Meeting 16: UAS — Tugas Besar 2

AI-40X: Generative AI & Large Language Models

Hendri Karisma, M.T.
Dosen Teknik Informatika STMIK Tazkia
VP Engineering at jejakin.com, 2026

Semester 4 — 2025/2026

Outline

Apa yang Sudah Kita Pelajari (Meeting 9–15)

1. **Meeting 9 — Alignment & RLHF:** Bagaimana membuat LLM mengikuti instruksi manusia dan berperilaku aman (SFT, RLHF, DPO).
2. **Meeting 10 — AI Safety & Security:** Prompt injection, jailbreaking, red teaming, dan strategi pertahanan.
3. **Meeting 11 — Vector DB & Hybrid Search:** Embeddings, cosine similarity, semantic search, ChromaDB/FAISS.
4. **Meeting 12 — RAG & LangChain:** Retrieval-Augmented Generation, prompt templates, chains, memory, agents.
5. **Meeting 13 — GRC & Green AI:** Governance, Risk, Compliance, carbon footprint, efisiensi komputasi.
6. **Meeting 14 — AI Agents:** Tool use, ReAct pattern, multi-step reasoning, LangChain agents.
7. **Meeting 15 — Production Engineering:** FastAPI, Docker, RAGAS evaluation, MLOps, CI/CD, cost management.

Konsep Kunci Fase 2

Teori & Etika

- Alignment: membuat AI selaras dengan nilai manusia
- Safety: melindungi sistem dari serangan adversarial
- GRC: tata kelola, risiko, kepatuhan regulasi
- Green AI: efisiensi energi dan keberlanjutan

Teknis & Praktis

- RAG: menggabungkan retrieval dengan generasi
- Vector DB: penyimpanan dan pencarian semantik
- Agents: LLM yang bisa menggunakan tools
- Production: API, Docker, monitoring, MLOps

Benang Merah

Fase 2 mempersiapkan kalian untuk membangun sistem AI yang **lengkap**: bukan hanya bisa bekerja, tapi juga **aman**, **etis**, **terukur**, dan **siap production**.

Skenario: Smart Knowledge Assistant

Skenario

Anda adalah **AI Engineer** di sebuah organisasi. Anda ditugaskan membangun **"Smart Knowledge Assistant"** — sebuah chatbot cerdas yang bisa menjawab pertanyaan berdasarkan dokumen internal organisasi.

Pilihan Domain (pilih salah satu):

1. **Pendidikan:** Panduan akademik, kurikulum, FAQ mahasiswa, peraturan kampus.
2. **UMKM:** Panduan perizinan usaha, strategi pemasaran, regulasi UMKM.
3. **Kesehatan:** Informasi penyakit umum, panduan kesehatan masyarakat, FAQ rumah sakit.
4. **Pemerintahan:** Layanan publik, prosedur administrasi, regulasi daerah.

Kumpulkan minimal 10–20 dokumen (PDF/TXT) yang relevan dengan domain pilihan.

Dua Opsi Pengerjaan

Opsi A: RAG System End-to-End

- Bangun pipeline RAG lengkap:
  1. Ingest dokumen → chunking → embedding → simpan di Vector DB
  2. User bertanya → retrieve konteks → generate jawaban
- Gunakan LLM via API (Groq/OpenAI) atau lokal (Ollama)
- **Cocok untuk:** yang ingin fokus pada arsitektur retrieval dan kualitas jawaban

Opsi B: Fine-tuned Chatbot + RAG

- Fine-tune model kecil (misal: Llama 3.2 1B atau Phi-3.5 Mini) menggunakan data domain spesifik
- Tambahkan RAG sebagai knowledge base pendukung
- **Cocok untuk:** yang ingin eksplorasi fine-tuning (LoRA/QLoRA) dan punya akses GPU (Google Colab)

Catatan

Kedua opsi dinilai dengan bobot yang sama. Pilih berdasarkan minat dan ketersediaan resource.

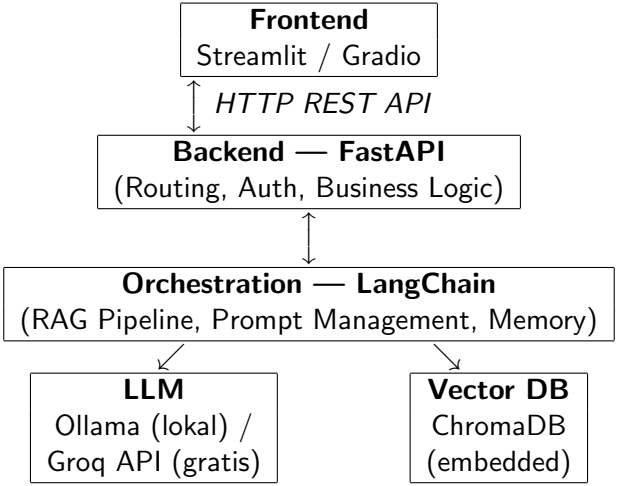
Komponen Teknis Wajib

1. **Backend API:**
   - Framework: FastAPI atau Flask.
   - Minimal 2 endpoint: `POST /chat` dan `POST /chat/stream` (streaming).
   - Integrasi dengan LLM (via API atau lokal).
2. **Knowledge Base (Vector DB):**
   - ChromaDB (recommended, paling mudah) atau FAISS.
   - Minimal 10 dokumen domain yang sudah di-index.
   - Chunking strategy yang jelas (chunk size, overlap).
3. **Frontend / User Interface:**
   - Streamlit (paling sederhana) atau Gradio.
   - Fitur: input pertanyaan, tampilkan jawaban, tampilkan sumber dokumen.
   - Bonus: streaming text (efek mengetik).

Komponen Laporan Wajib

1. **Ethics Report (Laporan Etika):**
   - **Bias Analysis:** Identifikasi potensi bias dalam data/model. Apakah ada kelompok yang dirugikan?
   - **Fairness:** Apakah sistem memberikan jawaban yang adil untuk semua tipe pertanyaan?
   - **Privacy:** Bagaimana data pengguna dilindungi? Apakah ada PII (Personally Identifiable Information) di knowledge base?
   - **Mitigasi:** Langkah apa yang diambil untuk mengurangi risiko etis?
2. **Sustainability Report (Laporan Keberlanjutan):**
   - **Carbon Footprint:** Estimasi emisi karbon dari training/inference (gunakan tools seperti CodeCarbon atau estimasi manual).
   - **Efisiensi:** Strategi untuk mengurangi konsumsi energi — model kecil, caching, batching.
   - **Green AI Practices:** Apa yang sudah dilakukan untuk membuat sistem lebih ramah lingkungan?

Diagram Arsitektur Sistem



Recommended Stack (Resource-Friendly)

Untuk mahasiswa dengan resource terbatas (no GPU, laptop biasa):

Komponen	**Rekomendasi**	**Alasan**
LLM	Groq API (free tier)	Gratis, cepat, mendukung Llama/Mixtral
LLM (alternatif)	Ollama (lokal)	Gratis, offline, privasi data
Vector DB	ChromaDB	`pip install chromadb`, tanpa server
Backend	FastAPI	Ringan, async, auto docs
Frontend	Streamlit	1 file Python = 1 web app
Orchestration	LangChain	Integrasi LLM + Vector DB mudah
Fine-tuning	Google Colab (free GPU)	T4 GPU gratis untuk LoRA

Penting

Kalian **tidak wajib** menggunakan GPU mahal atau API berbayar. Semua komponen di atas bisa dijalankan 100% gratis.

Rubrik Penilaian (Total 100%)

1. **Architecture & Technical Implementation (30%):**
   - Arsitektur sistem jelas dan modular
   - Backend API berfungsi dengan baik
   - Integrasi antar komponen (Frontend $\leftrightarrow$ API $\leftrightarrow$ LLM $\leftrightarrow$ VectorDB)
   - Error handling dan input validation
2. **RAG / Fine-tuning Quality (20%):**
   - Kualitas jawaban berdasarkan dokumen (faithfulness, relevancy)
   - Chunking strategy yang efektif
   - Evaluasi dengan metrik (RAGAS atau manual evaluation)
3. **Ethics & Sustainability Report (20%):**
   - Analisis bias, fairness, dan privacy yang mendalam
   - Estimasi carbon footprint dan strategi efisiensi
4. **Code Quality & Documentation (15%):**
   - Clean code, type hints, docstrings
   - README lengkap (cara install, cara jalankan, arsitektur)
   - Repository GitHub yang rapi
5. **Presentation & Live Demo (15%):**
   - Demo sistem berjalan lancar

Timeline Pengerjaan

Minggu	Fase	Aktivitas
9–12	Planning & Research	• Pilih domain dan kumpulkan dokumen • Riset arsitektur dan tools • Setup environment, initial prototype
13–14	Core Development	• Implementasi RAG pipeline / fine-tuning • Bangun backend API (FastAPI) • Bangun frontend (Streamlit)
15	Testing & Refinement	• Evaluasi kualitas RAG (RAGAS) • Tulis ethics & sustainability report • Perbaiki bug, optimasi performa
16	Presentation Day	• Presentasi + live demo • Tanya jawab teknis • Submit semua deliverables

Deliverables (Yang Dikumpulkan)

1. **Source Code:**
   - Repository GitHub (public atau private — invite dosen).
   - README.md yang lengkap: deskripsi, cara install, cara run, screenshot.
   - Struktur folder yang rapi.
2. **Laporan:**
   - Ethics Report (1–2 halaman).
   - Sustainability Report (1–2 halaman).
   - Bisa digabung dalam satu dokumen PDF.
3. **Slide Presentasi:**
   - Maksimal 15 slide.
   - Harus mengandung: arsitektur, demo screenshot, hasil evaluasi, analisis etika.
4. **Live Demo:**
   - Durasi: 10–15 menit (demo + tanya jawab).
   - Sistem harus bisa dijalankan saat presentasi.

Resource Gratis untuk Mahasiswa

LLM API (Gratis):

- **Groq:** `https://console.groq.com` — Llama 3, Mixtral, Gemma (sangat cepat, free tier generous).
- **Together AI:** `https://together.ai` — Banyak model open-source, free credits.
- **Google AI Studio:** `https://aistudio.google.com` — Gemini API gratis.

LLM Lokal (Offline, Privasi):

- **Ollama:** `https://ollama.com` — Install, lalu `ollama run llama3.2:1b`. Cukup ringan untuk laptop 8GB RAM.

GPU Gratis:

- **Google Colab:** T4 GPU gratis — cukup untuk fine-tuning LoRA model kecil.
- **Kaggle Notebooks:** P100 GPU gratis (30 jam/minggu).

Vector DB:

- **ChromaDB:** `pip install chromadb` — langsung jalan, tanpa setup server.

Tips Sukses Tugas Besar

1. **Mulai Sederhana, Iterasi:**
   - Minggu pertama: buat chatbot sederhana yang bisa menjawab 1 pertanyaan dari 1 dokumen.
   - Baru kemudian tambahkan fitur: multi-dokumen, streaming, UI yang lebih baik.
2. **Gunakan Stack yang Familiar:**
   - Jangan coba teknologi baru di menit terakhir.
   - Streamlit + FastAPI + ChromaDB + Groq API = stack yang sudah terbukti.
3. **Dokumentasi Sambil Jalan:**
   - Tulis README dan laporan sedikit demi sedikit, jangan menumpuk di akhir.
4. **Test dengan Pertanyaan Nyata:**
   - Siapkan 10–20 pertanyaan test dari domain kalian.
   - Evaluasi jawaban secara manual dan/atau dengan RAGAS.
5. **Jangan Lupa Ethics & Sustainability Report:**
   - Bobotnya 20% — jangan diabaikan!
   - Gunakan template yang sudah dipelajari di Meeting 13.

Perjalanan Semester Ini

Neural Networks → Deep Learning → Transformers → LLM → Prompt Engineering
↓
Fine-tuning → RAG → Agents → Safety & Ethics → Production

Pesan Akhir

Kalian sekarang memiliki fondasi untuk membangun aplikasi AI yang nyata. Dunia AI berkembang sangat cepat — kunci bertahan adalah **fundamental yang kuat** dan **kemauan untuk terus belajar**.

"The best way to learn AI is to build with AI."

Selamat mengerjakan Tugas Besar 2!