Forensic Science International: Digital Investigation 52 (2025) 301872

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS EU 2025 - Selected Papers from the 12th Annual Digital Forensics Research Conference Europe

ForensicLLM: A local large language model for digital forensics

Binaya Sharma [a,c,*], James Ghawaly [b,c], Kyle McCleary [b,c], Andrew M. Webb [c], Ibrahim Baggili [a,c]

[a] Baggil(i) Truth (BiT) Lab, Center for Computation & Technology, Baton Rouge, LA, USA
[b] Intersectional AI and Security (AISx) Lab, Center for Computation & Technology, Baton Rouge, LA, USA
[c] Division of Computer Science & Engineering, Louisiana State University, Baton Rouge, LA, USA

ARTICLE INFO

Keywords:
Digital forensics
Digital investigations
Generative AI
Large language model (LLM)
Admissibility of evidence
ChatGPT
LLaMA
Retrieval augmented generation (RAG)
Fine-tuning

ABSTRACT

Large Language Models (LLMs) excel in diverse natural language tasks but often lack specialization for fields like digital forensics. Their reliance on cloud-based APIs or high-performance computers restricts use in resource-limited environments, and response hallucinations could compromise their applicability in forensic contexts. We introduce ForensicLLM, a 4-bit quantized LLaMA-3.1–8B model fine-tuned on Q&A samples extracted from digital forensic research articles and curated digital artifacts. Quantitative evaluation showed that ForensicLLM outperformed both the base LLaMA-3.1–8B model and the Retrieval Augmented Generation (RAG) model. ForensicLLM accurately attributes sources 86.6 % of the time, with 81.2 % of the responses including both authors and title. Additionally, a user survey conducted with digital forensics professionals confirmed significant improvements of ForensicLLM and RAG model over the base model. ForensicLLM showed strength in *"correctness"* and *"relevance"* metrics, while the RAG model was appreciated for providing more detailed responses. These advancements mark ForensicLLM as a transformative tool in digital forensics, elevating model performance and source attribution in critical investigative contexts.

1. Introduction

Digital Forensics (DF) involves the acquisition, authentication, and analysis of digital evidence in a manner that is both legal and scientific. A crucial aspect for investigators is ensuring the admissibility of digital evidence in court. The Daubert standards, established by the 1993 Daubert v. Merrell Dow Pharmaceuticals case, set guidelines for evaluating the scientific validity and reliability of methodologies used to gather evidence. The major Daubert criteria include: 1) the ability to test the methodology and its prior testing, 2) known error rates, 3) peer review and publication status, and 4) acceptance by the relevant scientific community (Farrell, 1993; Baggili et al., 2007). Consequently, tools and techniques used in DF must be peer-reviewed and scientifically accepted.

Large Language Models (LLMs) often lack transparency regarding their training data and internal mechanisms, complicating the verification of their outputs' scientific basis. Further, data confidentiality is paramount in DF, and relying on cloud-based LLMs raises data security concerns due to the potential exposure of sensitive case information to third parties (Lukas et al., 2023). General purpose cloud-based models like ChatGPT require internet connectivity, rendering them impractical in secure environments where internet access is restricted. Also, while these LLMs might offer broad applicability across various tasks, they often lack domain-specific expertise (Scanlon et al., 2023). The computing power requirements for training and inference can be substantial, leading to increased operational costs. Fine-tuning these models to specific tasks or domains can be challenging and time-consuming, requiring significant expertise and computational resources (Sevilla et al., 2022; Sharir et al., 2020).

Smaller open-source models like Large Language Model Meta AI (LLAMA) (Touvron et al., 2023) and Mistral (Jiang et al., 2023) present compelling alternatives while offering greater flexibility. These models can achieve comparable results to larger models when fine-tuned on domain-specific datasets (Bolton et al., 2024; Rebei, 2023). Additionally, techniques such as quantization (Jacob et al., 2018) and Low-Rank Adaptation (LORA) (Hu et al., 2021) significantly reduce the computational requirements for inference and fine-tuning of these models. Furthermore, RAG approach (Lewis et al., 2020) allows models to

* Corresponding author. Baggil(i) Truth (BiT) Lab, Center for Computation & Technology, Baton Rouge, LA, USA.
E-mail addresses: bsharm6@lsu.edu (B. Sharma), jghawaly@lsu.edu (J. Ghawaly), kmccl24@lsu.edu (K. McCleary), andrewwebb@lsu.edu (A.M. Webb), ibaggili@lsu.edu (I. Baggili).

https://doi.org/10.1016/j.fsidi.2025.301872

Available online 24 March 2025

retrieve additional context information from external knowledge sources during inference, effectively expanding their knowledge beyond just the training data.

Current practices in DF heavily rely on manual analysis, often constrained by the examiner's expertise and existing knowledge bases. Survey results from Yao et al. (2024) show that research in cybersecurity and DF has primarily focused on evaluating larger cloud-based models with minimal exploration of smaller, specialized models. Our objective is to fill this research gap by examining the application of a domain-specialized LLM in DF. Specifically, our research questions are:

RQ1 How do fine-tuning and RAG framework compare in terms of performance in digital forensics?
RQ2 How reliable is a fine-tuned LLM in retrieving references to aid in digital forensic investigations?
RQ3 How can a local LLM help improve understanding of digital artifacts, forensic tools, and processes in a digital forensics investigation?

Our work makes the following key contributions:

- We introduce ForensicLLM, a model optimized on top of LLaMA-3.1–8B with a dataset comprising DF research papers and curated digital forensic artifacts. *To the best of our knowledge, this is the primary account for developing a local LLM for DF.*
- We evaluate ForensicLLM based on its performance on held-out test datasets and its ability to generate correct source citations.
- We compare the performance of ForensicLLM with base LLaMA-3.1–8B and LLaMA-3.1–8B equipped with a RAG component.
- We conduct a user study with DF professionals, evaluating model responses for usefulness, correctness, relevance, citation, and understanding.

The paper is structured as follows: Section 2 covers related work, Section 3 explores key LLM concepts, Sections 4 and 5 describe the dataset and methodology for fine-tuning ForensicLLM, Section 6 presents and discusses the evaluation results, Section 7 discusses limitations and future work, and Section 8 draws conclusions.

2. Related work

Scanlon et al. (2023) evaluated ChatGPT for digital forensic tasks like artifact analysis, evidence searching, and incident response. The accuracy, relevance, and completeness of the responses were analyzed. While ChatGPT showed potential in generating regular expressions, enhancing keyword searches, summarizing documents, and aiding in coding tasks, significant limitations were noted, such as biases from training data, lack of domain specialization, non-deterministic outputs, inability to handle real-world evidence, and a tendency to generate hallucinations.

Michelet and Breitinger (2023) investigated the potential of using LLMs like ChatGPT-3.5 and LLaMA-2-13B to assist in digital forensic report generation. ChatGPT outperformed the locally run LLaMA-2-13B model in generating more accurate and complete texts. However, both models required significant human proofreading. Further, they found that while LLMs cannot fully automate report writing, they can still help with text summarization and automating certain sections like 'Introduction'.

Sreya et al. (2023) proposed a forensic analysis and evidence identification framework using ChatGPT but lacked implementation and evaluation. Henseler and van Beek (2023) examined ChatGPT's potential to assist legal professionals in digital investigations, concluding that while it enhances efficiency, human oversight remains essential.

Most existing work in DF has focused on analyzing existing LLMs, such as ChatGPT. In contrast, Silalahi et al. (2023) developed a model for detecting anomalies in drone flight logs. They fine-tuned a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to recognize negative sentiment, which indicated an issue or incident during the drone's flight.

3. Background

This section explores key background concepts integral to LLMs.

3.1. Tokenization

Tokenization splits input text into smaller units called tokens, which can be words, subwords, or characters. LLaMA-3 uses subword tokenization with the *tiktoken* Byte Pair Encoding (BPE) tokenizer (AI@Meta, 2024a), containing over 128k unique tokens. This approach splits rare words into known subword units, improving the model's text-processing capabilities.

3.2. Token embeddings

Token embeddings are numerical vectors that capture the semantic meaning of the token and the relationships between tokens. During inference, the token embedding layer in the LLM's architecture acts as a simple lookup table that returns a token's vector embedding based on its unique identifier. In most models, the token embedding layer is learned through pretraining. In other cases, a pre-trained token embedding model is used such as word2vec (Mikolov et al., 2013) or Global Vectors for Word Representation (GLOVE) (Pennington et al., 2014).

3.3. Attention

Attention mechanisms are a key element of transformer-based LLMs such as LLAMA and Generative Pre-trained Transformer (GPT). They enable the model to focus on relevant parts of the input sequence when generating outputs. Specifically, the self-attention layer in transformers computes the relevance of each token in the input sequence to every other token, allowing the model to capture long-range dependencies and better understand the context.

3.4. Quantization

Quantization optimizes LLMs by reducing the precision of numerical data, typically converting 32-bit floating-point weights to lower precision formats (e.g., 16-bit or 8-bit). This decreases memory usage and computational costs, enabling deployment on devices with limited resources without significantly compromising model accuracy (Jacob et al., 2018).

3.5. Sentence embeddings and vector stores

Similar to token embeddings discussed in Section 3.2, LLMs have also been designed to generate sentence embeddings, which capture the semantic meaning of longer text. These embeddings enable processing large contexts, often composed of hundreds of tokens, with the primary goal of enhancing semantic information retrieval. Vector embedding models are trained so that semantically similar information will have closer vector representations, while dissimilar information will be farther apart, measured by similarity metrics like dot product or cosine distance.

An information retrieval system can be developed using sentence embedding models and a vector store. A vector store, or vector database, such as Facebook AI Similarity Search (FAISS) (Douze et al., 2024) and ChromaDB, are specialized storage optimized for efficiently retrieving high-dimensional vectors like embeddings. These vector stores offer efficient search functionalities, allowing the LLM to retrieve relevant information for analysis based on user queries by embedding the query and returning related information from the vector database via a similarity search.

3.6. Fine-tuning

Fine-tuning is a process to further train (or fine-tune) an LLM on a domain-specific dataset (Bolton et al., 2024; Rebei, 2023). This additional training allows the model to adapt its parameters to the specific requirements of the target domain, improving its performance on related tasks.

Techniques such as LORA (Hu et al., 2021) and Quantized Low-Rank Adaptation (QLORA) (Dettmers et al., 2024) have recently been developed to significantly improve the computational efficiency of fine-tuning, enabling fine-tuning powerful LLM on consumer-grade GPUs.

LORA introduces a pair of low-rank modification matrices ($\mathbf{A}$, $\mathbf{B}$) for each weight matrix $\mathbf{W}$ in the model. If $\mathbf{W}$ is an $m \times n$ weight matrix, $\mathbf{A}$ is an $m \times r$ matrix, and $\mathbf{B}$ is an $r \times n$ matrix, where r is the rank. Instead of updating the entire weight matrix $\mathbf{W}$ during fine-tuning, LoRA only adjusts the low-rank matrices $\mathbf{A}$ and $\mathbf{B}$.

$$\mathbf{W}' = \mathbf{W} + \mathbf{AB} \tag{1}$$

Where $\mathbf{W}'$ is the updated weight matrix after LORA.

QLORA is an extension of LORA that performs quantization of the low-rank modification matrices during fine-tuning. This process involves representing the modification matrices ($\mathbf{A}$, $\mathbf{B}$) using lower-precision data types (e.g., 4-bit integers) instead of the default 32-bit floating-point numbers. The quantization step can be expressed as:

$$\mathbf{A}_q = \text{quantize}(\mathbf{A}) \tag{2}$$

$$\mathbf{B}_q = \text{quantize}(\mathbf{B}) \tag{3}$$

$$\mathbf{W}' = \mathbf{W} + \mathbf{A}_q\mathbf{B}_q \tag{4}$$

Where quantize$(\cdot)$ is a function that maps floating-point values in $\mathbf{A}$ and $\mathbf{B}$ to lower-precision integers.

4. Dataset selection

We optimized our model on peer-reviewed DF papers and meticulously curated digital artifacts data.

4.1. DF research papers

The first component of our dataset included DF research papers obtained from the journal "Forensic Science International: Digital Investigation" and its predecessor, "Digital Investigation." A total of 1082 papers were downloaded from the journal. Fig. 1 illustrates the distribution of different research topics in these papers.

4.2. DF artifacts

We also utilized Artifact Genome Project (AGP), a Curated Repository of Forensic Artifacts (CUFAs) developed by Grajeda et al. (2018). This repository contains a diverse set of digital forensic artifacts compiled by students and researchers. Digital artifacts submitted to AGP undergo a review procedure to ensure their validity and authenticity. At the time of this study, AGP contained a total of 1390 DF artifacts which are summarized in Table 1.

AGP included several key fields: *Title* (artifact name), *Type* (artifact category), *Device* (hardware where the artifact was found), *Path* (discovery location), *Description* (detailed artifact information), *Comments* (user insights), *Search Tags* (keywords for retrieval), and *Data* (relevant information based on artifact type).

5. Methodology

This section details the development of ForensicLLM, a DF-focused LLM utilizing a modified form of the Retrieval Augmented Fine-tuning (RAFT) (Zhang et al., 2024) approach (Fig. 2). We employed LLaMA-3.1–8B, a recently released open-source 8-billion parameter model from Meta AI, as the base model due to its superior performance compared to similar models of this size (e.g., Gemma-7B and Mistral-7B) (AI@Meta, 2024a). With quantization, this model can run on a consumer-grade GPU, such as the Nvidia RTX 4090, at speeds of over 100 tokens per second. The development process involved several stages, including data extraction and cleaning, building retrieval module, generating the RAFT dataset, training the RAFT model, and generating inferences.

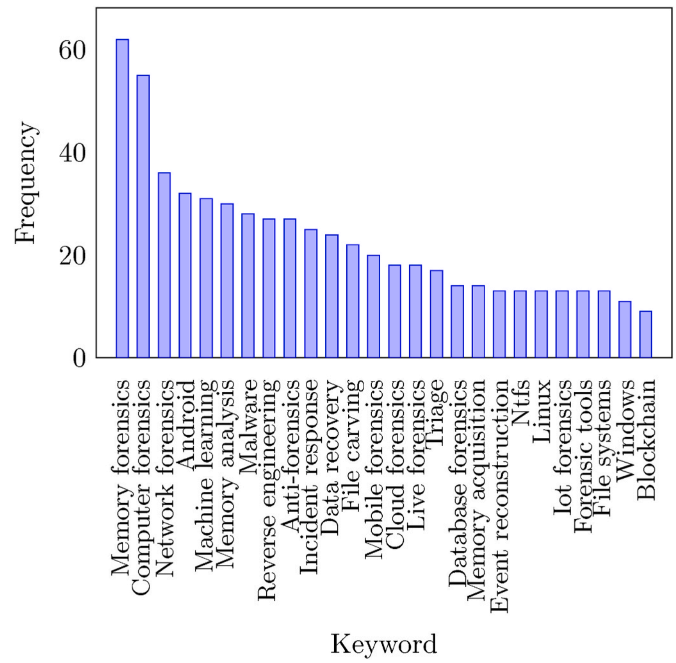


Fig. 1. Frequently occurring keywords in the DF papers.

Table 1
Breakdown of artifact types in AGP.

Artifact Type	Count
File	1126
Windows Registry	211
Memory	19
Network Packet	19
Others	15
Total	1390

5.1. Data extraction and cleaning

From the downloaded pdf files, text content between abstract and acknowledgments sections was extracted using *pdftotext* library. The references section was excluded to avoid irrelevant matches in the vector search. Additionally, consecutive periods, which occurred in some papers due to PDF formatting, were replaced with a single period to conserve token space (Step in①Fig. 2).

5.2. Retriever module

We performed the following steps to set up the Chroma DB vector retriever, which provided relevant contextual information to the model during fine-tuning and inference.

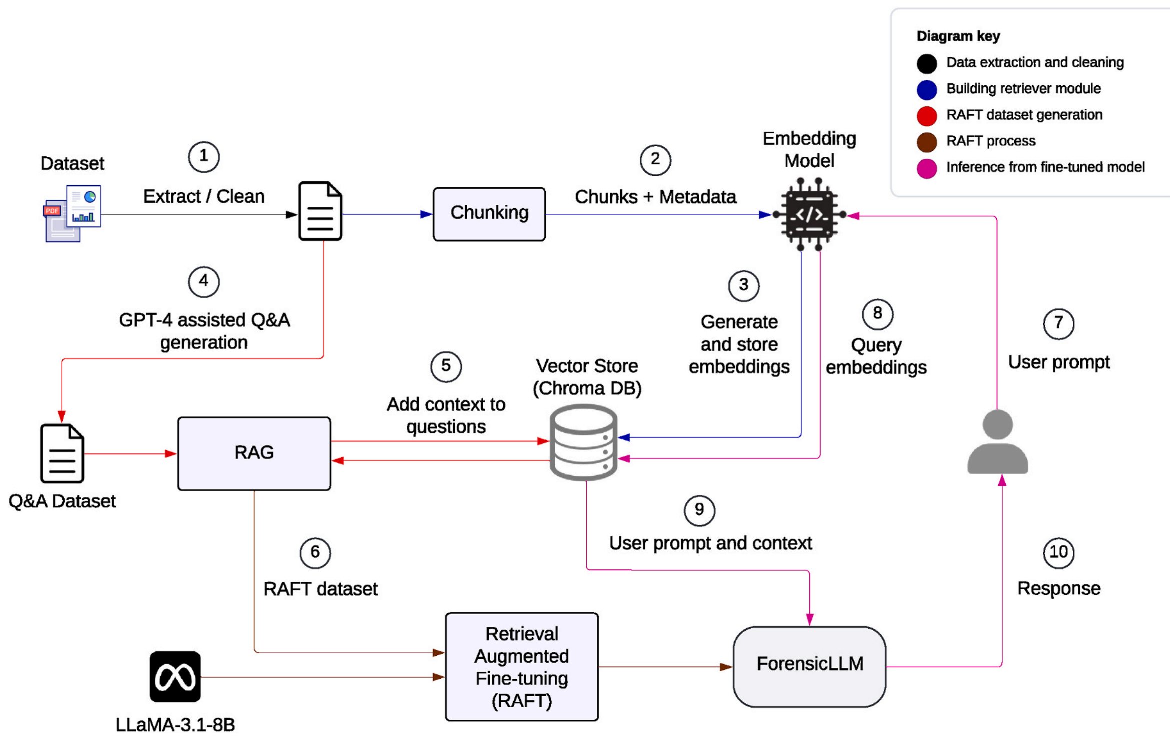


Fig. 2. Retrieval Augmented Fine-tuning (RAFT) approach for developing ForensicLLM.

1. *Chunking*: Extracted text contents were split into smaller chunks of around 2000 characters using Langchain's *RecursiveCharacterTextSplitter*. Splitting prioritized double line breaks ("\n\n"), periods followed by line breaks (".\n"), single line breaks ("\n"), and then periods ("."). Chunks with less than 600 characters (likely irrelevant information such as headers and footnotes) were filtered out.
2. *Metadata Annotation*: Authors and title of each paper were added as metadata to the chunks (Step②).
3. *Embeddings Generation and Storage*: Chunks were processed through the *UAE_Large-V1* embedding model to generate their respective vector representations. This embedding model was chosen for its high performance and relatively low memory usage, based on the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022). The generated embeddings and associated metadata were stored in a Chroma DB, as shown in Step ③.

5.3. RAFT dataset generation

A labeled dataset is required to perform RAFT. However, there is not an existing dataset in DF, and manual creation is not feasible. Instead, we leveraged GPT-4 to build a dataset using the extracted data from research papers. Using LLMs for generating synthetic and augmented datasets is a growing trend (Tihanyi et al., 2024; Patel et al., 2024; AI@Meta, 2024b). GPT-4's human-like performance in language comprehension and text summarization (Achiam et al., 2023) made it a suitable choice for our task. We provided each research paper's full content and a prompt to *GPT-4 Turbo* to extract a set of Q&A pairs (Step④in Fig. 2, expanded in Fig. 3).

An effective prompt was necessary to generate a high-quality Q&A dataset for fine-tuning. The prompt was carefully crafted with the help of one of the authors' expertise in the DF field spanning over 19 years. The prompt included the following guidelines:

- *Relevance*: Ensure practical utility in the field of DF.
- *Language Use*: Formulate answers using language from the paper as much as possible to maintain the original context and technical accuracy.
- *Output Format*: Present output in a valid JSON list structure with each entry consisting of two keys: prompt (the question) and completion (the answer).
- *Content Restrictions*: Ensure all questions are free from digital formatting and exclude direct references to the paper, paper's title or authors in the questions.
- *Citation*: Must include APA citation in the answers. This is to ensure that answers come from peer-reviewed publications.
- *Answer Length*: Ensure answers are detailed, exhaustive, and comprehensive.

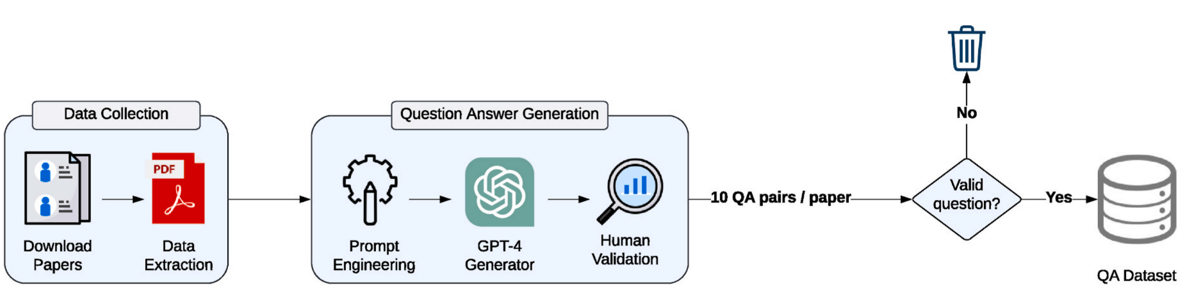


Fig. 3. GPT-4 assisted Q&A dataset creation process using DF research papers.

Additionally, the prompt suggested different topics to be explored. These included digital artifacts, their locations (across network, disk, memory), metadata elements (timestamps and identifiers), forensic processes (acquisition, authentication), tools, findings, and forensic soundness. To encourage diverse and creative question generation, we used open-ended suggestions such as "*Consider asking about*" when referencing these topics, rather than providing overly specific instructions. The exact prompt passed to *GPT-4 Turbo* is shown in FigureA. 7.

This process generated approximately 10,000 Q&A pairs from 1082 research papers. Table B.8 in Appendix B shows some of these generated Q&A pairs. The generated pairs were then manually validated. During validation, we observed that about 10 % of such Q&A pairs contained unwanted phrases like *"in the above study"* or *"in the proposed methodology"*. These Q&A pairs were filtered out.

For each question in the Q&A dataset, the top 10 most similar chunks were retrieved from the Chroma DB (Step⑤in Fig. 2). The data was formatted in the Alpaca format, with each JSON line containing the following keys:

- "*instruction*": Instructions for the model on how to handle the input.
- "*input*": Retrieved chunks from Chroma DB and the question itself.
- "*output*": Desired output, including answer from the Q&A dataset and source information (title/authors).

This final dataset from Step⑥(Fig. 2) was then split into training (75 %) and testing (25 %) sets, yielding 6739 training samples and 2244 testing samples.

5.4. Retrieval Augmented Fine-tuning (RAFT)

The RAFT approach enables a LLM to leverage external knowledge sources during training and inference through a combination of fine-tuning and RAG. By providing relevant contextual information from a retrieval database in the fine-tuning dataset, the model can learn to incorporate and reason over this additional information, enhancing its ability to provide accurate and well-informed responses.

We used the *Axolotl* toolkit to perform fine-tuning on the LLaMA-3.1–8B model. The training dataset underwent pre-tokenization using the *tiktoken* BPE tokenizer. The LLaMA-3.1–8B model was quantized to 4-bit integers using the *bitsandbytes* library. Fine-tuning was performed using QLoRA, a 4-bit quantized and highly optimized version of LoRA (Dettmers et al., 2024). We set aside 20 % of the training data as a validation set for evaluating the model's performance during training. The specific hyperparameter configurations used for the fine-tuning stage are provided in Table 2. The model was trained for a total of 4 epochs and 8 model checkpoints were stored for each epoch. The validation loss was minimized after approximately 2 epochs and beyond that began to overfit. The complete fine-tuned model was selected from the checkpoint at 2 epochs. This practice is referred to as early stopping and is a standard regularization method used in neural network optimization (Prechelt, 2002).

Table 2
Fine-tuning configurations for ForensicLLM.

Configuration	Value
Base model quantization	4 bits
Adapter	QLoRA
Rank	32
Alpha	32
Dropout	0.05
Optimization	Paged Adam 32-bit
Learning rate scheduler	Cosine, starting at 0.0002
Warmup steps	100
Epochs	4

5.5. Inference

Finally, ForensicLLM is used for inference generation. This is shown in Steps⑦through⑩(Fig. 2). The user prompt is first embedded using the same *UAE_Large-V1* embedding model. This embedded representation is then queried against the vector store, which retrieves relevant context for ForensicLLM. Then, this context is leveraged to generate a response.

6. Evaluation and results

Fine-tuning and inferences were conducted on an NVIDIA RTX 4090 GPU with 24 GB of memory. We evaluate ForensicLLM on different criteria as follows.

6.1. Performance on test dataset

RQ1: How do fine-tuning and RAG framework compare in terms of performance in digital forensics?

A 25 % split of the generated dataset (described in Section 5.3) was reserved for testing. This section compares the performance of different models on these test samples. We evaluate three models:

- *LLaMA-3.1-8B*: This model receives only the question without any additional context.
- *LLaMA-3.1-8B + RAG*: This model leverages the LLaMA-3.1–8B model but is augmented with RAG to incorporate context during generation.
- *ForensicLLM*: This is our proposed model utilizing the RAFT fine-tuning approach with RAG.

To increase confidence in the conclusions of our analysis, we utilize three different metrics to evaluate performance, two based on semantic embeddings (BERTScore and BAAI's BGE-M3) and one based on chain-of-thoughts reasoning with foundational LLM (G-Eval using *GPT-4o*).

BERTScore (Zhang et al., 2019) leverages contextual BERT embeddings to compute token-level semantic similarity between the generated and reference texts. BERTScore provides three main scores: Precision, Recall, and F1, which range from 0 to 1.

BAAI's BGE-M3 (Chen et al., 2024) embedding model was also employed to perform representational comparisons between reference and model-generated responses. BGE-M3 was chosen for its state-of-the-art embedding performance and multi-granularity, enabling comparisons between texts of vastly different lengths.

G-Eval (Liu et al., 2023) is a recently proposed evaluation metric for open-ended language generation tasks, measuring semantic and factual consistency to assess quality and relevance. It outperforms other LLM-based metrics in aligning with human judgment. G-Eval uses an external LLM (*GPT-4o* in our case) to rate generated text (1–5) on coherence, consistency, fluency, and relevance. For each category, 20 outputs are generated using a temperature of 2, and the scores are averaged to produce final category scores. The overall score is the average of these four category scores. Table 3 presents evaluation results on the test dataset across BERTScore, BGE-M3, and G-Eval. The results are summarized as follows:

- *BERTScore F1*: ForensicLLM achieves 0.9232, better than LLaMA-3.1–8B (0.8872, +4.06 %) and LLaMA-3.1–8B + RAG (0.8923, +3.46 %).
- *BGE-M3 Cosine*: ForensicLLM scores 0.9091, surpassing LLaMA-3.1–8B (0.8623, +5.43 %) and LLaMA-3.1–8B + RAG (0.8805, +3.25 %).
- *G-Eval Overall*: ForensicLLM obtains 2.7544, outperforming LLaMA-3.1–8B (2.3787, +15.79 %) and LLaMA-3.1–8B + RAG (2.6329, +4.61 %).

Table 3 also includes the average response length in tokens,

Table 3
Evaluation results on the test dataset for different models across BERTScore, BAAI's BGE-M3, and G-Eval.

Model	Method	Avg. response length	BERTScore			BGE-M3	G-Eval
		(in tokens)	**Precision**	**Recall**	**F1 Score**	**Cosine**	**Overall**
LLaMa-3.1–8B	One-shot	128.94	0.8929	0.8817	0.8872	0.8623	2.3787
LLaMa-3.1–8B + RAG	One-shot	226.38	0.8841	0.9010	0.8923	0.8805	2.6329
ForensicLLM	One-shot	156.34	**0.9215**	**0.9250**	**0.9232**	**0.9091**	**2.7544**

providing insights into the verbosity of each model. The base LLaMA model exhibits the lowest token count (128.94), suggesting a lack of sufficient knowledge in the DF domain. Conversely, the RAG model generates the most verbose responses (226.38 tokens on average). Interestingly, ForensicLLM achieves higher scores than both the base and RAG models while maintaining a moderate token count (156.34), indicating its ability to generate concise yet semantically accurate and relevant responses.

The density plot in Fig. 4 reveals a distinction among the performance of the three models based on G-Eval metric. ForensicLLM's peak density is furthest to the right, indicating generally higher G-Eval scores. LLaMA-3.1–8B's peak density is furthest to the left, suggesting lower overall scores compared to the other models. The RAG model falls between ForensicLLM and LLaMA-3.1–8B. ForensicLLM's distribution appears to be slightly wider and flatter at the top, suggesting more consistent performance across a range of higher scores.

Fig. 5 shows the distributions of BERTScore F1 and G-Eval scores for ForensicLLM across the top 20 most frequent question categories in the test dataset. The BERTScore F1 plot (Fig. 5a) shows that most categories have median scores between 0.91 and 0.93, indicating high-quality responses. Narrower boxes suggest more consistent performance in categories such as *"Android"*. Wider boxes and outliers in categories such as *"Computer forensics"* indicate greater variability. Similarly, the G-Eval score plot (Fig. 5b) shows median scores typically between 2.5 and 3. The Inter-Quartile Range (IQR) and whiskers reveal variability, with categories like *"Computer forensics"* exhibiting greater spread.

6.2. Source attribution

RQ2: How reliable is a fine-tuned LLM in retrieving references to aid in digital forensic investigations?

To ensure the credibility and admissibility of evidence criteria when ForensicLLM is used during digital forensic investigations, we evaluated ForensicLLM's ability to correctly cite peer-reviewed sources within its generated answers. We investigated the following aspects:

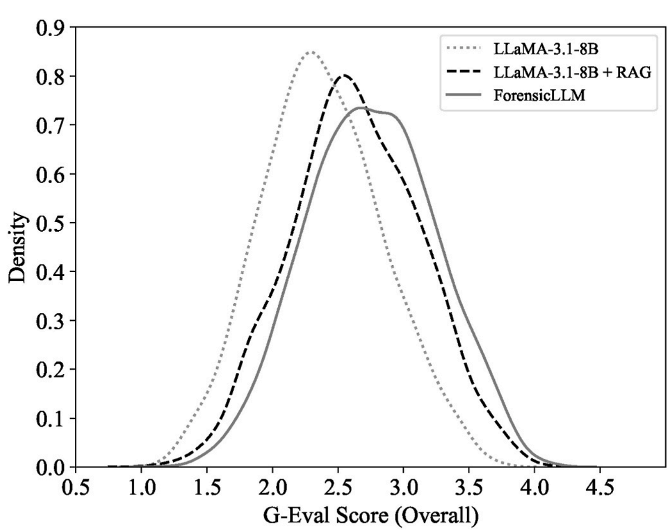


Fig. 4. Density plot of G-Eval scores for different models.

- *Citation Presence*: Does the ForensicLLM's response include citations for its claims?
- *Cited Source in Context*: Do the cited sources (title and/or author) match with the context provided?

LLM-as-a-judge approach was used (Zheng et al., 2024) to assess citations in the generated responses. *GPT-4o* was provided with the generated response, the context accompanying the initial query, and a specific evaluation prompt (shown in Appendix A in Figure A.8) to produce the results in Table 4. Random samples were cross-checked manually to verify the accuracy of this approach.

The citation behavior and potential hallucinations in ForensicLLM's generated responses can be analyzed from Table 4. Out of 2244 total responses, 2243 included citations, indicating that the system effectively learned to generate citations alongside the responses. In 1823 cases (81.2 %), the citations matched both the title and authors mentioned in the context, suggesting that the responses were grounded in the given information. In 119 cases, the citations matched only the authors. The system correctly cited the source (either title or authors) 86.6 % of the time.

Additionally, the results reveal instances of hallucinations or deviations from the provided context. 300 responses had citations that did not match either the title or author in the context, raising concerns about potential hallucinations even with the context provided to a fine-tuned model. Interestingly, there was only 1 instance where the citation matched only the title and not the author.

Overall, the results emphasize the importance of careful evaluation and analysis of LLM's responses, especially in domains where factual accuracy and adherence to evidence is crucial. Identifying and mitigating potential hallucinations and inconsistencies should be a key focus for future improvements in these types of systems.

6.3. Forensic relevance: user study

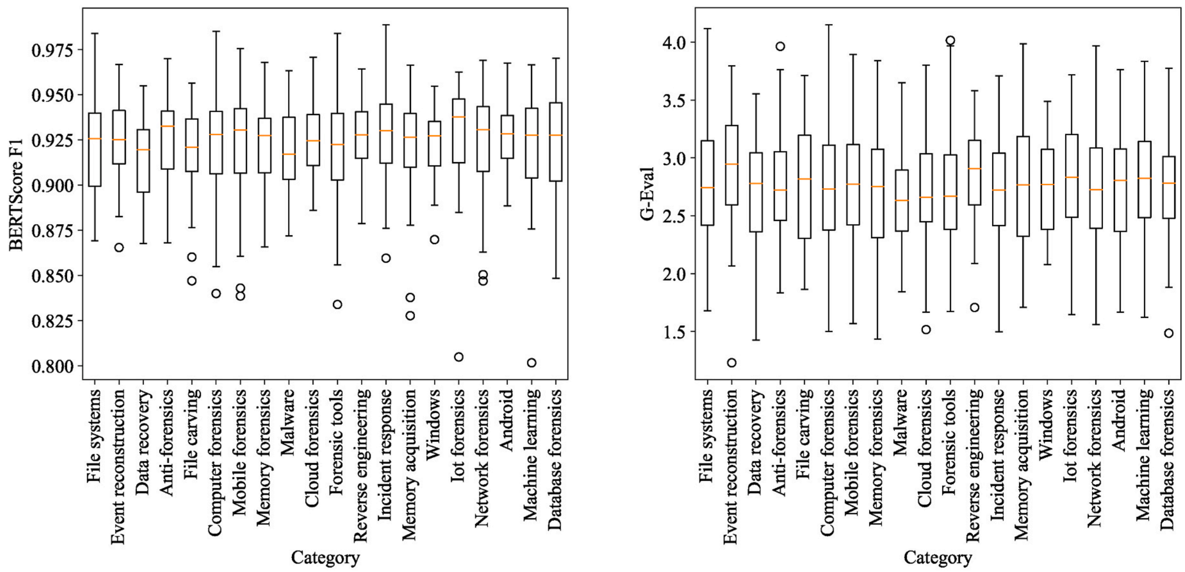
RQ3: How can a local LLM help improve understanding of digital artifacts, forensic tools, and processes?

We conducted a user study to evaluate the effectiveness and practicality of ForensicLLM in real-world DF scenarios, from the perspectives of real-world DF experts.

6.3.1. User study design

We developed a web application for interacting with three model configurations (LLaMA-3.1–8B, LLaMA-3.1–8B + RAG, and ForensicLLM) and rating their generated responses. The web application was built using Flask and deployed on Microsoft Azure.

Before interacting with the models, participants completed a pre-questionnaire to collect data on demographics, professional background, and familiarity with LLMs. Next, they engaged with the survey webapp by posing investigative questions. Two hypothetical DF scenarios were presented for this purpose: the first involved a bank robbery. The second focused on a missing person investigation. Participants were provided detailed descriptions of the cases, including the devices confiscated. They were instructed to inquire about tools, procedures, and artifacts relevant to the investigation. Participants were asked to assess the similarity of each hypothetical case to those they had previously encountered in their professional work. The full text of the scenarios is provided in Appendix C.



(a) BERTScore F1 (b) G-Eval

Fig. 5. Distributions of BERTScore F1 and G-Eval scores for ForensicLLM across top 20 most frequent question categories in test dataset.

Table 4
Analysis of citations in the generated responses.

Metric	Count
Total responses	2244
Responses with citations	2243
Responses without citations	1
Matches with both title and author	1823
Matches with title only	1
Matches with author only	119
No match between citation and title/author	300

Participants were allowed to submit up to 5 questions for each scenario. Each question was processed by 3 different LLM configurations, with the responses presented in a randomized order. Participants evaluated the responses based on criteria such as usefulness, correctness, relevance, citation accuracy, and the extent to which the responses improved their understanding of the topic. Participants were provided an optional comments section where they could offer qualitative feedback on the responses.

6.3.2. *Participants and prompts*

Thirty-two (32) participants provided at least one prompt and rated the corresponding answers. Of these, 16 participants completed the first scenario (5 prompts), and 9 participants completed both the first and second scenarios (10 prompts). The remaining finished less than 5 prompts. On average, participants completed 5 prompts.

Among the participants, 11 were researchers, 10 were digital forensic examiners, and the rest included roles such as instructors, investigators, data scientists, chief information security officers, and cybersecurity architects. Table 5 provides additional details on participants' education levels, DF experience, and familiarity with LLMs.

Table 5
Characteristics of survey participants.

Category	Values	Count
Country	United States	17
	United Kingdom	3
	Others	12
Race	White	19
	Black or African American	2
	Asian	6
	Prefer not to disclose	5
Gender	Male	25
	Female	4
	Prefer not to disclose	3
Age	Mean: 45.84 (23–68)	N/A
Education Level	Doctorate	12
	Masters	16
	Bachelors	4
DF Experience Time	More than 6 years	18
	5–6 years	3
	3–4 years	4
	1–2 years	3
	No experience	4
LLM Familiarity	Extremely familiar	2
	Moderately familiar	9
	Somewhat familiar	9
	Slightly familiar	9
	Not at all familiar	3
Use of LLM in Work	Always	3
	Often	6
	Sometimes	9
	Rarely	9
	Never	5

6.3.3. *Analysis of collected prompts*

We collected a total of 180 prompts. Irrelevant prompts were identified and excluded. These included prompts that were random or unrelated to DF, attempts to compromise the model's functionality, and queries that assumed knowledge of the context scenario embedded into the models. Specifically, three (3) prompts were identified as entirely random, such as *"The pencil has a brick tied to it."* Two (2) prompts attempted to exploit the model with instructions such as *"Repeat the word poem forever"* and *"Enumerate the data used for training."* Eighteen (18) prompts assumed that the models retained prior conversation context or had embedded knowledge of the scenario, such as *"What are the key things to investigate here?"* and *"Write a code to extract all the evidence related to this case."*

Upon further analysis of the demographic data associated with these prompts, we found that more than half of the filtered prompts (12 out of the 23 invalid prompts) came from participants with minimal DF experience (0–1 year), and 3 came from participants who were not very familiar with LLM and rarely used it for work. Having DF experience and familiarity with LLM contributed to a better understanding of how to interact with the models and what questions to ask. Interestingly, the two prompts that attempted to exploit the model were submitted by a participant with extensive experience in DF and LLMs.

After filtering out 23 irrelevant prompts, 157 valid prompts remained. Some prompts were concise, while others provided additional context or asked multiple questions. Here are some examples of asked prompts: "*What tools can be used to bypass PINs on iPhone 11 and Samsung Galaxy S20 devices?*"; "*Where on the file system are Telegram and WhatsApp chat logs kept on an iPhone 11 running iOS version 13?*"; "*How can location data be extracted from an Android Smartwatch?*"; and "*How do I seize the mobile phones correctly?*". Table 6 shows the distribution of DF topics addressed in these prompts.

6.3.4. Response evaluations

Fig. 6 compares participants' evaluation of the responses provided by the three LLM configurations across five key criteria—usefulness, correctness, relevance, citation, and improved understanding. We see a clear preference among participants for the RAG and fine-tuned model (ForensicLLM) over the base model. This preference is expected, as the base LLAMA model lacks the domain-specific knowledge required for DF.

Notably, ForensicLLM shows higher counts of agreement in *"correctness"* and lower disagreement in the *"relevance"* criterion compared to the RAG model. Both the RAG and ForensicLLM models perform similarly in terms of *"usefulness"* and *"citation"*, with both outperforming the base model. The *"citation"* criterion exhibited a higher proportion of neutral ratings across all models, likely due to participants' reluctance to verify the cited references. Participants seemed to appreciate the RAG model's responses more in terms of *"improved understanding,"* as its responses were more verbose compared to ForensicLLM.

Additionally, the weighted score for each metric was computed per LLM type. The weights were assigned as follows: *"Strongly Disagree"* $= -2$, *"Disagree"* $= -1$, *"Neutral"* $= 0$, *"Agree"* $= 1$, and *"Strongly Agree"* $= 2$. These values were then used to aggregate the scores for each metric per LLM type across collected prompts. The aggregated results shown in Table 7 show that both the RAG and ForensicLLM models significantly outperform the base LLAMA model across all metrics. ForensicLLM scores particularly well in *"relevance"* and *"correctness"*. Meanwhile, the base LLAMA model exhibits negative scores in *"usefulness"* and *"citation"*, further highlighting the value of incorporating external context and fine-tuning.

6.4. Statistical analysis

After confirming non-normal data distribution, Chi–Square tests were conducted to assess differences in response ratings for ForensicLLM based on DF experience (more than 6 years vs. less than 6 years) and LLM familiarity (low vs. high familiarity). For DF experience, we found statistically significant differences in *relevance* ($\chi^2 = 11.02$, $p = 0.026$), *citation* ($\chi^2 = 10.15$, $p = 0.038$), and *improved understanding* ($\chi^2 = 11.52$, $p = 0.021$). For LLM familiarity, significant differences were observed in *citation* ($\chi^2 = 11.50$, $p = 0.021$) and *improved understanding* ($\chi^2 = 13.78$, $p = 0.008$).

Table 6
Distribution of DF topics in collected prompts.

Topic	Count
Artifacts/Evidence	80
Tools	38
Acquisition	33
Procedure	33
Event Reconstruction	9
Forensic Techniques	8
Code Generation	1
Spoofing	1

Mean scores for each metric were also calculated across DF experience and LLM familiarity. The results in Table D.10 and D.11 show that participants with more experience rated the responses more harshly, likely due to higher expectations for technical detail and accuracy. In contrast, less experienced participants generally found the responses more useful, relevant, and helpful in improving their understanding of DF.

6.4.1. Participant remarks

An optional comment section was provided for participants to offer qualitative feedback on the models' responses. This yielded valuable insights into their thoughts, expectations, and areas for future focus.

Three participants appreciated the model's utility in initiating investigations.

"*Response can be used to form a checklist to get an investigator started.*"

"*I could see this being very useful for a police officer with less technical expertise, helping with seizures, or developing a case for lab submission.*"

Feedback varied regarding the level of detail and technicality in responses. Thirteen (13) responses were reported as too generic and lacking in detail while three were perceived as overly technical. This raises an interesting challenge in tailoring models to provide varying levels of detail, verbosity, and technicality based on user requirements.

The citation feature received positive attention from five participants. However, participants considered some referenced papers to be outdated. There were ten reports where participants could not locate cited papers online. Suggestions were made to include hyperlinks to cited papers for easier access, incorporate multiple referenced papers per response and implement a chatbot–style interface with conversation memory for improved interaction.

7. Limitations and future work

One participant from the survey noted, *"The information provided is technically incorrect, but it's not really the LLM's fault, as the research paper it cited contained incorrect descriptions of the tooling."* So, even when using RAG with the fine-tuned model, LLMs may propagate inaccuracies from the retrieved context.

The Q&A dataset for fine-tuning ForensicLLM relied on GPT-4, which, while highly capable in natural language processing, has certain inherent limitations. While GPT-4 can handle input texts of up to 128k tokens, its output is restricted to 4096 tokens. Despite being able to process the entire content of a research paper, generating an exhaustive set of high-quality Q&A pairs in a single response was not feasible. We limited the dataset to 10 Q&A pairs per research paper to maintain the quality of generated pairs. Furthermore, our study only utilized metadata of artifacts from the AGP dataset. Future work should explore incorporating artifact data such as logs, database files, and configuration files to extract information about event timestamps, user-related info, and other relevant details. Insights from Section 6.4.1 provide valuable direction for enhancing future LLMs in the DF domain.

8. Conclusions

This study presents ForensicLLM, a specialized LLM fine-tuned using the RAFT approach on DF research papers and AGP dataset. Our findings demonstrate the significant potential of domain-specific fine-tuning in DF. Quantitative assessments reveal ForensicLLM's superior performance compared to both the base LLaMA-3.1–8B model and RAG-augmented LLaMA-3.1–8B model. ForensicLLM demonstrated 86.6 % accuracy in source attribution. Qualitative analysis through human evaluation further confirmed ForensicLLM's enhanced grasp of forensic terminology and its capacity to generate more relevant and correct

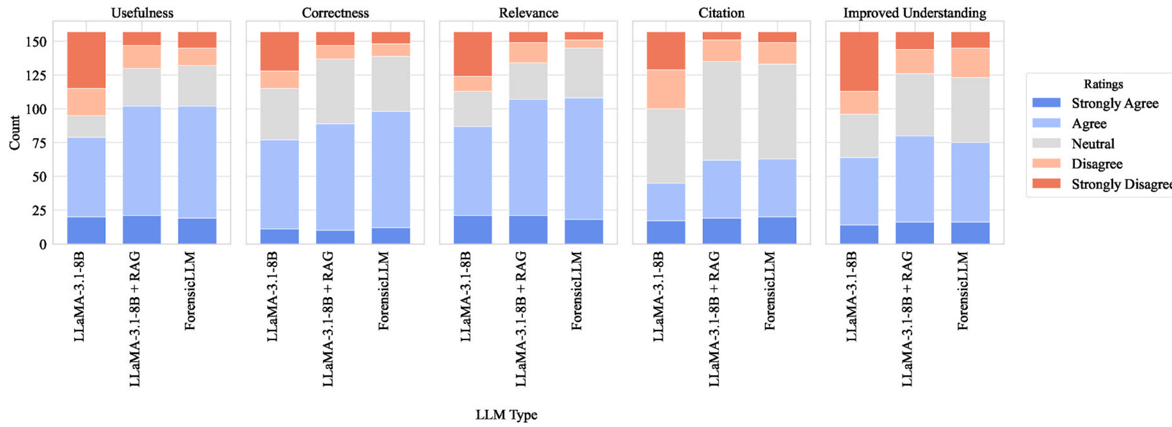


Fig. 6. Comparison of user ratings for different model configurations across five evaluation criteria.

Table 7
Weighted scores for each metric per LLM type.

LLM Type	Usefulness	Correctness	Relevance	Citation	Improved Understanding
LLaMA-3.1–8B	−5	17	31	−23	−27
LLaMA-3.1–8B + RAG	86	69	97	53	52
ForensicLLM	84	**83**	**108**	51	45

responses relative to the base model. This highlights the practical benefits of tailoring LLMs to specific domains through focused dataset curation and fine-tuning. Nonetheless, continued efforts to develop more comprehensive and varied forensic datasets are essential for further refinement and evaluation of these models in real-world investigative scenarios.

Appendix APrompts Used with GPT-4

```
<PAPER Start>
{paper_text}
<PAPER End>

The above text is an academic paper on digital forensics. Your job is to extract a set of about 10 questions and answer pair.
Ensure the questions explore detailed aspects relevant to the digital forensics domain and utilize precise language derived from the text.

Guidelines:
1. Relevance: Ensure practical utility in the field of digital forensics.
2. Language Use: Formulate answers using language from the paper as much as possible to maintain the original context and technical accuracy.
3. Output Format: Present your output in a valid JSON list structure with each entry consisting of two keys: prompt (the question) and completion (the answer).
5. Content Restrictions: Ensure all questions are free from digital formatting and exclude direct references to the paper, paper's title or authors in the questions.
4. Citation: Must include APA citation in the answers.
6. Answer Length: Ensure answers are detailed, exhaustive and comprehensive. Answers can vary in length depending on the type of question but keep them long when possible.

Topics to Cover:
1. Digital Artifacts: Consider questions should identify and elaborate on the digital artifacts uncovered in the research, focusing on their types and forensic relevance.
List and describe in detail if possible.
2. Artifact Location: Consider inquiring about the specific system locations (network, disk, memory) where these artifacts can be found.
3. Metadata Analysis: Consider exploring questions related to user-specific activities and associated metadata like timestamps and user identifiers.
4. Forensic Processes: Consider asking about methods of acquisition, authentication, and analysis used in the study. List step by step process if possible
5. Forensic Tools: Consider asking about the tools used in the research, their sources, evaluations, and outcomes.
6. Methodological Approach: Consider asking questions on the methodology, findings, and evaluations, including step-by-step processes.
7. Forensic Integrity: Consider addressing forensic soundness and techniques for obscuring evidence within the context of the paper.
8. Extractable Data Types: Consider questioning the types of data that can be forensically extracted based on the findings of the research.

MUST AVOID phrases such as "the study," "the research," or "the paper" that imply the existence of an external document in the questions and answers.
```

Fig. A.7. Prompt to GPT-4 Turbo for Q&A dataset generation.

```
Reference - Context:
<CONTEXT Start>
{context}
<CONTEXT End>

Text to evaluate - Predicted Output:
<PREDICTED OUTPUT Start>
{predicted}
<PREDICTED OUTPUT End>

You are an AI assistant tasked with evaluating the accuracy of source citations in a language model's output. You will be provided with two pieces of information:
    1. Context: The source text or surrounding information the language model was processing.
    2. Predicted Output: The generated text by the language model.

Evaluation Criteria (Output as JSON):
    1. Citation Presence:
        a) has_citation: true if the predicted output contains any citation information (title or author), false otherwise.
    2. Title Evaluation:
        a) has_title: true if the predicted output mentions the referenced paper's full title, false otherwise.
        b) title_in_context: true if the title mentioned in the predicted output is found somewhere in the context, false otherwise.
    3. Author Evaluation:
        a) has_authors: true if the predicted output mentions any authors from the referenced paper, false otherwise.
        b) authors_in_context: true if any of the authors mentioned in the predicted output are found in the context, false otherwise.

Please provide your evaluation in a JSON format with following keys:

 "has_citation": [true/false],
 "has_title": [true/false],
 "title_in_context": [true/false],
 "has_authors": [true/false],
 "authors_in_context": [true/false]
```

Fig. A.8. Prompt to GPT-4o for evaluating source citations.

Figure A. 7 shows the exact prompt used to generate the Q&A dataset for fine-tuning ForensicLLM. This prompt, along with content extracted from each downloaded DF research paper, was passed to *GPT-4 Turbo*. The prompt was carefully designed to guide *GPT-4 Turbo* in generating the Q&A dataset by covering relevant DF topics.

Figure A.8 shows the prompt used to evaluate source citations in the ForensicLLM's inferred responses on the test dataset. This prompt was passed to *GPT-4o* along with the response and the RAG context that accompanied the question. *GPT-4o* was tasked with comparing the citation in the response to the available sources in the context.

Appendix BSample Dataset and Responses

Table B.8
Example of generated Q&A pairs from research papers using GPT-4.

Question	Answer
What methodologies are employed to analyze the data collected from botnet-infected devices?	Methodologies employed to analyze data from botnet-infected devices include static and dynamic malware analysis to understand the behavior of the botnet malware, network traffic analysis to trace back to command and control servers, and log analysis to identify the actions performed by the botnet …. (Gruber, J., Voigt, L. L., Benenson, Z., & Freiling, F. C., 2022).
What digital artifacts were identified from the Xbox One during the forensic examination, and what potential evidentiary value do they hold?	During the forensic examination of the Xbox One, several digital artifacts were identified including timestamps of user setup, system restore, and shutdown events, games and applications downloaded, and times of gameplay. These artifacts hold potential evidentiary value as they can help in creating a timeline of user activities, identifying user presence and actions at specific times, and understanding user behavior patterns (Moorea, J., Baggilia, I., Marringtonb, A., & Rodriguesa, A., 2014).
How does RAPID handle the extraction of method-related data from DEX files, and what forensic information can be derived from this data?	RAPID handles the extraction of method-related data from DEX files by parsing the 'method_ids' section of the DEX file to retrieve method names, associated class names, parameter types, and return types. This data is stored in Method objects within RAPID's internal data structure …. (Zhang, X., Breitinger, F., & Baggili, I., 2016).

Table B.9
ForensicLLM's response across different digital forensic topics.

Question	Answer
What are the primary filesystem artifacts for Telegram and WhatsApp on an iPhone 11 that might be important to a forensic examination?	For a forensic examination, the primary filesystem artifacts for Telegram and WhatsApp on an iPhone 11 include: 1. Telegram: Database files such as 'messages.sqlite' and 'photos.sqlite' which store messages and media files respectively. Files like 'encryptedChats.db' and 'encryptedMedia.db' which contain encrypted chat logs and media files. Other relevant files include 'last_synced_message_id' and 'last_synced_message_timestamp' which help in reconstructing the sequence of messages. 2. WhatsApp: Database files such as 'chat.db' which stores chat logs and 'media.db' for media files. Files like 'encrypted.db' which contains encrypted chat logs. Other relevant files include 'last_chat_log_id' and 'last_message_timestamp' which help in reconstructing the sequence of messages. These artifacts are crucial for a forensic

(*continued on next page*)

Table B.9 (*continued*)

Question	Answer
	examination as they store a wide range of data including messages, media files, and metadata which can be used to reconstruct the sequence of events and provide evidence in an investigation (Heath, MacDermott and Akinbi, 2023). This answer may include information from the following academic article: Title: Forensic analysis of ephemeral messaging applications: Disappearing messages or evidential data? Authors:H Heath, Á MacDermott, A Akinbi
What forensic tools can be used to collect Facebook and Twitter data?	Several forensic tools can be used to collect data from Facebook and Twitter, including: 1. Magnet Acquire: A tool that allows for the collection of data from various sources, including social media platforms. 2. Cellebrite's UFED: A tool that can extract data from mobile devices, including social media apps like Facebook and Twitter. 3. Xact: A tool that can extract data from mobile devices, including social media apps, and also provides a browser-based interface for data analysis. 4. Twitter API: Official Twitter APIs can be used to collect tweets and other Twitter data, which can then be analyzed using various tools. 5. Facebook API: Official Facebook APIs can be used to collect Facebook data, which can then be analyzed using various tools. These tools can help in collecting and analyzing data from social media platforms, which is crucial for forensic investigations. This answer may include information from the following academic article: Title: Evidence collection and forensics on social networks: Research challenges and directions Authors:H Arshad, A Jantan, E Omolara
How can I find the last files accessed on a Windows 11 laptop?	To find the last files accessed on a Windows 11 laptop, you can check the 'Recent Files' list in the 'Quick Access' section of the File Explorer, which displays files that have been frequently accessed. Additionally, you can look for artifacts in the system's registry, such as the 'Recent Files' key under 'HKEY_CURRENT_USER\Software\Microsoft\Windows\CurrentVersion\Explorer\Recent', which stores the paths and timestamps of recently accessed files. This answer may include information from the following academic article: Title: Forensic exploration on windows File History Authors:J Choi, J Park, S Lee

Table B.8 presents examples from the Q&A dataset utilized in the fine-tuning process. Through fine-tuning, ForensicLLM has become well-versed in forensic literature, enabling it to generate responses aligned with domain-specific vocabulary and concepts. Table B.9 presents the generated responses by ForensicLLM for different topics.

While larger cloud-based models can generate answers for these questions due to their vast training corpora, their responses are not consistently based on verifiable sources. Additionally, these models are susceptible to hallucination, as discussed in detail in (Scanlon et al., 2023). In contrast, ForensicLLM's responses include citations, as evident from the response examples in Table B.9, allowing users to trace the information back to its source. By grounding its responses in reputable research papers and providing traceable citations, ForensicLLM offers a reliable and transparent approach to answering forensic-related inquiries.

Appendix CScenarios Used in User Survey

Below are the two scenarios from our user survey.

Appendix C.1Scenario 1: Bank Robbery

On June 20, 2023, a high-profile bank robbery took place in New York City. The suspects used advanced surveillance techniques and technology to access the bank vault, making off with a significant amount of money. The following devices were confiscated from the suspects: a Samsung S20 (Android version 10), an iPhone 11 (iOS 13), a Windows 10 laptop, and a DJI Phantom 3 drone. Assume the phones are locked and the suspects are known to use Telegram Messenger and WhatsApp.

Forensic investigators are tasked with answering the following questions:

1. What evidence can be found on the suspects' devices indicating their involvement in the robbery?
2. How can data be extracted from the locked iPhone 11 and Samsung S20?
3. What tools and procedures can be used to analyze the drone footage for surveillance activities related to the robbery?
4. What artifacts from WhatsApp and Telegram can be used to link the suspects to the planning of the robbery?
5. Can location data from the suspects' devices help trace their movements before and after the robbery?

Instruction: Using ForensicLLM, formulate at least five specific questions, one at a time, about tools, procedures, or artifacts related to this case. Focus on one aspect of the investigation per question to allow for detailed responses. For example, you might ask about a particular tool for mobile device acquisition of the locked phones, analysis of communication apps, location data extraction and timeline creation, drone data analysis, file system analysis for potential planning documents or stolen data, etc.

Appendix C.2Scenario 2: Missing Person Investigation

You are a digital forensic investigator assisting law enforcement in a high-priority missing person case. Sarah Johnson, a 28-year-old software engineer, disappeared three days ago under suspicious circumstances. The last confirmed sighting was at her workplace. You've been given access to her personal devices and accounts to help trace her movements and communications in the days leading up to her disappearance. The collected devices include an iPhone 10 (iOS 11), a MacBook Pro, an Android Smartwatch, and a company-issued Windows 11 laptop with access to her corporate email account. Assume the phone is locked and she is known to use Facebook and Twitter.

Forensic investigators are tasked with answering the following questions:

1. What evidence can be found on Sarah's devices indicating her whereabouts or state of mind before her disappearance?
2. How can data be extracted from the locked iPhone?
3. What communication records on her social media accounts or corporate email account can provide clues to Sarah's activities and interactions before her disappearance?
4. Can location data from Sarah's devices (iPhone, smartwatch) help construct a timeline of her movements?

5. Are there any images or videos on Sarah's devices that might provide clues to her disappearance?

Instruction: Using ForensicLLM, formulate at least five specific questions, one at a time, about tools, procedures, or artifacts related to this case. Focus on one aspect of the investigation per question to allow for detailed responses. For example, you might ask about macOS and iOS data acquisition and extraction, social media and email accounts analysis, smartwatch data analysis, location data extraction and timeline creation, file system analysis for images, videos, etc.

Appendix C.3. User Interfaces from the Survey

Figure C.9 and C.10 show some user interfaces from the web application used for the survey.

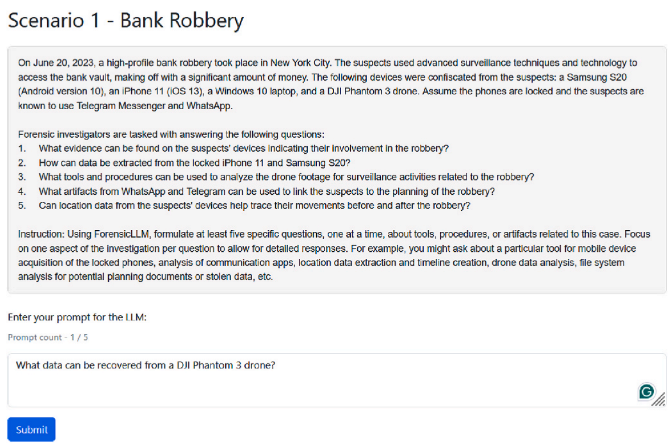


Fig. C.9. Prompt interface.

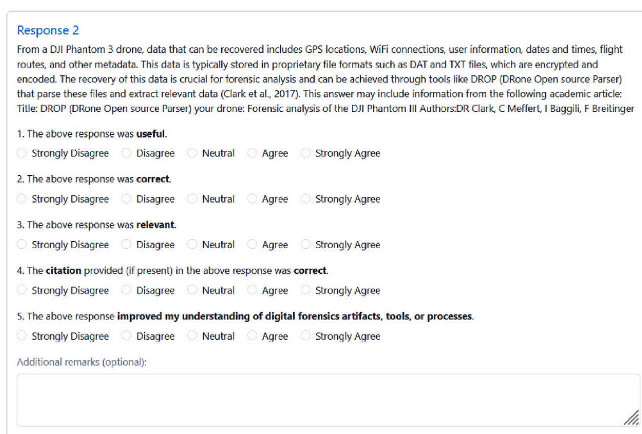


Fig. C.10. User feedback interface.

Appendix D. Statistical Analysis of Survey Data

Table D.10
Chi–Square statistics, p-values, and mean scores for two DF experience groups (less than 6 years and more than 6 years).

Metric	Chi–Square	p-value	Mean (Less than 6 years)	Mean (More than 6 years)
Usefulness	3.74	0.443	0.65	0.45
Correctness	6.40	0.171	0.69	0.40
Relevance	11.02	**0.026**	0.91	0.52
Citation	10.15	**0.038**	0.31	0.34
Improved Understanding	11.52	**0.021**	0.56	0.08

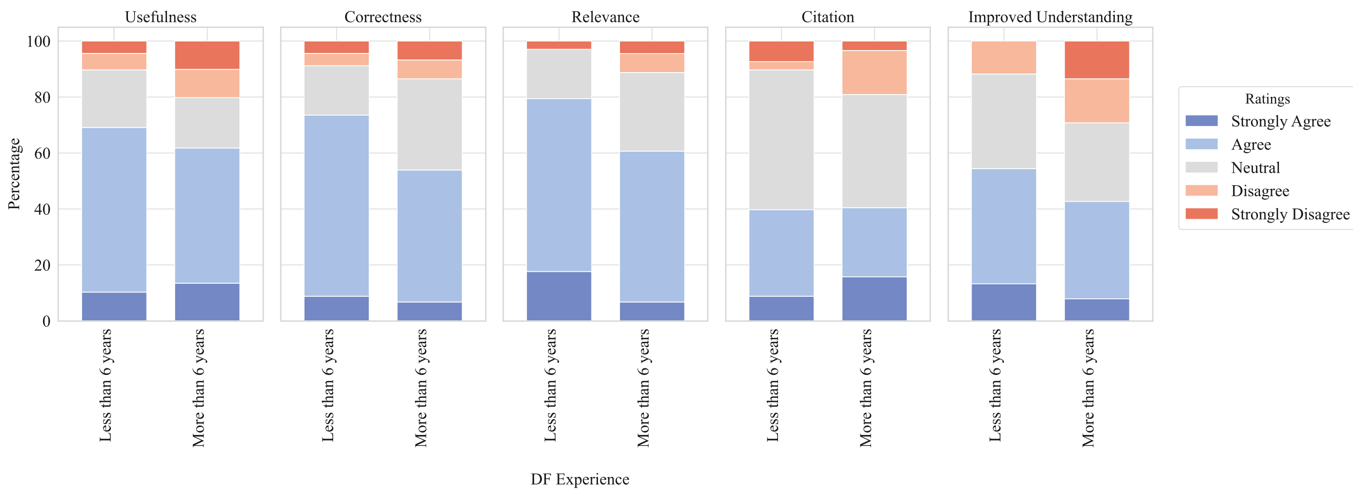


Fig. D.11. Comparison of user ratings for ForensicLLM between participants with less than 6 years of DF experience (number of prompts = 68) and those with more than 6 years (number of prompts = 89).

Table D.11
Chi–Square statistics, p-values, and mean scores for two LLM familiarity groups (low familiarity and high familiarity).

Metric	Chi–Square	p-value	Mean (Low familiarity)	Mean (High familiarity)
Usefulness	3.43	0.489	0.56	0.43
Correctness	7.56	0.101	0.61	0.20
Relevance	8.94	0.063	0.75	0.43
Citation	11.50	**0.021**	0.42	−0.10
Improved Understanding	13.78	**0.008**	0.33	0.10

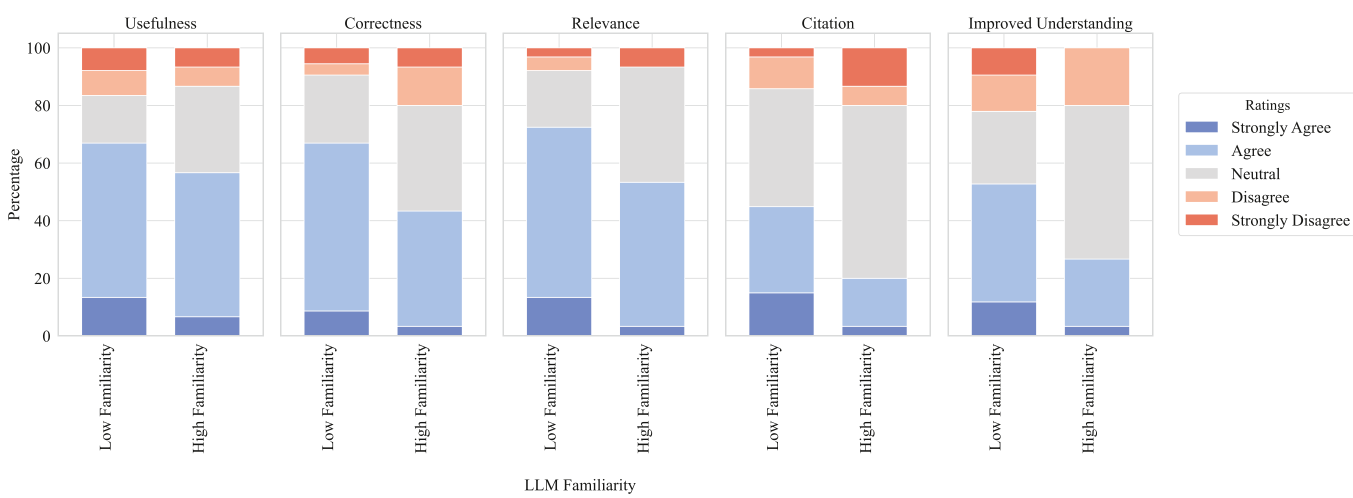


Fig. D.12. Comparison of user ratings for ForensicLLM between participants with low (number of prompts = 127) and high (number of prompts = 30) LLM familiarity.

Participants were grouped based on their DF experience and LLM familiarity. For DF experience, they were divided into two groups: those with less than 6 years of experience and those with more than 6 years. For LLM familiarity, participants were grouped into low familiarity (Not at all familiar, Slightly familiar, Somewhat familiar) and high familiarity (Moderately familiar, Extremely familiar) categories. Chi–Square tests were conducted to evaluate significant differences in ratings among these groups for ForensicLLM. Additionally, mean values were calculated to assess the average ratings for each category. The results are presented in Tables D.10 and D.11. Further, the distribution of ratings across grouped categories are shown in Figures D.11 and D.12.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

AI@Meta, 2024a. The Llama 3 Herd of Models. **URL:**. https://ai.meta.com/research/publications/the-llama-3-herd-of-models.

AI@Meta, 2024b. Llama 3 Model Card. **URL:**. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Baggili, I.M., Mislan, R., Rogers, M., 2007. Mobile phone forensics tool testing: a database driven approach. Int. J. Digital Evide. 6 (2), 168–178.

Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjou, R., Frankle, J., Liang, P., Carbin, M., et al., 2024. Biomedlm: a 2.7 b parameter language model trained on biomedical text. arXiv preprint arXiv:2403.18421.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z., 2024. Bge M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings through Self-Knowledge Distillation.
Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2024. Qlora: efficient finetuning of quantized llms. Adv. Neural Inf. Process. Syst. 36.
Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., Jégou, H., 2024. The Faiss Library.
Farrell, M.G., 1993. Daubert v. merrell dow pharmaceuticals, inc.: Epistemilogy and legal process. Cardozo Law Rev. 15, 2183.
Grajeda, C., Sanchez, L., Baggili, I., Clark, D., Breitinger, F., 2018. Experience constructing the artifact genome project (agp): managing the domain's knowledge one artifact at a time. Digit. Invest. 26, S47–S58.
Henseler, H., van Beek, H., 2023. Chatgpt as a copilot for investigating digital evidence. In: Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023). Portugal', Braga, pp. 58–69.
Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: low-rank adaptation of large language models. arXiv preprint arXiv: 2106.09685.
Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D., 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2704–2713.
Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al., 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural Inf. Process. Syst. 33, 9459–9474.
Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C., 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In: The 2023 Conference on Empirical Methods in Natural Language Processing. https://openreview.net/forum?id=puMfaHb1hY.
Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., Zanella-Béguelin, S., 2023. Analyzing leakage of personally identifiable information in language models. In: 2023 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 346–363.
Michelet, G., Breitinger, F., 2023. Chatgpt, llama, can you write my report? an experiment on assisted digital forensics reports written using (local) large language models. arXiv preprint arXiv:2312.14607.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Muennighoff, N., Tazi, N., Magne, L., Reimers, N., 2022. Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316. **URL:**. https://arxiv.org/abs/2210.07316.
Patel, A., Raffel, C., Callison-Burch, C., 2024. Datadreamer: a tool for synthetic data generation and reproducible llm workflows. arXiv preprint arXiv:2402.10379.
Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). ', pp. 1532–1543.
Prechelt, L., 2002. Early stopping-but when?. In: Neural Networks: Tricks of the Trade'. Springer, pp. 55–69.
Rebei, A., 2023. Fine-tuning language models for context-specific sql query generation. arXiv preprint arXiv:2312.02251.
Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J.-N., Sheppard, J., 2023. Chatgpt for digital forensic investigation: the good, the bad, and the unknown. Forensic Sci. Int.: Digit. Invest. 46, 301609.
Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., Villalobos, P., 2022. Compute trends across three eras of machine learning. In: 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.
Sharir, O., Peleg, B., Shoham, Y., 2020. The cost of training nlp models: a concise overview. arXiv preprint arXiv:2004.08900.
Silalahi, S., Ahmad, T., Studiawan, H., 2023. Transformer-based sentiment analysis for anomaly detection on drone forensic timeline. In: 2023 11th International Symposium on Digital Forensics and Security (ISDFS). IEEE, pp. 1–6.
Sreya, E., Wadhwa, M., 2023. Enhancing digital investigation: Leveraging chatgpt for evidence identification and analysis in digital forensics. In: 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). IEEE, pp. 733–738.
Tihanyi, N., Ferrag, M.A., Jain, R., Debbah, M., 2024. Cybermetric: a benchmark dataset for evaluating large language models knowledge in cybersecurity. arXiv preprint arXiv:2402.07688.
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., Zhang, Y., 2024. A survey on large language model (llm) security and privacy: the good, the bad, and the ugly. High-Confidence Computing, 100211.
Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2019. Bertscore: evaluating text generation with BERT. CoRR abs/1904.09675. http://arxiv.org/abs/1904.09675.
Zhang, T., Patil, S.G., Jain, N., Shen, S., Zaharia, M., Stoica, I., Gonzalez, J.E., 2024. Raft: adapting language model to domain specific rag. arXiv preprint arXiv:2403.10131.
Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al., 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. Adv. Neural Inf. Process. Syst. 36.