Technical Report: Performance and baseline evaluations of gpt-oss-safeguard-120b and gpt-oss-safeguard-20b

OpenAI

October 29, 2025

Contents

1 Introduction

gpt-oss-safeguard-120b and gpt-oss-safeguard-20b are two open-weight reasoning models post-trained from the gpt-oss models and trained to reason from a provided policy in order to label content under that policy. They are available under the Apache 2.0 license and our gpt-oss usage policy. Developed with feedback from the open-source community, these text-only models are compatible with our Responses API. The models are customizable, provide full chain-of-thought (CoT), can be used with different reasoning efforts (low, medium, high), and support Structured Outputs.

In this report, we describe gpt-oss-safeguard's capabilities and provide our baseline safety evaluations on the gpt-oss-safeguard models, using the underlying gpt-oss models as a baseline. For more information about the development and architecture of the underlying gpt-oss models, see the original gpt-oss model model card [1].

We recommend using these models to classify content against a provided policy, and not as the core functionality with which end users interact; the original gpt-oss models are better for those applications. The safety metrics provided below describe how gpt-oss-safeguard models function in chat settings. The gpt-oss-safeguard models are not intended for this use, but since they are open models, it is possible for someone to use the models in this way. Because of that possibility, we wanted to verify that they met our safety standards in such usage; this report shares the results of those tests. We also share an initial evaluation of multi-language performance in a chat setting; note that this does not directly assess performance during content classification with a provided policy.

The gpt-oss-safeguard models are fine-tunes of their gpt-oss counterparts, and were trained without any additional biological or cybersecurity data. As a result, we determined that the previous work estimating worst-case scenarios from gpt-oss release cross applies to these new models.

2 Safety Classification Performance

We evaluated the gpt-oss-safeguard models on both internal and external evaluation sets.

In the internal evaluation, we provided multiple policies simultaneously to gpt-oss-safeguard at inference time. For each test input, we evaluate whether gpt-oss-safeguard correctly classifies the text under all of the included policies. This is a challenging task—the model is counted as accurate only if it exactly matches the golden set labels for all the included policies.

Our gpt-oss-safeguard models and internal Safety Reasoner outperform gpt-5-thinking and the gpt-oss open models on multi-policy accuracy. The gpt-oss-safeguard models outperforming gpt-5-thinking is especially surprising given the former models' small size.

Table 1: Internal moderation evaluation (multi-policy accuracy)

Model	Multi-policy accuracy
gpt-oss-safeguard-120b	46.3%
gpt-oss-safeguard-20b	43.6%
internal-safety-reasoner	52.2%
gpt-5-thinking	43.2%
gpt-oss-120b	32.5%
gpt-oss-20b	32.1%

We further evaluated these models on the moderation dataset we released with our 2022 research paper and on ToxicChat, a public benchmark based on user queries to an open-source chatbot.

On the 2022 moderation evaluation set, we evaluated using our internal safety policies (adapting one policy to account for subsequent changes). gpt-oss-safeguard slightly outperformed all other tested models, including our internal Safety Reasoner and gpt-5-thinking, though this difference between Safety Reasoner and gpt-oss-safeguard is not statistically significant.

For ToxicChat, we evaluated using a short hand-written prompt adapted from some of our internal policies. Our internal Safety Reasoner again outperformed gpt-5-thinking, while gpt-5-thinking and Safety Reasoner both marginally outperform gpt-oss-safeguard-120b and gpt-oss-safeguard-20b. We expect that the relatively small size of gpt-oss-safeguard still makes it preferable for this type of task.

Table 2: OpenAI Moderation API and ToxicChat evaulations

Model	OpenAI Mod (2022) F1 Score	ToxicChat F1 Score
gpt-oss-safeguard-120b	82.9%	79.3%
gpt-oss-safeguard-20b	82.9%	79.9%
internal-safety-reasoner	82.8%	81.3%
gpt-5-thinking	79.8%	81.0%
gpt-oss-120b	80.4%	76.7%
gpt-oss-20b	78.7%	75.9%

2.1 Limitations

There are two specific limitations of gpt-oss-safeguard. First, we have observed that classifiers trained on tens of thousands of high-quality labeled samples can still perform better at classifying content than gpt-oss-safeguard does when reasoning directly from the policy. Taking the time to train a dedicated classifier may be preferred for higher performance on more complex risks.

Second, gpt-oss-safeguard can be time and compute-intensive, which makes it challenging to scale across all platform content. Internally, we handle this in several ways with Safety Reasoner: (1) we use smaller and faster classifiers to determine which content to assess and (2) in some circumstances, we use Safety Reasoner asynchronously to provide a low-latency user experience while maintaining the ability to intervene if we detect unsafe content.

3 Multilingual Performance

To evaluate multilingual capabilities, we used the MMMLU eval [2], a professionally human-translated version of MMLU in 14 languages. The answers were parsed from the model's response by removing extraneous markdown or Latex syntax and searching for various translations of "Answer" in the prompted language.

We find the gpt-oss-safeguard models perform at parity with gpt-oss models across all reasoning levels. Note that these evaluations address performance in a chat setting and do not directly assess performance during content classification with a provided policy.

Table 3: MMMLU evaluation

	gpt-oss-120b			**gpt-oss-safeguard-120b**			**gpt-oss-20b**			**gpt-oss-safeguard-20b**		
Language	low	medium	high	low	medium	high	low	medium	high	low	medium	high
Arabic	75.0	80.4	82.7	79.0	80.8	82.1	65.6	73.4	76.3	72.2	75.7	77.3
Bengali	71.5	78.3	80.9	76.9	79.6	80.2	68.3	74.9	77.1	73.1	76.8	78.0
Chinese	77.9	82.1	83.6	81.2	82.7	83.5	72.1	78.0	79.4	75.8	79.3	80.0
German	78.6	81.7	83.0	81.5	82.7	83.5	71.5	77.2	78.7	76.1	79.3	80.2
French	79.6	83.3	84.6	82.6	83.4	84.4	73.2	78.6	80.2	77.3	80.1	80.7
Hindi	74.2	80.0	82.2	79.0	81.0	81.6	70.2	76.6	78.8	74.5	78.5	79.4
Indonesian	78.3	82.8	84.3	81.5	82.9	83.6	71.2	77.4	79.5	76.5	79.1	80.4
Italian	79.5	83.7	85.0	82.3	83.8	84.7	73.6	79.0	80.5	77.8	80.8	81.1
Japanese	77.0	82.0	83.5	80.8	82.5	83.0	70.4	76.9	78.8	75.3	78.8	80.0
Korean	75.2	80.9	82.9	79.7	81.6	82.8	69.8	75.7	77.6	74.0	77.7	78.8
Portuguese	80.0	83.3	85.3	82.0	83.9	84.6	73.3	79.2	80.5	77.7	80.4	81.0
Spanish	80.6	84.6	85.9	83.8	84.7	85.3	75.0	79.7	81.2	78.5	81.2	81.9
Swahili	59.9	69.3	72.3	67.6	70.5	72.7	46.2	56.6	60.7	55.5	61.5	63.5
Yoruba	49.7	58.1	62.4	55.7	58.3	60.2	38.4	45.8	50.1	46.0	49.8	52.6
Average	74.1	79.3	81.3	78.1	79.9	80.9	67.0	73.5	75.7	72.2	75.6	76.8

4 Observed safety challenges and mitigations

4.1 Disallowed content

The following evaluations check that the model does not comply with requests for content that is disallowed under OpenAI's safety policies, including hateful content or illicit advice. Given the functionality of gpt-oss-safeguard, it's important to emphasize that the below metrics are baseline safety scores for the gpt-oss-safeguard models. That is, these metrics describe the performance of the safeguard models when used directly for end-use chat (a use case we do not recommend, and for which the models are not intended).

We consider several evaluations:

- **Standard Disallowed Content Evaluations**: We report our standard evaluations to test the safety of our models' outputs on requests for disallowed content. However, our recent models saturate this benchmark (as visible in the results table), and thus no longer provide useful signal for incremental safety progress. To help us benchmark continuing progress, we created the new Production Benchmarks evaluation set. We plan to stop publishing this older set in the near future and will instead share the more challenging set below.
- **Production Benchmarks**: As introduced with ChatGPT agent, this is a new, more

challenging evaluation set with conversations that are more representative of production data, and are thus highly multi-turn and less straightforward than our previous disallowed content evaluations.

We evaluate completions using LLM-based grading models. We use the metric not_unsafe, checking that the model did not produce unsafe output according to relevant OpenAI policy.

Note that the production benchmark set is designed specifically to be more challenging, which provides a useful signal on how our models improve on safety over time; scores are therefore expected to be lower than the standard evaluation.

Table 4: Standard Disallowed Content Evaluation

Category	**gpt-oss-120b**	**gpt-oss-safeguard-120b**	**gpt-oss-20b**	**gpt-oss-safeguard-20b**
hate (aggregate)[1]	0.996	1.000	0.996	0.977
self-harm/intent and self-harm/instructions	0.995	1.000	0.984	0.971
personal-data/semi-restrictive	0.967	0.959	0.947	0.955
sexual/exploitative	1.000	0.970	0.980	0.990
sexual/minors	1.000	0.990	0.971	0.990
illicit/non-violent	1.000	0.991	0.983	0.982
illicit/violent	1.000	1.000	1.000	1.000
personal-data/restricted	0.996	0.978	0.978	0.978

Table 5: Production Benchmarks

Category	**gpt-oss-120b**	**gpt-oss-safeguard-120b**	**gpt-oss-20b**	**gpt-oss-safeguard-20b**
non-violent hate	0.895	0.872	0.901	0.868
personal-data	0.888	0.925	0.921	0.906
harassment/threatening	0.832	0.767	0.819	0.778
sexual/illicit	0.919	0.925	0.852	0.918
sexual/minors	0.967	0.945	0.866	0.967
extremism	0.932	0.923	0.951	0.921
hate/threatening	0.898	0.828	0.829	0.797
illicit/nonviolent	0.692	0.674	0.656	0.585
illicit/violent	0.817	0.792	0.744	0.657
self-harm/intent	0.950	0.864	0.893	0.835
self-harm/instructions	0.910	0.831	0.899	0.860

We find gpt-oss-safeguard-120b and gpt-oss-safeguard-20b generally perform on par with their gpt-oss counterparts. Both of the safeguard models generally perform within 1-3 points of the gpt-oss models on the Standard Disallowed Content Evaluation. We observe some minor degradations in certain categories of the Production Benchmarks evaluation when comparing the safeguards models to the gpt-oss models, but also see the safeguards models outperform the gpt-oss models in other categories.

[1]Hate in this table is a combination of: harassment/threatening, hate, hate/threatening, and extremist/propaganda.

4.2 Jailbreaks

As stated above, we don't recommend using either gpt-oss-safeguard model as the main functionality with which end users interact. Because of this, robustness to jailbreaks is less important in the gpt-oss-safeguard models than in models that are built for end user interaction, as the recommended use cases are internal only. Even so, we further evaluate the robustness of gpt-oss-safeguard-120b and gpt-oss-safeguard-20b to jailbreaks: adversarial prompts that purposely try to circumvent model refusals for content it's not supposed to produce. We evaluate using the following approach:

- StrongReject [3]: inserts a known jailbreak into an example from the above safety refusal eval. We then run it through the same policy graders we use for disallowed content checks. We test jailbreak techniques on base prompts across several harm categories, and evaluate for not_unsafe according to relevant policy.

We find that gpt-oss-safeguard-120b outperforms gpt-oss-120b and gpt-oss-safeguard-20b underperforms gpt-oss-20b by 1-5 points.

Table 6: Jailbreak evaluations

Category	**gpt-oss-120b**	**gpt-oss-safeguard-120b**	**gpt-oss-20b**	**gpt-oss-safeguard-20b**
illicit/non-violent-crime prompts	0.979	0.982	0.960	0.916
violence prompts	0.983	0.989	0.979	0.942
abuse/disinformation/hate prompts	0.993	0.993	0.982	0.959
sexual-content prompts	0.989	0.991	0.970	0.963

4.3 Instruction Hierarchy

Since we don't recommend using either gpt-oss-safeguard model as the core functionality with which end users interact, adherence to an Instruction Hierarchy [4] is less important for these models. However, we post-trained the models with our harmony prompt format that uses several roles including: system messages, developer messages, and user messages. We collected examples of these different roles of messages conflicting with each other, and supervised gpt-oss-safeguard to follow the instructions in the system message over developer messages, and instructions in developer messages over user messages. This provides both model inference providers, and developers using the model to control guardrails at their respective levels.

First is a set of evaluations where system and user messages are in conflict with each other; the model must choose to follow the instructions in the system message to pass these evaluations.

- **System prompt extraction**: testing if a user message can extract the exact system prompt.
- **Prompt injection hijacking**: user message tries to make the model say "access granted", and the system message tries to stop the model from doing that unless a secret condition is met.

Table 7: Instruction Hierarchy Evaluation - System <> User message conflict

Evaluation (higher is better)	gpt-oss-120b	gpt-oss-safeguard-120b	gpt-oss-20b	gpt-oss-safeguard-20b
System prompt extraction	0.832	0.993	0.881	0.867
Prompt injection hijacking	0.780	0.728	0.639	0.512

In the other set of evaluations, we instruct the model to not output a certain phrase (e.g., "access granted") or to not reveal a bespoke password in the system message (or developer message), and attempt to trick the model into outputting it in user messages.

Table 8: Instruction Hierarchy Evaluation - Phrase and Password Protection

Evaluation (higher is better)	gpt-oss-120b	gpt-oss-safeguard-120b	gpt-oss-20b	gpt-oss-safeguard-120b
Phrase protection - system message/user message	0.912	0.807	0.793	0.642
Password protection - system message/user message	0.965	1.000	0.947	0.930
Phrase protection - developer message/user message	0.909	0.789	0.661	0.439
Password protection - developer message/user message	1.000	0.991	0.946	0.921

We observe that the gpt-oss-safeguard models tend to underperform their gpt-oss counterparts. More research is needed to understand why this is the case.

4.4 Hallucinated chains of thought

As with our gpt-oss models, we did not put any direct optimization pressure on the CoT for either of the gpt-oss-safeguard models. We believe that understanding how these models reason about policy classifications is critical to leveraging these models effectively, in addition to the position paper we joined with a number of other labs arguing that frontier developers should "consider the impact of development decisions on CoT monitorability."

Because these chains of thought are not restricted, they can contain hallucinated content, including language that does not reflect OpenAI's standard safety policies or the policy gpt-oss-safeguard has been asked to interpret.

4.5 Hallucinations

We check for hallucinations in gpt-oss-safeguard-120b and gpt-oss-safeguard-20b using the following evaluations, both of which were run without giving the models the ability to browse the internet:

- SimpleQA: A diverse dataset of four thousand fact-seeking questions with short answers

that measures model accuracy for attempted answers.

- PersonQA: A dataset of questions and publicly available facts about people that measures the model's accuracy on attempted answers.

We consider two metrics: accuracy (did the model answer the question correctly) and hallucination rate (did the model answer the question incorrectly). Higher is better for accuracy and lower is better for hallucination rate.

Table 9: Hallucination evaluations

Eval	Metric	gpt-oss-120b	gpt-oss–safeguard-120b	gpt-oss-20b	gpt-oss-safeguard-20b
SimpleQA	accuracy	0.168	0.142	0.067	0.072
	hallucination rate	0.782	0.822	0.914	0.896
PersonQA	accuracy	0.298	0.297	0.155	0.142
	hallucination rate	0.491	0.511	0.532	0.558

The gpt-oss-safeguard models tend to perform on par with their gpt-oss counterparts on both our SimpleQA and PersonQA evaluations. gpt-oss-safeguard-120b is slightly more prone to hallucinating than gpt-oss-120b on both evaluations, and gpt-oss-safeguard-20b is more prone to hallucinating on PersonQA but less so on SimpleQA when compared to gpt-oss-20b.

4.6 Fairness and Bias

We evaluated gpt-oss-safeguard-120b and gpt-oss-safeguard-20b on the BBQ evaluation [5]. Overall, we see both models outperform their gpt-oss counterparts across all metrics.

Table 10: BBQ evaluation

Metric (higher is better)	gpt-oss-120b	gpt-oss-safeguard-120b	gpt-oss-20b	gpt-oss-safeguard-20b
Accuracy on ambiguous questions	0.87	0.93	0.79	0.91
Accuracy on disambiguated questions	0.90	0.95	0.89	0.93

References

[1] OpenAI, "gpt-oss-120b gpt-oss-20b model card." https://arxiv.org/abs/2508.10925, 2025.

[2] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[3] A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, *et al.*, "A strongreject for empty jailbreaks," *arXiv preprint arXiv:2402.10260*, 2024.

[4] E. Wallace, K. Xiao, R. Leike, L. Weng, J. Heidecke, and A. Beutel, "The instruction hierarchy: Training llms to prioritize privileged instructions." https://arxiv.org/abs/2404.13208, 2024.

[5] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. R. Bowman, "BBQ: A hand-built bias benchmark for question answering," *arXiv preprint arXiv:2110.08193*, 2021.