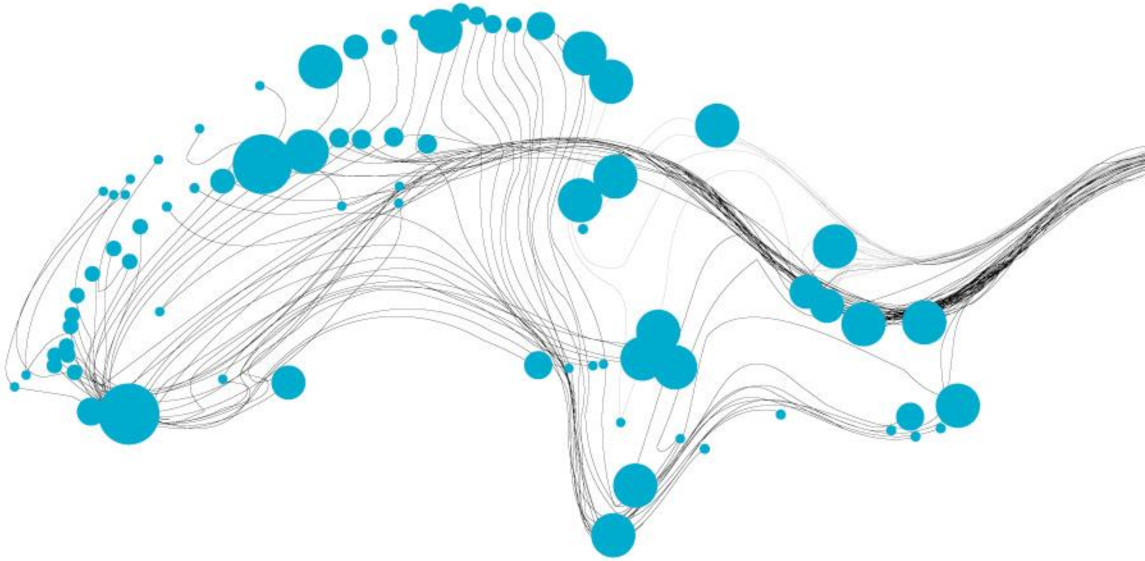
EEMCS / TCS

AUTOMATING THE CYBERSECURITY TRIAGE PROCESS

A COMPARATIVE STUDY ON THE PERFORMANCE OF LARGE LANGUAGE MODELS



PASCAL BAKKER
SUPERVISED BY JAIR SANTANNA
2024-07-05

UNIVERSITY
OF TWENTE.

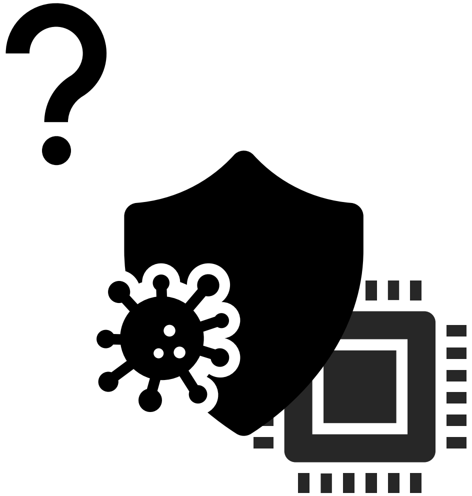
HOSPITAL – TRIAGE

1. Immediate
2. Emergent
3. Urgent
4. Less urgent
5. Non-urgent

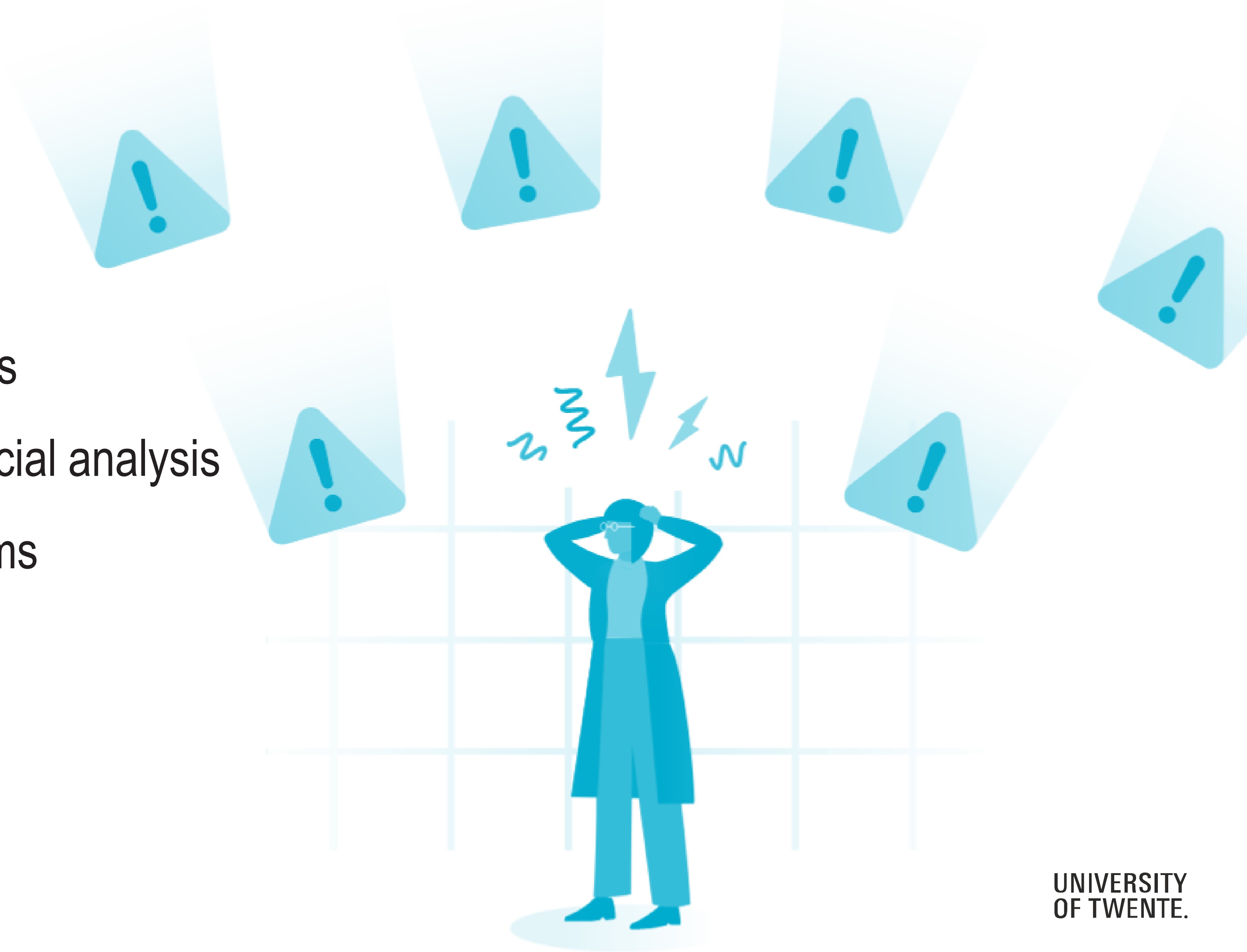
UNIVERSITY
OF TWENTE.

HOSPITAL – TRIAGE

1. Immediate
2. Emergent
3. Urgent
4. Less urgent
5. Non-urgent



THE PROBLEM WITH TRIAGE



Immense numbers of alarms

- Too little time → Superficial analysis
- Fatigue → Missed alarms
- Burnout / turnover
- Human error

UNIVERSITY
OF TWENTE.

LARGE LANGUAGE MODELS (LLMS)

- Natural Language Processing
- Much data → General understanding

1. Generating natural language
2. Identifying contextual relationships
3. Recognizing complex patterns
4. Analyzing semantics

- Many existing applications

→ Automate triage

RESEARCH QUESTIONS

How can LLMs be integrated into the existing incident response workflow to streamline the triage process?

What suitable evaluation metrics should be used to assess the performance of LLMs in cybersecurity triage?

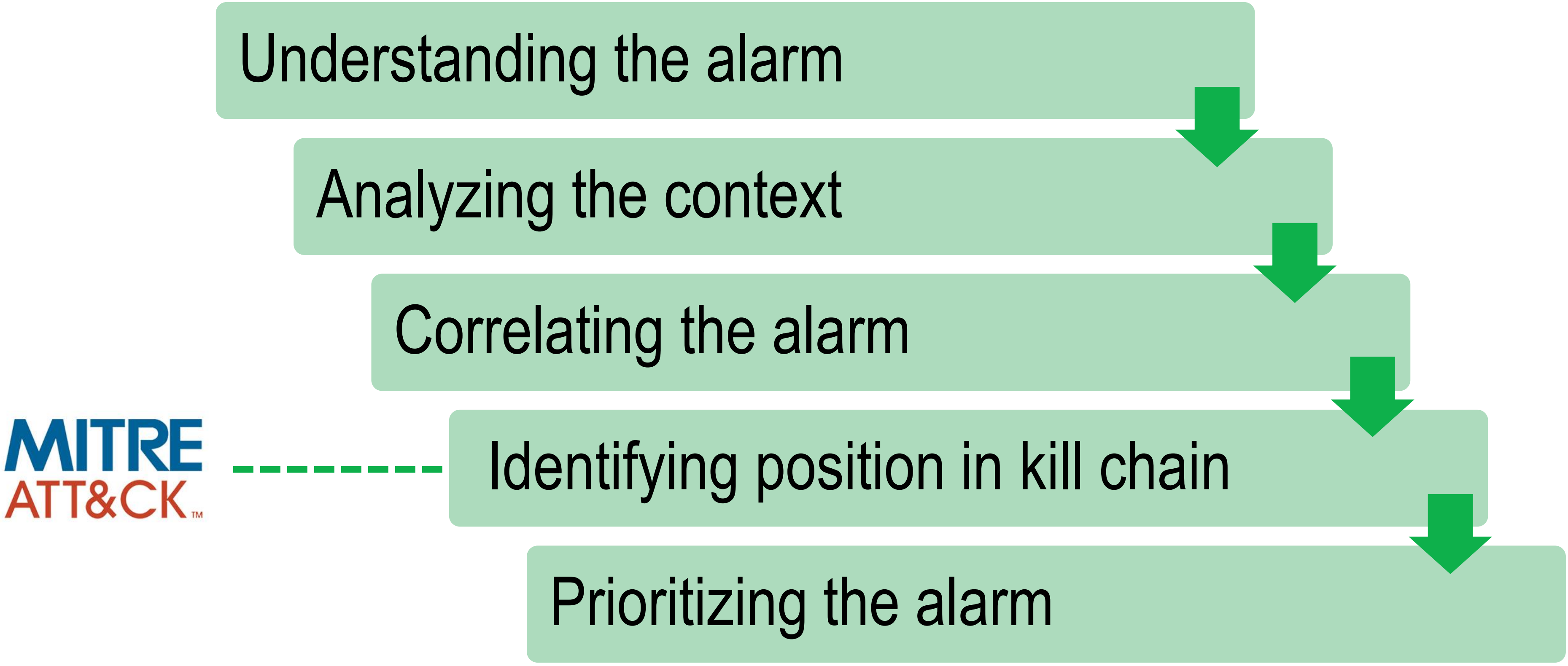
How do different LLMs compare in performance when optimizing the cybersecurity triage process?

DEFINING TRIAGE

Very little documentation → Interview

NORTHWAVE CYBER SECURITY

DEFINING TRIAGE



Understanding the alarm

Analyzing the context

Correlating the alarm

MITRE ATT&CK™ - - - - - - - - Identifying position in kill chain

Prioritizing the alarm

OPTIMIZING TRIAGE

USING LLMS

EXISTING OPTIMIZATIONS

- Organize alarms in trees [12]
- Integrate thread intelligence [41]
- Follow steps of senior analysts [25, 48]

OPTIMIZING TRIAGE

USING LLMS

EXISTING OPTIMIZATIONS

- Organize alarms in trees [12]
- Integrate thread intelligence [41]
- Follow steps of senior analysts [25, 48]

<u>Do not automate natural language tasks</u>

OPTIMIZING TRIAGE

USING LLMS

1. Detecting cybersecurity announcement emails
2. Detecting relation between email and alarm
3. Finding correlation between alarms
4. Determine position in kill chain
5. Determine priority of alarm

UNIVERSITY
OF TWENTE.

OPTIMIZING TRIAGE

USING LLMS

1. Detecting cybersecurity announcement emails
2. Detecting relation between email and alarm
3. Finding correlation between alarms
4. Determine position in kill chain
5. Determine priority of alarm

TASK 1

EVALUATING LLMS

ANNOUNCEMENT DETECTION

1. Give email to LLM
2. Is this a cybersecurity announcement?
3. True / false

TASK 1

EVALUATING LLMS

ANNOUNCEMENT DETECTION

1. Give email to LLM
2. Is this a cybersecurity announcement?
3. True / false

> **Customer X will add the user “sea_line” to the local administrators on the computer “FRLIM-IPC-0017”.**

“True” → ✅

“False” → ❌

TASK 2
EVALUATING LLMS

TACTIC DETECTION

1. Give email to LLM
2. What MITRE ATT&CK tactic can consequential alarms have?
3. e.g. “exfiltration”

TASK 2
EVALUATING LLMS

TACTIC DETECTION

1. Give email to LLM
2. What MITRE ATT&CK tactic can consequential alarms have?
3. e.g. “exfiltration”

```
Customer X will add the
user “sea_line” to the
local administrators on
the computer “FRLIM-IPC-
0017”.
```

“privilege escalation” → ✓

“persistence” → ✓

“reconnaissance” → ✗

EVALUATING AND COMPARING LLMS

EVALUATION METRICS

For different prompts:

- F1-score / accuracy
- Median time
- Error rate

EVALUATING AND COMPARING LLMS

EVALUATION METRICS

For different prompts:

- F1-score / accuracy
- Median time
- Error rate

DATASET

- 40 labeled emails
- 20 announcements with labeled tactics

EVALUATING AND COMPARING LLMS

EVALUATION METRICS

For different prompts:

- F1-score / accuracy
- Median time
- Error rate

DATASET

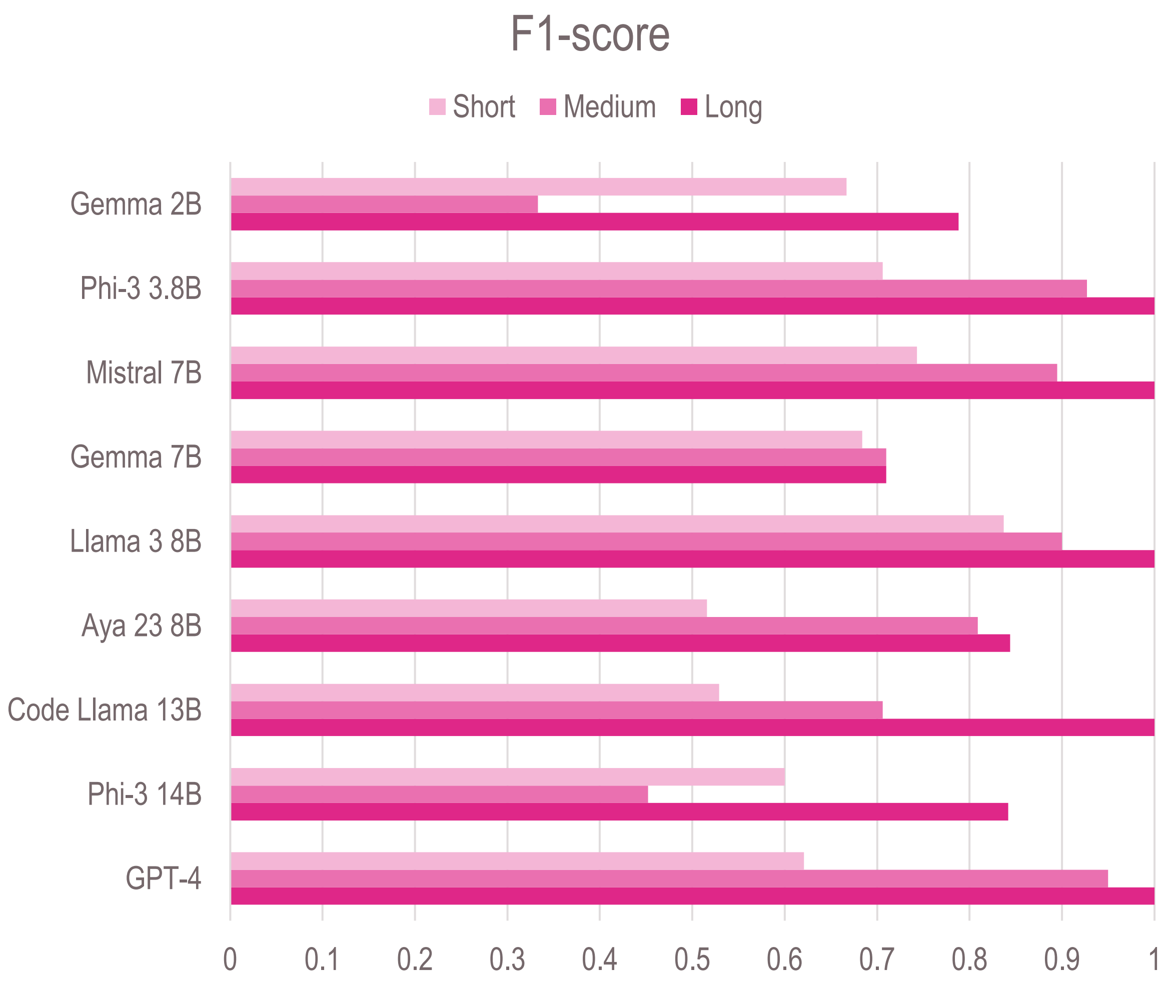
- 40 labeled emails
- 20 announcements with labeled tactics

LARGE LANGUAGE MODELS

(And variations)

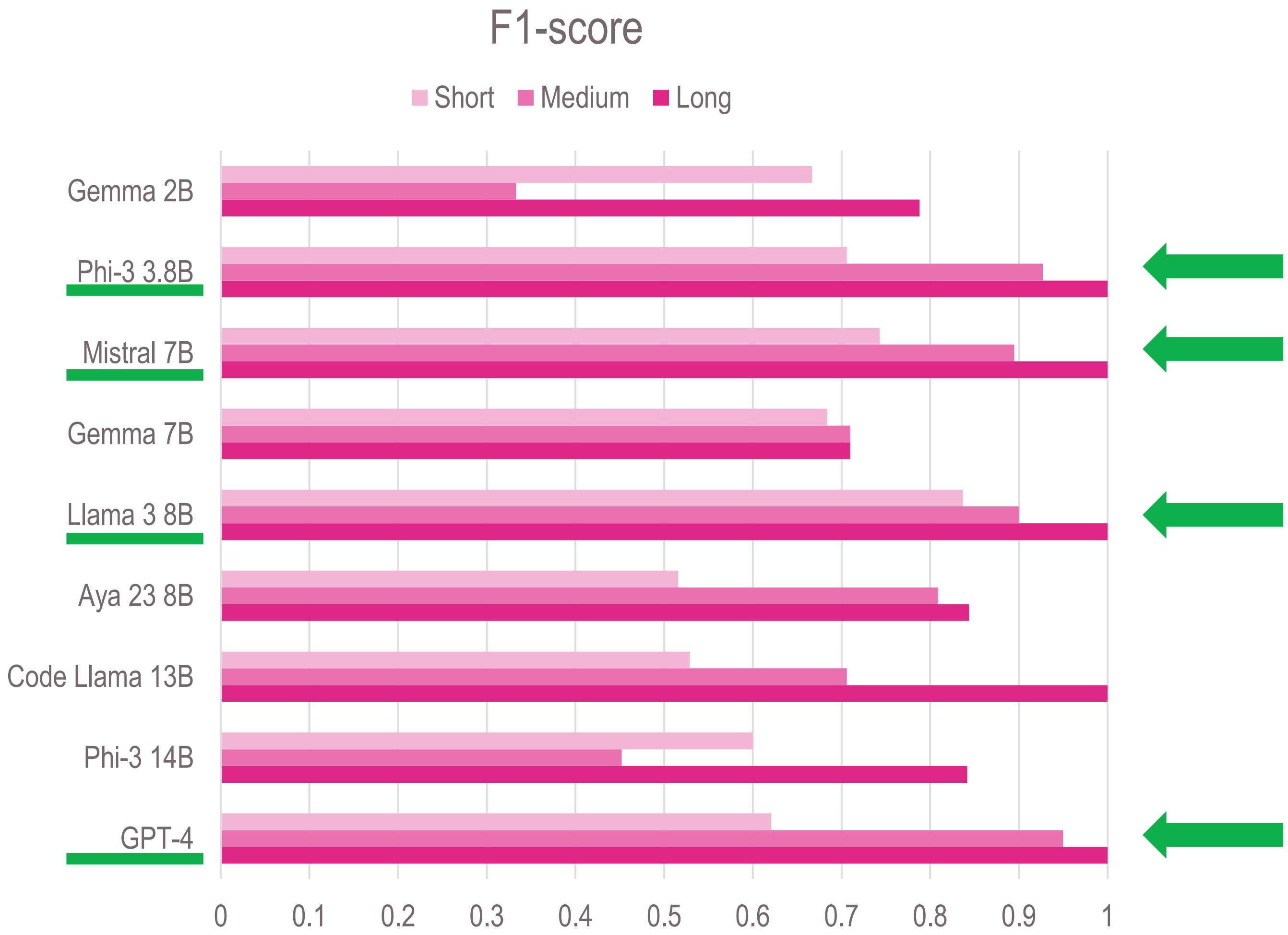
- GPT-4
- Llama 3
- Mistral
- Phi-3
- Gemma
- Aya 23
- Code Llama

RESULTS – ANNOUNCEMENT DETECTION



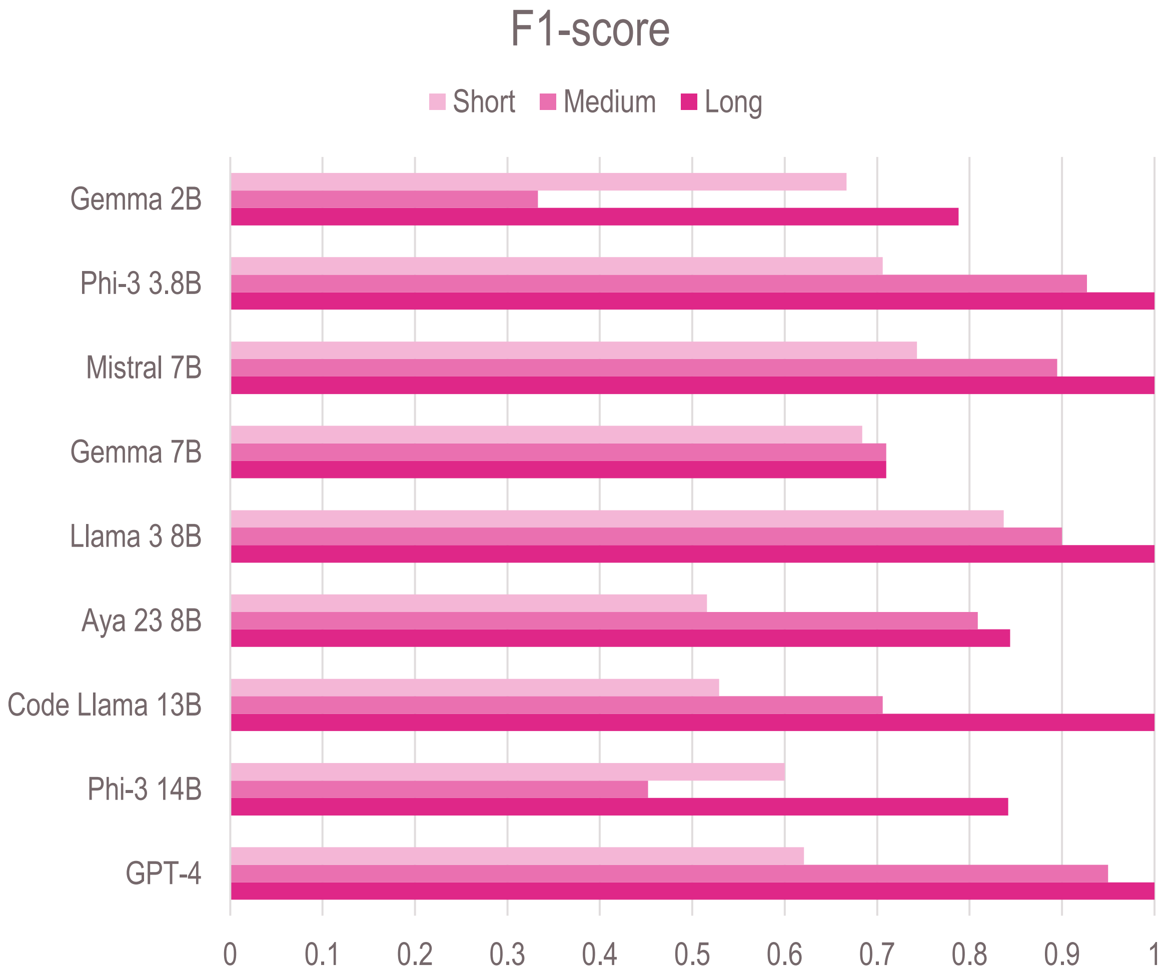
UNIVERSITY
OF TWENTE.

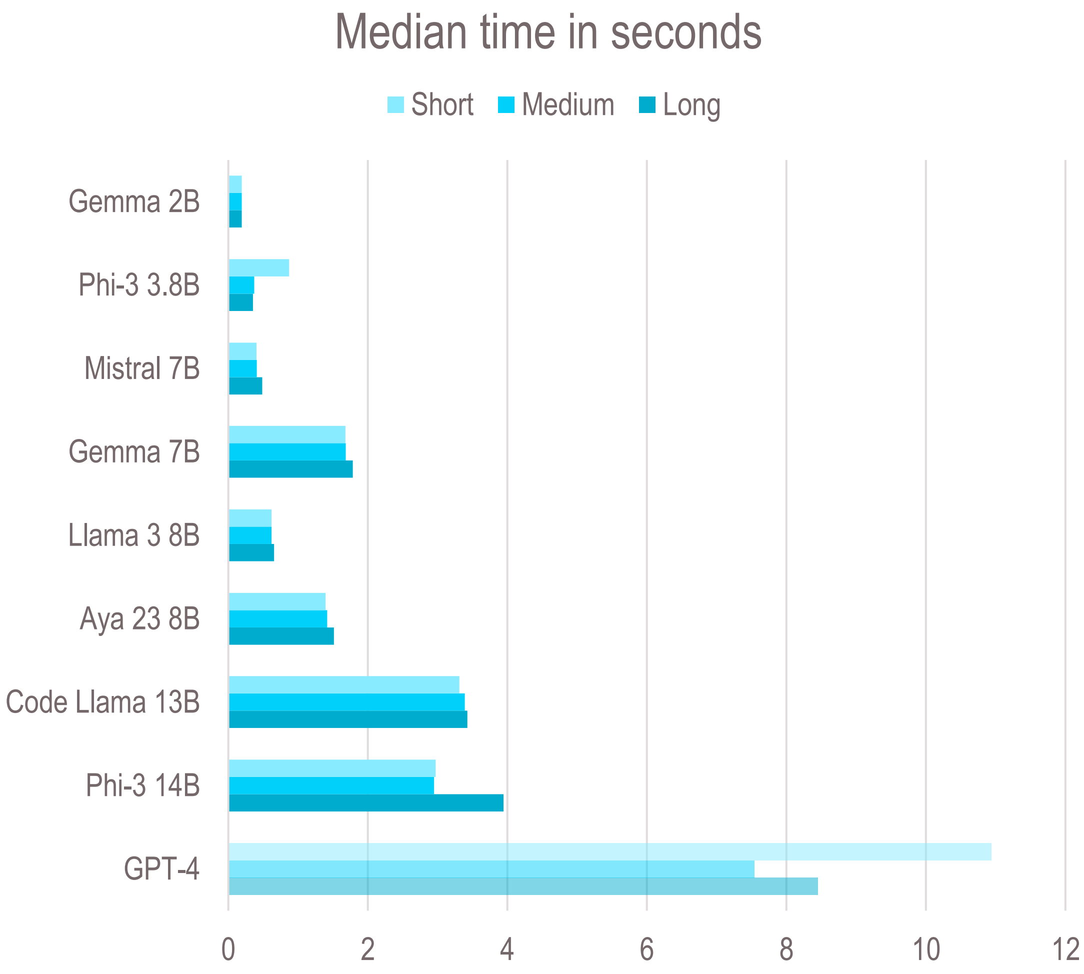
RESULTS – ANNOUNCEMENT DETECTION



UNIVERSITY
OF TWENTE.

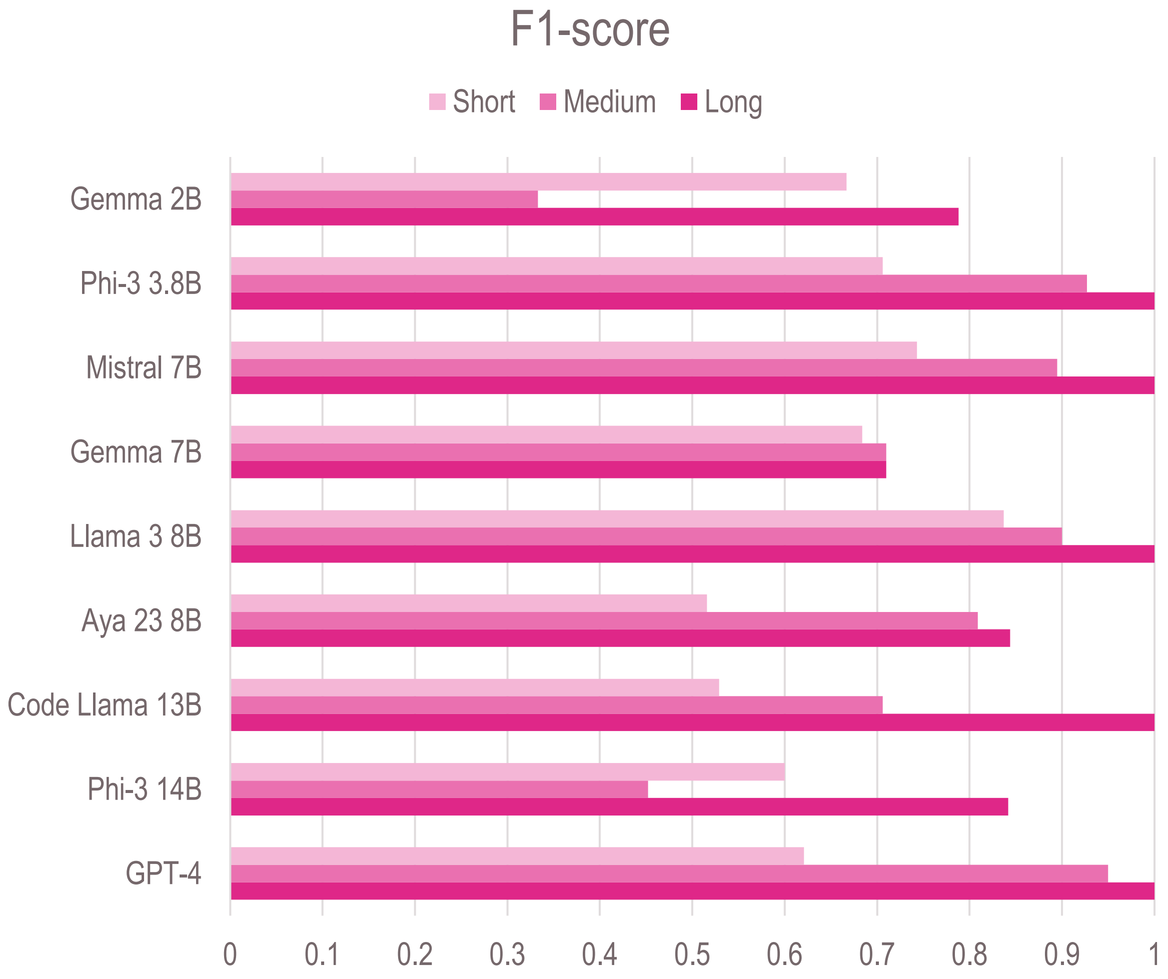
RESULTS – ANNOUNCEMENT DETECTION

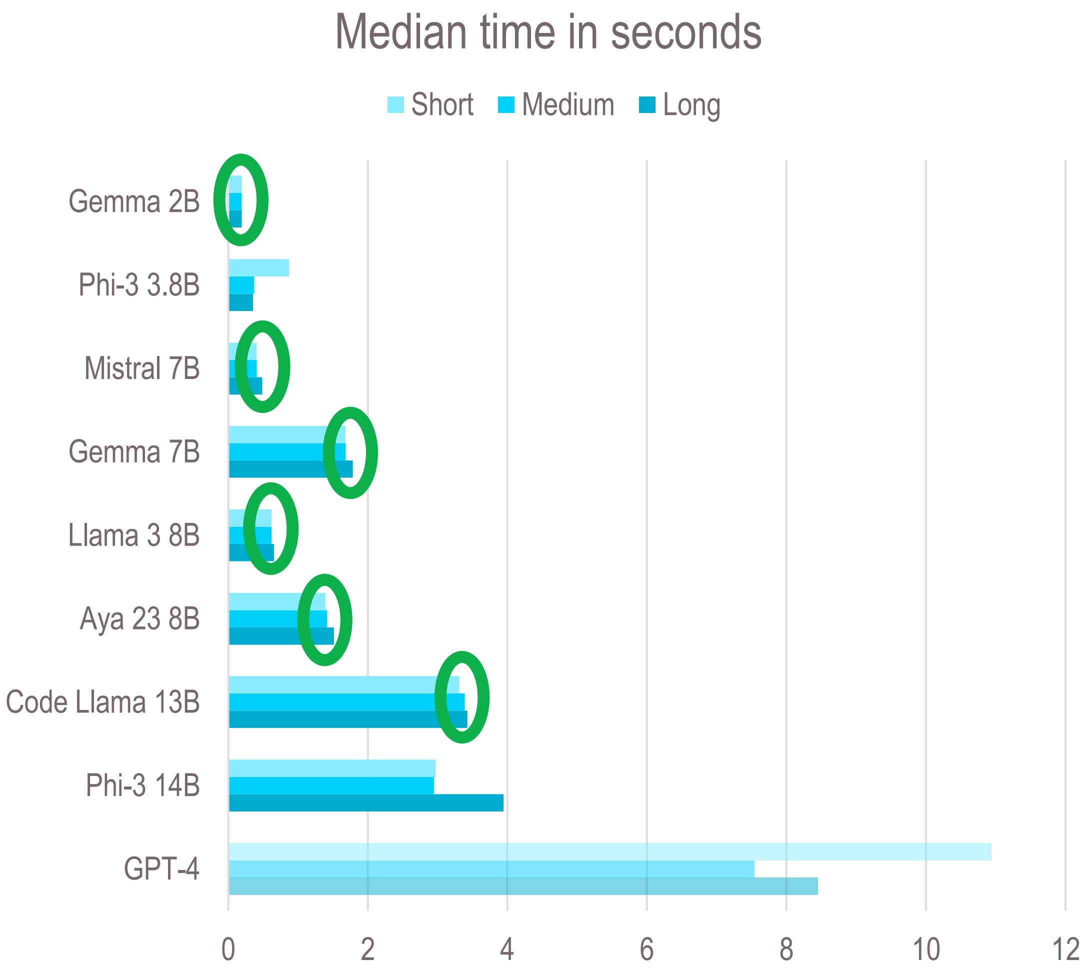




UNIVERSITY
OF TWENTE.

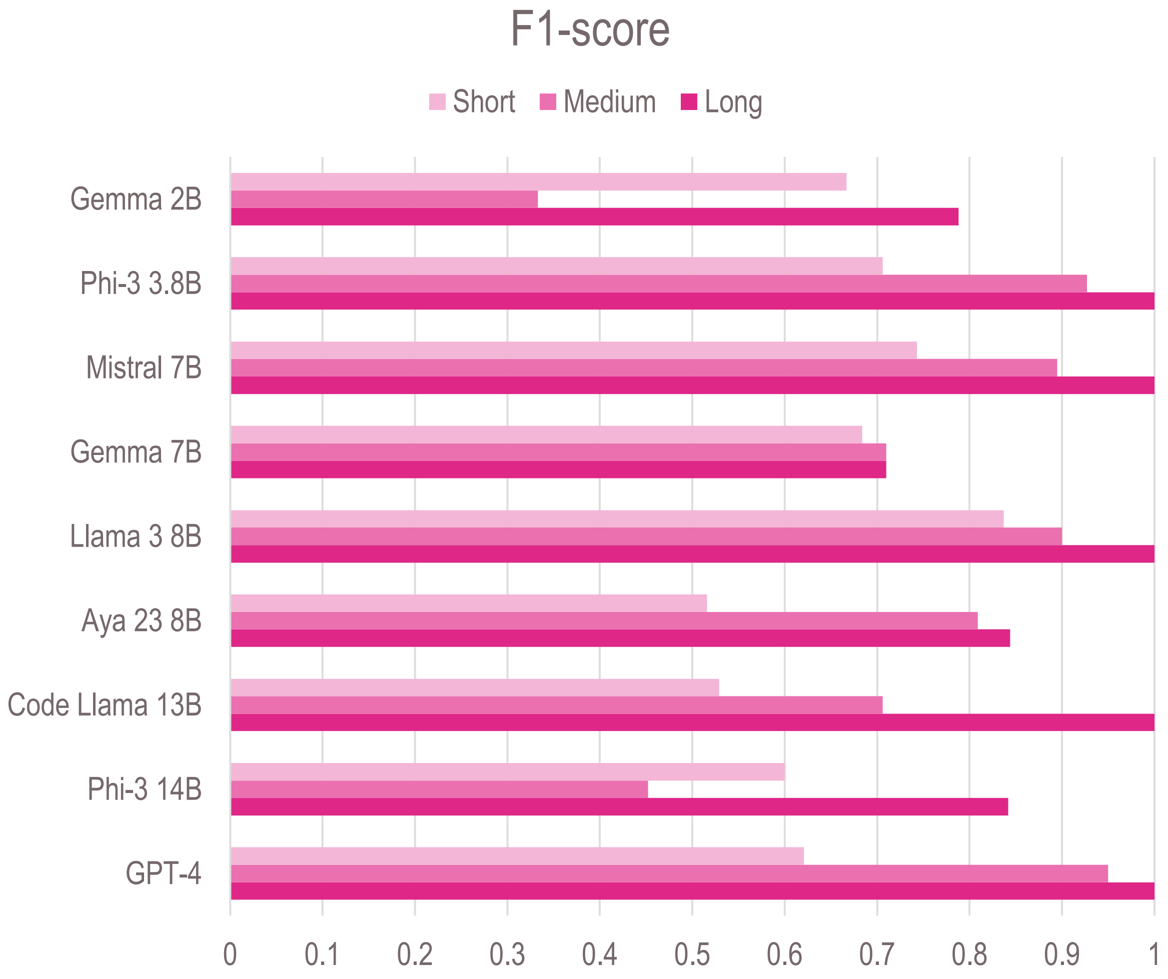
RESULTS – ANNOUNCEMENT DETECTION

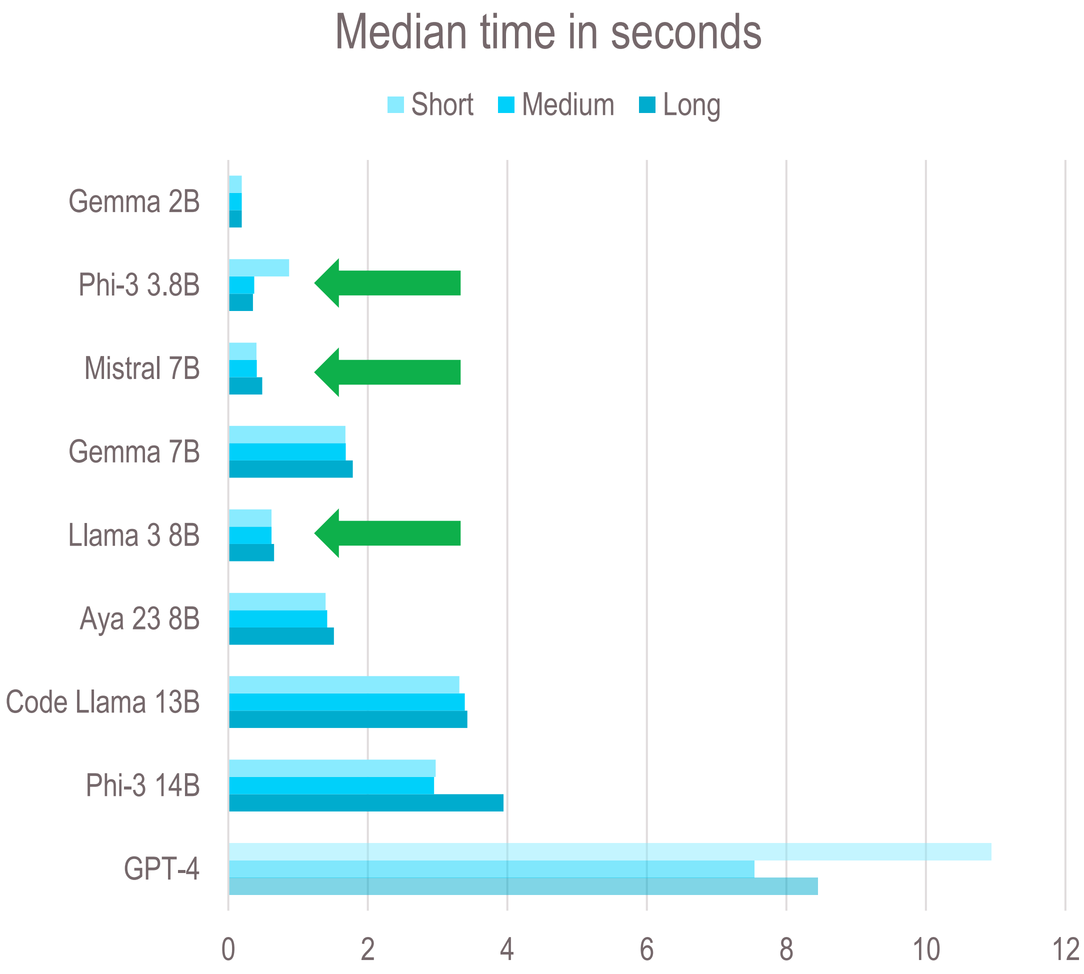




UNIVERSITY
OF TWENTE.

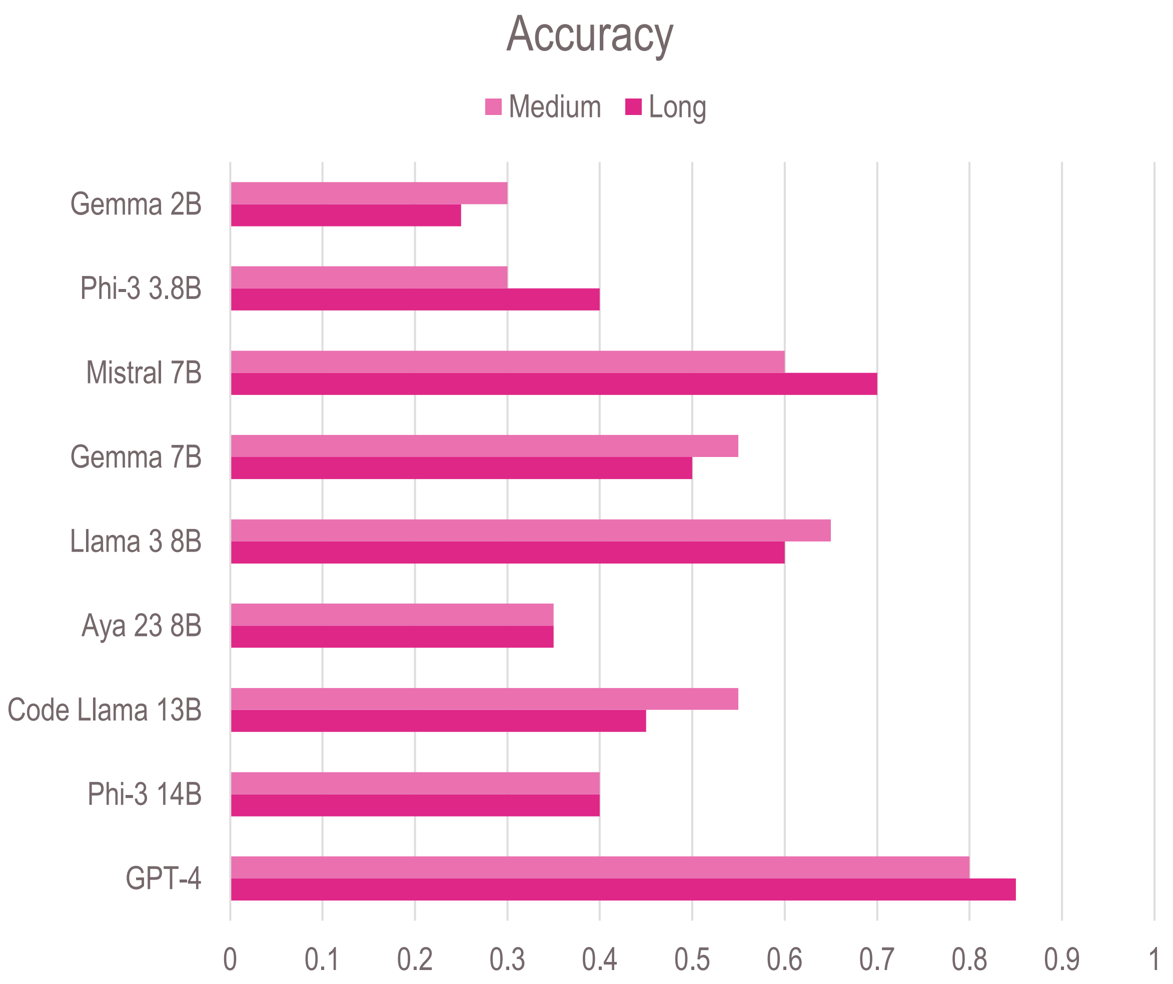
RESULTS – ANNOUNCEMENT DETECTION





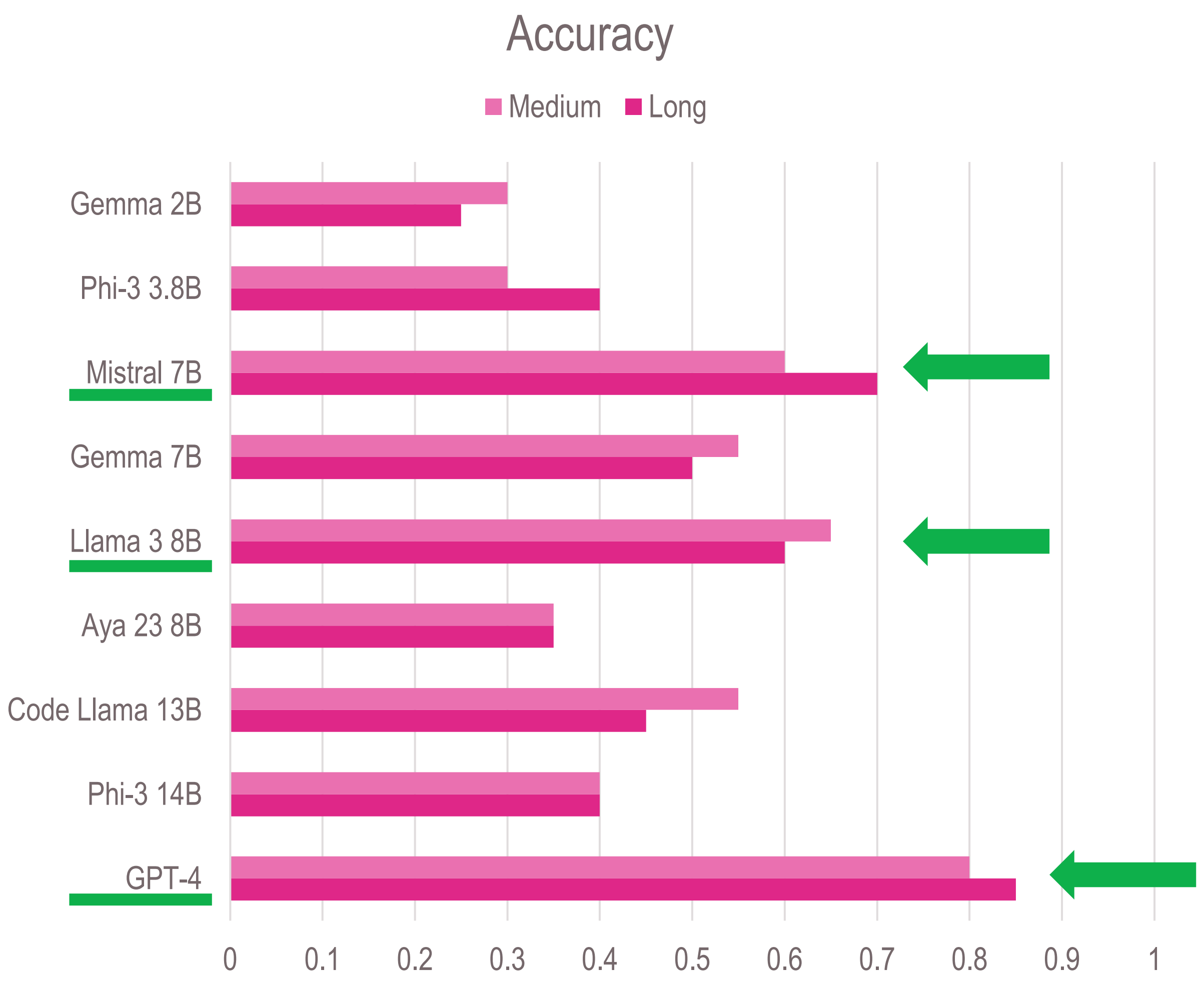
UNIVERSITY OF TWENTE.

RESULTS – TACTIC DETECTION



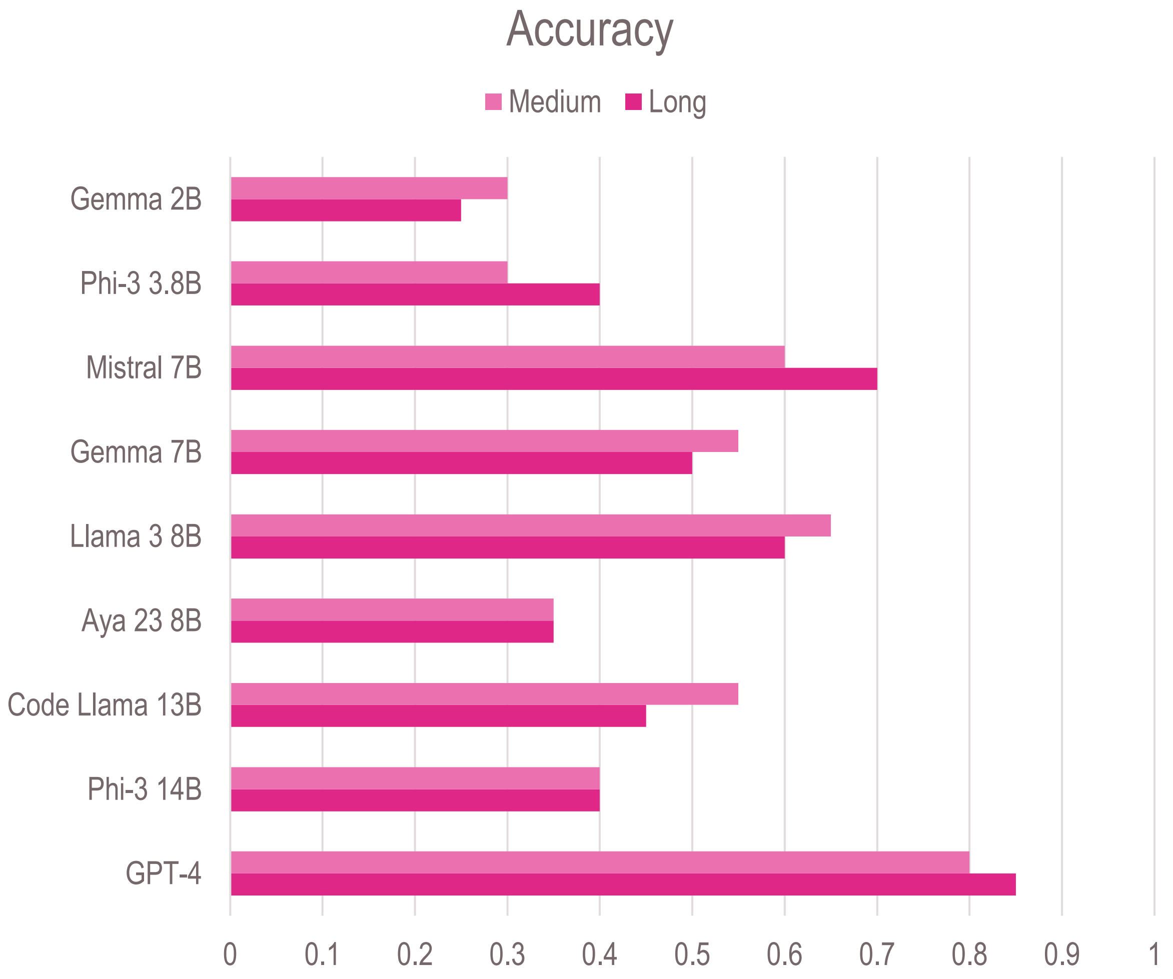
UNIVERSITY
OF TWENTE.

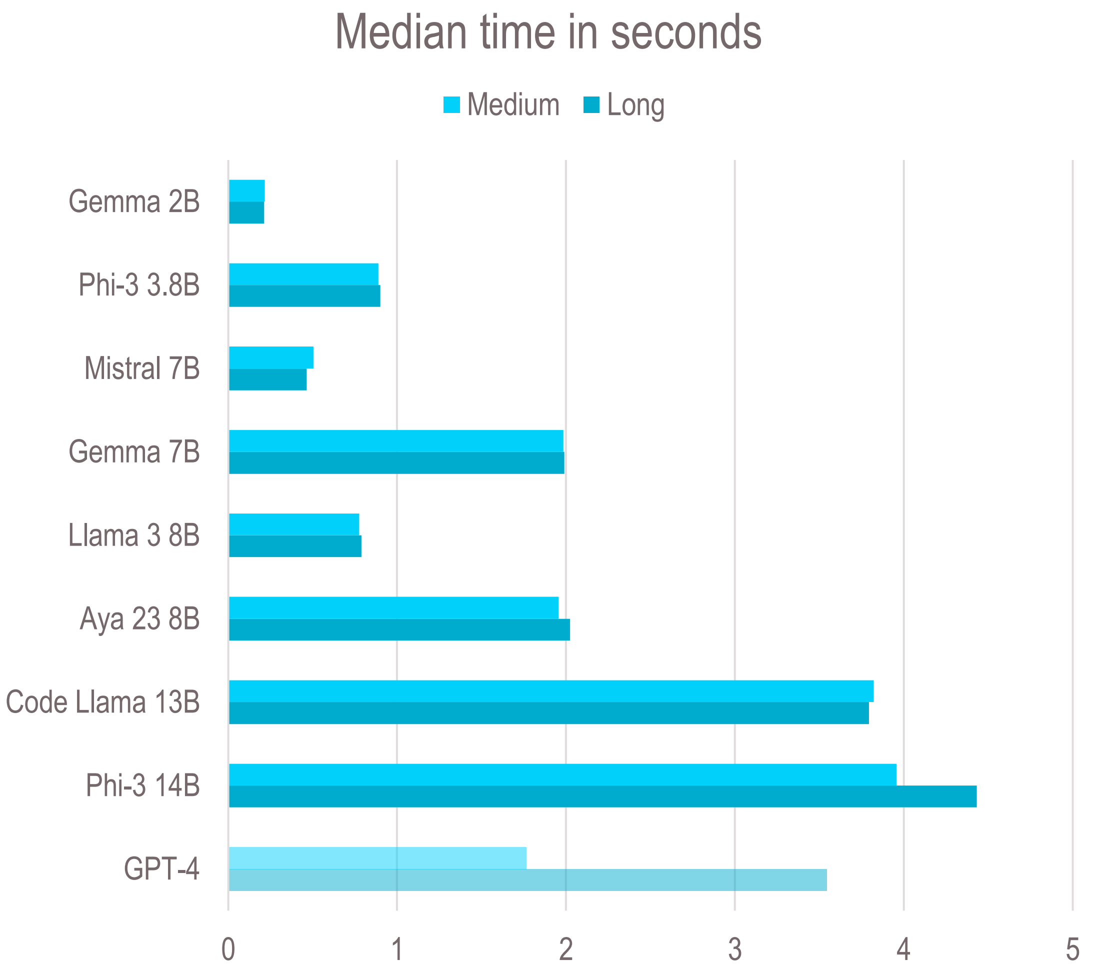
RESULTS – TACTIC DETECTION



UNIVERSITY
OF TWENTE.

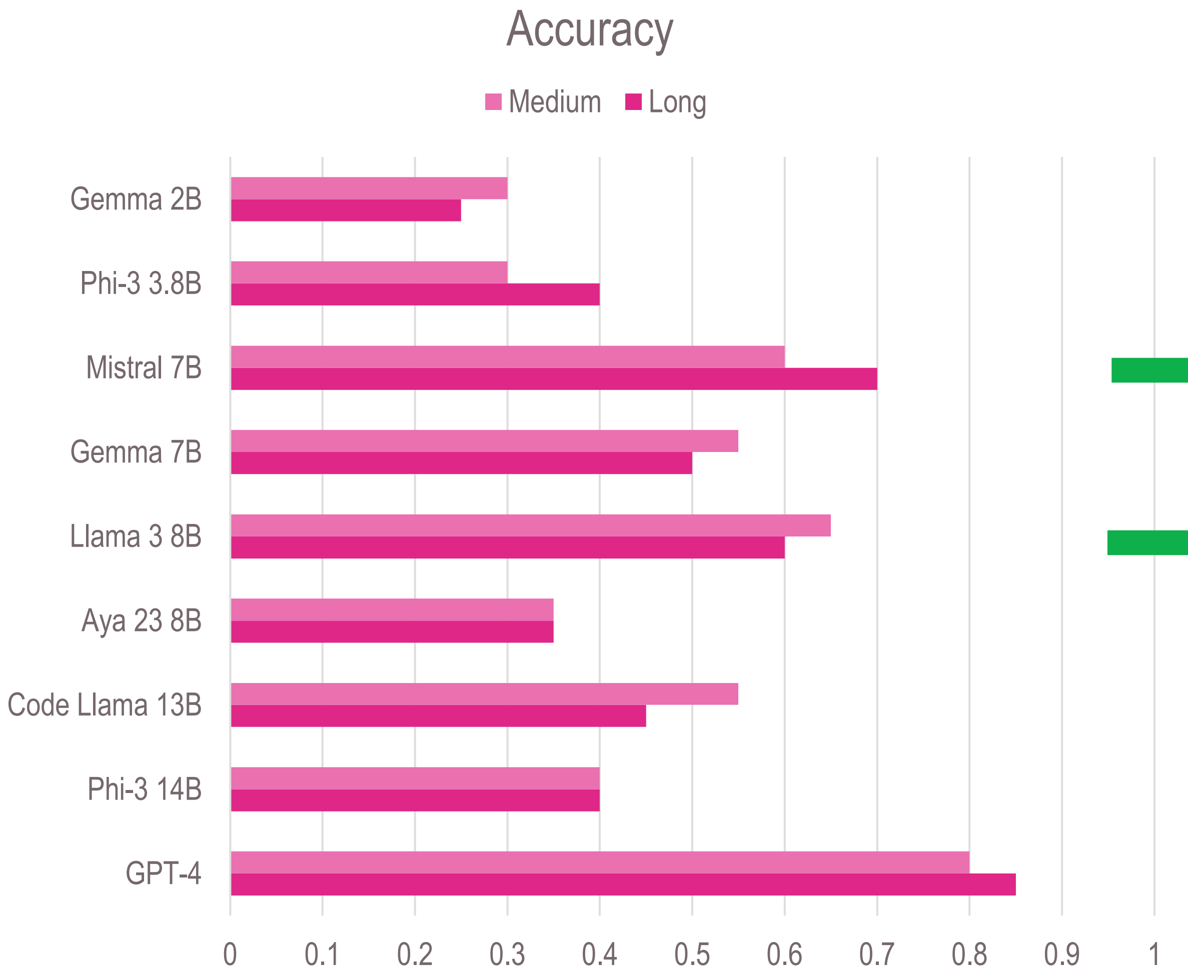
RESULTS – TACTIC DETECTION

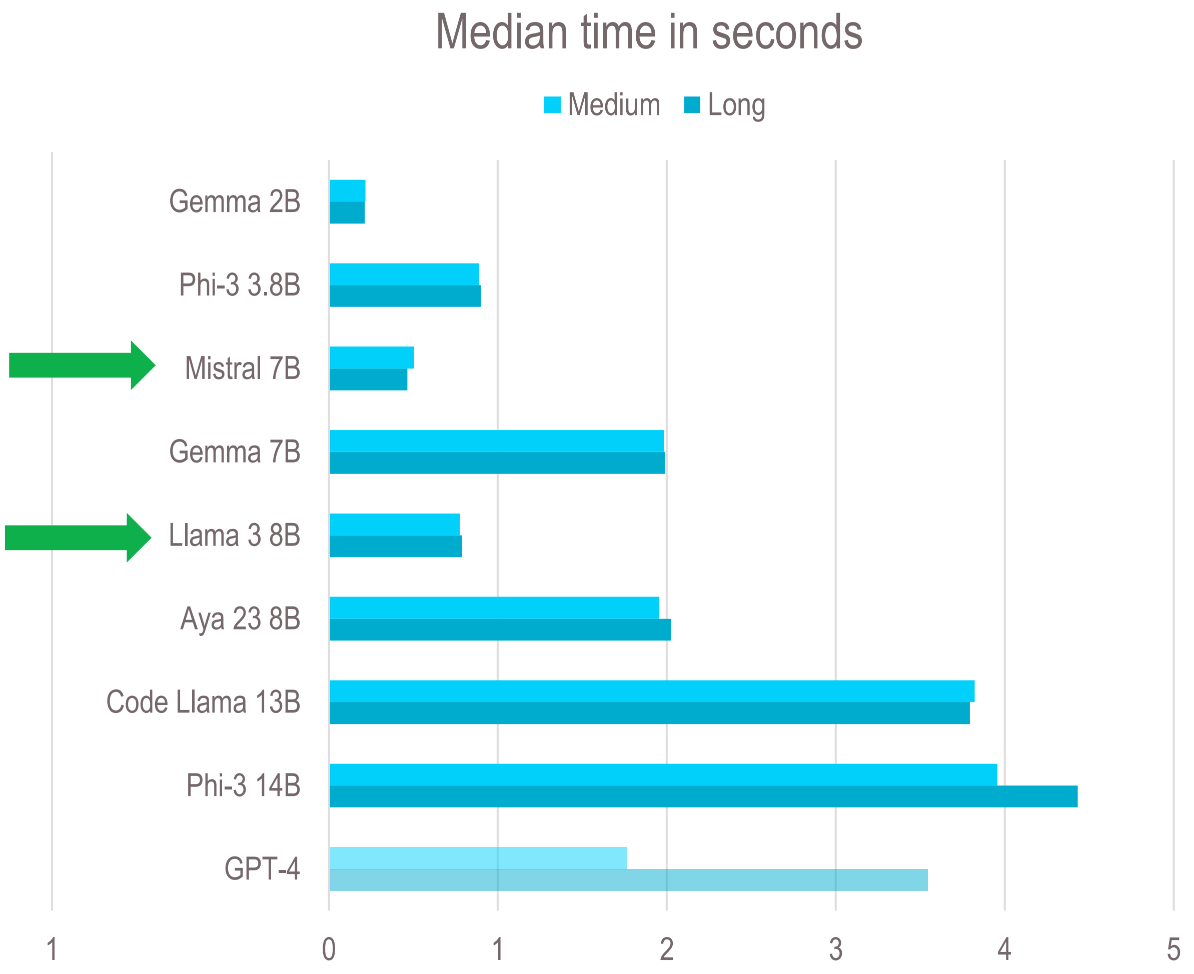




UNIVERSITY
OF TWENTE.

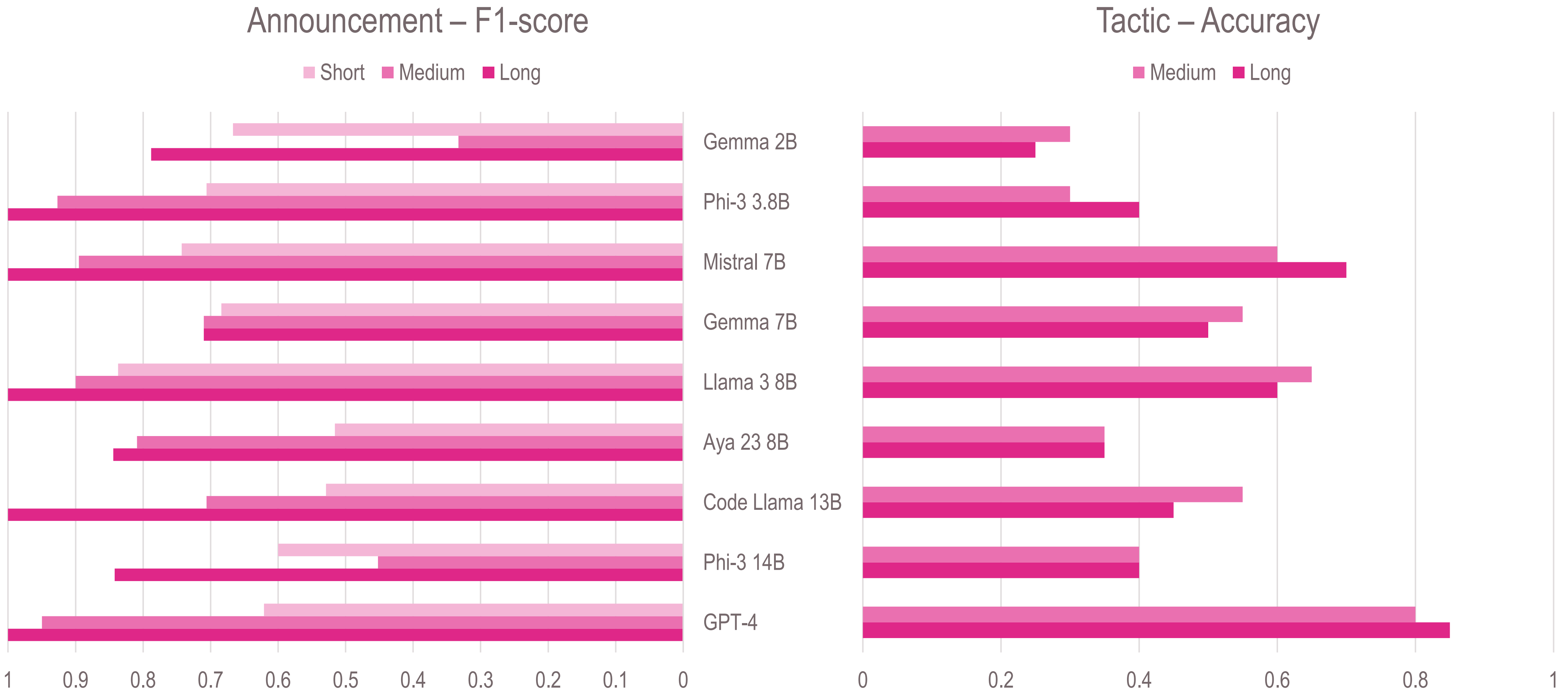
RESULTS – TACTIC DETECTION





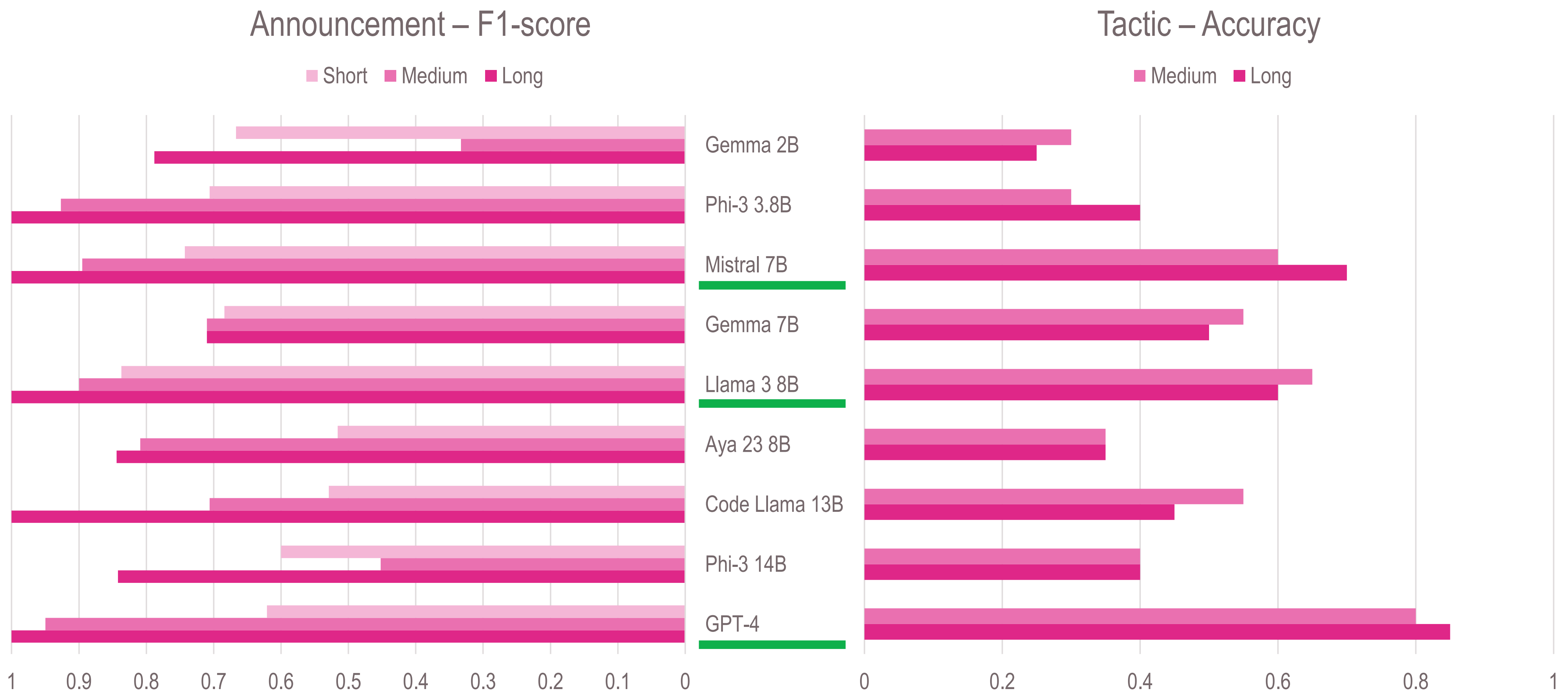
UNIVERSITY
OF TWENTE.

RESULTS – COMPARISON



UNIVERSITY
OF TWENTE.

RESULTS – COMPARISON



RESEARCH QUESTIONS

How can LLMs be integrated into the existing incident response workflow to streamline the triage process?

What suitable evaluation metrics should be used to assess the performance of LLMs in cybersecurity triage?

How do different LLMs compare in performance when optimizing the cybersecurity triage process?

RESEARCH QUESTIONS

- How can LLMs be integrated into the existing incident response workflow to streamline the triage process? ✓

- What suitable evaluation metrics should be used to assess the performance of LLMs in cybersecurity triage?

- How do different LLMs compare in performance when optimizing the cybersecurity triage process?

RESEARCH QUESTIONS

- How can LLMs be integrated into the existing incident response workflow to streamline the triage process? ✓

- What suitable evaluation metrics should be used to assess the performance of LLMs in cybersecurity triage? ✓

- How do different LLMs compare in performance when optimizing the cybersecurity triage process?

RESEARCH QUESTIONS

- How can LLMs be integrated into the existing incident response workflow to streamline the triage process? ✓

- What suitable evaluation metrics should be used to assess the performance of LLMs in cybersecurity triage? ✓

- How do different LLMs compare in performance when optimizing the cybersecurity triage process? ✓

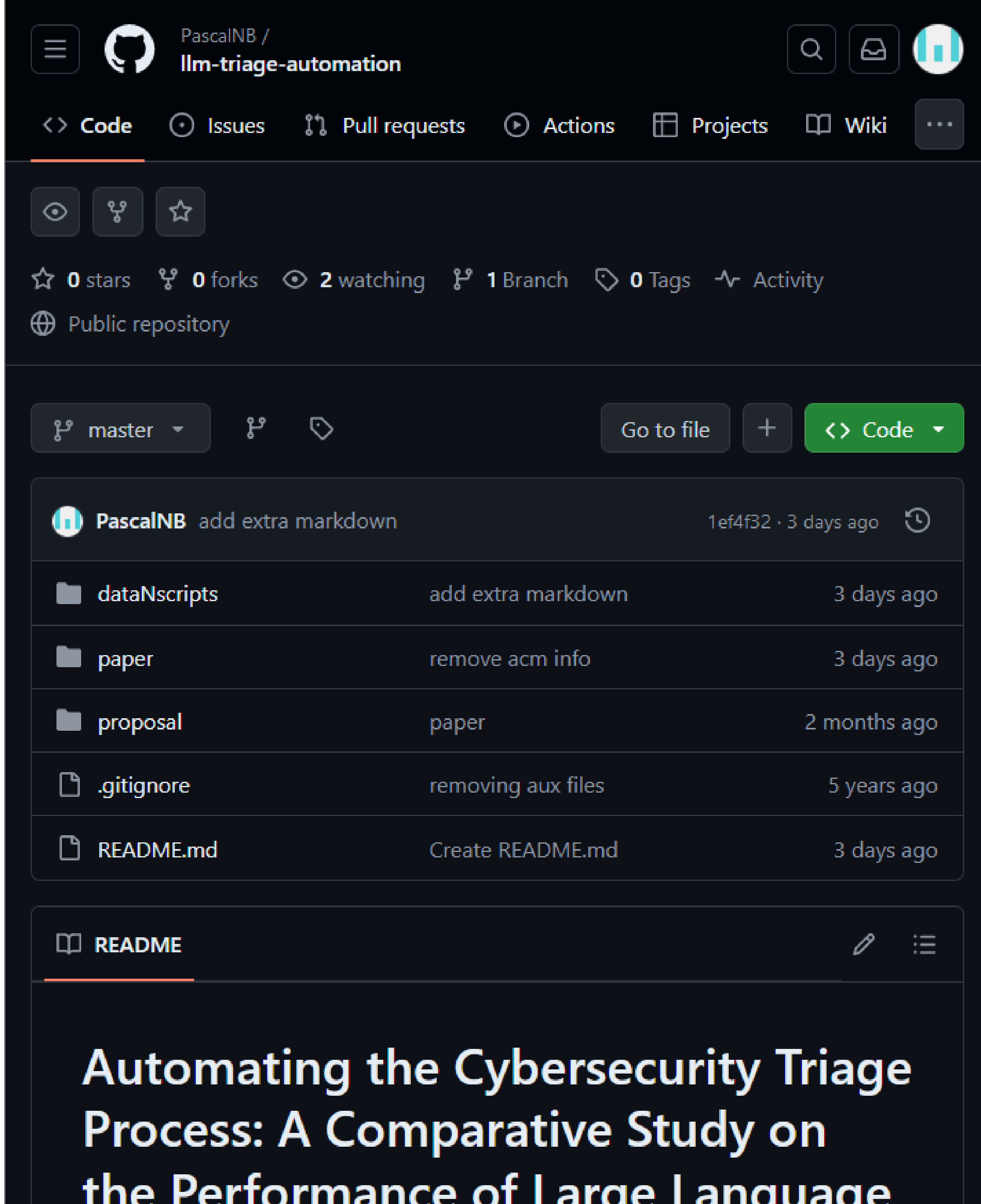
TAKEAWAYS

1. GPT-4 performed the best
   - Followed by Llama 3 and Mistral
   - Phi-3 3.8B good in simple tasks

2. Prompt size had no effect on time

3. Baseline towards further usage of LLMs
   - Defined key steps of triage
   - Identified optimizations
   - Evaluation framework

   → Other tasks and models



PascalNB add extra markdown 1ef4f32 · 3 days ago

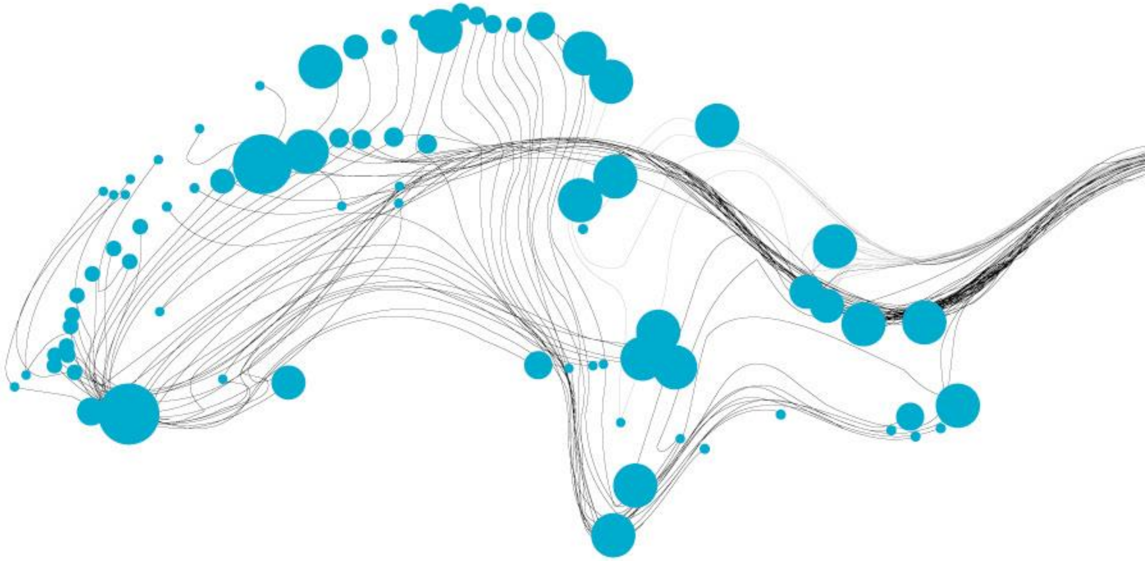
dataNscripts	add extra markdown	3 days ago
paper	remove acm info	3 days ago
proposal	paper	2 months ago
.gitignore	removing aux files	5 years ago
README.md	Create README.md	3 days ago

Automating the Cybersecurity Triage Process: A Comparative Study on the Performance of Large Language

EEMCS / TCS

AUTOMATING THE CYBERSECURITY TRIAGE PROCESS

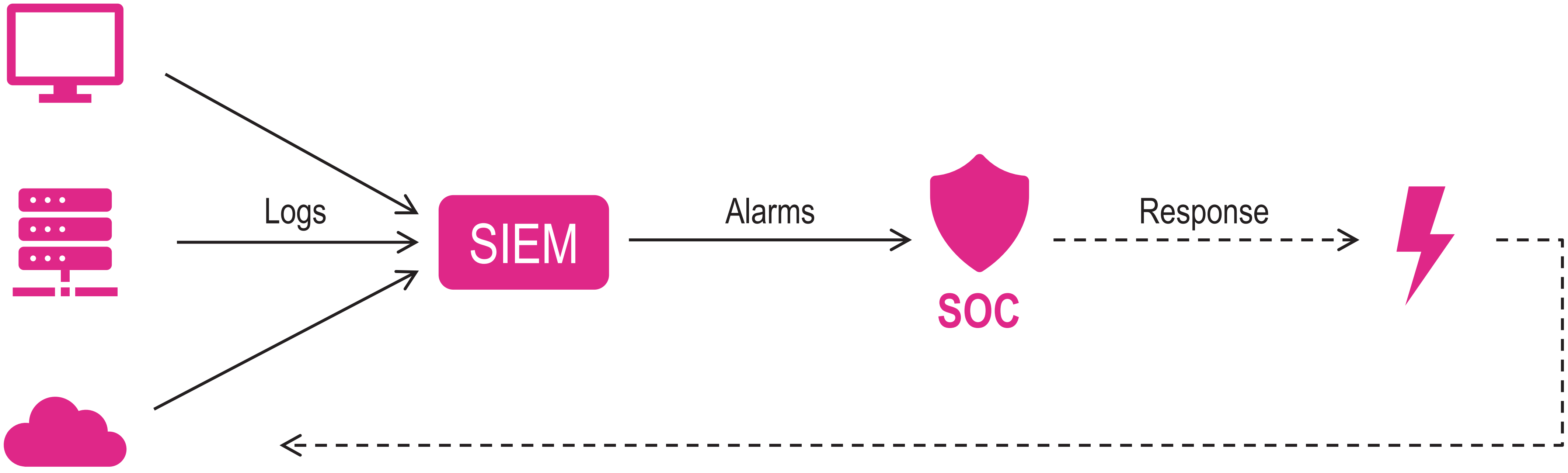
A COMPARATIVE STUDY ON THE PERFORMANCE OF LARGE LANGUAGE MODELS



PASCAL BAKKER
SUPERVISED BY JAIR SANTANNA
2024-07-05

UNIVERSITY
OF TWENTE.

THE INCIDENT RESPONSE WORKFLOW



Logs

SIEM

Alarms

SOC

Response

ERROR RATE

```
{
 “is_announcement”: True
}
```

✅

```
{
 “is_annoonment”: True
}
```

❌

EVALUATION METRICS

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$