Final Project & Optimization

DAY 3

Prepared by TARSOFT SDN BHD

TABLE OF CONTENTS

OBJECTIVE

Participants understand AI safety, privacy, optimization, and can present a mini project.

Day 2 Recap

What we learn yesterday?

- How to call Ollama via API.
- Integration with Python, JS, and mobile/web apps.
- Prompt engineering techniques.
- Why backend API is important

Million Dollar question:

How can we use this knowledge in our company projects?

Data Security & Privacy in Local AI

Data Security & Privacy in Local AI

- **Problem with Cloud AI:**
  - Data may leave your organization.
  - Regulatory concerns (GDPR, HIPAA, PDPA in Malaysia).
  - Risk of leaking **confidential documents** (contracts, medical records).
- **Ollama Advantage:**
  - Runs **locally** → no data leaves server.
  - Ideal for **banks, hospitals, government systems**.
- **Best Practices:**
  - Encrypt stored prompts & responses.
  - Use **access control** → only authorized users can send prompts.
  - Sanitize logs (remove PII).
  - Regularly update models to patch vulnerabilities.

Ollama Optimization Strategies

a. Model Quantization

- **Definition:** Reduce precision (e.g., FP32 → INT8) to make models smaller & faster.
- **Why use it?**
  - Saves memory.
  - Runs on lower hardware.
- **Example**:
  - LLaMA 2 FP16 → 13GB.
  - LLaMA 2 4-bit quantized → ~4GB.

b. Caching Strategies

- Store **frequent responses** to avoid recomputation.
- Example:
  - User asks: “What is RAG?”
  - Save response in DB → if same query, return cached answer.
- Saves **time** and **GPU** resources.b

c. Hardware Considerations

- **CPU-only mode:** Slower, but still usable.
- **GPU acceleration:** Needed for production scale.
- **RAM importance:** Large models (13B+) may need 16–32GB RAM.
- **Server setup suggestion**:
  - 1 × GPU (NVIDIA A100 or RTX 4090).
  - 64GB RAM.
  - SSD storage for fast access.

Hands-On 3

Integrate AI with a database (e.g., MySQL).

(points)

AI processes query → outputs natural language answer.

(points)

Mini Project

Project A: FAQ Bot

- AI answers customer FAQs from local database.
- Uses RAG with Ollama.

Project B: Document Analysis

- Upload PDF/Word.
- AI summarizes or extracts key info.

Project C: Auto-Reply System

- AI generates email/chat replies based on templates.
- Example: Customer complaint → AI suggests polite response.

Mini Project

Project Guidelines

- Work in teams of 2–3.
- Deliverables:
  - Working prototype (backend + Ollama).
  - Short demo (5–10 min).
- Criteria:
  - Functionality (does it work?).
  - Creativity (unique solution?).
  - Optimization (fast, lightweight?).

Mini Project

Project Presentation

- Each team presents:
  - Problem statement.
  - Demo (show working system).
  - Lessons learned.

Closing & Final Q&A

Final Q&A

- **Discuss:**
  - “What challenges did you face while building the project?”
  - “How can Ollama be used in your company’s workflow?”

Recap & Takeaways

You can now:

- Run Ollama locally.
- Integrate with applications.
- Optimize for performance.
- Build real projects with privacy in mind.

Closing Remarks

- AI is not magic → it’s **tools + data + creativity.**
- Keep experimenting with prompts & models.
- Encourage participants to **start a pilot project** in their workplace.

Thank You

Any Enquiries?

tarsoft.com.my