[Tensor , Risk , Consulting]

AI Model Transparency Report

Increasingly, global regulators are moving toward defined transparency standards for AI models: If models are not open source, then high levels of transparency are likely to be needed going forward. See our Transparency Report for a listing of some of the most popular AI models in use today and see where they come out in terms of transparency.

Are you meeting AI model transparency expectations? Is it clear to your customers and external stakeholders that certain decisions are made using AI? Do they understand how AI model reasoning takes place? Can you provide evidence that your models are not introducing bias or discrimination into your decision-making process?

For help improving transparency around AI model use and decision-making contact us at *michael@tensorrisk.com*.

Each AI model below is categorized as one of the following:

Open Source: Model code, model architecture, and parameter values are freely available; use and redistribution are not significantly restricted by license.

Note: This standard for open source is intentionally less strict than the standard employed by the Open Source Initiative and allows for restrictions pertaining to safety and/or acceptable use policy.

Open Source with Restrictions on Use: Model code, model architecture, and parameter values are freely available; license places significant limitations on use and/or redistribution for commercial purposes (including, but not necessarily limited to, royalties).

Closed Source- High Transparency: Neither open source standard is met. Transparency is provided such that users understand: (i) that they are interacting with a machine and understand how machine reasoning takes place; (ii) the types of data sources relied upon for training and training techniques; and (iii) the statistical accuracy of the model output.

Closed Source- Medium Transparency: Neither open source standard is met. Two of the three High Transparency standards are met.

Closed Source- Low or No Transparency: Neither open source standard is met. Zero or one of the three High Transparency standards are met.

January 2024

AI Model Provider	AI Model Name	Open Source	Open Source with Restrictions on Use	Closed Source-High Transparency	Closed Source-Medium Transparency	Closed Source-Low or No Transparency
Google	Gemini (mulit-modal)					✓
OpenAI	GPT-4					✓
Meta	Llama 2	✓				
Amazon	Titan					✓
Anthropic	Claude 2.1				✓	
Technology Innovation Institute	Falcon 180B		✓			
Mistral	Mixtral 8x7B	✓				
Stability AI	Stable Diffusion XL (image generation)		✓			
Alibaba Cloud	Tongyi Qianwen 2.0					✓
BigScience Project / Hugging Face	BLOOM	✓				
Open research collaboration and Microsoft	LLaVA		✓			
OpenAI	DALL-E 3					✓

Supporting Documentation

Google

Gemini

Source documents: gemini_1_report.pdf (storage.googleapis.com)

Categorization: Closed source- low transparency

Rationale: Google's own report accompanying the release of Gemini provides ample information about the statistical accuracy of the model's output. Other transparency factors (such as how reasoning takes place, specific data sources, and training techniques) are discussed at too high a level to be considered transparent.

(i) *Interacting with a machine and understand how machine reasoning takes place*- While high level types of reasoning are discussed, there is not a plain language, comprehensive description of how this reasoning takes place.
(ii) *Data sources and training techniques*- The description of data sources is provided at the "category" level (e.g., video, text, etc.), which is too high a level to be considered transparent. While some training insights are conveyed, there is no plain language, comprehensive description of the end-to-end training process.
(iii) *Statistical accuracy of model output*- Statistics are provided for various categories (e.g., holistic, math, python coding, reading comprehension, common sense, science, translation, health, medicine, technology, engineering)

Open AI

GPT-4

Source documents: gpt-4.pdf (openai.com)

Categorization: Closed source- low transparency

Rationale: Open AI's own report accompanying the release of GPT-4 provides ample information about the statistical accuracy of the model's output. Other transparency factors (such as how reasoning takes place, specific data sources, and training techniques) are intentionally sparse, with Open AI citing competitive and safety considerations.

(i) *Interacting with a machine and understand how machine reasoning takes place*- While reference is made to certain reasoning techniques, there is not a plain language, comprehensive description of how reasoning takes place.
(ii) *Data sources and training techniques*- "GPT-4 is a Transformer-style model pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF). Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
(iii) *Statistical accuracy of model output*- Significant, detailed information is provided on a wide variety of subjects.

Meta

Llama 2

Source documents: Llama 2 - Resource Overview - Meta AI

Categorization: Open source

Rationale: " With each model download you'll receive:

- Model code
- Model Weights
- README (User Guide)
- Responsible Use Guide
- License
- Acceptable Use Policy
- Model Card "

"Our model and weights are licensed for both researchers and commercial entities, upholding the principles of openness. Our mission is to empower individuals, and industry through this opportunity, while fostering an environment of discovery and ethical AI advancements."

"Llama 2 is also available under a permissive commercial license, whereas Llama 1 was limited to non-commercial use."

"

1. This is a bespoke commercial license that balances open access to the models with responsibility and protections in place to help address potential misuse.
2. Our license allows for broad commercial use, as well as for developers to create and redistribute additional work on top of Llama 2.
3. We want to enable more innovation in both research and commercial use cases, but believe in taking a responsible approach to releasing AI technologies.
4. For more details, our license can be found here.

"

"a. Grant of Rights. You are granted a non-exclusive, worldwide, non-transferable and royalty-free limited license under Meta's intellectual property or other rights owned by Meta embodied in the Llama Materials to use, reproduce, distribute, copy, create derivative works of, and make modifications to the Llama Materials."

Amazon

Titan Text

Source documents: Responsible AI – AWS AI Service Cards: Amazon Titan Text – Amazon Web Services

Categorization: Closed source- low transparency

Rationale: "Amazon Titan Text is a family of proprietary large language models (LLMs) designed for enterprise use cases." Detailed, public-facing information is not readily available.

(i) *Interacting with a machine and understand how machine reasoning takes place*- Some high level information is provided.
(ii) *Data sources and training techniques*- Some high level information is provided.
(iii) *Statistical accuracy of model output*- Some test examples are provided.

Anthropic

Claude 2.1

Source documents: ModelCardClaude2_with_appendix.pdf (anthropic.com)

Categorization: Closed source- medium transparency

Rationale: Medium level transparency is provided for categories i and ii. High level transparency is provided in category iii.

(i) *Interacting with a machine and understand how machine reasoning takes place*- Some information has been provided which enables an understanding of how reasoning takes place in accordance with Athropic's HHH principles and Constitution-based approach.
(ii) *Data sources and training techniques*- Some publicly available information has been provided with respect to data sources and training techniques.
(iii) *Statistical accuracy of model output*- Detailed information is provided about accuracy on a range of topics.

Technology Innovation Institute (Abu Dhabi Government Funded Research Institution)

Falcon 180B

Source documents: Falcon LLM (tii.ae)

Categorization: Open Source with restrictions

Rationale: "We are launching our latest Falcon 180B LLM as an open access model for research and commercial use. Falcon-180B is accessible to developers through a royalty-free license, based on Apache 2.0. The license includes restrictions on illegal or harmful use, and requires those intending to provide hosted access to the model to seek additional consent from TII. The model is free of charge to download, use and integrate into applications and end user products. Those hosting providers wishing to provide shared instances of the model or its derivatives as a managed service (for inference or fine tuning) are not covered by the proposed license, and are invited to e nter into a separate license arrangement with TII."

Mistral AI

Mixtral 8x7B

Source documents: Mixtral of experts | Mistral AI | Open-weight models

Categorization: Open source

Rationale: "Licensed under Apache 2.0", which allows users to use the software for any purpose, to distribute it, to modify it, and to distribute modified versions of the software under the terms of the license, without concern for royalties.

Stability AI

Stable Diffusion XL (Image Generation)

Source documents: Terms of Use — Stability AI

Categorization: Open source with restrictions on use

Rationale: Free for non-commercial use. Fee structure exists for commercial use.

Alibaba Cloud

Tongyi Qianwen 2.0

Source documents: Alibaba Cloud Launches Tongyi Qianwen 2.0 and Industry-specific Models to Support Customers Reap Benefits of Generative AI - Alibaba Cloud Community

Categorization: Closed source- low transparency

Rationale: While Alibaba Cloud has a compelling history of open sourcing older and/or smaller versions of its models, this relatively new LLM model appears to remain closed source-- at least for the time being. An exhaustive online search failed to obtain evidence of open source technology for this model or any significant information regarding model transparency. We recognize that additional information may exist in non-English source documents.

BigScience Project

BLOOM Model

Source documents: Bloom AI

Release of largest trained open-science multilingual language model ever | CNRS

BLOOM (huggingface.co)

Categorization: Open Source

Rationale: "**BLOOM is the largest multilingual language model to be trained 100% openly and transparently." "BLOOM—by virtue of its open-science ethos—lets scientists from all horizons freely explore how language models work, in order to improve them."** "BLOOM is freely distributed under its Responsible AI Licence explicitly prohibiting use for malicious purposes."

Open research collaboration and Microsoft

LLaVA

Source documents: LLaVA (llava-vl.github.io)

LLaVA: Large Language and Vision Assistant: People - Microsoft Research

Categorization: Open source with restrictions on use

Rationale: "The service is a research preview intended for non-commercial use only, subject to the model License of LLaMA. We make GPT-4 generated visual instruction tuning data, our model and code base publicly available."

Open AI

Dall-E 3 Text to Image

Source documents: DALL·E 3 (openai.com)

Categorization: Closed source- low transparency

Rationale: Open AI indicates that they are presently working toward a tool that lets users know when an image has been generated; this tool is not presently available. "DALL·E 3 is now available to all ChatGPT Plus and Enterprise users, and will be available via the API and in Labs later this fall." "As with DALL·E 2, the images you create with DALL·E 3 are yours to use and you don't need our permission to reprint, sell or merchandise them."