

Membaca Indeks AI 2025: Menjembatani Prinsip dan Praktik dalam AI yang Bertanggung Jawab

Penulis : Tjok Istri Sintawati
Editor : Ayom Mratita Purbandani



Pendahuluan

Pesatnya penggunaan kecerdasan artifisial (AI) sepanjang tahun 2012 hingga 2026 dibarengi dengan berbagai dampak yang mendisrupsi kehidupan manusia, mulai dari penggunaan deepfake untuk menghasilkan gambar intim hingga eksploitasi identitas individu yang sudah meninggal melalui AI chatbot. Rentetan peristiwa ini menegaskan urgensi satu hal penting, bahwa pengembangan AI mesti dilakukan dengan bertanggung jawab. Ide ini kemudian menghidupi prinsip **Responsible Artificial Intelligence (RAI)**, yang belakangan kian menjadi topik penting dalam AI. Urgensi ini diangkat dalam salah satu bab dari **The AI Index 2025 Annual Report** yang disusun oleh **Stanford Institute for Human-Centered Artificial Intelligence**. Report breakdown ini bermaksud untuk menyediakan ringkasan singkat dari bab 'Responsible AI' yang menjadi salah satu bahasan dalam Artificial Intelligence Index Report 2025. Adapun, tahun 2024 menandai babak bersejarah dalam RAI dengan terjadinya peningkatan kerja sama global dalam tata kelola AI yang berbasis pada prinsip bertanggung jawab (Tabel 1). Dimensi utama yang membentuk RAI sendiri terdiri dari privasi dan tata kelola data, transparansi dan keterjelasan, keamanan dan keselamatan, serta keadilan yang dijabarkan dalam Tabel 2.

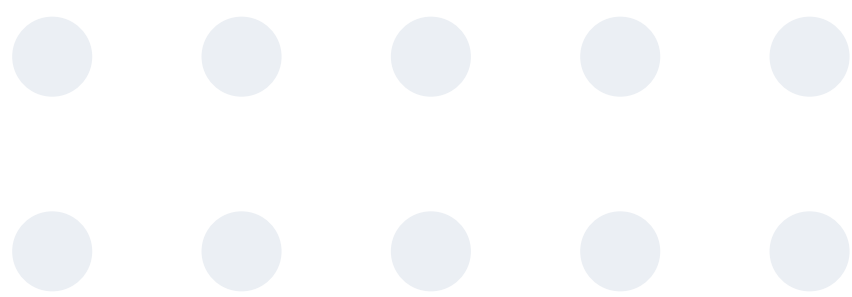
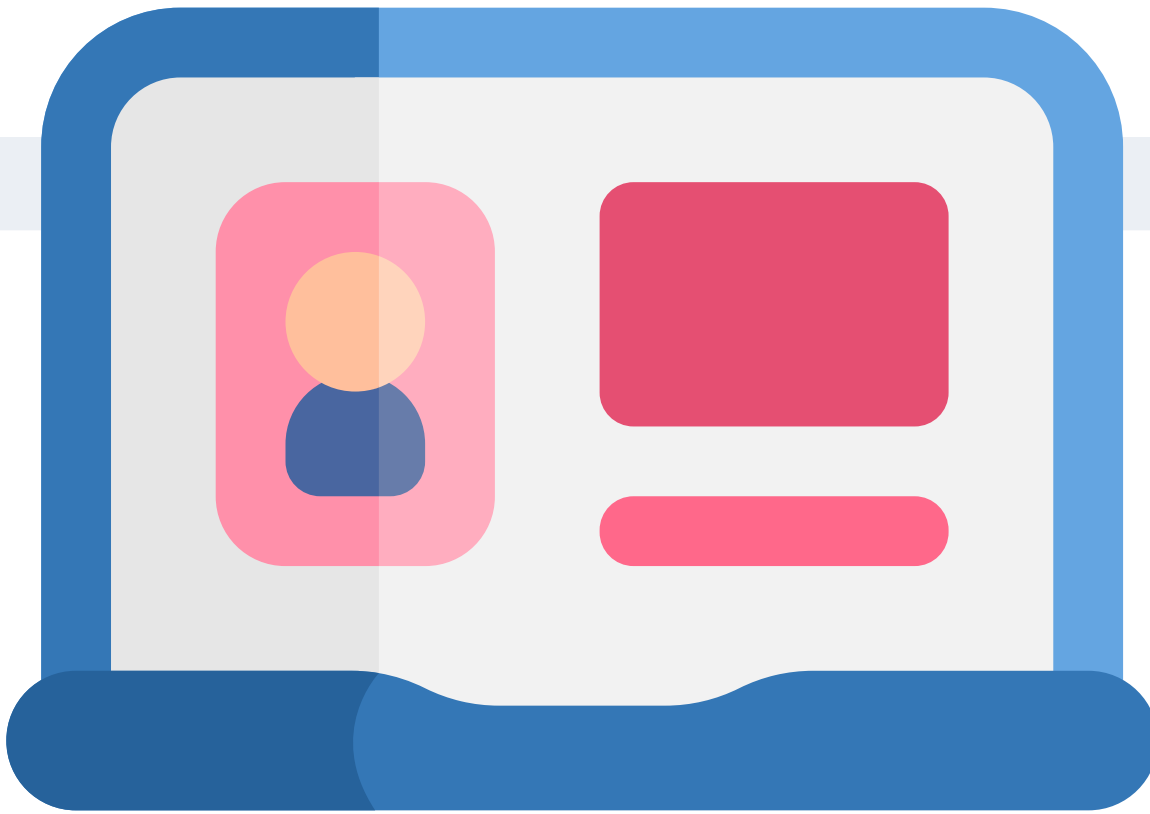


Table 1. Beberapa Tonggak Kebijakan RAI dalam Lingkup Global

Waktu	Stakeholder	Penjelasan
Mei 2024	OECD	Prinsip-prinsip AI-nya telah disempurnakan untuk mencerminkan kemajuan terbaru, khususnya melalui perkembangan teknologi AI baru-baru ini termasuk AI tujuan umum dan generatif, dalam tata kelola AI. Prinsip-prinsip ini menekankan pertumbuhan inklusif, transparansi, kemampuan menjelaskan, serta menghormati supremasi hukum, hak asasi manusia, dan nilai-nilai demokrasi.
Sep 2024	United Nations	Laporan <i>Governing AI for Humanity</i> telah diperbarui dengan menyoroti upaya untuk membangun mekanisme tata kelola AI global. Laporan tersebut merekomendasikan pengembangan cetak biru untuk mengatasi risiko terkait AI dan menyerukan kepada organisasi standar nasional dan internasional, perusahaan teknologi, masyarakat sipil, dan pembuat kebijakan untuk berkolaborasi dalam standar AI.
Oct 2024	G7	Kerja Sama AI G7 menegaskan kembali komitmen terhadap pasar AI yang adil dan terbuka di tengah pertumbuhan AI yang pesat. Mereka menyoroti kebutuhan mendesak akan pendekatan regulasi yang terkoordinasi.
Nov 2024	International Network of AI Safety Institutions	Jaringan ini menyatukan organisasi-organisasi teknis yang berkomitmen untuk memajukan keamanan AI dengan membantu pemerintah dan masyarakat memahami risiko sistem AI canggih sekaligus mengusulkan solusi.

Table 2. Explanation of RAI Dimensions

Dimensi RAI	Definisi
Privasi	Hak individu atas kerahasiaan, anonimitas, dan perlindungan keamanan data pribadi mereka.
Tata kelola data	Penetapan kebijakan, prosedur, dan standar untuk memastikan kualitas, akses, dan perizinan data.
Keadilan dan bias	Memastikan bahwa algoritma dirancang dan diterapkan dengan desain yang mengurangi bias atau diskriminasi, sembari mempertimbangkan ragam kebutuhan seluruh pemangku kepentingan.



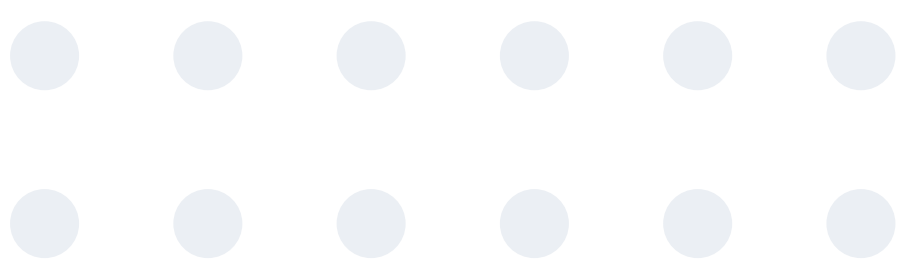
Dimensi RAI	Definisi
Transparansi	Keterbukaan tentang bagaimana sistem AI bekerja dan bagaimana sistem AI diterapkan, dipantau, dan dikelola mulai dari fase pembuatan hingga fase operasional.
Keterjelasan	Kemampuan untuk memahami dan mengartikulasikan alasan di balik keluaran sistem AI.
Keamanan dan keselamatan	Koherensi sistem AI untuk memerangi ancaman, meminimalkan kerugian akibat penyalahgunaan, dan mengatasi risiko keamanan yang melekat.

Privacy and Data Governance

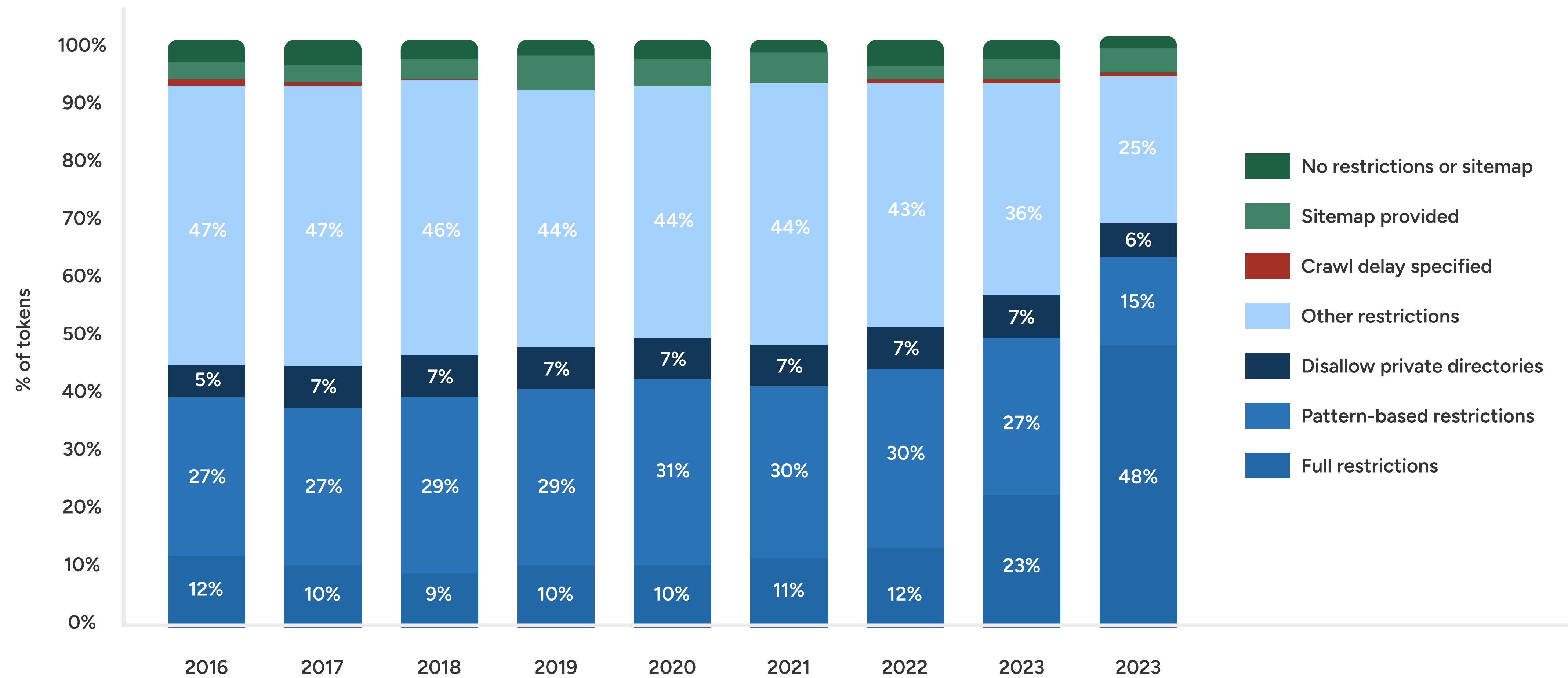
Aspek privasi memberi penekanan pada hak individu pada kerahasiaan, anonimitas, proteksi data pribadi, serta informed consent terkait penggunaan data. Tata kelola data, di lain sisi, berkaitan erat dengan kebijakan, prosedur, serta standar yang ditetapkan oleh organisasi untuk menjamin penggunaan data dari segi kualitas, keamanan, dan penggunaan etis. Keduanya memiliki kesinambungan terkait dengan proses pengolahan serta penggunaan data secara etis. Kondisi dari kedua aspek ini tergambar, di antaranya, melalui audit terhadap lisensi dan atribusi dataset serta persetujuan data protokol.

Penelitian dari Longpre dkk¹ yang dikutip dalam laporan ini menguji 1.800 dataset teks yang umum digunakan dalam pelatihan model AI lalu menemukan bahwa lebih dari 70% dataset dari situs-situs hosting dataset populer tidak menyediakan informasi lisensi yang memadai. sementara 50% dari lisensi justru salah dikategorikan. Kesalahan atribusi terhadap lisensi dari dataset memuat dampak negatif untuk pengembangan AI. Hal ini meliputi risiko pelanggaran undang-undang hak cipta, kebijakan penggunaan data, hingga regulasi privasi.

Protokol persetujuan data pada tahun 2024 menjadi lebih ketat dibandingkan tahun sebelumnya. Penelitian lain oleh Longpre dkk.² berusaha melakukan audit longitudinal terkait dengan protokol konsen terhadap 14.000 domain website yang digunakan dalam pelatihan dataset AI seperti C4, RefinedWeb, dan Dolma. Banyak di antaranya yang menerapkan protokol konsen lebih ketat misalnya melalui pembaruan ke robots.txt files serta ketentuan layanan secara eksplisit. Proporsi token dari web teratas C4 dengan pembatasan penuh misalnya meningkat dari angka 10% di tahun 2017 menjadi 48% pada tahun 2024 sebagaimana tertera dalam Gambar 1. Eskalasi tren terhadap protokol konsen diasumsikan berhubungan erat dengan masalah hukum khususnya fair use.



Gambar 1. Persentase di Web Teratas C4 Berdasarkan Pembatasan Penuh Robots.Txt Pada Tahun 2016-2024



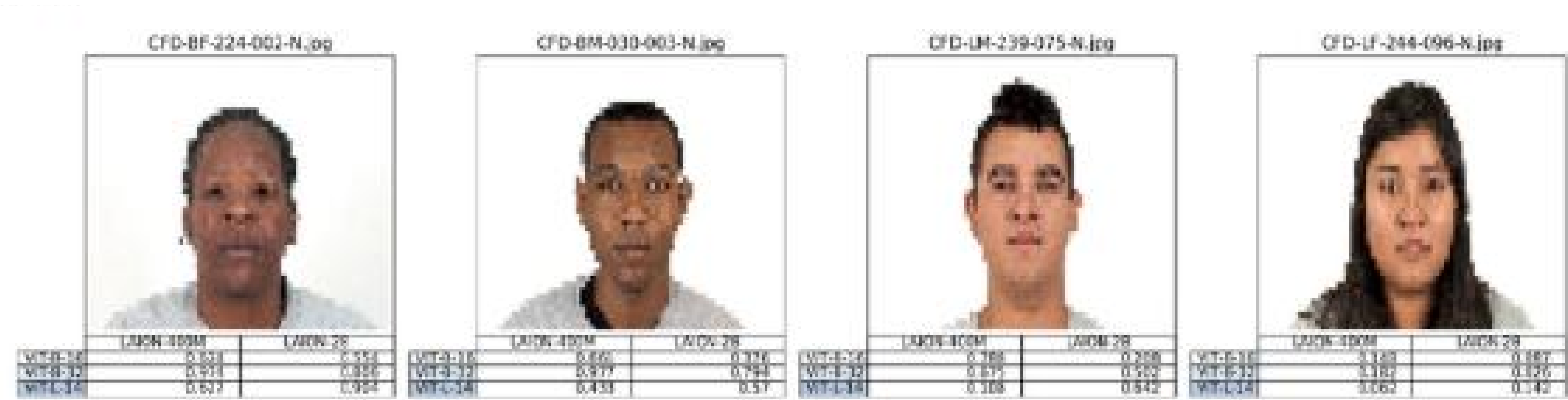
Sumber: Longpre et al., 2025 | Chart: 2025 AI Index Report

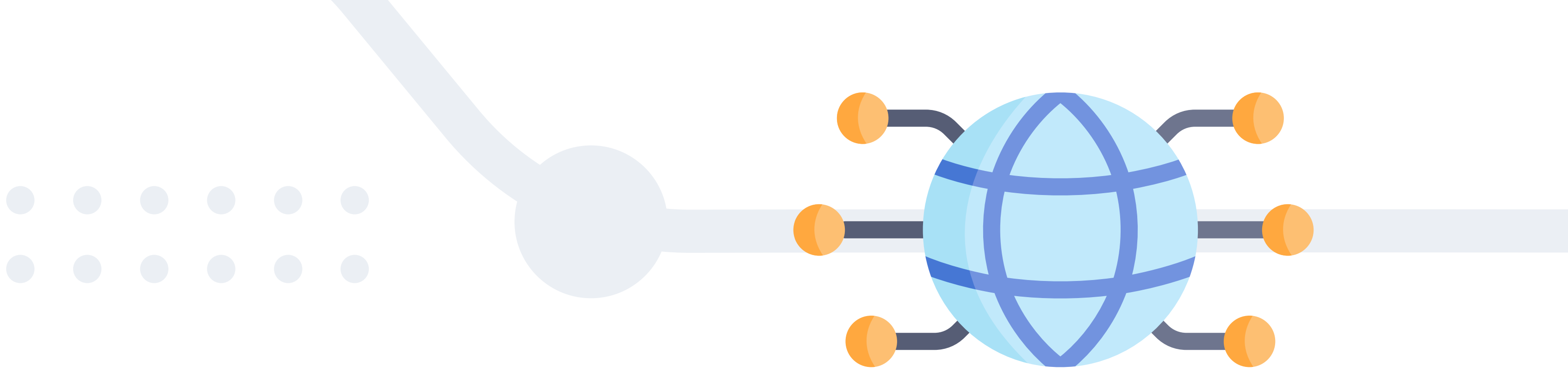
Keadilan dan Bias

Keadilan menjadi tujuan yang berusaha dicapai dalam RAI. Sejalan dengan hal tersebut, memastikan AI terhindar dari bias dan diskriminasi terhadap sekumpulan individu atau kelompok tertentu menjadi sebuah langkah penting. Dalam mengukur hal tersebut, laporan ini melakukan asesmen terhadap 2 hal yakni validitas terhadap klasifikasi ras pada Vision-Language Models (VLMs) serta bias implisit dalam Large-Language Models (LLMs) yang secara eksplisit tidak bias.

Pengujian terhadap 14 VLMs yang dilatih melalui dataset LAION-400M dan LAION-2B menggunakan Chicago Face Dataset (CFD) menghasilkan temuan bahwa dataset lebih besar justru memperkuat bias rasial. Kendatipun, dataset lebih besar mengurangi misidentifikasi dari model manusia terhadap entitas bukan manusia seperti gorila dan orangutan. Dalam Gambar 2, disajikan data berupa probabilitas dari wajah yang diidentifikasi untuk memenuhi beberapa label spesifik berdasarkan kelompok demografis berubah secara simultan dari berbagai model. Laki-laki berkulit hitam dan Latin secara tidak proporsional diklasifikasikan sebagai kriminal dengan probabilitas meningkat hingga sekitar 69% saat ukuran datanya bertambah dari 400 juta menjadi 2 miliar.

Gambar 2. Kumpulan Wajah dan Kemungkinannya Diklasifikasikan Sebagai "Kriminal" Berdasarkan Ukuran Model dan Dataset.





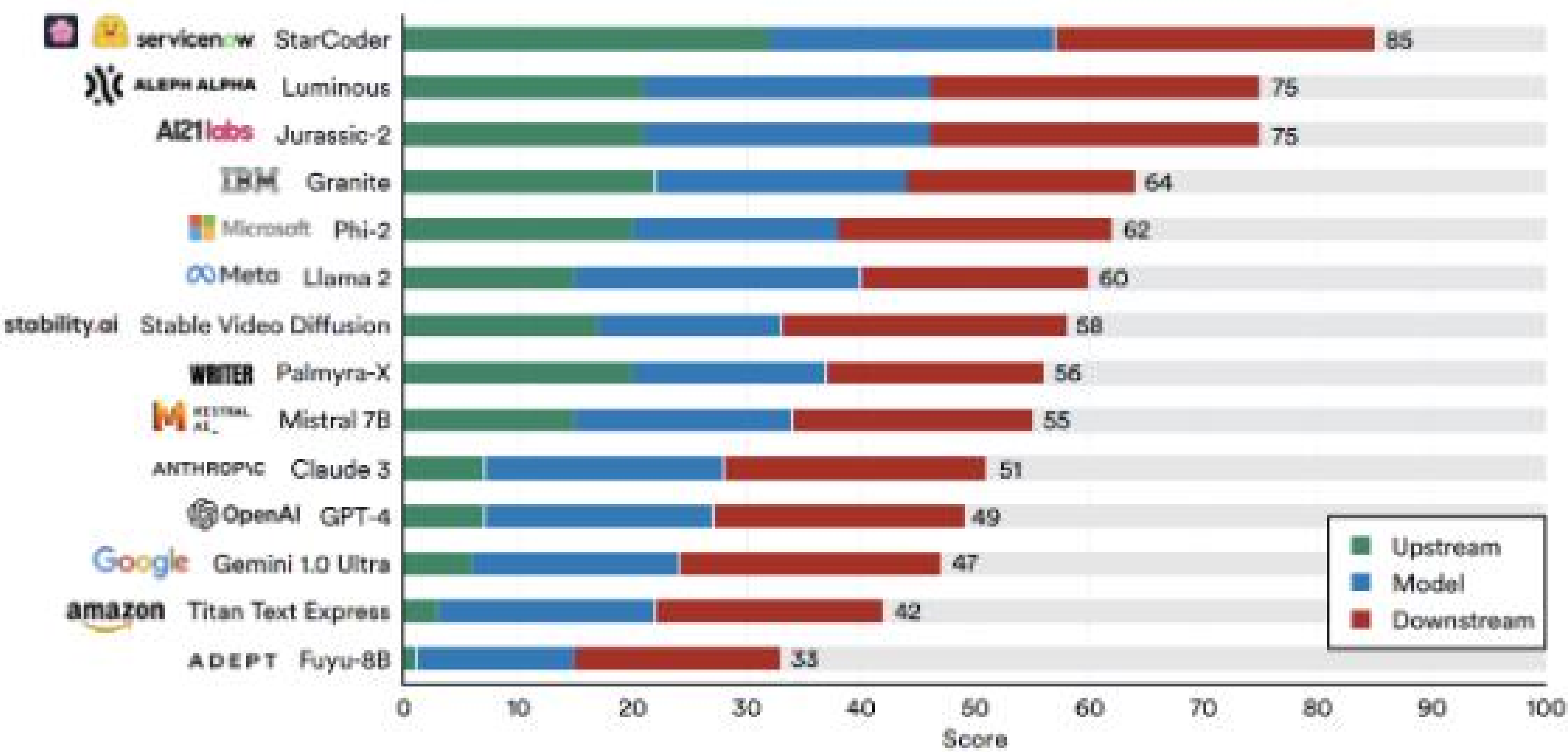
Sementara itu, investigasi mengenai bias implisit yang dilakukan peneliti terhadap beberapa model LLM memperoleh temuan bahwa terdapat implicit bias dari bias implisit dalam LLM yang secara eksplisit tidak bias yang selaras dengan stereotip masyarakat. Metode penelitian dilakukan dengan melewati 2 tahap yakni mendeteksi bias dalam LLM, peneliti mengembangkan 2 metode yakni bias implisit LLM (asosiasi otomatis antara kata atau konsep) dan bias pengambilan keputusan LLM (perilaku model yang mencerminkan bias implisit). Selanjutnya, mereka menyelidiki pola diskriminasi relatif ketika AI mengambil keputusan dengan mengujinya pada beberapa model AI. Beberapa bias yang ditemukan dalam penelitian ini yaitu misalnya keterkaitan antara gender dengan kualifikasi pekerjaan tertentu seperti LLM memiliki kecenderungan untuk mengasosiasikan perempuan dengan bidang ilmu humaniora dibandingkan STEM.

Transparansi dan Keterjelasan

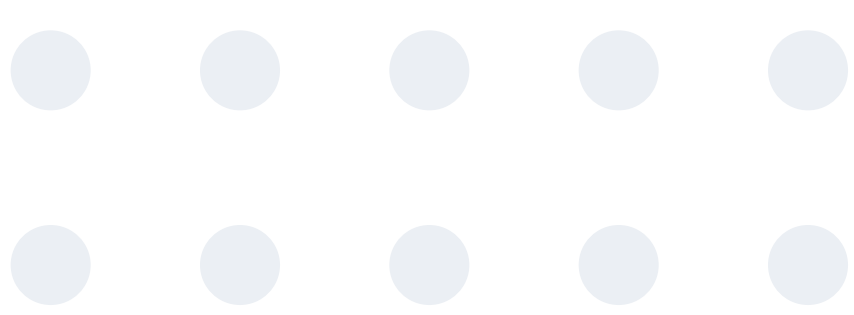
Transparansi dalam tata kelola AI terbagi menjadi beberapa lapisan. Ini termasuk transparansi data dan model, yang menyangkut sumber data dan keputusan algoritmik, serta transparansi operasional, yang mencakup bagaimana AI dipantau dan diimplementasikan dalam praktik. Melengkapi hal ini juga adalah transparansi penjelasan, yang berfokus pada bagaimana AI memahami makna dan mengartikulasikan alasan di balik keputusannya.

HAI Stanford melalui proyek Foundation Model Transparency Index (FMTI) v1.1 mengembangkan indeks untuk mengevaluasi transparansi dalam pengembangan dan penerapan model AI. Indikator yang dinilai secara garis besar terdiri atas 3 domain yaitu hulu (data & komputasi untuk pelatihan); model (sistem inti AI); serta hilir (aplikasi dan penerapan) dan diujikan terhadap beberapa model AI sebagaimana tergambar dalam tabel 6. Di tahun 2024, terjadi peningkatan signifikan terhadap transparansi rata-rata model AI yakni dari 37% menjadi 58%. Meskipun begitu, kurangnya transparansi masih terjadi di beberapa bidang seperti akses data, status hak cipta, dan dampak hilir.

Gambar 3. Skor FMTI 2024 Menurut Domain

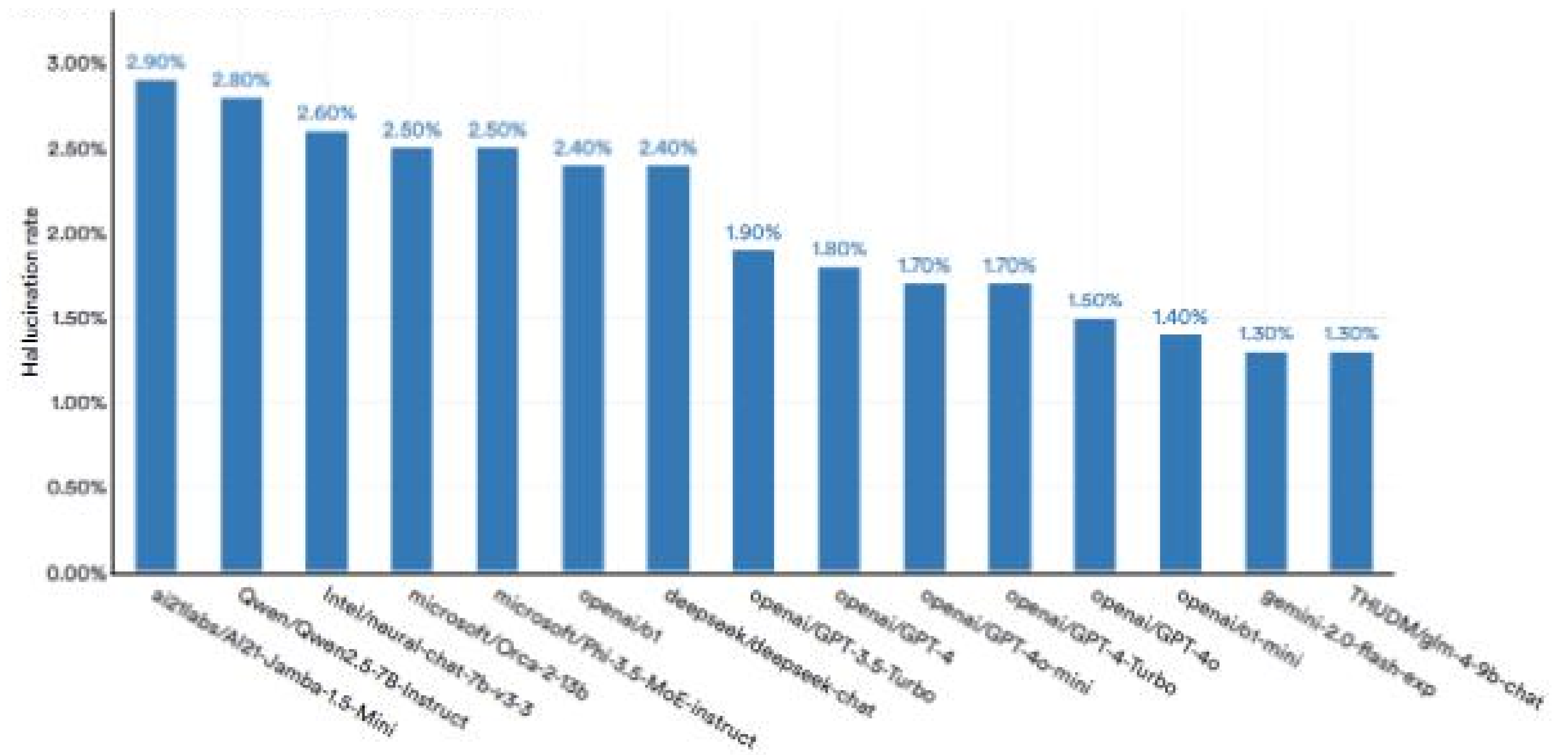


Sumber: May 2024 Foundation Model Transparency Index



Tantangan utama lainnya dalam pengembangan AI adalah adanya ketidakakuratan faktual dan halusinasi yang terus berlanjut. Implikasinya cukup serius di dunia nyata akibat kepiawaian AI dalam menghasilkan informasi yang tampaknya kredibel tetapi justru menyesatkan. Salah satu benchmark yang dikembangkan untuk mengukur derajat halusinasi dari model AI yaitu The Hughes Hallucination Evaluation Model (HHEM). Benchmark yang dikembangkan oleh Vectara ini menilai bagaimana tingkat halusinasi dari LLM ketika meringkas isi dari suatu dokumen. Hasil pengujian menggunakan HHEM terhadap beberapa model AI menghasilkan kesimpulan bahwa model yang mempunyai tingkat halusinasi terendah/terbaik yaitu GLM-4-9b-Chat dan Gemini-2.0-Flash-Exp (Gambar 4).

Gambar 4. Tingkat Halusinasi HHEM



Sumber: HHEM Leaderboard, 2025 | Chart: 2025 AI Index Report

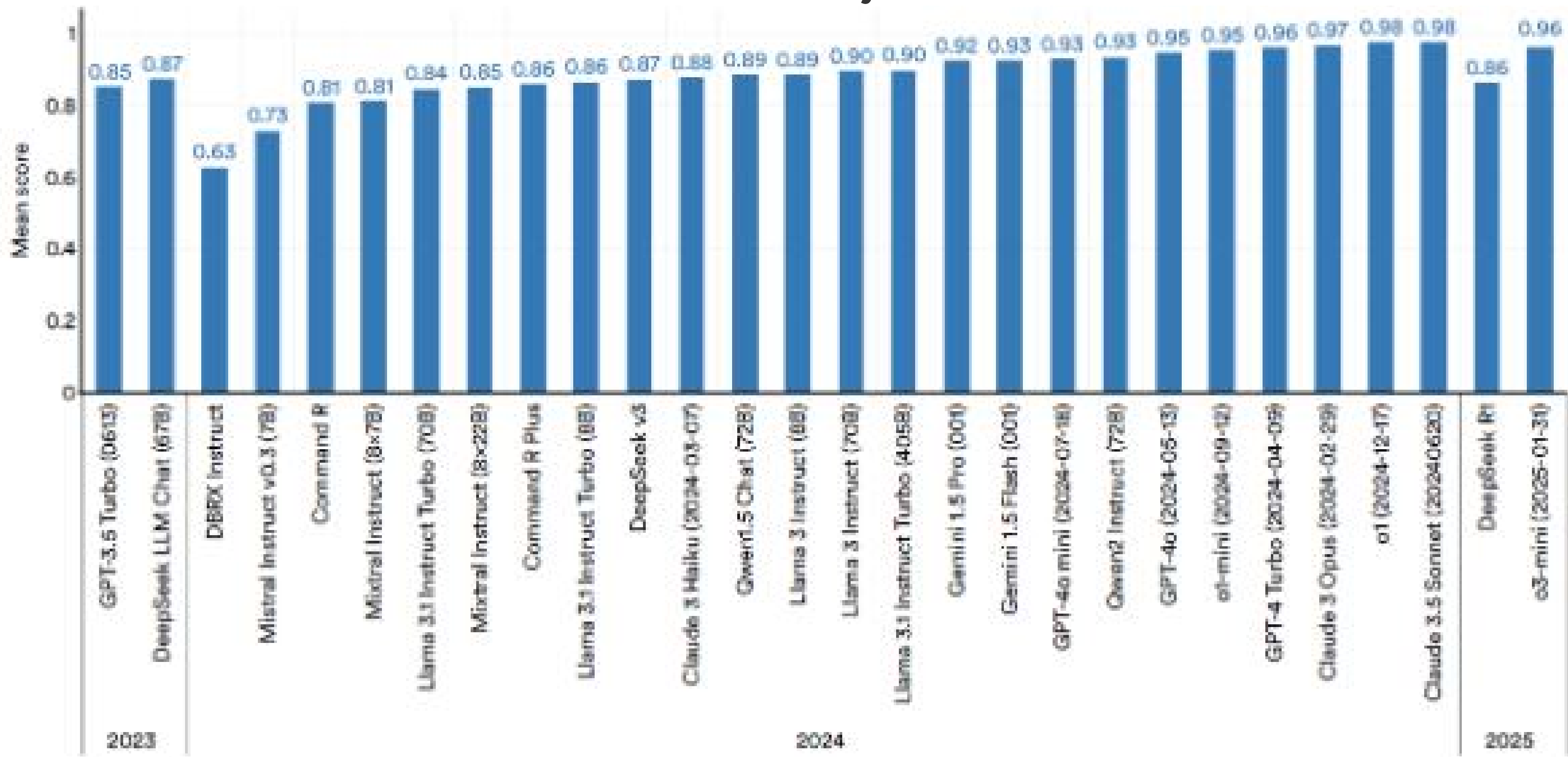
Keamanan dan keselamatan

Dalam laporan ini, dimensi keamanan dan keselamatan diuraikan ke dalam 3 aspek utama: 1) jaminan bahwa sistem AI mampu menjaga komponen intinya dari ancaman eksternal; 2) mekanisme untuk meminimalkan kerugian dari penyalahgunaan AI secara sengaja maupun tidak; and 3) risiko inheren dari AI seperti halnya halusinasi. Salah satu alat ukur yang dikembangkan untuk mengevaluasi tanggung jawab dan keselamatan dari model AI yaitu HELM Safety. Benchmark yang diperkenalkan oleh Stanford’s Center for Research on Foundation Models (CRFM) ini merupakan rangkaian dari beberapa tolak ukur AI yang menguji aspek safety AI seperti BBQ, SimpleSafetyTests,HarmBench, AnthropicRedTeam, and XSTest (Tabel 3). Berdasarkan hasil pengujian menggunakan HELM Safety terhadap berbagai model AI, yang paling aman yakni Claude 3.5 Sonnet (0.977) dan o1 (0.976) (Figure 5).

Tabel 3. Masing-Masing Benchmark HELM Safety

Benchmark	Fungsi
BBQ	Mengukur bias sosial, melindungi kelas di bawah undang-undang anti-diskriminasi AS.
SimpleSafetyTests	Menilai risiko yang terkait dengan perilaku melukai diri sendiri, cedera fisik, dan materi pelecehan seksual anak.
HarmBench	Mengevaluasi respons terhadap pertanyaan terkait pelecehan, produksi senjata kimia, dan disinformasi.
AnthropicRedTeam	Menilai tingkat bahaya dengan memeriksa bagaimana model menangani percakapan yang bersifat konfrontatif.
XSTest	Mengukur keseimbangan antara kebermanfaatan dan ketidakberbahayaan.

Gambar 5. HELM Safety: Skor Rata-Rata

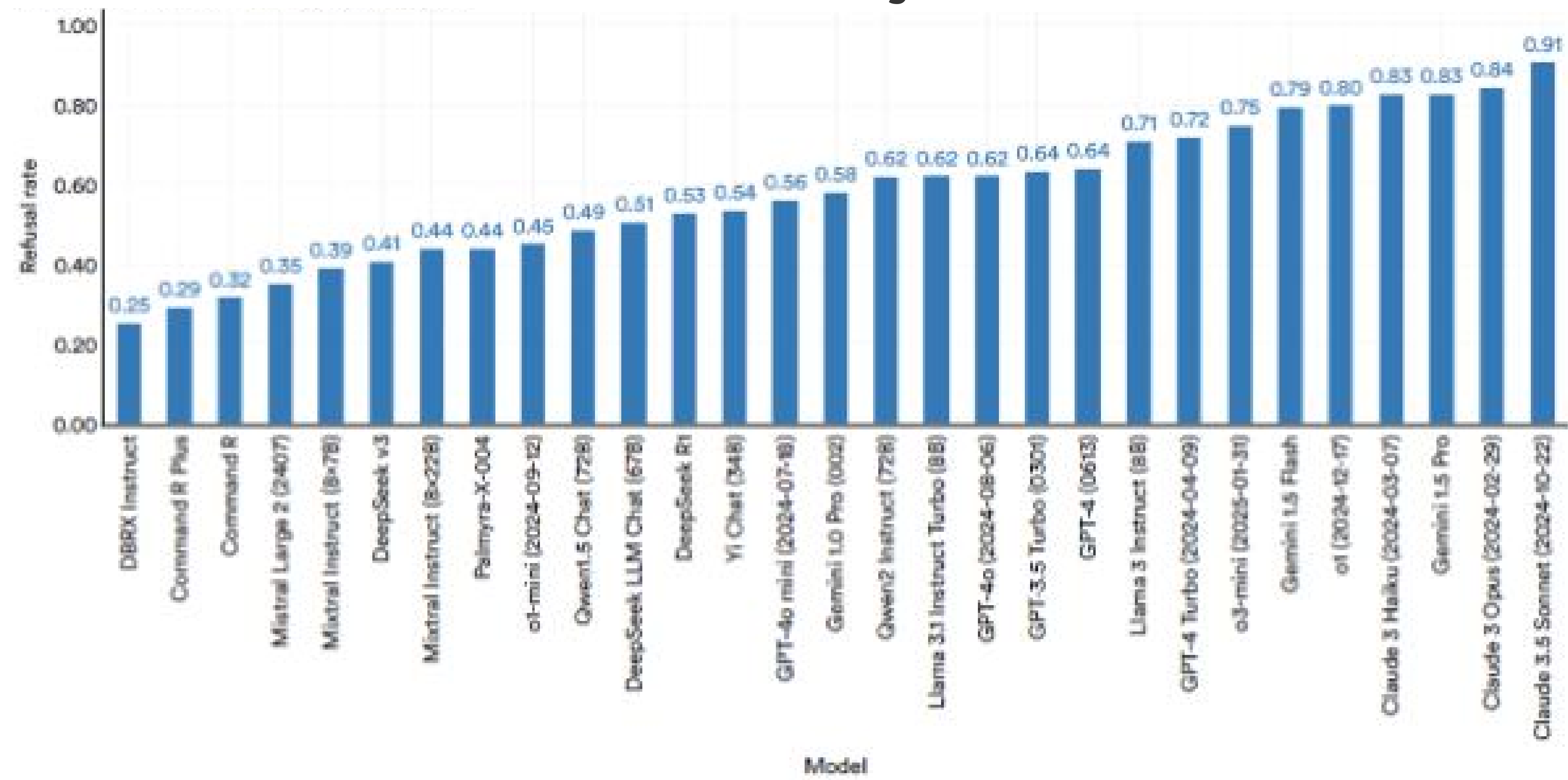


Source: HELM, 2025 | Chart: 2025 AI Index Report

Benchmark lainnya yang digunakan untuk mengevaluasi dimensi keselamatan dalam model AI yaitu AIR-Bench. Keselarasan antara pengembangan AI dengan kerangka kerja regulasi dan korporasi di dunia nyata menjadi acuan utama dari asesmen yang dijalankan oleh AIR-Bench. Pada dasarnya, AIR-Bench menilai frekuensi model AI dalam menolak suatu permintaan oleh karena risiko seperti keamanan, etika, maupun kepatuhan. Risiko tersebut terdiri atas 4 tingkat yang meliputi risiko sistem dan operasional, risiko keamanan konten, risiko sosial, dan risiko hukum dan hak. Klasifikasi ini didasarkan pada analisis delapan

pemerintah dan 16 kebijakan perusahaan. Hasil evaluasi dari AIR-Bench terhadap 22 model AI menemukan bahwa tingkat penolakan terbesar terjadi pada seri Claude dari Anthropic dengan nilai 91%. Terjadinya inkonsistensi dalam beberapa bidang menunjukkan bahwa perbaikan yang ditargetkan menjadi penting.

Gambar 6. AIR-Bench: Tingkat Penolakan



Untuk meningkatkan keamanan pada model AI, para peneliti telah mengusulkan beberapa saran untuk pengembangan model. Salah satu pendekatan utama adalah mengembangkan strategi penyesuaian yang lebih kuat untuk mencegah manipulasi mekanisme keamanan AI. Hal ini dimaksudkan untuk mencegah AI menanggapi permintaan pengguna yang dianggap berbahaya, seperti prosedur perakitan bom, bahkan jika permintaan tersebut telah dirancang sedemikian rupa sehingga AI akan menjawabnya. Secara paralel, teknik-teknik baru seperti latent adversarial training (LAT) yang ditargetkan menawarkan jalan baru untuk meningkatkan ketahanan model. LAT meningkatkan keamanan model AI selama pelatihan terhadap serangan adversarial, menghilangkan kerentanan backdoor, dan meningkatkan kemampuan untuk menghapus pengetahuan berbahaya atau yang dilindungi hak cipta sambil mengurangi biaya komputasi.

Tren terhadap Penerapan RAI dalam Berbagai Bidang Organisasi dan Bisnis

AI Index bersama McKinsey & Company melakukan survei terhadap 759 pemimpin industri dari 30 negara pada tahun 2024 mengenai penerapan RAI dalam bisnisnya. Investasi yang dikeluarkan perusahaan untuk mengoperasikan RAI, termasuk pengeluaran modal dan operasional, lebih besar dikeluarkan oleh perusahaan dengan pendapatan tinggi. Anggaran terhadap tata kelola AI umumnya difokuskan pada bidang information security (21%) dan data analytics (17%). Implementasi terhadap RAI membawa pengaruh positif bagi organisasi, terlebih dalam hal peningkatan operasional bisnis (42%) (Tabel 4). Meskipun demikian, masih terjadi ketimpangan antara risiko yang diobservasi dengan upaya mitigasi untuk mengatasinya seperti dalam hal keamanan siber, kepatuhan terhadap peraturan, dan privasi pribadi/individu. Pemimpin bisnis menyebutkan bahwa salah satu kendala utama yang dihadapi dalam pengaplikasian RAI yakni kesenjangan pengetahuan dan pelatihan (51%).

Tabel 4 Dampak RAI & Hambatan Utama dalam Penerapan RAI

Impact of RAI	Peningkatan operasional bisnis (42%), Peningkatan kepercayaan pelanggan (34%), Peningkatan reputasi merek (29%), Peningkatan hasil bisnis (28%), Penurunan jumlah insiden (22%)
Hambatan Utama Implementasi RAI	Kesenjangan pengetahuan dan pelatihan (51%), Keterbatasan sumber daya atau anggaran (45%), Ketidakpastian regulasi (40%), Keterbatasan teknis (32%), Resistensi organisasi (22%)

Politik (Pemilu)

Perhatian mengenai urgensi RAI dalam menghasilkan informasi yang valid dan dapat dipertanggungjawabkan turut meningkat di ranah politik. Tahun 2024 menandai perhelatan demokrasi yang berlangsung serentak di berbagai belahan dunia sekaligus periode manakala disinformasi yang dikeluarkan oleh AI secara ekstensif terjadi. Kekhawatiran etis yang berpotensi dilanggar ketika AI menjadi alat bagi manipulasi informasi terdiri atas liar’s dividend hingga pelanggaran privasi (Tabel 5). Dalam survei yang dilakukan oleh Rest of World terhadap Konten pemilu yang dihasilkan oleh AI, terdapat sebanyak 60 insiden yang terjadi di 15 negara melalui 10 platform media sosial yang berbeda. Beberapa kasus yang dimaksud diringkas dalam Tabel 6.

Tabel 5. Kekhawatiran Etis Seputar AI dan Manipulasi Informasi

Kekhawatiran Etis	Deskripsi
Liar’s Dividend	Kepercayaan masyarakat terkikis karena adanya teknologi deepfake. Fenomena ini menciptakan keraguan publik bahkan terhadap informasi terverifikasi dan bukti asli.
Blackmail	Teknologi deepfake kerap disalahgunakan oleh blackmailers untuk mendapat keuntungan dari korban seperti melalui eksploitasi seksual, pemerasan finansial, dan sabotase reputasi.
Erosi Kepercayaan pada Bukti	Konten yang dihasilkan AI menyebabkan pemalsuan yang hiperrealistis dan mengikis kepercayaan publik terhadap integritas informasi.
Pengurangan Otonomi Kognitif	Kemampuan AI berpeluang dalam memanipulasi proses pengambilan keputusan individu misalnya dengan menghasilkan pesan yang dipersonifikasikan berdasarkan preferensi, perilaku, dan kerentanan individu.

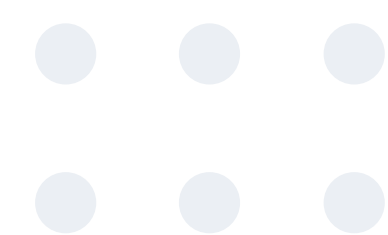
Kekhawatiran Etis	Deskripsi
Eksplorasi Merek Pribadi	Teknologi deepfake dapat menciptakan narasi palsu melalui pencurian merek pribadi dan memalsukan dukungan dari individu populer untuk menipu individu.
Penguatan Ujaran Kebencian	AI juga dapat memperkuat bias yang sudah ada dan mengamplifikasi ujaran kebencian dengan cara menciptakan echo chambers dan filter bubbles dari metrik keterlibatan pengguna.
Pengurangan Keterlacakan Operasi Asing	Kemampuan AI dalam menuliskan teks secara fasih seperti meminimalkan kesalahan tata bahasa yang berpotensi memberdayakan pelaku asing yang jahat dan menjadi semakin tidak dapat dilacak.
Pelanggaran Privasi	Cara kerja AI dalam memperoleh kumpulan dataset ekstensif untuk pelatihan juga menimbulkan kekhawatiran terkait penyalahgunaan dan paparan informasi pribadi.

Tabel 6. Kasus-Kasus Penting Konten Pemilu yang Dihasilkan AI

Kasus	Kronologi
Lubang jalan hasil AI berupaya mempengaruhi pemilih di Afrika Selatan. (South Africa, image, X/Twitter, Facebook, Instagram, Reddit, May 4, 2024)	Sebuah gambar AI-generated menampilkan jalanan panjang dipenuhi lubang dengan latar menuju Table Mountain yang ikonik di Cape Town. Foto itu diunggah di Facebook pada 4 Mei oleh seorang pengguna dengan keterangan bahwa di bawah pemerintahan Democratic Alliance (DA), pemerintah kota telah gagal mempertahankan layanan dasar dan berkontribusi pada kerusakan infrastruktur. Banyak pengguna yang membagikan gambar tersebut untuk mencegah pemilih di Western Cape untuk mendukung DA. Setelah ditelusuri oleh AFP Fact Check, teruji bahwa gambar tersebut dihasilkan oleh AI dan bersumber dari seorang pengguna Instagram yang membuat karya seni AI.

Kasus	Kronologi
Deepfake yang menampilkan para pemimpin partai Pakistan menyerukan boikot pemilu. (Pakistan, audio and video, X/Twitter, Feb. 7, 2024)	Sebuah video yang menampilkan para pemimpin politik dari Pakistan yakni masing-masing Imran Khan, mantan perdana menteri sekaligus pendiri Pakistan Tehree-e-Insaf (PTI) party dan Yasmin Rashid, pemimpin PTI terkemuka lainnya, tersebar luas di media sosial. Dalam video yang berbeda ini, keduanya sama-sama menyatakan boikot terhadap pemilu. Akun X official dari PTI kemudian menyampaikan bahwa audio dalam video itu palsu.
Pemilu Amerika Serikat terpengaruh oleh kampanye 'spamouflage'.(China and US, image, X/Twitter, Facebook, YouTube, TikTok, Medium, Feb. 15, 2024)	Kampanye spamouflage yang telah aktif dilakukan sejak tahun 2017 semakin bertambah parah dengan maraknya penggunaan gambar AI-generated. Pada tahun 2024, The Institute for Strategic Dialogue (ISD), sebuah organisasi think-tank berbasis di UK, menyebutkan bahwa pemerintah Cina dicurigai telah menyusup ke dalam pemilu US 2024 dengan menggunakan gambar AI-generated untuk menyebarkan misinformasi yang kemudian memperburuk polarisasi politik.





Strengths and Weaknesses of the Report

Strength



Contains rich and complex data

The report's biggest asset is its longitudinal, data-driven consistency. Because HAI has been running this since 2017, readers can track genuine trends over time. This report obtains data through various methods, including benchmark testing of AI models and surveys of several organizations worldwide, making it sharp and well-evidenced. For instance, responsible AI chapters tracking the gap between corporate recognition of RAI risks and actual action. That framing is analytically honest and politically careful at the same time. This makes this report stand out for its rich and complex information.



Providing numerous tables, case studies, and diagrams to illustrate data

The comprehensive data presented in this report is complemented by illustrations in the form of tables and diagrams, making it easier for readers to digest the information. In addition, several case study boxes are also included to visualize notable points from the RAI dimensions.

Weakness



Geopolitical framing stays surface-level on the Global South

The US-China competition narrative dominates the policy chapter. Countries in the Global South are mostly positioned as recipients of frameworks designed elsewhere (OECD, EU, UN) rather than as actors with their own governance agendas.



Gender and intersectionality are absent

There is no dedicated analysis of how AI harms or benefits vary by gender, or how algorithmic systems produce gendered and racialized outcomes. The responsible AI chapter gestures at "fairness and bias" but does not operationalize it along these axes.

Reference

¹Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K. and Wu, X., 2024. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8), pp.975-987.

²Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C. and Klyman, K., 2024. Consent in crisis: The rapid decline of the ai data commons, 2024. URL <https://arxiv.org/abs/2407.14933>.