228K messages and 40K raw phishing emails ;
- SHAP-based explanations to support analyst validation and user awareness ;
- Deployment oriented recommendations on model.

3. Methodology

Operational LLM Pipeline

Pipeline :

- Extraction
- Preprocessing
- Training
- Prediction
- Explanation

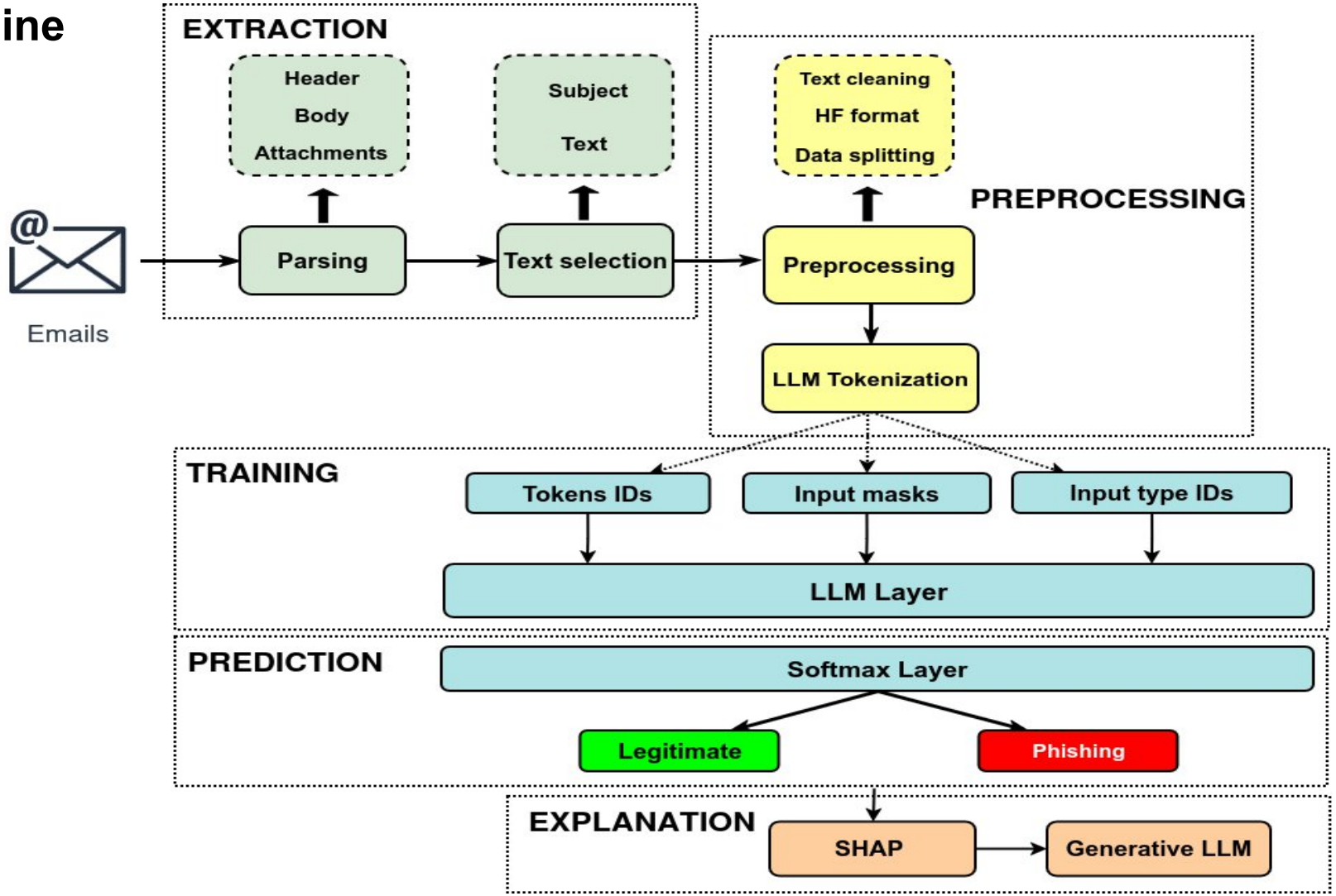


Figure 1 : Operational framework for fine-tuned LLM phishing email detection

3. Methodology

Characteristics of pre-trained LLMs for phishing detection

Table 2 : Summary of characteristics of models fine-tuned for phishing detection

Characteristics	CamemBERT-base	XLM-RoBERTa	Mistral-7B
Language	French	Multilingual	Multilingual
Pre-training Corpus	OSCAR (138 GB)	Common Crawl (2.5 TB)	-
Tokenization	SentencePiece	SentencePiece	Byte-level BPE
Max Input Sequence Length	512	1,024	8,192
Embedding Dimension	768	768	4,096
Attention Heads	12	12	32
Parameters	~110 million	~270 million	~7 billion
Size on Disk	422 MB	1.1 GB	28 GB
Vocabulary Size	32k	250k	32k
Architecture	Encoder	Encoder	Decoder

Remarks

- CamemBERT-base
- XLM-RoBERTa
- Mistral-7B

4. Experimental Protocol

DATABASES

Table 3 : Summary of email datasets characteristics. Language encoding (EN is english and FR is French)

Characteristic	Chiper Ionescu	Signal Spam	téïcée
Source	Public sources	Signal Spam	téïcée
Format	CSV (text)	Raw files	Raw files
Total emails	227,973	41,896	803
Main languages	EN	EN/FR	EN/FR
Phishing emails	108,032	41,896	377
Legitimate emails	119,941	0	426
Training split	162,938 (70%)	Not used	Not used
Validation split	34,915 (15%)	-	-
Test split	34,916 (15%)	41,896 (100%)	803 (100%)

4. Experimental Protocol

DATABASES

Table 3 : Summary of email datasets characteristics. Language encoding (EN is english and FR is French)

Characteristic	Chiper Ionescu	Signal Spam	téïcée
Source	Public sources	Signal Spam	téïcée
Format	CSV (text)	Raw files	Raw files
Total emails	227,973	41,896	803
Main languages	EN	EN/FR	EN/FR
Phishing emails	108,032	41,896	377
Legitimate emails	119,941	0	426
Training split	162,938 (70%)	Not used	Not used
Validation split	34,915 (15%)	-	-
Test split	34,916 (15%)	41,896 (100%)	803 (100%)

Dataset	Phishing Emails	Clean Emails	Total Emails
CEAS 2008 Live Spam Challenge	21,842	17,312	39,154
Enron Email Dataset	13,976	15,791	29,767
Ling-Spam Dataset	458	2,401	2,859
Nazario Phishing Email Dataset	1,565	0	1,565
SpamAssassin Email Dataset	1,718	4,091	5,809
TREC 2005 Public Spam Corpus	22,932	32,278	54,210
TREC 2006 Public Spam Corpus	3,989	12,393	16,382
TREC 2007 Public Spam Corpus	29,392	24,353	53,745
Nigerian 5 Phishing Email Dataset	1,500	0	1,500
Nigerian Fraud Email Dataset	3,332	0	3,332
Phishing Email Dataset	7,328	11,322	18,650
Total	108,032	119,941	227,973

4. Experimental Protocol

For LLMs Fine-tuning

- Dataset : Chiper Ionescu (Enron, SpamAssassin, TREC, ...)
- Parts of emails used : subject + body
- train/val/test split : 70% (162938) / 15%(34915) / 15% (34916)
- LLMs : CamemBERT, XLM-RoBERTa, Mistral7B
- Features : tokens embeddings
- Hyperparameter configurations (with QLoRA for Mistral7B)
- Training hyperparameter configurations
- Evaluation metrics: accuracy, precision, recall et f1
- Generalization capability : Signal Spam, téïcée company

Ressources :

- Research lab computing ressource : processors (2 Intel Gold 5218R 2.1GHz processors), core (80), RAM (384GB), GPU (NVIDIA Quadro RTX 6000 24GB RAM)
- Laptop specifications : 32GB RAM, 16 CPUs, AMD Ryzen 7 2.8GHz avec pour système d'exploitation Linux Debian 12

5. Results

téïcée® la raison de votre confiance

GREYC Electronics and Computer Science Laboratory

Performance evaluation and generalization capability

Table 4: Performance for different evaluation scenarios for LLMs (CamemBERT-base, XLM-RoBERTa, and Mistral 7B) when trained with Chiper Ionescu dataset. Training (min) is total wall-clock fine-tuning time (10 epochs for encoders, ≈45 min/epoch; 3 epochs for Mistral with QLoRA, ≈40 h/epoch). Throughput (emails/s) averages end-to-end rates including parsing and preprocessing.

LLM	Training (min)	Test data	Accuracy (%)	Precision (%)	Recall (%)	emails/s (CPU)	emails/s (GPU)
CamemBERT-base	448	Chiper Ionescu Signal Spam téïcée	98.64 89.54 100.00	98.78 – 100.00	98.42 89.54 100.00	7	438
XLM-RoBERTa	449	Chiper Ionescu Signal Spam téïcée	99.46 95.76 99.13	99.25 – 99.73	99.68 95.76 98.41	7	308
Mistral 7B	7260	Chiper Ionescu Signal Spam téïcée	99.57 98.55 95.02	99.41 – 91.40	99.69 98.55 98.67	0.06	3

5. Results

téïcée® la raison de votre confiance

GREYC Electronics and Computer Science Laboratory

Performance evaluation and generalization capability

Table 4: Performance for different evaluation scenarios for LLMs (CamemBERT-base, XLM-RoBERTa, and Mistral 7B) when trained with Chiper Ionescu dataset. Training (min) is total wall-clock fine-tuning time (10 epochs for encoders, ≈45 min/epoch; 3 epochs for Mistral with QLoRA, ≈40 h/epoch). Throughput (emails/s) averages end-to-end rates including parsing and preprocessing.

LLM	Training (min)	Test data	Accuracy (%)	Precision (%)	Recall (%)	emails/s (CPU)	emails/s (GPU)
CamemBERT-base	448	Chiper Ionescu	98.64	98.78	98.42	7	438
		Signal Spam	89.54	–	89.54		
		téïcée	100.00	100.00	100.00		
XLM-RoBERTa	449	Chiper Ionescu	99.46	99.25	99.68	7	308
		Signal Spam	95.76	–	95.76		
		téïcée	99.13	99.73	98.41		
Mistral 7B	7260	Chiper Ionescu	99.57	99.41	99.69	0.06	3
		Signal Spam	98.55	–	98.55		
		téïcée	95.02	91.40	98.67		

- Strong in-distribution : all models achieve high accuracy when fine-tuned with Chiper Ionescu

5. Results

téïcée® la raison de votre confiance

GREYC Electronics and Computer Science Laboratory

Performance evaluation and generalization capability

Table 4: Performance for different evaluation scenarios for LLMs (CamemBERT-base, XLM-RoBERTa, and Mistral 7B) when trained with Chiper Ionescu dataset. Training (min) is total wall-clock fine-tuning time (10 epochs for encoders, ≈45 min/epoch; 3 epochs for Mistral with QLoRA, ≈40 h/epoch). Throughput (emails/s) averages end-to-end rates including parsing and preprocessing.

LLM	Training (min)	Test data	Accuracy (%)	Precision (%)	Recall (%)	emails/s (CPU)	emails/s (GPU)
CamemBERT-base	448	Chiper Ionescu	98.64	98.78	98.42	7	438
		Signal Spam	89.54	–	89.54		
		téïcée	100.00	100.00	100.00		
XLM-RoBERTa	449	Chiper Ionescu	99.46	99.25	99.68	7	308
		Signal Spam	95.76	–	95.76		
		téïcée	99.13	99.73	98.41		
Mistral 7B	7260	Chiper Ionescu	99.57	99.41	99.69	0.06	3
		Signal Spam	98.55	–	98.55		
		téïcée	95.02	91.40	98.67		

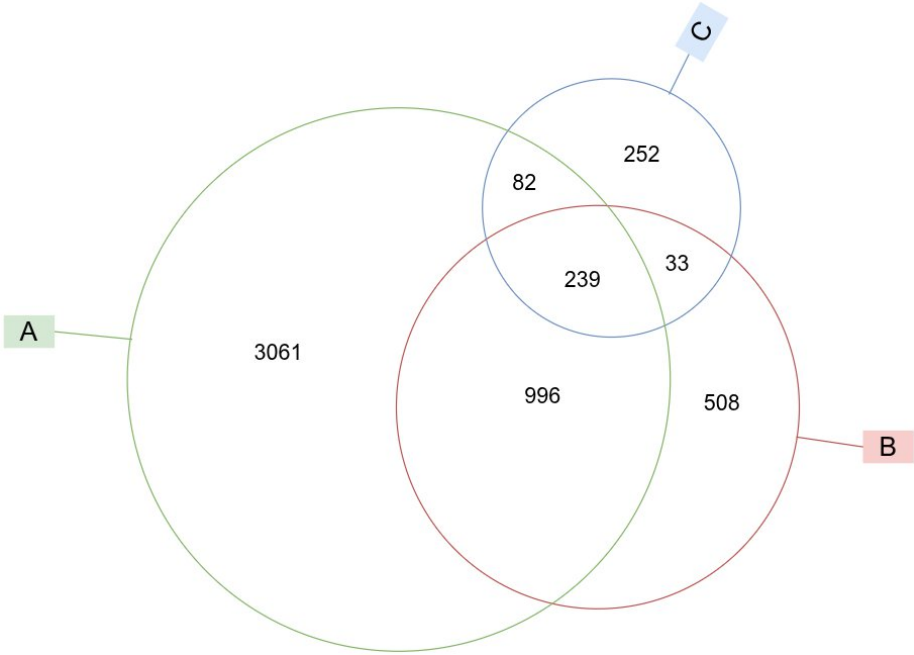
- Strong in-distribution : all models achieve high accuracy when fine-tuned with Chiper Ionescu
- Good generalization capability
- XLM-RoBERTa offers the best balance of accuracy, cross-dataset robustness, and deployment cost

5. Results

Errors analysis on signal spam

Table 5: Qualitative analysis of errors made by the fine-tuned LLMs. FN stands for False Negatives.

Set	Cardinality	Interpretation
A	4378	Set of FN for the CamemBERT model
B	1776	Set of FN for the XLM-RoBERTa model
C	606	Set of FN for the Mistral7B model
$A \cap B$	1235	Intersection of FN for CamemBERT and XLM-RoBERTa
$A \cap C$	321	Intersection of FN for CamemBERT and Mistral7B
$B \cap C$	272	Intersection of FN for XLM-RoBERTa and Mistral7B
$A \cap B \cap C$	239	Intersection of FN for CamemBERT, XLM-RoBERTa and Mistral7B
$A \cup B$	4919	Union of FN for CamemBERT and XLM-RoBERTa
$A \cup C$	4663	Union of FN for CamemBERT and Mistral7B
$B \cup C$	2110	Union of FN for XLM-RoBERTa and Mistral7B
$A \cup B \cup C$	5171	Union of FN for CamemBERT, XLM-RoBERTa and Mistral7B



This analysis show that, ensemble methods combining Mistral7B with either smaller model could potentially reduce the overall FNR

5. Results

Explaining phishing detection

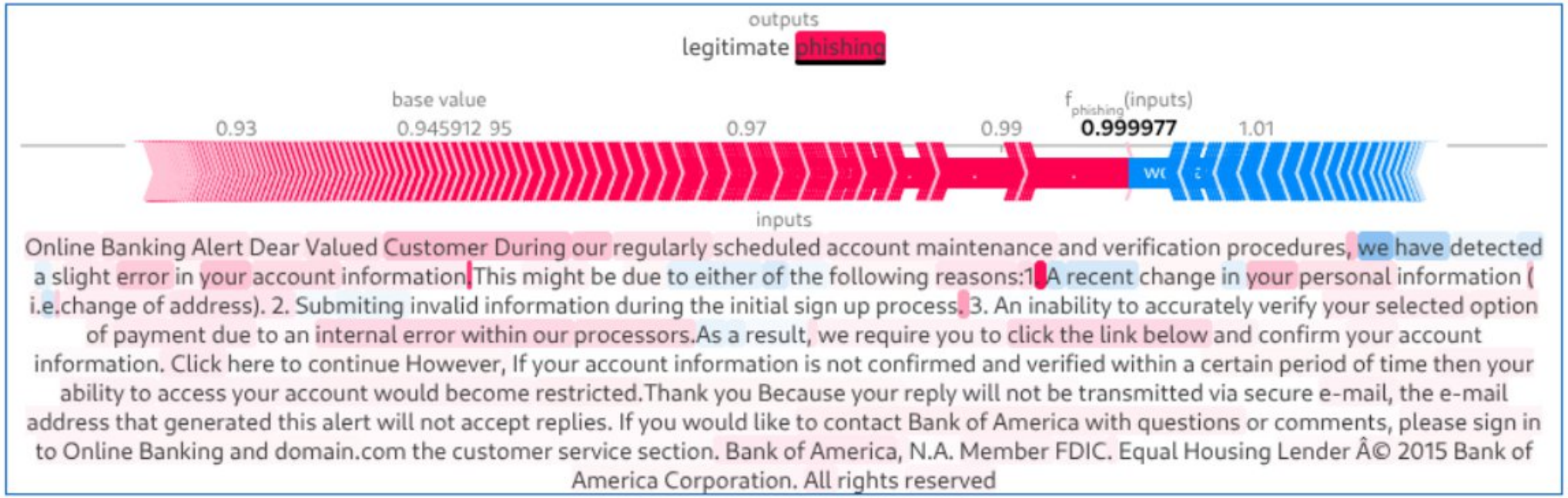


Figure 2: SHAP text plot explaining the model's prediction for a phishing email. Tokens in red increase the probability of the phishing class, tokens in blue decrease it. The visualization reveals a nuanced decision pattern: "alert"-fake urgency, "verify your account"-invitation to click, "click here"-suspicious link, "restricted"-threat of losing access, "Bank of America"-brand impersonation, "do not reply"-attempt to avoid reporting

5. Results

téicée® la raison de votre confiance

GREYC Electronics and Computer Science Laboratory

Comparison with litterature

Table 6: Comparison on Chiper Ionescu (this work, same test split) and reported LLM-based phishing detectors (other datasets; not directly comparable).

Source	Model	Accuracy (%)	Precision (%)	Recall (%)	GPU time (ms)
Character-level baselines (Chiper and Ionescu, 2025) (Chiper Ionescu)					
(Chiper and Ionescu, 2025)	CharCNN	93.71	83.73	74.57	—
(Chiper and Ionescu, 2025)	CharGRU	97.69	97.65	97.71	9
(Chiper and Ionescu, 2025)	CharBiLSTM	96.62	96.61	96.60	—
Our fine-tuned models (Chiper Ionescu)					
This work	CamemBERT-base	98.64	98.78	98.42	2
This work	XLM-RoBERTa	99.46	99.25	99.68	3
This work	Mistral 7B	99.57	99.41	99.69	142
Published LLM-based methods (other datasets)					
(Atawneh and Aljehani, 2023)	BERT	99.6	—	—	—
(Gogoi and Ahmed, 2022)	BERT / DistilBERT	99.0	—	—	—
(Kuikel et al., 2025)	BERT	98.5	—	—	—
(Uddin and Sarker, 2024)	RoBERTa	97.9	—	—	—
(Koide et al., 2024)	GPT-4 (prompting)	99.7	—	—	—

5. Results

Comparison with litterature

Table 6: Comparison on Chiper Ionescu (this work, same test split) and reported LLM-based phishing detectors (other datasets; not directly comparable).

Source	Model	Accuracy (%)	Precision (%)	Recall (%)	GPU time (ms)
Character-level baselines (Chiper and Ionescu, 2025) (Chiper Ionescu)					
(Chiper and Ionescu, 2025)	CharCNN	93.71	83.73	74.57	—
(Chiper and Ionescu, 2025)	CharGRU	97.69	97.65	97.71	9
(Chiper and Ionescu, 2025)	CharBiLSTM	96.62	96.61	96.60	—
Our fine-tuned models (Chiper Ionescu)					
This work	CamemBERT-base	98.64	98.78	98.42	2
This work	XLM-RoBERTa	99.46	99.25	99.68	3
This work	Mistral 7B	99.57	99.41	99.69	142
Published LLM-based methods (other datasets)					
(Atawneh and Aljehani, 2023)	BERT	99.6	—	—	—
(Gogoi and Ahmed, 2022)	BERT / DistilBERT	99.0	—	—	—
(Kuikel et al., 2025)	BERT	98.5	—	—	—
(Uddin and Sarker, 2024)	RoBERTa	97.9	—	—	—
(Koide et al., 2024)	GPT-4 (prompting)	99.7	—	—	—

- Our fine-tuned models outperform character-level models (testing with more than 30K emails)
- But slower than character-level models even with GPU
- Other LLMs studies reported ~99% on other corpora (in general with very small testing set)

Conclusion

- A global pipeline for processing raw emails and LLMs Fine-tuning
- A comparison of three pretrained LLMs under the same protocol, datasets, and metrics
- Good generalization capability
- XLM-RoBERTa offers the best balance of accuracy, cross-dataset robustness, and deployment cost

Future works

- Errors analysis could be made with a much bigger labeled dataset
- Rigorous quantitative evaluation of explanation faithfulness, adversarial robustness and concept drift
- Sharing benchmarks with raw MIME emails to improve comparability across LLM-based detectors

téicée®
la raison de votre confiance

GREYC
Electronics and Computer Science Laboratory

Thanks for your kind attention!

Questions ?