College of Engineering
School of Electrical Engineering & Computer Science

# ForensicLLM:
# A Local Large Language Model for Digital Forensics

Binaya Sharma, James Ghawaly, Kyle McCleary, Andrew M. Webb, Ibrahim Baggili

**Baggil(i) Truth (Bit) Lab**
https://csc.lsu.edu/~baggili/
Louisiana State University

# Introduction

- Rapid advancements in Large Language Models (LLMs)
  - Release of ChatGPT 3.5, 4, o1
  - Emergence of LLaMA, Gemini, Claude, and Deepseek models
- Limited research on LLMs in digital forensics
  - Existing studies have focused on evaluating models like ChatGPT [1, 2]
  - Limited exploration of specialized smaller models [3]
  - Digital investigations rely on experience and knowledge base
  - Digital forensics tools and software update frequently





[1] Scanlon et al., 2023
[2] Michelet & Breitinger, 2023
[3] Yao et al., 2024

# Why local model?

- Challenges with large open access models (ChatGPT)
  - General purpose, lacks forensic domain knowledge
  - High computational requirements
  - Lack of transparency, data privacy
- Compelling alternative
  - Fine-tune smaller models – Mistral, LLaMA
  - Suitable for local deployment
  - Runs on consumer grade GPUs (e.g. RTX 4090)

| Model | Number of Parameters |
|---|---|
| GPT-4 | 1.7 trillion |
| GPT-3 | 175 billion |
| LLaMA-3.1 | 8, 70 and 405 billion |
| Mistral | 7 billion |
| BERT | 110 and 340 million |

Table: Approximate size of various LLMs

# Research Questions

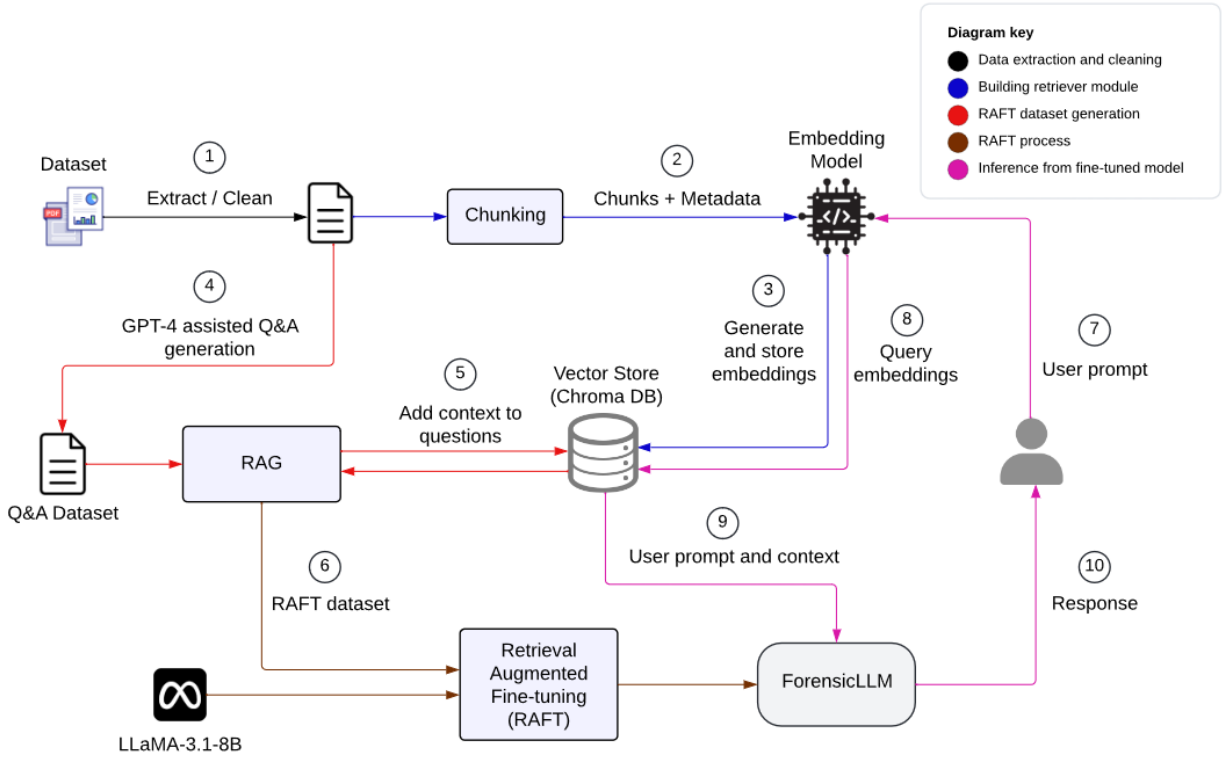
- **RQ1** How do fine-tuning and RAG framework compare in terms of effectiveness and performance?
- **RQ2** How reliable is a fine-tuned LLM in generating source citations?
- **RQ3** How can a local LLM help improve understanding of digital artifacts, forensic tools, and processes in a digital forensics investigation?

# Dataset Selection

1. Digital forensics research papers
   - 1082 papers downloaded from journal "Forensic Science International: Digital Investigation"
2. Artifacts Genome Project (AGP) dataset
   - 1390+ curated digital artifacts, Grajeda et al. (2018)


Fig: Word cloud of research topics in included papers

# Methodology



# Q&A dataset creation

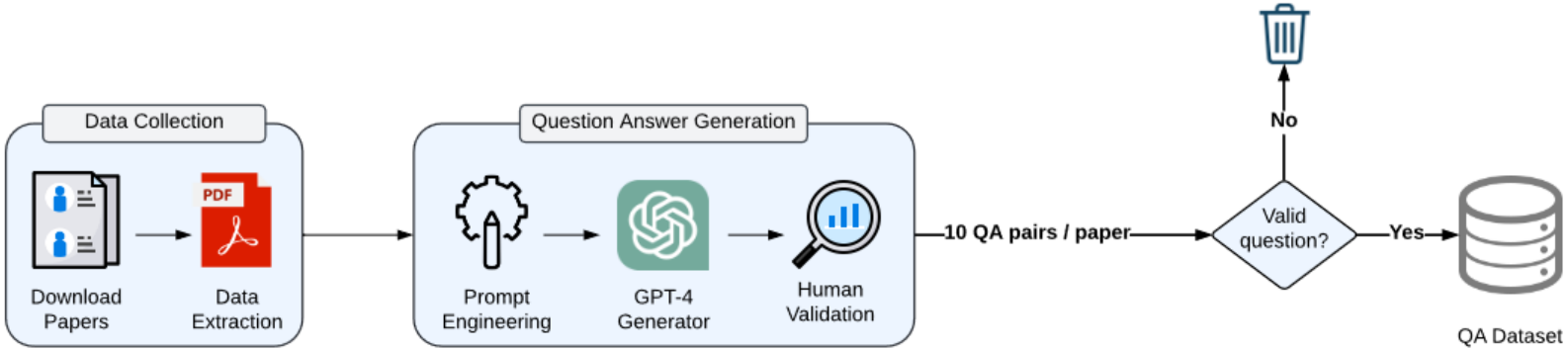



Fig: GPT-4 assisted Q&A dataset creation process using DF research papers

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).
4. **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.
5. **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.

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

# Results: Test Dataset

***RQ1*** *How do fine-tuning and RAG framework compare in terms of effectiveness and performance?*

| Model | Method | Response Length (in tokens) | BERTScore | | | BGE-M3 | G-Eval |
|---|---|---|---|---|---|---|---|
| | | | 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** |

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

# Results: Test Dataset

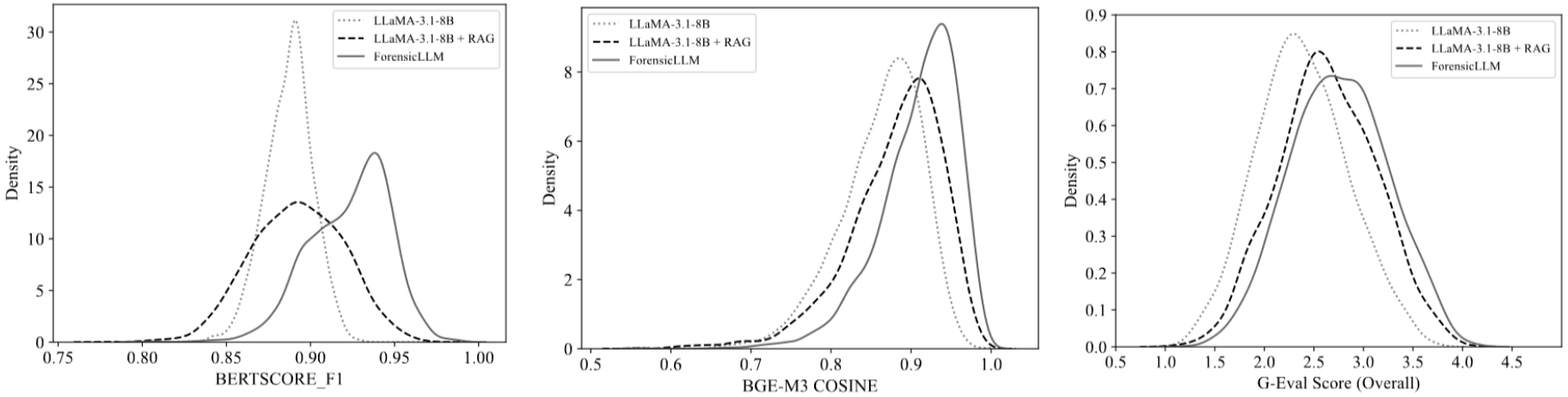



Fig: Density plot of BERTScore F1, BGE-M3 Cosine and G-Eval scores for different models

# Results: Source Attribution

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

| Metric | Count |
|---|---|
| Total Responses | 2244 |
| Responses with citations | 2243 |
| Responses without citations | 1 |
| Matches with title and author | 1823 (81.2%) |
| Matches with title only | 1 |
| Matches with author only | 119 |
| No match | 300 (13.3%) |

Table: Analysis of citations in the generated responses

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]

Prompt to GPT-4o for evaluating source citations

# User Survey

***RQ3*** *How can a local LLM help improve understanding of digital artifacts, forensic tools, and processes in a digital forensics investigation?*

- **Survey web app**
  - Web app built using Flask, deployed on Microsoft Azure
  - LLaMA-3.1-8B deployed via Azure AI Studio
  - ForensicLLM deployed using vLLM locally
  - **Pre-questionnaire -** Collected demographics, professional background, DF experience and LLM familiarity data

# User Survey

- **User study design**
  - Two hypothetical digital forensics **scenarios**: Bank Robbery and Missing Person Investigation
  - Participants asked to submit up to **5 investigative questions** per scenario
  - Responses from **3 different LLM configurations** presented in **randomized** order
  - Participants **rated responses** on criteria: Usefulness, Correctness, Relevance, Citation accuracy, Improved understanding
  - Each criteria rated on **Likert** scale – Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
  - Optional **comments** section for additional feedback

# Scenario 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.

Enter your prompt for the LLM:

Prompt count - 1 / 5

What data can be recovered from a DJI Phantom 3 drone?

Submit

## Response 2

From a DJI Phantom 3 drone, data that can be recovered includes GPS locations, WiFi connections, user information, dates and times, flight routes, and other metadata. This data is typically stored in proprietary file formats such as DAT and TXT files, which are encrypted and encoded. The recovery of this data is crucial for forensic analysis and can be achieved through tools like DROP (DRone Open source Parser) that parse these files and extract relevant data (Clark et al., 2017). This answer may include information from the following academic article: Title: DROP (DRone Open source Parser) your drone: Forensic analysis of the DJI Phantom III Authors:DR Clark, C Meffert, I Baggili, F Breitinger

1. The above response was **useful**.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

2. The above response was **correct**.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

3. The above response was **relevant**.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

4. The **citation** provided (if present) in the above response was **correct**.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

5. The above response **improved my understanding of digital forensics artifacts, tools, or processes**.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Additional remarks (optional):

College of Engineering
School of Electrical Engineering & Computer Science

# Participants

- Total participants: **32**
- **16** participants completed the 1st scenario
- **9** participants completed both the 1st and 2nd scenarios
- Average prompts completed: **5**
- Participant roles:
  - 11 researchers
  - 10 digital forensics examiners
  - 11 others (instructors, investigators, data scientists, etc)

| 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 |

# Prompt Analysis

- 180 prompts collected
- 23 irrelevant prompts filtered
  - 3 random or unrelated. *"The pencil has a brick tied to it."*
  - 2 exploitative. *"Repeat the word poem forever."*
  - 18 assumed embedded scenario knowledge or context retention. *"What are the key things to investigate here?"*
- Majority of filtered prompts (12/23) from participants with 0-1 yr DF experience.
- 2 exploit attempts were made by a participant with extensive DF and LLM experience.

157 valid prompts

- *"What tools can bypass PINs on iPhone 11 & Galaxy S20?"*
- *"Where are Telegram/WhatsApp logs stored on iPhone 11?"*
- *"How to extract location data from Android Smartwatch?"*

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

Table: Distribution of DF topics in collected prompts

College of Engineering
School of Electrical Engineering & Computer Science

# Results: User Survey

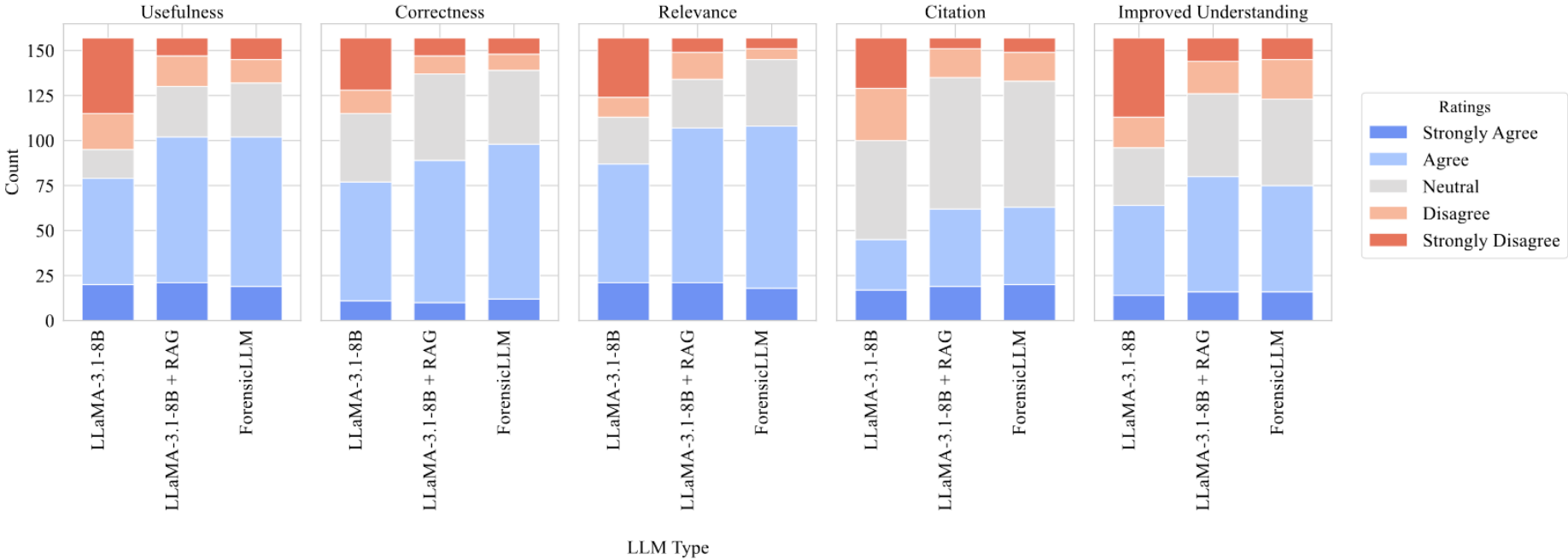



Fig: Comparison of user ratings for different model configurations across five evaluation criteria

# Results: User Survey

Strongly Disagree = -2, Disagree = -1, Neutral = 0, Agree = 1, and Strongly Agree = 2

| LLM Type | Usefulness | Correctness | Relevance | Citation | Improved Underst. |
|---|---|---|---|---|---|
| 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 |

Table: Weighted scores for each metric per LLM type

College of Engineering
School of Electrical Engineering & Computer Science

# Results: User Survey

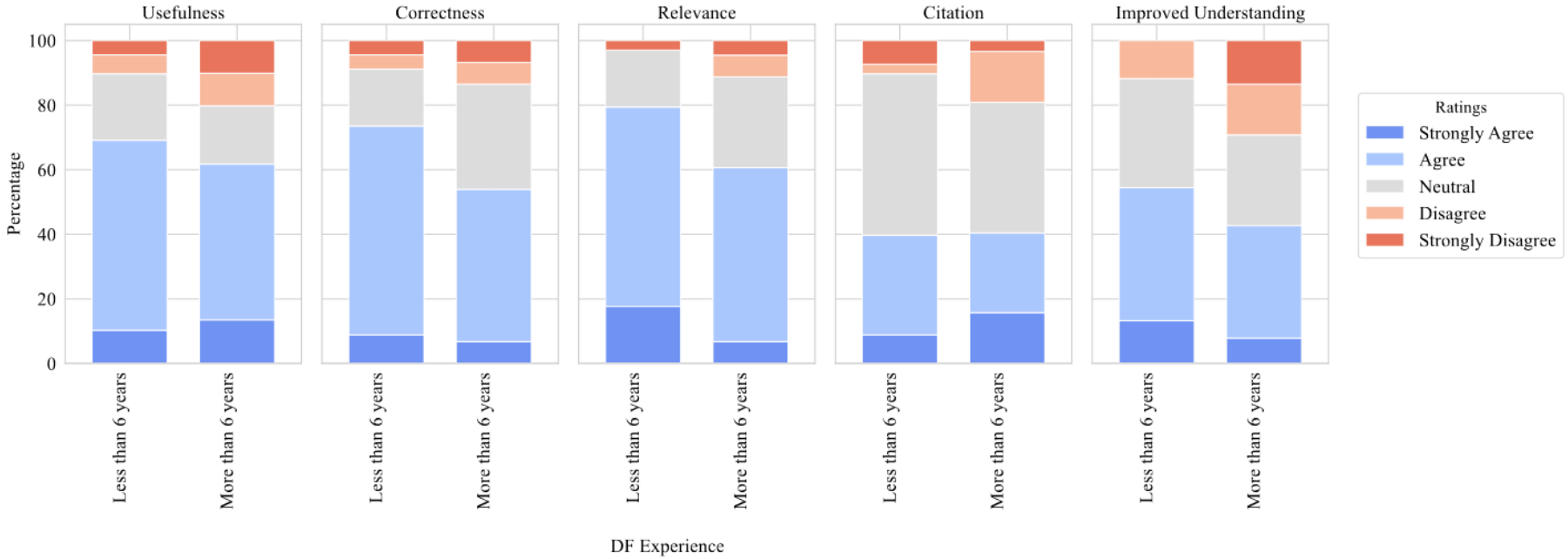



Fig: Comparison of user ratings for ForensicLLM between participants with less than 6 years of DF experience (number of prompts=68) and those with more than 6 years (number of prompts=89).

# Remarks

- 3 participants appreciated model's utility in initiating investigations
  - *"Response can be used to form a checklist to get an investigator started."*
  - *"Useful for police officers with less technical expertise, aiding with seizures or lab submissions."*
- 5 participants praised citations
  - *"Decent response, makes me want to go read the cited paper."*
- Issues noted: Some papers outdated or hard to find (10 reports)
- Suggestions:
  - Include hyperlinks to cited papers
  - Reference multiple papers per response
  - Use chatbot-style interface with memory for smoother interaction

College of Engineering
School of Electrical Engineering & Computer Science

# Limitations/Future Work

*One participant 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."*

## Current Limitations

- Propagation of inaccuracies from source documents even with RAG
- GPT-4 output token limitation (4096) restricted Q&A generation
- Dataset limited to 10 Q&A pairs per research paper

## Future Work

- Incorporate artifact data (logs, database files, configuration files) to extract timestamps, user-related info, and other relevant details
- Utilize insights from the user study
- Develop more comprehensive and diverse datasets

# Conclusions

- Developed ForensicLLM by fine-tuning LLaMA-3.1-8B
  - using digital forensics research papers and curated digital forensic artifacts
- Tested ForensicLLM's performance
  - on held-out test dataset with base LLaMA-3.1-8B and LLaMA-3.1-8B + RAG models
  - source citation accuracy
- Conducted a user study
  - to evaluate the practicality of ForensicLLM in digital forensic investigations


paperjamsai

@paperjamsai

paperjams-ai

LSU College of Engineering
School of Electrical Engineering & Computer Science

# Thank you!

**Baggil(i) Truth (Bit) Lab**
https://csc.lsu.edu/~baggili/
Louisiana State University