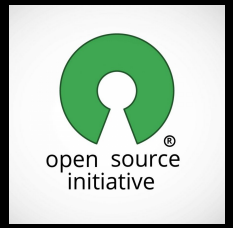
The State of the “Open” AI:
Exploring Data on AI Model Releases

Gabriel Toscano,
Open Source Initiative &
Duke Sanford School of Public Policy



Study: The state of “open” AI

Goals

- Understand how “open” models are being released
- Analyze key trends in “open” AI model releases

Study: The state of “open” AI

Goals

- Understand how “open” models are being released
- Analyze key trends in “open” AI model releases

Not

- Evaluate models for “openness”
- Evaluate the Open Source AI Definition (OSAID 1.0)

Study: The state of “open” AI

Procedure

- Download AI model metadata from Hugging Face

Study: The state of “open” AI

Procedure

- Download AI model metadata from Hugging Face
  - Using **keyword search**
    - “Open” (N = 20,076)
    - “Open Source” (N=68)

Study: The state of “open” AI

Procedure

- Download AI model metadata from Hugging Face
  - Using **keyword search**
    - “Open” (N = 20,076)
    - “Open Source” (N=68)
  - Using **author search**
    - Large labs (Alibaba, Deepseek, Google, Meta, Microsoft, Mistral, XAI, Open AI)
    - (N=2,028)

Study: The state of “open” AI

Procedure

- Download AI model metadata from Hugging Face
  - Using **keyword search**
    - “Open” (N = 20,076)
    - “Open Source” (N=68)
  - Using **author search**
    - Large labs (Alibaba, Deepseek, Google, Meta, Microsoft, Mistral, XAI, Open AI)
    - (N=2,028)
- **Quantitative and qualitative** analysis using Python

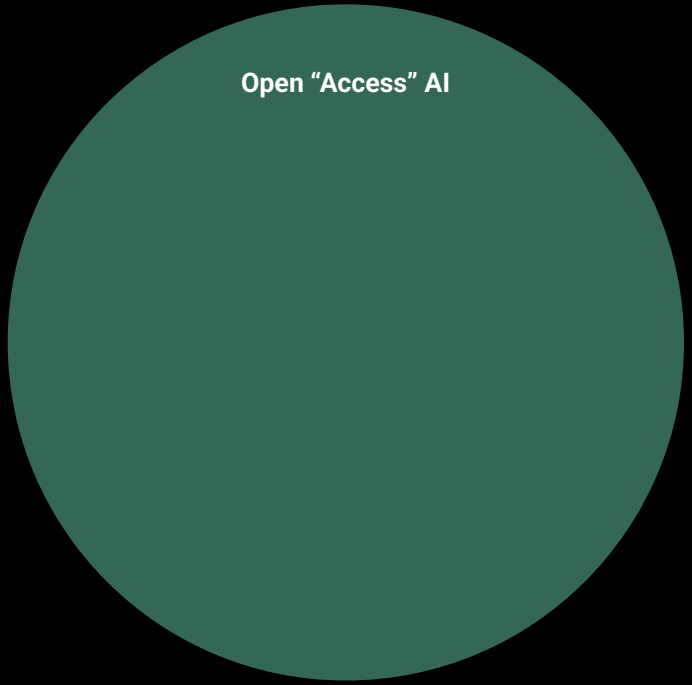
Key Findings

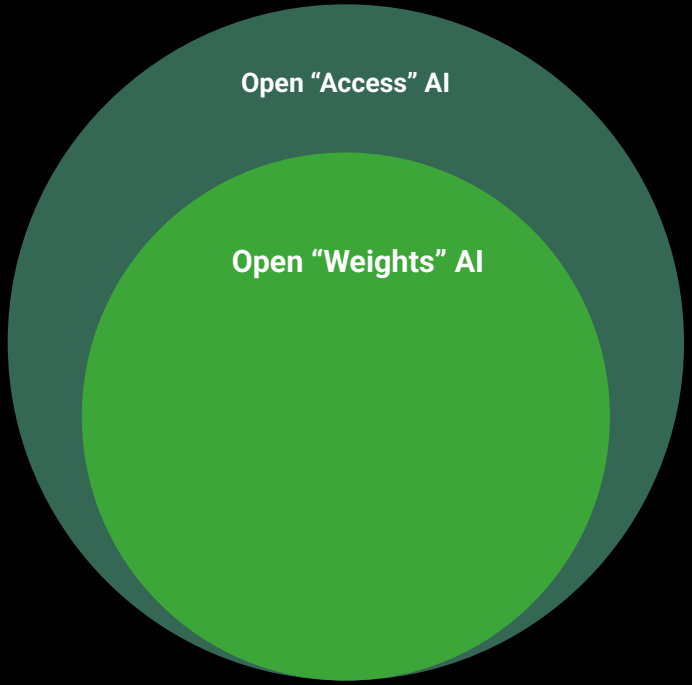
- The overwhelming majority of “open” models are based on larger models
- Apache 2.0 is the most popular OSI-approved license, followed by MIT
- CC-by licenses are prevalent, despite Creative Commons’ recommendations against using CC licenses for software
- The vast majority, over 50% of all models in in this sample, are released with an “unknown license”
- Alibaba’s Qwen family of models are the most popular base model in this sample
- **Custom licenses like Qwen, Llama, Gemma, Grok, OpenRAIL are becoming increasingly common, specially for flagship models, yet impose usage restrictions**

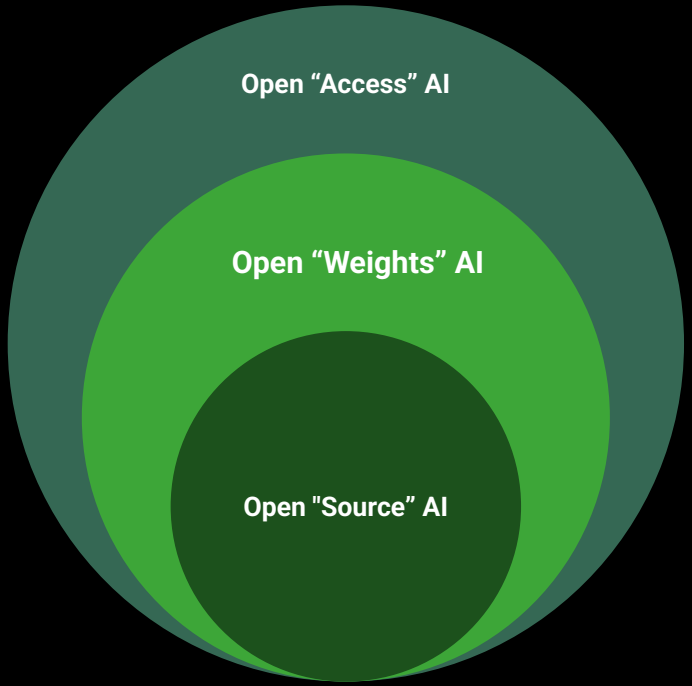
The Open Source AI Definition (OSAID) 1.0

OSAID 1.0

1	**Data Information**	• Under OSI-approved terms • Complete description of all data • Description of where to obtain public and private data
2	**Code**	• Under OSI-approved licenses • Training as well as data cleaning & prep code • Model architecture, supporting libraries
3	**Parameters**	• Under OSI-approved terms • Ideal both, final and intermediate states • Weights and configuration settings


Open “Access” AI


Open “Access” AI
Open “Weights” AI


Open “Access” AI
Open “Weights” AI
Open "Source” AI

Open Weights vs OSAID 1.0

Feature	Open Weights	Open Source AI
Weights & Biases	Released	Released
Training Code	Not Shared	Fully Shared
Intermediate Checkpoints	Withheld	Nice to have
Training dataset	Not Shared/Not disclosed	Released*
Training Data Composition	Partially/Not Disclosed	Fully Disclosed

*when legally allowed

Key Findings

Validated Systems as of 2024

- Pythia (Eleuther AI), OLMo (AI2), Amber and CrystalCoder (LLM360), and T5 (Google).

Systems that could pass if licenses were different

- BLOOM (BigScience), Starcoder2 (BigCode), Falcon (TII).

Systems that did not pass:

- Llama2 (Meta), Grok (X/Twitter), Phi-2 (Microsoft), Mixtral (Mistral).

Licenses

“Open” Models

License	Model Count	Proportion (%)	Cumulative (% of total)
unknown	11786	58.72	58.72
apache-2.0	4697	23.4	82.13
mit	1086	5.41	87.54
other	814	4.06	91.59
llama-family	660	3.29	94.88
cc-by-4.0	229	1.14	96.02
cc-by-nc-4.0	222	1.11	97.13
creativeml-openrail-m	111	0.55	97.68
openrail	72	0.36	98.04
cc-by-nc-sa-4.0	67	0.33	98.38

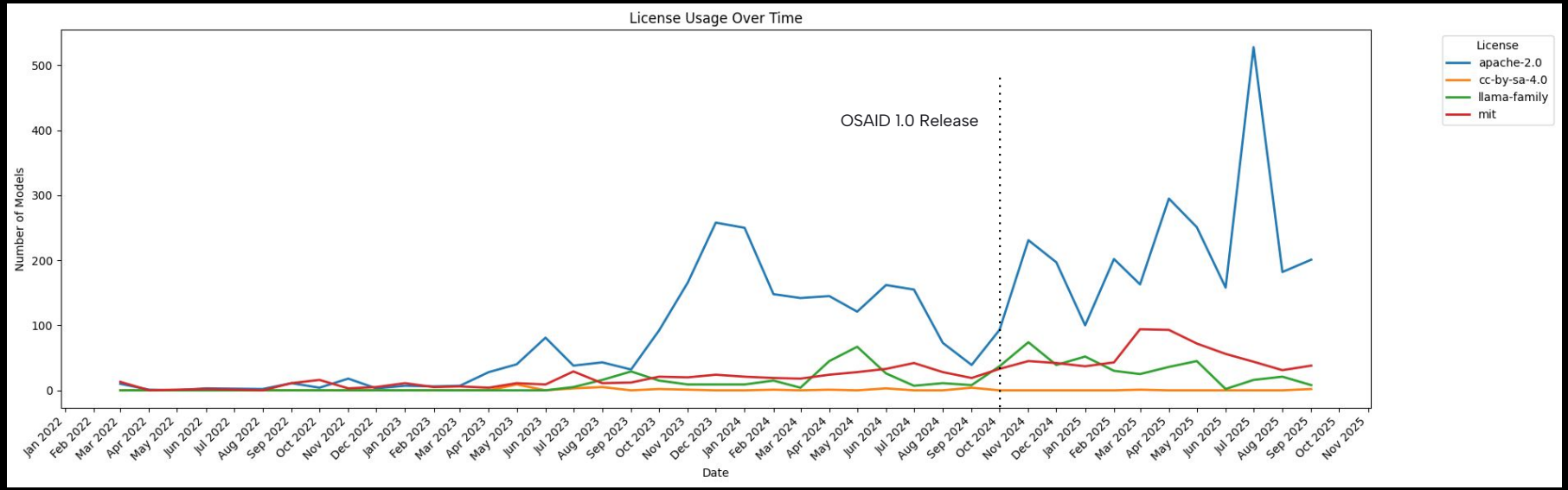
“Open” Models

License	Model Count	Proportion (%)	Cumulative (% of total)
unknown	11786	58.72	58.72
apache-2.0	4697	23.4	82.13
mit	1086	5.41	87.54
other	814	4.06	91.59
llama-family	660	3.29	94.88
cc-by-4.0	229	1.14	96.02
cc-by-nc-4.0	222	1.11	97.13
creativeml-openrail-m	111	0.55	97.68
openrail	72	0.36	98.04
cc-by-nc-sa-4.0	67	0.33	98.38

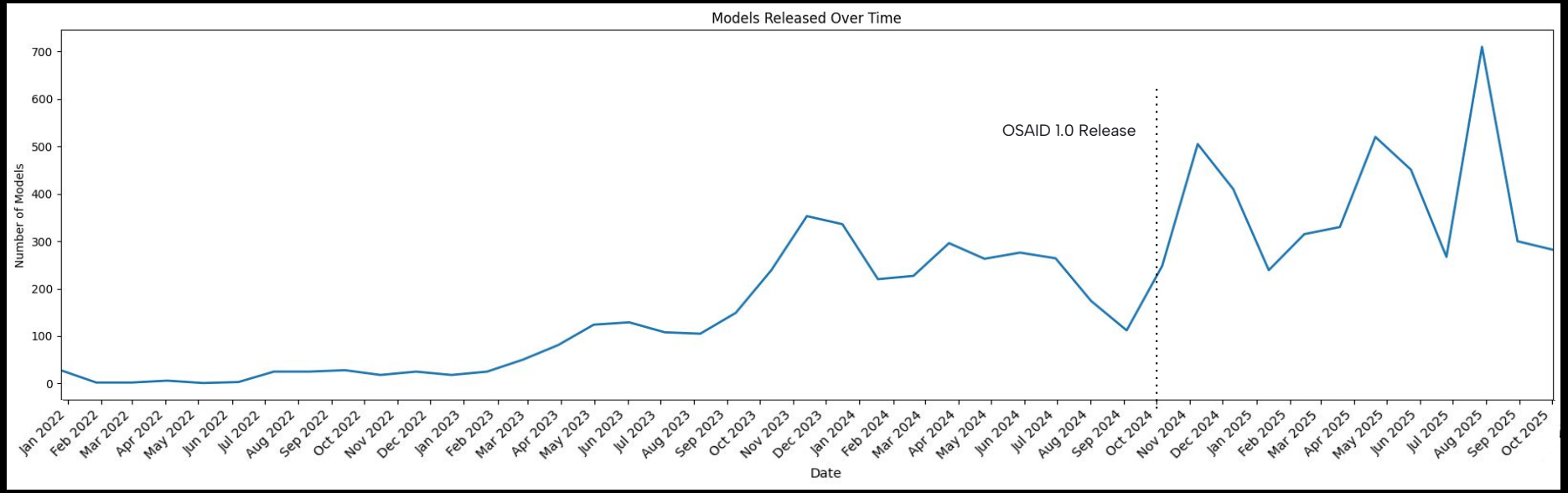
“Open Source” Models

License	Model Count	Proportion (%)	Cumulative (% of total)
unknown	36	52.94	52.94
apache-2.0	19	27.94	80.88
mit	8	11.76	92.65
llama-family	3	4.41	97.06
cc-by-nc-sa-4.0	2	2.94	100

License Use Over time
January 2022 – October 2025



Number of Model Releases Over time
January 2022 – October 2025



Prominent Organizations

Prominent Organizations

Alibaba (Qwen)

- +300 models
- Apache 2.0, Qwen

DeepSeek

- 78 models
- MIT for code, Deepseek for model

Meta (Llama)

- 70 models
- Llama-family of licenses

Mistral

- 39 models
- Apache 2.0

Prominent Organizations

Alibaba (Qwen)

- +300 models
- Apache 2.0, Qwen

DeepSeek

- 78 models
- MIT for code, Deepseek for model

Meta (Llama)

- 70 models
- Llama-family of licenses

Mistral

- 39 models
- Apache 2.0

Microsoft

- +400 models
- MIT, Apache, cc-by-*, Microsoft Research License

Google

- + 1000 models
- Apache 2.0, and Gemma licenses

XAI (Grok)

- 2 models
- Grok-1 under Apache 2.0, Grok-2 under Grok-2 license

OpenAI

- 30 models
- Apache 2.0

Prominent Organizations

Model Name	Number of Child Models
Qwen/Qwen2.5-1.5B-Instruct	342
Qwen/Qwen2.5-7B-Instruct	253
beomi/Llama-3-Open-Ko-8B-Instruct-preview	113
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	88
mistralai/Mistral-7B-v0.1	87
meta-llama/Meta-Llama-3-8B	70
microsoft/swin-tiny-patch4-window7-224	53
black-forest-labs/FLUX.1-dev	41
lerobot/smolvl4_base	39
openchat/openchat_3.5	36
openai/whisper-tiny	35

Custom Licenses

Custom Licenses

- **DeepSeek**
- **Qwen**
- **Llama**
- **Grok**
- **Gemma**

Custom Licenses
Qwen

2. Grant of Rights

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

Custom Licenses
Qwen

2. Grant of Rights

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

4. Restrictions

If you are commercially using the Materials, and ***your product or service has more than 100 million monthly active users, You shall request a license from Us.*** *You cannot exercise your rights under this Agreement without our express authorization.*

Custom Licenses

DeepSeek (model license)

2. Grant of Copyright License. Subject to the terms and conditions of this License, DeepSeek hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to ***reproduce, prepare, publicly display, publicly perform, sublicense, and distribute*** *the Complementary Material, the Model, and Derivatives of the Model.*

Custom Licenses

DeepSeek (model license) continued...

Use Restrictions You agree not to use the Model or Derivatives of the Model:

- *In any way that violates any applicable national or international law or regulation or infringes upon the lawful rights and interests of any third party;*
- *For military use in any way;*
- *For the purpose of exploiting, harming or attempting to exploit or harm minors in any way;*
- *To generate or disseminate verifiably false information and/or content with the purpose of harming others;*
- *To generate or disseminate inappropriate content subject to applicable regulatory requirements;*
- *To generate or disseminate personal identifiable information without due authorization or for unreasonable use;*
- *To defame, disparage or otherwise harass others;*
- *For fully automated decision making that adversely impacts an individual's legal rights or otherwise creates or modifies a binding, enforceable obligation;*
- *For any use intended to or which has the effect of discriminating against or harming individuals or groups based on online or offline social behavior or known or predicted personal or personality characteristics;*
- *To exploit any of the vulnerabilities of a specific group of persons based on their age, social, physical or mental characteristics, in order to materially distort the behavior of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;*
- *For any use intended to or which has the effect of discriminating against individuals or groups based on legally protected characteristics or categories.*

Custom Licenses
Llama

a. Grant of Rights. You are granted a non-exclusive, worldwide, non-transferable and royalty-free limited license under Meta's intellectual property…to ***use, reproduce, distribute, copy, create derivative works of, and make modifications*** *to the Llama Materials.*

Custom Licenses
Llama

a. Grant of Rights. You are granted a non-exclusive, worldwide, non-transferable and royalty-free limited license under Meta's intellectual property…to ***use, reproduce, distribute, copy, create derivative works of, and make modifications*** *to the Llama Materials.*

*iv. Your use of the Llama Materials must…****adhere to the Acceptable Use Policy*** *for the Llama Materials (available at https://llama.com/llama3/use-policy), which is hereby incorporated by reference into this Agreement.*

Custom Licenses
Llama

a. Grant of Rights. You are granted a non-exclusive, worldwide, non-transferable and royalty-free limited license under Meta's intellectual property…to ***use, reproduce, distribute, copy, create derivative works of, and make modifications*** *to the Llama Materials.*

*iv. Your use of the Llama Materials must…****adhere to the Acceptable Use Policy*** *for the Llama Materials (available at https://llama.com/llama3/use-policy), which is hereby incorporated by reference into this Agreement.*

2. Additional Commercial Terms. If, on the Meta Llama 3 version release date, the monthly active users of the products or services made… ***is greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta…***

Custom Licenses
Llama 4

With respect to any multimodal models included in Llama 4, the ***rights granted under Section 1(a)*** *of the Llama 4 Community License Agreement are* ***not being granted to you if you are an individual domiciled in, or a company with a principal place of business in, the European Union.***

Custom Licenses

Grok2

a. Permitted Uses: xAI grants you a non-exclusive, worldwide, revocable license to ***use, reproduce, distribute, and modify*** *the Materials: For* ***non-commercial and research purposes; and for commercial use solely if you and your affiliates abide by all of the guardrails provided in xAI's Acceptable Use Policy*** *(https://x.ai/legal/acceptable-use-policy), including 1. Comply with the law, 2. Do not harm people or property, and 3. Respect guardrails and don't mislead.*

Custom Licenses
Grok2

a. Permitted Uses: xAI grants you a non-exclusive, worldwide, revocable license to ***use, reproduce, distribute, and modify*** *the Materials: Fo****r*** ***non-commercial and research purposes; and for commercial use solely if you and your affiliates abide by all of the guardrails provided in xAI's Acceptable Use Policy*** *(https://x.ai/legal/acceptable-use-policy), including 1. Comply with the law, 2. Do not harm people or property, and 3. Respect guardrails and don't mislead.*

b. Restrictions:

You may not use the Materials, derivatives, or outputs (including generated data) to ***train, create, or improve any foundational, large language, or general-purpose AI models****, except for modifications or fine-tuning of Grok 2 permitted under and in accordance with the terms of this Agreement.*

Custom Licenses
Grok2

a. Permitted Uses: xAI grants you a non-exclusive, worldwide, revocable license to ***use, reproduce, distribute, and modify*** *the Materials: For* ***non-commercial and research purposes; and for commercial use solely if you and your affiliates abide by all of the guardrails provided in xAI's Acceptable Use Policy*** *(https://x.ai/legal/acceptable-use-policy), including 1. Comply with the law, 2. Do not harm people or property, and 3. Respect guardrails and don't mislead.*

b. Restrictions:

You may not use the Materials, derivatives, or outputs (including generated data) to ***train, create, or improve any foundational, large language, or general-purpose AI models****, except for modifications or fine-tuning of Grok 2 permitted under and in accordance with the terms of this Agreement.*

5. Acceptable Use

You are ***responsible for implementing appropriate safety measures, including filters and human oversight, suitable for your use case****. You must comply with xAI's Acceptable Use Policy (AUP), as well as all applicable laws. You may not use the Materials for illegal, harmful, or abusive activities.*

Custom Licenses
Gemma

2.2 Use

*You **may use, reproduce, modify, Distribute, perform or display** any of the Gemma Services only in accordance with the terms of this Agreement, and must not violate (or encourage or permit anyone else to violate) any term of this Agreement.*

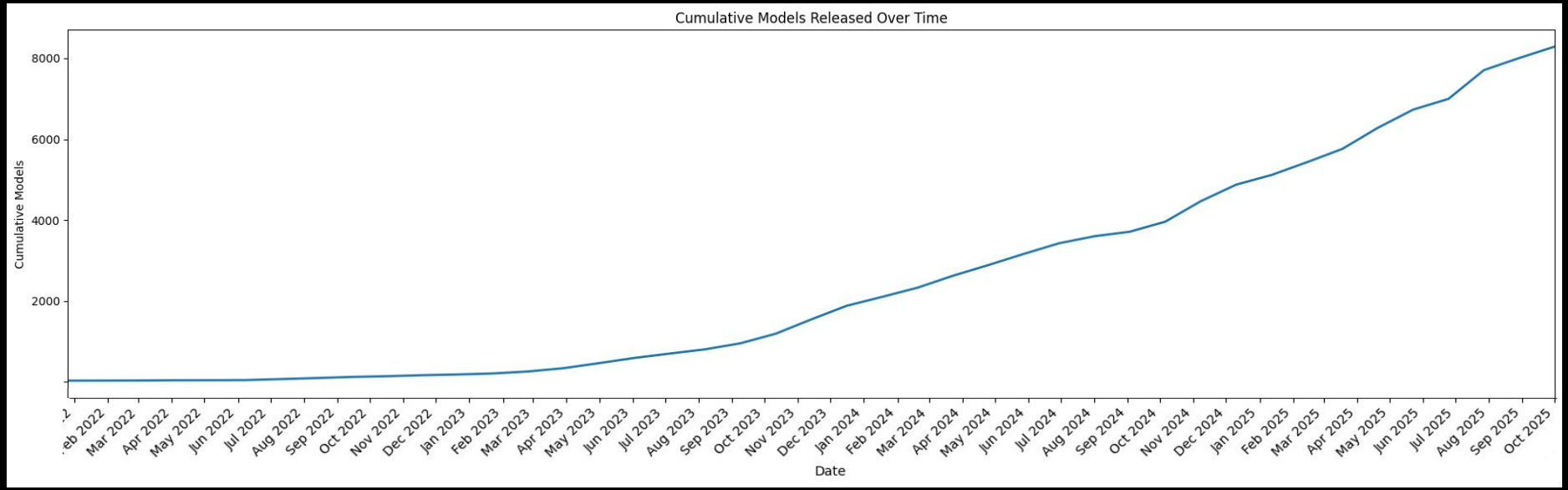
Custom Licenses
Gemma

2.2 Use

*You **may use, reproduce, modify, Distribute, perform or display** any of the Gemma Services only in accordance with the terms of this Agreement, and must not violate (or encourage or permit anyone else to violate) any term of this Agreement.*

3.2 Use Restrictions

*You **must not use any of the Gemma Services: for the restricted uses set forth in the Gemma Prohibited Use Policy at** ai.google.dev/gemma/prohibited_use_policy ("Prohibited Use Policy"), which is hereby incorporated by reference into this Agreement; or in violation of applicable laws and regulations.*


Cumulative Models Released Over Time
Cumulative Models
8000
6000
4000
2000
Date
Feb 2022
Mar 2022
Apr 2022
May 2022
Jun 2022
Jul 2022
Aug 2022
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Aug 2023
Sep 2023
Oct 2023
Nov 2023
Dec 2023
Jan 2024
Feb 2024
Mar 2024
Apr 2024
May 2024
Jun 2024
Jul 2024
Aug 2024
Sep 2024
Oct 2024
Nov 2024
Dec 2024
Jan 2025
Feb 2025
Mar 2025
Apr 2025
May 2025
Jun 2025
Jul 2025
Aug 2025
Sep 2025
Oct 2025

Key Findings

- The overwhelming majority of “open” models are based on larger models
- Apache 2.0 is the most popular OSI-approved license, followed by MIT
- CC-by licenses are prevalent, despite Creative Commons’ recommendations against using CC licenses for software
- The vast majority, over 50% of all models in in this sample, are released with an “unknown license”
- Alibaba’s Qwen family of models are the most popular base model in this sample
- **Custom licenses like Qwen, Llama, Gemma, Grok, OpenRAIL are becoming increasingly common, specially for flagship models yet impose usage restrictions**

Next Steps

- Network analysis
  - Models and models
  - Licenses and models

Next Steps

- Network analysis
  - Models and models
  - Licenses and models

- Download trends

Next Steps

- Network analysis
  - Models and models
  - Licenses and models
- Download trends
- Data documentation
  - Datasets and documentation

Next Steps

- Network analysis
  - Models and models
  - Licenses and models
- Download trends
- Data documentation
  - Datasets and documentation
- U.S. Policies in Open Source Software and Open Source AI
  - Existing definitions and provisions

Next Steps

- Network analysis
  - Models and models
  - Licenses and models
- Download trends
- Data documentation
  - Datasets and documentation
- U.S. Policies in Open Source Software and Open Source AI
  - Existing definitions and provisions
- **Crowdsourced “open” AI model evaluation**

Thank you

Gabriel Toscano

gabriel.toscano@duke.edu

sites.duke.edu/gabrieltoscano

Submit an “open” AI model for the study