Scores for Meta 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 Meta, we evaluate Llama 2.
8. In advance of releasing the Foundation Model Transparency Index, we reached out to Meta 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: 1
   - Justification: Meta says it trained Llama 2 on 2 trillion tokens.
   - Source: https://arxiv.org/pdf/2307.09288.pdf.

2. Upstream → Data → Data Sources
   - Score: 0
   - Justification: Data sources are not disclosed beyond "Our training corpus includes a new mix of data from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals."
   - 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: Data creators are not characterized in the technical report (note: in contrast, data about how people are mentioned in the data used to the model is provided, for example in Table 9 on demography).
   - 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: No information found related to selection protocols for including and excluding data sources disclosed.
   - 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: Data is filtered to remove PII, but otherwise Section 4 of the technical report explicitly indicates no further filtration is done.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

6. Upstream → Data → Data Augmentation
    - Score: 1
    - Justification: Data augmentation is clearly discussed in the description of the Ghost Attention method in section 3.3 the technical report.
    - Source: https://arxiv.org/pdf/2307.09288.pdf

7. Upstream → Data → Harmful Data Filtration
    - Score: 1
    - Justification: Section 4.1 explicitly indicates no such filtration is done deliberately: "No additional filtering was conducted … to allow Llama 2 to be more widely usable across tasks … while avoiding the potential for the accidental demographic erasure sometimes caused by over-scrubbing".
    - Source: https://arxiv.org/pdf/2307.09288.pdf

8. Upstream → Data → Copyrighted data
    - Score: 0
    - Justification: No information found regarding the copyright status of the data used to build Llama 2.
    - 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 information found regarding the license status of the data used to build Llama 2.
    - 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: Meta discloses that it deliberately avoids sources that are likely to contain high volumes of personal information. However, in the chosen data, the presence of personal information is not characterized.
    - 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: 1
    - Justification: The use of human labor is discussed extensively in the technical report, both in relation to the fine-tuning stage, and the red teaming stage, and evaluations of safety.
    - Source: https://arxiv.org/pdf/2307.09288.pdf

12. Upstream → Data Labor → Employment of Data Laborers
    - Score: 0
    - Justification: No information found about the organization that directly employs the people involved in data labor.
    - 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 about 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 about the wages for people who perform data labor.
    - 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: 1
    - Justification: Instructions given for red teaming and SFT are disclosed in the technical report. “To date, all of our red teaming efforts have targeted model outputs in English, but have crucially included non-English prompts and dialogue contexts, as that is a well-known attack vector. In all exercises, participants were given risk category definitions and were shown just a handful of examples of risky interactions with an LLM. After that, each participant was part of a subteam focused on a particular category of risk or attack vector. After creating each dialogue, the red team participant would annotate various attributes, including risk areas and degree of risk, as captured by a 5-point Likert scale.” pp 29 in the Llama 2 technical report; “The annotators are instructed to initially come up with prompts that they think could potentially induce the model to exhibit unsafe behavior, i.e., perform red teaming, as defined by the guidelines. Subsequently, annotators are tasked with crafting a safe and helpful response that the model should produce” pp 24.
    - Source: https://arxiv.org/pdf/2307.09288.pdf

16. Upstream → Data Labor → Labor Protections
    - Score: 0
    - Justification: No information found about labor protections for data laborers.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

17. Upstream → Data Labor → Third Party Partners
    - Score: 0
    - Justification: Contract workers and external vendors are explicitly mentioned in the context of red teaming in section 4.3 of the technical report, but no information is provided on these individuals. Including Meta employees they are collectively described as "350 people, including domain experts in cybersecurity, election fraud, social media misinformation, legal, policy, civil rights, ethics, software engineering, machine learning, responsible AI, and creative writing. They also included individuals representative of a variety of socioeconomic, gender, ethnicity, and racial demographics" but these third party partners are not described in more detail.
    - 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 about whether external entities are provided with queryable access to the data used to build LLAMA2.
   - 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 about whether external entities are provided with direct access to the data used to build Llama 2.
   - 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: The amount of computation performed is not reported (e.g. in FLOPs).
   - 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: 1
   - Justification: The duration in GPU hours is a cumulative 3.3M hours on Nvidia A100s as reported in the model card.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

22. Upstream → Compute → Compute Hardware
   - Score: 0
   - Justification: The type of hardware is disclosed to be Nvidia A100s but the number of hardware units is not disclosed.
   - 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: The hardware is disclosed to be from the Meta research cluster and additional Meta-internal production clusters.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

24. Upstream → Compute → Energy Usage

- Score: 1
- Justification: Meta reports the Thermal Design Power of the NVIDIA A100 GPUs they train (350 - 400 mW) and the number of A100 GPU hours (3.3 million hours). From this, the (maximal) energy usage in mWh can be computed to at least 1 significant figure as required by this indicator as 1 * 10^9.
- Source: https://arxiv.org/pdf/2307.09288.pdf

25. Upstream → Compute → Carbon Emissions

- Score: 1
- Justification: The carbon emissions are disclosed as 539tC02.
- Source: https://arxiv.org/pdf/2307.09288.pdf

26. Upstream → Compute → Broader Environmental Impact

- Score: 0
- Justification: No information found any broader environmental impacts from building the model, though the use of carbon offsets is discussed in the Llama 2 technical report.
- 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: The stages of the model pipeline are clearly described in Sections 2-4 of the technical report.
- Source:

28. Upstream → Methods → Model Objectives

- Score: 1
- Justification: The model objectives are fairly clear: pretraining (next word prediction as described in LLaMA 1), fine-tuning use an autoregressive objective only for answer tokens, reward modeling describes details of RLHF.
- Source: https://arxiv.org/pdf/2307.09288.pdf

29. Upstream → Methods → Core Frameworks
- Score: 0
- Justification: The code for inference using Llama 2 is released, which discloses the dependencies and frameworks (e.g. PyTorch), but the code for building the model is not provided. In addition, the paper discusses some aspects but the core frameworks are not disclosed (there is a mention of using FSDP and citation of the PyTorch implementation of FSDP, but this is not made fully clear.)
- 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: 1
- Justification: Given the thoroughness of describing how the model is built with all fine-grained stages, objectives and training details disclosed, it is reasonably clear there are no additional dependencies. This is also indicated by the status in the model card: "This is a static model trained on an offline dataset."
- Source: https://arxiv.org/pdf/2307.09288.pdf

31. Upstream → Data Mitigations → Mitigations for Personally Identifiable Information
- Score: 1
- Justification: During data sourcing, sources are avoided that are known to include high volume of personal information about private individuals: "We excluded data from certain sites known to contain a high volume of personal information about private individuals" in section 4.1 of the technical report.
- Source: https://arxiv.org/pdf/2307.09288.pdf

32. Upstream → Data Mitigations → Mitigations for Copyright
- Score: 0
- Justification: No information found about steps Meta takes to mitigate the presence of copyrighted information in the data.
- 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 model card states that the input modality is text.
- Source: https://arxiv.org/pdf/2307.09288.pdf

34. Model → Model Basics → Output Modality
   - Score: 1
   - Justification: The model card states that the output modality is text.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

35. Model → Model Basics → Model Components
   - Score: 1
   - Justification: That the model is a single component based on the Transformer architecture (e.g. no retrieval module) is made clear from the description of the architecture in the model card and the technical report.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

36. Model → Model Basics → Model Size
   - Score: 1
   - Justification: The model sizes are disclosed, with the largest Llama 2 model being a dense model with 70B parameters. Each Llama 2 model is composed of a single component.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

37. Model → Model Basics → Model Architecture
   - Score: 1
   - Justification: The model card gives a clear description of the model architecture as does the technical report.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

38. Model → Model Basics → Centralized Model Documentation
   - Score: 1
   - Justification: The technical report and model card therein provides centralized documentation.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

39. Model → Model Access → External Model Access Protocol
   - Score: 1
   - Justification: Access requests can be made via a form, which states criteria for granting (e.g. fewer than 700m monthly active users in the previous month) access.
   - Source: http://web.archive.org/web/20230914171744/https://ai.meta.com/resources/models-and-libraries/llama-downloads/

40. Model → Model Access → Black Box External Model Access

- Score: 1
- Justification: The model weights are openly accessible, permitting blackbox querying as well.
- Source: http://web.archive.org/web/20230914171744/https://ai.meta.com/resources/models-and-libraries/llama-downloads/

41. Model → Model Access → Full External Model Access

- Score: 1
- Justification: The model weights are openly accessible.
- Source: http://web.archive.org/web/20230914171744/https://ai.meta.com/resources/models-and-libraries/llama-downloads/

42. Model → Capabilities → Capabilities Description

- Score: 0
- Justification: While many aspects of Llama 2's capabilities are discussed and evaluated, there is no clear description. In the introduction to the Llama 2 paper, capabilities of LLMs generally are described, but the capabilities of Llama 2 are not clearly stated. There is a pointer to section 4.1 of the paper for additional details on capabilities, but section 4.1 does not contain a description; as the section reads, "Benchmarks give a summary view of model capabilities and behaviors that allow us to understand general patterns in the model, but they do not provide a fully comprehensive view of the impact the model may have on people or real-world outcomes" -- benchmarks like those contained in section 4.1 give only a summary view of model capabilities, not a clear description. Meta's Llama 2 Responsible Use Guide says that the model card provides details on capabilities, but it does not include a clear description of capabilities.
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

43. Model → Capabilities → Capabilities Demonstration

- Score: 1
- Justification: Several capabilities are demonstrated in the appendix of the report.
- Source: https://arxiv.org/pdf/2307.09288.pdf.

44. Model → Capabilities → Evaluation of Capabilities

- Score: 1
- Justification: Extensive evaluations of capabilities on standard reproducible benchmarks (e.g. MMLU, HellaSwag, Human-Eval).
- Source: https://arxiv.org/pdf/2307.09288.pdf

45. Model → Capabilities → External Reproducibility of Capabilities Evaluation

- Score: 1
- Justification: Extensive evaluations of capabilities on standard reproducible benchmarks (e.g. MMLU, HellaSwag, Human-Eval). Note hyperparameters and prompting information are provided, which exceeds the standard we impose for this version of the Index.
- Source: https://arxiv.org/pdf/2307.09288.pdf

46. Model → Capabilities → Third Party Capabilities Evaluation

- Score: 0
- Justification: No information found indicating third parties have evaluated Llama 2’s capabilities.
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

47. Model → Limitations → Limitations Description

- Score: 1
- Justification: Limitations are described in 5.2 and in model card: ". Testing conducted to date has been in English, and has not covered, nor could it cover all scenarios. For these reasons, as with all LLMs, Llama 2’s potential outputs cannot be predicted in advance, and the model may in some instances produce inaccurate, biased or other objectionable responses to user prompts. Therefore, before deploying any applications of Llama 2, developers should perform safety testing and tuning tailored to their specific applications of the model."
- Source: https://arxiv.org/pdf/2307.09288.pdf

48. Model → Limitations → Limitations Demonstration

- Score: 0
- Justification: Limitations are not clearly demonstrated, though some examples are provided in the technical report's appendix of various failure modes (for example incorrect refusals).
- 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: Since weights are accessible, third-party evaluation should be possible and no restrictions imposed in license.
- Source: https://arxiv.org/pdf/2307.09288.pdf

50. Model → Risks → Risks Description

- Score: 1
- Justification: Risks are described extensively, spanning both many forms of unintentional harm (e.g. bias, toxicity) and intentional harm (e.g. disinformation). See 4.2.1 for safety categories.
- Source: https://arxiv.org/pdf/2307.09288.pdf

51. Model → Risks → Risks Demonstration

- Score: 1
- Justification: There are a few examples that demonstrate risks throughout the Llama 2 technical report, including for scams (e.g. Table 12, 35) and conspiracy theories (Table 13).
- Source: https://arxiv.org/pdf/2307.09288.pdf

52. Model → Risks → Unintentional Harm Evaluation

- Score: 1
- Justification: Several evaluations of unintentional harms (e.g. ToxiGen, BOLD, TruthfulQA) in paper and in Appendix A.
- Source: https://arxiv.org/pdf/2307.09288.pdf

53. Model → Risks → External Reproducibility of Unintentional Harm Evaluation

- Score: 1
- Justification: Not all evaluations are reproducible, but some are on standard benchmarks (e.g. ToxiGen, BOLD, TruthfulQA).
- Source: https://arxiv.org/pdf/2307.09288.pdf

54. Model → Risks → Intentional Harm Evaluation

- Score: 0
- Justification: Some evaluations of intentional harm, but since all results are presented in terms of safety scores, there is not sufficient disaggregation to clarify if safety scores relate to malicious use or other unintentional forms of harm. For example, if scores had been disaggregated with respect to categories in Table 42, this would clarify the issue.
- 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: 0
   - Justification: There are no clear evaluations of intentional harm, and therefore there are not multiple reproducible evaluations.
   - 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: 0
   - Justification: While red teaming evaluations are conducted by external entities in part, no external assessment of risk is provided.
   - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

57. Model → Mitigations → Mitigations Description
   - Score: 1
   - Justification: Supervised Safety Fine Tuning, Safety RLHF, and Safety Context Distillation are labeled as mitigations and applied to a wide range of "safety risks"
   - Source: https://arxiv.org/pdf/2307.09288.pdf

58. Model → Model Mitigations → Mitigations Demonstration
   - Score: 1
   - Justification: Safety RLHF demonstrated via prompt and response (table 12) and supervised safety fine tuning demonstrated in same way (table 5), safety context distillation (table 39).
   - Source: https://arxiv.org/pdf/2307.09288.pdf

59. Model → Model Mitigations → Mitigations Evaluation
   - Score: 1
   - Justification: Quantitative evaluations of each mitigation given in section 4.2 of Llama paper.
   - Source: https://arxiv.org/pdf/2307.09288.pdf

60. Model → Model Mitigations → External Reproducibility of Mitigations Evaluation
   - Score: 0
   - Justification: No information found to indicate mitigations evaluations are externally reproducible.
   - 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.
    - 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: 0
    - Justification: No information found related to evaluations of robustness, reliability, hallucinations, uncertainty, calibration, causality, interpretability, or explainability.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

63. Model → Trustworthiness → External Reproducibility of Trustworthiness Evaluation
    - Score: 0
    - Justification: No information found related to trustworthiness evaluations, and so no information was found about their reproducibility.
    - 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: 1
    - Justification: Inference duration evaluated in figure 24 of technical report
    - Source: https://arxiv.org/pdf/2307.09288.pdf

65. Model → Inference → Inference Compute Evaluation
    - Score: 0
    - Justification: Meta discloses the number of GPU hours and the GPU hardware--which can produce an upper bound on the FLOPs--but does not disclose FLOPs.
    - 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: No information found related to a Meta's protocol for deciding whether to release Llama 2.
    - 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: Section 5.3 of paper discusses release strategy in sufficient detail to satisfy this indicator.
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

68. Downstream → Distribution → Distribution Channels

- Score: 1
- Justification: Meta discloses its preferred distribution channel as Azure.
- Source: https://web.archive.org/web/20230914013454/https://about.fb.com/news/2023/07/llama-2/

69. Downstream → Distribution → Products and Services

- Score: 1
- Justification: Meta discloses that Llama Chat and Code Llama are powered by Llama 2.
- Source: https://web.archive.org/web/20230914052546/https://ai.meta.com/llama/ and https://web.archive.org/web/20230914052503/https://ai.meta.com/blog/code-llama-large-language-model-coding/

70. Downstream → Distribution → Detection of Machine-Generated Content

- Score: 0
- Justification: No information found related to a mechanism for the detection of content generated by Llama 2.
- 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: 1
- Justification: "Custom commercial license" for Llama 2 is included at the bottom of the access request form.
- Source: https://ai.meta.com/resources/models-and-libraries/llama-downloads/

72. Downstream → Distribution → Terms of Service

- Score: 1
- Justification: As Meta's preferred distribution channel for Llama 2, Azure's Terms of Service satisfy this indicator.
- Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

73. Downstream → Usage Policy → Permitted, Restricted, and Prohibited Users

- Score: 1
- Justification: The custom commercial license for Llama 2 discloses restrictions on competitors - "If, on the Llama 2 version release date, the monthly active users of the products or services made available by or for Licensee, or Licensee's affiliates, is greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta, which Meta may grant to you in its sole discretion, and you are not authorized to exercise any of the rights under this Agreement unless or until Meta otherwise expressly grants you such rights."
- Source: https://ai.meta.com/resources/models-and-libraries/llama-downloads/

74. Downstream → Usage Policy → Permitted, Restricted, and Prohibited Uses

- Score: 1
- Justification: Acceptable Use Policy outlines prohibited uses, Model Card states "Intended Use Cases Llama 2 is intended for commercial and research use in English. Tuned models are intended for assistant-like chat, whereas pretrained models can be adapted for a variety of natural language generation tasks. Out-of-scope Uses Use in any manner that violates applicable laws or regulations (including trade compliance laws). Use in languages other than English. Use in any other way that is prohibited by the Acceptable Use Policy and Licensing Agreement for Llama 2."
- Source: https://ai.meta.com/resources/models-and-libraries/llama-downloads/ and https://arxiv.org/pdf/2307.09288.pdf

75. Downstream → Usage Policy → Usage Policy Enforcement

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

76. Downstream → Usage Policy → Justification for Enforcement Action
    - Score: 0
    - Justification: No information found related to whether users receive a justification when they are subject to an enforcement action for violating the usage policy.
    - 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: 0
    - Justification: No information found related to a mechanism for appealing the developer's decision that the usage policy has been violated.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

78. Downstream → Model Behavior Policy → Permitted, Restricted, and Prohibited Model Behaviors
    - Score: 0
    - Justification: High-level information is provided about prohibited model behaviors, but the information is not sufficient to satisfy this indicator. Llama 2 is trained to refuse to generate harmful, offensive, or biased content (analogue to Llama 2 chat on pp 58) - "Like other LLMs, Llama 2 may generate harmful, offensive, or biased content due to its training on publicly available online datasets. We attempted to mitigate this via fine-tuning". Permitted behavior is unclear, though the model card's reference to "natural language generation tasks" paired with the examples in the technical report is a weak version of this.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

79. Downstream → Model Behavior Policy → Model Behavior Policy Enforcement
    - Score: 0
    - Justification: No information found related to the enforcement protocol for the model behavior policy.
    - 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: 0
   - Justification: No information found about how the usage policy and the model behavior policy interoperate.
   - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

81. Downstream → User Interface → User Interaction with AI System
   - Score: 1
   - Justification: The Azure environment makes clear that the user is interacting with an AI system and with Llama 2 specifically. From Azure the user chooses from a "model catalog."
   - Source: https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/introducing-llama-2-on-azure/ba-p/3881233

82. Downstream → User Interface → Usage Disclaimers
   - Score: 1
   - Justification: Users via all distribution channels are required to agree to Meta's license, which includes the usage policy.
   - Source: https://ai.meta.com/resources/models-and-libraries/llama-downloads/

83. Downstream → User Data Protection → User Data Protection Policy
   - Score: 1
   - Justification: The Llama 2 access request page points to Meta's privacy policy, which includes protocols for how Meta stores, accesses, and shares user data.
   - Source: https://www.facebook.com/privacy/policy/

84. Downstream → User Data Protection → Permitted and Prohibited Use of User Data
   - Score: 1
   - Justification: Azure's privacy policy outlines permitted and prohibited uses of user data.
   - Source: https://www.microsoft.com/licensing/terms/product/PrivacyandSecurityTerms/all

85. Downstream → User Data Protection → Usage Data Access Protocol
   - Score: 0
   - Justification: No information found related to a protocol for granting external entities access to usage data.

- 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: Meta released the complete model weights to any developer who accepts their license, giving them access to a precise version of the model.
    - Source: https://ai.meta.com/resources/models-and-libraries/llama/ to download the model

87. Downstream → Model Updates → Change Log
    - Score: 1
    - Justification: Updates doc in GitHub serves as de facto change log; to update its model, Meta would need to introduce a new version of the model weights with accompanying documentation, which would serve as a change log.
    - Source: https://github.com/facebookresearch/llama/blob/main/UPDATES.md

88. Downstream → Model Updates → Deprecation Policy
    - Score: 1
    - Justification: Meta released the complete model weights to any developer who accepts their license, meaning there is no risk of deprecation.
    - Source: https://ai.meta.com/resources/models-and-libraries/llama/

89. Downstream → Feedback → Feedback Mechanism
    - Score: 1
    - Justification: The acceptable use policy and the custom commercial license point to the feedback page.
    - Source: https://developers.facebook.com/llama_output_feedback/

90. Downstream → Feedback → Feedback Summary
    - Score: 0
    - Justification: No information found related to any summary of user feedback.
    - 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 in connection with a summary of government inquiries related to the model.

- 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: 0
    - Justification: No information found related to a monitoring mechanism for tracking model use.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

93. Downstream → Impact → Downstream Applications
    - Score: 1
    - Justification: Hugging Face's tab outlining spaces using Llama 2 models is a sufficient proxy for downstream applications to satisfy this indicator.
    - Source: https://huggingface.co/meta-llama/Llama-2-7b-hf

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: Search protocol produced no 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: The GitHub centralizes a substantial amount of relevant documentation for downstream use for Llama 2, such as the Acceptable Use Policy and details about how to deploy Llama 2.
    - Source: https://github.com/facebookresearch/llama/blob/main/README.md

100. Downstream → Documentation for Deployers → Documentation for Responsible Downstream Use
    - Score: 1
    - Justification: Llama 2 responsible use guide provides such documentation
    - Source: https://web.archive.org/web/20230914013513/https://ai.meta.com/llama/responsible-use-guide/