INDICATORS - FMTI CroissantLLM

FMTI

Methodology:
We apply the FMTI indicators to the CroissantLLM base model.

Disclaimers:
The FMTI grid is meant to assess Foundation Models, but base models and models that were fine-tuned on instruction or chat datasets imply different training, evaluation and data curation protocols, thus largely modifying their assessment through the FMTI. Training an instruction or chat model from a base model is a process that has recently been completely democratized through the use of crowdsourced or synthetic datasets, and individuals are now fully capable of finetuning their own model variants in a variety of manners. As such, we consider this work's contribution mainly lies in the base model training, and are aware SFT finetuning of the Croissant model will be done outside of the author's control; whether on proprietary data, synthetic chat datasets, crowdsourced chat instructions - leading to different legal and copyright implications for the finetuned models. We thus focus on the base model in our evaluation, and give the complete criteria list as detailed in the appendix.
Transparency evaluation should ideally be done by an independent third party as there are obvious biases in auto-evaluating a model, and point attribution is not always trivial for certain criterias. As such, we take a rather conservative approach to point attribution and detail our process in an open document. Efforts have consciously been made within the technical report to include information not initially given to validate certain criterias, which puts us at a clear advantage with respects to work published before the index's release. We are open to discussions for potential scoring modifications, and consider these FMTI scores to be the reflection of our compliance efforts with respects to the listed transparency principles, rather than scores fairly comparable to the larger foundation models with vastly different usage objectives.

Criterias

1. Upstream → Data → Data size
   • Definition: For the data used in building the model, is the data size disclosed? • Notes: Data size should be reported in appropriate units (e.g. bytes, words, tokens, images, frames) and broken down by modality. Data size should be reported to a precision of one significant figure (e.g. 4 trillion tokens, 200 thousand images). No form of decomposition into data phases is required.
   • References: Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science, Datasheets for Datasets

   Yes - 3T tokens, 1.1T unique tokens

2. Upstream → Data → Data sources
   • Definition: For all data used in building the model, are the data sources disclosed? • Notes: To receive this point, a meaningful decomposition of sources must be listed in an understandable way (e.g. named URLs/domains/databases/data providers). It does not suffice to say data is “sourced from the Internet" or comes from "licensed sources”. • References: Datasheets for Datasets, Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure

   Yes - all sources are given, including internet data where URLs can be retrieved

3. Upstream → Data → Data creators
   • Definition: For all data used in building the model, is there some characterization of the people who created the data?
   • Notes: While information about data creators may not be easily discernible for some data scraped from the web, the general sources (URLs/domains) should be listed, and, for other data that is bought, licensed, or collected, a reasonable attempt at characterizing the under lying people who provided the data is required to receive this point. The relevant properties of people can vary depending on context: for example, relevant properties could include demographic information like fraction of Black individuals contributing to the dataset, geographic information like fraction of European individuals contributing to the dataset, language information like fraction of L1 English speakers, or occupational information like the fraction of professional artists.
   • References: Datasheets for Datasets, Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure

   No - difficult to do at this scale, although attempts to diversify the corpus in terms of geographical and language distribution is made

4. Upstream → Data → Data source selection
   • Definition: Are the selection protocols for including and excluding data sources disclosed? • Notes: Selection protocols refer to procedures used to choose which datasets or subsets of datasets will be used to build a model. We will award this point even if the selection protocols are non-exhaustive.
   • References: Datasheets for Datasets, Towards Accountability for Machine Learning

Datasets: Practices from Software Engineering and Infrastructure

Yes, and code is given

5. Upstream → Data → Data curation
   • Definition: For all data sources, are the curation protocols for those data sources disclosed? • Notes: Curation protocols refer to steps taken to further modify data sources, such as procedures to manage, annotate, and organize data. The aims of curation might include improving the quality, relevance, and representativeness of the data. We will award this point if the developer reports that it does not perform any further curation beyond the data sources.
   • References: Datasheets for Datasets, Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure

   Yes and code is given

6. Upstream → Data → Data augmentation
   • Definition: Are any steps the developer takes to augment its data sources disclosed? • Notes: Such steps might include augmenting data sources with synthetic data. We will award this point if the developer reports that it does not take any steps to augment its data. • References: Datasheets for Datasets, Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure

   No data augmentation, it is said in the report

7. Upstream → Data → Harmful data filtration
   • Definition: If data is filtered to remove harmful content, is there a description of the associ ated filter?
   • Notes: Such harmful content might relate to violence or child sexual abuse material. We will award this point if the developer reports that it does not perform any harmful data filtration.
   • References: Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, Toxicity

   yes - Data filtration is done, and to ensure minimal risks, we rely on internet datasets created through open code, and that have been extensively used - instead of doing everything in house.

8. Upstream → Data → Copyrighted data
   • Definition: For all data used in building the model, is the associated copyright status dis closed?
   • Notes: To receive this point, the copyright status (e.g. copyrighted, public domain) must relate to some decomposition of the data. We will award this point if there is some meaningful decomposition of the data, even if the decomposition is insufficient to receive the Data Creators point or if the disclosure is not comprehensive relative to legal copyright standards.
   • References: Addressing "Documentation Debt" in Machine Learning Research: A

Retro spective Datasheet for BookCorpus, Machine Learning and Artificial Intelligence: Legal Concepts

Yes - everything is public domain data

9. Upstream → Data → Data license
   - Definition: For all data used in building the model, is the associated license status disclosed? • Notes: To receive this point, the license status must relate to some decomposition of the data. We will award this point if there is some meaningful decomposition of the data, even if the decomposition is insufficient to receive the Data Creators point.
   - References: Addressing "Documentation Debt" in Machine Learning Research: A Retro spective Datasheet for BookCorpus, Machine Learning and Artificial Intelligence: Legal Concepts

   No - although best efforts are done to characterize the licences for all data sources.

10. Upstream → Data → Personal information in data
    - Definition: For all data used in building the model, is the inclusion or exclusion of personal information in that data disclosed?
    - Notes: To receive this point, the disclosure of personal information must relate to some decomposition of the data. We will award this point if there is some meaningful decomposi tion of the data, even if the decomposition is insufficient to receive the Data Creators point. Additionally, we will award this point if the developer reports the inclusion of personal information, independent of if and how they mitigate related privacy concerns.
    - References: Data Capitalism: Redefining the Logics of Surveillance and Privacy, What Does it Mean for a Language Model to Preserve Privacy?

    No - not possible at this scale

11. Upstream → Data labor → Use of human labor
    - Definition: Are the phases of the data pipeline where human labor is involved disclosed? • Notes: Phases of the data pipeline that involve human labor include activities and tasks performed by people to collect, annotate, clean, or validate data. This indicator is inclusive of all data that is created by or on behalf of the developer. We will award this point if the developer gives a reasonable best-effort description of the use of human labor in their data pipeline.
    - References: The future of crowd work, AI Is a Lot of Work: As the technology becomes ubiquitous, a vast tasker underclass is emerging — and not going anywhere

    Yes - all sources are collected by authors of the papers. sources are themselves credited or described.

12. Upstream → Data labor → Employment of data laborers
    - Definition: Is the organization that directly employs the people involved in data labor disclosed for each phase of the data pipeline?
    - Notes: Phases of the data pipeline that involve human labor include activities and tasks performed by people to collect, annotate, clean, or validate data. This

indicator is inclusive of all data that is created by or on behalf of the developer. We will award this point if the developer provides the name of the organization that employs data laborers, even if other details about the employment relationship are not disclosed.
- References: The future of crowd work, AI Is a Lot of Work: As the technology becomes ubiquitous, a vast tasker underclass is emerging — and not going anywhere

Yes - affiliations for the authors

13. Upstream → Data labor → Geographic distribution of data laborers • Definition: Is geographic information regarding the people involved in data labor disclosed for each phase of the data pipeline?
    - Notes: This indicator is inclusive of all data that is created by or on behalf of the developer. We will award this point if the developer gives a reasonable best-effort description of the geographic distribution of labor at the country-level.
    - References: Cleaning Up ChatGPT Takes Heavy Toll on Human Workers, Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass

    Yes - affiliations for the authors

14. Upstream → Data labor → Wages
    - Definition: Are the wages for people who perform data labor disclosed? • Notes: This indicator is inclusive of data labor at all points of the model development process, such as training data annotation or red teaming data used to control the model. We will award this point if the developer reports that it does not compensate workers. For all data that is created by or on behalf of the developer,
    - References: The future of crowd work, AI Is a Lot of Work: As the technology becomes ubiquitous, a vast tasker underclass is emerging — and not going anywhere

    Yes - info is given (no additional work for base model)

15. Upstream → Data labor → Instructions for creating data
    - Definition: Are the instructions given to people who perform data labor disclosed? • Notes: This indicator is inclusive of all data that is created by or on behalf of the developer. We will award this point if the developer makes a reasonable best-effort attempt to disclose instructions given to people who create data used to build the model for the bulk of the data phases involving human labor.
    - References: Everyone wants to do the model work, not the data work, The future of crowd work

    Yes - info is given (no additional work for base model)

16. Upstream → Data labor → Labor protections
    - Definition: Are the labor protections for people who perform data labor disclosed? • Notes: This indicator is inclusive of data labor at all points of the model development process, such as training data annotation or red teaming data used to control the model. It is also inclusive of all data that is created by or on behalf of the developer. As an example, labor protections might include protocols to reduce the harm to

workers' mental health stemming from exposure to violent content when annotating training data. We will award this point if the developer reports that it does not protect workers or if it does not use data laborers and therefore has no labor protections.

- References: The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence, Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass

Yes - info is given (no additional work for base model)

17. Upstream → Data labor → Third party partners
    - Definition: Are the third parties who were or are involved in the development of the model disclosed?
    - Notes: This indicator is inclusive of partnerships that go beyond data labor as there may be third party partners at various stages in the model development process. We will award this point if the developer reports that it was the sole entity involved in the development of the model.
    - References: The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence, Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass

Yes - info is given (no additional entities)

18. Upstream → Data access → Queryable external data access
    - Definition: Are external entities provided with queryable access to the data used to build the model?
    - Notes: We will award this point for any reasonable mechanism for providing access: direct access to the data, an interface to query the data, a developer-mediated access program where developers can inspect requests, etc. Developers may receive this point even if there are rate-limits on the number of queries permitted to an external entity and restrictions on which external entities are given access, insofar as these limits and restrictions are transparent and ensure a reasonable amount of external access. We may accept justifications for prohibiting queries of specific parts of the data.
- References: Datasheets for Datasets, The ROOTS Search Tool: Data Transparency for LLMs

Yes - Direct access to the data

19. Upstream → Data access → Direct external data access
    - Definition: Are external entities provided with direct access to the data used to build the model?
    - Notes: We will award this point if external entities can directly access the data without any form of gating from the developer. With that said, we may award this point if the developer provides justifications for prohibiting access to specific parts of the data or to unauthorized external entities.
- References: Datasheets for Datasets, The ROOTS Search Tool: Data Transparency for LLMs

Yes - Direct access to the data

20. Upstream → Compute → Compute usage
    • Definition: Is the compute required for building the model disclosed? • Notes: Compute should be reported in appropriate units, which most often will be floating point operations (FLOPS). Compute should be reported to a precision of one significant figure (e.g. 5 x 10^{25} FLOPS). We will award this point even if there is no decomposition of the reported compute usage into compute phases, but it should be clear whether the reported compute usage is for a single model run or includes additional runs, or hyperparameter tuning, or training other models like reward models, or other steps in the model development process that necessitate compute expenditure.
    • References: Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning, Energy and Policy Considerations for Deep Learning in NLP*

    Yes - given in the tech report (training)

21. Upstream → Compute → Development duration
    • Definition: Is the amount of time required to build the model disclosed? • Notes: The continuous duration of time required to build the model should be reported in weeks, days, or hours to a precision of one significant figure (e.g. 3 weeks). No form of decomposition into phases of building the model is required for this indicator, but it should be clear what the duration refers to (e.g. training the model, training and subsequent evaluation and red teaming).
    • References: Compute Trends Across Three Eras of Machine Learning, Training Compute Optimal Large Language Models

    Yes - given in the tech report (training)

22. Upstream → Compute → Compute hardware
    • Definition: For the primary hardware used to build the model, is the amount and type of hardware disclosed?
    • Notes: In most cases, this indicator will be satisfied by information regarding the number and type of GPUs or TPUs used to train the model. The number of hardware units should be reported to a precision of one significant figure (e.g. 800 NVIDIA H100 GPUs). We will not award this point if (i) the training hardware generally used by the developer is disclosed, but the specific hardware for the given model is not, or (ii) the training hardware is disclosed, but the amount of hardware is not. We will award this point even if information about the interconnects between hardware units is not disclosed.
    • References: Compute Trends Across Three Eras of Machine Learning, Training Compute Optimal Large Language Models

    Yes - given in the tech report (trainig,)

23. Upstream → Compute → Hardware owner
    • Definition: For the primary hardware used in building the model, is the owner of

the hardware disclosed?

- Notes: For example, the hardware owner may be the model developer in the case of a self owned cluster, a cloud provider like Microsoft Azure, Google Cloud Platform, or Amazon Web Services, or a national supercomputer. In the event that hardware is owned by multiple sources or is highly decentralized, we will award this point if a developer makes a reasonable effort to describe the distribution of hardware owners.
- References: Compute Trends Across Three Eras of Machine Learning, Training Compute Optimal Large Language Models

Yes - Jean Zay given by the tech report (training)

24. Upstream → Compute → Energy usage

• Definition: Is the amount of energy expended in building the model disclosed? • Notes: Energy usage should be reported in appropriate units, which most often will be megawatt-hours (mWh). Energy usage should be reported to a precision of one significant figure (e.g. 500 mWh). No form of decomposition into compute phases is required, but it should be clear whether the reported energy usage is for a single model run or includes additional runs, or hyperparameter tuning, or training other models like reward models, or other steps in the model development process that necessitate energy usage. • References: Quantifying the Carbon Emissions of Machine Learning, Carbon Emissions and Large Neural Network Training

Yes - given in the report (eco impact)

25. Upstream → Compute → Carbon emissions

- Definition: Is the amount of carbon emitted (associated with the energy used) in building the model disclosed?
- Notes: Emissions should be reported in appropriate units, which most often will be tons of carbon dioxide emitted (tCO2). Emissions should be reported to a precision of one significant figure (e.g. 500 tCO2). No form of decomposition into compute phases is required, but it should be clear whether the reported emissions is for a single model run or includes additional runs, or hyperparameter tuning, or training other models like reward models, or other steps in the model development process that generate emissions.
- References: Quantifying the Carbon Emissions of Machine Learning, Carbon Emissions and Large Neural Network Training

Yes - given in the tech report (eco impact)

26. Upstream → Compute → Broader environmental impact

- Definition: Are any broader environmental impacts from building the model besides carbon emissions disclosed?
- Notes: While the most direct environmental impact of building a foundation model is the energy used and, therefore, the potential carbon emissions, there may be other environmen tal impacts. For example, these may include the use of other resources such as water for cooling data centers or metals for producing specialized hardware. We recognize that there does not exist an authoritative or consensus list of broader environmental factors. For this reason, we will award this

point if there is a meaningful, though potentially incomplete, discussion of broader environmental impact.
- References: Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning, Energy and Policy Considerations for Deep Learning in NLP

No - Not disclosed

27. Upstream → Methods → Model stages
    - Definition: Are all stages in the model development process disclosed? • Notes: Stages refer to each identifiable step that constitutes a substantive change to the model during the model building process. We recognize that different developers may use different terminology for these stages, or conceptualize the stages differently. We will award this point if there is a clear and complete description of these stages.
    - References: Model Cards for Model Reporting, Scaling Instruction-Finetuned Language Models

    Yes - detailed in the report

28. Upstream → Methods → Model objectives
    - Definition: For all stages that are described, is there a clear description of the associated learning objectives or a clear characterization of the nature of this update to the model? • Notes: We recognize that different developers may use different terminology for these stages, or conceptualize the stages differently. We will award this point if there is a clear description of the update to the model related to each stage, whether that is the intent of the stage (e.g. making the model less harmful), a mechanistic characterization (e.g. minimizing a specific loss function), or an empirical assessment (e.g. evaluation results conducted before and after the stage).
    - References: Model Cards for Model Reporting, Scaling Instruction-Finetuned Language Models

    Yes - CLM

29. Upstream → Methods → Core frameworks
    - Definition: Are the core frameworks used for model development disclosed? • Notes: Examples of core frameworks include Tensorflow, PyTorch, Jax, Hugging Face Trans formers, Seqio, T5X, Keras, SciKit, and Triton. If there are significant internal frameworks, there should be some description of their function and/or a reasonably similar publicly available analogue. We recognize that there does not exist an authoritative or consensus list of core frameworks. For this reason, we will award this point if there is a meaningful, though potentially incomplete, list of major frameworks for the first version of the index. • References: Model Cards for Model Reporting, Scaling Instruction-Finetuned Language Models

    Yes - In tech report

30. Upstream → Methods → Additional dependencies

- Definition: Are any dependencies required to build the model disclosed besides data, com pute, and code?
- Notes: For example, if the model depends on an external search engine, programmable APIs, or tools, this should be disclosed. We recognize that there is not widespread consensus regarding what constitutes key dependencies beyond the data, compute, and code. We will award this point only if developers give a reasonable best-effort description of any additional dependencies or make clear that no additional dependencies are required.
- References: Analyzing Leakage of Personally Identifiable Information in Language Models, ProPILE: Probing Privacy Leakage in Large Language Models

Yes - no additional deps

31. Upstream → Data Mitigations → Mitigations for privacy
    - Definition: Are any steps the developer takes to mitigate the presence of PII in the data disclosed?
    - Notes: Such steps might include identifying personal information in the training data, filtering specific datasets to remove personal information, and reducing the likelihood that models will output personal information. We will award this point if the developer reports that it does not take steps to mitigate the presence of PII in the data.
    - References: Deduplicating Training Data Mitigates Privacy Risks in Language Models, Machine Learning and Artificial Intelligence: Legal Concepts

    Yes - PII filters are implemented on the internet dataset. we also integrate canaries to assess the risks of PII leakage which we find to be low to non-existing.

32. Upstream → Data Mitigations → Mitigations for copyright
    - Definition: Are any steps the developer takes to mitigate the presence of copyrighted information in the data disclosed?
    - Notes: Such steps might include identifying copyrighted data, filtering specific datasets to remove copyrighted data, and reducing the likelihood that models will output copyrighted information. We will award this point if the developer reports that it does take steps to mitigate the presence of copyrighted information in the data.
    - References: Addressing "Documentation Debt" in Machine Learning Research: A Retro spective Datasheet for BookCorpus, Machine Learning and Artificial Intelligence: Legal Concepts

    Yes - listed in the paper

33. Model → Model basics → Input modality
    - Definition: Are the input modalities for the model disclosed?
    - Notes: Input modalities refer to the types or formats of information that the model can accept as input. Examples of input modalities include text, image, audio, video, tables, graphs.
    - References: Model Cards for Model Reporting, Interactive Model Cards: A Human-Centered Approach to Model Documentation

Yes - text

34. Model → Model basics → Output modality
    - Definition: Are the output modalities for the model disclosed?
    - Notes: Output modalities refer to the types or formats of information that the model can accept as output. Examples of output modalities include text, image, audio, video, tables, graphs.
    - References: Model Cards for Model Reporting, Interactive Model Cards: A Human-Centered Approach to Model Documentation

Yes - text

35. Model → Model basics → Model components
    - Definition: Are all components of the model disclosed?
    - Notes: Model components refer to distinct and identifiable parts of the model. We recognize that different developers may use different terminology for model components, or conceptu alize components differently. Examples include: (i) For a text-to-image model, components could refer to a text encoder and an image encoder, which may have been trained separately. (ii) For a retrieval-augmented model, components could refer to a separate retriever module.
    - References: Model Cards for Model Reporting, Interactive Model Cards: A Human-Centered Approach to Model Documentation

Yes - llama architecture

36. Model → Model basics → Model size
    - Definition: For all components of the model, is the associated model size disclosed? • Notes: This information should be reported in appropriate units, which generally is the number of model parameters, broken down by named component. Model size should be reported to a precision of one significant figure (e.g. 500 billion parameters for text encoder, 20 billion parameters for image encoder).
    - References: Model Cards for Model Reporting, Interactive Model Cards: A Human-Centered Approach to Model Documentation

Yes - described in paper

37. Model → Model basics → Model architecture
    - Definition: Is the model architecture disclosed?
    - Notes: Model architecture is the overall structure and organization of a foundation model, which includes the way in which any disclosed components are integrated and how data moves through the model during training or inference. We recognize that different develop ers may use different terminology for model architecture, or conceptualize the architecture differently. We will award this point for any clear, though potentially incomplete, description of the model architecture.
    - References: Model Cards for Model Reporting, Interactive Model Cards: A Human-Centered Approach to Model Documentation

Yes - llama architecture which is open

38. Model → Model basics → Centralized model documentation
- Definition: Is key information about the model included in a centralized artifact such as a model card?
- Notes: We recognize that different developers may share this information through different types of documentation, such as a system card or several clearly interrelated documents. We will award this point for the disclosure of any such centralized artifact that provides key information typically included in a model card, though the artifact may be longer-form than a standard model card (e.g. a technical report).
- References: Model Cards for Model Reporting, Interactive Model Cards: A Human-Centered Approach to Model Documentation

Yes - llama architecture which is open

39. Model → Model access → External model access protocol
- Definition: Is a protocol for granting external entities access to the model disclosed? • Notes: A model access protocol refers to the steps, requirements, and considerations involved in granting authorized model access to external entities. We will award this point if the developer discloses key details of its protocol, including (i) where external entities can request access (e.g. via an access request form); (ii) explicit criteria for selecting external entities; and (iii) a transparent decision on whether access has been granted within a specified, reasonable period of time.
- References: The Gradient of Generative AI Release: Methods and Considerations, Structured access: an emerging paradigm for safe AI deployment

Yes - direct access

40. Model → Model access → Blackbox external model access
- Definition: Is black box model access provided to external entities?
- Notes: Black box model access refers to the ability to query the model with inputs and receive outputs, potentially without further access. Examples of external entities that might be granted access include researchers, third-party auditors, and regulators. We will award this point for any reasonable access level: direct access to the model weights, an interface to query the model, a developer-mediated access program where developers can inspect requests, etc. Developers may receive this point even if there are rate-limits on the number of queries permitted to an external entity and restrictions on the external entities that are permitted access, insofar as these limits and restrictions are transparent.
- References: The Gradient of Generative AI Release: Methods and Considerations, Structured access: an emerging paradigm for safe AI deployment

Yes - direct access

41. Model → Model access → Full external model access
- Definition: Is full model access provided to external entities?
- Notes: Full model access refers to the ability to access the model via the release

of model weights. Developers may receive this point even if there are some restrictions on the external entities that are permitted access (e.g. geographic restrictions), insofar as these restrictions are transparent (e.g. via some high-level description of who has been granted access to the foundation model).
- References: The Gradient of Generative AI Release: Methods and Considerations, Structured access: an emerging paradigm for safe AI deployment

Yes - direct access

42. Model → Capabilities → Capabilities description
- Definition: Are the model's capabilities described?
- Notes: Capabilities refer to the specific and distinctive functions that the model can perform. We recognize that different developers may use different terminology for capabilities, or conceptualize capabilities differently. We will award this point for any clear, but potentially incomplete, description of the multiple capabilities.
- References: Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, Holistic Evaluation of Language Models

Yes -extensive testing + direct access

43. Model → Capabilities → Capabilities demonstration
- Definition: Are the model's capabilities demonstrated?
- Notes: Demonstrations refer to illustrative examples or other forms of showing the model's capabilities that are legible or understandable for the general public, without requiring specific technical expertise. We recognize that different developers may use different termi nology for capabilities, or conceptualize capabilities differently. We will award this point for clear demonstrations of multiple capabilities.
- References: Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, Holistic Evaluation of Language Models

Yes - extensive testing + direct access

44. Model → Capabilities → Evaluation of capabilities
- Definition: Are the model's capabilities rigorously evaluated, with the results of these evaluations reported prior to or concurrent with the initial release of the model?
- Notes: Rigorous evaluations refer to precise quantifications of the model's behavior in relation to its capabilities. We recognize that capabilities may not perfectly align with evaluations, and that different developers may associate capabilities with evaluations differ ently. We will award this point for clear evaluations of multiple capabilities. For example, this may include evaluations of world knowledge, reasoning, state tracking or other such proficiencies. Or it may include the measurement of average performance (e.g. accuracy, F1) on benchmarks for specific tasks (e.g. text summarization, image captioning). We note that evaluations on standard broad-coverage benchmarks are likely to suffice for this indicator, though they may not if the model's capabilities are presented as especially unusual such that standard evaluations will not suffice.
- References: Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, Holistic Evaluation of Language Models

Yes - extensive testing + direct access

12

45. Model → Capabilities → External reproducibility of capabilities evaluation • Definition: Are the evaluations of the model's capabilities reproducible by external entities? • Notes: For an evaluation to be reproducible by an external entity, we mean that the associated

    data is either (i) publicly available or (ii) described sufficiently such that a reasonable facsimile can be constructed by an external entity. In addition, the evaluation protocol should be sufficiently described such that if the evaluation is reproduced, any discrepancies with the developer's results can be resolved. We recognize that there does not exist an authoritative or consensus standard for what is required for an evaluation to be deemed externally reproducible. Evaluations on standard benchmarks are assumed to be sufficiently reproducible for the purposes of this index. We will award this point for reproducibility of multiple disclosed evaluations. In the event that an evaluation is not reproducible, a justification by the model developer for why it is not possible for the evaluation to be made reproducible may be sufficient to score this point.

- References: Leakage and the reproducibility crisis in machine-learning-based science, Holis tic Evaluation of Language Models

Yes - evaluation is open-sourced and transparent

46. Model → Capabilities → Third party capabilities evaluation
    - Definition: Are the model's capabilities evaluated by third parties?
    - Notes: By third party, we mean entities that are significantly or fully independent of the developer. We will award this point if (i) a third party has conducted an evaluation of model capabilities, (ii) the results of this evaluation are publicly available, and (iii) these results are disclosed or referred to in the developer's materials.
    - References: Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Gover nance, Holistic Evaluation of Language Models

No - not to this day but the evaluation framework is available and everyone can run it to reproduce so this will change

47. Model → Limitations → Limitations description
    - Definition: Are the model's limitations disclosed?
    - Notes: Limitations refer to the specific and distinctive functions that the model cannot perform (e.g. the model cannot answer questions about current events as it only contains data up to a certain time cutoff, the model is not very capable when it comes to a specific application). We recognize that different developers may use different terminology for limitations, or conceptualize limitations differently. We will award this point for any clear, but potentially incomplete, description of multiple limitations.
    - References: The Fallacy of AI Functionality, Holistic Evaluation of Language Models

Yes - Todo

48. Model → Limitations → Limitations demonstration
    - Definition: Are the model's limitations demonstrated?
    - Notes: Demonstrations refer to illustrative examples or other forms of showing the limita tions that are legible or understandable for the general public, without requiring specific technical expertise. We recognize that different developers may use different terminology for limitations, or conceptualize the limitations differently. We will award this point for clear demonstrations of multiple limitations.
    - References: The Fallacy of AI Functionality, Holistic Evaluation of Language Models

    Yes - in the model limitations section

49. Model → Limitations → Third party evaluation of limitations
    - Definition: Can the model's limitations be evaluated by third parties? • Notes: By third parties, we mean entities that are significantly or fully independent of the model developers. In contrast to the third party evaluation indicators for capabilities and risks, we will award this point if third party evaluations are possible even if no third party has yet conducted them. Such evaluations are possible if, for example, the model is deployed via an API (or with open weights) and there are no restrictions on evaluating limitations (e.g. in the usage policy).
    - References: Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Gover nance, Holistic Evaluation of Language Models

    Yes - Open model

50. Model → Risks → Risks description
    - Definition: Are the model's risks disclosed?
    - Notes: Risks refer to possible negative consequences or undesirable outcomes that can arise from the model's deployment and usage. This indicator requires disclosure of risks that may arise in the event of both (i) intentional (though possibly careless) use, such as bias or hallucinations and (ii) malicious use, such as fraud or disinformation. We recognize that different developers may use different terminology for risks, or conceptualize risks differently. We will award this point for any clear, but potentially incomplete, description of multiple risks.
    - References: Evaluating the Social Impact of Generative AI Systems in Systems and Society, Ethical and social risks of harm from Language Models

    Yes - in the model limitations section

51. Model → Risks → Risks demonstration
    - Definition: Are the model's risks demonstrated?
    - Notes: Demonstrations refer to illustrative examples or other forms of showing the risks that are legible or understandable for the general public, without requiring specific technical expertise. This indicator requires demonstration of risks that may arise in the event of both (i) intentional (though possibly careless) use, such as biases or hallucinations and (ii) malicious use, such as fraud or disinformation. We recognize that different developers may use different terminology for risks, or conceptualize risks differently. We will award this point for clear demonstrations of multiple risks.

- References: Evaluating the Social Impact of Generative AI Systems in Systems and Society, Ethical and social risks of harm from Language Models

Yes - to show with limitations (hallucinations)

52. Model → Risks → Unintentional harm evaluation
    - Definition: Are the model's risks related to unintentional harm rigorously evaluated, with the results of these evaluations reported prior to or concurrent with the initial release of the model?
    - Notes: Rigorous evaluations refer to precise quantifications of the model's behavior in relation to such risks. Unintentional harms include bias, toxicity, and issues relating to fairness. We recognize that unintended harms may not perfectly align with risk evaluations, and that different developers may associate risks with evaluations differently. We will award this point for clear evaluations of multiple such risks. We note that evaluations on standard broad-coverage benchmarks are likely to suffice for this indicator, though they may not if the model's risks related to unintentional harm are presented as especially unusual or severe.
    - References: Evaluating the Social Impact of Generative AI Systems in Systems and Society, Ethical and social risks of harm from Language Models

Yes -Crows dataset scores

53. Model → Risks → External reproducibility of unintentional harm evaluation • Definition: Are the evaluations of the model's risks related to unintentional harm repro ducible by external entities?
    - Notes: For an evaluation to be reproducible by an external entity, we mean that the associated data is either (i) publicly available or (ii) described sufficiently such that a reasonable facsimile can be constructed by the external entity. In addition, the evaluation protocol should be sufficiently described such that if the evaluation is reproduced, any discrepancies with the developer's results can be resolved. We recognize that there does not exist an authoritative or consensus standard for what is required for an evaluation to be deemed externally reproducible. Evaluations on standard benchmarks are assumed to be sufficiently reproducible for the purposes of this index. We will award this point for reproducibility of multiple disclosed evaluations. In the event that an evaluation is not reproducible, a justification by the developer for why it is not possible for the evaluation to be made reproducible may suffice.
    - References: Leakage and the reproducibility crisis in machine-learning-based science, Ethical and social risks of harm from Language Models

Yes - Crows dataset is available publicly

54. Model → Risks → Intentional harm evaluation
    - Definition: Are the model's risks related to intentional harm rigorously evaluated, with the results of these evaluations reported prior to or concurrent with the initial release of the model?.
    - Notes: Rigorous evaluations refer to precise quantifications of the model's behavior in relation to such risks. Intentional harms include fraud, disinformation, scams, cybersecurity attacks, designing weapons or pathogens, and uses of the

model for illegal purposes. We recognize that unintentional harms may not perfectly align with risk evaluations, and that different developers may associate risks with evaluations differently. We will award this point for clear evaluations of multiple such risks. We note that evaluations on standard broad-coverage benchmarks are likely to suffice for this indicator, though they may not if the model's risks related to unintentional harm are presented as especially unusual or severe.

- References: Evaluating the Social Impact of Generative AI Systems in Systems and Society, Ethical and social risks of harm from Language Models

No

55. Model → Risks → External reproducibility of intentional harm evaluation
    - Definition: Are the evaluations of the model's risks related to intentional harm reproducible by external entities?
    - Notes: For an evaluation to be reproducible by an external entity, we mean that the associated data is either (i) publicly available or (ii) described sufficiently such that a reasonable facsimile can be constructed by the external entity. In addition, the evaluation protocol should be sufficiently described such that if the evaluation is reproduced, any discrepancies with the developer's results can be resolved. We recognize that there does not exist an authoritative or consensus standard for what is required for an evaluation to be deemed externally reproducible. Evaluations on standard benchmarks are assumed to be sufficiently reproducible for the purposes of this index. We will award this point for reproducibility of multiple disclosed evaluations. In the event that an evaluation is not reproducible, a justification by the model developer for why it is not possible for the evaluation to be made reproducible may suffice.
    - References: Leakage and the reproducibility crisis in machine-learning-based science, Ethical and social risks of harm from Language Models

No

56. Model → Risks → Third party risks evaluation
    - Definition: Are the model's risks evaluated by third parties?
    - Notes: By third party, we mean entities that are significantly or fully independent of the developer. A third party risk evaluation might involve the developer allowing a third party to choose a methodology for evaluating risks that differs from that of the developer. We will award this point if (i) a third party has conducted an evaluation of model risks, (ii) the results of this evaluation are publicly available, and (iii) these results are disclosed or referred to in the developer's materials. If the results are not made public (but are disclosed to have been conducted) and/or the results are not discoverable in the developer's materials, we will not award this point. We may accept a justification from either the third party or the developer for why part of the evaluation is not disclosed in relation to risks.
    - References: Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Gover nance, Ethical and social risks of harm from Language Models

No

57. Model → Model Mitigations → Mitigations description
- Definition: Are the model mitigations disclosed?
- Notes: By model mitigations, we refer to interventions implemented by the developer at the level of the model to reduce the likelihood and/or the severity of the model's risks. We recognize that different developers may use different terminology for mitigations, or conceptualize mitigations differently. We will award this point for any clear, but poten tially incomplete, description of multiple mitigations associated with the model's risks. Alternatively, we will award this point if the developer reports that it does not mitigate risk.
- References: Evaluating the Social Impact of Generative AI Systems in Systems and Society, Ethical and social risks of harm from Language Models

Yes - no mitigation apart from data curation in pretraining and SFT dataset which is aligned

58. Model → Model Mitigations → Mitigations demonstration
- Definition: Are the model mitigations demonstrated?
- Notes: Demonstrations refer to illustrative examples or other forms of showing the mitiga tions that are legible or understandable for the general public, without requiring specific technical expertise. We recognize that different developers may use different terminology for mitigations, or conceptualize mitigations differently. We will award this point for clear demonstrations of multiple mitigations. We will also award this point if the developer reports that it does not mitigate the risks associated with the model.
- References: Evaluating the Social Impact of Generative AI Systems in Systems and Society, Ethical and social risks of harm from Language Models

Yes - example (refusal to answer medical question)

59. Model → Model Mitigations → Mitigations evaluation
- Definition: Are the model mitigations rigorously evaluated, with the results of these evalu ations reported?
- Notes: Rigorous evaluations refer to precise quantifications of the model's behavior in relation to the mitigations associated with its risks. We will award this point for clear evaluations of multiple mitigations.
- References: Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation, Ethical and social risks of harm from Language Models

No

60. Model → Model Mitigations → External reproducibility of mitigations evaluation • Definition: Are the model mitigation evaluations reproducible by external entities? • Notes: For an evaluation to be reproducible by an external entity, we mean that the associated

data is either (i) publicly available or (ii) described sufficiently such that a reasonable facsimile can be constructed by the external entity. In addition, the evaluation protocol should be sufficiently described such that if the evaluation is reproduced, any discrepancies with the developer's results can be resolved. In the

case of mitigations evaluations, this will usually involve details about a comparison to some baseline, which may be a different, unmitigated version of the model. We recognize that there does not exist an authoritative or consensus standard for what is required for an evaluation to be deemed externally reproducible. We will award this point for reproducibility of multiple disclosed evaluations. In the event that an evaluation is not reproducible, a justification by the model developer for why it is not possible for the evaluation to be made reproducible may suffice.

- References: Leakage and the reproducibility crisis in machine-learning-based science, Ethical and social risks of harm from Language Models

No

61. Model → Model Mitigations → Third party mitigations evaluation
    - Definition: Can the model mitigations be evaluated by third parties?
    - Notes: By third party, we mean entities that are significantly or fully independent of the model developers. This indicator assesses whether it is possible for third parties to assess mitigations, which is not restricted to the methods the developer uses to assess mitigations. In contrast to the third party evaluation indicators for capabilities and risks, we will award this point if third party evaluations are possible even if no third party has yet conducted them.
    - References: Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Gover nance, Ethical and social risks of harm from Language Models

No

62. Model → Trustworthiness → Trustworthiness evaluation
    - Definition: Is the trustworthiness of the model rigorously evaluated, with the results of these evaluations disclosed?
    - Notes: Rigorous evaluations refer to precise quantifications of the model's behavior in relation to its trustworthiness. For example, this may include evaluations of the model's robustness or reliability, its uncertainty, calibration, or causality, or its interpretability or explainability. We recognize that trustworthiness may not perfectly align with evaluations, and that different developers may associate trustworthiness with evaluations differently. We will award this point for a clear evaluation of the trustworthiness of the model.
    - References: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models

No

63. Model → Trustworthiness → External reproducibility of trustworthiness evaluation
    - Definition: Are the trustworthiness evaluations reproducible by external entities?
    - Notes: For an evaluation to be reproducible by an external entity, we mean that the associated data is either (i) publicly available or (ii) described sufficiently such that a reasonable facsimile can be constructed by the external entity. In addition, the evaluation protocol should be sufficiently described such that if the evaluation is

reproduced, any discrepancies with the developer's results can be resolved. We recognize that there does not exist an authoritative or consensus standard for what is required for an evaluation to be deemed externally reproducible. Evaluations on standard benchmarks are assumed to be sufficiently reproducible for the purposes of this index. We will award this point for reproducibility of at least one evaluation. In the event that an evaluation is not reproducible, we may accept a justification by the model developer for why it is not possible for the evaluation to be made reproducible.

- References: Leakage and the reproducibility crisis in machine-learning-based science, Bridg ing the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems

No

64. Model → Inference → Inference duration evaluation
    - Definition: Is the time required for model inference disclosed for a clearly-specified task on a clearly-specified set of hardware?
    - Notes: The duration should be reported in seconds to a precision of one significant figure (e.g. 0.002 seconds). We recognize that no established standard exists for the standardized reporting of inference evaluation. Therefore, we permit the developer to specify the task and hardware setup, as long as both are disclosed. The hardware in this evaluation need not be the hardware the developer uses for inference if it in fact does any inference itself. For example, the specific task might be generating 100,000 tokens as 5,000 sequences of length 20 and the fixed set of hardware might be 8 NVIDIA A100s. The hardware in this evaluation need not be the hardware the developer uses for inference if it in fact does any inference itself.
    - References: MLPerf Inference Benchmark, Cheaply Evaluating Inference Efficiency Metrics for Autoregressive Transformer APIs

    Yes

65. Model → Inference → Inference compute evaluation
    - Definition: Is the compute usage for model inference disclosed for a clearly-specified task on a clearly-specified set of hardware?
    - Notes: Compute usage for inference should be reported in FLOPS to a precision of one significant figure (e.g. 5 x 10^{25} FLOPS). We recognize that no established standard exists for the standardized reporting of inference evaluation. Therefore, we permit the developer to specify the task and hardware setup, as long as both are clear. For example, the specific task might be generating 100k tokens as 5k sequences of length 20 and the fixed set of hardware might be 8 NVIDIA A100s. The hardware in this evaluation need not be the hardware the developer uses for inference if it in fact does any inference itself.
    - References: MLPerf Inference Benchmark, Cheaply Evaluating Inference Efficiency Metrics for Autoregressive Transformer APIs

    Yes

66. Downstream → Distribution → Release decision-making
    - Definition: Is the developer's protocol for deciding whether or not to release a

model disclosed?

- Notes: We recognize that the release of a foundation model falls along a spectrum, with many forms of partial release, and that different developers may conceptualize release differently. We will award this point for any clear protocol that discusses the decision-making process, including if the protocol is more general to the developer rather than the specific foundation model under consideration.
- References: The Gradient of Generative AI Release: Methods and Considerations, The Time Is Now to Develop Community Norms for the Release of Foundation Models

Yes - in risks

67. Downstream → Distribution → Release process

• Definition: Is a description of the process of how the model was released disclosed? • Notes: A description of the release process might include information about who received access to the model at what stage of the release of the model. For example, a developer might conduct a staged release where it releases the model to a select group at first and subsequently makes the model more widely available. We recognize that the release of a foundation model falls along a spectrum, with many different forms of release, and that different developers may conceptualize release differently. We will award this point for any detailed discussion of the release process, including if the discussion is more general to the developer rather than the specific foundation model under consideration. • References: The Gradient of Generative AI Release: Methods and Considerations, The Time Is Now to Develop Community Norms for the Release of Foundation Models

Yes - In risks

68. Downstream → Distribution → Distribution channels

- Definition: Are all distribution channels disclosed?
- Notes: By distribution channel, we mean any pathway by which the model is made accessible to entities beyond the developer. We recognize that distribution channels may arise without the knowledge of the model developer. For example, the weights of a model may be released through one distribution channel and then be distributed through other channels. We will award this point if the developer discloses all of the distribution channels of which it is aware.
- References: Understanding accountability in algorithmic supply chains, Thinking Upstream: Ethics and Policy Opportunities in AI Supply Chains

Yes - HuggingFace

69. Downstream → Distribution → Products and services

- Definition: Does the developer disclose whether any products and services offered by the developer are dependent on the model?
- Notes: We recognize that a developer may provide many products and services that depend on a foundation model or internal derivatives of the model. We will award this point for a reasonable best-effort description of any ways the developer makes internal use of the model in its products or services.

• References: Understanding accountability in algorithmic supply chains, On AI Deployment: AI supply chains (and why they matter)

Yes - No products as of yet are based on Croissant

70. Downstream → Distribution → Detection of machine-generated content • Definition: Are any mechanisms for detecting content generated by this model disclosed? • Notes: Such a mechanism might include storing a copy of all outputs generated by the
model to compare against, implementing a watermark when generating content using the model, or training a detector post-hoc to identify such content. We will award this point if any such mechanism is disclosed or if the developer reports that it has no such mechanism.
• References: A Watermark for Large Language Models, Robust Distortion-free Watermarks for Language Models

No - Not yet

71. Downstream → Distribution → Model License
• Definition: Is a license for the model disclosed?
• Notes: In the event that licenses are written more generally, it should be clear which assets they apply to. We recognize that different developers may adopt different business models and therefor have different types of model licenses. Examples of model licenses include responsible AI licenses, open-source licenses, and licenses that allow for commercial use.
• References: Stronger Together: on the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML, An investigation of licensing of datasets for machine learning based on the GQM model

Yes - MIT license

72. Downstream → Distribution → Terms of service
• Definition: Are terms of service disclosed for each distribution channel? • Notes: We will award this point if there are terms-of-service that appear to apply to the bulk of the model's distribution channels.
• References: Terms-we-Serve-with: a feminist-inspired social imaginary for improved trans parency and engagement in AI, Identifying Terms and Conditions Important to Consumers using Crowdsourcing

Yes MIT license

73. Downstream → Usage policy → Permitted and prohibited users
• Definition: Is a description of who can and cannot use the model disclosed? • Notes: Such restrictions may relate to countries (e.g. US-only), organizations (e.g. no competi tors), industries (e.g. no weapons industry users) or other relevant factors. These restrictions on users are often contained in multiple policies; we group them here for simplicity. We will awarded this point for a clear description of permitted, restricted, and prohibited users of the model.
• References: Best Practices for Deploying Language Models, Meta Platform Terms

Yes - no restrictions

74. Downstream → Usage policy → Permitted, restricted, and prohibited uses • Definition: Are permitted, restricted, and prohibited uses of the model disclosed? • Notes: We will award this point if at least two of the following three categories are disclosed:

(i) permitted uses, (ii) restricted uses, and (iii) prohibited uses. By restricted uses, we mean uses that require a higher level of scrutiny (such as permission from or a separate contract with the developer) to be permitted. These uses are generally included in an acceptable use policy, model license, or usage policy.

• References: Best Practices for Deploying Language Models, Meta Platform Terms

Yes - No restrictions so all uses are possible

75. Downstream → Usage policy → Usage policy enforcement

- Definition: Is the enforcement protocol for the usage policy disclosed?
- Notes: By enforcement protocol, we refer to (i) mechanisms for identifying permitted and prohibited users, (ii) mechanisms for identifying permitted/restricted/prohibited uses, (iii) steps the developer takes to enforce its policies related to such uses, and (iv) the developer's procedures for carrying out these steps. We will award this point for a reasonable best-effort attempt to provide the bulk of this information, though one line indicating the developer reserves the right to terminate accounts is insufficient. Alternatively, we will award this point if the developer reports that it does not enforce its usage policy.
- References: Best Practices for Deploying Language Models, Meta Platform Terms

Yes - no usage policy so nothing to enforce

76. Downstream → Usage policy → Justification for enforcement action • Definition: Do users receive a justification when they are subject to an enforcement action for violating the usage policy?

- Notes: For example, does the developer disclose a protocol for telling users which part of the usage policy they violated, when they did so, and what specifically was violative? Enforcement actions refer to measures to limit a user's ability to use the model, such as banning a user or restricting their ability to purchase tokens. We will award this point if the developer discloses that it gives justification for enforcement actions or, alternatively, if it discloses that it does not provide justification for enforcement actions or that it does not enforce its usage policy.
- References: Best Practices for Deploying Language Models, Meta Platform Terms

Yes - no usage policy so nothing to enforce

77. Downstream → Usage policy → Usage policy violation appeals mechanism • Definition: Is a mechanism for appealing potential usage policy violations disclosed?

• Notes: We will award this point if the developer provides a usage policy violation appeals

mechanism, regardless of whether it is provided via a user interface or distribution channel. • References: Best Practices for Deploying Language Models, Meta Platform Terms

Yes - no usage policy so nothing to enforce

78. Downstream → Model behavior policy → Permitted, restricted, and prohibited model behaviors

- Definition: Are model behaviors that are permitted, restricted, and prohibited disclosed? • Notes: We refer to a policy that includes this information as a model behavior policy, or a developer's policy on what the foundation model can and cannot do (e.g. such a policy may prohibit a model from generating child sexual abuse material). We recognize that different developers may adopt different business models and that some business models may make enforcement of a model behavior policy more or less feasible. We will award this point if at least two of the three categories (i.e. permitted, restricted, and prohibited model behaviors) are disclosed. Alternatively, we will award this point if the developer reports that it does not impose any restrictions on its model's behavior.
- References: I'm Afraid I Can't Do That: Predicting Prompt Refusal in Black-Box Generative Language Models, Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

Yes - no model behavior policy so nothing to enforce

79. Downstream → Model behavior policy → Model behavior policy enforcement • Definition: Is the enforcement protocol for the model behavior policy disclosed? • Notes: By enforcement protocol, we refer to mechanisms for identifying whether model

behavior is permitted or prohibited and actions that may arise in the event the model behavior policy is violated. For example, the developer may make updates to the model in response to issues with the model's adherence to the model behavior policy. We will award this point if there is a clear description of the enforcement protocol, or if the developer reports that it does not enforce its model behavior policy or that it has no such restrictions on the model's behavior.

- References: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

Yes - no model behavior policy so nothing to enforce

80. Downstream → Model behavior policy → Interoperability of usage and model behavior policies

- Definition: Is the way that the usage policy and the model behavior policy interoperate disclosed?
- Notes: For example, if a user attempts to use the model for a prohibited use such as spam, how does the model behavior policy apply if at all? We will also award this point if the developer reports that it does not impose any restrictions on its model's behavior in the event of usage policy violation.

- References: I'm Afraid I Can't Do That: Predicting Prompt Refusal in Black-Box Generative Language Models, Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

Yes - no model behavior policy so nothing to enforce

81. Downstream → User Interface → User interaction with AI system • Definition: For distribution channels with user-facing interfaces, are users notified (i) that they are interacting with an AI system, (ii) of the specific foundation model they are interacting with, and (iii) that outputs are machine-generated?
    - Notes: A user-facing interface refers to the means by which the user interacts with the foundation model, including how the user can observe outputs from the foundation model and other notifications. We will award this point if, for all distribution channels with user facing interfaces, the user is provided adequate transparency as to the foundation model being distributed and the potential presence of any model outputs.
    - References: Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges, Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness

Yes on HF platform

82. Downstream → User Interface → Usage disclaimers
    - Definition: For distribution channels with user-facing interfaces, are users provided with disclaimers involving model use?
    - Notes: A user-facing interface refers to the means by which the user interacts with the foundation model, including how the user can observe outputs from the foundation model and other notifications. Usage disclaimers could include information about what constitutes a usage policy violations or how users should interpret model outputs. We will award this point if, for all distribution channels with user-facing interfaces, the user is provided with usage disclaimers.
    - References: Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges, Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness

Yes - Disclaimers are added to the model card on the HF plaform

83. Downstream → User data protection → User data protection policy • Definition: Are the protocols for how the developer stores, accesses, and shares user data disclosed?
    - Notes: We will also award this point if the developer reports that it has no user data protection policy.
    - References: Privacy as Contextual Integrity, Redesigning Data Privacy: Reimagining Notice Consent for human technology interaction

Yes - No user facing channel associated with model explictky

84. Downstream → User data protection → Permitted and prohibited use of user

data • Definition: Are permitted and prohibited uses of user data disclosed?

- Notes: Developers use user data for a range of purposes such as building future models, updating existing models, and evaluating both existing and future models. We will award this point if a developer discloses its policy on the use of user data from interactions associated with this model, including both permitted and prohibited uses. This may span different distribution channels if multiple channels supply user data to the developer. Alternatively, we will award this point if the developer reports it does not impose any limits on its use of user data.
- References: Privacy as Contextual Integrity, Redesigning Data Privacy: Reimagining Notice Consent for human technology interaction

Yes - No user data is stored

85. Downstream → User data protection → Usage data access protocol • Definition: Is a protocol for granting external entities access to usage data disclosed? • Notes: Usage data refers to the data created through user interaction with the model, such as user inputs to the model and associated metadata such as the duration of the interaction. A usage data access protocol refers to the steps, requirements, and considerations involved in granting external entities access to usage data; this goes beyond stating the conditions under which related personal information may be shared with external entities. We will award this point for a clear description of the usage data access protocol or if the developer reports it does not share usage data with external entities.
    - References: How Cambridge Analytica Sparked the Great Privacy Awakening, Redesigning Data Privacy: Reimagining Notice Consent for human technology interaction

86. Downstream → Model Updates → Versioning protocol
    - Definition: Is there a disclosed version and versioning protocol for the model? • Notes: By versioning, we mean that each instance of the model is uniquely identified and that the model is guaranteed to not change when referring to a fixed version number; alternatively, the version clearly indicating a specific instance of the model may be able to change by noting that it is the "latest" or an "unstable" version. We recognize that different developers may adopt different versioning practices that may differ from standard semantic versioning practices used elsewhere in software engineering.
    - References: How is ChatGPT's behavior changing over time?, Putting the Semantics into Semantic Versioning

Yes - version control is explicited in the paper

87. Downstream → Model Updates → Change log
    - Definition: Is there a disclosed change log for the model?
    - Notes: By change log, we mean a description associated with each change to the model (which should be indicated by a change in version number). We recognize that different developers may adopt different practices for change logs that may differ from practices used elsewhere in software engineering. We will award this point if the change log provides a clear description of changes that is legible to a technical audience.

- References: How is ChatGPT's behavior changing over time?, Watch out for This Commit! A Study of Influential Software Changes

Yes - Change log in the model card on HF

88. Downstream → Model Updates → Deprecation policy
    - Definition: Is there a disclosed deprecation policy for the developer?
    - Notes: By deprecation policy, we refer to a description of what it means for a model to be deprecated and how users should respond to the deprecation (e.g. instructions to migrate to a newer version). We will award this point for a clear disclosure of a deprecation policy or if there is no risk of deprication (e.g. if the developer openly releases model weights).
    - References: How is ChatGPT's behavior changing over time?,Automatic Android Deprecated API Usage Update by Learning from Single Updated Example

Yes - the open source nature of the model implies no deprecation risk

89. Downstream → Feedback → Feedback mechanism
    - Definition: Is a feedback mechanism disclosed?
    - Notes: By feedback mechanism, we refer to a means for external entities to report feedback or issues that arise in relation to the foundation model. Such entities may include but are not necessarily limited to users. We will award this point if the developer discloses a feedback mechanism that has been implemented.
    - References: Ecosystem Graphs: The Social Footprint of Foundation Models, Outsider Over sight: Designing a Third Party Audit Ecosystem for AI Governance

Yes - email to authors, flagging on HF. Authors indicate this in the report.

90. Downstream → Feedback → Feedback summary
    - Definition: Is a report or summary disclosed regarding the feedback the developer received or, alternatively, the way the developer responded to that feedback?
    - Notes: We recognize that there does not exist an authoritative or consensus standard for what is required in a feedback report. For this reason, we will award this point if there is a meaningful, though potentially vague or incomplete, summary of feedback received.
    - References: Achieving Transparency Report Privacy in Linear Time, Evaluating a Method ology for Increasing AI Transparency: A Case Study

No - Not yet

91. Downstream → Feedback → Government inquiries
    - Definition: Is a summary of government inquiries related to the model received by the developer disclosed?
    - Notes: Such government inquiries might include requests for user data, requests that certain content be banned, or requests for information about a developer's business practices. We recognize that there does not exist an authoritative or consensus standard for what is required for such a summary of government inquiries. For this reason, we will award this point if (i) there is a meaningful,

though potentially vague or incomplete, summary of government inquiries, or (ii) a summary of government inquiries related to user data.
- References: Transparency Report: Government requests on the rise, Ecosystem Graphs: The Social Footprint of Foundation Models

Yes - No government inquiries as of yet

92. Downstream → Impact → Monitoring mechanism
- Definition: For each distribution channel, is a monitoring mechanism for tracking model use disclosed?
- Notes: By monitoring mechanism, we refer to a specific protocol for tracking model use that goes beyond an acknowledgement that usage data is collected. We will also award this point for a reasonable best-effort attempt to describe monitoring mechanisms, or if a developer discloses that a distribution channel is not monitored.
- References: Progressive Disclosure: Designing for Effective Transparency, Ecosystem Graphs: The Social Footprint of Foundation Models

Yes - Distribution channels are not monitored

93. Downstream → Impact → Downstream applications
- Definition: Across all forms of downstream use, is the number of applications dependent on the foundation model disclosed?
- Notes: We recognize that there does not exist an authoritative or consensus standard for what qualifies as an application. We will award this point if there is a meaningful estimate of the number of downstream applications, along with some description of what it means for an application to be dependent on the model.
- References: Market concentration implications of foundation models: The Invisible Hand of ChatGPT, Ecosystem Graphs: The Social Footprint of Foundation Models

94. Downstream → Impact → Affected market sectors
- Definition: Across all downstream applications, is the fraction of applications corresponding to each market sector disclosed?
- Notes: By market sector, we refer to an identifiable part of the economy. While established standards exist for describing market sectors, we recognize that developers may provide vague or informal characterizations of market impact. We will award this point if there is a meaningful, though potentially vague or incomplete, summary of affected market sectors.
- References: Market concentration implications of foundation models: The Invisible Hand of ChatGPT, Ecosystem Graphs: The Social Footprint of Foundation Models

No

95. Downstream → Impact → Affected individuals
- Definition: Across all forms of downstream use, is the number of individuals affected by the foundation model disclosed?
- Notes: By affected individuals, we principally mean the number of potential users of appli cations. We recognize that there does not exist an authoritative or consensus standard for what qualifies as an affected individual. We will award this point if there is a meaningful estimate of the number of affected individuals along with a

clear description of what it means for an individual to be affected by the model.
- References: Market concentration implications of foundation models: The Invisible Hand of ChatGPT, Ecosystem Graphs: The Social Footprint of Foundation Models

No - Although the model is usable and interesting for a large share of the french speakers of the world with a mobile phone's worth of compute.

96. Downstream → Impact → Usage reports
    - Definition: Is a usage report that gives usage statistics describing the impact of the model on users disclosed?
    - Notes: We recognize that there does not exist an authoritative or consensus standard for what is required in a usage report. Usage statistics might include, for example, a description of the major categories of harm that has been caused by use of the model. We will award this point if there is a meaningful, though potentially vague or incomplete, summary of usage statistics.
    - References: Expert explainer: Allocating accountability in AI supply chains, Ecosystem Graphs: The Social Footprint of Foundation Models

    No, not at the moment

97. Downstream → Impact → Geographic statistics
    - Definition: Across all forms of downstream use, are statistics of model usage across geogra phies disclosed?
    - Notes: We will award this point if there is a meaningful, though potentially incomplete or vague, disclosure of geographic usage statistics at the country-level.
    - References: Expert explainer: Allocating accountability in AI supply chains, Ecosystem Graphs: The Social Footprint of Foundation Models

    No, not at the moment and difficult to compute since usage and users are not tracked. However, we posit french-speaking countries would be the main geographical user bases.

98. Downstream → Impact → Redress mechanism
    - Definition: Is any mechanism to provide redress to users for harm disclosed? • Notes: We will also award this point if the developer reports it does not have any such redress mechanism.
    - References: Computational Power and AI, Ecosystem Graphs: The Social Footprint of Foundation Models

    Yes - No such mechanism as indicated in the report

99. Downstream → Documentation for Deployers → Centralized documentation for down stream use
    - Definition: Is documentation for downstream use centralized in a centralized artifact? • Notes: Centralized documentation for downstream use refers to an artifact, or closely-linked artifacts, that consolidate relevant information for making use of or repurposing the model. Examples of these kinds of artifacts include a website with dedicated documentation infor mation, a github repository with dedicated documentation information, and an ecosystem card. We recognize that different developers may take different approaches to centralizing information. We will award

this point if there is a clearly-identified artifact(s) that contains the majority of substantive information (e.g. capabilities, limitations, risks, evaluations, distribution channels, model license, usage policies, model behavior policies, feedback and redress mechanisms, dependencies).
• References: Datasheets for Datasets, Model Cards for Model Reporting

Yes - Code and documentation is available on Github and HuggingFace

100. Downstream → Documentation for Deployers → Documentation for responsible down stream use

• Definition: Is documentation for responsible downstream use disclosed? • Notes: Such documentation might include details on how to adjust API settings to promote responsible use, descriptions of how to implement mitigations, or guidelines for responsible use. We will also award this point if the developer states that it does not provide any such documentation. For example, the developer might state that the model is offered as is and downstream developers are accountable for using the model responsibly.
• References: Ecosystem Graphs: The Social Footprint of Foundation Models, Expert explainer: Allocating accountability in AI supply chains

Yes - The model is offered as is and downstream developers are accountable for using the model responsibly, although usage examples are provided.