PERTEMUAN 08 · TEORI

LLM & *Generative AI*

Cara kerja Transformer, GPT, dan Claude — dar
summarizer untuk bisnis. Pintu masuk ke Blok

Teori | 100 menit | CPMK: C3, C4, C5

Tujuan Pembelajaran

Setelah UTS, kita masuk Blok 2: AI yang *menghasilkan* konten, bukan hanya mengklasifikasi atau memprediksi. Setelah sesi ini, mahasiswa diharapkan mampu:

1. Menjelaskan evolusi pemrosesan bahasa dari model statistik, RNN/LSTM, hingga Transformer
2. Menguraikan alur kerja Transformer: tokenisasi, embedding, self-attention, dan prediksi token berikutnya
3. Membedakan tahap pelatihan LLM: pre-training, instruction tuning, dan alignment
4. Menjelaskan parameter generasi (context window, temperature, top-p) dan dampaknya terhadap output
5. Mengidentifikasi keterbatasan dan risiko LLM (halusinasi, bias, privasi) beserta mitigasinya
6. Memetakan aplikasi LLM untuk bisnis, terutama chatbot dan summarizer, termasuk kapan LLM *bukan* pilihan yang tepat

🧭 **Peta Blok 2:** P-08 (konsep LLM) → P-09 (prompt engineering & API) → P-10/P-11 (RAG & chatbot dokumen) → P-12 (etika & regulasi) → P-13/P-14 (workshop proyek UAS).

1. Dari NLP Klasik ke LLM

LLM (*Large Language Model*) bukan teknologi yang muncul tiba-tiba. Ia adalah hasil dari rangkaian terobosan dalam memodelkan bahasa — masing-masing memecahkan keterbatasan pendahulunya.

Era statistik

N-gram & Bag-of-Words / TF-IDF

Menghitung frekuensi kata. Sederhana dan cepat, tetapi tidak memahami urutan panjang maupun makna ("bank" sungai vs "bank" keuangan sama saja).

2013

Word embedding (Word2Vec)

Kata direpresentasikan sebagai vektor padat; kata bermakna mirip berdekatan di ruang vektor. Namun satu kata tetap punya satu vektor, apa pun konteksnya.

2014–2016

RNN / LSTM + mekanisme attention

Memproses teks berurutan (seperti di P-06). Kelemahan: lambat karena harus berurutan, dan informasi di awal kalimat panjang mudah "terlupakan".

2017

Transformer — "Attention Is All You Need"

Membuang rekurensi dan hanya mengandalkan self-attention. Seluruh token diproses paralel, sehingga pelatihan di GPU jauh lebih efisien dan bisa diskalakan.

2018–2020

BERT, GPT-2, GPT-3: era pre-training skala besar

Model dilatih pada korpus teks raksasa lalu diadaptasi ke banyak tugas. GPT-3 (175 miliar parameter) menunjukkan kemampuan few-shot: cukup diberi contoh di prompt tanpa melatih ulang.

2022 →

Model berbasis instruksi & asisten percakapan

> *LLM pada dasarnya adalah mesin prediksi token berikutnya yang dilatih pada skala sangat besar. Kemampuan "menalar", merangkum, dan menulis kode muncul dari tugas sederhana ini.*
>
> — *Inti konsep yang akan kita bedah hari ini*

2. Cara Kerja Transformer

Ketika Anda mengetik pertanyaan ke chatbot, teks melewati rangkaian tahap berikut sebelum satu token jawaban dihasilkan. Proses ini diulang token demi token (*autoregressive*) sampai jawaban selesai.

LANGKAH 1
Tokenisasi
Teks dipecah menjadi token (potongan kata/sub-kata) lalu diubah menjadi ID angka

→

LANGKAH 2
Embedding + Posisi
Tiap token jadi vektor; informasi urutan ditambahkan (positional encoding)

→

LANGKAH 3
Self-Attention
Setiap token menimbang relevansi token lain dalam konteks (multi-head)

→

LANGKAH 4
Feed-Forward × N lapis
Blok attention + feed-forward (dengan residual & normalisasi) ditumpuk puluhan lapis

→

LANGKAH 5
Softmax → Token
Distribusi probabilitas atas seluruh kosakata; satu token dipilih

Self-attention: inti Transformer

Untuk setiap token, model menghitung tiga vektor: **Query (Q)** — "apa yang sedang saya cari?", **Key (K)** — "apa yang saya tawarkan?", dan **Value (V)** — "informasi yang saya bawa". Kecocokan Q dan K menentukan seberapa besar V dari token lain diambil.

Q·K^T: skor kecocokan antar token · **$\sqrt{d_k}$**: skala agar nilai stabil · **softmax**: mengubah skor menjadi bobot yang berjumlah 1 · **V**: informasi yang dijumlahkan berdasarkan bobot

🔍 **Contoh intuitif:** pada kalimat *"Budi memberikan buku itu kepada Ani karena dia membutuhkannya"*, self-attention membantu model menghubungkan kata "dia" dengan "Ani" atau "Budi" berdasarkan konteks seluruh kalimat — sesuatu yang sulit dilakukan RNN pada kalimat panjang. *Multi-head* berarti ada banyak "kepala" attention paralel yang masing-masing menangkap jenis hubungan berbeda (sintaks, referensi, makna).

Tiga varian arsitektur

VARIAN	CARA MEMBACA TEKS	CONTOH	CENDERUNG UNGGUL UNTUK
Encoder-only	Dua arah (melihat kata kiri dan kanan)	BERT	Klasifikasi, pencarian, embedding teks
Decoder-only	Searah (hanya melihat token sebelumnya) dan memprediksi token berikutnya	Keluarga GPT; mayoritas LLM percakapan modern	Generasi teks, chatbot, penulisan kode
Encoder-decoder	Encoder memahami input, decoder menghasilkan output	T5	Terjemahan, ringkasan

Catatan: detail arsitektur internal model komersial tidak selalu dipublikasikan oleh penyedianya; tabel di atas menggambarkan pola umum yang dikenal publik.

3. Bagaimana LLM Dilatih

LLM yang Anda pakai sebagai asisten dibentuk melalui beberapa tahap. Memahami tahap ini menjelaskan mengapa model bisa fasih berbahasa namun tetap keliru soal fakta.

Pre-training

- Tugas: memprediksi token berikutnya pada korpus teks sangat besar (web, buku, kode)
- Self-supervised: label berasal dari teks itu sendiri, tanpa anotasi manual
- Biaya komputasi sangat besar; hasilnya *base model* yang fasih tetapi belum "patuh"

Instruction tuning (SFT)

- Fine-tuning pada pasangan instruksi → jawaban yang dibuat/dikurasi manusia
- Mengubah model dari "melanjutkan teks" menjadi "menjawab permintaan"
- Jauh lebih murah daripada pre-training

TAHAP 3

Alignment

- Menyelaraskan perilaku model dengan preferensi dan nilai manusia: membantu, jujur, tidak berbahaya
- Pendekatan umum: RLHF (umpan balik manusia), pelatihan berbasis preferensi, atau prinsip tertulis seperti *Constitutional AI*
- Menekan output berbahaya, namun tidak menjamin akurasi fakta

⚠ **Implikasi penting:** karena objektif dasarnya adalah *memprediksi teks yang masuk akal*, bukan *memverifikasi kebenaran*, LLM bisa menghasilkan jawaban yang meyakinkan tetapi salah. Inilah akar masalah halusinasi yang dibahas di bagian 5.

4. Mengendalikan Output: Token, Context & Sampling

TOKEN

Satuan teks LLM

CONTEXT WINDOW

Memori kerja model

Batas jumlah token (prompt + riwayat + jawaban) yang bisa diproses sekaligus. Informasi di luar jendela ini tidak "diketahui" model dalam percakapan.

TEMPERATURE

Tingkat keacakan

Rendah → output konsisten dan deterministik (cocok untuk ekstraksi data, klasifikasi). Tinggi → lebih bervariasi dan kreatif (brainstorming, copywriting).

TOP-P & MAX TOKENS

Batas sampling & panjang

Top-p hanya memilih dari token dengan probabilitas kumulatif teratas. Max tokens membatasi panjang jawaban sehingga biaya dan latensi terkendali.

Coba sendiri: efek temperature

Model memberi skor pada setiap kandidat token berikutnya, lalu softmax dengan temperature *T* mengubahnya menjadi probabilitas. Geser slider dan perhatikan distribusinya. Angka di bawah hanya ilustrasi, bukan keluaran model sungguhan.

Ibu kota Indonesia adalah ___

Jakarta	82.1%
Nusantara	13.6%
Surabaya	2.5%
Bandung	1.2%
Bali	0.7%

Temperature 1.0

Temperature sedang: token teratas masih dominan, tetapi alternatif punya peluang muncul.

5. Lanskap Model & Memilih LLM

Beberapa keluarga model yang umum digunakan: **GPT** (OpenAI), **Claude** (Anthropic), **Gemini** (Google), serta model *open-weight* seperti **Llama** (Meta) dan **Mistral** yang bisa dijalankan sendiri. Versi dan kemampuan model berubah cepat — selalu cek dokumentasi resmi penyedia, jangan bergantung pada daftar yang sudah usang.

KRITERIA	PERTANYAAN PENENTU
Kualitas & bahasa	Seberapa baik model menangani Bahasa Indonesia dan domain Anda? Uji dengan data nyata, bukan hanya benchmark.
Biaya	Harga per token input/output × volume. Model lebih kecil sering cukup untuk tugas sederhana.
Latensi	Apakah pengguna menunggu jawaban real-time (chatbot) atau diproses batch (ringkasan malam hari)?
Privasi & kepatuhan	Apakah data boleh dikirim ke API pihak ketiga? Bagaimana kebijakan retensi data? Relevan dengan UU PDP (dibahas di P-12).
Kontrol & lisensi	Model berbayar via API (mudah dipakai) vs open-weight (kontrol penuh, tetapi perlu infrastruktur dan perhatikan lisensi).

6. Keterbatasan & Risiko LLM

Menggunakan LLM di bisnis berarti mengelola risikonya. Berikut risiko utama beserta mitigasi yang akan kita praktikkan di sesi-sesi berikutnya.

RISIKO	PENJELASAN	MITIGASI
Halusinasi	Jawaban fasih namun salah atau mengarang (termasuk sumber/sitasi fiktif).	RAG (P-10), instruksi "jawab hanya dari konteks", verifikasi manusia untuk keputusan penting.

Pengetahuan usang	Model hanya tahu data sampai batas waktu pelatihannya (*knowledge cutoff*).	Memberi dokumen terkini di prompt, RAG, atau tool pencarian.
Bias	Mewarisi bias dari data latih; dapat memperlakukan kelompok tertentu secara tidak adil.	Pengujian pada berbagai skenario, panduan di system prompt, audit berkala (P-12).
Prompt injection	Teks dari pengguna atau dokumen berisi instruksi jahat yang membajak perilaku model.	Pisahkan instruksi dan data, batasi hak akses tool, validasi output.
Kebocoran data	Data pribadi atau rahasia perusahaan terkirim ke layanan eksternal.	Anonimisasi, pilih penyedia dengan kebijakan data jelas, patuhi UU PDP.
Biaya & latensi	Panggilan API per token bisa membengkak pada volume tinggi.	Pilih ukuran model yang sesuai, batasi max tokens, cache jawaban umum.

7. Aplikasi LLM untuk Bisnis

APLIKASI 1

Chatbot Layanan Pelanggan / Internal

- Alur: pertanyaan pengguna → (ambil dokumen relevan) → LLM + system prompt → jawaban
- Nilai bisnis: respons 24/7, menurunkan beban tim support
- Kunci: jawaban berbasis dokumen resmi (RAG), jalur eskalasi ke manusia, batasan topik
- Metrik: tingkat penyelesaian, kepuasan pengguna, tingkat eskalasi, akurasi jawaban

APLIKASI 2

Summarizer

- Alur: dokumen panjang → pecah (chunking) → ringkas per bagian → gabungkan
- Contoh: ringkasan rapat, ulasan pelanggan, laporan, kontrak
- Kunci: format output terstruktur, panjang terkontrol, cek kesetiaan terhadap sumber
- Metrik: kelengkapan, kesesuaian fakta, penilaian pengguna

Ekstraksi & Klasifikasi Teks

- Mengubah teks bebas menjadi data terstruktur (JSON): faktur, email, tiket
- Mengklasifikasi sentimen atau kategori tanpa melatih model baru
- Gunakan temperature rendah untuk konsistensi

Pembuatan Konten & Kode

- Draf email, deskripsi produk, materi pembelajaran, soal ujian
- Asisten penulisan kode dan dokumentasi
- Kunci: hasil dianggap *draf* yang ditinjau manusia

⚖️ **LLM bukan jawaban untuk semua masalah.** Untuk prediksi pada data tabular (churn, harga properti, fraud), model ML klasik seperti Random Forest atau XGBoost dari Blok 1 biasanya lebih akurat, lebih murah, dan lebih mudah dijelaskan. Gunakan LLM ketika inputnya *bahasa alami* atau outputnya perlu berupa teks.

8. Latihan Pemahaman

Klik opsi untuk memeriksa jawaban; penjelasan akan muncul setelah Anda memilih.

No. 1 Komponen Transformer manakah yang memungkinkan setiap token menimbang relevansi token lain dalam kalimat, sekaligus membuat pemrosesan bisa paralel?

A Lapisan rekuren (LSTM) yang memproses token satu per satu

B Lapisan Max Pooling seperti pada CNN

C Mekanisme self-attention

D Fungsi aktivasi ReLU pada output layer

A Karena model sengaja berbohong untuk menyenangkan pengguna

B Karena model dilatih memprediksi token yang paling mungkin muncul, bukan memverifikasi kebenaran fakta

C Karena context window selalu terlalu kecil

D Karena temperature selalu diatur ke 0

No. 3 Anda membangun sistem yang mengekstrak nomor faktur dan total tagihan dari email ke format JSON. Pengaturan temperature mana yang paling sesuai?

A Rendah (mendekati 0) agar output konsisten dan dapat diprediksi

B Tinggi (≥ 1,5) agar hasil lebih kreatif

C Tidak berpengaruh pada tugas ekstraksi

D Harus diacak per permintaan agar model tidak bias

No. 4 Urutan tahap pelatihan LLM asisten yang paling tepat adalah...

A Alignment → Pre-training → Instruction tuning

B Instruction tuning → Alignment → Pre-training

C Pre-training → Alignment → Instruction tuning

No. 5 Sebuah perusahaan ingin chatbot HR menjawab pertanyaan karyawan berdasarkan kebijakan internal yang sering diperbarui. Pendekatan yang paling tepat dan hemat biaya adalah...

A Melatih LLM dari nol setiap kali kebijakan berubah

B Menggunakan RAG: mengambil bagian dokumen kebijakan yang relevan lalu memberikannya ke LLM sebagai konteks

C Mengandalkan pengetahuan bawaan LLM tanpa dokumen apa pun

D Menaikkan temperature agar jawaban lebih lengkap

No. 6 Sebuah perusahaan logistik ingin memprediksi apakah pelanggan akan berhenti berlangganan dari tabel berisi 20 fitur numerik dan kategorikal. Pilihan mana yang paling masuk akal?

A Memasukkan seluruh tabel ke LLM sebagai teks dan meminta prediksi

B Melatih LLM baru dengan miliaran parameter

C Menggunakan model ML klasik (misalnya Random Forest / XGBoost) dengan evaluasi F1/AUC; LLM opsional untuk menjelaskan hasil

D Menggunakan CNN karena data berisi banyak angka

9. Proyek Akhir AI

dengan demo langsung. Dua dari 11 topik proyek berbasis Generative AI dan langsung memakai materi hari ini: **Chatbot Akademik berbasis RAG** dan **Generator Soal Otomatis dari PDF**.

PROYEK UAS · BOBOT 35%

Panduan Proyek Akhir Kecerdasan Buatan

11 topik proyek, pembagian peran kelompok, format laporan dan panduan per bab, rubrik penilaian, timeline, serta ketentuan integritas akademik.

Buka Panduan Proyek Akhir →

📖 PERSIAPAN SESI BERIKUTNYA

🔧 **P-09: Prompt Engineering & API Integration**
Zero-shot, few-shot, chain-of-thought; integrasi LLM via API ke aplikasi Python. Persiapan: daftar akun OpenAI atau Anthropic (ada free tier) dan siapkan API key.

📄 **Vaswani et al. (2017) — Attention Is All You Need**
Makalah asli Transformer. Cukup baca abstrak dan bagian arsitektur model (Bab 3) untuk memperdalam materi hari ini.

📚 **Géron Ch. 16 — Natural Language Processing with RNNs and Attention**
Menghubungkan RNN/LSTM dari P-06 dengan mekanisme attention dan Transformer.

ISI HALAMAN

Tujuan Pembelajaran

1. Dari NLP Klasik ke LLM
2. Cara Kerja Transformer

4. Token, Context & Sampling

5. Lanskap & Memilih Model

6. Keterbatasan & Risiko

7. Aplikasi Bisnis

8. Latihan Pemahaman

9. Proyek Akhir AI

NAVIGASI SESI

← P-07: Review & Persiapan UTS

→ P-09: Prompt Engineering & API

🏠 Kembali ke Beranda

🚀 Panduan Proyek Akhir

SKOR LATIHAN ANDA

0 / 6

soal dijawab benar

Mulai kerjakan soal latihan!

INFO PERTEMUAN

Tipe
Teori

Durasi

C3, C4, C5

Topik

Transformer, GPT, Claude; chatbot & summarizer untuk bisnis

Referensi

Vaswani et al. (2017) · Géron Ch. 16