ChatGPT

Pertempuran Raksasa AI di Dunia Maya

Bayangkan sekelompok **raksasa supercerdas** sedang berkumpul di arena maya: ada *GPT* (OpenAI) dengan tubuh robot Transformer, *Gemini* (Google DeepMind) si robot futuristik bersenjata MoE (Mixture of Experts), *Claude* (Anthropic) si ksatria bijak berbaju Transformer, dan puluhan model lainnya seperti LLaMA (Meta) dan Mistral. Mereka saling pandang, siap bertarung dalam pertarungan tak kasat mata yang penuh fakta teknis. Setiap raksasa ini punya kemampuan uniknya sendiri–satu meriam Multi-Head Attention di kepala, yang lain panel-panel pakar MoE di balik punggungnya, semua berlomba unjuk kekuatan pemahaman bahasa, penalaran, dan coding. Pertempuran ini bukan pertarungan pisau atau pedang sungguhan, tapi **benchmark dan evaluasi** sebagai medali kemenangan. Mari kita cermati masing-masing lawannya dalam cerita seru nan lucu ini, dengan analogi imajinatif tapi tetap fokus fakta.

Arsitektur Para Juara

Bayangkan **GPT** si robot idola kita sebagai prajurit Transformer klasik. Namanya memang *Generative Pre-trained Transformer*, artinya otaknya adalah arsitektur *Transformer* (dengan lapisan self-attention) yang pernah diperkenalkan oleh Vaswani dkk. [1]. GPT dilatih besar-besaran dengan memprediksi kata berikutnya dalam korpus teks raksasa, lalu **finetune** lagi pakai RLHF (Reinforcement Learning from Human Feedback) untuk jadi lebih "baik dan patuh" [2].

Di sisi lain, **Claude** sang ksatria juga bertopang pada senjata Transformer. [3] Dia pakai pelatihan serupa: belajar dulu tanpa pengawas (unsupervised), lalu dibimbing **RLHF** dan **Constitutional AI** untuk membuat "aturan" sendiri lewat konstitusi model [3]. Konstitusi model ini ibarat kitab suci yang mengatur tingkah laku Claude agar tetap aman. Jadinya Claude mahir bicara sopan santun dan kurang banyak menolak permintaan berguna—hasil riset terbaru mereka [3] [4].

Sementara itu, **Gemini** muncul dengan teknologi baru ala Google DeepMind. Versi terbaru (Gemini 1.5 dan 2.0) menerapkan **arsitektur campuran**: bukan hanya Transformer biasa, tapi dioptimalkan dengan *Mixture-of-Experts (MoE)* [5]. Artinya, otak Gemini terdiri atas banyak *pakar* berbeda, dan hanya sebagian yang aktif setiap kali ia berpikir. Teknik MoE ini memungkinkan jumlah parameter luar biasa besar (bayangkan tentara-pakar tersembunyi) tanpa memperlambat kecepatan. Gemini bahkan melatih modelnya untuk ingat konteks lebih panjang—sampai jutaan token! [5]. Contohnya, Google berhasil membuat Gemini Ultra dengan jendela konteks hingga 1 juta token sekaligus MoE canggih [5] [6].

Model-model terbaru lain juga ikut tren MoE. Misalnya **Mistral AI** meluncurkan *Mixtral 8×7B*, yang total punya 46,7 miliar parameter—tapi hanya ~12,9 miliar digunakan per token karena sifat sparsennya [7]. Menurut mereka, Mixtral ini bisa mengalahkan Llama 2 70B dari Meta pada banyak tes sambil jauh lebih irit inference [8] [7]. Intinya, sebagian besar juara saat ini masih menggunakan struktur *decoder-only Transformer*, namun inovasi MoE dan peningkatan konteks membuat mereka seperti pahlawan super dengan kekuatan ekstra.

Ukuran Pasukan: Berapa Banyak Parameter?

Setiap raksasa tentu menghitung banyak tentara (parameter) di tubuhnya. **GPT-3** pernah punya sekitar **175 miliar** parameter—sudah riuh sekali di dunia AI. GPT-4 diperkirakan *sepuluh kali lipat* lebih besar.

Rumornya, GPT-4 awalnya punya hingga ~1,8 triliun parameter [9], meski OpenAI tak pernah konfirmasi resmi. Analisis pihak ketiga seperti Klu.ai mengisyaratkan angka sekitar 1-1.8 triliun pula [10]. Versi terbaru GPT-4.1 sendiri dirancang agar lebih efisien, tapi tetap punya miliaran neuron tersembunyi di balik layarnya.

Gemini juga mewarisi tradisi Google yang membuat model sangat masif. Informasi publik agak terbatas, tapi mereka bahkan melatih model yang disebut "Secret Project" dengan triliunan token latihan. Misalnya, Gemini 1.5 diperkenalkan dengan "jendela konteks 1 juta token" berkat arsitektur MoE [5]. Pengumuman Google terkini menyebut versi Pro dan Ultra, yang bisa melibatkan triliunan parameter tersembunyi. Intinya, Gemini Ultra didesain untuk "tugas kompleks", artinya ukurannya bisa menandingi atau melampaui GPT terkuat.

Claude 3 (keluarga Haiku, Sonnet, Opus) tidak mengumumkan secara langsung jumlah parameter, tapi petunjuknya ada. Claude 3 Sonnet "diperkirakan" sekitar 175 miliar parameter [11] —sebanding dengan GPT-3—sementara Opus, sang andalan, mungkin melampaui itu. Anthropic lebih menekankan kemampuan dan efisiensi daripada nominal parameter, tapi dari tes-tes benchmark (lihat nanti) kita tahu mereka juga melatih model-model besar.

Di luar nama-nama utama tadi, **Meta LLaMA** punya ukuran pasukan berbeda: Llama 2 diluncurkan dalam varian 7B, 13B, dan 70B parameter [12]. Kemudian Llama 3 (rilis 2024) hadir dalam versi 8B dan 70B [13]. Yang mengejutkan, edisi Llama 3.1 terbaru memiliki tambahan model raksasa 405B parameter [14]! Ini menandakan Meta sangat serius mengangkangi para pesaing. Model-model kecil seperti Mistral asli hanya 7B (untuk model utamanya) [15], tapi mereka bisa "menangkam" performa model raksasa dengan desain pintar. Intinya, arena ini dipenuhi pasukan parameter berkali lipat triliun di jajaran teratas, sedang pemain lain memilih gerilya pakar demi kecepatan.

Lumbung Data dan Pelatihan

Pertempuran memang butuh bekal: apa yang dipelajari model sebelum "bertarung"? **GPT** dilatih di lautan data teks. Wikipedia menyebut GPT-4 belajar memprediksi token berikutnya pada kumpulan teks raksasa (data publik dan *berlisensi* pihak ketiga) [2]. Setelah fase awal itu, baru deh OpenAI menjalani pelatihan lanjutan dengan *RLHF* untuk memperhalus perilaku model [2] [16]. Jadi GPT pertama mahir tulis sihir bahasa secara umum, lalu diberikan koreksi langsung oleh tim manusia demi hasil lebih aman dan berguna.

Google **Gemini** juga minum pelatihan besar—dari web, buku, kode, dan sejumlah data internal Google (seperti halnya pendahulunya PaLM). Meskipun detail pasti tidak diungkap, mereka menyebut punya dataset yang lebih luas daripada PaLM2 [17]. Selain itu, Google serius menggunakan RLHF dan evaluasi internal ketat agar jawabannya sesuai kebijakan perusahaan. Mereka juga eksperimen dengan *Multimodal Live API*, meaning Gemini dilatih untuk mengerti gambar, suara, video; dan sanggup menggunakan alat seperti Google Search saat menjawab [18]. Sederhananya, Gemini punya bekal data super banyak, termasuk percakapan internal (Bard), kode (AlphaCode 2), dan segalanya di alam Google.

Sementara **Claude** menonjol dengan "Constitutional AI"-nya. Setelah fase pelatihan awal (unsupervised learning di data publik besar), mereka menyuntikkan latihan dari preferensi manusia (RLHF) sambil mengikuti nilai-nilai konstitusi model [3]. Jadi Claude "hati-hati berbicara" karena ia punya buku aturan internal. Data latihnya tidak diumumkan satu-satu, tapi Anthropic sering menekankan etika dan keamanan: seperti melatih model yang tidak mudah diajak bicara jahat (hasil red-teaming mereka) atau melatih agar lebih cerdas dan patuh tanpa melewati batas [3] [4]. Hasilnya, Claude unggul menjaga nada sopan dan lebih jarang menolak pertanyaan berguna daripada sebelumnya [19].

Model lain: Meta LLaMA memilih data publik saja (model open-ish), dan Meta tak memberi akses penuh ke RLHF. Mistral melatih modelnya di data web terbuka dan repositori kode, lalu menyediakan model instruksi lebih kecil (like Mixtral Instruct) hasil fine-tuning khusus. Pendeknya, semua "prajurit" ini mewarisi lahan data besar. Perbedaan cara latih (RLHF, CAI, atau sebaliknya *open training*) lah yang membuat taktik masing-masing berbeda. Tapi satu kesamaan: semua dilatih dengan *teknik self-supervised* dulu (belajar sendiri dari teks/data), barulah disempurnakan manual lewat feedback.

Arena Benchmark: Siapa Paling Cerdas?

Pertarungan sesungguhnya terjadi di arena benchmark. Tiap model bertarung lewat angka-akurasi di MMLU, BIG-bench, HumanEval (coding), dan sejumlah tes lain. Hasilnya seperti piala yang dipajang: siap menang, siap kalah.

Pada **MMLU (Massive Multitask Language Understanding)**—ujian kognitif campuran dari pelajaran sekolah hingga perguruan tinggi—**Gemini 3 Pro** terbaru memimpin dengan akurasi 90,10% [20]! Peringkat kedua ada Gemini 3 Flash ~88,6%, diikuti *Claude Opus 4.1* sekitar 87,9% [20]. OpenAI mengklaim *GPT-4.1* meraih 90,2% [21], jadi sebenarnya masih seimbang dengan jawara Google dan Anthropic (semua di kisaran 90%). Sisi serunya: tes MMLU jadi makin sulit ditingkahi, tampaknya semua model top sudah amat paham pengetahuan umum.

Di **GPQA (Graduate-level Google-Proof Q&A)** untuk pemecahan masalah rumit, Google Geminis unggul lagi. Misalnya, GPT-4.1 cuma ~66% di tingkatan soal Sulit (Diamond) [22]. Sebaliknya, Gemini 2.5 Pro sempat dikabarkan meraih *state-of-the-art* di GPQA. Artinya, Google si robot raksasa tampak jago sains lanjutan. Begitu pula tes matematika kompleks seperti GSM8K: Claude 3 Opus juaranya (katanya, dua kali lipat akurasi dari Claude 2.1), sementara Google dan OpenAI saling dorong ke angka 80-an persen [4].

Pada bidang **coding**, perbedaan mencolok muncul. Tes SWE-Bench (Soal Coding Tingkat Menengah) menunjukkan GPT-4.1 raih ~55% [22], sementara **Gemini 2.5 Pro** 63,8% dan **Claude 3.7 Sonnet** sekitar 62–63% [22]. Begitu pula *HumanEval* (tes Python). Berdasarkan perbandingan komunitas, *Claude 3 Opus* menoreh ~84,9% di HumanEval, jauh lebih tinggi daripada GPT-4 klasik yang 67,0% [23]. (Catatan: GPT-4 Turbo/4o terbaru memang mengejar dengan ~91% di bench serupa [24], tetapi pada saat yang sama Opus mengejar, jadi duel kode masih ketat.) Intinya, Claude tampaknya dilatih ekstra coding —mungkin karena banyak contoh dan penyempurnaan persoalan algoritmik—sehingga menang besar atas GPT-4 generasi awal. Sementara Google bersaing dengan alat multi-modal (bisa membaca kode dan web secara dinamis), yang membantu meraih hasil tinggi juga.

Secara keseluruhan, setiap jagoan punya arena keahliannya. **GPT-4.1** misalnya, mendapat sorotan karena ekspansi *1-juta token context*, sehingga bisa mengintip dokumen superpanjang (berkilometer), mendekati kemampuan Gemini [6]. **Claude 3** unggul di banyak tugas kompleks karena dilatih khusus, terutama coding dan reasoning, tapi memang agak lebih fokus tata krama jawaban. **Gemini** banyak tools (terhubung Google Search) dan MoE-nya memberi nilai apik di hampir semua benchmark. Sebut saja **LLaMA 3.1-405B** Meta; meski tidak komersial luas, katanya bisa mengalahkan Gemini Pro dan Claude Sonnet di banyak tes [13] —menunjukkan model terbuka pun tetap jempolan kalau cukup besar.

Kunci: Pertarungan ini sangat kompetitif dan dinamis. Setiap saat ada model baru meluncur (GPT-4.1, Gemini 3, Claude 4, LLaMA 3.1, dll.) dengan skor lebih tinggi. Misalnya, OpenAI bilang GPT-4.1 capai *90,2% MMLU* [21], sementara Google dan Anthropic sudah lari mendekati angka itu. Pada coding, klaim GPT-4 Turbo vs Claude berbalik-balik, menunjukan latih-latihan intensif masih berlangsung. Bersama-

sama, mereka membuat standar kecerdasan buatan terus naik—seperti laga akhir musim sepak bola di mana semua tim memperlihatkan taktik dan kekuatan paling mutakhir.

Kesimpulan Cerita

Di kisah pertarungan *super-Large Language Models* ini, setiap tokoh muncul layaknya pahlawan super di medan laga maya. Arsitekturnya bersaing antara seragam Transformer klasik dengan inovasi Mixture-of-Experts, jumlah pasukannya (parameter) dihitung hingga triliunan, dan dataset latihannya seluas samudra. Ketika ranah benchmark berbicara, Gemini, Claude, dan GPT saling salip di indikator berbeda. Yang pasti, setiap langkah mereka bertangan-tangan manusia (RLHF) atau isi konstitusi model demi output lebih "cerdas" dan aman.

Bagi mahasiswa IT dan profesional AI, nonton aksi ini seru sekaligus informatif. Kita bisa membayangkan empat sahabat kompetitif (GPT, Gemini, Claude, LLaMA/Mistral) saling lempar guyonan teknologi: "Aku Transformer sejati!", "Tidak, aku Mix-of-Experts modern!", "Aku punya instruksi konstitusi, nih!"—sambil tak lupa menunjuk grafik benchmark. Di balik gaya "bahasa bayi lucu" ini, tersembunyi fakta serius: pembangunan model makin cepat, algoritma dan data terus berkembang. Arena *pertempuran raksasa* ini menandai bahwa kita berada di era kemajuan AI yang sangat menarik dan gegap gempita [3] [21].

Referensi: Informasi teknis di atas didukung oleh dokumen dan laporan resmi: model card Anthropic (Claude) [3], Wikipedia dan riset terkini tentang GPT-4 [2] [16], berita peluncuran Claude 3 [4], dan hasil benchmark terbaru seperti MMLU dan coding benchmarks [20] [23]. Semua angka dan fakta telah dikutip dari sumber tersebut.

---

[1] [2] [9] [10] [16] GPT-4 - Wikipedia
https://en.wikipedia.org/wiki/GPT-4

[3] anthropic.com
https://www.anthropic.com/claude-2-model-card

[4] [19] Introducing the next generation of Claude \ Anthropic
https://www.anthropic.com/news/claude-3-family

[5] [18] Gemini (language model) - Wikipedia
https://en.wikipedia.org/wiki/Gemini_(language_model)

[6] [21] [22] OpenAI says GPT-4.1 sets new 90%+ standard in MMLU reasoning benchmark
https://www.rdworldonline.com/openai-claims-gpt-4-1-sets-new-90-standard-in-mmlu-reasoning-benchmark/

[7] [8] [15] Mixtral of experts | Mistral AI
https://mistral.ai/news/mixtral-of-experts

[11] The Number of Parameters of GPT-4o and Claude 3.5 Sonnet
https://aiexpjourney.substack.com/p/the-number-of-parameters-of-gpt-4o

[12] [13] [14] Llama (language model) - Wikipedia
https://en.wikipedia.org/wiki/Llama_(language_model)

[17] Is this public knowledge? Size of dataset: PaLM2 = 3.6 trillion tokens ...
https://www.reddit.com/r/Bard/comments/18mv4bk/is_this_public_knowledge_size_of_dataset_palm2_36/

[20] MMLU Pro
https://www.vals.ai/benchmarks/mmlu_pro

23 24 Gpt4 comparison to anthropic Opus on benchmarks - Community - OpenAI Developer Community
https://community.openai.com/t/gpt4-comparison-to-anthropic-opus-on-benchmarks/726147