**OpenAI (GPT-4) - total: 27/48**

1. Data sources
   * 1 point: Very generic and vague description: “GPT-4 is a Transformer-style model pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers….Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.”
   * Source: <https://arxiv.org/pdf/2303.08774.pdf> section 1
     1. Note: Could give 0 points for being maximally vague.
2. Data governance
   * 2 points: Grounded discussion of data governance related to erotic content in training data: “At the pre-training stage, we filtered our dataset mix for GPT-4 to specifically reduce the quantity of inappropriate erotic text content. We did this via a combination of internally trained classifiers and a lexicon-based approach to identify documents that were flagged as having a high likelihood of containing inappropriate erotic content. We then removed these documents from the pre-training set.” No other discussion of protocols related to data governance, bias mitigation, or exclusion of datasets with inadequately mitigated biases.
   * Source: <https://arxiv.org/pdf/2303.08774.pdf> system card section 3.1
3. Copyrighted data
   * 0 points: No description.
   * Source: <https://arxiv.org/pdf/2303.08774.pdf>
4. Compute
   * 0 points: Model size, training time, number and type of hardware units, and FLOPs are not measured. Microsoft is implied to be the compute provider.
   * Source: <https://arxiv.org/pdf/2303.08774.pdf>
5. Energy
   * 0 points: Energy usage, emissions, measurement strategy, and energy reduction measures are not described.
   * Source: <https://arxiv.org/pdf/2303.08774.pdf>
6. Capabilities and limitations
   * 4 points: Fine-grained discussion of a wide variety of capabilities (answering multiple choice questions across subjects and languages, commonsense reasoning, coding, arithmetic) tied to standard public benchmarks as well as specific examples in the form of standardized tests. Limitations include hallucinations (evaluated by an internal factual test for accuracy across different subjects and tasks as well as standard public benchmarks) no knowledge of events after September 2021, simple reasoning errors, gullibility, and overconfidence in incorrect answers; one incorrect example on TruthfulQA is given.
   * Source: <https://arxiv.org/pdf/2303.08774.pdf> sections 4 and 5
7. Risks and mitigations
   * 4 points: Extensive list of risks (hallucinations, harmful content, harms of representation, allocation, and quality of service, disinformation and influence operations, proliferation of conventional and unconventional weapons, privacy, cybersecurity, potential for risky emergent behaviors, interactions with other systems, economic impacts, acceleration, overreliance) and detailed subsections on each in the system card. Mitigations include model refusals “aimed to reduce the tendency of the model to produce such harmful content”, red teaming, “documentation and hedging within the model”, “we have trained models to refuse malicious cybersecurity requests, and scaled our internal safety systems, including in monitoring, detection and response”, usage policies that prohibit the use of OpenAI’s models in high-risk contexts, and investing in efforts to further study risks. The methodology for implementing mitigations through filtering, RHLF, steering GPT-4 using reward models, and other measures is disclosed. A large number of examples of how mitigations impacted responses to specific prompts are provided, as is measurement of how incorrect behavior rate fell (per internal testing) in the wake of mitigation measures and accuracy on adversarial questions improved (per TruthfulQA) after mitigations. Some justification is provided with respect to difficulties in mitigating certain risks.
   * Source: <https://arxiv.org/pdf/2303.08774.pdf> System card
8. Evaluations
   * 4 points: Accuracy is measured on multiple benchmarks (MMLU, HellaSwag, ARC, WinoGrade, HumanEval, DROP, GSM-8K), unintentional and intentional harms are measured and described in the system card (though there is little quantitative detail regarding potential for malicious use), and calibration and user experience were evaluated as well.
   * Source: <https://arxiv.org/pdf/2303.08774.pdf>
     1. Note: was unable to find evaluation of robustness, but that could be my oversight
9. Testing
   * 3 points: Disclosure of results and process of significant internal testing (e.g., internal factual evaluation by category such as learning, technology, writing, history, math, science, recommendation, code, business) as well as external red teaming, though no external evaluations due to external access (e.g., HELM).
   * Source: <https://arxiv.org/pdf/2303.08774.pdf>
10. Machine-generated content
    * 3 points: GPT-4 often reminds users it is a language model, helping to highlight that its responses are machine-generated. OpenAI’s Usage Policy contains some provisions requiring disclosure of machine-generated content: “Consumer-facing uses of our models in medical, financial, and legal industries; in news generation or news summarization; and where else warranted, must provide a disclaimer to users informing them that AI is being used and of its potential limitations. Automated systems (including conversational AI and chatbots) must disclose to users that they are interacting with an AI system.”
    * Sources: <https://chat.openai.com/> ; <https://openai.com/policies/usage-policies> ; <https://arxiv.org/pdf/2303.08774.pdf>
      1. Note: could have easily given 2 points here. Usage policy is weaker than others, and technical paper does not address machine-generated content.
11. Member states
    * 2 points: Disclosure of EU member states where API access is available. No discussion of specific deployment practices in Europe or specific EU member states.
    * Source: <https://platform.openai.com/docs/supported-countries>
12. Downstream documentation
    * 4 points: Substantial documentation for ensuring downstream providers comply with the AIA, with significant documentation related to fine tuning, safety, production, and plugins.
    * Source: <https://platform.openai.com/docs/introduction>
      1. Note: could have given 3 points. For consistency with PaLM 2, gave 4 though I think PaLM may be slightly better on this count