Meeting 8: UTS — Tugas Besar 1

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?

1. **Meeting 1:** Pengantar — Evolusi AI, dari rule-based ke neural, era LLM.
2. **Meeting 2:** Arsitektur Transformer — Self-Attention, Multi-Head Attention, Positional Encoding.
3. **Meeting 3:** Tokenisasi — BPE, WordPiece, SentencePiece; vocabulary vs. OOV.
4. **Meeting 4:** GPT dari Nol — Decoder-only Transformer, causal masking, next-token prediction.
5. **Meeting 5:** Training & Decoding — Loss function, optimizer, sampling (top-k, top-p, temperature).
6. **Meeting 6:** Scaling Laws — Kaplan, Chinchilla, emergent abilities.
7. **Meeting 7:** Fine-Tuning Efisien — LoRA, QLoRA, PEFT, pengantar etika AI.

Konsep Kunci yang Harus Dikuasai

Arsitektur & Teori:

- Transformer (Encoder vs. Decoder)
- Self-Attention:
  $\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$
- Causal Masking
- Positional Encoding
- Scaling Laws & Chinchilla

Praktik & Implementasi:

- Tokenisasi (BPE)
- Training loop (forward, loss, backward, step)
- Decoding strategies
- LoRA: $W_{\text{new}} = W + BA$
- QLoRA & quantization (NF4)
- Library: `transformers`, `peft`, `bitsandbytes`

Opsi A: NanoGPT from Scratch

Deskripsi

Bangun dan latih model GPT kecil dari nol menggunakan PyTorch.

- **Dataset** (pilih salah satu):
  - Shakespeare (klasik, bahasa Inggris)
  - Kumpulan Pantun Indonesia
  - Lirik Lagu Indonesia
- **Arsitektur:** Decoder-only Transformer (2–6 layer, 4–8 head, embedding 128–256).
- **Target:**
  - Training loss yang **konvergen** (plot loss curve).
  - Teks output yang **koheren**: ejaan benar, struktur baris wajar, gaya mirip data.
- **Referensi:** Karpathy's NanoGPT, video "Let's Build GPT".

Opsi B: Fine-Tune Model Kecil dengan LoRA

Deskripsi

Fine-tune pre-trained small model menggunakan LoRA/QLoRA pada dataset instruksi bahasa Indonesia.

- **Base Model** (pilih salah satu):
  - `HuggingFaceTB/SmolLM-360M` atau `SmolLM-1.7B`
  - `TinyLlama/TinyLlama-1.1B-Chat-v1.0`
- **Dataset Instruksi** (pilih/buat sendiri):
  - FAQ kampus STMIK Tazkia
  - Panduan UMKM / bisnis kecil
  - Dataset instruksi bahasa Indonesia (Hugging Face Hub)
- **Teknik:** QLoRA (4-bit quantization + LoRA adapter) untuk efisiensi memori.
- **Target:** Model bisa menjawab pertanyaan sesuai domain dataset dengan koheren.

Deliverables

1. **Kode Program** — Jupyter Notebook atau GitHub repository.
2. **Laporan Dokumentasi & Analisis** — PDF, 5–10 halaman:
   - Penjelasan arsitektur/metode yang dipilih.
   - Proses training: hyperparameter, hardware, waktu training.
   - Analisis hasil: loss curve, contoh output, evaluasi kualitas.
3. **Contoh Output Model** — Minimal 5 contoh generasi teks / jawaban.
4. **Analisis Etika & Bias** — Refleksi tertulis tentang bias potensial dan limitasi model.
5. **Presentasi Singkat** — 5–7 menit per kelompok (demo model).

Kriteria Penilaian

Komponen	Bobot	Keterangan
Kode Program	30%	Berjalan, bersih, terdokumentasi
Dokumentasi & Analisis	25%	Lengkap, sistematis, insightful
Kualitas Output Model	25%	Loss konvergen, output koheren
Analisis Etika & Bias	10%	Reflektif, kritis, solutif
Presentasi	10%	Jelas, demo berjalan, Q&A
Total	**100%**	

Catatan Penting

- Plagiarisme = nilai 0. Kode boleh adaptasi dari referensi, tapi **harus dipahami dan didokumentasikan**.
- Kerja kelompok (maks. 2 orang) atau individu.

Tips Pengerjaan & Sumber Daya

Infrastruktur (Gratis):

- **Google Colab** — GPU T4 gratis (15 GB VRAM). Cukup untuk QLoRA model $\leq$ 3B.
- **Kaggle Notebooks** — GPU P100 / T4, 30 jam/minggu.
- **Lightning AI Studios** — Alternatif dengan GPU gratis.

Pre-trained Models:

- Hugging Face Hub
- Cari model dengan tag `id` (Indonesian) atau model kecil (`SmolLM`, `TinyLlama`).

Library Penting:

- `torch` — Framework deep learning.
- `transformers` — Model & tokenizer dari Hugging Face.
- `peft` — LoRA, QLoRA, dan metode PEFT lainnya.
- `bitsandbytes` — Quantization 4-bit/8-bit.
- `datasets` — Loading & preprocessing data.
- `trl` — `SFTTrainer` untuk fine-tuning chat model.

Referensi Video:

- Karpathy — "Let's Build GPT"

Timeline & Pengumpulan

Tahap	Deadline
Pemilihan opsi (A/B) & pembentukan kelompok	Akhir Meeting 8
Progress report (dataset & arsitektur)	Meeting 9
Pengumpulan kode + laporan + output	Meeting 13
Presentasi & demo	Meeting 14

Format Pengumpulan:

- Upload ke **Google Classroom / LMS** (link akan dibagikan).
- Struktur folder:
  - `code/` — Notebook (.ipynb) atau link GitHub repo.
  - `report/` — Laporan (PDF).
  - `outputs/` — Contoh output model (screenshot / teks).

Pertanyaan?

Silakan ajukan pertanyaan sekarang atau hubungi dosen/asisten melalui forum kelas.