**EleutherAI (GPT-NeoX) - total:**

**Rubric.**

1. Data sources
   * 4 points: GPT-NeoX is trained on The Pile, which is described in minute detail in the technical paper on The Pile and its datasheet.
   * Sources: <https://arxiv.org/abs/2201.07311>; <https://arxiv.org/pdf/2101.00027.pdf>; <https://huggingface.co/EleutherAI/gpt-neox-20b#training-dataset>
2. Data governance
   * 3 points: Several data governance issues in The Pile dataset are raised in the technical paper, including sexually explicit content, bias related to gender, religion, and
   * Sources: <https://arxiv.org/pdf/2101.00027.pdf> ; <https://arxiv.org/pdf/2201.07311.pdf>
     1. Note: could have given 2 points. Unclear how effective the mitigations are, and its not clear that every dataset is subject to appropriate data governance measures
3. Copyrighted data
   * 4 points: The Pile datasheet gives a good description of which datasets have copyrighted materials, suggests that the use of this material for training LLMs is permitted under US law, and suggests users in other jurisdictions remix the datasets to avoid violating copyright law. The Pile technical paper gives an excellent breakdown page 14-15 of the different types of consent for each dataset.
   * Sources: <https://arxiv.org/pdf/2101.00027.pdf> ; <https://arxiv.org/pdf/2201.07311.pdf>
     1. Note: Could have been a 4+: other sections of The Pile datasheet go into painstaking detail while this one only gives a couple examples of datasets with copyrighted material and does not give any numerical estimates of copyrighted materials (e.g., the Washington Post [used](https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/) keywords).
4. Compute
   * 4 points: model size (20 billion parameters), training time (1830 hours), number of hardware units (96 A100s, 24 AMD EPYC 7532 CPUs), FLOPs (117 teraFLOPs per GPU) are disclosed.
   * Source: <https://arxiv.org/pdf/2204.06745.pdf>
5. Energy
   * 4 points: energy use (66.25MWh), emissions (31.73 metric tons of CO2), and measurement strategy (systems were based in Illinois, carefully tracked emissions throughout the process) are disclosed in the technical paper. No details on mitigation steps, but due to elaborate detail on other dimensions and use of “a codebase that builds on Megatron (Shoeybi et al., 2020) and DeepSpeed (Rasley et al., 2020) to facilitate efficient and straightforward training of large language models”, this warrants full points.
   * Sources: <https://arxiv.org/pdf/2204.06745.pdf>
     1. Note: this could have been a 3. They don’t give details on the methodology behind “Throughout the development and training of our model, we tracked our energy usage and carbon emissions.”
6. Capabilities and limitations
   * 3 points: Capabilities and limitations are described in detail, though there is room for improvement in grounding them in evaluations and specific examples. While the model card for GPT-NeoX is crystal clear about what the model can and cannot do, the technical paper for GPT-NeoX and The Pile do not evaluate the impact of limitations or provide a great number of examples.
   * Sources: [https://huggingface.co/EleutherAI/gpt-neox-20b](https://huggingface.co/EleutherAI/gpt-neox-20b#training-dataset) ; <https://arxiv.org/pdf/2204.06745.pdf>
7. Risks and mitigations
   * 1 point: risks are listed (implicit reference to factual inaccuracy in the model card, biases in The Pile listed in its technical paper) but there is no detail on mitigations. There is no usage policy, the model card simply says “please note that you need to conduct your own risk and bias assessment”
   * Source: <https://arxiv.org/pdf/2204.06745.pdf> ; <https://arxiv.org/pdf/2101.00027.pdf> ; [https://huggingface.co/EleutherAI/gpt-neox-20b](https://huggingface.co/EleutherAI/gpt-neox-20b#training-dataset)
8. Evaluations (additive)
   * 1 point: Evaluation for accuracy using EleutherAI Language Model Evaluation Harness, but no evaluation for bias, malicious use, robustness, or other factors.
   * Source: <https://arxiv.org/pdf/2204.06745.pdf>
9. Testing (additive)
   * 0 points: No disclosure of internal testing, no mention of external evaluations or red teaming.
   * Source: <https://arxiv.org/pdf/2204.06745.pdf>
10. Machine-generated content
    * 2 points: Model card states “Please inform your audience that you are using artificially generated text.” No watermarking, usage policy, mechanism for enforcement, or guidelines are provided.
      1. Note: I am tempted to give only 1 point. This provision is not part of a broader usage policy, making it even more unenforceable, it does not give examples of unacceptable machine-generated uses, it does not specify how to inform the audience, etc.
11. Member states
    * 0 points: No description of deployment practices in relation to the EU.
    * Source: <https://arxiv.org/pdf/2204.06745.pdf> ; [https://huggingface.co/EleutherAI/gpt-neox-20b](https://huggingface.co/EleutherAI/gpt-neox-20b#training-dataset)
12. Downstream documentation
    * 3 points: Extensive documentation on the model and data is provided via HuggingFace, technical papers, and an external [website](https://the-eye.eu/). No usage policy, however.
    * Source: [https://huggingface.co/EleutherAI/gpt-neox-20b](https://huggingface.co/EleutherAI/gpt-neox-20b#training-dataset)
      1. Note: I do not believe that there is adequate documentation for compliance with the provisions of the AIA without a usage policy.