TRANSPARENCY IN FRONTIER AI

What Leading Labs Are (And Aren't) Telling Us

RESEARCH BY
DAVID ROBUSTO

AMERICANS *for* RESPONSIBLE INNOVATION

TRANSPARENCY IN FRONTIER AI

What Leading Labs Are (and Aren't) Telling Us

Research by
DAVID ROBUSTO
Policy Analyst, ARI

ABSTRACT

As artificial intelligence systems become increasingly embedded in critical infrastructure, transparency in development and deployment is crucial for effective oversight. This paper presents a novel analysis of transparency practices across seven leading frontier AI models, evaluating 21 key metrics in four categories: User-Facing Documentation, Technical Transparency, Risk & Safety Information, and Evaluation & Impact. Our findings reveal a concerning trend: closed AI models are providing minimal public information about their technical underpinnings. When evaluated under our framework, the six closed models averaged a 0.95 out of 4.00 on Technical Transparency metrics.

Additionally, this research identifies a critical "documentation drift" problem, where significant model updates occur with minimal disclosure or documentation updates, creating an expanding gap between documented and deployed capabilities. This study also exposes a stark divide in industry practices: established technology companies with diverse revenue streams maintain higher transparency, while new AI startups trend toward opacity. Our transparency scoring system shows Meta's open-weight Llama 3.2 (88.9/100) and Google's Gemini 1.5 (62.5/100) leading in disclosure practices, while newer models from startup firms like OpenAI's o1 Preview (44.7/100) and xAI's Grok-2 (19.4/100) are significantly more opaque. These findings highlight the urgent need for robust, standardized disclosures to bolster transparency - disclosures that balance meaningful oversight with competitive innovation. We conclude by broadly calling for industry standards and legislative frameworks that could address this transparency deficit without inadvertently consolidating power among large tech incumbents.

INTRODUCTION

Frontier AI models are rapidly becoming embedded in critical infrastructure – from the financial sector to power grids to educational tools. As these models grow more powerful and ubiquitous, it is imperative that researchers, policymakers, and the public understand how they function. However, our new analysis of transparency metrics across seven frontier AI models reveals this is not the case: the closed models reviewed are decidedly opaque, with the top five closed models averaging 55.1 points out of 100.[1] The picture is even worse for information about these models' technical details, with those five models averaging 1.11 out of 4 on all Technical Transparency metrics. This opacity comes at a particularly concerning moment, with safety-conscious employees exiting leading labs and external reviewers uncovering "significant gaps" in safety measures.

This analysis exposes stark divides in industry disclosure practices. Established tech giants like Google and Meta, which possess greater resources and diverse revenue streams beyond AI, provide more robust disclosures. Meta's Llama 3.2 leads with the highest transparency score (88.9/100), while Google's Gemini 1.5 sets the standard among closed models (62.5/100). In contrast, AI-native companies – startups whose primary business model depends on selling access to their models – are trending toward increased opacity. This trend is compounded by the "documentation drift" problem – the growing gap between initial model documentation and the capabilities of deployed systems which have received significant but under-reported updates.

Through our systematic evaluation of 21 key metrics across four major categories – User-Facing Documentation, Technical Transparency, Risk & Safety, Evaluation & Impact – this paper examines the state of AI transparency, analyzes emerging patterns in disclosure practices, and explores the implications for effective oversight and governance.

WHY IS TRANSPARENCY IMPORTANT FOR FRONTIER MODELS?

> "*Transparency is an essential precondition for public accountability, scientific innovation, and effective governance of digital technologies. Without adequate transparency, stakeholders cannot understand foundation models, who they affect, and the impact they have on society.*" (Rishi Bommasani, Kevin Klyman, et. al, 2023)

As AI systems have grown more powerful and impactful in the last decade, transparency has emerged as a key principle for AI development. In a 2019 paper, three researchers at

[1] This excludes xAI's Grok-2, an outlier which scored 19.4 points total. Averaging its score on all Technical Transparency metrics, it scored 0.17 out of 4.

ETH Zurich overviewed 84 examples of AI principles and guidelines. Among them, transparency was the most frequently occurring suggestion, appearing in over 85% of analyzed documents.

Transparency is particularly important for frontier models, which are the most advanced systems being deployed today. These systems are increasingly implemented in critical domains like healthcare and defense, where their complex architectures, vast parameter spaces, and gigantic training datasets make it difficult to understand their behavior, often even for their own developers.

While the benefits of frontier models are vast, they also risk causing significant harm. These include, among others, discrimination (Claude 3 Model Card, pg. 28), hallucinations (GPT-4's Technical Report, pg. 46), or even the ability to assist experts in the reproduction of a biological threat (OpenAI o1 System Card, pg. 17). There is a current lack of standardized pre-deployment testing procedures for frontier systems. This means deployment decisions and domains marked for acceptable use rely on internal corporate assessments and ad hoc external tests.[2] As a result, adequate transparency is the only thing providing meaningful oversight as these systems are rapidly integrated into sensitive domains that impact people's livelihoods.[3]

For additional detail on transparency's importance in this context, the Center for Research on Foundation Models' relevant research paper – where they introduce the Foundation Model Transparency Index (FMTI) – does an excellent job overviewing the history and scholarship of AI transparency in Section 2.3. This work is meaningfully influenced by the FMTI, which was published in October 2023 and updated in May 2024 and assessed 10 powerful models on 100 transparency indicators.

METHODOLOGY

Our analysis examines frontier models including GPT-4, GPT-4o, and the o1 Preview by OpenAI, Claude 3/3.5 by Anthropic, Llama 3.2 by Meta, Gemini 1.5 by Google, and Grok-2 by xAI. The transparency of these seven major frontier AI models is assessed across 21 key metrics.

[2] While companies like OpenAI have solicited external assessments from organizations like Apollo Research and METR, these evaluations primarily test for systemic risks and do not verify things like benchmark performance or a model's technical details. Additionally, assessments by different companies are unstandardized and so can only be viewed as individual artefacts.
[3] Importantly, this type of transparency does not require the open release of model weights. An analogy here would be APIs. Developers can understand APIs through clear documentation, performance characteristics, and technical specifications, but don't have access to the underlying technical implementation.

These metrics span crucial aspects of AI development across four groups: *User-Facing Documentation* like prohibited uses and input modalities, *Technical Transparency* like model size and training data composition, Information on *Risk & Safety* like the alignment principles that guide development, and *Evaluation & Impact* like bias and environmental impact. Each metric was scored on a scale of 0–4, with higher scores indicating more comprehensive disclosure, adjusted for factors like a significant delay in disclosure after a model's release. Further details on the categories examined, metrics used, scoring system, and score modifiers are available in Appendix A.

Category	GPT-4	GPT-4o	o1 Preview	Llama 3.2	Gemini 1.5	Claude 3/3.5 Sonnet	Grok-2
Model Use Guidelines	4	3	1	4	3.5	3.25	0
Capabilities & Limitations	4	3.5	3.5	4	4	4	1
Changes From Previous Distinct Model	2	N/A	N/A	4	4	2	1
Access Methods	4	4	4	4	3	3.25	4
Input/Output Formats	4	4	3	4	4	3.25	4
External Tool Integration	N/A	1	2	3	4	3.25	1
Training Data Composition	1	2	1	2	1	1	0
Knowledge Cutoff	4	4	3	4	0	3.25	1
Model Architecture	1	0	0	4	4	0	0
Model Size	0	0	0	4	0	0	0
Training Time	0	0	0	4	0	0	0
Post-Training Enhancements	4	1	1	4	3	1.75	0
Interpretability and Explainability Techniques	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alignment Principles	3	3	2	3	4	3.25	0
Security	1	0	0	N/A	2	1.25	0
Privacy Controls	2	2	1	N/A	2	1	2
Model Release Criteria	1	4	4	2	2	2.25	0
Environmental Impact	0	1	0	4	1	1	0
Industry Benchmarks	2.5	1.5	1.5	4	2.5	2.5	1.5
Systemic Risk Evaluations	4	4	4	4	3	4	0
Direct Risk Evaluations	4	3	3	2	3	3.25	0
Total (out of 100)	59.9	53.9	44.7	88.9	62.5	54.4	19.4

Figure 1. Numeric transparency scores of seven models across 21 categories. Scores adjusted for penalties. Total ignores N/A values and scaled to 100.

For the full analysis of how each model was assessed across the 21 metrics, see ***Appendix B***.

RESULTS

Key Findings

The results paint a clear but concerning picture of industry transparency, with three key observations:

1. Closed models are especially opaque around their technical details, which is further compounded by the 'documentation-drift' problem.
2. Established big tech players (in this case Meta and Google) are currently more transparent than AI startups.
3. Despite the use of industry benchmarks, information disclosed on assessing model capabilities is not sufficient for meaningful comparison between models.

Discussion

Total transparency scores varied significantly across models. Meta's open-weight Llama 3.2 leads by a significant margin, scoring 88.9 out of 100 points – nearly 30 points higher than its closest competitor. Among closed models, Google's Gemini 1.5 achieved the highest transparency score of 62.5. OpenAI's GPT-4 came close with 59.9 points, while their two successive models became more opaque, with GPT-4o scoring 54.5 points and o1 Preview scoring 44.7. Anthropic's Claude 3/3.5 was in the middle, scoring 54.4 points. The least transparent model reviewed is xAI's Grok-2, which provides little public information, scoring just 19.4 points. The long-form analysis of each model is available in Appendix B.

Before discussing the results of the analysis, there is an important piece of context worth mentioning. While many companies release detailed documentation about their AI models at launch, these same models often receive substantial updates that can meaningfully change their behavior and capabilities. These updates typically come with minimal documentation – in many cases just a few sentences describing general improvements. A recent notable example is Claude 3.5 Sonnet, which was updated in October, resulting in significant increases in benchmark scores and anecdotal performance, but with hardly any details of changes made to the model. Much of the AI community has noted that this causes confusion and has started calling the updated model Claude 3.6 Sonnet for clarity.

This documentation gap – which we will refer to as the "documentation drift" problem, borrowing a term from software development – creates a moving target and complicates efforts to systematically evaluate, understand, and compare AI models. Our analysis of Claude, for example, is no longer reflective of the model users interface with today. It also highlights the challenge of unknown-unknowns in this space. Researchers and the public rely entirely on these companies to disclose when these models change. Both of these facts emphasize the need for better guidelines or requirements for ongoing documentation post-deployment.

With that context in mind, our analysis uncovers several concerning patterns about the state of frontier model transparency.

FINDING 1: Closed Models are Particularly Opaque About Technical Details

Despite variations in overall transparency, our analysis reveals that closed models are consistently opaque around their technical details. Every closed model in our sample, regardless of company size or overall transparency score, maintains near-complete opacity around Technical Transparency metrics such as model size (all scoring 0) and model architecture (all scoring 0–1). This category also features the largest gap between Meta's score (3.50) and the average score of the closed models (0.95) at 2.55.[4] This systematic opacity in technical details significantly hampers independent verification of capabilities and limitations, external risk assessments, and meaningful comparison between models. While some evaluation can be done through black-box testing, rigorous technical analysis requires greater insight into a model's internal and development details.

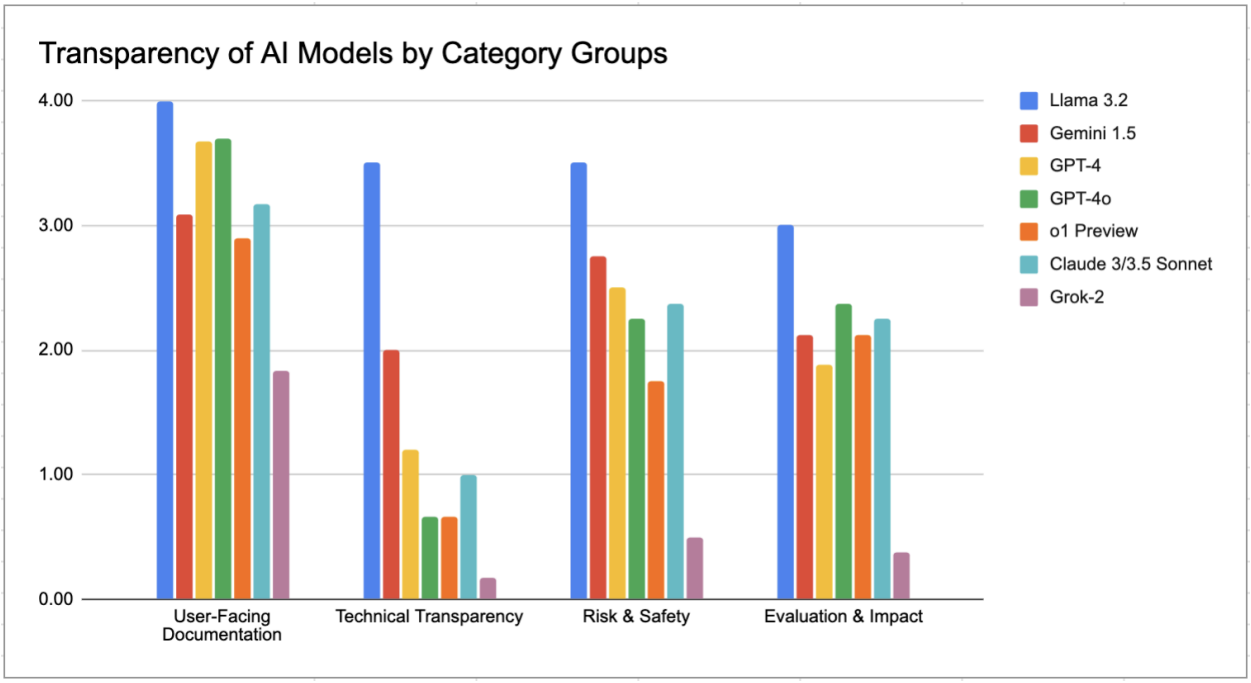

Figure 3. Category group scores for frontier models.

The contrast with Meta's Llama 3.2 is particularly telling. Meta provides detailed information about model architecture, training procedures, and computational requirements,

[4] The next closest category is Risk & Safety, with a discrepancy of 1.48.

demonstrating that such technical transparency is feasible.[5] This suggests that opacity around technical details may be driven by factors such as business strategy rather than technical constraints or oversight challenges.

This opacity, combined with the documentation-drift problem, can lead to a situation where the public has a very poor understanding of the deployed models they interact with. Even if a model had solid documentation on its initial release, like in the case of Claude 3, successive model updates with minimal additional documentation[6] can mean public understanding quickly becomes out of date. Other examples of this problem can be seen with GPT-4o, Gemini 1.5, and Grok-2.

Without access to these technical details, external researchers and oversight bodies cannot effectively assess these systems' capabilities, limitations, and potential risks. This technical opacity ultimately undermines the public's ability to understand, find evaluations of, and hold accountable the systems making high-impact decisions about their finances, health, and safety.

FINDING 2: Established Players are More Transparent than AI-native Companies

Our analysis also highlighted a divergence in disclosures between startup AI companies and established technology giants. Meta and Google demonstrate higher levels of transparency in their AI releases, with Meta's open-weight Llama 3.2 model achieving the highest transparency score (88.9) and Google's Gemini 1.5 leading among closed models (62.5). For comparison, the AI startups scored between 59.9 (GPT-4) and 19.4 (Grok-2).

We do not purport to know for certain that this trend would hold if more models were included or why it might be the case. One theory is that these startups simply lack the resources to create robust documentation. However, this theory becomes less plausible when considering documentation on older models from these startups like GPT-2, which contains significantly more detail on technical information like model architecture and training data than GPT-4's technical report. GPT-4's technical report is also even longer than Meta's Llama 3 report.

Another possibility is that this might be partially explained by these companies' business incentives. The startups analyzed are mostly AI-native companies,[7] which primarily generate revenue through selling access to their models. These companies' market positions ostensibly rest on capability advantages. Established big tech companies, on the

[5] While outside the scope of this analysis, reviewing older papers like OpenAI's technical report on GPT-2 further emphasizes that more robust disclosures of technical details are possible.
[6] The Claude 3.5 Sonnett "Model Card Addendum" is just 8 pages, and the October 2024 update to 3.5 Sonnett only received a short blog post.
[7] With the arguable exception of xAI, being a broader part of Elon Musk's interconnected corporations.

other hand, have more diversified revenue streams. Google and Meta have both been accused of lagging on capabilities, but their thriving digital advertising businesses could mean that they can choose to compete instead on transparency. Meta in particular has argued forcefully in favor of openness, partially because, as CEO Mark Zuckerberg noted, "Selling access to AI models isn't our business model."[8]

This pattern raises important questions about whether an AI company's financial position impacts its ability and desire to be transparent. If a link is identified, it would open further questions about how to align commercial incentives with the public interest when it comes to AI transparency, particularly as these systems become further integrated into society.

FINDING 3: Capability Evaluation is Apples-to-Oranges

The documentation drift problem is not the only challenge for meaningfully comparing frontier models. Any attempt to compare these seven models using standard benchmarks faces systematic obstacles. Currently, each company essentially operates its own evaluation framework with varying levels of transparency, making it difficult for researchers, policymakers, and users to make informed comparisons between models or track progress in the field. To combat this, there is a need for standardization in how capabilities are measured and reported.

The first challenge is that the benchmark landscape itself is rapidly evolving – major model releases often evaluate against a somewhat different set of benchmarks than its predecessors, making direct comparisons difficult. For example, when comparing GPT-4 and o1 Preview, the former uses benchmarks like HellaSwag and WinoGrande, which have been removed, while the latter uniquely uses MATH and MathVista.[9]

This shifting evaluation surface is compounded by a deeper challenge: the lack of reproducibility in benchmark results. With the notable exception of Meta's Llama 3.2, for which complete model weights and significant evaluation details are available, there is no reasonable way to reproduce the benchmark results reported by the other AI companies, which impacted all of their scores. Even in cases where documentation is relatively detailed, crucial information is often omitted – such as exactly how questions were presented to the model, what sampling temperature was set to, and what the criteria were for scoring.

[8] Notably, while Meta has been effectively transparent about many aspects of their Llama models, its choice to release Llama's model weights publicly comes with significant risks. Researchers affiliated with the People's Liberation Army have already used Meta's models to develop an AI tool for the Chinese military. This patently harmful result further emphasizes the importance of separating principles like transparency from practices like the unrestricted release of model weights.

[9] While OpenAI might argue these o1's reasoning focus makes it a different type of model which should be evaluated for different capabilities, there is little evidence that users do not view or use the o1 models simply as GPT-4o's successor, considering the shared interface.

Without the ability to independently verify benchmark claims or run identical evaluations across models, the field lacks a true apples-apples comparison of model capabilities. Additionally, there is the elephant in the room: most of these benchmarks are not very effective at measuring the capabilities they are meant to.

ADDITIONAL ANALYSIS

Other distinct but notable takeaways from this analysis include:

- User-Facing Documentation was the best scoring category, with an average score of 3.19 out of 4 and even Grok-2 scoring modestly (1.83).
- With the exception of Grok-2, the models all scored well (3 or above) on systemic risk evaluations, and in general much of the released documentation is through this safety and risk lens.
- Security saw a generally poor performance as well, with many companies evaluating their models' ability to engage in malicious cyber behavior, but not providing much information on how they are protecting these systems.
- There is general opacity surrounding environmental impact (all closed models score a 1 or below), although Meta (4) shows that this can be done.

LIMITATIONS

As with all analyses of emerging technologies, this analysis is constrained in several ways:

- **Few data points** – While covering many of the most widely used frontier models today, this analysis is still far from comprehensive. Additionally, the sample size of seven models limits our ability to draw statistically significant conclusions about broader industry trends.
- **Temporal constraints** – Our analysis captures a specific snapshot in time during a period of rapid development in AI. Given the fast-paced nature of AI advancement, some findings may quickly become outdated as new models are released or existing ones are updated. For example, since this analysis was completed, the full version of OpenAI's o1 model and its documentation was released.
- **Scoring subjectivity** – Despite efforts to establish clear criteria, the scoring process inevitably involves some subjective judgment, particularly when evaluating the completeness and quality of disclosures. This is especially true for categories requiring qualitative assessment, such as risk evaluations.
- **Documentation inconsistency** – The varying formats, depth, location, and organization of model documentation across companies makes direct comparisons challenging. Additionally, it is possible some documentation was overlooked as there are no common repositories.

TOWARDS GREATER TRANSPARENCY

As today's frontier AI models become more powerful and broadly deployed, the urgent need for straightforward and consistent transparency standards is clear. Our analysis finds three critical challenges: first, transparency remains middling in closed models, with especially poor disclosures on technical details. The "documentation drift" problem compounds these opacity concerns, as significant model updates routinely occur with minimal disclosure, creating a widening gap between what we know about these systems and their actual capabilities.

Second, established tech companies like Google and Meta tend to provide more detail than AI startups. This highlights a fundamental tension in the field: while producing comprehensive disclosures is technically feasible, as demonstrated by Llama 3.2, startups have not done so possibly because of either economic disincentives or a lack of administrative capacity. Finally, while many companies evaluate their models against similar benchmarks, there remains insufficient disclosure of the underlying methodology and data, making meaningful, apples-to-apples comparisons between models effectively impossible.

These challenges underscore that the complexity and rapidly evolving nature of frontier models make transparency even more critical. Transparent documentation enables independent experts, investigative journalists, and oversight bodies to identify when a system's capabilities drift away from what was initially described, catch deployments in ill-suited contexts, and detect the emergence of unforeseen risks.

These findings help make the case for decisive, coordinated action. Researchers, industry leaders, and policymakers must work together to develop standards and legislative frameworks for transparency, ensuring that disclosures are both rigorous and practical. Such frameworks would make it feasible to track model changes, verify benchmark claims, and maintain an ecosystem that safeguards development without stifling innovation or legitimate competitive advantage. Without such balanced intervention, we risk a future where the public and regulatory bodies alike remain in the dark, increasingly subject to decisions guided by powerful AI systems whose construction, limitations, and guiding principles are hidden from view.

Appendix A

Score by Release Date

Score and Release Date		
Model	**Release Date**	**Final Score**
Gemini 1.5	2/15/2024	62.5
GPT-4	3/14/2023	59.9
GPT-4o	5/14/2024	53.9
Claude 3.5	6/20/2024	54.4
o1 Preview	9/12/2024	44.7
Grok-2	8/13/2024	19.4

Category Group Descriptions

Category Group Title	**Data Points Contained**
User-Facing Documentation	Model Use Guidelines, Capabilities & Limitations, Changes from Previous Distinct Model, Access Methods, Input/Output Formats, Knowledge Cutoff
Technical Transparency	External Tool Integration, Training Data Composition, Model Architecture, Model Size, Training Time, Post-Training Enhancements, Interpretability and Explainability Techniques (although these were not scored across the board)
Risk & Safety	Alignment Principles, Security, Privacy Controls, Systemic Risk Evaluations
Evaluation & Impact	Model Release Criteria, Environmental Assessment, Industry Benchmarks, Direct Risk Evaluations

Category Group Scores

Subcategories	Llama 3.2	Gemini 1.5	GPT-4	GPT-4o	o1 Preview	Claude 3/3.5 Sonnet	Grok-2
User-Facing Documentation	4.00	3.08	3.67	3.70	2.90	3.17	1.83
Technical Transparency	3.50	2.00	1.20	0.67	0.67	1.00	0.17
Risk & Safety	3.50	2.75	2.50	2.25	1.75	2.38	0.50
Evaluation & Impact	3.00	2.13	1.88	2.38	2.13	2.25	0.38

Metric Descriptions

Metric Name	Description
Model Use Guidelines	A description of the intended, unintended, and prohibited uses of the model.
Capabilities & Limitations	A description of the model's capabilities and limitations.
Changes From Previous Distinct Model	A high level overview of how the model differs from the previous (major named) version (e.g., GPT-3 vs GPT-4).
Access Methods	A list of available ways for customers and the public to use (/access, in the case of open models) the model from the developer's systems/infrastructure (e.g., a platform like ChatGPT or an API).
Input/Output Formats	A list of modalities the model can parse and produce (e.g., text, images, audio).
External Tool Integration	A description of the model's ability to connect to and utilize external software, APIs, databases, etc.
Training Data Composition	Information about the model's training data, centered around key points such as the data's sources, size, creators, selection/filtration criteria, etc.
Knowledge Cutoff	The time after which the model has not been trained on any new data.
Model Architecture	Information on the basic model type (e.g., transformer) and how the model was trained (e.g., supervised learning, reinforcement learning).
Model Size	The number of parameters in the model.
Training Time	The amount of time and computational power required to train the model, typically measured in GPU hours.
Post-Training Enhancements	A description of any additional changes made to the model after initial pretraining (e.g., fine tuning, or the addition of capabilities like chain-of-thought reasoning).
Interpretability and Explainability Techniques	An overview of methods taken to make the

	model's behavior and internal processes more understandable to humans (e.g., mechanistic interpretability techniques or post-hoc explanation).
Alignment Principles	A description of the developer's procedures/guiding principles for determining desirable model behavior, preventing undesirable model behavior, and resolving conflicts either between or within desirable and undesirable behavior.
Security	Information about what steps the developer takes to institute: 1. traditional cybersecurity protocols (e.g., access controls, etc.) and 2. protections unique to AI systems, such ways to prevent model weight exfiltration, data poisoning, and prompt injections.
Privacy Controls	An overview of measures taken to protect user privacy when interacting with the model. This could include data retention practices, encryption standards, data sharing practices, breach notifications, treatment of sensitive data, etc.
Model Release Criteria	Under what criteria and circumstances should a model be/not be cleared for public release? How does the model perform against those criteria?
Environmental Impact	What impact does the model's training and ongoing use have on the environment? How much energy does operating the model require? How much water (or water byproduct) does it require? What carbon offsetting is being done? Etc.
Industry Benchmarks	Standardized performance metrics from established industry and academic benchmarks, including methodology used, raw scores, comparative results against baseline models, and information about benchmark versions and potential training data contamination.
Systemic Risk Evaluations	Assessment of model risks and testing conducted to evaluate potential catastrophic harms, including systematic testing for CBRN

	capabilities, autonomous replication, large-scale manipulation, etc. Encompasses both internal and external testing (including specialized red teaming), automated safety evaluations, long-term mitigation strategies, and ongoing monitoring plans for emerging systemic risks.
Direct Risk Evaluations	Assessment of model risks and testing conducted to evaluate user-facing and operational concerns, including systematic testing for harmful outputs, bias, hallucination rates, jailbreak resistance, etc. Encompasses both internal and external testing (including red teaming), automated evaluations, immediate mitigation strategies, and routine monitoring protocols.

Scoring System

Base Rating	**Brief Description**
0	Nothing mentioned regarding the given item of disclosure, or information is explicitly withheld
1	Extremely high-level mentions without meaningful detail (e.g., "the model has been trained on public and private data from the internet, filtered for quality" or "we use standard cybersecurity protocols")
2	Some specifics are given, but significant gaps remain (e.g., training data includes filtered web content, CommonCrawl, and academic papers, with basic quality filters applied" or "safety testing included red teaming and automated evaluations")
3	Mostly detailed information with minor to moderate gaps (e.g., "training data includes [vague mention to public and private data], filtered using [some technique names] with [x%] removed due to [y]" or "safety evaluation included [x] rounds of red teaming by [y] external experts finding [z] issues which were addressed" with no mention of testing methodology).
4	Comprehensive information meeting all reasonable disclosure expectations.

Score Modifiers

Modifier Label*	Additional Considerations	Score Impact	Applies to:	Notes
A	Item of disclosure is not present in the most recent model documentation** in a form that would score over a 1	-0.75	All categories	Cannot bring a score below 0. Only applies if previous documentation's disclosure item would have scored over a 1. This deduction will be applied to a disclosure item unless the most recent model documentation explicitly states that item (or any/all unnamed items) has not changed from previous documentation.
B	Information not disclosed but the organization has significant other work on the topic	1.00	All categories	
C	Information not disclosed in model documentation*** but is disclosed elsewhere	2.00	All categories	
D	Specific category's information is spread over multiple documents	-0.50	All categories	
E	Information disclosed over 2 months after model release	-0.50	All categories	
F	Information must be inferred, either from unclear language or from research papers and other non-model documentation	-1.00	All categories	
G	Does not allow for full reproducibility	-0.50	Industry Benchmarks	Cannot bring a score below 0

* For use in detailed disclosure evaluations
** 'Most recent model documentation' refers to documentation about the latest named update to a model (e.g., Claude 3.5, Gemini 1.5, Llama 3.3, etc.)
*** 'Model documentation' refers to blog posts, Model Cards, System Cards, and other documentation primarily focused on one model/family of models

Appendix B

USER-FACING DOCUMENTATION METRICS

	Gemini 1.5 (Pro & Flash)	GPT-4	GPT-4o	Claude 3/3.5 Sonnet	o1 Preview	Grok-2	Llama 3.2
1 **Model Use Guidelines**	Intended uses specified on pg. 105 of Gemini 1.5 Technical Report and prohibited uses are outlined in the Gemini app policy guidelines and prohibited use policy 4 - 0.5 (D)	Discusses general positive uses on page 42 of the Technical Report. Explains harmful content meant to be mitigated on page 47. Discusses system safety including moderation and usage policies on page 66, link to OpenAI's broader usage policies 4	Intended uses are inferred through general descriptions of capabilities, such as in the intro of the System Card. Mentions mitigating information harms, bias and discrimination, and content that violates usage policies on pg. 2 of the System Card, along with further details throughout 3	Properly detailed sections for each on the Claude 3 Model Card pg.2, also referencing the Anthropic Acceptable Use Policy. Missing from 3.5 documentation 4 - 0.75 (A)	Minor discussion of this being a reasoning model for hard STEM problems in the release blog, with only a single mention of the org's usage policies in the System Card 1	Not disclosed at present 0	Llama 3.2 repository contains the majority of relevant documentation. Intended and unintended uses detailed in Acceptable Use Policy and the Model Card 4
2 **Capabilities & Limitations**	Capabilities are discussed throughout the Technical Report, including in the intro, and in sec. 6 (pg. 28) which evaluates core capabilities Pg. 107 of the Technical Report explicitly points to 1.0's Technical Report as describing continually relevant limitations 4	Capabilities and limitations sufficiently discussed on pgs. 4-11 of the 4 Technical Report. Limitations further discussed starting on page 44. 4	Capabilities are most clearly laid out in the release blog while limitations are discussed on System Card pg. 12 4 - 0.5 (E)	Core capability evaluations and results described in 3's Model Card pgs. 4-23. Limitations discussed on pgs. 31-32. Updated capabilities overviewed in 3.5's Model Card Addendum pgs. 1-5 4	Capabilities discussed throughout the o1 release blog, such as in the "whom it's for" section, as well as a series of videos on the blog Limitations are discussed through the release blog ("How it works"), evaluations blog, and System Card, although ideally information would be more aggregated and explicitly labeled limitations 4 - 0.5 (E)	Release blog mentions "frontier capabilities in chat, coding and reasoning" without much more detail. No evaluation or formal discussion of limitations 1	Discusses capabilities at a high level throughout the release blog. Discusses limitations at the end of the Model Card 4 - 0.5 (E)
3 **Knowledge Cutoffs**	External sources report this to be November 2023, but not verified from any Google documentation 0	Up to September 2021, as per Technical Report pg. 10 4	October 2023 per System Card pg. 1 4	August 2023 cutoff, as per 3's Model Card pg. 1. Not mentioned in 3.5's Model Card Addendum 4 - 0.75 (A)	October 2023, as per API documentation 4 - 1 (F)	Unspecified, but does indicate access to real-time data via X (formerly Twitter) 1	December 2023, as per Model Card 4

4 **Changes From Previous Distinct Model**	The existence of this new robust Technical Report for 1.5 -- a model within the same family as 1.0 -- is helpful to this point in its own right. This is primarily covered in the introduction, pgs. 1-5, which covers both performance improvements from 1.0, as well as the way 1.5 has improved since its release to the public a month prior 4	Metric comparisons through the Technical Report, but no high level overview of capability changes outside brief references in the release blog 2	Debatable whether this is considered a new version of GPT-4, but since it given the novel description of an "omni" model we will treat it as separate. Several metric comparisons are still made throughout the documentation, such as in the release blog N/A	No high level overview of capability changes relative to Claude 2. Throughout 3's Model Card, authors compare 3 to Claude 2, including on human evaluated head-to-head tests, but it is done so selectively. Importantly, Authors do not list Claude 2 performance on any of the benchmark tables. 3.5's Model Card Addendum compares the model's performance with Claude 3 both in the capabilities descriptions and in benchmark results 2	Comparisons in evaluation metrics throughout the evaluations blog and the System Card, but in general this is the first generation of these reasoning-focused models N/A	Not meaningfully summarized within a single document, but the Grok-2 benchmarks found in the release blog offer a comparison to Grok 1.5 1	Model card has a section on "new capabilities" albeit a slightly high level one 4
5 **Access Methods**	The Gemini 1.0 Technical Report overviews Gemini, Gemini Advanced, and Gemini APIs. The release blog does mention AI Studio and Vertex AI, but it's unclear whether this is comprehensive 3	ChatGPT Plus and API as specified in release blog 4	ChatGPT and the API, as per the release blog 4	Overviewed on 3's Model Card pg. 1 but omitted from 3.5's Card 4 - 0.75 (A)	ChatGPT and API, as specified in release blog 4	Grok-2 and Grok-2 mini are stated in the release blog as available through the X platform (website and phone application). Additionally, both models were stated to be made available through xAI's enterprise API "later this month" 4	Weights are openly distributed via the Llama website 4
6 **Input/Output Formats**	In the Gemini 1.5 Model Card on Technical Report pg. 105 4	Text I/O and image input as noted in release blog 4	Text, vision, and voice, as per the release blog 4	Overviewed on 3's Model Card pg. 1 but omitted from 3.5's Card 4 - 0.75 (A)	Not explicitly stated, but inferable through the disclaimer in the release blog "how it works" section 4 - 1 (F)	In the Grok-2 release blog it is written that the model possesses "advanced capabilities in both text and vision understanding [...] across a wide range of tasks, whether you're seeking answers, collaborating on writing, or solving coding tasks." Could be more clear/formalized 4	Multimodal capabilities explained in model cards and the release blog 4

TECHNICAL TRANSPARENCY METRICS

	Gemini 1.5 (Pro & Flash)	GPT-4	GPT-4o	Claude 3/3.5 Sonnet	o1 Preview	Grok-2	Llama 3.2
7 **External Tool Integration**	The section on Function calling (Gemini 1.5 Technical Report pgs. 32-33) addresses this 4	GPT-4 did not have the capacity for function calling at the time of its release, but the API documentation now has a relevant section. ChatGPT Plugins were also released shortly following the release of GPT-4 N/A	Capacity for tool use acknowledged throughout the System Card, such as on pg. 20, but not explained 1	3's Model Card pg. 1 mentions the model excels at tool use. In their release statement, Anthropic indicated this feature would come at a later date. Two months later Anthropic released more information on this stipulating that users can provide Claude tool access, but there are none provided server-side. 3.5's Model Card Addendum does not discuss this 4 - 0.75 (A)	Tool use is involved in evaluations several times throughout the System Card, such as pgs 29-31, but not clear whether the public version has access 2	Release blog mentions significant improvements in "its tool use capabilities" without providing further details 1	Release blog mentions tool use several times, but only the 3.1 release blog actually links to a site on the topic/broader system integrations. Information in the model documentation is relatively sparse 3 - 0.75 (A) - 0.5 (D)
8 **Training Data Composition**	Very little information in the Technical Report or elsewhere beyond "data sourced across many different domains, including web documents and code, and incorporates image, audio, and video content" 1	Techincal Report pg. 2 explains the model is trained on public data found on the internet and private data from partnerships 1	System Card pgs. 1-2 overviews the pre-training dataset with moderate specificity and explains several filtering steps 2	Little to no information. 3's Model Card describes a mix of public and non-public data on pg. 3. Also has a high level overview of data crawling policies (follow robots.txt, don't bypass CAPTCHA, transparent crawling). 3.5's Model Card Addendum also does not discuss this 1	System card pg. 1 describes the standard use of public data and proprietary data. Also briefly describes data filtering and reinforcement learning without providing details 1	Not disclosed at present 0	Strong overview of filtering and cleaning techniques, info on multilingual data, other domain-specific datatypes in Llama 3's Technical Report. Better overview of the pretraining corpus in 3.2's Model Card than other models, but still significantly opaque. Uniquely, has a discussion of how 3.1's data was used to guide 3.2's training in the Model Card 2 - 0.75 (A)

9 **Model Architecture**	Architectural overview starts on pg. 5 of Gemini 1.5 Technical Report and also has an overview on pg. 105 4	Techincal Report pg. 2 explains this is a "Transformer-style model pre-trained to predict the next token in a document" 1	Not disclosed at present 0	No infromation beyond being a "new family of large multimodal models" in 3's Model Card 0	Only informaiton comes from high level distinctions from existing models. Pg. 1 of the System Card states that it uses a "large-scale reinforcement learning" algorithm to engage in chain-of-thought reasoning post-query. Does not mention anything along the lines of it using the transformer architecture 0	Not disclosed at present 0	Llama 3's Technical Report starting on pg. 6, explains how it deviates from previous versions' architecture. Provides crucial details such as # of layers 4 - 0.75 (A)
10 **Model Size**	Not disclosed at present 0	Not disclosed at present 0	Not disclosed at present 0	Not disclosed at present 0	Not disclosed at present 0	Not disclosed at present 0	Three sizes available for use: 1B, 3B, 11B, and 90B, with the latter two being for the vision models, as per model cards 4
11 **Training Time**	Not disclosed at present 0	Not disclosed at present 0	Not disclosed at present 0	Not disclosed at present 0	Not disclosed at present 0	Not disclosed at present 0	Calculation in GPU hours for each size and version overviewed in model cards 4
12 **Post-Training Enhancements**	Discussion of SFT and RLHF on 1.5's Technical Report pg. 52 - relatively high level but still useful 3	Discussion begins on pg. 12 of the Technical Report with reference to RLHF and mentions of other safety fine tuning. Primary discussion is found on pgs. 61-64 and extensively discussing fine-tuning and RLHF. Throughout the report, the primary purpose of fine-tuning is explicitly stated here to be aligning responses with user intent. 4	Mentions that the model is aligned to human preferences on System Card pg. 2 1	3's Model Card pg. 3 explains that Constitutional AI (CAI) is used and, at a high level, how, which Anthropic has expanded on in separate research papers. Not mentioned in 3.5's Model Card Addendum 3 - 0.75 (A) - 0.5 (D)	One mention of an RLHF step on System Card pg. 6 and notes that the model uses RL to engage on chain-of-thought reasoning 1	Not disclosed at present 0	Technical Report section on post-training (pg. 15) contains a solid overview, effectively reiterated in the model cards 4

13 **Interpretability and Explainability Techniques*** ***Points will not be deducted for this as it is evolving science with few best practices**	Google DeepMind has previously completed work on interpretability, such as building a compiler for the RASP language, but does not mention any of it in the context of Gemini N/A	None presented at GPT-4's launch, but OpenAI has presented work on interpretability including on extracting features from GPT-4 and on using LLMs to explain themselves/others N/A	None presented but as otherwise noted, OpenAI has helped develop this field N/A	Nothing disclosed at the time of release, but Anthropic has been a pioneer of technical interpretability. They released a paper on analyzing "features" in Claude 3 (2 months post-release) via dictionary learning that provided unparalleled insight into a SOTA LLM N/A	The fact that the model reasons aloud using chain-of-thought inherently makes it more transparent. However, OpenAI has explicitly chosen to hide the true chains of thought for "user experience, competitive advantage, and the option to pursue the chain of thought monitoring" ("Hiding the Chains of Thought" in the evaluations blog), making it hard to know whether what you see is actually reflective of the model's processes N/A	Not disclosed at present N/A	Inherently more interpretable than others by releasing model weights, but no special tools released N/A

RISK AND SAFETY METRICS

	Gemini 1.5 (Pro & Flash)	GPT-4	GPT-4o	Claude 3/3.5 Sonnet	o1 Preview	Grok-2	Llama 3.2
14 **Alignment Principles**	Technical Report section 9.2 -- Policies and Desiderata -- starting on pg. 50 overviews this effectively. The following section explains how development details are conducted to align with those goals 4	Alignment is discussed throughout the Technical Report, including pgs. 11-14 and more extensively discussed on 61-69. There is not much explicitly stated about positive alignment and trade-offs between different types of behavior. This is further articulated by OpenAI later with their Preparedness Framework. 3	Discusses alignment throughout, such as on System Card pgs. 5-6. The section on OpenAI's Preparedness Framework on pgs. 12-13 and the assessments that follow provide good detail, but are mostly in the context of unwanted behavior and risks 3	Claude 3 Model Card pg. 4 points to their work on CAI, referencing the "set of ethical and behavioral principles that the model uses to guide its outputs." Discussion of this topic is missing from 3.5's Addendum 4 - 0.75 (A)	Despite significant safety testing done, and a moderate discussion of undesirable model behavior in the System Card's section on Preparedness Framework Evaluations (pg. 13), there are no documented guiding principles that answer the question "what are we aligning to" 2	Not disclosed at present 0	Llama 3's technical Report pgs. 15-19 discuss various alignment techniques under the theme of aligning with human preferences. This is more of a discussion of what they do (SFT, DPO, reward modeling, etc.) than a philosophical overview. The model cards have a broad overview the model's values, which helps round out the picture, although these two elements should be better connected (the alignment goals and the techniques to achieve them) 3 - 0.5 (D)
15 **Security**	Good discussion in the 1.5 Technical Report about LLM specfic attacks like prompt injection on pg. 60. Not so much information on the company's practices to secure the model and their infrastructure 2	Discussion on Technical Report pgs. 53-54 of the model's cybersecurity capabilities as a threat vector, but little information on how the model and surrounding systems are secured. OpenAI Security Portal has a better general discussion of internal company security protocols and certifications 1	Pg. 13 of the System Card details how the model was evaluated for dangerous cybersecurity capabilities, but no notable mentions of protecting the model itself or surrounding infrastructure 0	Pg. 4 in Claude 3's Model Card lists a variety of security measures (e.g., MFA, two-party controls). Additionally, pg. 26 points to specific security commitments for similar models in their RSP. Further security and compliance information in their Trust Portal. The Model Card does not contain robust information on AI-specific attacks like data poisoning. The 3.5 Addendum does not include information on either 2 - 0-75 (A)	Pg. 14 of the System Card details how the model was evaluated for dangerous cybersecurity capabilities. It is not discussed how the model/system is itself secured 0	Not disclosed at present 0	Apples to oranges enough as an open weight model. Meta releases system guards (Llama Guard, Prompt Guard, and Code Shield) alongside their models to attempt to bolster these open system, as described in the model cards N/A

16 **Privacy Controls**	Google's Privacy Policy and the Gemini Apps Privacy Notice mostly cover this, although it is not well articulated in model-specific documentation 0 + 2 (C)	Minor discussion of privacy details and mitigations on pg. 53 of Technical Report. More information in the OpenAI Privacy Portal 2	Privacy risk acknowledged on pg. 7 of the System Card, along with the impact a mitigation had on the risk, as well as pg. 10. "Advanced data filtering" mentioned to reduce personal information in the training data on pg. 2 without more detail 2	Refers to their privacy policy on pg. 3 in 3's Model Card but does not go into detail on model-specific protections/processes 1	System Card pg. 2 briefly mentions filtering personal data out of training data 1	xAI's privacy policy offers an overview of this 0 + 2 (C)	Also apples to oranges enough. Privacy is a category addressed in "Llama Guard". Separately, Meta does have a generative AI privacy guide which gives more information N/A
17 **Systemic Risk Evaluations**	Pgs. 68-72 of the Technical Report overviews dangerous capability evaluations. Some subsections are lacking in detail or evaluation results but does overview methodology and distinguishes between in-house and external testing approaches 3	Strong discussion of testing for risks across CBRN, autonomous replication, and manipulation, discussed across both internal and external testers in the Technical Report beginning pg. 11-14 and continuing 44-60 4	Preparedness framework evaluations run in the System Card from pgs. 12-17, detailing systemic risks. Additional external assessments run on 19-20, with additional info in the appendix 4	Pgs. 23-28 in Claude 3's Model Card overview testing and evaluations, for catastrophic risk, trust and safety, and elections integrity. Particularly the catastrophic risk work points to guidelines in their RSP. Pg. 6 in 3.5's Addendum provides a solid but lacking in detail overview of Anthropic's safety evaluations and commitments regarding the new model 4	Detailed discuss throughout the System Card's sections on "observed safety challenges and evaluations" (pgs. 2-13) and "Preparedness Framework evaluations" (pgs. 13-32) with additional information in the Appendix, including discussions of internal and external testing 4	Not disclosed at present 0	Model cards have a section on critical risks, but a more robust overview can be found on pgs. 40-50 in 3's Technical Report which overviews safety evals 4

EVALUATION AND IMPACT METRICS

	Gemini 1.5 (Pro & Flash)	GPT-4	GPT-4o	Claude 3/3.5 Sonnet	o1 Preview	Grok-2	Llama 3.2
18 **Model Release Criteria**	Pg. 49 of the Technical Report mentions evaluation by the "Responsibility and Safety Council" who makes release decisions, largely based on alignment with the company's AI principles. However, the specifics of that process and how this apply to Gemini 1.5 specifically are somewhat unclear 2	Not presented with GPT-4's launch, but this has since been improved with the public development of their Preparedness work 0 + 1 (B)	Overviewed in the System Card's section on "Preparedness Framework Evaluations" on pg. 12 4	3's Model Card pg. 3 has a section for their release decision. In it, they mention their RSP, as well as having their decision guided by the NIST AI RMF, including using red teaming, incremental rollouts, etc. This section is strong but lacks detail. There is only a brief mention of the RSP on the 3.5 Addendum pg. 6 3 - 0.75 (A)	Overviewed in the System Card's section on "Preparedness Framework Evaluations" on pg. 13 4	Not disclosed at present 0	Describes a three-pronged strategy in the model cards as part of their "responsible release approach." Not a great evaluation of how this model fares when evaluated against any specific criteria in that approach 2
19 **Environmental Impact**	While not in Gemini's documentation, Google does release an annual environmental report in which they mention a recent 13% YoY increase in energy consumption due largely to data center costs 0 + 1 (B)	Not disclosed at present 0	Environmental harms mentioned on System Card pg. 19 and references to OpenAI's other work, however no details of this specific model's impact provided 1	Anthropic does not disclose energy used by their models, but they do provide a high level overview of their efforts to hit net zero climate impact through offsets pg. 4 of the Model Card. No data or details are provided. 1	Not disclosed at present 0	Not disclosed at present 0	Model cards provide information on both energy use and carbon emissions. They also provide a separate document detailing their methodology 4

20 **Industry Benchmarks**	Uses standard benchmarks to evaluate core capabilities (pg. 28 of Technical Report) but only provides moderate evaluation details and in the paper, only compares benchmarks to other Gemini models and lacks information for reproducibility 3 - 0.5 (G)	A table of benchmark results appears on pg. 7 of the Technical Report, including a methodology for multi-shot prompting and benchmark selection. Correctly compares results to SOTA, but is lacking on methodological details and are not fully reproducible 3 - 0.5 (G)	Some standard benchmarks provided in release blog, with the System Card pg. 8 providing slightly more detail, but very little context for decisions like reporting on "a subset of MMLU" 2 - 0.5 (G)	Claude 3's Model Card gives standard tables of evaluation metrics on pg. 6-8. Throughout the document, there is moderate discussion of evaluation details but only for some benchmarks. 3.5's Addendum presents comparative standard benchmarks starting on pg. 2, but lacks methodological details and is not fully reproducible 3 - 0.5 (G)	The evaluations blog appendix provides scores on a number of standard benchmarks. However, very little information beyond the raw score is given 2 - 0.5 (G)	Grok-2's release blog presents standard industry benchmarks, but lacks any methodological detail or information required for reproducibility 2 - 0.5 (G)	Standard benchmarks in results section from the model cards. Overall does a good job with a large number of benchmarks. Also, importantly, provides information on how they arrived at their results and their methodology. Eval information further detailed here 4
21 **Direct Risk Evaluations**	Sectio on Technical Report pg. 51 overviews training for safety and direct risks. Section beginning on pg. 52 overviews the model's levels of risk through things like violations of toxicity policies and robustness to jailbreaks. Separate section on representational harms on pg. 65. Missing robust discussion of hallucinations 3	Strong discussion throughout the Technical Report's section on "Observed Safety Challenges" (pg. 44) including hallucinations on pg. 46, harmful outputs, bias, and other harms of representation on pgs. 47-50 4	Strong discussion from System Card pgs. 19-24 on societal impacts including hallucinations, anthropomorphization, health, etc. Bias mentioned as a topic of external testing and given acknowledgement throughout but without detail 3	Pgs. 27-32 in 3's Model Card overviews risks like discrimination, elections integrity, hallucinations, etc. These include quantitative metrics, red teaming, etc. The section on areas for improvement starting on pg. 31 is particularly valuable. Most of these elements are absent from 3.5's Addendum 4 - 0.75 (A)	Moderate to strong discussion throughout the System Card including disallowed content and jailbreak evaluations on pgs. 3-4, hallucination evaluations on pgs. 4-5, and bias evaluations on 5-6. However, a number of the evaluations lacked detail either in methodology, results, or both, such as the section on "regurgitation evaluation" 3	Not disclosed at present 0	Moderate discussion on pgs. 50-51 of 3's Technical Report. However, especially considering this model is open, they could do with a more robust discussion/evaluation of potential risks. Llama Guard is advertised to help in some cases, such as with "hate" related inputs and outputs. Very little discussion in 3.2 specifically 2 - 0.75 (A)