Harnessing AI Cognition:

A Study on LLMs for Spam Email Classification

Zhen Ze Ong

CST--Math and Computer Science, Southern Arkansas University

CSCI4223: Cyber Forensics

Dr. Rami Alroobi

May 13, 2024

Abstract:

Spam tactics are always on the rise. Therefore, traditional machine learning solutions for spam detection are increasingly challenged by the need for adaptive and intelligent models that will be able to mimic elements of human reasoning. This study tries to evaluate the efficacy of five important large language models (LLMs)—Llama2, Mistral, Synthia, Zephyr, and CausalLM—in classifying spam emails. These models were selected for their diverse architectures and were tested on a uniform dataset to assess their spam detection capabilities without prior specific training. The ground truth is known and used for assessment. The models operated under a zero-shot learning framework, and tested mainly with limited computing infrastructure. The results show that while Mistral, Synthia, and Zephyr achieved high accuracy rates of 85-90%, Llama2 and CausalLM underperformed, highlighting issues related to task comprehension and response generation considering the smaller models–-7b models—tested. Additionally, response time analysis revealed that although effective, the processing speed of these models could be a limitation for real-time applications and that more computational resources are essential for effective deployment. In conclusion, this study suggests that LLMs can be integrated as a core layer for more complicated spam detection processes and could potentially enhance the adaptability and accuracy of spam detection. However, more future research to optimize LLMs responsiveness and real-time learning capabilities in cybersecurity contexts will be necessary.

1. Introduction

1.1 Objectives

In the modern era of communication, spam emails are becoming more and more prevalent. According to a blog in 2023, the daily total number of spam emails grew from 269 billion in 2017 to 333.2 billion to 2022 (Ellis and Brandl). Malicious attackers have also started using new techniques like spoofing to trick unsuspecting victims. The objective of this research is to test the capabilities as well as limitations of artificial intelligence (AI) to distinguish spam email in a zero-shot environment.

1.2 Importance of spam detection and the role of AI

AI's capabilities to understand the context of human language are increasing every year. Attackers are finding new ways to obscure their malicious intent, and their intentions may not be detected by classic spam filters. Technologically unsavvy individuals may also be targeted by attackers to break into corporate systems, by creating personalized malicious emails. In the future, AI could be implemented to detect the harmful intent of emails based on the context of the emails.

1.3 Brief overview of LLMs

Large Language Models (LLMs)—such as OpenAI's ChatGPT and Google's Gemini—have caught public interest in recent years. Their ability to answer unique questions, write long lines of code, and summarize long documents have quickly become noticed by many. By using multi-dimensional vectors, LLMs can determine a 'mathematical' relationship between words and link them sequentially to produce a logically sound sentence (“What are Large Language Models? - LLM AI Explained - AWS”). They can also seemingly understand human language by using similar logic. This allows them to understand the content of emails, which could be used to detect malicious emails.

2. Background and Related Work

2.1 Traditional spam detection techniques

Traditional spam detection relies on predetermined rules and features to identify malicious emails. Keyword filtering and pattern analysis look for specific words or phrases commonly found in spam emails, like "free money" or "click here to win." Excessive capitalization and high ratio of URLs to text are also indicators of illegitimate emails. It is also possible to blacklist known spammer email addresses and domains by analyzing spam complaints and suspicious activities.

These techniques are simple and straightforward, but they can be fooled by spammers who can adapt their tactics to work around known detection techniques. Clever attackers can also target specific individuals with carefully crafted emails, which can go unnoticed by traditional detection techniques.

2.2 Overview of large language models and their general applications

Large language models (LLMs) are a type of artificial intelligence (AI) that are trained on massive amounts of text data, allowing them to understand and respond to human language to a large extent. They are often used for text generation, translation between different languages, and chatbots. LLMs are under constant development, but they have the potential to revolutionize many fields that rely on text processing and communication.

2.3 Modern AI spam detection techniques

Machine learning and deep learning have shown promise in spam detection. These techniques can learn from labeled data (spam vs. not spam) and identify complex patterns that traditional methods might miss.

Supervised machine learning algorithms like Naive Bayes and Support Vector Machines (SVM) have been used to classify emails based on features extracted from the content and sender information. Naive Bayes uses conditional probability to estimate the likelihood of certain events based on other external events (Garvey). SVM plots data of different labels and tries to find the hyperplane, or best fit, between the two labels. (Shreyak) It plots new data and determines which side of the hyperplane it lies, and uses that for classification. Deep learning models like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) can learn more intricate patterns from the raw text data, potentially leading to higher accuracy.

However, although these techniques seem more effective than traditional techniques, they require large amounts of labeled data for training. This causes them to fall behind as spam methods continue to evolve.

3. Methodology

3.1 Selected large language models

In this experiment, five models—Llama2, Mistral, Synthia, Zephyr, and CausalLM—were tested on the same dataset of spam and ham emails. They were all given the same prompt of classifying whether each email was spam or ham.

Llama2, released by Meta AI, is a transformer-based autoregressive causal language model. Llama2 models are open source and free for research and commercial use. It is pretrained on publicly available online data and iteratively refined using Reinforcement Learning from Human Feedback (RLHF) (*Meta Llama 2*, n.d.).

Mistral was developed by Mistral AI, a French company composed of previous Meta and Google employees. It uses grouped-query attention (GQA) and sliding window attention (SWA). The developers claim that Mistral 7B surpasses Llama2 13B in its ability to follow instructions (Jiang et al., 2023, 1).

Synthia, developed by Determ, is used to summarize large texts and analyze reports. Determ claims that it can perform advanced sentiment and tone in speech while having the ability to get a big picture of patterns and trends (Homer, n.d.).

Zephyr was developed to identify patterns in healthcare. Its algorithms are built to detect biological signals and insights from real-world data (*Harnessing Real World Data to Democratize Precision Medicine*, n.d.).

CausalLM (Causal Language Modeling) by HuggingFace is a decoder-only model, which means that it takes a sequence of previous tokens, and outputs a single output of the next token. They only look at past tokens to predict the following token. It is opposed to encoder models, which look at an entire sequence of inputs (Lebryk, 2024).

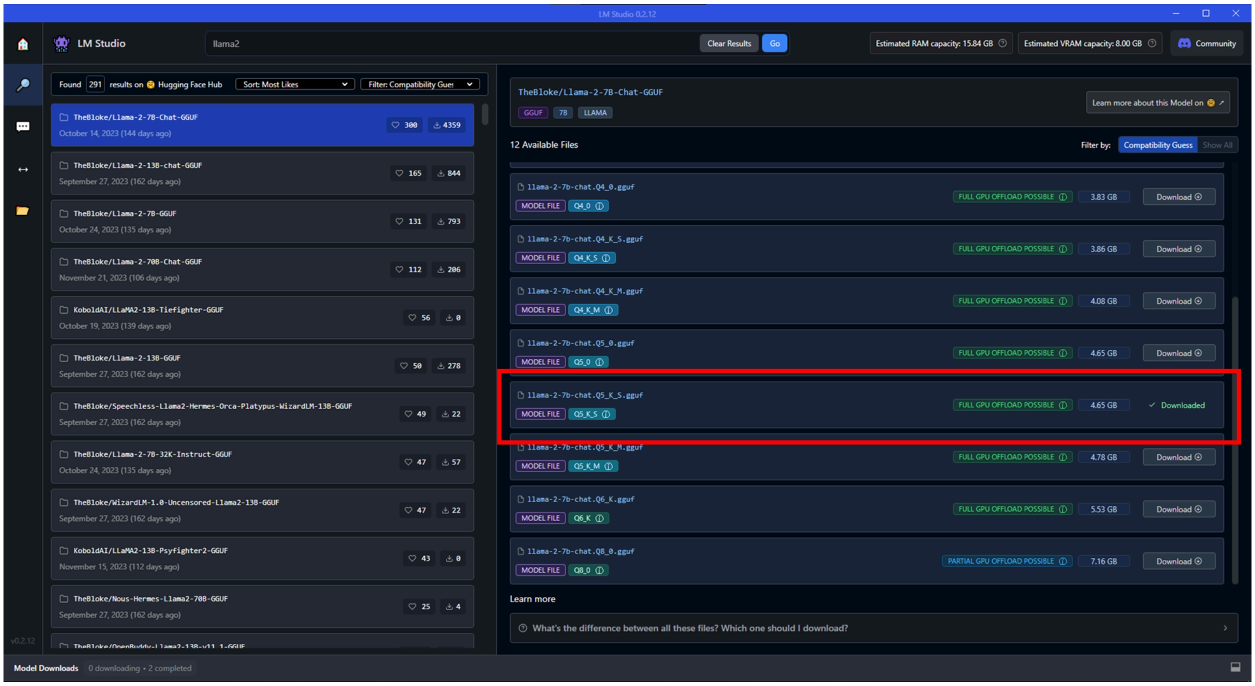
3.2 Dataset and preprocessing steps

The models were tested using a set of 100 randomly selected emails taken from two datasets one with 4993 emails and another with 5729 emails. The datasets consisted of email content along with labels of either ‘spam’ or ‘ham’. As the experiment was done to compare the capabilities of different LLM models, it was crucial to test them on the same 100 emails. The same 100 emails were run through each model ten times, to test the average performance and response speed.

Preprocessing was done manually on Microsoft Excel. To avoid inconsistencies in the results, rows with empty or duplicate values were removed using Excel’s search and replace function.

3.3 Experimental setup and configuration

The LLM models were downloaded and hosted through LM Studio, and connected to a Jupyter Notebook using the OpenAI API. The models can operate through the OpenAI API the same way as OpenAI’s LLM. They were tested with a temperature—a factor of randomness—of 0.0, so that the results will have no variation between multiple tests. Only if the result was invalid—due to error in the LLM—the temperature will be increased to 0.7, to attempt to get a different and valid output.



Found 291 results on Hugging Face Hub

TheBloke/Llama-2-7B-Chat-GGUF 300 4359
October 14, 2023 (144 days ago)

TheBloke/Llama-2-13B-chat-GGUF 165 844
September 27, 2023 (162 days ago)

TheBloke/Llama-2-7B-GGUF 131 793
October 24, 2023 (135 days ago)

TheBloke/Llama-2-70B-Chat-GGUF 112 206
November 21, 2023 (106 days ago)

KoboldAI/LLaMA2-13B-Tiefighter-GGUF 56 0
October 19, 2023 (139 days ago)

TheBloke/Llama-2-13B-GGUF 50 278
September 27, 2023 (162 days ago)

TheBloke/Speechless-Llama2-Hermes-Orca-Platypus-WizardLM-13B-GGUF 49 22
September 27, 2023 (162 days ago)

TheBloke/Llama-2-7B-32K-Instruct-GGUF 47 57
October 24, 2023 (135 days ago)

TheBloke/WizardLM-1.0-Uncensored-Llama2-13B-GGUF 47 22
September 27, 2023 (162 days ago)

KoboldAI/LLaMA2-13B-Psyfighter2-GGUF 43 0
November 15, 2023 (112 days ago)

TheBloke/Nous-Hermes-Llama2-70B-GGUF 25 4
September 27, 2023 (162 days ago)

TheBloke/OpenBuddy-Llama2-13B-v11.1-GGUF

TheBloke/Llama-2-7B-Chat-GGUF

12 Available Files

File	Type	Quantization	GPU offload	Size	Status
llama-2-7b-chat.Q4_0.gguf	MODEL FILE	Q4_0	FULL GPU OFFLOAD POSSIBLE	3.83 GB	Download
llama-2-7b-chat.Q4_K_S.gguf	MODEL FILE	Q4_K_S	FULL GPU OFFLOAD POSSIBLE	3.86 GB	Download
llama-2-7b-chat.Q4_K_M.gguf	MODEL FILE	Q4_K_M	FULL GPU OFFLOAD POSSIBLE	4.08 GB	Download
llama-2-7b-chat.Q5_0.gguf	MODEL FILE	Q5_0	FULL GPU OFFLOAD POSSIBLE	4.65 GB	Download
llama-2-7b-chat.Q5_K_S.gguf	MODEL FILE	Q5_K_S	FULL GPU OFFLOAD POSSIBLE	4.65 GB	✓ Downloaded
llama-2-7b-chat.Q5_K_M.gguf	MODEL FILE	Q5_K_M	FULL GPU OFFLOAD POSSIBLE	4.78 GB	Download
llama-2-7b-chat.Q6_K.gguf	MODEL FILE	Q6_K	FULL GPU OFFLOAD POSSIBLE	5.53 GB	Download
llama-2-7b-chat.Q8_0.gguf	MODEL FILE	Q8_0	PARTIAL GPU OFFLOAD POSSIBLE	7.16 GB	Download

Learn more

What's the difference between all these files? Which one should I download?

Model Downloads 0 downloading • 2 completed

Downloading Llama2 7B model through LM Studio

Another set of tests were run using an initial temperature of 0.7, to test the consistency—whether a model responds differently for the same email—of each model. If an output is invalid, the test is run again with the same parameters, since the output is already random.

The models were given a prompt to only return ‘THIS IS SPAM’ or 'THIS IS NOT SPAM’ as their outputs. Any other output is considered invalid and disregarded. The models’ stream parameters are set to true, meaning that they output every token sequentially instead of waiting for the entire output to be completed before returning. By using a streaming, the output can be checked for errors and stopped before the LLM starts hallucinating and returns an extremely long output, taking minutes to generate a result.

3.4 Evaluation metrics

The accuracy of each model is evaluated by counting the number of matching results out of the total 100 emails. If the model classifies 80 out of 100 emails correctly either as spam or not spam, it is considered 80% accurate.

Precision is the calculation of $\frac{tp}{tp + fp}$ where True Positive (TP) is when an email that is actually spam and is correctly identified as spam by the model. False Positive (FP) is when an email that is not spam but is mistakenly flagged as spam by the model. Other classifications can be True Negative (TN) which indicates that an email that is not spam and is correctly identified as not spam and False Negative (FN) where an email that is actually spam but is missed by the model. High precision is crucial when the cost of a false positive is high. In the context of email classification, a false positive (wrongly identifying a legitimate email as spam) could lead to removing a good email.

Recall is calculated using $\frac{tp}{tp + fn}$. High recall is crucial when missing a positive case is expensive. A false negative (identifying a spam email as legitimate email) could lead to spam emails slipping through the filter.

An F1 score is calculated as $\frac{2 \times Precision \times Recall}{Precision + Recall}$. It is a balanced view when both precision and recall are important. This is the best indicator of how well a model performs in determining if an email is spam or not.

4. Results

4.1 Detailed presentation of the results for each LLM

Llama2:

```
167.34124 total seconds
1.67341 seconds per query
CPU utilization: 56.30%
Memory utilization: 82.80%
Number of matches: 29
Number of mismatches: 71
Accuracy: 28.999999999999996%
TPs: 26
TNs: 3
FPs: 71
FNs: 0
Precision: 0.26804123711340205
Recall: 1.0
F1 Score: 0.42276422764227645
```

A run of the Llama2 model produced the above results. It has a low precision of 0.268 and high recall of 1.0, meaning that it identifies a majority of emails as spam while never incorrectly identifying a spam email as non-spam. Its low F1 score of 0.423 indicates that Llama2 is not an optimal model to identify spam emails from non-spam emails. It has an average response time of 1.673 seconds per email.

Mistral:

```
179.90660 total seconds
1.79907 seconds per query
CPU utilization: 62.80%
Memory utilization: 85.60%
Number of matches: 88
Number of mismatches: 12
Accuracy: 88.0%
TPs: 26
TNs: 62
FPs: 12
FNs: 0
Precision: 0.6842105263157895
Recall: 1.0
F1 Score: 0.8125000000000001
```

Mistral had several false positives, meaning that it identified legitimate emails as spam. This is seen as precision is 0.684. However, it has a recall of 1.0, which means that it never identified spam as legitimate email. It has the highest F1 score among the models of 0.813 while having an average response time of 1.799 seconds per email.

Sythia:

```
179.00845 total seconds
1.79008 seconds per query
CPU utilization: 59.70%
Memory utilization: 86.30%
Number of matches: 90
Number of mismatches: 10
Accuracy: 90.0%
TPs: 16
TNs: 74
FPs: 0
FNs: 10
Precision: 1.0
Recall: 0.6153846153846154
F1 Score: 0.761904761904762
```

Synthia had a precision of 1.0, meaning that it never identified non-spam email as spam. However, all of its inaccurate results were several spam emails identified as non-spam emails, causing its recall to be 0.615 and the final F1 score to be 0.762. This indicates that Synthia is more trusting of emails, leading it to accepting spam as non-spam, but is unlikely to falsely identify legitimate emails as spam. It has an average response time of 1.790 seconds per email.

Zephyr:

```
274.04934 total seconds
2.74049 seconds per query
CPU utilization: 57.60%
Memory utilization: 83.40%
Number of matches: 86
Number of mismatches: 14
Accuracy: 86.0%
TPs: 20
TNs: 66
```

```
FPs: 8
FNs: 6
Precision: 0.7142857142857143
Recall: 0.7692307692307693
F1 Score: 0.7407407407407408
```

Zephyr has the most similar precision and recall values—0.714 and 0.769 respectively—among the tested models. This indicates that it was not too biased in determining whether an email is spam or not. However, it has the longest response time of 2.740 seconds, which can be due to the lack of bias in its conclusions.

CausalLM:

```
167.34298 total seconds
1.67343 seconds per query
CPU utilization: 64.50%
Memory utilization: 87.90%
Number of matches: 38
Number of mismatches: 62
Accuracy: 38.0%
TPs: 26
TNs: 12
FPs: 62
FNs: 0
Precision: 0.29545454545454547
Recall: 1.0
F1 Score: 0.456140350877193
```

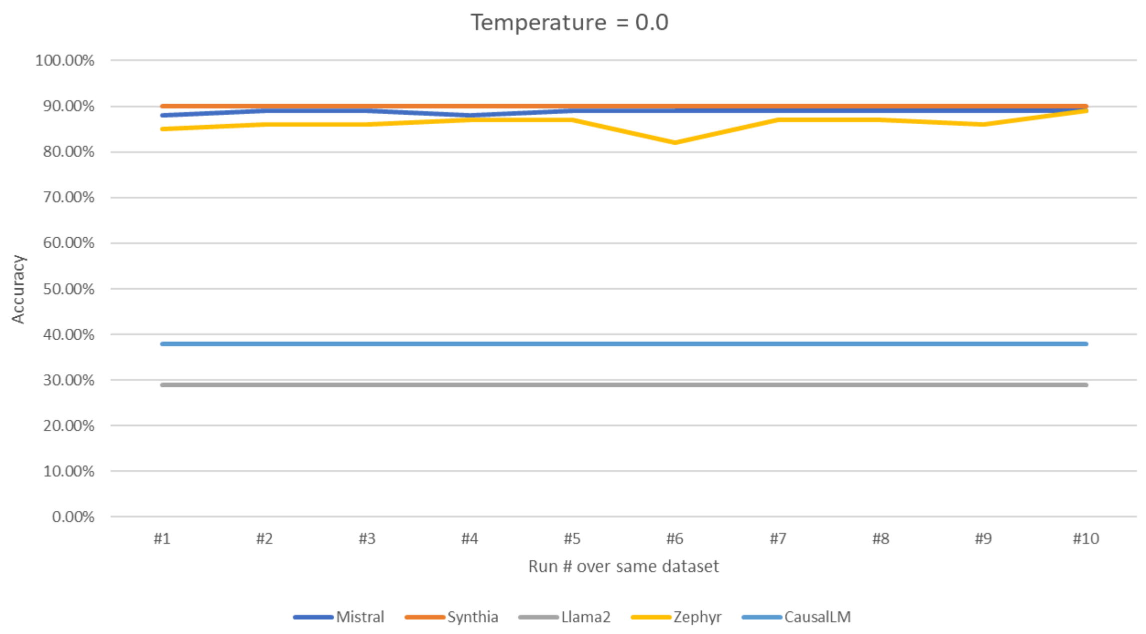
CausalLM was highly inaccurate, with a large number of false positives. Similarly to Llama2, it identifies most emails as spam emails, causing a low precision of 0.295. It has no false negatives, meaning that it never identified spam emails as non-spam emails. It has an average response time of 1.673 seconds per email.

4.2 Comparative analysis of the models based on performance metrics

Mistral had the highest F1 score of 0.813, with Synthia and Zephyr—0.762 and 0.741 respectively—slightly behind. CausalLM and Llama2 both had lackluster scores of 0.423 and

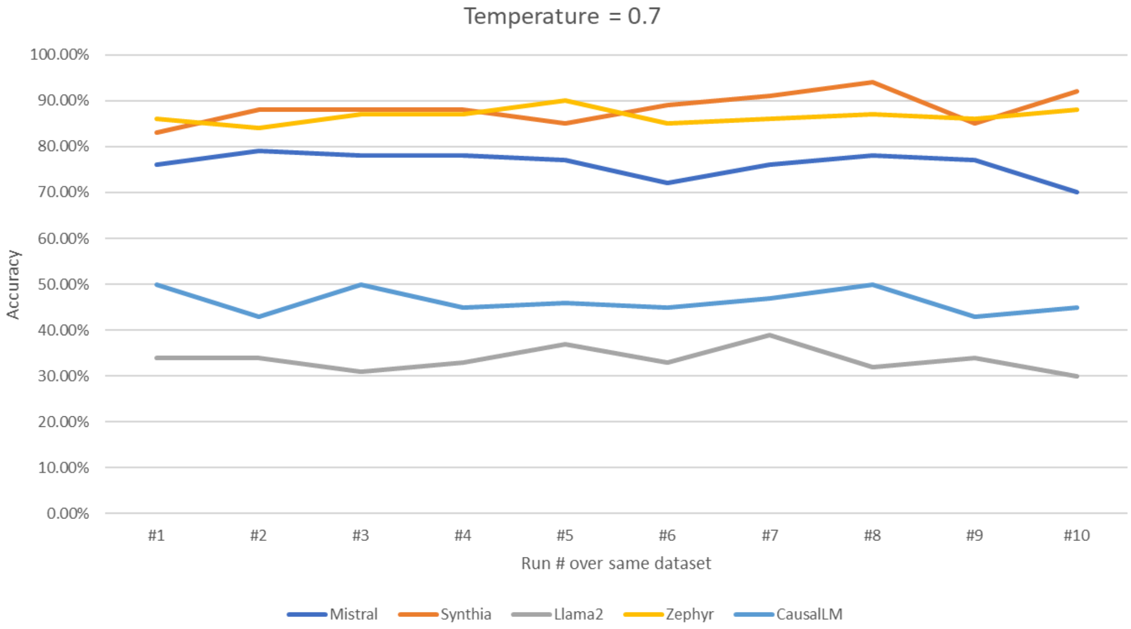
0.456 respectively. Synthia had the highest accuracy, correctly identifying 90 out of 100 emails. Zephyr had the highest response time of 2.740 seconds per email. There was little variation of CPU and RAM usage between all the models.

4.3 Visuals



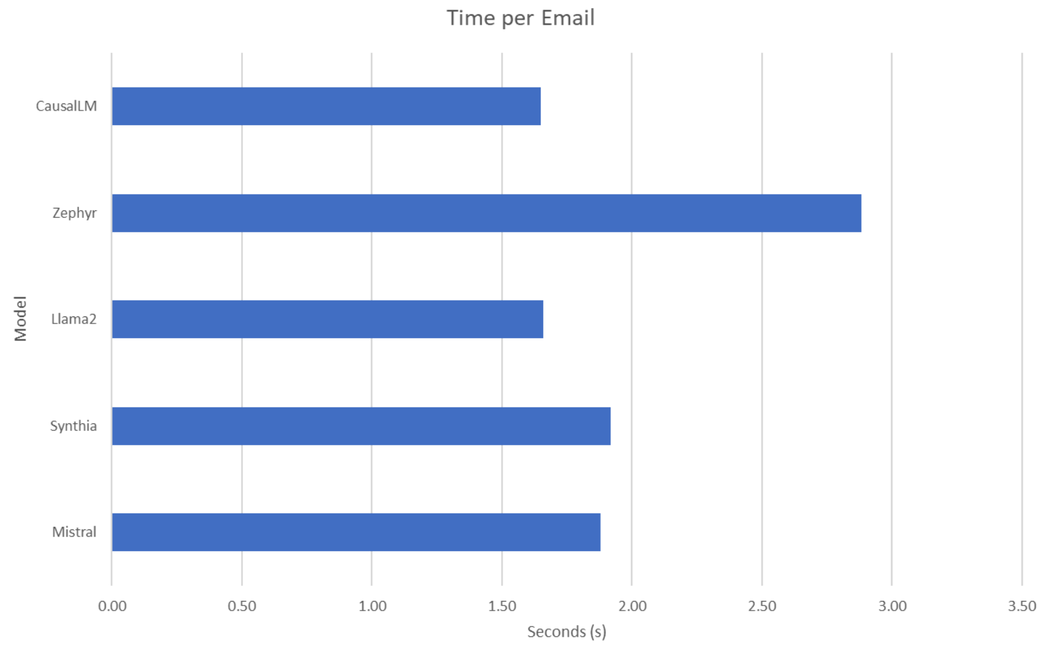
Accuracy consistency of models at 0.0 temperature

While the temperature was set to 0.0, the most significantly different number of accurate results were given by Zephyr, since it had the most initial invalid responses. Subsequent responses were returned at a temperature of 0.7. Other models had little to no variations in response over 10 runs, as expected.



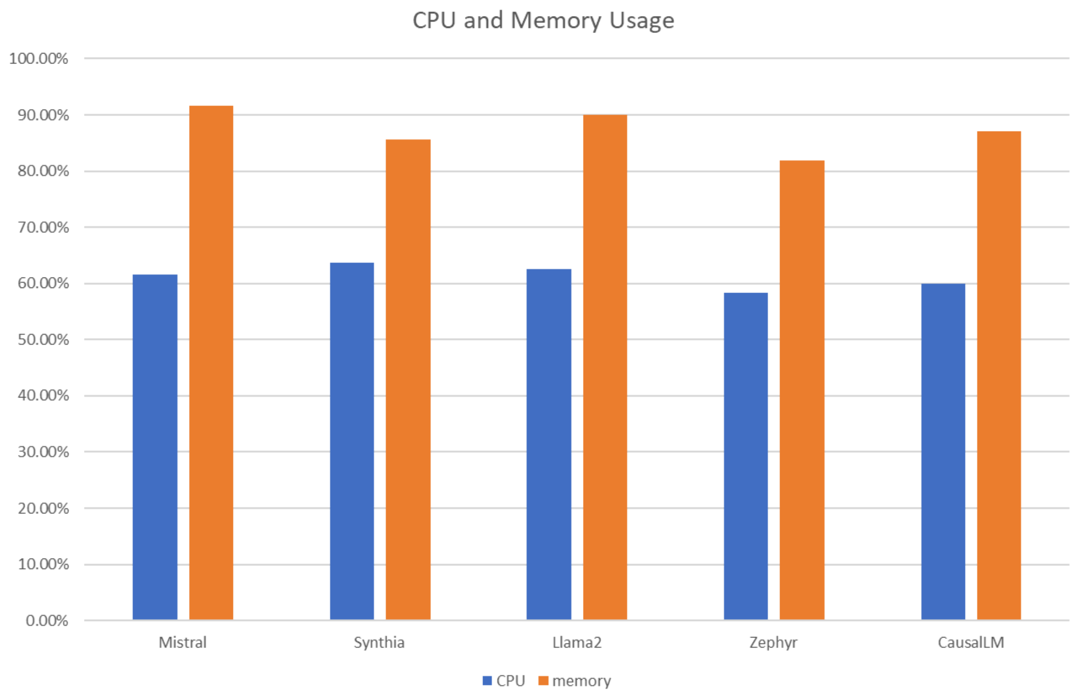
Accuracy consistency of models at 0.7 temperature

When the temperature was increased to 0.7, all models had varying numbers of accurate responses. The better performing models—Mistral, Synthia, and Zephyr—began having lower accuracy numbers with the most significant drop being Mistral, falling below 80% accuracy, which indicates that models might perform classification better at a lower temperature. Llama2 and CausalLM both rose in performance slightly, but still had accuracy numbers below 50%.



Average classification time of models

As previously mentioned, Zephyr had the highest average response time per email of roughly 2.80 seconds. CausalLM and Llama2 had the lowest average response times around 1.65 seconds, which might indicate their poor performance in classification. Synthia and Mistral took an average of about 1.80 seconds.



Average CPU and memory usage of models

The CPU and memory usage between models both had no outstanding differences. Any slight differences could be a cause of background applications affecting the system. This might change if a model has an increase or decrease in parameters, as all the models tested here were 7B.

5. Discussion

5.1 Interpretation of the results

As seen in the results, some models significantly outperform others in classification. The results might not be an indicator of how good an LLM is overall, as they might only be ineffective in classification. The better performing models—Mistral, Synthia, and Zephyr—had average response times ranging from 1.80 to 2.80 seconds per email. These long response times are due to the nature of LLMs, since they have to be fed with input through tokens and return an output through tokens as well.

5.2 Analysis of the performance differences among the models

The different performances between the models could be a result of the different ways the models were trained and reinforced. A shorter response time by Llama2 and CausalLM seems to be correlated with their poor performances, but Zephyr has slightly worse results than Mistral and Sythia despite having a significantly longer response time.

5.3 Assessment of the practical implications of the findings

Since Mistral had the highest F1 score in the results, it might indicate that the use of grouped-query attention and sliding window attention could have some benefit to LLM classification. CausalLM and Llama2 seem to have a stronger focus on causal language modeling, which might cause them to provide unintended responses when returning shorter outputs like ‘THIS IS SPAM’ and ‘THIS IS NOT SPAM’. Further research into these aspects might be beneficial to improving classification using LLMs.

5.4 Evaluation of the models' limitations and strengths

Mistral has the best performance statistically, based on its F1 score. It also has a similar average response time to Sythia, which had the highest accuracy of 90 correct classifications out of 100. Despite having a longer response time than the other two, Zephyr does not lean heavily on classifying most emails either as spam or not spam, which all the other models seem to have a bias on either. In the context of email classification, incorrectly identifying spam as non-spam and legitimate emails as spam both have negative consequences, which mean that a model with a high F1 score like Mistral is the most beneficial.

6. Future Work

6.1 Suggestions for improving the models' performance

As these experiments were conducted on a personal desktop, the models should see an increased performance in speed if hosted on a more powerful computer or cloud platform. Additional tunings of the temperatures of individual models could improve their classification accuracy. The models were using 7b parameters, so exploring an increase or decrease of parameters could also have a positive impact on the accuracy.

6.2 Potential areas for further research

Further research of LLMs in spam detection could be conducted. Traditional techniques, and other AI methods—machine learning and deep learning—could be compared alongside LLMs to find differences between them. LLMs could also have the potential to break down and describe why it thinks an email is spam, and be tested for its accuracy in that. LLMs can act as a personal assistant for individuals, while filtering their email inboxes.

6.3 Discussion of upcoming technologies that could enhance LLM capabilities

In the future, other AI technologies could be integrated with LLMs, which can potentially improve its classification capabilities. One of them could be computer vision, allowing the LLM to view any attached images that may seem suspicious. It may also be able to detect visual differences of structure or details between real and phishing emails. There is also research on self-improving AI and models that can perform causal learning, allowing them to reason and make better decisions. Allowing AI to mimic closer to human reasoning will improve its chances of detecting spam similar to a person with proper training.

7. Conclusion

7.1 Summary of the research outcomes

LLM capabilities for spam detection vary between models, with some being more accurate than others. Different models may approach the content of the emails differently. Mistral, Synthia, and Zephyr had good results of 85-90% accuracy, but Zephyr had a far longer average response time of 2.9 seconds, likely due to its repetitive invalid outputs. CausalLM and Llama2 were extremely inaccurate having 30-40% accuracy while having the shortest response times of 1.6 seconds. Mistral's accuracy dipped slightly when the temperature was increased to 0.7. These models operated under a zero-shot learning environment, with the best results of 85-90% accuracy.

7.2 Final thoughts on the impact of LLMs on spam detection

The main limitation of LLMs is the response time, with an average time of 1.8 seconds per email for accurate responses. This is not optimal in real world scenarios, where millions of emails are being sent and received every second. However, LLMs could be used as an additional layer of protection in specific cases, like in corporations or systems with a high need of security. The processing speed of LLMs may also improve as the field of AI technology continues to grow.

7.3 Reflections on the future of AI in cybersecurity

AI is a modern tool that can have both a positive and negative impact on security. AI is starting to be used by attackers to automate and adapt their attacks. Attackers might also develop techniques that can break AI-based security defenses, like manipulating data to trick AI models. AI might also be biased in their methods, which can lead to unfair or discriminatory security actions. Despite those, AI can analyze a large amount of data to look for malicious activity. It could potentially detect zero-day attacks that traditional pattern-based detection techniques might miss. AI might be able to prioritize alerts and predict future attacks, allowing for a more proactive approach to cybersecurity.

References

Ellis, C., & Brandl, R. (2023, October 19). *Spam Statistics 2024: Survey on Junk Email, AI Scams & Phishing*. Email Tool Tester. Retrieved May 9, 2024, from https://www.emailtooltester.com/en/blog/spam-statistics/

Garvey, M. (2020, November 30). *Naïve Bayes Spam Filter — From Scratch | by Mark Garvey*. Towards Data Science. Retrieved May 9, 2024, from https://towardsdatascience.com/na%C3%AFve-bayes-spam-filter-from-scratch-12970ad3dae7

Harnessing Real World Data to Democratize Precision Medicine. (n.d.). Zephyr AI. Retrieved May 10, 2024, from https://www.zephyrai.bio

Homer, N. (n.d.). *AI Assistant Synthia*. Determ. Retrieved May 10, 2024, from https://www.determ.com/features-synthia-ai/

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lauchaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & William El Sayed. (2023, October 10). Mistral 7B. https://arxiv.org/abs/2310.06825v1

Lebryk, T. (2024, March 4). *Training CausalLM Models Part 1: What Actually Is CausalLM?* Towards Data Science. Retrieved May 10, 2024, from https://towardsdatascience.com/training-causallm-models-part-1-what-actually-is-causallm-6c3efb2490ec

Meta Llama 2. (n.d.). Meta Llama. Retrieved May 10, 2024, from https://llama.meta.com/llama2/

Shreyak. (2020, August 5). *Spam Mail Detection Using Support Vector Machine*. Becoming Human: Artificial Intelligence Magazine. Retrieved May 9, 2024, from https://becominghuman.ai/spam-mail-detection-using-support-vector-machine-cdb57b0d62a8

What are Large Language Models? - LLM AI Explained - AWS. (n.d.). Amazon AWS. Retrieved May 9, 2024, from https://aws.amazon.com/what-is/large-language-model/