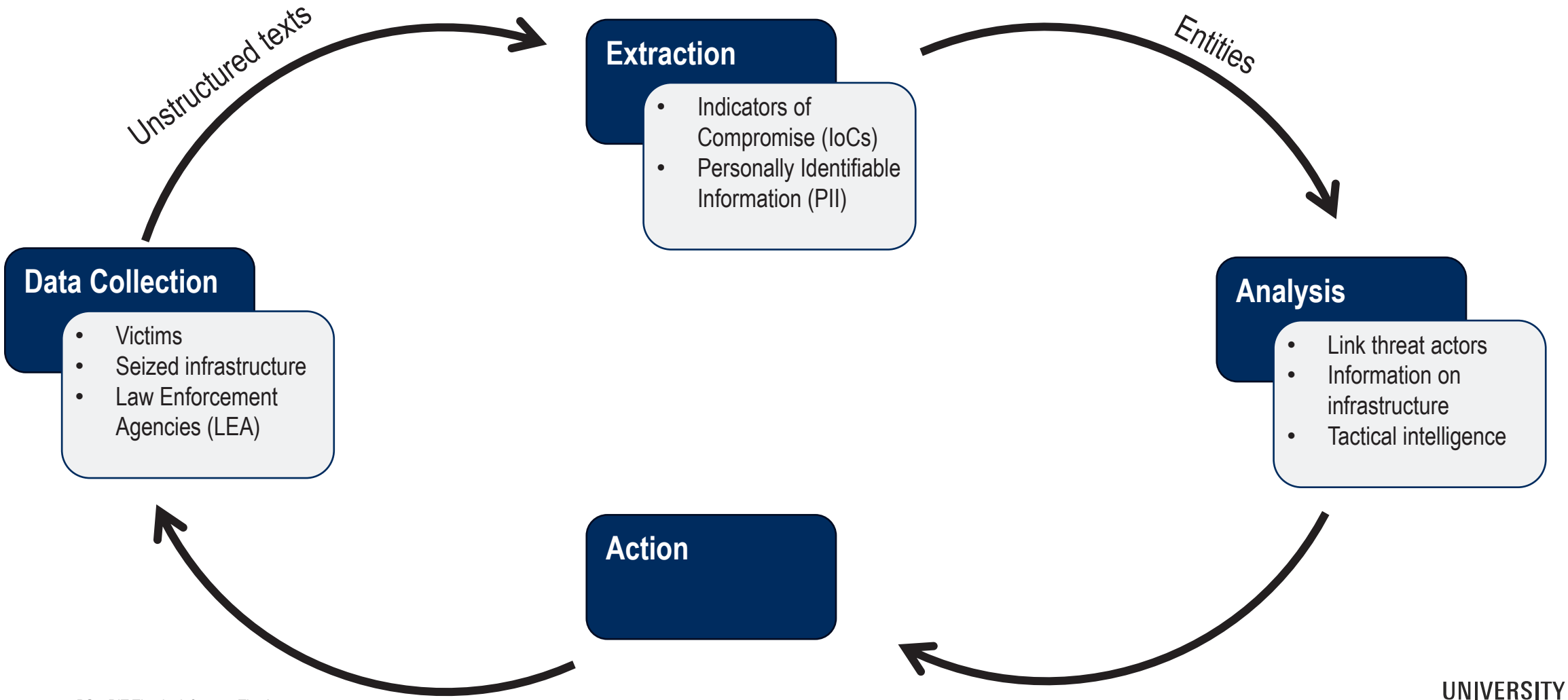
# NAMED ENTITY RECOGNITION IN CYBERCRIME INTELLIGENCE ANALYSIS: A HYBRID LLM APPROACH TO REDUCE MANUAL ETL WORKLOAD

TIM ANGEVARE – S2744007



UNIVERSITY OF TWENTE.

# CYBERCRIME INVESTIGATION INTELLIGENCE CYCLE

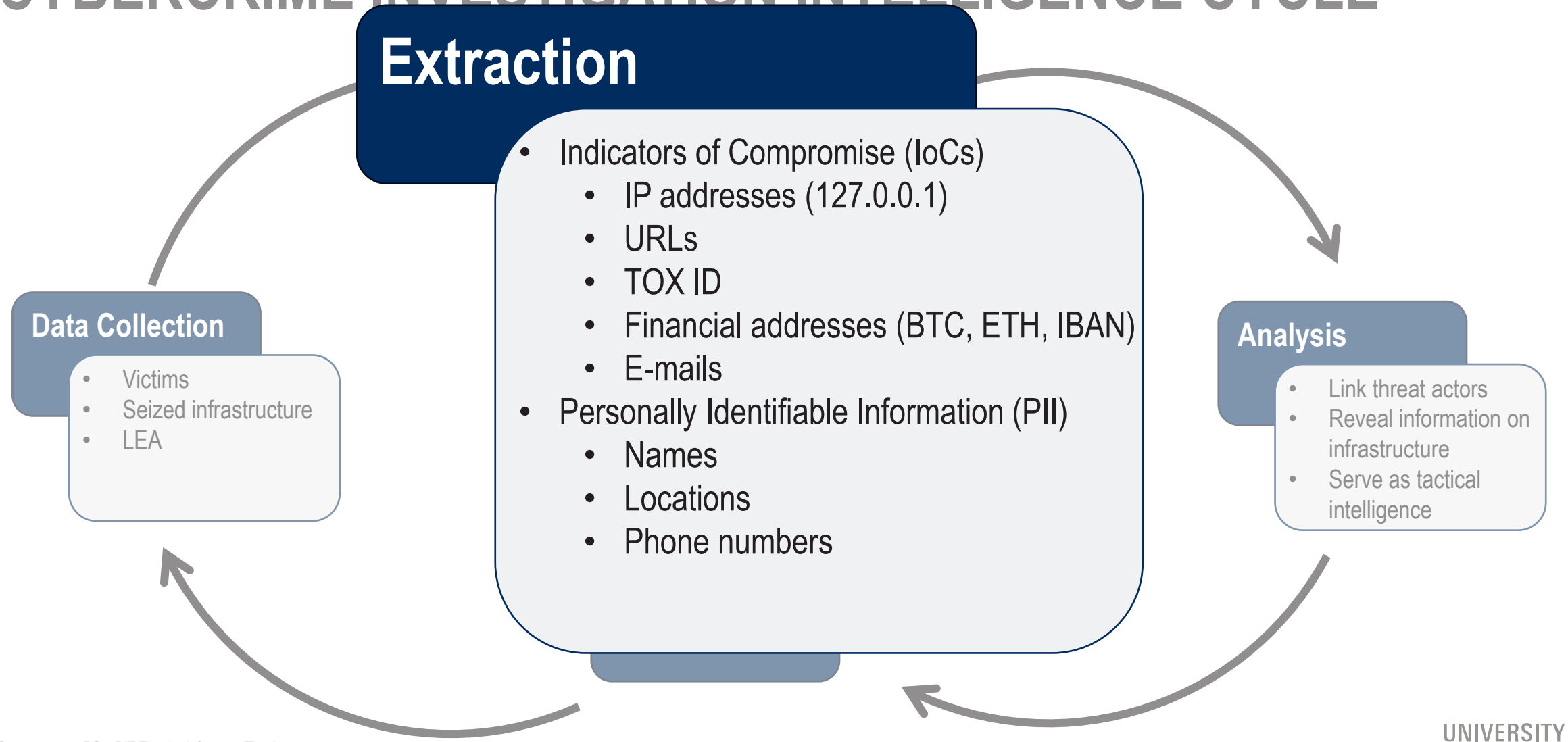





UNIVERSITY OF TWENTE.

[1]:Europol. 2025. *Cyber Intelligence – Sharing knowledge, know-how and updates to enhance the fight against cybercrime*. https://www.europol.europa.eu/howwe-work/services-support/intelligence-analysis/cyber-intelligence Content type:page, last updated 18 Aug 2025

# CYBERCRIME INVESTIGATION INTELLIGENCE CYCLE



# CRITICAL BOTTLENECK

## Increasing amount of data

- 33% increase in financial losses caused by cybercrime (USA, 2024) [2]
- modern devices and cloud storage confiscated in cybercrime investigations generate data volumes that exceed the analytical and storage capacity of many agencies. [3]

## Resource constraints

- INTERPOL Africa cyberthreat report [4]
  - 90% respondents: LEA need some or significant improvement
  - 95% of countries: resource constraints as limitation
- A lack of resources to support all requested investigations at Europol [5]

## Complexity of data

- Various formats
  - +31-70-356-1220
  - +31 (70) 3561220
  - 0703561220
  - <a href="tel:+31703561220">
- “Defanging” techniques
  - hxxps[:]//anonfiles[.]io
  - 91[.]198[.]174[.]192



UNIVERSITY OF TWENTE.

[2]: Internet Crime Complaint Center. 2024. *Internet Crime Report 2024*. Technical Report. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf [3]: Alexis B. Grochmal. 2025. *Challenges Faced by Law Enforcement Collecting and Using Digital Evidence in Cybercrime Investigations*. Ph. D. Dissertation. Marymount University. https://www.proquest.com/dissertations-theses/challengesfaced-law-enforcement-collecting-using/docview/3193985896/se-2 [4]: INTERPOL. 2025. *INTERPOL AFRICA CYBERTHREAT ASSESSMENT REPORT 2025*. Technical Report. INTERPOL. https://www.interpol.int/content/download/23094/ file/25COM009248%20-%20Cybercrime_Africa%20Cyberthreat%20Assessment%20Report_Design_2025-05%20v11.pdf [5]: Europol. (2025, December 18). The future of Europol [Audio podcast episode]. In The Europol Podcast. Spotify

# CRITICAL BOTTLENECK

## Resource constraints

- INTERPOL Africa cyberthreat report [4]
  - 90% respondents: LEA need some or significant improvement
  - 95% of countries: resource constraints as limitation
- A lack of resources to support all investigations at Europol [5]

**Manual extraction is not a solution**

## Complexity of data

- Various formats
  - +31-70-356-1220
  - +31 (70) 3561220
  - 0703561220
  - <a href="tel:+31703561220">
- “Defanging” techniques
  - hxxps[:]//anonfiles[.]io
  - “91[.]198[.]174[.]192

**Current pattern matching & NLP methods are inaccurate**

# CRITICAL BOTTLENECK

**What about Large Language Models (LLMs)?**

## Resource constrai

- INTERPOL Africa cyberthr
  - 90% respondents: LEA ne significant improvement
  - 95% of countries: resource constraints as limitation
- A lack of resources to support all investigations at Europol

## Complexity of data

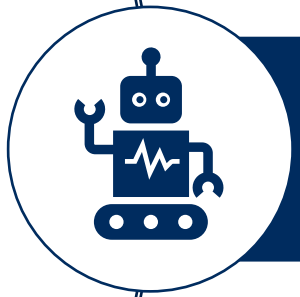
- formats
  - 31-70-356-1220
  - +31 (70) 3561220
  - 0703561220
  - <a href="tel:+31703561220">
- "Defan chniques
  - h les[.]io
  - "192

**Manual Extraction not a solution**

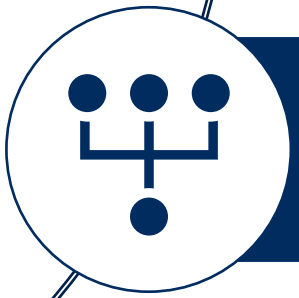
**Current pattern matching & NLP methods inaccurate**

# RESEARCH QUESTIONS

RQ1: How do pattern-based methods and Natural Language Processing (NLP) models perform across IoC and PII entity types in unstructured cybercrime texts?



RQ2: Under what conditions can LLMs achieve competitive or superior performance in terms of precision, recall, and F1-scores compared to NLP methods for entity extraction from unstructured cybercrime texts?



RQ3: How can we integrate NLP methods with LLMs to optimize IoC and PII extraction accuracy for operational LEA contexts?

# **EXPERIMENT 1:** PATTERN-BASED METHODS AND NLP MODELS.

# METHODOLOGY

- Literature review
- Test on synthetic dataset of 13 E-mails created by senior Europol analysts
- Count number of True Positives (TP), False Positives (FP), and False Negatives (FN)
- $precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$
- 1.00 highest score

Bob sent 1.2 BTC to Tim from the adress:
1A1zP1eP5QGefi2DMPTfTL5SLmv7DivfNa

```
{
'entities' : [
{'entity': 'BTC','type' : 'BTC', 'start_pos': 13, 'end_pos': 16}
{'entity': 'Tim', 'type' : 'PERSON', 'start_pos': 20,
'end_pos': 23 }
  ]
}
```

# RESULTS: PATTERN-BASED METHODS AND NLP MODELS

| Entity type | RegEx | SpaCy | GLiNER-PII |
|---|---|---|---|
| Person | 0.0 | 0.59 | 0.93 |
| E-mail | 0.89 | 1.0 | 1.0 |
| Location | 0.0 | 0.63 | 0.8 |
| BTC | 1.0 | 0.0 | 0.2 |
| IP | 0.75 | 0.0 | 0.75 |
| TOX ID | 1.0 | 0.0 | 1.0 |
| URL | 0.67 | 0.28 | 0.56 |
| Phone | 0.75 | 0.0 | 1.0 |

F1 score on synthetic e-mail dataset n=13

# **EXPERIMENT 2:** LLM-BASED EXTRACTION

# METHODOLOGY

- 19 open-source Large Language Models
  - Various origins, families, sizes
  - Fine-tuned and generalized
- GPU cluster at University of Twente
- Python with Jupyter Notebooks and Ollama
- Fine-tuning
  - Prompt optimization
  - Context chunking
  - Low Rank Adaptation (LoRA) fine-tuning

# OPTIMIZING THE PROMPT

- Role and Task
- Role, Task and Format
- One-shot example
- Chain of Verification

You are a cyber intelligence analyst with 20 years of experience in the field.

Your task is to extract any entity from the input text. For each entity found you MUST indicate the type in UPPERCASE. ONLY extract entities if literal entity is present in input text.
The expected entity types are the following:
- EMAIL: email addresses format (user@domain.tld)
- IP: IP addresses (IPv4 x.x.x.x or IPv6)
- BTC: ONLY Bitcoin wallet addresses (26-35 alphanumeric, starting with 1, 3, or bc1) EXCLUDE the word bitcoin or values (for example 2.0 BTC)
- IBAN: iban bank account number
- PERSON: Human names (John Smith, John, Catalina) EXCLUDE initials (for example A.H.)
- LOCATION: cities, countries, geographic locations
- PHONE: phone numbers in any format
- URL: URLs and web addresses EXCLUDE filenames
- TOX: Tox messenger IDs

**Output**:
The output MUST be in a JSON object with key 'entities' and the value a list of dictionaries including every entity found. For each entity you MUST indicate the type in UPPERCASE.

**OUTPUT EXAMPLE**:

```
{
  "entities": [
    {"entity": "target123@darkmail.org", "type": "EMAIL"},
    {"entity": "10.45.67.89", "type": "IP"},
    {"entity": "Thompson", "type": "PERSON"},
    {"entity": "Helsinki", "type": "LOCATION"},
    {"entity": "Tim", "type": "PERSON"}
  ]
}
```

**Verification**:
1. verify that your answer in valid JSON format.
2. verify that all extracted entities are present in the input text.
3. verify that no entities from the example or system prompt are included in your answer.
4. verify that extracted entities match the expected formats for their types.
5. provide your final revised answer based on the verifications above.

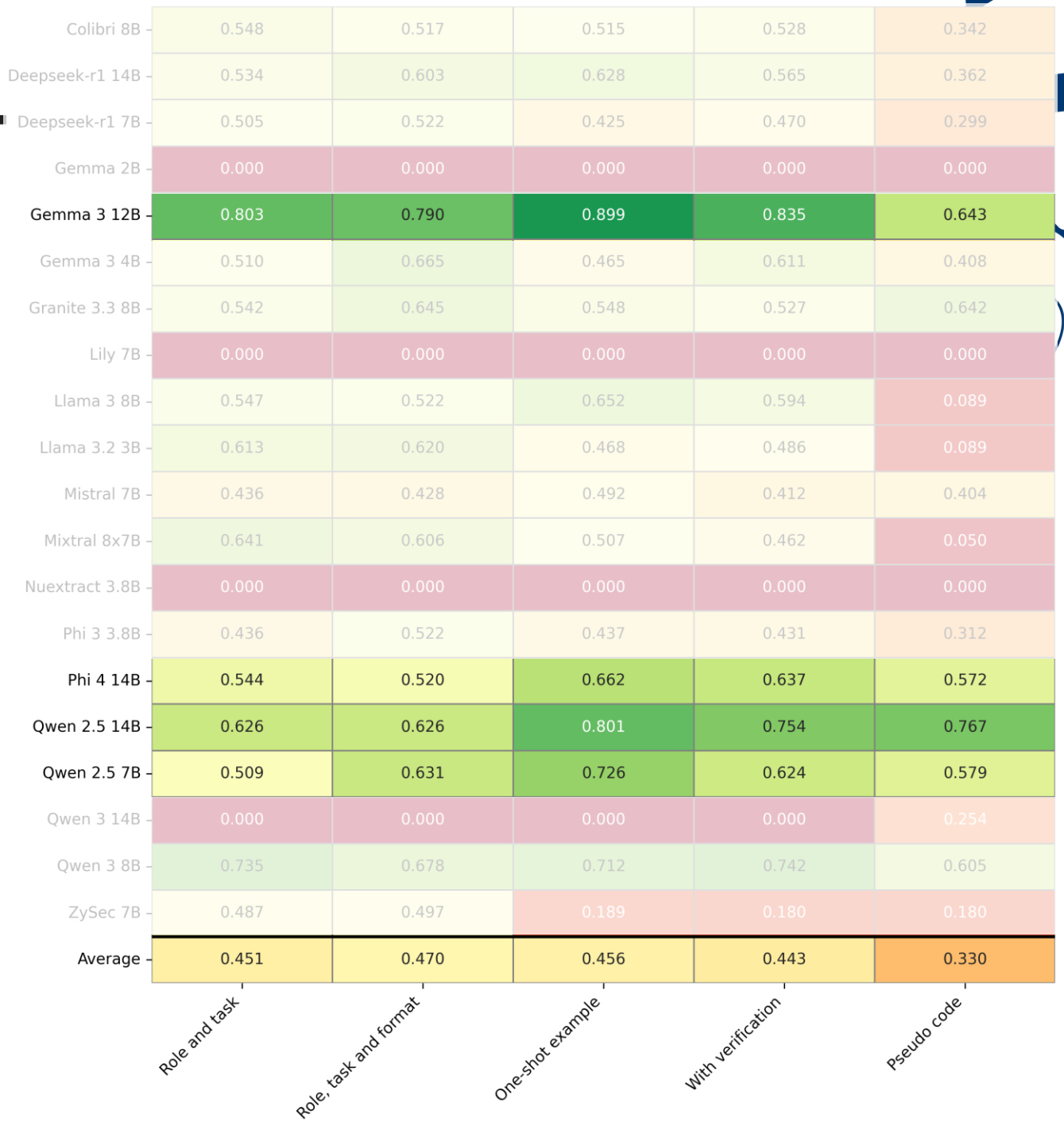
# OPTIMIZING THE PROMPT

- Role and Task
- Role, Task and Format
- One-shot example
- Chain of Verification
- Pseudo-code

```
EXECUTE the following function logic strictly.

CONST ENTITY_TYPES = {{
  {ENTITY_TYPES}
}}

DEF analyze_document(input_text):
    CONST SCOPE = [EMAIL, IP, BTC, IBAN, PERSON, LOCATION, PHONE, URL, TOX]
    extracted_artifacts = []

    # EXTRACTION LOOP
    ITERATE through input_text searching for SCOPE:
      MATCH criteria:
        IF type is PERSON:
          full or partial (First, Last, Initials)
        ELSE:
          MATCH standard pattern

      # DUPLICATE HANDLING
      IF entity exists in extracted_artifacts:
        ADD it again (We need frequency/location data)

      APPEND {{
        "entity": match.text,
        "type": match.type.UPPER(),
      }} to extracted_artifacts


    # RETURN
    response_object = {{
      "entities": extracted_artifacts,
    }}

    RETURN JSON(response_object)
```

# OPTIMIZING THE PROMPT

- Role and Task
- Role, Task and Format
- One-shot example
- Chain of Verification
- Pseudo-code



# CONTEXT CHUNKING

- Top 3 model families
- Chunking by:
  - Document
  - Section
  - Paragraph
  - Sentence

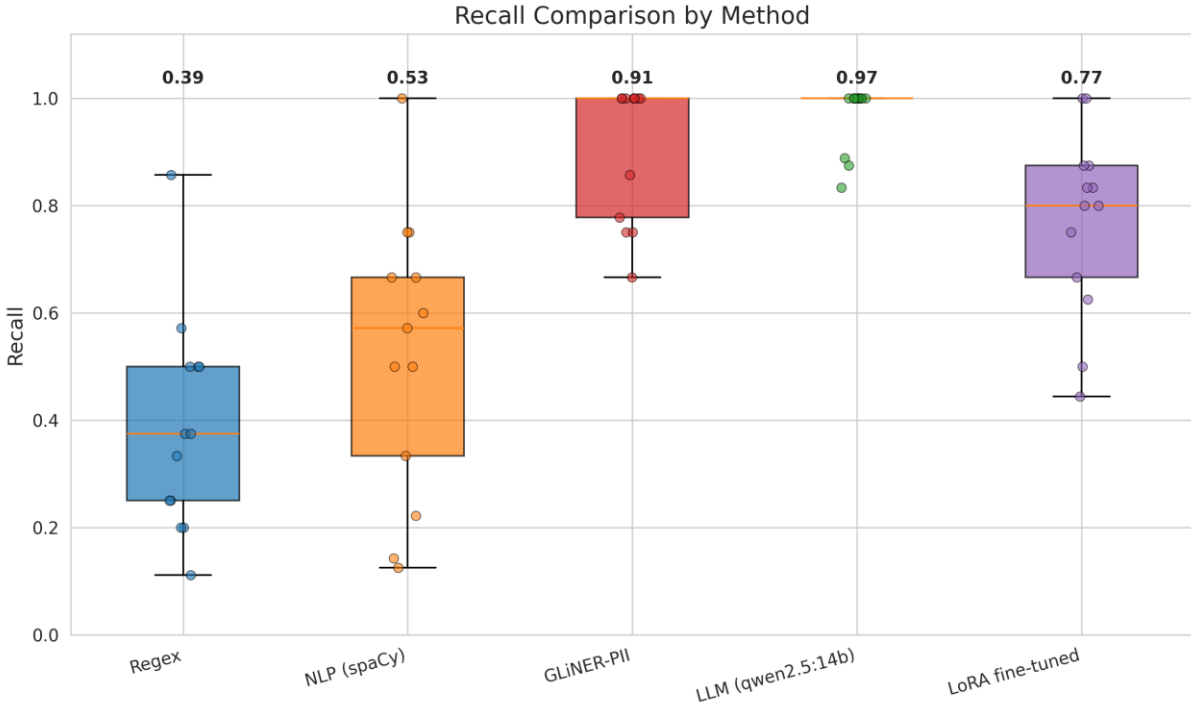
| Model | Parameters | Document | Paragraph | Sentence |
|---|---|---|---|---|
| Phi 4 | 14B | 0.730 | 0.820 | 0.928 |
| Gemma 3 | 12B | 0.899 | 0.801 | 0.829 |
| Qwen 2.5 | 14B | 0.809 | 0.881 | 0.945 |
| Qwen 2.5 | 7B | 0.776 | 0.692 | 0.857 |

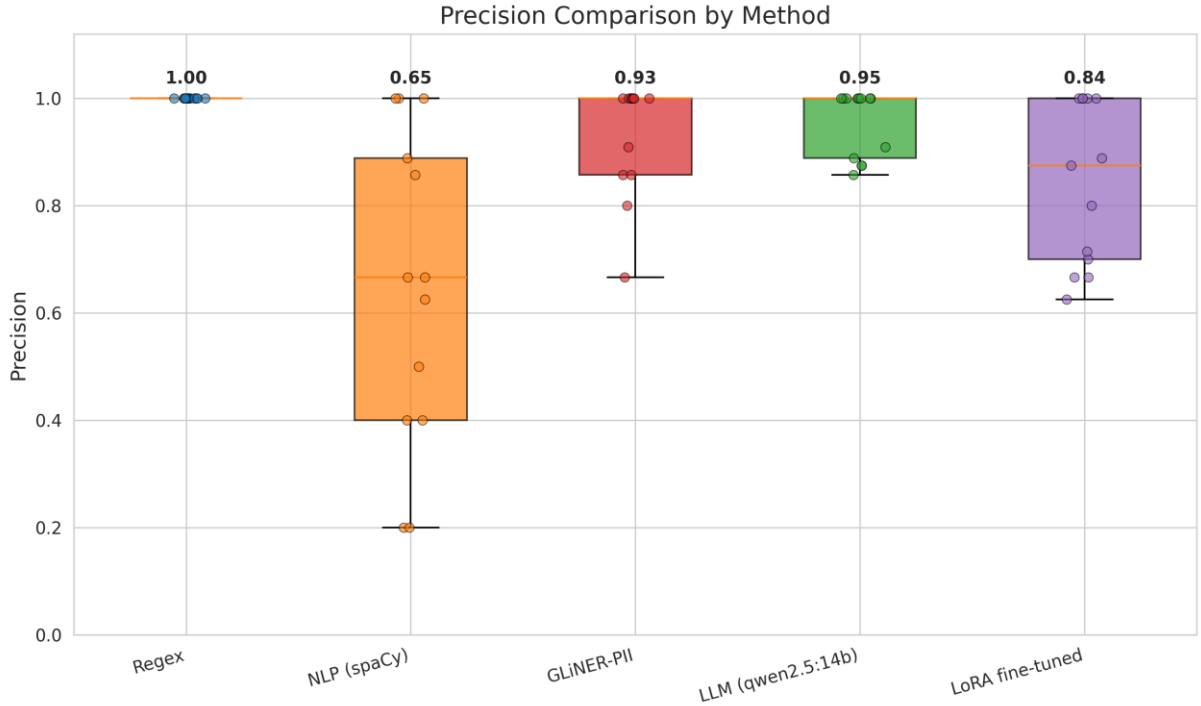
F1 score on synthetic e-mail dataset n=13

# LOW RANK ADAPTATION (LORA) FINE-TUNING

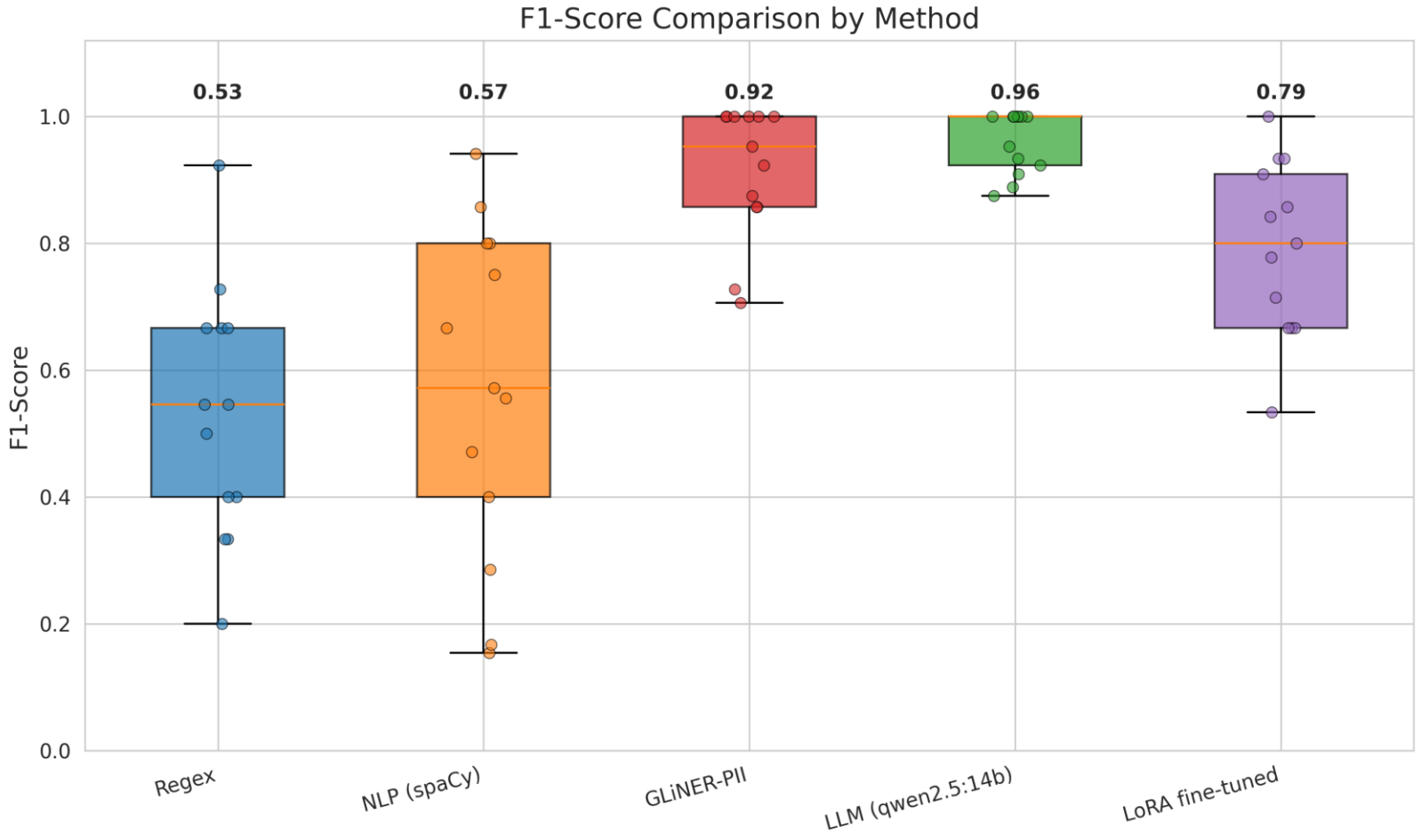
- Parameter efficient fine-tuning
- AI4Privacy dataset (PII)
- Difference in dataset negatively impacted results

# RESULTS: LLM FINE-TUNING (1)



# RESULTS: LLM FINE-TUNING (2)



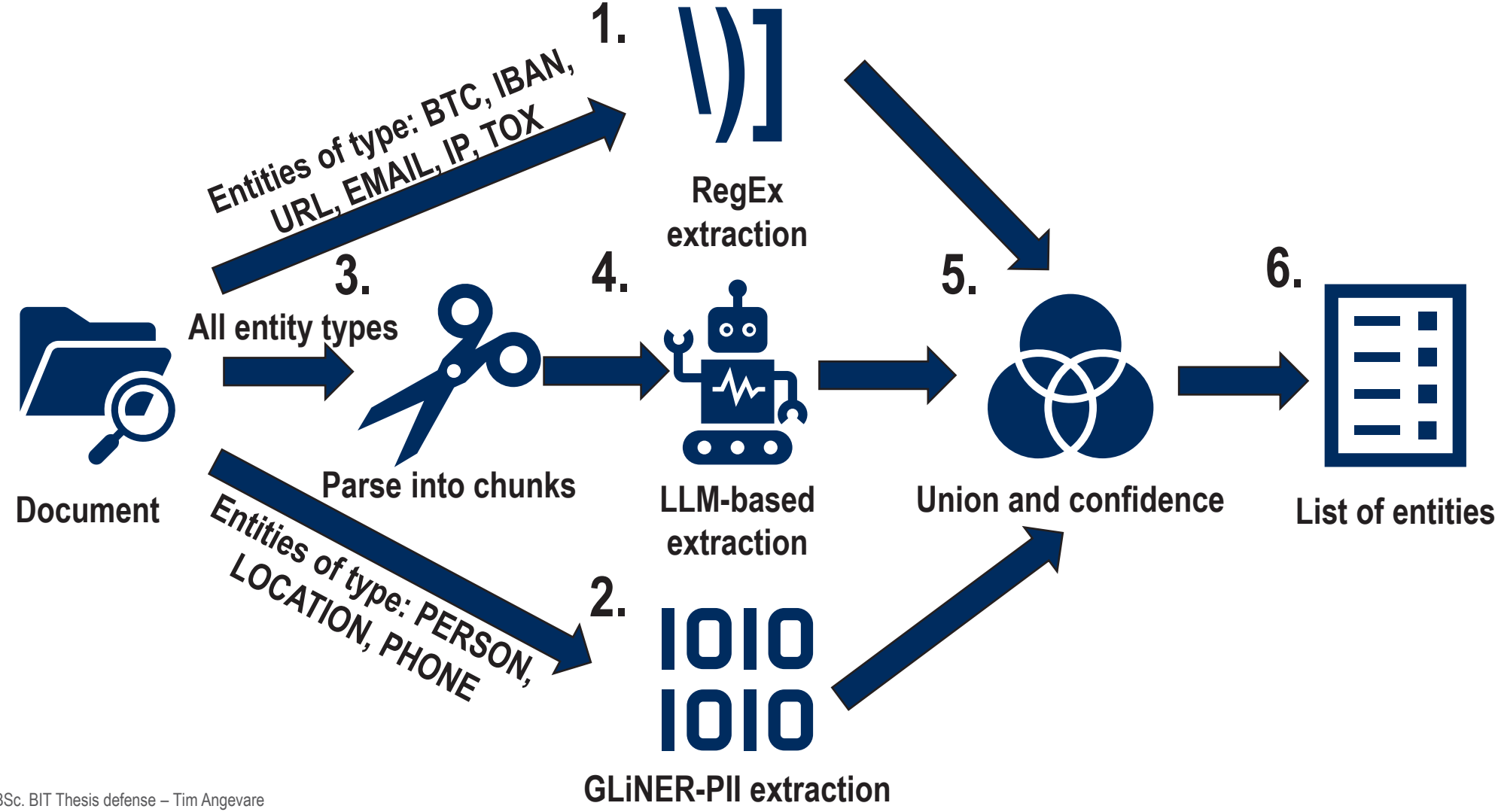
# **EXPERIMENT 3:** HYBRID PIPELINE

# RESULTS: PATTERN-BASED METHODS AND NLP MODELS

| Entity type | RegEx | GLiNER-PII | Qwen 2.5 14B |
|---|---|---|---|
| Person | 0.0 | 0.93 | 0.98 |
| E-mail | 0.89 | 1.0 | 1.0 |
| Location | 0.0 | 0.80 | 0.83 |
| BTC | 1.0 | 0.2 | 0.5 |
| IP | 0.75 | 0.75 | 1.0 |
| TOX ID | 1.0 | 1.0 | 1.0 |
| URL | 0.67 | 0.56 | 0.67 |
| Phone | 0.75 | 1.0 | 1.0 |

F1 score on synthetic e-mail dataset n=13

# HYBRID PIPELINE DESIGN






UNIVERSITY OF TWENTE.

# RESULTS: HYBRID PIPELINE

- Synthetic e-mail dataset

- 500 Conti ransomware leaked messages
- Annotated by author
- ‘real dataset’

| Metric | ‘High’ confidence | Total |
|---|---|---|
| Precision | 0.987 | 0.903 |
| Recall | 0.805 | 1.000 |
| F1 score | 0.879 | 0.945 |

| Metric | ‘High’ confidence | Total |
|---|---|---|
| Precision | 0.94 | 0.79 |
| Recall | 0.94 | 0.82 |
| F1 score | 0.94 | 0.80 |

# CONCLUSION

This work contributes:

- a validated systematic methodology for LLM-based NER designed for LEA contexts
- the design of a hybrid pipeline that maximizes accuracy while reducing manual ETL workload.

# FUTURE WORK

- Collaboration for implementation in LEA contexts
- Attempt LoRA finetuning with high quality annotated dataset containing ALL target entity types and cybercrime context
- Research the feasibility of a LoRA feedback loop where annotations are used to retrain the LLM
- Expand to support extraction of multiple languages

# NAMED ENTITY RECOGNITION IN CYBERCRIME INTELLIGENCE ANALYSIS: A HYBRID LLM APPROACH TO REDUCE MANUAL ETL WORKLOAD

TIM ANGEVARE – S2744007



UNIVERSITY OF TWENTE.

# FINAL PROMPT

You are a cyber intelligence analyst with 20 years of experience in the field.

Your task is to extract any entity from the input text. For each entity found you MUST indicate the type in UPPERCASE. ONLY extract entities if literal entity is present in input text.
The expected entity types are the following:
- EMAIL: email addresses format (user@domain.tld)
- IP: IP addresses (IPv4 x.x.x.x or IPv6)
- BTC: ONLY Bitcoin wallet addresses (26-35 alphanumeric, starting with 1, 3, or bc1) EXCLUDE the word bitcoin or values (for example 2.0 BTC)
- IBAN: iban bank account number
- PERSON: Human names (John Smith, John, Catalina) EXCLUDE initials (for example A.H.)
- LOCATION: cities, countries, geographic locations
- PHONE: phone numbers in any format
- URL: URLs and web addresses EXCLUDE filenames
- TOX: Tox messenger IDs

**Output**:
The output MUST be in a JSON object with key 'entities' and the value a list of dictionaries including every entity found. For each entity you MUST indicate the type in UPPERCASE.

**OUTPUT EXAMPLE**:

```
{
  "entities": [
    {"entity": "target123@darkmail.org", "type": "EMAIL"},
    {"entity": "10.45.67.89", "type": "IP"},
    {"entity": "Thompson", "type": "PERSON"},
    {"entity": "Helsinki", "type": "LOCATION"},
    {"entity": "Tim", "type": "PERSON"}
  ]+
}
```

Role

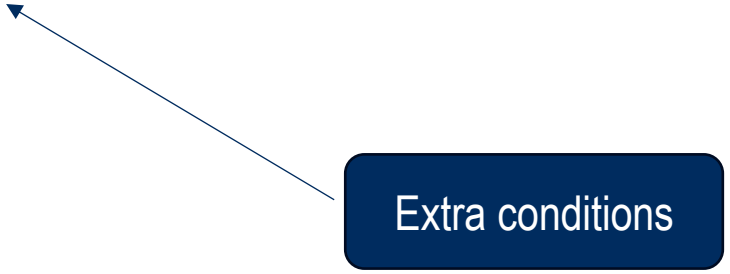
Task

Format

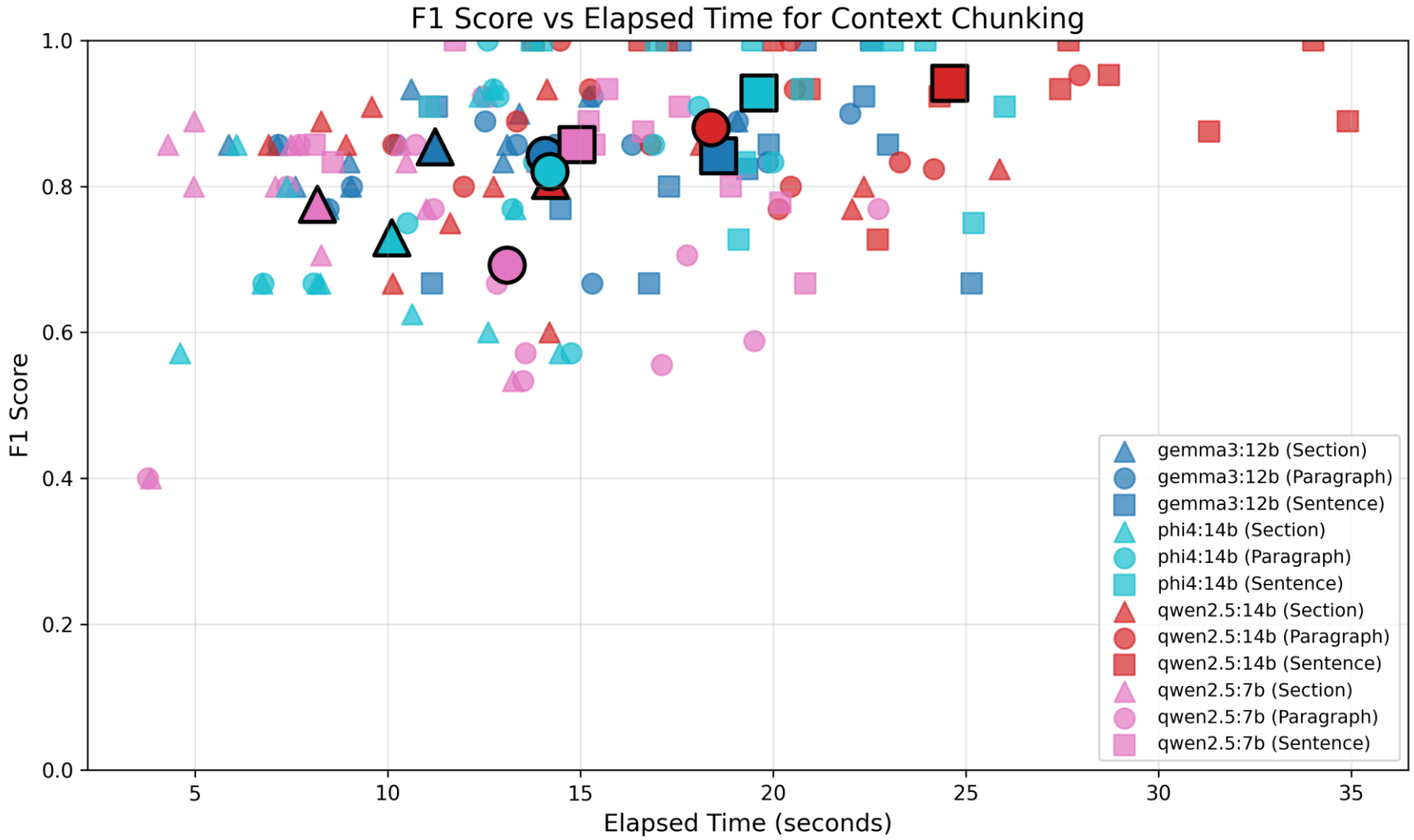
One-shot Example

# FINAL PROMPT (2)

Return empty array if no entities found in the input text.
PAY ATTENTION to sentences that begin with entity type PERSON, for example Anna.
PAY ATTENTION to when the sentences begin with possesive forms of entity type PERSON, for example Catalina's
PAY ATTENTION to when the sentences contain a FULL NAME, the FULL NAME MUST be extracted as ONE entity.
PAY ATTENTION to when entities follow a different format, extract them EXACTLY like in the source text.
DO NOT include any entities from the example or the system prompt in your answer.



# CONTEXT CHUNKING TIME IMPACT

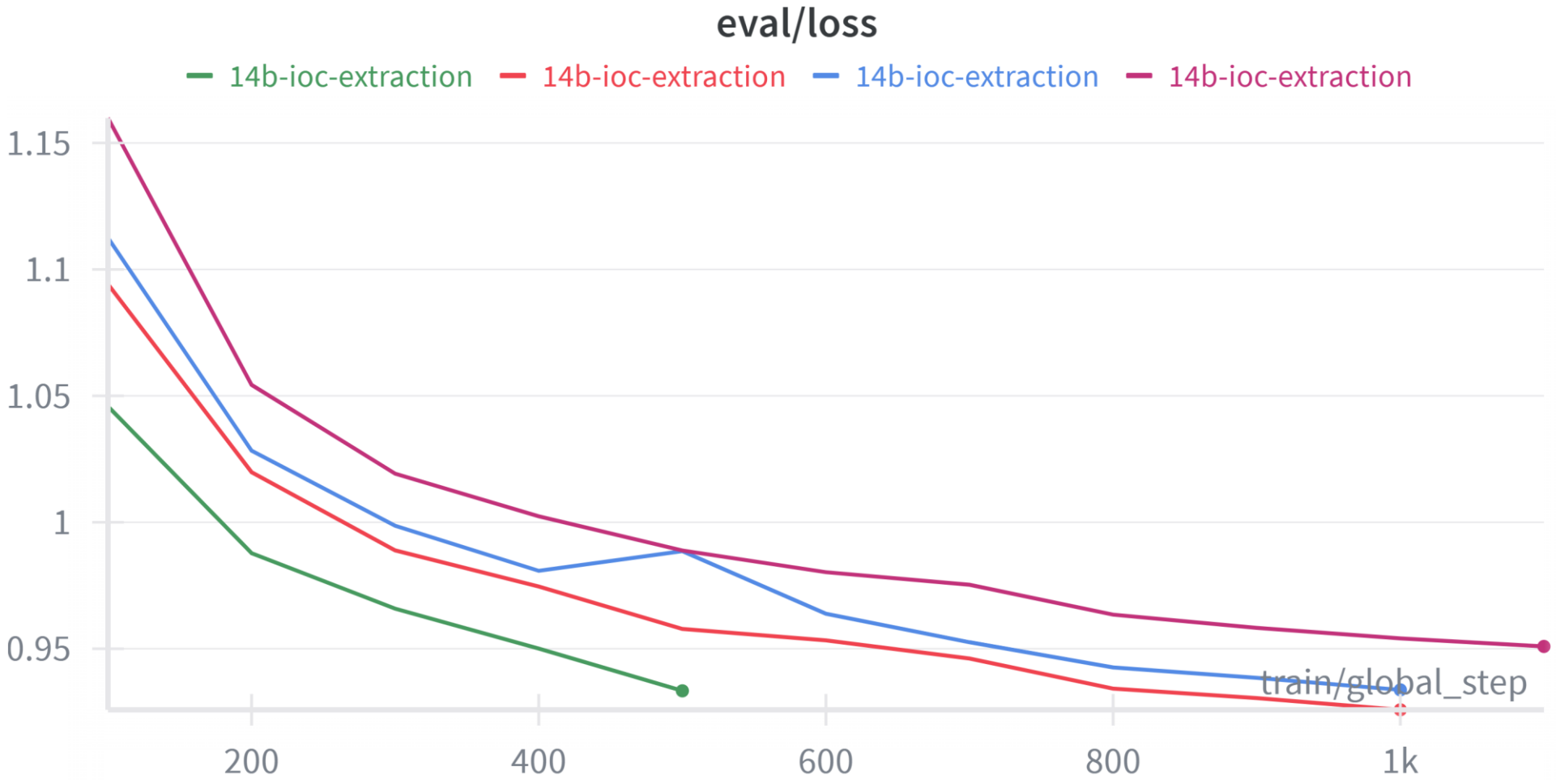


# LORA FINE-TUNING




UNIVERSITY OF TWENTE.

# LORA FINE-TUNING






UNIVERSITY OF TWENTE.

# LLM F1 PER ENTITY TYPE

| Entity type / method | Qwen2.5 14B | LoRA fine-tuned |
|---|---|---|
| Person | 0.98 | 0.68 |
| E-mail | 1.00 | 1.00 |
| Location | 0.83 | 0.54 |
| BTC | 0.50 | 0.40 |
| IP | 1.00 | 1.00 |
| TOX ID | 0.89 | 1.00 |
| URL | 0.67 | 0.50 |
| Phone | 1.0 | 1.00 |

# HYBRID PIPELINE PER ENTITY TYPE

| Entity Type | "High" Confidence | "Low" Confidence | Total |
|---|---|---|---|
| BTC | 3 | 1 | 4 |
| E-mail | 15 | 0 | 15 |
| IP | 5 | 0 | 5 |
| Location | 9 | 4 | 13 |
| Person | 27 | 4 | 31 |
| PHONE | 4 | 1 | 5 |
| TOX ID | 3 | 0 | 3 |
| URL | 5 | 0 | 5 |
| Total | 71 | 10 | 81 |

# HYBRID PIPELINE ENTITY EXTRACTION


[
{'text': 'marcellus.grant@protonmail.com', 'type': 'EMAIL', 'start_pos': 233, 'end_pos': 263,
'methods': ['regex', 'llm'], 'confidence': 'high', 'num_methods': 2},
{'text': 'Anna Hargrove', 'type': 'PERSON', 'start_pos': 5, 'end_pos': 18, 'methods': ['llm'],
'confidence': 'low', 'num_methods': 1}
]