Universität Augsburg
Fakultät für Angewandte Informatik

Fine-Tuning Large Language Models for Digital Forensics:

Case Study and General Recommendations

Gaëtan Michelet, Hans Henseler, Harm van Beek, Mark Scanlon, and Frank Breitinger

IMF – 16.09.2025

About me

Gaëtan Michelet

- Ph.D Student - University of Augsburg
- Supervised by Frank Breitinger
- gaetan.michelet@uni-a.de

A big thanks to:

- The Netherlands Forensic Institute (NFI)
- The University of Lausanne (Mobi.Doc)

Identified gap

Large Language Models (LLMs) for Digital Forensics (DF)

- Discussion of their capabilities for DF
- Tests of ready-to-use LLMs
- ForensicLLM
  - LLM Fine-tuned for DF purposes (but not for a specific task)

Can we do it for a specific task?

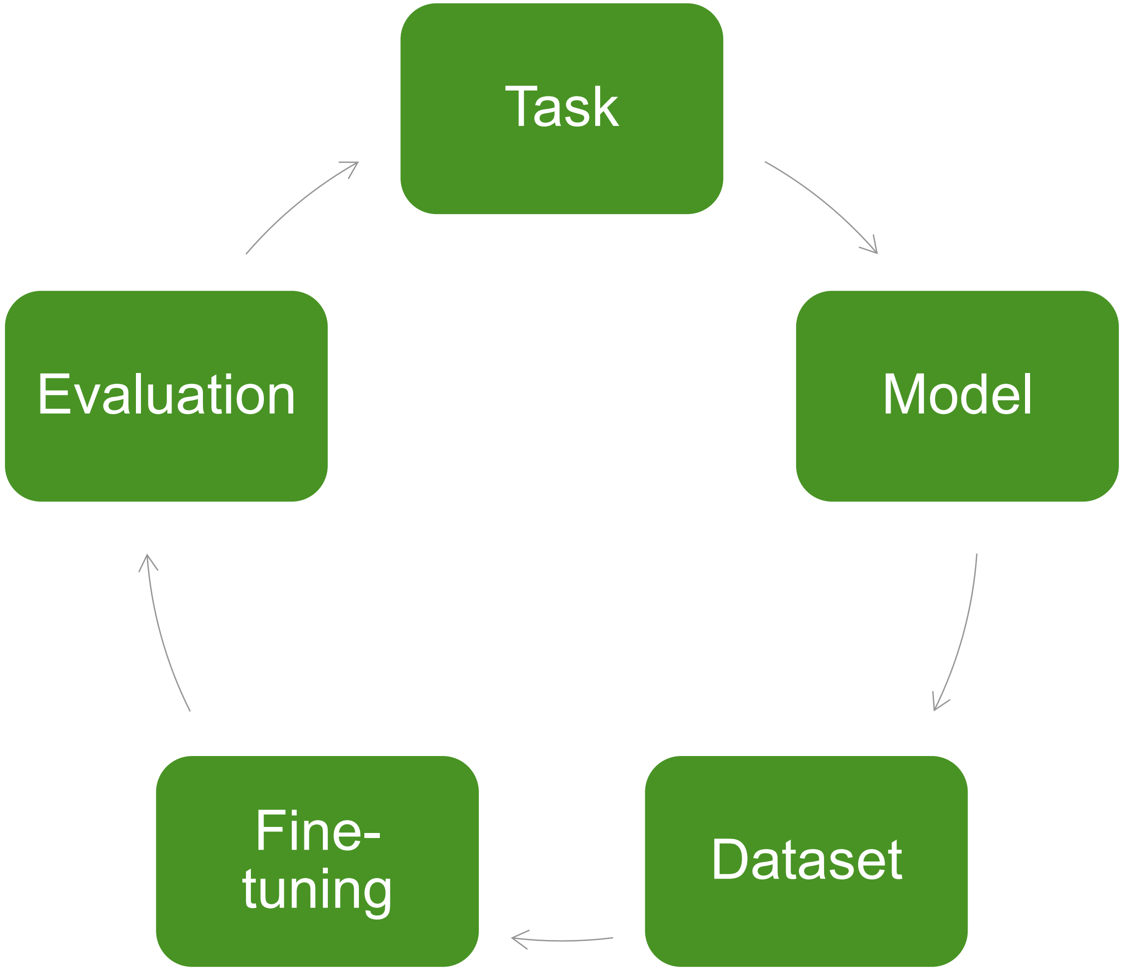
RQ and contributions

1. How can small local LLMs be fine-tuned to address the unique challenges and requirements of digital forensics?
2. How can we evaluate the results to ensure that the fine-tuning process was successful and identify the best-performing model?
3. How beneficial is the fine-tuning process for improving model performance and should practitioners adopt it in forensic workflows?

Contributions:

- Recommendations
- Case study
- Insights, models and datasets

The fine-tuning process



Task

Recommendations

Time-consuming and common

Difficult to automate using «regular» automation

Already tested on LLMs

Choices

Chat summarization tasks

Three levels of complexity

Previoulsy tested on GPT 3.5 and Llama 2-13B

Task 1

Messages:
Alice: Hi Bob, did you complete the report?
Bob: Not yet, I’ll finish it by tonight.
Alice: Please make sure it’s done before the meeting tomorrow.

Please provide a detailed summary of the conversation.

Alice asked Bob about the completion of the report. Bob replied that he would finish it by that night. Alice reminded him to have it done before the meeting the next day.

Task 2 (1/2)

Messages:
[10:00] Alice: Hey, are we still on for lunch?
[10:05] Bob: Yes, see you at 12.
[11:00] Charlie: Don’t forget the meeting at 3 PM.
[11:15] Alice: Thanks for the reminder.

Please provide short summaries of each topic discussed, including the timestamps of the relevant messages.

[10:00 - 10:05] Alice and Bob confirm lunch plans.
[11:00 - 11:15] Charlie reminds Alice about the 3 PM meeting, and Alice acknowledges.

Task 2 (2/2)

Please provide a detailed summary of the topic starting at 11:00 and ending at 11:15.

Charlie informed Alice about a meeting scheduled at 3 PM. Alice thanked Charlie for the reminder.

Task 3

Crime Investigated: Unauthorized Access
Messages:
[09:00] Dave: Did you get into the system?
[09:05] Eve: Yes, I bypassed the firewall.
[09:10] Dave: Excellent. Download the files and delete the logs.
[09:15] Eve: Will do.
[10:20] Dave: By the way, are you coming to the office party tonight?
[10:25] Eve: Yes, looking forward to it!
[10:30] Dave: Great, see you there.

Please provide a detailed summary related to the crime of Unauthorized Access.

The topic of interest for the investigation started at 09:00 and ended at 09:15. Eve informed Dave that she successfully bypassed the firewall to access the system. Dave instructed her to download the files and delete the logs, indicating activities related to unauthorized access.

Model

Recommendations	Choices
Size and version	Llama 3.1-8B-Instruct
Recently released in open weight	Gemma 2-2B-Instruct
With information about training process and dataset	Mistral 7B-Instruct-v0.3
	To compare + generalize

Datasets

Requirements

Sample of quality in quantity

Open access with information about creation

If none available, it must be created

Choices

Training with 60 / 120 / 180 samples

Testing with 36 samples

Combination of GPT4 generated and SAMSum samples

Datasets

Popular chat summarization datasets (SAMSum…)

- Too short and not related to crimes

GPT 4 generated chats (single topic)

- Chat (crime or chitchat)
- GPT 4 detailed + short summary
- Manual detailed + short summary

Mixed with SAMSum for task 2 and 3 (No SAMSum sample in the testing dataset)

Fine-tuning

Requirements

According to available resources

Loss computation method

Ideally several configurations

Choices

SFT with QLoRA and cross-entropy

Training batch of 8 and 16

Two variations of loss computation:
«answer only» and «prompt + answer»

Evaluation

Requirements	Choices
Auto and/or Manual	ROUGE-1, ROUGE-2, ROUGE-L
Standard and/or Custom	BLEU-1, BLEU-2
	BERTScore, RoBERTaScore
	No manual evaluation

Results

Combining all the variables (tasks, datasets, models, configurations)

- 216 fine-tuned models

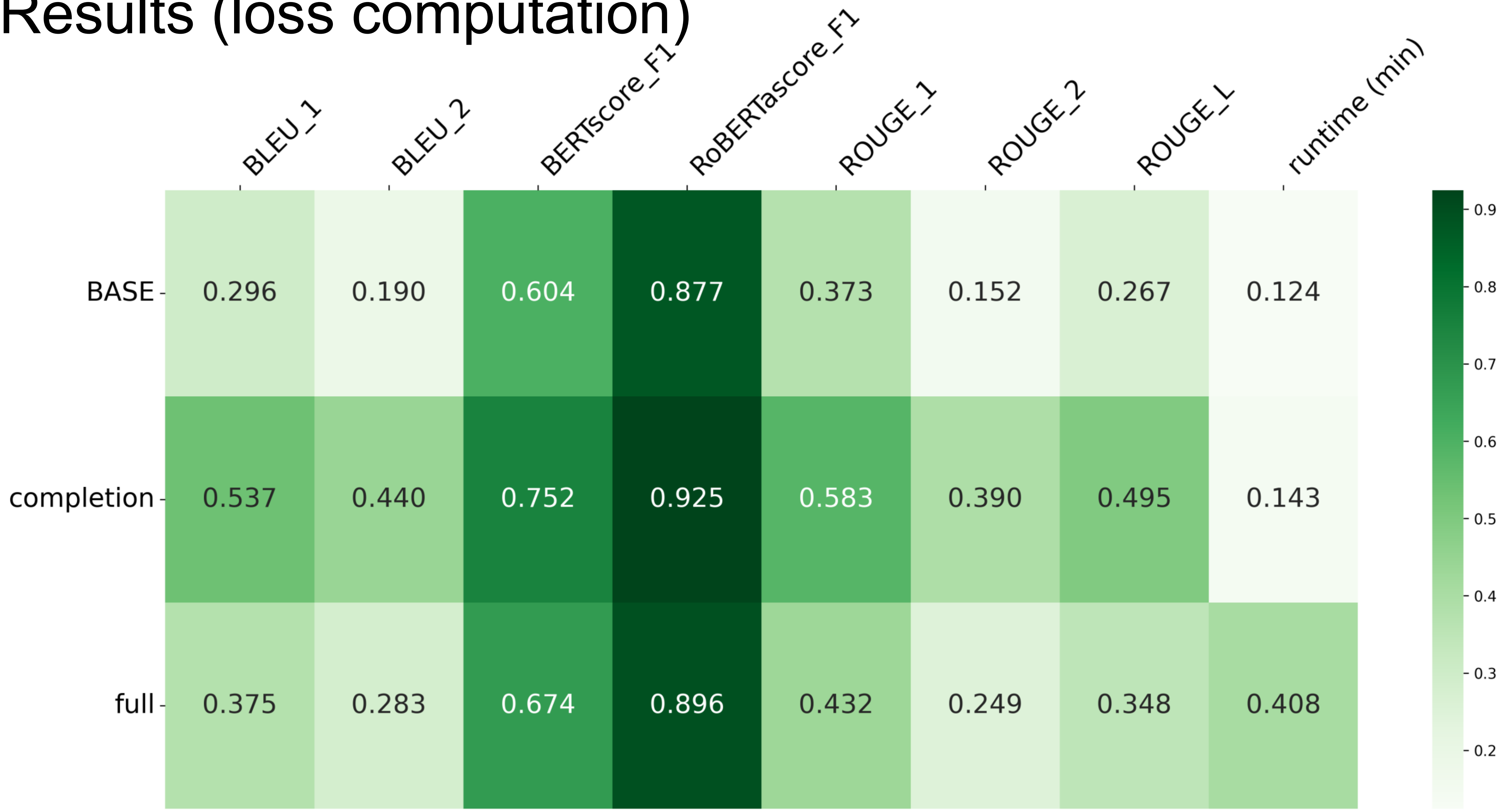
Figures displayed

- All models fine-tuned on the « answer only » loss
- Comparison against the manually generated testing samples

Keep in mind

- Models trained on automatic/manual performed better when compared to automatic/manual (respectively)

Results (loss computation)



Results (Fine-tuned vs Base models)

base_model-type_model	BLEU_1	BLEU_2	BERTscore_F1	RoBERTascore_F1	ROUGE_1	ROUGE_2	ROUGE_L	runtime (min)
('Llama-3.1-8B', 'BASE')	0.316	0.203	0.615	0.879	0.388	0.160	0.272	0.171
('Llama-3.1-8B', 'FT')	0.557	0.455	0.767	0.927	0.599	0.403	0.506	0.176
('Mistral-0.3-7B', 'BASE')	0.300	0.196	0.608	0.880	0.373	0.154	0.276	0.125
('Mistral-0.3-7B', 'FT')	0.549	0.450	0.758	0.926	0.593	0.396	0.502	0.145
('gemma-2-2b', 'BASE')	0.273	0.170	0.589	0.873	0.359	0.143	0.252	0.077
('gemma-2-2b', 'FT')	0.504	0.416	0.732	0.921	0.557	0.372	0.476	0.107

Results (Nb of samples)

	BLEU_1	BLEU_2	BERTscore_F1	RoBERTascore_F1	ROUGE_1	ROUGE_2	ROUGE_L	runtime (min)
BASE	0.296	0.190	0.604	0.877	0.373	0.152	0.267	0.124
60-samples	0.517	0.423	0.742	0.922	0.568	0.377	0.482	0.131
120-samples	0.537	0.442	0.750	0.924	0.581	0.391	0.494	0.140
180-samples	0.556	0.456	0.764	0.927	0.600	0.403	0.508	0.157

Discussion

Improvements are there

- Loss computation is important
- Small number of samples is sufficient
- Dataset can be (partially) synthetic

Insights

Difficulties encountered

- Preliminary tests
- Lack of datasets in DF
- Guided by computational resources

Our opinion

- Very costly
- Not beneficial yet (LLMs keep evolving)

Limits and future work

1. **Task**
   Testing tasks unrelated to summarization
2. **Model**
   Testing bigger/smaller models
3. **Dataset**
   Using more samples with complex chats
4. **Fine tuning**
   Testing other hyper-parameters
5. **Evaluation**
   Evaluating manually

Conclusion

Fine-tuning LLMs for DF tasks is possible

- Improvements
- With a small dataset

Not sufficiently beneficial given the costs

Future research should focus on running more tests