Evaluating open- and closed-source LLM-based chatbots to combat cybercrime targeting senior citizens

Anonymous ACL submission

Abstract

This study evaluates state-of-the-art open-source and closed-source models that we trained on a cybersecurity Q&A task to assist senior citizens in recognizing and responding to cybercrimes. We evaluate five LLMs and their finetuned variants using automatic evaluation metrics such as F1, BertScore, n-gram based overlap, and human evaluation of response quality through seven criteria, including accuracy, relevance, and usefulness. Our evaluation results show that several open-source models, particularly fine-tuned variants, outperform the closed-source model, with *Mistral3-LoRA* leading on nearly all automatic evaluation metrics and *LLaMA3.1-LoRA* achieving the highest recall. However, *ChatGPT-4o* slightly outperforms in the human evaluation task, with annotators preferring its responses for their formatting and polished language. Our chatbot application, code, and data is available at https://anonymous.4open.science/r/SeniorSafeAI-36F4/.

1 Introduction

Annually, over 100,000 senior fraud complaints result in total losses exceeding $3 billion (Internet Crime Complaint Center, 2023). In 2022, approximately 6,000 senior citizens reported losses of more than $100,000 (Internet Crime Complaint Center, 2023). The total losses for the over-60 age group are five times higher than those for the 20-29 age group. Implementing best-practice policies to fight senior cybercrime is a complex challenge. One major barrier is the lack of actionable information, as many victims feel embarrassed or unsure about whom to contact. Law enforcement also struggels with limited data on hacker methods and victim profiles. Meanwhile, cybersecurity education has not kept pace with rapidly evolving offender tactics and AI-assisted social engineering. As a result, effective resolutions to senior cybercrime remain inadequately measured and responded. AI chatbots using Large Language Models (LLMs) can address these issues by providing victims with a stigma-free reporting site and up-to-date advice based on best practices. However, there are two major challenges in developing AI chatbots for such critical tasks.

First, the increasing reliance on closed-source (often proprietary) LLMs raises significant concerns about transparency, security, and privacy. Users have consistently expressed concerns about data privacy, bias, and ethical implications (Fondrie-Teitler and Jayanti, 2023). Yet, the opaque nature of closed-source models makes effective auditing difficult. In academic contexts, growing reliance on closed-models contributes a "Reproducibility Crisis", as researchers lack control over model behaviors (i.e., frequent model updates) (Ollion et al., 2024). These concerns have renewed interest in open-source LLMs, which offer a more transparent and collaborative alternative. Open-source LLMs are "freely accessible, open for modification and distribution", allowing for public scrutiny, knowledge sharing, and fostering innovation (Kukreja et al., 2024). Although LLMs developed by large companies often achieve high average performance, studies indicate that open-source models can perform nearly as well as closed-source models in tasks such as clinical Q&As (Adams et al., 2024), summarization of medical dialogues and climate fact checking (Wolfe et al., 2024).

Second, another remaining challenge is the issue of evaluation. The evaluation of chatbots measures real-world performance in achieving the defined purpose and further ensuring safety and trustworthiness by demonstrating reliability and generalizability. However, agreed-upon and robust evaluations are lacking (Abeysinghe and Circi, 2024). On the one hand, automatic evaluation evaluation is popular, because of easy implementation and repeatability. However, the caveat is that they may

not always be correlated with human evaluation. On the other hand, human evaluation is widely accepted, but it does not consistently agree with automatic evaluation evaluation results, and is often limitted to small subset of human annotators leading to underpowered results (Clark et al., 2021).

Motivated by these challenges, we formulate the following two research questions:

RQ1: How does the performance of open-source and closed-source LLMs compare across automatic and human evaluation metrics?

RQ2: How does the performance of LLMs compare to human ground truth responses in terms of automatic and human evaluation metrics?

We contribute the following to the EMNLP community in the following three ways: First, we introduce a working chatbot prototype trained on 589 original Q&A pairs developed by individuals who were trained on how to find evidence-based information on cybersecurity and victimization prevention practices. The dataset covers major crime types affecting older adults, including identity theft, romance scams, credit card fraud, investment fraud, fake tech support, and best practices for cyber hygiene. Second, we conduct a mixed-methods evaluation framework that combines lexical, semantic, and entity-level evaluation metrics with human annotator ratings to assess LLM performance in cybercrime content. Third, we empirically show that fine-tuning affects open-source models differently: for *Mistral3* and *LLaMA3.1*, low-rank adaptation improves performance, with higher precision and recall compared to closed-source models like *ChatGPT-4o*, while other models show limited improvement.

2 Related Work

2.1 LLM-based solutions for mitigating cybercrimes

Recent empirical research reveals that when applied to cybersecurity, LLMs exhibit both substantial promise and critical vulnerabilities. These vulnerabilities fall into two broad categories: technical weaknesses and ethical risks. While LLMs can assist in tasks such as cyber threat intelligence and secure code generation, their performance remains highly context-dependent and demands continuous human oversight to ensure accuracy (Clairoux-Trepanier et al., 2024; Charan et al., 2023). Existing operational assessments echo this pattern: Models often generate plausible but misleading outputs when processing real-world cybercrime data, especially without adversarial defenses or expert supervision (Clairoux-Trepanier et al., 2024; Islam and Sandborn, 2023).

The literature reaches a clear consensus: current LLM architectures are not resilient enough, technically or ethically, for safe integration into cybersecurity workflows. They perform reliably only in controlled conditions and degrade rapidly under adversarial pressure. In particular, general-purpose models underperform in specialized contexts without fine-tuning, and their decision-making remains unstable without supplemental verification mechanisms (Motlagh et al., 2024; Islam and Sandborn, 2023). Even promising results in related fields, such as software system reliability, confirm the same pattern: LLMs assist, but cannot yet be trusted to operate autonomously. Beyond technical failure, ethical risks emerge as equally serious. Empirical evaluations show that LLMs can propagate biased, harmful, or deceptive content, particularly when exposed to ambiguous, manipulated, or hostile environments (Liu et al., 2025; Xu et al., 2024). Based on these findings, we anticipate that finetuning and domain adaptation, features more readily available to open-source than closed-source models, are needed before they can be integrated into existing cybersecurity workflows.

2.2 Evaluating LLMs

The major challenge of LLM evaluation is that the tasks go beyond choosing from predefined categories. Instead, it involves interpreting the generated language, by assessing how the generated language is "coherent, relevant, and contextually accurate" (Iusztin and Labonne, 2024). Evaluation of LLMs heavily depends on, often only, automatic metrics (Van der Lee et al., 2021). Automatic evaluation is popular, easy, and replicable since it measures diverse attributes in texts quantitatively. Based on the chatbot's intended purpose, automatic LLM evaluation can adopt domain- or task- specific evaluation approach (Abeysinghe and Circi, 2024). *Domain-specific LLM evaluations* use previous benchmarks to evaluate a model performance. These domain specific benchmarks are designed to be reproducible, by capturing domain specific performance more accurately. Open Medical-LLM Leaderboard and Hallucination Leaderboard are some examples of domain specific LLM evaluation. Domain-specific evaluations, however, requires well-curated and tested benchmarks, and

cannot guide a given specific tasks. *Task-specific LLM evaluation* use metrics to evaluate narrow-focused tasks, bypassing pre-existing evaluation datasets. ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and classic metrics such as accuracy, precision, recall, and F1 score are good examples of task-specific LLM evaluation. However, these metrics are criticized because they tend to be underinformative, too simplistic, and do not correlate well with human evaluations (Van der Lee et al., 2021).

Human evaluations are considered as "gold standard" because human-written texts include diverse aspects that cannot easily be encoded computationally (Clark et al., 2021). Human evaluators can assess a text based on given dimensions such as text quality, naturalness, or humanlikeness (Van der Lee et al., 2021). However, due to its costly nature, human evaluations can only use limited number of evaluators, leading to underpowered results (Abeysinghe and Circi, 2024; Van der Lee et al., 2021). In addition, human evaluators are influenced by framing of questions (Schoch et al., 2020), and level of training for the assessment (Clark et al., 2021). To leverage the strengths of both automatic and human evaluation approaches, researchers advocate for a "mixed-methods" approach to evaluate LLMs (Abeysinghe and Circi, 2024).

3 Method: SeniorSafeAI

SeniorSafeAI's architecture (Figure 1) integrates a FastAPI backend with Uvicorn implementation for asynchronous request handling, a ReactJS frontend for user interaction, and LangChain to structure conversational flows for our fine-tuned LLMs. We utilized Firebase's Cloud Firestore for real-time database management, which includes user authentication for login and account creation, as well as storing chatlogs for each user account. This allows users to securely access their own conversation history, and manage their chat sessions (as shown in Figure 2). Our chatbot application, code, and data is available at anonymous GitHub repo, `https://anonymous.4open.science/r/SeniorSafeAI-36F4/`.

3.1 Training data

First, our training data included 589 question-answer (Q&A) pairs (the "ground truth") that comprehensively cover key cybercrime types and topics relevant to older adults. To our knowledge, there is

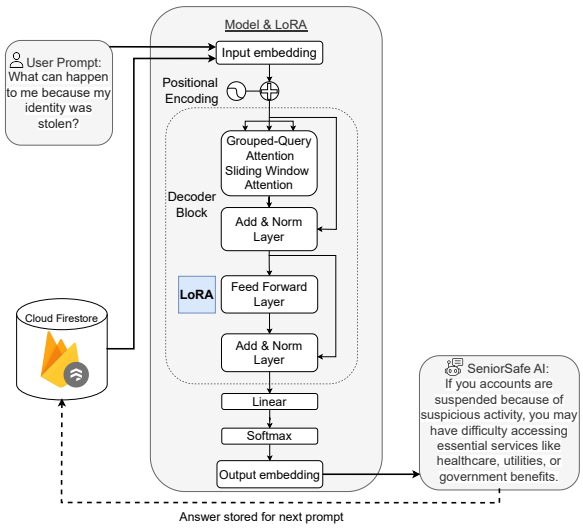


Figure 1: SeniorSafeAI chatbot workflow with LoRA.

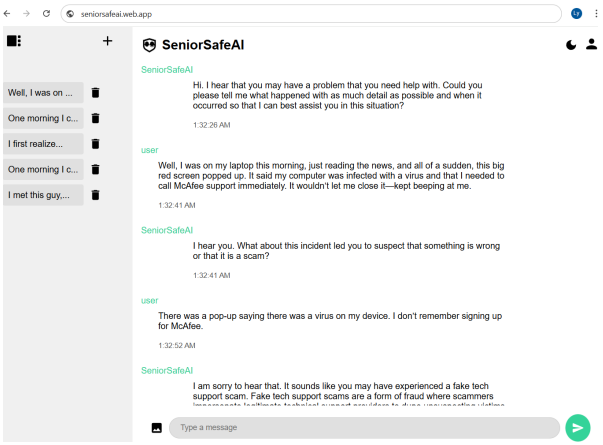


Figure 2: SeniorSafeAI user interface.

currently no gold standard Q&A dataset for cybersecurity guidance for seniors. The closest dataset to ours is CyberQ (Agrawal et al., 2024), but the task is different from ours as the Q&As contain general cybersecurity education for students and novice learners, not age-specific, scenario-based guidance for seniors.

Our Q&As were distributed across seven major categories: general questions (70 Q&A pairs), cyber hygiene (52 pairs), investment fraud (93 pairs), fake tech support (96 pairs), credit card fraud (123 pairs), identity theft (101 pairs), and romance scams (54 pairs). Each Q&A was developed through iterative co-design sessions with cybersecurity researchers and student coders, guided by domain experts to reflect realistic and high-risk scenarios encountered by seniors. This dataset is grounded in up-to-date law enforcement guidelines (e.g., Consumer Financial Protection Bureau, Inter-

net Crime Complaint Center (IC3), Federal Trade Commission) and cybersecurity best practices (e.g., password hygiene, fraud reporting protocols). We also ensured that the factual accuracy of each answer was supported by at least one verified document, totaling 543 unique support PDFs. These documents included government advisories, consumer protection materials, and official reporting instructions.

For model training and evaluation, the dataset was split into 80% for training and 20% for held-out testing. The training data was used to fine-tune our LLM-based response system, and the test set was for benchmarking across accuracy, entity precision, and language similarity against model-generated outputs. The structure of training data followed Alpaca prompt template (`User: %instruction%: %input%. Assistant: %output%`). The `instruction` specifies the crime type (i.e. Identity Theft; Romance Scams; Credit Card Fraud; Investment Fraud; Fake Tech Support; Cyber Hygiene; General Questions) and the question category (e.g., Reporting and Taking Action; Financial and Legal Concerns; Prevention and Moving Forward). An example is shown below:

```
{
  "instruction": "You are an expert on cybercrime and senior
    citizens' safety, designed to answer questions on
    Romance Scams in terms of Reporting and Taking Action,
    including how it impacts seniors and practical steps
    they can take to protect themselves.",

  "input": "Should I report this to local law enforcement
    or a cybercrime unit?",

  "output": "Yes, the most common law enforcement agency
    is the FBI and you may file a report with their
    Internet Crime Complaint Center (IC3.gov)."
}
```

Figure 3: Example training instance for the Romance Scam crime type with $instruction, input, output$.

3.2 LLMs and specifications

We selected five state-of-the-art models for training and evaluation: LLaMA3.1-8B, Gemma-7B, Qwen2.5-7B, and Mistral3v0.3-7B, all open-source, and ChatGPT-4o, a closed-source model. This selection allowed us to compare how well open-source models performed against closed-source models. If their performance was comparable, we prioritized open-source models for their transparency in implementation and design, as well as the flexibility for us to finetune and maintain complete control over data processing and storage. With the open-source models, we fine-tuned them using Low-Rank Adaptation (LoRA) on the train-test Q&A dataset so the models can learn from domain-specific knowledge. For all models, we adopted the same system prompt to establish the chatbot's role. Each model was provided with the context: *"You are an expert on cybercrime and senior citizens' safety, designed to answer all questions truthfully, including how it impacts seniors and practical steps they can take to protect themselves"*.

With the combination of base models and their LoRA-variants (with the exception of ChatGPT-4o as we cannot finetune the proprietary model), we evaluated a total of nine models. Each LoRA-tuned model was trained with the following hyperparameters: LoRA Rank = 128, Alpha = 256, Dropout = 0.05, Batch Size = 128, Cutoff Length = 256, Learning Rate = 3e-5, Repetition Penalty = 1.11, Temperature = 0.1, and min_p = 0.05. We trained each model for five epochs. For instance, the Mistral-7B-v0.3 LoRA model showed steady performance improvements across epochs, with evaluation loss dropping from 1.86 (epoch 1) to 1.29 (epoch 3). The final training loss was 50.73 after 479 steps ($\approx$ 3.8 epochs), with training runtime of 298 seconds and an average throughput of 7.9 samples per second. LoRA training and evaluation were conducted using a high-performance workstation equipped with a single NVIDIA RTX 6000 Ada Generation GPU (48 GB GDDR6) and a Dell Precision 5860 Tower featuring an Intel® Xeon® W5-2545 processor (12 cores, 24 threads). On this setup, training took $\approx$ 3.96 minutes for Mistral-7B-v0.3, 4.71 minutes for Qwen2.5-7B-Instruct, 5.13 minutes for LLaMA3.1-8B, and 5.65 minutes for Gemma-7B, based on end-to-end training runtimes tracked using Weights & Biases (WandB).

3.3 Evaluation

We used a mixed-methods approach to learn about users' preferences as well as to compare the performance of different models. The initial evaluation set includes ten ground truth Q&As which are unseen in the training data (conversational snippets in Appendix A1-A3). This initial set is used to establish agreement between the coders as well as to refine our evaluation criteria. Once finalized, we apply these criteria to a held-out test set of 118 Q&As for full evaluation. The details of our automatic evaluation metrics and human evaluation metrics are presented in the following subsections.

3.3.1 Automatic evaluation metrics

We used automatic evaluation metrics to compare chatbot responses with ground truth text involve both lexical similarity (word match) using entity precision, recall, ROUGE-L, BLEU, and semantic (meanings of words) similarity using BertScore F1, and log-odds with a Dirichlet prior.

Token-level Precision and Recall measure the overlap between ground truth and generated answers after lowercasing and tokenizing. Given token sets T_{gt} and T_{gen}:

$$\begin{aligned} \text{Precision} &= \frac{|T_{\text{gt}} \cap T_{\text{gen}}|}{|T_{\text{gen}}|} \\ \text{Recall} &= \frac{|T_{\text{gt}} \cap T_{\text{gen}}|}{|T_{\text{gt}}|} \end{aligned} \tag{1}$$

Entity Precision and Recall extend token-level metrics to named entities and noun phrases, extracted using spaCy's pipeline and pattern-based matching (Honnibal and Montani, 2017).

ROUGE-L retrieves the longest common subsequence (LCS) between model and ground truth responses, where order of the words matter. The F1 version used in our script combines both LCS-based recall and precision:

$$\text{ROUGE-L} = \frac{2 \cdot PLCS \cdot R_{LCS}}{P_{LCS} + R_{LCS}} \tag{2}$$

where P_{LCS} and R_{LCS} denote LCS precision and recall relative to generated and ground truth.

BLEU evaluates n-gram precision with smoothing (method 1 in NLTK). While BLEU typically uses 4-gram precision, we adapt it for sentence-level comparison:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}$$

where BP is the brevity penalty, and p_n is the clipped n-gram precision, which means we count each n-gram in the generated response only up to the number of times it appears in the ground truth.

String Similarity is measured using normalized Levenshtein distance. This metric measures how similar two texts are by counting the fewest character changes (insertions, deletions, or substitutions) needed to make them match.

Length Ratio is computed to check verbosity of the generated text in comparison to the ground truth length. Measured as $\frac{\text{len(gen)}}{\text{len(gt)}}$.

For further lexical and semantic comparison between the ground truth and models' responses, we used Monroe et al. (2008)'s approach to calculate log-odds ratio with a Dirichlet prior, which identifies words that are significantly overrepresented in one corpus compared to another. This approach has been validated in recent works on domain-specific social media text [citations redacted for review]. Positive log-odds values (z-scores) indicate words more prevalent in the ground truth answers, whereas negative values indicate words more prevalent in the model-generated responses.

BERTScore uses contextual embeddings to evaluate semantic similarity. It computes cosine similarity between each token in the generated and ground truth answers using pre-trained RoBERTa-large model. The final values are normalized via baseline rescaling, consistent with (Zhang et al., 2019)'s best practices. We calculate the F1 value:

$$F_{\text{BERT}} = \frac{2 \cdot P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{4}$$

3.3.2 Human evaluation metrics

For human evaluation, we asked annotators to rate model responses based on seven criteria: Clarity, accuracy, relevance, error prevention and education, consistency, usefulness, and intelligence (Cheng et al., 2024; Balaji et al., 2024). Definitions and how each criterion is operationalized are detailed in Appendix A.2, and on our GitHub Repo[1]. Each criterion was scored on a 3-point Likert scale, with an optional "Not Applicable" category.

Annotators also provided open-ended rationales to explain their scores, which allows us to triangulate their ratings with their qualitative reasons. All evaluations were conducted on a held-out set of 11-12 unseen Q&A pairs from the 118-question test set. Annotators reported that they each took about two hours to complete the annotation and evaluation task.

4 Results

4.1 Automatic evaluation results

Automatic evaluation results reveal performance differences across open- and closed-source models, with tradeoff between precision and recall (see Table 1). Among all models, *Mistral3-LoRA* performs best across multiple metrics, achieving the highest values for precision (0.41), BERTScore F1 (0.31),

[1] https://anonymous.4open.science/r/SeniorSafeAI-36F4/, *Human evaluation* folder

Model	Precision	Recall	BERTScore F1	Entity P	Entity R	ROUGE-L	BLEU	String Sim.	Len. Ratio
Open-source models									
LLaMA3.1-Base	0.21	0.30	0.18	0.13	0.20	0.18	0.03	0.26	1.99
Qwen2.5-Base	0.23	0.30	0.24	0.15	0.20	0.20	0.04	0.27	1.75
Mistral33-Base	0.21	0.33	0.21	0.14	0.22	0.19	0.04	0.26	2.37
Gemma-Base	0.20	0.24	0.15	0.10	0.14	0.16	0.02	0.25	1.89
Mistral33-LoRA	**0.41**	0.20	**0.31**	**0.31**	0.16	**0.23**	**0.05**	**0.28**	**0.60**
LLaMA3.1-LoRA	0.11	**0.43**	0.10	0.07	**0.27**	0.13	0.02	0.16	7.20
Qwen2.5-LoRA	0.12	0.36	0.09	0.07	0.23	0.13	0.02	0.18	5.62
Gemma-LoRA	0.04	0.04	-0.25	0.05	0.02	0.04	0.00	0.21	1.29
Closed-source models									
ChatGPT-4o	0.21	0.25	0.14	0.11	0.15	0.17	0.02	0.26	1.63

Table 1: Results of automatic evaluation metrics across models. Highest values for each metric are bolded.

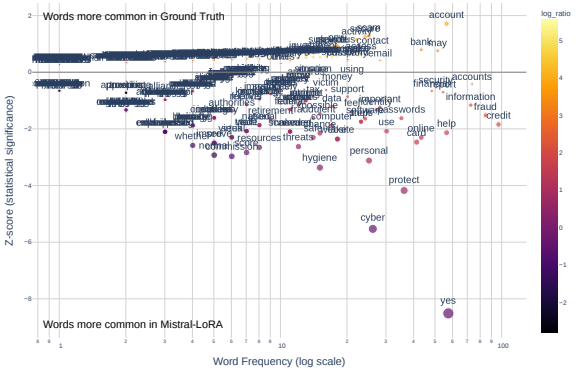
Model	Clarity	Accuracy	Relevance	Error Prev. & Educ.	Consistency	Usefulness	Intelligence
Ground Truth	2.8	2.8	2.8	2.6	2.8	2.7	2.7
Open-source models							
LLaMA3.1-Base	2.8	2.7	2.8	2.5	2.7	2.6	2.7
Qwen2.5-Base	2.8	2.7	2.7	2.4	2.7	2.6	2.6
Mistral3-Base	2.7	2.7	2.8	2.3	2.7	2.5	2.6
Gemma-Base	2.7	2.6	2.6	2.3	2.6	2.4	2.5
Mistral3-LoRA	2.6	2.6	2.6	2.2	2.6	2.4	2.5
LLaMA3.1-LoRA	2.7	2.7	2.7	2.4	2.6	2.5	2.6
Qwen2.5-LoRA	2.4	2.5	2.5	2.3	2.3	2.2	2.2
Gemma-LoRA	1.4	1.5	1.4	1.3	1.4	1.3	1.4
Closed-source models							
ChatGPT-4o	**2.9**	**2.9**	**2.9**	**2.7**	**2.9**	**2.8**	**2.8**

Table 2: Human evaluation scores across seven criteria. Each evaluator's scores were averaged first, then these averages were combined to ensure equal weighting per evaluator. Highest values for each criterion are bolded.

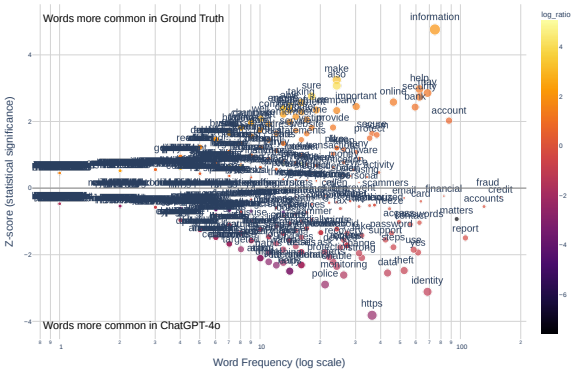
entity precision (0.31), ROUGE-L (0.23), BLEU (0.05), and string similarity (0.28). These results suggest that it generates responses that are both textually and semantically close to the ground truth. In contrast, *LLaMA3.1-LoRA* yields the highest recall (0.43) and entity recall (0.27) but produces overly long responses (length ratio: 7.20), resulting in lower precision (0.11).

Base models tend to outperform their LoRA variants, except for Mistral3, though overall performance remains modest. *Qwen2.5-Base* achieves the highest BERTScore F1 (0.24) among the base models, along with high ROUGE-L (0.20) and precision (0.23). *Mistral3-Base* has the highest recall (0.33) among base models, though its overall similarity metrics are slightly lower. These findings indicate that while no one model outperforms across all dimensions, *Mistral3-LoRA* offers the best overall alignment with ground truth in terms of accuracy. In contrast, *LLaMA3.1-LoRA* generates longer responses, which increases recall but at the cost of precision. Notably, *ChatGPT-4o*, despite being a closed-source model, does not outperform on any metrics compared to the top-performing open-source models. *ChatGPT*'s performance is comparable to open-source base models, and is lower than *Mistral3-LoRA* on key measures such as precision, BERTScore F1, and entity precision.

Log-odds distribution shows open-source models have more similar word profiles to the ground truth than closed-source model. When analyzing the distribution of all words in both corpora, *Mistral3-LoRA* (Figure 4a) has the highest percentage of words with z-scores between -1 and 1 (92.28% of all words), followed by *ChatGPT-4o* (89.42%) (Figure 4b) and *Qwen2.5-Base* (88.19%). Words within this range appear across both corpora with comparable frequency of usage. Examples of words within this range for Mistral3-LoRA includes "national", "specialists", "incident", and "losing". Words used in *ChatGPT-4o* that falls within this range includes "cards", "prevention", and "enabling". As shown in Figure 4, *Mistral3-LoRA* uses words more similar to those in the ground truth, which explains the narrow z-score range in the "Words more common in Ground Truth" (top quadrant). On the other hand, the wider z-score range in *ChatGPT-4o*'s responses indicates that ground truth words are used with less consistency compared to the ground truth responses.



(a) Mistral3-LoRA vs. Ground Truth



(b) ChatGPT-4o vs. Ground Truth

Figure 4: Comparison of word usage in ground truth (top quandrants) versus model responses (bottom quadrants), weighted by log-odds with a Dirichlet prior.

4.2 Human evaluation results

Human evaluation results (Table 2) indicate that *ChatGPT-4o* is rated most frequently (across 53 Q&As) as the top model across all coders, followed by the ground truth answers, *LLaMA3.1-Base*, and *Qwen2.5-Base*. *ChatGPT-4o*'s responses were preferred by coders because it was "easy to understand", "consistent", and "well-rounded". One annotator noted that "the model has been consistent with every response", and another stated that "the response is well-written and provides relevant information". *LLaMA3.1-Base* is also rated top for providing "clear and useful information" and "provides good resources", with examples included in the responses. One annotator indicated that "The response's recommendations also incorporate organizations like the AARP and FTC into encouraging the user to share their story, which is a unique perspective", showing that the model integrated external resources. Another coder mentioned that this model's response contains "strong focus on preventative advice like 2FA or fraud alerts", which lists the actionable steps that seniors can take to protect themselves. One drawback of the model, compared to *ChatGPT-4o*, is the formatting and tone of language. In particular, one annotator noted that "with all the text aligned to the left with a lack of strong visual separation, the response may be hard for the user to read". Another noted that, " this is a good response but it's a bit wordy", meaning that while the content is relevant and informative, the response could benefit from more concise phrasing. This is also observed in *Qwen2.5-Base*'s response, where four different annotations discussed that tone of the language and formatting could be improved. The comments were "missing the emotional support side", "ramble and give too many details", and "formatting could be a bit better", "the structure makes the response hard to read".

Human annotations also reveal that base models' responses are slightly preferred over LoRA-finetuned models'. While base models' responses are considered "concise" and "straight to the point", LoRA models' responses are sometimes overly detailed, with one annotator stating this about *Qwen2.5-LoRA*: "I like that there are additional steps but perhaps, that shouldn't be included /with/ the response because it takes away the focus of the /current/ situation.". Similarly, *LLaMA3.1-LoRA* is perceived as "very detailed", but "could be organized a little better but gives great tools and resources and explanations".

5 Discussion

Some open-source models outperform closed-source models in automatic evaluation metrics, whereas a closed-source model receive higher qualitative ratings from text formatting and language. These findings have important implications for the future development of chatbots using open-source LLMs, particularly in contexts where user security and privacy are paramount. We provide evidence and best practices demonstrating that open-source models can achieve high performance in tasks that demand strong protections for sensitive user data. By highlighting these findings, we aim to draw greater attention to the value and importance of leveraging open-source LLMs.

This emphasis is particularly timely given the growing trend of over-reliance on closed-source models developed by major corporations such as

OpenAI. To illustrate this trend, we conducted a small-scale analysis of 50 LLM-related articles from the ACM Digital Library. We found that studies using closed-source models outnumbered thos using open-source models by a factor of two. Notably, OpenAI accounted for 41 of the 42 proprietary LLMs used, highliting its near-total dominance in the current research landscape. While reported performance is high, closed-source LLMs require expensive subscriptions to access, limiting their use by small organizations and individual researchers. The growing reliance on closed-source LLMs increase concerns about the reproducibility of research. Moreover, studies that involve stringent requirements around privacy, security, or intellectual property are left with limited options, as closed-source models often lack transparency and control needed for this area of research.

We also find that **LoRA fine-tuning has different effects depending on the opens-source model**. *Mistral3-LoRA* benefits the most, outperforming all models on nearly every automatic evaluation metric, including precision, BERTScore, and log-odds similarity. *LLaMA3.1-LoRA* shows the largest improvement in recall. In contrast, fine-tuned variants like *Gemma-LoRA* and *Qwen2.5-LoRA* show minimal gains, with base models performing just as well in many cases. These differences likely reflect how well each architecture supports low-rank adaptation. This is reflected in studies showing that LoRA outperforms traditional fine-tuning on base models for LLaMA (Gajulamandyam et al., 2025; Dettmers et al., 2023) and Mistral (Zhao et al., 2024). For other models such as Qwen, and especially Gemma (Maatouk et al., 2024), LoRA and other types of parameter-efficient finetuning (PEFT) techniques because these models may not fully incorporate low-rank adapters within their attention and feedforward layers (Wang et al., 2025) (details in Figure 1). We should consider this in real-world application on which models to use for best LoRA (and other PEFT methods) finetuning.

Taken together, our results show that open-source models are reliable for the task of providing accurate, domain-aligned cybersecurity guidance for seniors, and can be safely integrated into real-world cybersecurity workflows.

We find that human annotators prefer ground truth and *ChatGPT* over open-source models due to better formatting, clarity, and tone. While *Mistral3-LoRA* outperforms all other models in terms of lexical and semantic similarity to ground truth responses, *ChatGPT-4o* is rated highest by human annotators for its clean formatting and fluent language. This finding is consistent with Clark et al. (2021), stressing the importance of developing systematic evaluation methods.

Based on the ratings (Table 2) and comments from the annotators, we learn that *ChatGPT-4o* is often preferred, especially on *error prevention & education* content, because the model outlines clear steps for mitigating risks based on a particular crime type. On the other hand, *Mistral3-LoRA* and *LLaMA-LoRA* contained fewer actionable steps in their responses, which may explain why they score slightly lower on *usefulness* category. This finding shows that our finetuning can be improved by including more structured training data that provides clear, step-by-step guidance on how to respond to different cybercrime threats.

6 Conclusion and Future Work

This study is the first to systematically evaluate the performance of open-source and closed-source LLMs in helping seniors protect themselves against cybercrimes. To address RQ1, our results show that open-source, fine-tuned models, namely Mistral3-LoRA, outperform closed-source models like ChatGPT-4o across all automatic evaluation metrics, including lexical and semantic similarity, entity overlap, and log-odds alignment with ground truth. For RQ2, while ChatGPT-4o was consistently preferred by human annotators for its polished language and clean formatting, fine-tuned open-source models demonstrated closer alignment with ground truth in terms of terminology used and perceived informativeness of the content.

These findings suggest that open-source models, when fine-tuned on domain-specific data, can provide accurate and contextually relevant responses, while also offering advantages in transparency, customizability, and data privacy. This has implications for developing accessible and trustworthy chatbot solutions to support older adults in navigating cybercrime threats, without relying on opaque models. With the goal to further develop our transparent and trustworthy chatbot, we will finetune these models using Retrieval Augmented Generation (RAG) with a larger training dataset. Future work will also involve user testing of the chatbot with senior users to assess interface accessibility, response flow, and overall user experience.

7 Acknowledgments

Anonymized for Review.

8 Limitations

This study has several areas of improvement. First, we plan to expand the dataset to over 1,000 examples to improve generalizability of our novel cybercrime Q&A training dataset. Second, beyond fine-tuning with LoRA, we plan to compare performance of open-source models trained with newer methods such as Quantized LoRA (QLoRA), and Retrieval-Augmented Generation (RAG). Finally, our study does not yet include user feedback from the senior population. We will conduct a follow-up study with 10–15 senior citizens, in collaboration with a county-level senior center, to collect data on chatbot usability, perceived trust, and the users' likelihood to act on the chatbot's advice.

References

Bhashithe Abeysinghe and Ruhan Circi. 2024. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches. *arXiv preprint arXiv:2406.03339*.

Lisa C Adams, Daniel Truhn, Felix Busch, Felix Dorfner, Jawed Nawabi, Marcus R Makowski, and Keno K Bressem. 2024. Llama 3 challenges proprietary state-of-the-art large language models in radiology board–style examination questions. *Radiology*, 312(2):e241191.

Garima Agrawal, Kuntal Pal, Yuli Deng, Huan Liu, and Ying-Chih Chen. 2024. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23164–23172.

Ananya Balaji, Lea Duesterwald, Ian Yang, Aman Priyanshu, Costanza Alfieri, and Norman Sadeh. 2024. Generating effective answers to people's everyday cybersecurity questions: an initial study. In *International Conference on Web Information Systems Engineering*, pages 363–379. Springer.

PV Charan, Hrushikesh Chunduri, P Mohan Anand, and Sandeep K Shukla. 2023. From text to mitre techniques: Exploring the malicious use of large language models for generating cyber attack payloads. *arXiv preprint arXiv:2305.15336*.

Xusen Cheng, Liyang Qiao, Bo Yang, and Zikang Li. 2024. An investigation on the influencing factors of elderly people's intention to use financial ai customer service. *Internet research*, 34(3):690–717.

Vanessa Clairoux-Trepanier, Isa-May Beauchamp, Estelle Ruellan, Masarah Paquet-Clouston, Serge-Olivier Paquette, and Eric Clay. 2024. The use of large language models (llm) for cyber threat intelligence (cti) in cybercrime forums. *arXiv preprint arXiv:2408.03354*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Sam Fondrie-Teitler and Avi Jayanti. 2023. Consumers are voicing concerns about ai. Technical report, Federal Trade Commission.

Deva Kumar Gajulamandyam, Sainath Veerla, Yasaman Emami, Kichang Lee, Yuanting Li, Jinthy Swetha Mamillapalli, and Simon Shim. 2025. Domain specific finetuning of llms using peft techniques. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00484–00490. IEEE.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7.

Internet Crime Complaint Center. 2023. FBI Elder Fraud Report 2023. Technical report, FBI.

Ray Islam and Peter Sandborn. 2023. Demonstration of a response time based remaining useful life (rul) prediction for software systems. *arXiv preprint arXiv:2307.12237*.

Paul Iusztin and Maxime Labonne. 2024. Llm engineer's handbook: Master the art of engineering large language models from concept to production.

S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, and D. Guha. 2024. A literature survey on open source large language models. In *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, pages 133–143.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Feng Liu, Jiaqi Jiang, Yating Lu, Zhanyi Huang, and Jiuming Jiang. 2025. The ethical security of large language models: A systematic review. *Frontiers of Engineering Management*, pages 1–13.

Ali Maatouk, Kenny Chirino Ampudia, Rex Ying, and Leandros Tassiulas. 2024. Tele-llms: A series of specialized large language models for telecommunications. *arXiv preprint arXiv:2409.05314*.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Farzad Nourmohammadzadeh Motlagh, Mehrdad Hajizadeh, Mehryar Majd, Pejman Najafi, Feng Cheng, and Christoph Meinel. 2024. Large language models in cybersecurity: State-of-the-art. *arXiv preprint arXiv:2402.00891*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16.

Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. 2025. Lora-pro: Are low-rank adapters properly optimized? *ICLR*.

Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, and 1 others. 2024. Laboratory-scale ai: Open-weight models are competitive with chatgpt even in low-resource settings. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1199–1210.

HanXiang Xu, ShenAo Wang, Ningke Li, Kailong Wang, Yanjie Zhao, Kai Chen, Ting Yu, Yang Liu, and HaoYu Wang. 2024. Large language models for cyber security: A systematic literature review. *arXiv preprint arXiv:2405.04760*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. Lora land: 310 fine-tuned llms that rival gpt-4, a technical report. *arXiv preprint arXiv:2405.00732*.

A Appendix

A.1 Conversational snippets from SeniorSafeAI

Figures A1-A3 show the actual conversational snippets between the user and *Mistral-LoRA*, *ChatGPT-4o*, *Llama3.1-Base*, respectively.

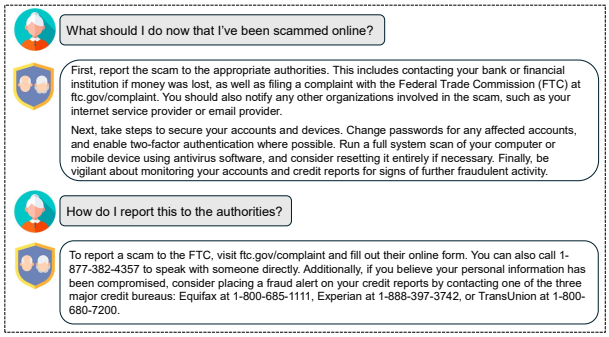


Figure A1: Conversational snippet between user and *Mistral-LoRA*

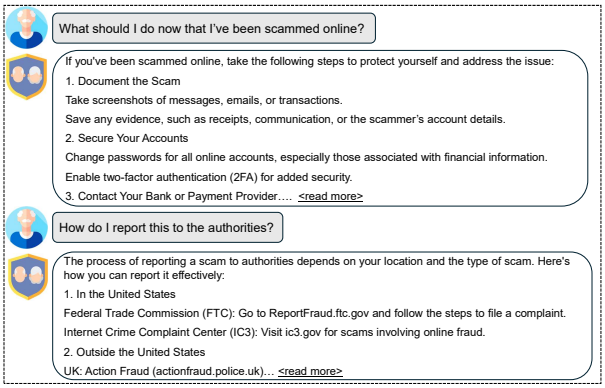


Figure A2: Conversational snippet between user and *ChatGPT-4o*

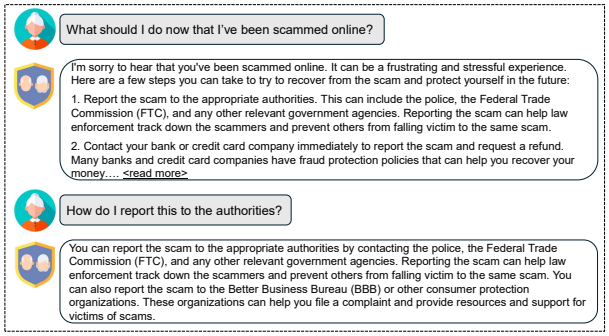


Figure A3: Conversational snippet between user and *Llam3.1-Base*

A.2 Human evaluation metrics

The following lists seven criteria used for human evaluation, including their definitions and operationalizations:

- *Clarity* evaluates whether the response is easy to follow and provides clear directions without unnecessary complexity.
- *Accuracy* assesses whether the information is factually correct, up-to-date, and aligned with best security practices (e.g., encryption,

multi-factor authentication, regular software updates).

- *Relevance* examines whether the response directly addresses the user's question.
- *Error prevention and education* evaluates the extent to which the response helps users avoid common cybersecurity mistakes (e.g., phishing scams or unsafe browsing).
- *Consistency* measures whether the chatbot's response maintains a stable tone, terminology, and structure.
- *Usefulness* assesses how effective, clear, and actionable the instructions are for helping users accomplish a task or understand a concept.
- *Intelligence* evaluates the overall quality of reasoning and contextual appropriateness of the response.