Scores for OpenAI on the 2023 Foundation Model Transparency Index

Background

1. Please see the paper describing the Foundation Model Transparency Index in order to understand what this document includes. The paper provides necessary background on (i) what these indicators are and why they were chosen, (ii) our standardized process for scoring the transparency of foundation model developers, and (iii) what these scores mean in context.
2. This document contains only information that was publicly available before September 15, 2023. It has not been updated and should be interpreted as a snapshot of transparency as of September 15, 2023.
3. In order to assess the transparency of foundation model developers, we used a rigorous, standardized search protocol to find publicly available information related to these indicators. You can find more information about this search protocol in the paper describing the Foundation Model Transparency Index.
4. We evaluate every company in this same way–you can find scoring documents for the other companies here.
5. We evaluate each company on 100 indicators of transparency. You can find the definition of each indicator and additional information about how each indicator was scored here.
6. Scores for each indicator are either 0 or 1. If the score is a 0, we do not provide a source for the score because our standardized search protocol (which includes many relevant sources) did not yield enough information to award a point. If the score is a 1, we provide a source that includes the information we cite in the justification for the score.
7. We evaluate each company on the basis of its flagship foundation model; in the case of OpenAI, we evaluate GPT-4.
8. In advance of releasing the Foundation Model Transparency Index, we reached out to OpenAI for comment (along with the 9 other companies we evaluated) and offered an opportunity to provide feedback on the index and the organization's scores.

Scores for Each Indicator

1. Upstream → Data → Data Size
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

2. Upstream → Data → Data Sources
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

3. Upstream → Data → Data Creators
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

4. Upstream → Data → Data Source Selection
    - Score: 0
    - Justification: Not information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

5. Upstream → Data → Data Curation

- Score: 1
- Justification: OpenAI provides sufficient information on data curation in section 3.1 of the system card.
- Source: https://arxiv.org/abs/2303.08774

6. Upstream → Data → Data Augmentation
   - Score: 0
   - Justification: Data description does not make clear whether or not augmentation is performed, given many details on the data are withheld as disclosed in "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details … dataset construction …"
   - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

7. Upstream → Data → Harmful Data Filtration
   - Score: 1
   - Justification: As stated in the system card, "At the pre-training stage, we filtered our dataset mix for GPT-4 to specifically reduce the quantity of inappropriate erotic text content. We did this via a combination of internally trained classifiers[37] and a lexicon-based approach to identify documents that were flagged as having a high likelihood of containing inappropriate erotic content. We then removed these documents from the pre-training set. …."
   - Source: https://arxiv.org/abs/2303.08774

8. Upstream → Data → Copyrighted data
   - Score: 0
   - Justification: No information found related to copyrighted data.
   - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

9. Upstream → Data → Data License
   - Score: 0
   - Justification: No decomposition of the data, though the GPT-4 technical report does say "GPT-4 is a Transformer-style model pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers."
   - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

10. Upstream → Data → Personal Information in Data

- Score: 0
- Justification: No decomposition of the data, though a generic sentence overall in the technical report: “GPT-4 has learned from a variety of licensed, created, and publicly available data sources, which may include publicly available personal information.”
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

11. Upstream → Data Labor → Use of Human Labor
    - Score: 0
    - Justification: OpenAI does describe the use of human labor, but it appears clear it is incomplete due to their discussion of red teaming, manual data collection, and so forth, where labor is not clearly disclosed: “We collect demonstration data (given an input, demonstrating how the model should respond) and ranking data on outputs from our models (given an input and several outputs, rank the outputs from best to worst) from human trainers.”
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

12. Upstream → Data Labor → Employment of Data Laborers
    - Score: 0
    - Justification: While the involvement of human labor is disclosed, the employer of the data laborers is not disclosed.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

13. Upstream → Data Labor → Geographic Distribution of Data Laborers
    - Score: 0
    - Justification: No information found related to geographic distribution of data laborers.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

14. Upstream → Data Labor → Wages
    - Score: 0
    - Justification: No information found related to wages.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

15. Upstream → Data Labor → Instructions For Creating Data

- Score: 0
- Justification: No information found related to instructions for creating data.
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

16. Upstream → Data Labor → Labor Protections
    - Score: 1
    - Justification: Footnote 28 in GPT-4 technical report reads "With all workers, we follow industry-best practices by ensuring every annotator retains the right to opt out of any task they find unpleasant, receive a market wage commensurate with the work they deliver, and have opportunities and channels through which they can discuss their work and raise objections. We generally implement two distinct sets of guidelines tailored to whether our annotators work with sensitive or unwanted content. For non-sensitive annotation, we have built technical features (in part with OpenAI's moderation endpoint) into our data pipeline to filter our sensitive content. For sensitive content annotation, we use vendor-provided features like mandated breaks, blurring or grayscale of materials, and clearly delineated project categories such that no contractor is surprised by the nature of the material. Additionally, for vendor-managed workers, we have implemented ongoing workers' wellness surveys and support procedures that we regularly discuss with our vendors"
    - Source: https://arxiv.org/abs/2303.08774

17. Upstream → Data Labor → Third Party Partners
    - Score: 0
    - Justification: Insufficiently clear in light of "GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

18. Upstream → Data Access → Queryable External Data Access
    - Score: 0
    - Justification: No information found
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

19. Upstream → Data Access → Direct External Data Access

- Score: 0
- Justification: No information found
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

20. Upstream → Compute → Compute Usage
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

21. Upstream → Compute → Development Duration
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

22. Upstream → Compute → Compute Hardware
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

23. Upstream → Compute → Hardware Owner
    - Score: 1
    - Justification: OpenAI discloses that GPT-4 was trained on Microsoft Azure supercomputers.
    - Source: https://arxiv.org/abs/2303.08774

24. Upstream → Compute → Energy Usage

- Score: 0
- Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

25. Upstream → Compute → Carbon Emissions
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

26. Upstream → Compute → Broader Environmental Impact
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

27. Upstream → Methods → Model Stages
    - Score: 1
    - Justification: Pretraining, RLHF, and further model-based steering are described in Section 3.1 of System Card.
    - Source: https://arxiv.org/abs/2303.08774

28. Upstream → Methods → Model Objectives
    - Score: 1
    - Justification: OpenAI provides some characterization of the purpose of different stages in the system card, especially in Section 3.
    - Source: https://arxiv.org/abs/2303.08774

29. Upstream → Methods → Core Frameworks

- Score: 0
- Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

30. Upstream → Methods → Additional Dependencies
    - Score: 0
    - Justification: Not clear based on model description in paper/system card.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

31. Upstream → Data Mitigations → Mitigations for Personally Identifiable Information
    - Score: 1
    - Justification: "We take a number of steps to reduce the risk that our models are used in a way that could violate a person's privacy rights. These include fine-tuning models to reject these types of requests, removing personal information from the training dataset where feasible, creating automated model evaluations, monitoring and responding to user attempts to generate this type of information, and restricting this type of use in our terms and policies. Our efforts to expand context length and improve embedding models for retrieval may help further limit privacy risks moving forward by tying task performance more to the information a user brings to the model. We continue to research, develop, and enhance technical and process mitigations in this area."
    - Source: https://arxiv.org/abs/2303.08774

32. Upstream → Data Mitigations → Mitigations for Copyright
    - Score: 0
    - Justification: No information found
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

33. Model → Model Basics → Input Modality
    - Score: 1
    - Justification: The input modality is text and images. The GPT-4 technical report says "a large multimodal model capable of processing image and text inputs and producing text outputs:"
    - Source: https://arxiv.org/abs/2303.08774
34. Model → Model Basics → Output Modality

- Score: 1
- Justification: The output modality is text. The GPT-4 technical report says "a large multimodal model capable of processing image and text inputs and producing text outputs:"
- Source: https://arxiv.org/abs/2303.08774

35. Model → Model Basics → Model Components
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

36. Model → Model Basics → Model Size
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

37. Model → Model Basics → Model Architecture
    - Score: 0
    - Justification: No information as indicated by "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

38. Model → Model Basics → Centralized Model Documentation
    - Score: 1
    - Justification: The technical report and system card centralize key information about the model.
    - Source: https://arxiv.org/abs/2303.08774

39. Model → Model Access → External Model Access Protocol

- Score: 1
- Justification: OpenAI provides a clear research access program where researchers and other external entities can request access, with some discussion of criteria used to make decisions. The time frame for a decision of 4-6 weeks is clearly disclosed.
- Source: https://web.archive.org/web/20230717232903/https://openai.com/form/researcher-access-program

40. Model → Model Access → Black Box External Model Access
    - Score: 1
    - Justification: Black box access is provided via the OpenAI API to GPT-4.
    - Source: https://web.archive.org/web/20230906044201/https://platform.openai.com/docs/models/gpt-4 ;

41. Model → Model Access → Full External Model Access
    - Score: 0
    - Justification: Weights are not made available.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

42. Model → Capabilities → Capabilities Description
    - Score: 1
    - Justification: Many capabilities described in the GPT-4 product page such as advanced reasoning capabilities, more useful responses, general knowledge and problem solving, and so forth.
    - Source: https://web.archive.org/web/20230912150319/https://openai.com/gpt-4

43. Model → Capabilities → Capabilities Demonstration
    - Score: 1
    - Justification: Many capabilities demonstrations on the GPT-4 product page.
    - Source: https://web.archive.org/web/20230912150319/https://openai.com/gpt-4

44. Model → Capabilities → Evaluation of Capabilities
    - Score: 1
    - Justification: Many capabilities evaluations on standard public benchmarks (e.g. MMLU, HellaSwag) in Section 4 of the GPT-4 technical report
    - Source: https://arxiv.org/abs/2303.08774

45. Model → Capabilities → External Reproducibility of Capabilities Evaluation

- Score: 1
- Justification: Many evals on standard public benchmarks (e.g. MMLU, HellaSwag) in Section 4; new evals on exams include some methodology in Appendix A of paper. While some details of sourcing exams are unclear/incomplete, the public benchmarks are assumed to be sufficiently reproducible.
- Source: https://arxiv.org/abs/2303.08774

46. Model → Capabilities → Third Party Capabilities Evaluation
    - Score: 1
    - Justification: OpenAI provides evaluations conducted by "qualified third-party contractors" for their evaluations on AP Exams spanning a range of subjects as well as for the Uniform Bar Exam by “CaseText and Stanford CodeX”.
    - Source: https://arxiv.org/abs/2303.08774

47. Model → Limitations → Limitations Description
    - Score: 1
    - Justification: According to the GPT-4 technical report, “it is not fully reliable (e.g. can suffer from “hallucinations”), has a limited context window, and does not learn from experience. Care should be taken when using the outputs of GPT-4, particularly in contexts where reliability is important."
    - Source: https://arxiv.org/abs/2303.08774

48. Model → Limitations → Limitations Demonstration
    - Score: 0
    - Justification: The GPT-4 system card extensively describes risks and limitations. However, these limitations are not clearly demonstrated as required for this indicator.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

49. Model → Limitations → Third Party Evaluation of Limitations
    - Score: 1
    - Justification: Limitations can be studied via black box model access; no restrictions on usage in usage policy that constrain this.
    - Source:

50. Model → Risks → Risks Description

- Score: 1
- Justification: Both intentional and unintentional harms are described in system card as seen in this quote: "GPT-4 can generate potentially harmful content, such as advice on planning attacks or hate speech. It can represent various societal biases and worldviews that may not be representative of the users intent,4 or of widely shared values. It can also generate code that is compromised or vulnerable. The additional capabilities of GPT-4 also lead to new risk surfaces."
- Source: https://arxiv.org/abs/2303.08774

51. Model → Risks → Risks Demonstration
    - Score: 1
    - Justification: Both intentional and unintentional harms are demonstrated in the system card (e.g. Figure 1).
    - Source: https://arxiv.org/abs/2303.08774

52. Model → Risks → Unintentional Harm Evaluation
    - Score: 0
    - Justification: There is some evaluation (e.g. of model behavior on sensitive content in Figure 7 of the system card), but it is not sufficiently rigorous (no substantive details on underlying prompts besides “e.g. regulated medical advice”).
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

53. Model → Risks → External Reproducibility of Unintentional Harm Evaluation
    - Score: 0
    - Justification: Evaluations are not reproducible: for example, prompts involved in evaluation in Figure 7 of the system card are not described with sufficient precision to be reproducible.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

54. Model → Risks → Intentional Harm Evaluation
    - Score: 0
    - Justification: There is some evaluation (e.g. Appendix E of system card), but it is neither quantitative nor sufficiently rigorous.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

55. Model → Risks → External Reproducibility of Intentional Harm Evaluation

- Score: 1
- Justification: The evaluations in Appendix E are reproducible (prompts are provided in Figure 10).
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

56. Model → Risks → Third Party Risk Evaluation
    - Score: 1
    - Justification: There is an extensive evaluation by the Alignment Research Center that is discussed in the system card, pointing to more extensive information on the ARC website.
    - Source: https://arxiv.org/abs/2303.08774

57. Model → Model Mitigations → Mitigations Description
    - Score: 1
    - Justification: GPT-4 technical report reads "Recommendations and training data gathered from these experts fed into our mitigations and improvements for the model; for example, we've collected additional data to improve GPT-4's ability to refuse requests on how to synthesize dangerous chemicals (Table 5). Model-Assisted Safety Pipeline: As with prior GPT models, we fine-tune the model's behavior using reinforcement learning with human feedback (RLHF) to produce responses better aligned with the user's intent. … To steer our models towards appropriate behaviour at a more fine-grained level, we rely heavily on our models themselves as tools. Our approach to safety consists of two main components, an additional set of safety-relevant RLHF training prompts, and rule-based reward models (RBRMs). Our rule-based reward models (RBRMs) are a set of zero-shot GPT-4 classifiers. These classifiers provide an additional reward signal to the GPT-4 policy model during RLHF fine-tuning that targets correct behavior, such as refusing to generate harmful content or not refusing innocuous requests."
    - Source: https://arxiv.org/abs/2303.08774

58. Model → Model Mitigations → Mitigations Demonstration
    - Score: 1
    - Justification: In section 6 of the body of the technical report, three examples of disallowed prompts related to synthesizing a dangerous chemical, making a bomb, and finding cheap cigarettes. In the system card, section 2 provides a clear demonstration of the majority of mitigations.
    - Source: https://arxiv.org/abs/2303.08774

59. Model → Model Mitigations → Mitigations Evaluation
    - Score: 1
    - Justification: RBRM and RLHF both evaluated quantitatively in section 3 of the system card.
    - Source: https://arxiv.org/abs/2303.08774

60. Model → Model Mitigations → External Reproducibility of Mitigations Evaluation
    - Score: 0
    - Justification: No information found related to the data associated with mitigations evaluations.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

61. Model → Model Mitigations → Third Party Mitigations Evaluation
    - Score: 0
    - Justification: No information found to indicate mitigations can be evaluated by third parties; neither RBRM mitigation nor RLHF mitigation can be evaluated by third parties as there is no access to GPT-4 early.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

62. Model → Trustworthiness → Trustworthiness Evaluation
    - Score: 1
    - Justification: Calibration is evaluated in section 6 of technical report
    - Source: https://arxiv.org/abs/2303.08774

63. Model → Trustworthiness → External Reproducibility of Trustworthiness Evaluation
    - Score: 0
    - Justification: No information found related to the data associated with the calibration evaluation.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

64. Model → Inference → Inference Duration Evaluation
    - Score: 0
    - Justification: No information found related to an evaluation of inference duration.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

65. Model → Inference → Inference Compute Evaluation
    - Score: 0
    - Justification: No information found related to an evaluation of inference compute
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

66. Downstream → Distribution → Release Decision-Making
    - Score: 0
    - Justification: Some information related to release is described, but it does not clearly articulate the decision-making behind how GPT-4 was released: "OpenAI has been iterating[21] on GPT-4 and our deployment plan since early August to prepare for a safer launch. We believe this has reduced the risk surface, though has not completely eliminated it. Today's deployment represents a balance between minimizing risk from deployment, enabling positive use cases, and learning from deployment. Our work during the period consisted of the
    - following interrelated steps…"
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

67. Downstream → Distribution → Release Process
    - Score: 1
    - Justification: OpenAI involved 50 external researchers prior to the core public release, using their feedback to improve the model (e.g. add data to improve refusal rate for requests around synthesizing dangerous chemicals).
    - Source: https://arxiv.org/abs/2303.08774

68. Downstream → Distribution → Distribution Channels
    - Score: 1
    - Justification: Distribution channels for GPT-4 include ChatGPT Plus, Playground, and OpenAI's API.
    - Source: https://web.archive.org/web/20230901133254/https://openai.com/research/gpt-4

69. Downstream → Distribution → Products and Services
    - Score: 1
    - Justification: OpenAI discloses GPT-4 powers ChatGPT Plus.
    - Source: https://web.archive.org/web/20230901133254/https://openai.com/research/gpt-4

70. Downstream → Distribution → Detection of Machine-Generated Content
    - Score: 0
    - Justification: No information found about a mechanism to detect machine-generated content for GPT-4.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

71. Downstream → Distribution → Model License
    - Score: 0
    - Justification: No information found related to a license for GPT-4.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

72. Downstream → Distribution → Terms of Service
    - Score: 1
    - Justification: OpenAI has a broad Terms-of-Service that applies to its API, playground, and ChatGPT Plus.
    - Source: https://web.archive.org/web/20230825000505/https://openai.com/policies/service-terms

73. Downstream → Usage Policy → Permitted, Restricted, and Prohibited Users
    - Score: 1
    - Justification: Supported list of countries describes permitted geographies and implies restricted geographies.
    - Source: https://web.archive.org/web/20230825000505/https://openai.com/policies/service-terms

74. Downstream → Usage Policy → Permitted, Restricted, and Prohibited Uses
    - Score: 1
    - Justification: Usage policy lays out many disallowed uses as well as “further requirements for certain uses”
    - Source: https://web.archive.org/web/20230912150420/https://openai.com/policies/usage-policies

75. Downstream → Usage Policy → Usage Policy Enforcement
    - Score: 1
    - Justification: From the GPT-4 system card, “We use a mix of reviewers and automated systems to identify and enforce against misuse of our models. Our automated systems include a suite of machine learning and rule-based classifier detections that identify content that might violate our policies. When a user repeatedly prompts our models with

policy-violating content, we take actions such as issuing a warning, temporarily suspending, or in severe cases, banning the user. Our reviewers ensure that our classifiers are correctly blocking violative content and understand how users are interacting with our systems. These systems also create signals that we use to mitigate abusive and inauthentic behavior on our platform. We investigate anomalies in API traffic to learn about new types of abuse and to improve our policies and enforcement."
    - Source: https://arxiv.org/abs/2303.08774

76. Downstream → Usage Policy → Justification for Enforcement Action
    - Score: 0
    - Justification: No information found regarding justification for enforcement actions.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

77. Downstream → Usage Policy → Usage Policy Violation Appeals Mechanism
    - Score: 1
    - Justification: Clear violations of the "content policy" cause the ChatGPT UI to bring up a mechanism for appeal
    - Source: http://web.archive.org/web/20230912151047/https://docs.google.com/forms

78. Downstream → Model Behavior Policy → Permitted, Restricted, and Prohibited Model Behaviors
    - Score: 1
    - Justification: Model behavior guidelines lay out such restrictions. Also GPT-4 technical report discusses a large amount of work on model refusal and says the aim of this work was "making it more stringent in rejecting requests that go against our content policy, while being more open to requests it can safely fulfill." Examples include refusing malicious cybersecurity requests. E.g., "Our work on model refusals (described in Section 2) aimed to reduce the tendency of the model to produce such harmful content. Below we provide some examples from GPT-4-early compared to GPT-4-launch, the version we are launching with"; "To mitigate potential misuses in this area, we have trained models to refuse malicious cybersecurity requests, and scaled our internal safety systems, including in monitoring, detection and response"; "To tackle overreliance, we've refined the model's refusal behavior, making it more stringent in rejecting requests that go against our content policy, while being more open to requests it can safely fulfill. One objective here is to discourage users from disregarding the model's refusals."
    - Source: http://web.archive.org/web/20230912151036/https://cdn.openai.com/snapshot-of-chatgpt-model-behavior-guidelines.pdf and https://arxiv.org/pdf/2303.08774.pdf

79. Downstream → Model Behavior Policy → Model Behavior Policy Enforcement
    - Score: 0
    - Justification: The GPT-4 system card clarifies how it enforces its usage policy, but not how it enforces its model behavior policy (e.g. how it specifically translates its model behavior policy into refusals, how it tweaks future versions in response to issues that surface).
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

80. Downstream → Model Behavior Policy → Interoperability of Usage and Model Behavior Policies
    - Score: 1
    - Justification: GPT-4 system card reads "We reduced the prevalence of certain kinds of content that violate our usage policies (such as inappropriate erotic content) in our pre-training dataset, and fine-tuned the model to refuse certain instructions such as direct requests for illicit advice."
      Source: https://arxiv.org/pdf/2303.08774.pdf

81. Downstream → User Interface → User Interaction with AI System
    - Score: 1
    - Justification: ChatGPT Plus references at the top of the page that the user is interacting with GPT-4 and at the bottom of the page links to the release notes with details about the specific version.
    - Source: https://chat.openai.com/ ; https://help.openai.com/en/articles/6825453-chatgpt-release-notes

82. Downstream → User Interface → Usage Disclaimers
    - Score: 1
    - Justification: Signing up for ChatGPT points to TOS which in turn points to Usage Policies
    - Source: https://platform.openai.com/onboarding?app=chat

83. Downstream → User Data Protection → User Data Protection Policy
    - Score: 1
    - Justification: Privacy policy discloses protocols for storing, accessing, and sharing user data
    - Source: https://web.archive.org/web/20230912150325/https://openai.com/policies/privacy-policy

84. Downstream → User Data Protection → Permitted and Prohibited Use of User Data
    - Score: 1
    - Justification: Privacy policy and API data privacy statement disclose permitted and prohibited uses of user data
    - Source: https://openai.com/api-data-privacy and https://openai.com/policies/privacy-policy and https://web.archive.org/web/20230912150303/https://openai.com/enterprise-privacy

85. Downstream → User Data Protection → Usage Data Access Protocol
    - Score: 0
    - Justification: No information found related to usage data access protocol
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

86. Downstream → Model Updates → Versioning Protocol
    - Score: 1
    - Justification: Model upgrades page says "With the release of gpt-3.5-turbo, some of our models are now being continually updated. gpt-3.5-turbo, gpt-4, and gpt-4-32k point to the latest model version. You can verify this by looking at the response object after sending a ChatCompletion request. The response will include the specific model version used (e.g. gpt-3.5-turbo-0613). We also offer static model versions that developers can continue using for at least three months after an updated model has been introduced."
    - Source: https://web.archive.org/web/20230913184747/https://platform.openai.com/docs/models/continuous-model-upgrades and https://platform.openai.com

87. Downstream → Model Updates → Change Log
    - Score: 1
    - Justification: OpenAI's deprecations page includes a changelog for GPT-4.
    - Source: https://platform.openai.com/docs/deprecations

88. Downstream → Model Updates → Deprecation Policy
    - Score: 1
    - Justification: OpenAI's downstream documentation includes a page on deprecations with instructions on migrating to newer versions.
    - Source: https://platform.openai.com/docs/deprecations

89. Downstream → Feedback → Feedback Mechanism
    - Score: 1
    - Justification: OpenAI has a "Chat Model Feedback form" tied to ChatGPT Plus as a distribution channel for GPT-4.
    - Source: https://web.archive.org/web/20230912153627/http://web.archive.org/screenshot/https://openai.com/form/chat-model-feedback

90. Downstream → Feedback → Feedback Summary
    - Score: 0
    - Justification: No information found related to any summary of user feedback. This is distinct from the use of human feedback in building models, such as through reinforcement learning from human feedback and other forms of pre-deployment red-teaming.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

91. Downstream → Feedback→ Government Inquiries
    - Score: 0
    - Justification: No information found related to a summary of government inquiries.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

92. Downstream → Impact → Monitoring Mechanism
    - Score: 1
    - Justification: OpenAI discloses that it monitors the use of its model, including anomalies in API traffic. "We use a mix of reviewers and automated systems to identify and enforce against misuse of our models. Our automated systems include a suite of machine learning and rule-based classifier detections that identify content that might violate our policies. When a user repeatedly prompts our models with policy-violating content, we take actions such as issuing a warning, temporarily suspending, or in severe cases, banning the user. Our reviewers ensure that our classifiers are correctly blocking violative content and understand how users are interacting with our systems. These systems also create signals that we use to mitigate abusive and inauthentic behavior on our platform. We investigate anomalies in API traffic to learn about new types of abuse and to improve our policies and enforcement."
    - Source: https://arxiv.org/abs/2303.08774

93. Downstream → Impact → Downstream Applications
    - Score: 0
    - Justification: No information found related to the number of applications dependent on the foundation model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

94. Downstream → Impact → Affected Market Sectors
    - Score: 0
    - Justification: No information found related to the fraction of applications corresponding to each market sector.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

95. Downstream → Impact → Affected Individuals
    - Score: 0
    - Justification: No information found related to the number of individuals affected by the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

96. Downstream → Impact → Usage Reports
    - Score: 0
    - Justification: No information found related to usage statistics describing the impact of the model on users.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

97. Downstream → Impact → Geographic Statistics
    - Score: 0
    - Justification: No information found regarding statistics of model usage across geographies.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

98. Downstream → Impact → Redress Mechanism
    - Score: 0
    - Justification: No information found regarding any mechanism to provide redress to users for harm caused by the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

99. Downstream → Documentation for Deployers → Centralized Documentation for Downstream Use
    - Score: 1
    - Justification: Several such artifacts make for centralized documentation, such as OpenAI's API documentation and its centralized page for policies.
    - Source: https://web.archive.org/web/20230912150310/https://openai.com/policies and https://web.archive.org/web/20230913184451/https://platform.openai.com/docs/api-reference/introduction

100. Downstream → Documentation for Deployers → Documentation for Responsible Downstream Use
    - Score: 1
    - Justification: Safety best practices page in the API documentation provides documentation for responsible downstream use.
    - Source: https://web.archive.org/web/20230912150452/https://platform.openai.com/docs/guides/safety-best-practices