**Meta (LLaMA) - total: 25/48**

1. Data sources
   * 4 points: Dataset pipeline is disclosed: training on the corpus described in Section 2.1 and Table 1, with fine-grained attribution to sources (e.g. Wikipedia, GitHub, Stack Exchange) and associated data sizes (relative and absolute)
   * Sources: <https://arxiv.org/pdf/2302.13971.pdf> Section 2.1
2. Data governance
   * 3 points: The subject of data governance and curation is raised (“we use data sources that has been leveraged to train other LLMs, with the restriction of only using data that is publicly available, and compatible with open sourcing”) with some exclusion described (i.e. licensing for GitHub repositories), though not fully indicating that governance practices are required for data inclusion.
   * Sources: <https://arxiv.org/pdf/2302.13971.pdf> Section 2.1
3. Copyrighted data
   * 0 points: Copyright is not discussed.
   * Source: <https://arxiv.org/pdf/2302.13971.pdf>
4. Compute (additive)
   * 4 points: The model size, compute requirements (training time, number of A100s), FLOPs, and associated measurement methodology are described.
   * Source: <https://arxiv.org/pdf/2302.13971.pdf> Section 6
5. Energy (additive)
   * 4 points: The energy usage is disclosed, measurement methodology, and the emissions if the model were trained using the carbon intensity factor aligned with the US national average are disclosed. While no mitigations are described to explicitly reduce energy usage, efficiency improvements are described (e.g. reducing runtime through improved implementation of causal multi-head attention and reducing number of activations computed in backward pass).
   * Source: <https://arxiv.org/pdf/2302.13971.pdf> Section 6 and 2.4
6. Capabilities and limitations
   * 3 points: Capabilities are described (“As a foundation model, LLaMA is designed to be versatile and can be applied to many different use cases”) with specific use cases (“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”) . Several limitations (e.g. “There is still more research that needs to be done to address the risks of bias, toxic comments, and hallucinations in large language models. Like other models, LLaMA shares these challenges.”), largely aligning with the literature on language model limitations. Lacks grounding to evaluation/use cases.
   * Sources: <https://ai.facebook.com/blog/large-language-model-llama-meta-ai/>, <https://arxiv.org/pdf/2302.13971.pdf>, and <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md>
7. Risks and mitigations (additive)
   * 1 point: Risks are listed (e.g. “our model has not been trained with human feedback, and can thus generate toxic or offensive content, incorrect information or generally unhelpful answers”) though no significant mitigations (exception: minimal data filtering) are taken as disclosed in the model card.
   * Source: <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md>
8. Evaluations (additive)
   * 2 points: Extensive evaluation for accuracy across many tasks/datasets. Evaluation of unintentional harm (toxicity on RealToxicityPrompts, bias on CrowS-Pairs and WinoGender, truthfulness/unintended misinformation on TruthfulQA). No evaluation of malicious use or other desiderata (e.g. robustness, calibration).
   * Source: <https://arxiv.org/pdf/2302.13971.pdf> and <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md>
9. Testing (additive)
   * 0 points: No discussion of internal testing beyond standard public benchmarks. No discussion of external evaluation (e.g. HELM) or redteaming (e.g. ARC).
   * Source: <https://arxiv.org/pdf/2302.13971.pdf> and <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md>
10. Machine-generated content
    * 0 points: No discussion of disclosure of content being machine-generated. Given model is openly released, no enforcement mechanism is discussed as well, including in LLaMA license.
    * Source: <https://arxiv.org/pdf/2302.13971.pdf> and <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md>
11. Member states
    * 2 points: Restrictions on US export control countries specified in license but no further restrictions relevant for EU member states. No discussion of how the models are deployed across different EU member states.
    * Sources: <https://docs.google.com/forms/d/e/1FAIpQLSfqNECQnMkycAp2jP4Z9TFX0cGR4uf7b_fBxjY_OjhJILlKGA/viewform>
12. Downstream documentation
    * 2 points: Some documentation in the form of a model card, though no discussion of providing downstream and the model card may not all have all relevant information for downstream compliance (i.e. fairly minimal content, much of which is generic, beyond the evaluation results).
    * Source: <https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md>