**Meta (LLaMa) - total: 20/48 (likely change to 21/48 bc compute oversight)**

1. Data sources
   * 4 points: Specific datasets and their absolute sizes are disclosed (CommonCrawl, C4, Github, Wikipedia, Gutenberg and Books3, ArXiv, StackExchange) in section 2.1 of the LLaMa technical paper on pre-training data.
   * Sources: <https://arxiv.org/pdf/2302.13971.pdf>
2. Data governance
   * 2 points: The LLaMa technical paper underscores that it is trained using only publicly available datasets, but does not outline biases in each dataset, steps to mitigate bias in each dataset, or exclusion of data without adequate governance. Lines relevant to data governance include: (1) “Large language models have been showed to reproduce and amplify biases that are existing in the training data (Sheng et al., 2019; Kurita et al., 2019), and to generate toxic or offensive content (Gehman et al., 2020). As our training dataset contains a large proportion of data from the Web, we believe that it is crucial to determine the potential for our models to generate such content.” (2) “Our model is particularly biased in the religion category (+10% compared to OPT-175B), followed by age and gender. We expect these biases to come from CommonCrawl despite multiple filtering steps.” (3) “We only kept [GitHub] projects that are distributed under the Apache, BSD and MIT licenses.”
   * Source: <https://arxiv.org/pdf/2302.13971.pdf>
     1. Note: with more details on filtering with an eye towards mitigating bias, this would be 3 points. Closest is “We filtered the data from the Web based on its proximity to Wikipedia text and references. For this, we used a Kneser-Ney language model and a fastText linear classifier” in the Mitigation section of the model card, but no details are provided as to how or why or what this mitigates.
3. Copyrighted data
   * 0 points: No description of copyrighted training data.
   * Source: <https://arxiv.org/pdf/2302.13971.pdf> ; <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md>
     1. Note: could give 1 point for filtering GitHub data “We only kept projects that are distributed under the Apache, BSD and MIT licenses.” But no acknowledgement of copyright considerations.
4. Compute (additive)
   * 3 points: Model sizes (7B, 13B, 33B, 65B), hardware units (A100-80GB), and training time (GPU hours per model size) are disclosed. Was unable to find details on FLOPs, but likely my oversight.
   * Source: <https://arxiv.org/pdf/2302.13971.pdf> Section 6
     1. Note: paper focuses on single GPU implementation, so does not give number of hardware units.
     2. Note: Can’t find detail on of FLOPs in technical paper. Unclear to me if the table on page 3 is sufficient detail.
5. Energy (additive)
   * 4 points: energy usage, power consumption, and measurement strategy for each model is disclosed. Only mitigation to reduce energy use mentioned is models of reduced size, but other areas are sufficiently detailed to warrant full points.
   * Source: <https://arxiv.org/pdf/2302.13971.pdf> Section 6
6. Capabilities and limitations
   * 2 points: Some grounded discussion of specific capabilities and limitations. Press release indicates assessments of bias, toxicity, and misinformation in the technical paper are indicative of limitations: “We also provide in the paper a set of evaluations on benchmarks evaluating model biases and toxicity to show the model’s limitations and to support further research in this crucial area.” Model card states “The primary use of LLaMA is research on large language models, including: exploring potential applications such as question answering, natural language understanding or reading comprehension, understanding capabilities and limitations of current language models, and developing techniques to improve those, evaluating and mitigating biases, risks, toxic and harmful content generations, hallucinations….Risks and harms of large language models include the generation of harmful, offensive or biased content. These models are often prone to generating incorrect information, sometimes referred to as hallucinations. We do not expect our model to be an exception in this regard.” Evaluations and specific examples of capabilities and limitations are lacking, as are detailed descriptions of out of scope uses beyond “downstream applications without further risk evaluation and mitigation”
   * Sources: <https://ai.facebook.com/blog/large-language-model-llama-meta-ai/> ; <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md> ; <https://arxiv.org/pdf/2302.13971.pdf>
7. Risks and mitigations (additive)
   * 1 point: Risks are listed, but mitigations (other than some filtering, which is not described in sufficient detail), effectiveness of mitigations, and non-mitigated risks are not listed. Model card states “Risks and harms: Risks and harms of large language models include the generation of harmful, offensive or biased content. These models are often prone to generating incorrect information, sometimes referred to as hallucinations. We do not expect our model to be an exception in this regard.”
   * Sources: <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md> ; <https://arxiv.org/pdf/2302.13971.pdf>
8. Evaluations
   * 2 points: Extensive evaluation of accuracy with respect to common sense reasoning, closed-book question answering, mathematical reasoning, reading comprehension, code generation, and other tasks. From the model card: “The model was evaluated on the following benchmarks: BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC, OpenBookQA, NaturalQuestions, TriviaQA, RACE, MMLU, BIG-bench hard, GSM8k, RealToxicityPrompts, WinoGender, CrowS-Pairs.” Unintentional harm is evaluated using RealToxicityPrompts for toxicity, WinoGender and CrowS-Pairs for bias, and TruthfulQA for unintended misinformation, per the technical paper. Unintentional harm, robustness, and other factors are not evaluated.
   * Sources: <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md> ; <https://arxiv.org/pdf/2302.13971.pdf>
9. Testing (additive)
   * 0 points: No description of internal testing other than using standard public benchmarks. No description of external evaluation or red teaming.
   * Sources: <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md> ; <https://arxiv.org/pdf/2302.13971.pdf>
10. Machine-generated content
    * 0 points: No discussion of machine-generated content or policy indicating Meta opposes or prohibits users from misleading others about whether LLaMa-generated content is in fact machine generated. No attempt at watermarking; model is openly released.
    * Source: LLaMa license, model card, and technical paper. <https://docs.google.com/forms/d/e/1FAIpQLSfqNECQnMkycAp2jP4Z9TFX0cGR4uf7b_fBxjY_OjhJILlKGA/viewform> ; <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md> ; <https://arxiv.org/pdf/2302.13971.pdf>
11. Member states
    * 0 points: No description of deployment practices in relation to the EU.
    * Sources: <https://docs.google.com/forms/d/e/1FAIpQLSfqNECQnMkycAp2jP4Z9TFX0cGR4uf7b_fBxjY_OjhJILlKGA/viewform>
12. Downstream documentation
    * 2 points: Model card provides relevant documentation, though it provides relatively generic guidance for downstream providers. No other clear mechanism for providing appropriate documentation to downstream providers, and no usage policy.
    * Sources: <https://docs.google.com/forms/d/e/1FAIpQLSfqNECQnMkycAp2jP4Z9TFX0cGR4uf7b_fBxjY_OjhJILlKGA/viewform> ; <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md>