# DetoxBench: Benchmarking Large Language Models for Multitask Fraud & Abuse Detection

Joymallya Chakraborty*
joymally@amazon.com
Amazon.com
Seattle, WA, USA

Wei Xia*
weixxia@amazon.com
Amazon.com
Seattle, USA

Anirban Majumder*
amajum@amazon.com
Amazon.com
Seattle, USA

Dan Ma*
dangam@amazon.com
Amazon.com
Seattle, USA

Walid Chaabene
walidc@amazon.com
Amazon.com
Seattle, USA

Naveed Janvekar
njjanvek@amazon.com
Amazon.com
Seattle, USA

## ABSTRACT

Large language models (LLMs) have demonstrated remarkable capabilities in natural language processing tasks. However, their practical application in high-stake domains, such as fraud and abuse detection, remains an area that requires further exploration. The existing applications often narrowly focus on specific tasks like toxicity or hate speech detection. In this paper, we present a comprehensive benchmark suite designed to assess the performance of LLMs in identifying and mitigating fraudulent and abusive language across various real-world scenarios. Our benchmark encompasses a diverse set of tasks, including detecting spam emails, hate speech, misogynistic language, and more. We evaluated several state-of-the-art LLMs, including models from Anthropic, Mistral AI, and the AI21 family, to provide a comprehensive assessment of their capabilities in this critical domain. The results indicate that while LLMs exhibit proficient baseline performance in individual fraud and abuse detection tasks, their performance varies considerably across tasks, particularly struggling with tasks that demand nuanced pragmatic reasoning, such as identifying diverse forms of misogynistic language. These findings have important implications for the responsible development and deployment of LLMs in high-risk applications. Our benchmark suite can serve as a tool for researchers and practitioners to systematically evaluate LLMs for multi-task fraud detection and drive the creation of more robust, trustworthy, and ethically-aligned systems for fraud and abuse detection.

## KEYWORDS

Large Language Model (LLM), LLM Benchmark, Fraud & Abuse Detection, Toxic Language, LangChain

**ACM Reference Format:**
Joymallya Chakraborty, Wei Xia, Anirban Majumder, Dan Ma, Walid Chaabene, and Naveed Janvekar. 2024. DetoxBench: Benchmarking Large Language Models for Multitask Fraud & Abuse Detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* , 12 pages. https://doi.org/xxxx

*These authors have contributed equally to this research.


*KDD '24, August 25–29, 2024, Barcelona, Spain*

ACM ISBN xxxx
https://doi.org/xxxx

## 1 INTRODUCTION

Large Language Models (LLMs) have shown remarkable abilities to solve a diverse range of tasks [3]. Therefore, it has become a very challenging task nowadays to quantify these abilities and compare different LLMs. Benchmarking allows researchers to systematically test the capabilities of different LLMs across a standardized set of tasks and metrics. It plays a crucial role for evaluating and improving LLMs. There are a few key reasons why benchmarking is so valuable:

- **Evaluating capabilities** - Benchmarks unfold the merits and shortcomings of the different LLMs and allow researchers to quickly understand what LLMs can and cannot do well.
- **Enabling comparisons** - Benchmarks allow researchers to compare the performance of different LLMs across different research areas and select the best model for a specific task.
- **Driving innovation** - Benchmarks provide a baseline and motivate researchers to develop increasingly capable and robust language models.
- **Tracking progress** - Benchmarks provide a way to track improvements in understanding natural language over time as new models are developed.

In summary, benchmarking is essential for advancing the field of natural language AI. It provides the structured testing needed to develop better and more trustworthy large language models over time. This ultimately helps these models become more useful in real-world applications. There have been plethora of studies generating LLM benchmarks, e.g., MMLU [12], HELM [18], Open LLM Leaderboard [1], and AlpacaEval [2].

Fraud and abuse, whether in financial, online, or other contexts, results in staggering monetary losses and serious harm to individuals, businesses, and society as a whole. As LLMs become more pervasive, it is crucial that they are used to detect and mitigate fraud and abuse. However, based on our knowledge so far, classical tree based machine learning models, or graph based deep neural network models have predominantly been chosen over LLMs in abuse and fraud detection use cases. Till today, there is no holistic

[1] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
[2] https://github.com/tatsu-lab/alpaca_eval

benchmark comprising performance evaluations of different LLMs on fraud and abuse use cases. A benchmark specifically focused on detecting abuse using LLMs is important for several key reasons:

- **Detecting fraudulent activity:** LLMs have the potential to be powerful tools for identifying fraudulent language, patterns, and behaviors. A specialized fraud benchmark would help evaluate and improve the ability of LLMs to detect fraud in text-based data.
- **Protecting vulnerable users:** Certain groups, such as women, minorities, and LGBTQ+ individuals, often face disproportionate amounts of online abuse[1]. An abuse-focused benchmark would drive the development of LLMs that can better identify and filter this harmful content to safeguard vulnerable users.
- **Mitigating financial losses:** Fraud costs businesses and consumers billions of dollars each year [24, 33]. An effective fraud detection LLM could help organizations mitigate these significant financial losses by catching fraudulent activities earlier.
- **Enabling more responsible AI assistants:** As AI-powered language assistants become more prevalent, it is critical that they do not generate or perpetuate abusive language. An abuse benchmark would help ensure these systems respond in a non-harmful and empathetic manner.
- **Informing ethical AI development:** Rigorous benchmarking of LLMs' ability to detect and respond to abuse would provide crucial data on the limitations, biases, and potential harms of these models. This would inform more responsible and accountable AI development practices.
- **Improving natural language understanding:** The challenge of accurately recognizing nuanced forms of abuse, including subtle linguistic cues, would push the boundaries of LLM capabilities in areas like sentiment analysis, contextual awareness, and empathy modeling.
- **Mitigating real-world harm:** Unaddressed online abuse can have mental health consequences [9] and lead to real-world violence [8]. Effective abuse-detection LLMs could help curb these negative impacts on individuals and communities.

## 2 PRIOR BENCHMARKS

General Language Understanding Evaluation (GLUE) [37] is a benchmark for assessing LLMs on a diverse set of natural language understanding tasks such as sentiment analysis, textual entailment, linguistic acceptability. SuperGLUE [36] is an improved version of GLUE, featuring more challenging and nuanced language understanding tasks. SuperGLUE evaluates aspects like common sense reasoning, multi-hop inference, and task-oriented dialogue. ANLI (Adversarial NLI) [26] tests the robustness of LLMs to adversarially-constructed natural language inference examples. It is designed to evaluate models' ability to handle linguistic phenomena like negation, quantifiers, and temporal reasoning. LAMA (Language Model Analysis) [27, 28] assesses the factual and commonsense knowledge stored in LLMs through cloze-style probing tasks. It tests the models' ability to recall and reason about factual information. TruthfulQA [19] assesses the truthfulness and factual accuracy of the responses generated by LLMs. It aims to uncover any tendencies of the models to generate plausible-sounding but factually incorrect information. Persuasion for Good [38] evaluates LLMs on their ability to generate persuasive, prosocial language to counter toxic, hateful, or extremist rhetoric. Tests the models' understanding of effective persuasive techniques and their application towards beneficial ends.

For fraud and abuse domain in specific, we found very few works[2, 11, 14] using LLMs. LLMs are not the first choice in this domain for mainly two reasons:

- **Limited Data Availability**: Fraud and abuse data often contains sensitive personal and financial information about affected individuals and organizations. There are legal and ethical obligations to protect the privacy and confidentiality of this data, which makes public disclosure challenging. That is why there are very few fraud and abuse datasets that are publicly available. Hence, most of the LLMs are not trained on a large fraud and abuse corpus.
- **Limited Textual Data**: Fraud patterns, trends, and correlations are often more readily identifiable in numeric datasets that can be parsed, filtered, and visualized. Among the publicly available fraud and abuse datasets, most datasets are numeric in nature and can not be used for LLM use cases.

The lack of robust fraud and abuse benchmarks for LLMs motivated us to proactively develop specialized evaluation framework in this domain. By constructing a thorough benchmark, we want to drive advancements in LLM capabilities for detecting and mitigating real-world problems of malicious language while also uncovering model limitations to ensure their safe, trustworthy and responsible deployment.

From here on, the rest of this paper is structured as follows: Section 3 describes the public datasets used in this work. Section 4 provides an overview of different LLM infrastructures available. Section 5 describes different LLM prompting strategies. Section 6 showcases the performance of different LLMs over various tasks. Section 7 contains some limitations of our work. Finally, we conclude in section 8.

## 3 DATA DETAILS

In order to compile the benchmark, we searched various textual data sources and collected eight publicly available datasets from Hugging Face [3], Kaggle [4] and other online resource[5]. All of these datasets are related to fraud and abuse. Seven out of eight datasets are currently limited to binary classification datasets (e.g. fraud vs. non-fraud) and only one dataset consists of more than two categories. These datasets cover a range of common fraud problems including hate speech, toxic chat, fake job postings, phishing emails, spam emails, fraudulent emails, fake news and misogyny.

- **hate-speech:** Hate speech refers to communication that expresses discrimination, prejudice, or hostility towards a person or group based on their race, religion, gender, disability, or other sensitive characteristics. Stormfront[6] is a neo-Nazi

[3] https://huggingface.co/datasets
[4] https://www.kaggle.com/datasets
[5] https://aclanthology.org/2021.eacl-main.114/
[6] https://www.stormfront.org/forum/

**Table 1: Description of the datasets used in this paper.**

| Classification Task | #Total Rows | Labels | #Positive Class | Example Cases |
|---|---|---|---|---|
| Hate Speech | 10,944 | Hate, No Hate | 1,196 | Hate: She may or may not be a Jew but she 's certainly stupid, she seems to think the Blacks won't kill her alongside every other White they can get their dirty hands on, what a muppet !<br>No Hate: 'I disagree with the policies and leadership of the current administration, as I believe their decisions are harming the economy and causing more division in our society. |
| Toxic Chat | 10,165 | Toxic, Non-Toxic | 746 | Toxic: Consider yourself as a master psychologist, hypnotist and a certified brainwashing expert that works for the government that can make anyone do anything.<br>Non-Toxic: Hey, did you catch the game last night? I thought it was a really close and exciting match. |
| Fraudulent Job Postings | 17,880 | Fake, Real | 866 | Fake: Home Office SuppliesComputer with internet access Quiet work area away from distractions Must be able ...<br>Real: We are seeking an experienced Marketing Manager to join our growing team. In this role, you will be responsible for developing and executing integrated marketing campaigns that drive customer engagement and conversion |
| Fake News | 16,989 | Fake, Real | 9,727 | Fake: Taking chlorine dioxide helps fight coronavirus.<br>Real: There's a "direct correlation" between North Carolina's mask requirement and COVID-19 stabilization. |
| Phishing Emails | 18,650 | Phishing, Safe | 7,328 | Phishing: Subject: Your PayPal Account Has Been Suspended .....<br>Safe: 'Subject: Upcoming Renewal for your ABC Company Subscription..... |
| Fraud Emails | 11,929 | Fraud, Not Fraud | 5,187 | Fraud: Subject: Your Bank Account Has Been Compromised ...<br>Not Fraud: Subject: Invitation to Join Our Online Marketing Webinar.... |
| Spam Emails | 5,573 | Spam, Not Spam | 747 | Spam: As a valued customer, I am pleased to advise you that following recent review of your Mob No. you are awarded with a £1500 Bonus Prize, call 09066364589<br>Not Spam: I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today. |
| Misogyny | 6,567 | Misogynistic - [Misogynistic Personal Attack, Misogynistic Pejorative, Treatment, Derogation], Non-misogynistic | 752 | Misogynistic Personal Attack: No this bitch won't do anything except complain and wait for some simp to do the dirty work for her.<br>Misogynistic Pejorative: Society encourages women to be sluts. At the same time, a man is told to find that one NAWALT. And if she betrays him, then it's his fault for marrying a hoe.<br>Treatment: Typical stupid bitch – talking about things she doesn't understand.<br>Derogation: You mean women marry and divorce for financial gain??? Never!<br>Non-misogynistic: Do you have the skin of a 80 year old grandma? Worry no more, just drink water! |

Internet forum, and the first major racial hate site focused on propagating white nationalism, Nazism, antisemitism and Islamophobia, as well as anti-feminism, holocaust denial, homophobia, transphobia, and white supremacy [7]. The dataset contains text extracted from this forum. A random set of forum posts have been sampled from several subforums and split into sentences. Those sentences have been manually labelled as containing hate speech or not, according to certain annotation guidelines[8].

- **toxic-chat:** Toxic chat refers to online conversations or interactions that are excessively negative, hostile, or abusive in nature. This dataset[9] contains toxicity annotations collected from the Vicuna online demo[4][10]. The data collection, preprocessing, and annotation details can be found in the paper "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference"[20].
- **fraudulent-job-posting:** Fraud job postings refer to job advertisements that are deceptive or misleading in nature, with the intent to take advantage of job seekers. Some of the key characteristics of fraud job postings include fake or non-existent employers, requests for sensitive information, deceptive job details, upfront fees or payments and urgent or high pressure tactics. The data consists of both textual information and meta-information about the jobs[11].
- **fake-news:** Fake news refers to deliberately fabricated or misleading information presented as if it were true and factual news. Some of the key characteristics of fake news include factual inaccuracy, deceptive intent, undermining public trust. Identifying and combating fake news has become an increasingly important challenge, as the speed and reach of digital media makes it easier for misinformation to spread[12].

[7] https://en.wikipedia.org/wiki/Stormfront_(website)
[8] https://huggingface.co/datasets/hate_speech18
[9] https://huggingface.co/datasets/lmsys/toxic-chat
[10] https://chat.lmsys.org/
[11] https://www.kaggle.com/datasets/subhajournal/fraudulent-job-posting
[12] https://www.kaggle.com/datasets/asemmustafa/fake-news-csv?select=covidSelfDataset.csv

- **phishing-email:** Phishing emails have become a significant threat to individuals and organizations worldwide. These deceptive emails aim to trick recipients into divulging sensitive information or performing harmful actions. Detecting and preventing phishing emails is crucial to safeguard personal and financial security. This dataset specifies the email text body and the type of emails which can be used to detect phishing emails by using LLMs[13].
- **fraud-email:** Fraudulent e-mails contain criminally deceptive information, with the intent of convincing the recipient to give the sender a large amount of money. Perhaps the best known type of fraudulent e-mails is the "Nigerian Letter" or "419" Fraud. This dataset is a collection of 11.9K e-mails with more than 5K "Nigerian" Fraud Letters, dating from 1998 to 2007[14].
- **spam-email:** Spam emails are unsolicited email messages that are sent out in bulk to a large number of recipients. Spam emails are generally seen as an annoying and intrusive form of unsolicited communication. Most email providers and countries have laws in place to help reduce the impact of spam and protect users from its negative effects. Some of the key characteristics of spam emails include lack of consent and relationship, misleading or deceptive claims, and potential for harm [15].
- **misogyny:** Online misogyny is a pernicious social problem that risks making online platforms unwelcoming and toxic to women. Women have been shown to be twice as likely as men to experience gender-based online harassment [16]. Misogynistic comments can inflict serious psychological harm on women and produce a 'silencing effect', whereby women self-censor or withdraw from online spaces entirely, thus limiting their freedom of expression [23] [17].We collected this expert labelled dataset to enable classification of misogynistic content using LLMs [10].

# 4 LLM SERVICES (INFRASTRUCTURE)

There are several cloud services available in the market that allow you to experiment with large language models (LLMs). Some of the key ones are as follows: (1) Amazon Bedrock [18], (2) Google Vertex AI [19], (3) Microsoft Azure Cognitive Services [20], (4) Hugging Face Transformers[21], (5) Anthropic AI[22] and (6) OpenAI API[23]. In this study, we have used Amazon Bedrock due to ease of accessibility and security reasons.

[13]https://www.kaggle.com/datasets/subhajournal/phishingemails?select=Phishing_Email.csv
[14]https://www.kaggle.com/datasets/rtatman/fraudulent-email-corpus?select=fradulent_emails.txt
[15]https://www.kaggle.com/datasets/ashfakyeafi/spam-email-classification?select=email.csv
[16]https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/
[17]https://www.amnesty.org/en/latest/press-release/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/
[18]https://aws.amazon.com/bedrock
[19]https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform
[20]https://azure.microsoft.com/en-us/
[21]https://huggingface.co/
[22]https://www.anthropic.com/
[23]https://platform.openai.com/docs/overview

## 4.1 Amazon Bedrock:

Amazon Bedrock is a fully-managed service that offers access to foundational models (FMs) from AI21 Labs, Cohere, Anthropic, Mistral AI and so on. This service provides users with a wide range of FMs, enabling them to choose the model that best suits their specific use case. With Bedrock's serverless experience, users can quickly get started, customize the FMs with their own data, and easily integrate and deploy them into applications using Amazon Web Services (AWS) tools, without the need for infrastructure management. This streamlined approach accelerates the development of generative AI applications. The selected FMs are particularly well-suited for text generation tasks and can be employed as classifiers to detect fraud and abuse by leveraging publicly available datasets.

## 4.2 AI21 Labs Jurassic:

Jurassic models are provided by AI21 Labs which are suitable for various sophisticated language generation tasks such as question answering, text generation, search, and summarization. The models were launched on March 2023 and trained with data updated to mid 2022.

*4.2.1 Jurassic-2 Ultra:* Jurassic-2 Ultra is a model with 60 billion paratemeters This model has a 8,191 token context window (i.e. the length of the prompt + completion should be at most 8,192 tokens) and supports multiple languages

*4.2.2 Jurassic-2 Mid:* Jurassic-2 Mid is less powerful than Ultra with 17 billion parameters. It also supports 8,191 token context window and multiple languages.

## 4.3 Cohere:

Cohere models are text generation models provided by Cohere Inc. The size of the models is not officially disclosed and there are four models available in AWS Bedrock services.

*4.3.1 Command and Command Light:* Command models are Cohere's flagship text generation model. It is trained to follow user commands and can be used for applications like chat and text summarization. Both of the models support 4,000 token context window and only supports English language. Command Light is a smaller and faster version of command.

*4.3.2 Command R and R+:* Command R and R+ models are a generative language model optimized for long-context tasks and large scale production workloads. Both of them have 128k token context window and support multiple languages.

## 4.4 Anthropic:

We also test Anthropic's Claude family of models which are Claude 2 and Claude 2.1, both of which were launched early 2023 and are able to complete tasks like text generation, conversation, complex reasoning and analysis. Both of the models support multiple languages.

*4.4.1 Claude 2:* Claude 2 have a context window of 100k tokens and is estimated to have over 130 billion parameters. [24].

[24]https://textcortex.com/post/claude-2-parameters

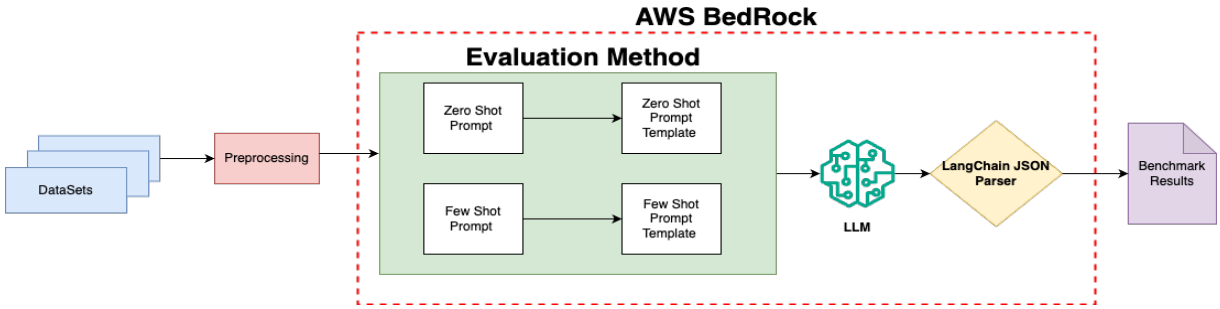


**Figure 1: The DetoxBench Pipeline**

*4.4.2 Claude 2.1.* : An update to Claude 2 that features double the context window (200k), plus improvements across reliability, hallucination rates, and evidence-based accuracy in long document and RAG contexts. [25].

### 4.5 Mistral AI:

The Mixture of Experts (MoE) models are also within the scope of our testing. Bedrock provides an MoE model of this type from Mistral AI. We will test the Mistral 8x7B, which is an MoE model combining 8 smaller models, each with 7 billion parameters. Concurrently, we will also test their flagship model, Mistral Large. Both of the have a context window of 32k and support mulitple lanuages.

*4.5.1 Mixtral 8x7B Instruct:* Mixtral 8x7B, a Sparse Mixture of Experts (SMoE) language model. Mixtral has the same architecture as Mistral 7B, with the difference that each layer is composed of 8 feedforward blocks (i.e. experts) [13].

*4.5.2 Mistral Large:* The Mistral Large is Mistral AI's flagship language model, boasting superior benchmark performance compared to the Claude 2 models. Additionally, it features a JSON format mode that ensures the model's output adheres to valid JSON syntax. This functionality is particularly beneficial for our benchmark task result collection, as we require the model to generate output in the JSON format.[26].

## 5 PROMPTING

A prompt serves as an input to large language models (LLMs) and plays a crucial role in determining the output of the model. Prompts generally consists of instructions, questions, input data and examples. The prompting techniques for LLMs refer to the strategies used to effectively instruct the model to generate desired outputs. This involves carefully crafting the prompt to elicit the desired response from the LLM and includes structuring the prompt, using specific phrasing, and incorporating relevant context. The pre-defined templates are very good starting point for prompts. These templates can be customized for different tasks and can ensure consistency in prompting. The effectiveness of these prompting techniques can vary depending on the specific task, the LLM being used, and the desired output. Experimentation and iterative refinement are often necessary to find the most effective prompting strategies for a given use case. Below are some of most popular prompting strategies used across wide range of use cases.

### 5.1 Zero-shot prompting:

Zero-shot prompting means that the prompt will not contain any examples or demonstrations. The zero-shot prompt directly instructs the model to perform a task without any additional examples to steer it. One sample prompt template used in zero-shot prompting:

> **Text:** "Please decide if the following text is hate speech or not:"
> **Output:** "She may or may not be a Jew but she 's certainly stupid , she seems to think the Blacks won't kill her."

### 5.2 Few-shot prompting:

Few-shot prompting can be used as a technique to enable in-context learning where we provide examples or demonstrations in the prompt to steer the model to better performance.

> **Text:** "Check the following job post: The title is Agile Project Manager. The location is NZ, , Wellington. The requirements is Not Mentioned. The employment type is Full-time. The industry is Not Mentioned. The function is Not Mentioned. "
> **Output:** "Fake is 1, Real is 0. The job post is 0."
> **Text:** "Check the following job post: The title is RN PACU . The location is US, GA, . The requirements is Not Mentioned. The employment type is Full-time. The industry is Hospital & Health Care. The function is Not Mentioned."
> **Output:** " Fake is 1, Real is 0. The job post is 1."
> **Text:** "Answer the following job post: The title is Credit and Collections Clerk. The location is US, WA, . The requirements is #NAME?. The employment type is Full-time. The industry is Hospital & Health Care. The function is Financial Analyst."

In this work, we experimented with zero-shot, and few-shot prompting. In our future iterations, we will include advanced prompting techniques such as Chain-Of-Thought, Tree of Thoughts, Prompt Chaining, and Self-Consistency.

[25] https://www.anthropic.com/news/claude-2-1
[26] https://mistral.ai/news/mistral-large/

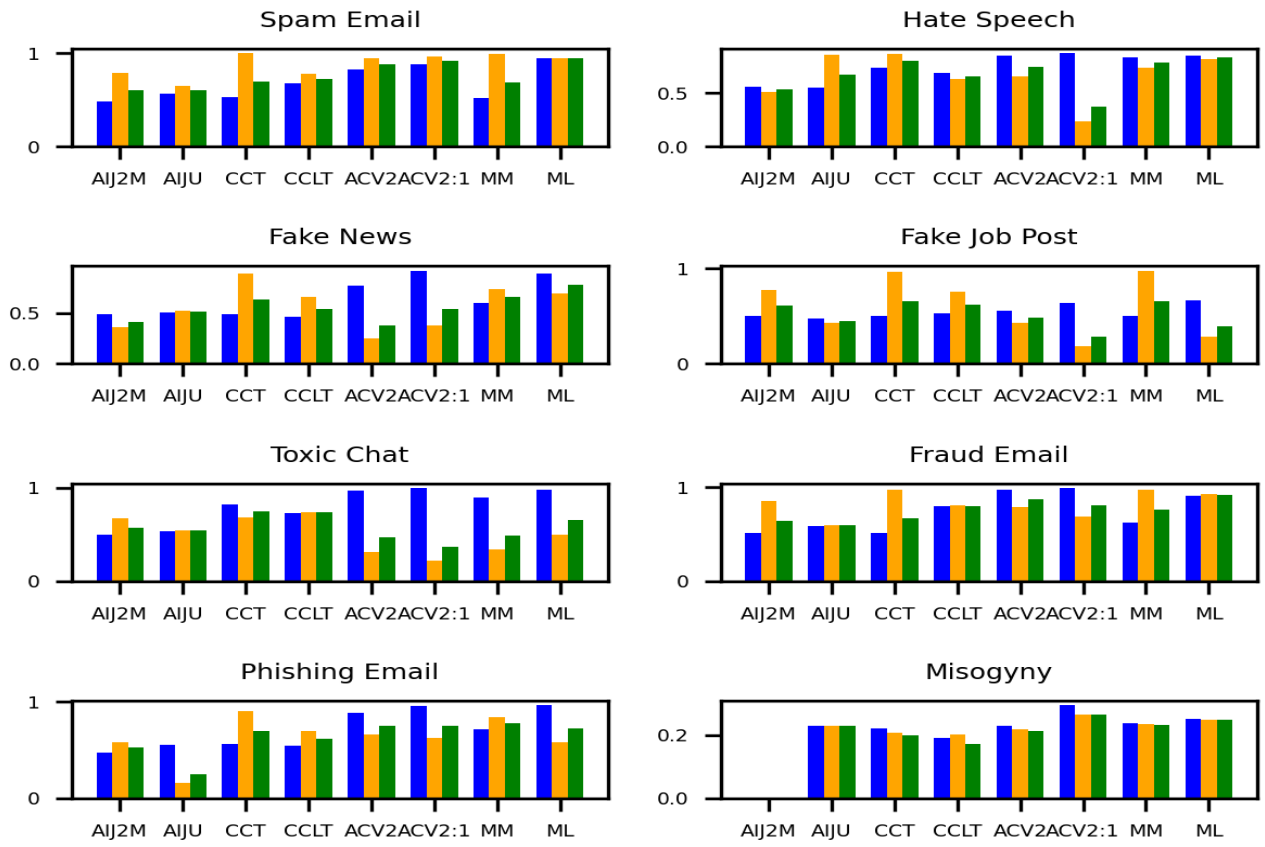


**Figure 2: Classification performance of 8 LLMs for different tasks with zero shot prompting strategy. The results are shown in the form of three metrics - Precision (Blue), Recall (Yellow), and F1 Score (Green). We referred the LLMs using these abbreviations - AIJ2M - AI21.J2-Mid; AIJU - AI21.J2-Ultra; CCT - Cohere.Command-Text; CCLT - Cohere.Command-Light-Text; ACV2: Anthropic.Claude-V2; ACV2:1 - Anthropic.Claude-V2:1; MM - Mixtral MOE; ML - Mistral Large.**

## 6 RESULTS

We have experimented with eight different large language models (LLMs) across eight different datasets for classification task. For all tasks except Misogyny, the label is binary i.e. consists two categories. Like most of the fraud datasets, our datasets are also hugely imbalanced i.e. the number of abusive cases is much lower than the number of non-abusive cases. Hence, we used random undersampling of the majority class (non-abusive) to create a balanced test set. For the Misogyny task, we transformed it into a multiple-choice problem, allowing the model to choose one correct option out of four. We tested all tasks using LLMs from the AWS Bedrock services, which provide a convenient API and a variety of LLMs for selection. To structure the output of the LLMs in a JSON format, we used LangChain parser framework[27].

We have represented the classification results in form of three widely known classification metrics - *precision*, *recall*, and *F1 score*. Figure 2 and 3 shows precision, recall, and F1 score for zero shot and few shot prompting respectively. To get a better summary, we show the F1 score for all the experiments in the form of heat map in Figure 4. Here we discuss our key observations:

- **F1 Score:** For zero shot and few shot prompting, out of eight datasets, in five cases (spam email, hate speech, fake news, fraud email, misogyny) Mistral family (ML - Mistral Large) performs the best. After Mistral family, the Anthropic Claude models achieve the second best F1 score.
- **High Precision but Low Recall:** Anthropic Claude family models are highly precise in most of the cases but the recall is significantly low. For toxic chat and hate speech detection (few-shot), the precision is over 90% but the recall is less than 10%. Though these models suffer from low recall, but they can be used for proprietary use cases where highly precise results are preferred.
- **High Recall but Low Precision:** Cohere family models provide high recall i.e. can detect most of the fraud cases. However, the precision is not that great i.e. the false positive rate is high. For fake job detection (few-shot), the recall

[27] https://python.langchain.com/v0.1/docs/modules/model_io/output_parsers/types/json/

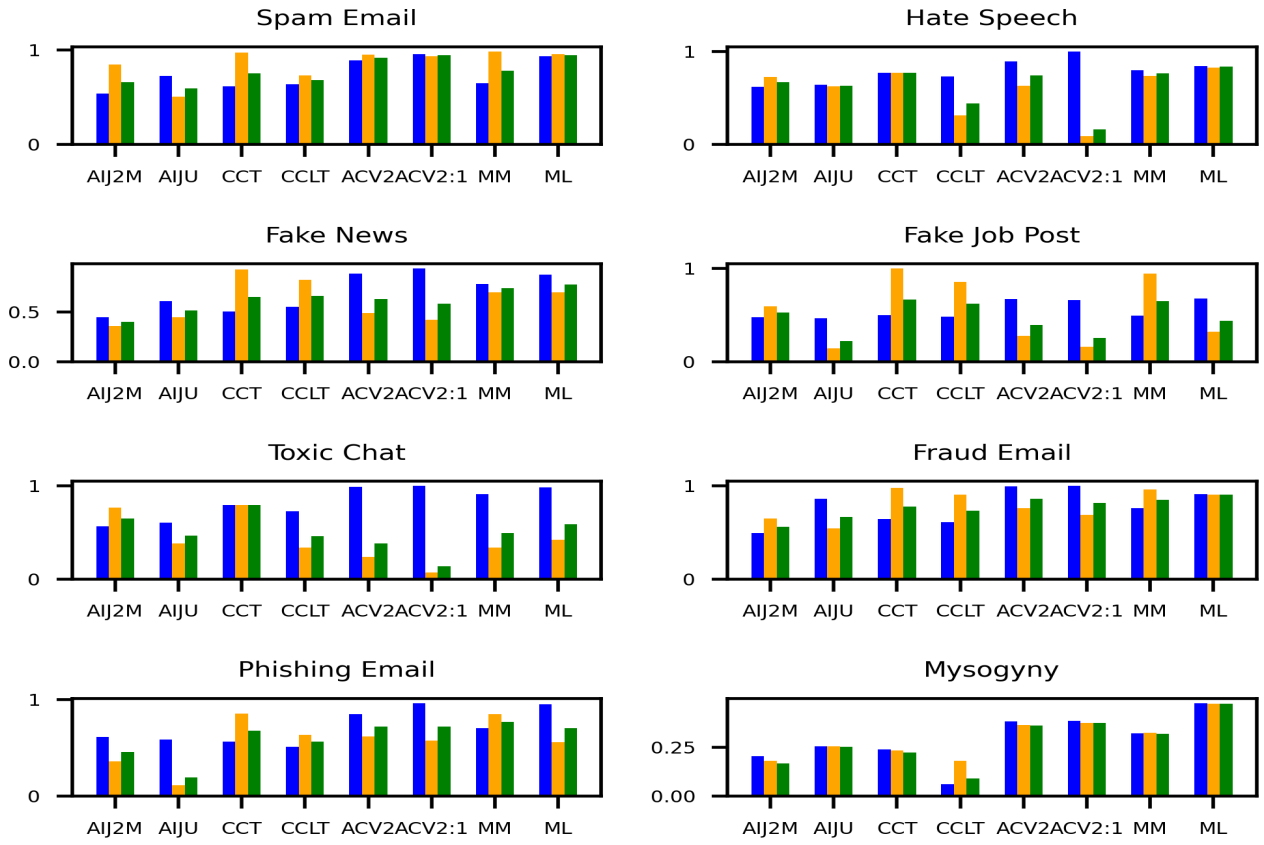


**Figure 3: Classification performance of 8 LLMs for different tasks with few shot prompting strategy. The results are shown in the form of three metrics - Precision (Blue), Recall (Yellow), and F1 Score (Green).**

is 85% and the precision is 48%. For fraud email detection (few-shot), the recall is 98% and the precision is 64%.

- **Inference Time:** AI21 family models are the fastest for inference (1.5 seconds/instance), while Mistral Large, and Anthropic Claude models are the slowest (10 seconds/instance).
- **Effect of Prompts:** For most of the cases, few-shot prompting does take more time for inference but does not improve the performance of the LLMs significantly. For example, for Mistral Large model, few-shot prompting improves performance in only two tasks - fake job detection, and misogyny detection.
- **Format Compliance:** There are some LLMs (Command R, Command R+ from Cohere family) that could not follow the LangChain JSON parser output formats. Hence, we removed them from our final benchmark results. These models are not the best choice for production use cases as they require additional post processing of outputs which might be expensive.
- **Multi-Class Classification:** In the misogyny dataset, there are five different categories. Out of five categories, four of them are different kinds of misogynistic categories (Misogynistic_personal_attack, Misogynistic_pejorative, Treatment, Derogation), and one of them is non-misogynistic category. We transformed this dataset into a multiple-choice problem, allowing the model to choose one correct option out of four. While reporting the metrics, we take the "weighted" recall, precision, and F1 across all the classes. A121 Jurassic-2 Mid (AIJ2M) model could only classify 2.1% instances and rest of them resulted as "undecided". Similar to binary classification, here also we got the best results from Mistral, and Anthropic Claude family models.

## 7 LIMITATIONS

There are several limitations of our work.

- The datasets to the best of our knowledge, are the most representative among publicly available datasets of fraud/abuse detection problems. We do not claim these datasets to be comprehensive, but hopefully with time the collection will grow to cover more business scenarios and dataset variations.
- While these datasets are useful for research and development of fraud detection algorithms, they do not carry any

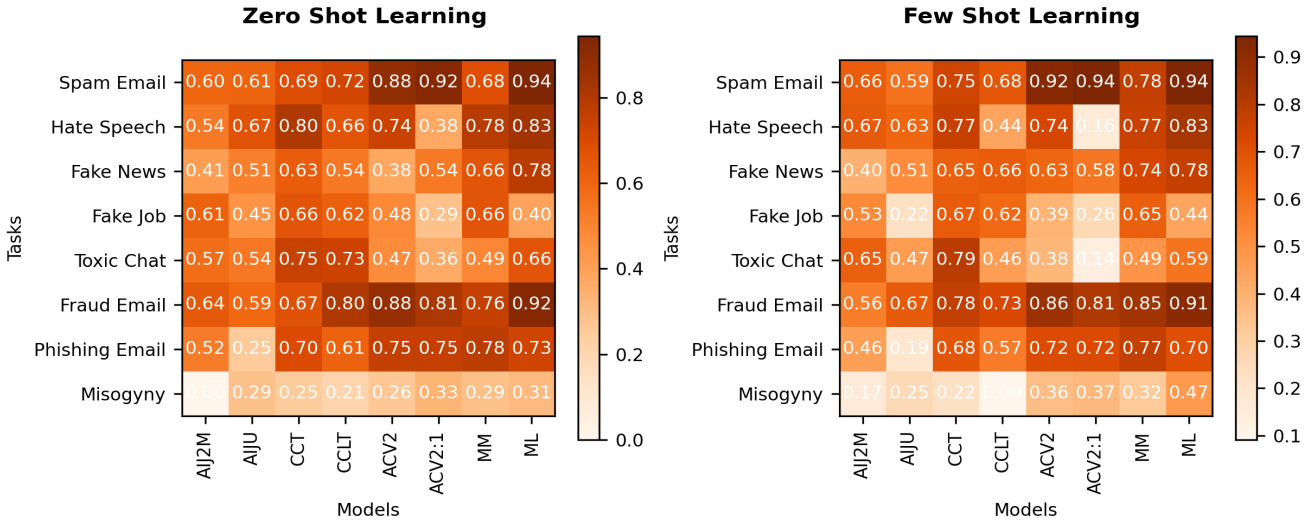


**Figure 4: F1 score of 8 LLMs for different tasks with zero shot prompting and few shot prompting strategy respectively.**

information about real fraud. If someone uses any models trained on these datasets to directly make decisions about fraud, it would cause a negative bias and could lead to false accusations.

- We started experimenting with relatively smaller size LLMs such as FLAN-T5, RoBERTa, mT0, etc. We decided to not include those models in our study as they require an additional Verbalizer[28] to assign a label (abusive/non-abusive) to every textual input based on higher likelihood. On the other hand, comparatively larger LLMs that we used were capable enough to classify most of the inputs without using Verbalizer.
- In this study, we used zero shot and few shot prompting techniques, and followed the best practices of LLM prompting as guided by huggingface. The reason behind that is we wanted to capture the strength of the models using minimal prompt intelligence. In future version, we will also experiment with other advanced prompting techniques.
- There are some advanced LLMs that we could not include in our work such as GPT family (not available on Bedrock), Llama family (Data privacy issues). The inclusion of these models could have given different set of results.
- This work only focuses on English language texts. We decided to use English only for two reasons. First, as most of the advanced LLMs support English and majority of prior benchmark works also focused on English only. Second, there are very few non-English textual datasets related to fraud and abuse domain that are publicly available .

[28] https://thunlp.github.io/OpenPrompt/modules/verbalizer.html

## 8 DATASET DISCLAIMERS AND TERMS

Firstly, the datasets used in our work were collected from the online data sources and are completely anonymous. We have carefully gone through the data and taken out anything that could have personal information in it. However, there is still a chance that some personal information might be left in the data. If anyone come across anything in the data that should not be made public please inform the authors. Secondly, these datasets may contain racism, sexuality, or other undesired content. The statements or opinions made in this dataset do not reflect the views of researchers or institutions involved in the data collection effort. Thirdly, the users of these datasets are responsible for ensuring its appropriate use and should not be utilized for training dialogue agents, or any other applications, in manners that conflict with legal and ethical standards. Finally, the users of these datasets must not attempt to determine the identity of individuals in this dataset.

## 9 CONCLUSION & FUTURE WORK

In this paper, we systematically evaluate eight leading LLMs on eight different fraud and abuse categories. To the best of our knowledge, this is the first LLM benchmark specifically focused on fraud and abuse detection tasks. Based on our experiments, we make some key observations such as: a) We find larger size leads to better performance (200 Billion Anthropic family, and 176 Billion Mistral AI family are performing the best). b) Some LLMs (Mistral Large, Anthropic Claude) are much better than other LLMs (AI21, Cohere) to understand the contextual meaning for fraud and abuse classifications. c) Few-shot prompting does not always improve results over zero-shot prompting used in our experiments. For future work, we plan to experiment with fine-tuning LLM models to enhance their ability in handling fraud and abuse tasks. Additionally, in this paper, we utilized a simple LLM chain without a memory buffer.

In future work, we will explore different chain structures, such as applying the Sequential Chain along with Chain-of-Thought (COT) prompts. We will also experiment with the Router Chain to process mixed types of fraud and abuse classifications.

## REFERENCES


[1] belongto.org. 2023. 87% of LGBTQ+ youth report hate and harassment online. https://www.belongto.org/87-of-lgbtq-youth-report-hate-and-harassment-online/

[2] Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, David Molnar, Spencer Whitman, and Joshua Saxe. 2024. CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models. arXiv:2404.13161 [cs.CR]

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[4] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132 [cs.AI]

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416 [cs.LG]

[6] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv:1901.02860 [cs.LG]

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[8] International Woemn's Media Foundation. 2024. A Mental Health Guide for Journalists Facing Online Violence. https://www.iwmf.org/mental-health-guide/

[9] Michael Greenwood. 2023. The Impact of Cyberbullying on Mental Health. https://www.news-medical.net/health/The-Impact-of-Cyberbullying-on-Mental-Health.aspx

[10] Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 1336–1350. https://doi.org/10.18653/v1/2021.eacl-main.114

[11] Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An Investigation of Large Language Models for Real-World Hate Speech Detection. arXiv:2401.03346 [cs.CY]

[12] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY]

[13] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG]

[14] Liming Jiang. 2024. Detecting Scams Using Large Language Models. arXiv:2402.03147 [cs.CR]

[15] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. arXiv:1909.05858 [cs.CL]

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv:1909.11942 [cs.CL]

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs.CL]

[18] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL]

[19] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958 [cs.CL]

[20] Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation. arXiv:2310.17389 [cs.CL]

[21] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. arXiv:2001.08210 [cs.CL]

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[23] Karla Mantilla. 2013. Gendertrolling: Misogyny Adapts to New Media. *Feminist Studies* 39, 2 (2013), 563–570. https://doi.org/10.1353/fem.2013.0039

[24] Jay Mayfield. 2024. New FTC Data Shed Light on Companies Most Frequently Impersonated by Scammers. https://www.ftc.gov/news-events/news/press-releases/2024/05/new-ftc-data-shed-light-companies-most-frequently-impersonated-scammers

[25] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. arXiv:2402.06196 [cs.CL]

[26] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. arXiv:1910.14599 [cs.CL]

[27] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How Context Affects Language Models' Factual Predictions. In *Automated Knowledge Base Construction*. https://openreview.net/forum?id=025X0zPfn

[28] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[29] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. https://api.semanticscholar.org/CorpusID:49313245

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. https://api.semanticscholar.org/CorpusID:160025533

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[32] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]

[33] Kylee Sibilia. 2024. Financial fraud growth slowed in 2023. https://www.prnewswire.com/news-releases/financial-fraud-growth-slowed-in-2023-but-losses-remained-high-alloy-report-finds-25-of-companies-lost-over-1m-to-fraud-302044017.html

[34] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]
[36] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. arXiv:1905.00537 [cs.CL]
[37] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv:1804.07461 [cs.CL]
[38] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2020. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. arXiv:1906.06725 [cs.CL]
[39] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652 [cs.CL]

# 10 APPENDIX

## 10.1 LLM Architecture

Language modeling is a long-standing research topic, dating back to the 1950s with Shannon's application of information theory to human language, where he measured how well simple n-gram language models predict or compress natural language text. Since then, statistical language modeling became fundamental to many natural language understanding and generation tasks, ranging from speech recognition, machine translation, to information retrieval. The recent advances on large language models (LLMs), pretrained on Web-scale text corpora, significantly extended the capabilities of language models [25]. LLMs are transformer based neural language models that contain tens to hundreds of billions of parameters which are pretrained on massive training data. The most widely used LLM architectures are encoder only[7], decoder only[6] and encoder-decoder[17].

*10.1.1* ***The Transformer - model architecture:*** The transformer architecture was originally designed for sequence transduction or neural machine translation to convert an input sequence to an output sequence. It is a simple network architecture based solely on attention mechanisms, dispensing with recurrence and convolutions entirely[35]. Transformer architecture consists of seven key components. A demonstration of each of the components is shown below.

- **Input Embedding:** The ML models use user-entered tokens as training data, while it can only process numeric information. Thus, it is necessary to transform these textual inputs into a numerical format known as "input embeddings". These embeddings function similarly to a dictionary, assisting the model in understanding the meaning of words by arranging them in a mathematical space where comparable phrases are situated close together.
- **Positional Embedding**: The order of words in a sentence is essential in the NLP field for identifying the statement's meaning. The positional encoding is utilized to encode each word's location in the input sequence as a collection of integers which allows the model to grasp sentence word order better and provide grammatically accurate and semantically relevant output. The original transformer architecture uses sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = sin(pos/10000^{(2i/d_{model})}) \quad (1)$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{(2i/d_{model})}) \quad (2)$$

- **Encoder:** The encoder is composed of a stack of N = 6 identical layers. Each layer has two sub-layers.The first is a multi-headed self-attention layer and the second is a simple position-wise fully connected feed-forward network. It processes the input text and generates a series of hidden states. This consists of two linear transformations with a ReLU activation in between.

$$FFN(x) = max(0, xW_1 + b1)W_2 + b2 \quad (3)$$

- **Outputs (shifted right):** During the training process, the decoder acquires the ability to predict the next word by analyzing the previous words. In this case, the output sequence is shifted by one position to the right. Consequently, the decoder is able to use the words that came before it.
- **Output Embedding:** The output is converted to a format known as "output embedding." Like input embeddings, output embeddings also undergo positional encoding, enabling the model to understand the order of words in a sentence.
- **Decoder:** The decoder is composed of a stack of N = 6 identical transformer layers. In addition to the two sub-layers in each encoder layer, the decoder has a third sub-layer, which performs multi-head attention over the output of the encoder stack. It creates output sequences from positionally encoded input sequences while learns to predict the next word from previous words in positionally encoded output embedding during the training period.
- **Linear Layer and Softmax:** The linear layer maps to the higher-dimensional space once the decoder has generated the output embedding. This step is required to convert the output embedding into the original input space. The softmax function generates a probability distribution for each output token in the developed vocabulary, allowing us to generate probabilistic output tokens.

*10.1.2* ***Transformer: Encoder.*** Encoder models use only the encoder of a Transformer model [29]. At each stage, the attention layers can access all the words in the initial sentence. These models are often characterized as having "bi-directional" attention, and are often called auto-encoding models. The pretraining of these models usually revolves around somehow corrupting a given sentence (for instance, by masking random words in it) and tasking the model with finding or reconstructing the initial sentence. Encoder models are best suited for tasks requiring an understanding of the full sentence, such as sentence classification, named entity recognition (and more generally word classification), and extractive question answering. The representatives of this family of include: ALBERT[16], BERT[7], DistilBERT[32], ELECTRA [30], RoBERTa[22].

*10.1.3* ***Transformer: Decoder.*** Decoder models use only the decoder of a Transformer model[31]. At each stage, for a given word the attention layers can only access the words positioned before it in the sentence. These models are often called auto-regressive models. The pretraining of decoder models usually revolves around predicting the next word in the sentence. These models are best

[29] https://huggingface.co/learn/nlp-course/en/chapter1/5
[30] https://openreview.net/pdf?id=r1xMH1BtvB
[31] https://huggingface.co/learn/nlp-course/en/chapter1/6

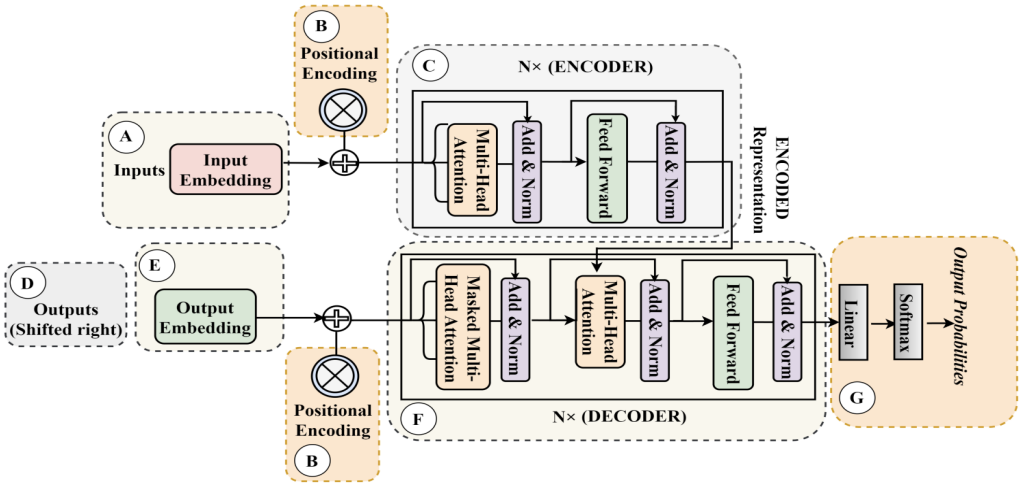


**Figure 5: The Transformer - model architecture [35]**

**Table 2: F1 score of 8 LLMs with 95% confidence interval for different tasks with zero shot prompting.**

| | AIJ2M | AIJU | CCT | CCLT | ACV2 | ACV2:1 | MM | ML |
|---|---|---|---|---|---|---|---|---|
| **Spam Email** | 0.60 ± 0.03 | 0.61 ± 0.03 | 0.69 ± 0.03 | 0.72 ± 0.05 | 0.88 ± 0.02 | 0.92 ± 0.02 | 0.68 ± 0.03 | 0.94 ± 0.01 |
| **Hate Speech** | 0.54 ± 0.04 | 0.67 ± 0.03 | 0.80 ± 0.03 | 0.66 ± 0.05 | 0.74 ± 0.03 | 0.38 ± 0.05 | 0.78 ± 0.03 | 0.83 ± 0.02 |
| **Fake News** | 0.41 ± 0.05 | 0.51 ± 0.04 | 0.63 ± 0.03 | 0.54 ± 0.05 | 0.38 ± 0.05 | 0.54 ± 0.04 | 0.66 ± 0.03 | 0.78 ± 0.03 |
| **Fake Job** | 0.61 ± 0.04 | 0.45 ± 0.04 | 0.66 ± 0.03 | 0.62 ± 0.05 | 0.48 ± 0.04 | 0.29 ± 0.05 | 0.66 ± 0.03 | 0.40 ± 0.05 |
| **Toxic Chat** | 0.57 ± 0.04 | 0.54 ± 0.04 | 0.75 ± 0.03 | 0.73 ± 0.04 | 0.47 ± 0.05 | 0.36 ± 0.08 | 0.49 ± 0.05 | 0.66 ± 0.04 |
| **Fraud Email** | 0.64 ± 0.04 | 0.59 ± 0.04 | 0.67 ± 0.03 | 0.80 ± 0.03 | 0.88 ± 0.02 | 0.81 ± 0.03 | 0.76 ± 0.03 | 0.92 ± 0.03 |
| **Phishing Email** | 0.52 ± 0.05 | 0.25 ± 0.05 | 0.70 ± 0.03 | 0.61 ± 0.05 | 0.75 ± 0.03 | 0.75 ± 0.03 | 0.78 ± 0.03 | 0.73 ± 0.03 |

**Table 3: F1 score of 8 LLMs with 95% confidence interval for different tasks with few shot prompting.**

| | AIJ2M | AIJU | CCT | CCLT | ACV2 | ACV2:1 | MM | ML |
|---|---|---|---|---|---|---|---|---|
| **Spam Email** | 0.66 ± 0.03 | 0.59 ± 0.03 | 0.75 ± 0.03 | 0.68 ± 0.05 | 0.92 ± 0.02 | 0.94 ± 0.02 | 0.78 ± 0.03 | 0.94 ± 0.01 |
| **Hate Speech** | 0.67 ± 0.04 | 0.63 ± 0.03 | 0.77 ± 0.03 | 0.44 ± 0.05 | 0.74 ± 0.03 | 0.16 ± 0.05 | 0.77 ± 0.03 | 0.83 ± 0.02 |
| **Fake News** | 0.40 ± 0.04 | 0.51 ± 0.04 | 0.65 ± 0.03 | 0.66 ± 0.03 | 0.63 ± 0.04 | 0.58 ± 0.04 | 0.74 ± 0.03 | 0.78 ± 0.03 |
| **Fake Job** | 0.53 ± 0.04 | 0.22 ± 0.04 | 0.67 ± 0.03 | 0.62 ± 0.03 | 0.39 ± 0.05 | 0.26 ± 0.05 | 0.65 ± 0.03 | 0.44 ± 0.04 |
| **Toxic Chat** | 0.65 ± 0.04 | 0.47 ± 0.04 | 0.79 ± 0.03 | 0.46 ± 0.04 | 0.38 ± 0.05 | 0.14 ± 0.08 | 0.49 ± 0.05 | 0.59 ± 0.04 |
| **Fraud Email** | 0.56 ± 0.05 | 0.67 ± 0.04 | 0.78 ± 0.03 | 0.73 ± 0.03 | 0.86 ± 0.02 | 0.81 ± 0.03 | 0.85 ± 0.02 | 0.91 ± 0.02 |
| **Phishing Email** | 0.46 ± 0.05 | 0.19 ± 0.05 | 0.68 ± 0.03 | 0.57 ± 0.05 | 0.72 ± 0.03 | 0.72 ± 0.03 | 0.77 ± 0.03 | 0.70 ± 0.03 |

suited for tasks involving text generation. The representatives of this family of include: CTRL[15], GPT [29], GPT-2 [30], Transformer XL[6], Llama 2[34].

*10.1.4* ***Transfomer: Encoder-Decoder:*** Encoder-decoder models (also called sequence-to-sequence models) use both parts of the Transformer architecture[32]. At each stage, the attention layers of the encoder can access all the words in the initial sentence, whereas the attention layers of the decoder only access the words positioned before a given word in the input. The pretraining of these models can be done using the objectives of encoder or decoder models, but usually involves something a bit more complex. For instance, some models are pretrained by replacing random spans of text that can contain several words with a single mask special word, and the objective is then to predict the text that this mask word replaces.

[32]https://huggingface.co/learn/nlp-course/en/chapter1/7?fw=pt

Sequence-to-sequence models are best suited for tasks revolving around generating new sentences depending on a given input, such as summarization, translation, or generative question answering. The representatives of this family of models include: BART[17], mBART[21], Marian[33], T5[31]

*10.1.5* ***Instruction Fine-tuned Models:*** Instruction tuning is a simple method that combines appealing aspects of both the pre-train–finetune and prompting paradigms by using supervision via finetuning to improve the ability of language models to respond to inference-time text interactions. The supervision teaches the model to perform tasks described via instructions. Recent empirical results[39]demonstrate promising abilities of language models to perform tasks described purely via instructions. Finetuning on groups of language tasks has been shown to significantly boost this zero-shot task generalization of language models [5, 39].

## 10.2 LangChain Framework

In this paper, we leverage LangChain for in-context learning. LangChain is a robust Python library designed to simplify interactions with various LLM providers. Chains are a vital component of LangChain, allowing multiple elements to seamlessly integrate. In this work, we specifically used the LLMChain[34].

## 10.3 Additional Results

We have reported precision, recall, and F1 score for various experiments in the Results section above. In Table 2, and Table 3, we are showing the 95% confidence interval along with the F1 score for different tasks and different models. The key observation here is the variance is low that means LLMs are stable in abuse detection.

[33]https://huggingface.co/docs/transformers/model_doc/marian
[34]https://api.python.langchain.com/en/latest/chains/langchain.chains.llm.LLMChain.html