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

1. Data sources
   * 1 point: Dataset is extremely minimally described as “publicly available data (such as internet data) and data licensed from third-party providers”. The system card vaguely says, but does not confirm, that “GPT models are *often* trained in two stages” (training via next-word prediction, then fine-tuning via RLHF), so we do not give the point (there are similar insinuations elsewhere in the paper), especially given the model is disclosed as text and image, but this does not discuss images. No fine-grained sourcing or sizes.
   * Sources: <https://cdn.openai.com/papers/gpt-4.pdf> Section 2 and System Card Section 1
2. Data governance
   * 2 points: The subject of data governance and curation is raised, with specific practices to mitigate harms (e.g. filtration practices) described in the System card around violations of usage practices (e.g. inappropriate erotic content).
   * Sources: <https://cdn.openai.com/papers/gpt-4.pdf> System Card Section 1.1
3. Copyrighted data
   * 0 points: Copyright is not discussed.
   * Source: <https://cdn.openai.com/papers/gpt-4.pdf>
4. Compute (additive)
   * 0 points: The model size, training time, and compute requirements are not disclosed (this is acknowledged in the paper).
   * Source: <https://cdn.openai.com/papers/gpt-4.pdf>
5. Energy (additive)
   * 0 points: The energy usage, emissions, energy sources, and measurement methodology are not disclosed
   * Source: <https://cdn.openai.com/papers/gpt-4.pdf>
6. Capabilities and limitations
   * 4 points: Capabilities are described (“GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks.”), with several more specific use cases (concrete deployments with clients at https://openai.com/gpt-4). Several limitations (e.g. “it is not fully reliable (e.g. can suffer from “hallucinations”), has a limited context window, and does not learn from experience. Care should be taken when using the outputs of GPT-4, particularly in contexts where reliability is important.”) with an extensive system card and concrete measurements (e.g. comparisons to other models).
   * Sources: <https://cdn.openai.com/papers/gpt-4.pdf> and <https://openai.com/gpt-4> and <https://openai.com/research/gpt-4>
7. Risks and mitigations (additive)
   * 3 points: Risks are listed extensively in System Card (e.g. hallucinations, harmful content, harms of representation/allocation/quality of service, disinformation, weapons, privacy, cybersecurity, emergent behaviors, interactions with other systems, economic impacts, acceleration, overreliance), mitigations are provided in some on a per-category basis in Section 2 (e.g. training on ChatGPT data to reduce hallucinations, refusals to address harmful content and stereotypes). In many cases, examples or evaluations are provided or described. Some risks (e.g. disinformation) are discussed, but no mitigation is given and no clear justification is provided to justify the lack of mitigation, though it is acknowledged that further testing is needed to understand the extent of risk.
   * Source: <https://cdn.openai.com/papers/gpt-4.pdf> System Card
8. Evaluations (additive)
   * 4 points: Extensive evaluation for accuracy across many tasks/datasets. Evaluation of unintentional harm (e.g. Figure 9 on unsafe generations of disallowed and sensitive content). Evaluation of calibration (Figure 8). Evaluation of intentional harm (e.g. Table 6).
   * Source: <https://cdn.openai.com/papers/gpt-4.pdf> Sections 4-6 and System Card
9. Testing (additive)
   * 3 points: Extensive discussion of various forms of internal testing beyond standard public benchmarks (e.g. factuality evaluations in Figure 6). No discussion of external evaluation (e.g. HELM). Discussion of external redteaming (ARC evals).
   * Source: <https://cdn.openai.com/papers/gpt-4.pdf> Sections 4-6 and System Card
10. Machine-generated content
    * 4 points: Clear on ChatGPT interface and if using API that content is machine-generated. Usage policy has a requirements stating “where else warranted, must provide a disclaimer to users informing them that AI is being used and of its potential limitations.” and “Automated systems (including conversational AI and chatbots) must disclose to users that they are interacting with an AI system. With the exception of chatbots that depict historical public figures, products that simulate another person must either have that person's explicit consent or be clearly labeled as “simulated” or “parody.” with a further Sharing and Publication Policy.
    * Source: <https://openai.com/policies/usage-policies>
11. Member states
    * 2 points: Restricted to US embargo countries, with no specific discussion of EU. ChatGPT interface has specific supported countries and iOS app has specific access to 11 listed countries, some of which are in Europe
    * Sources: <https://openai.com/policies/terms-of-use> and <https://help.openai.com/en/articles/6825453-chatgpt-release-notes> and <https://help.openai.com/en/articles/7947663-chatgpt-supported-countries>
12. Downstream documentation
    * 3 points: Extensive documentation of the model in the form of the system card and of its use in API documentation, though mechanism for providers encountering all aspects (e.g. system card from API use) are unclear.
    * Source: <https://cdn.openai.com/papers/gpt-4.pdf>and<https://platform.openai.com/docs/models/gpt-4>