**Stability (Stable Diffusion v2) - 22/48**

1. Data sources
   * 4 points: The dataset (LAION-5B), its size (5.85 billion CLIP-filtered image-text pairs), classifier (NSFW) with sensitivity score ("p\_unsafe" score of 0.1), and number of steps (150k steps then another 140k steps) are disclosed. Further details on the dataset are provided in a technical paper.
   * Source: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training) and <https://arxiv.org/pdf/2210.08402.pdf>
2. Data governance
   * 2 points: Both the Stable Diffusion v2 model card and the LAION-5B paper discuss biases in training data, though only LAION-5B discusses mitigations and those are relatively thin.
   * Sources: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training) and <https://arxiv.org/pdf/2210.08402.pdf>
3. Copyrighted data
   * 0 points: Was unable to find any description of copyrighted data.
   * Sources: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training) and <https://arxiv.org/pdf/2210.08402.pdf>
     1. Note: several examples of misuse listed in the model card are related to copyright infringement
4. Compute
   * 4 points: The model size, number and type of GPUs, training time, compute provider, and methodology for measurement are disclosed in the model card.
   * Source: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training)
5. Energy
   * 3 points: emissions (15,000kg CO2 eq.), measurement strategy (ML Impact Calculator), and compute region (US-east) are disclosed. No mitigation measures are discussed. This section says “Stable Diffusion v1 Estimated Emissions”, potentially casting doubt on its accuracy, though the emissions reported are greater than those reported in the Stable Diffusion v1 [model card](https://github.com/CompVis/stable-diffusion/blob/main/Stable_Diffusion_v1_Model_Card.md).
   * Sources: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training)
6. Capabilities and limitations
   * 4 points: Grounded discussion of both capabilities and limitations. Capabilities: “a model that can be used to generate and modify images based on text prompts…intended for research purposes only” with use cases including “Safe deployment of models which have the potential to generate harmful content. Probing and understanding the limitations and biases of generative models. Generation of artworks and use in design and other artistic processes. Applications in educational or creative tools. Research on generative models.” Limitations: “The model does not achieve perfect photorealism The model cannot render legible text The model does not perform well on more difficult tasks which involve compositionality, such as rendering an image corresponding to “A red cube on top of a blue sphere” Faces and people in general may not be generated properly. The model was trained mainly with English captions and will not work as well in other languages. The autoencoding part of the model is lossy The model was trained on a subset of the large-scale dataset [LAION-5B](https://laion.ai/blog/laion-5b/), which contains adult, violent and sexual content. To partially mitigate this, we have filtered the dataset using LAION's NFSW detector (see Training section).” Room for improvement in terms of referencing specific evaluations of these capabilities and limitations.
   * Source: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training)
7. Risks and mitigations
   * 1 points: Risks are listed (bias toward “white and western cultures” due to training on data with English language descriptions, “adult, violent and sexual content” in LAION-5B, “generating images that people would foreseeably find disturbing, distressing, or offensive; or content that propagates historical or current stereotypes”) though there is no list of mitigations, description of how well mitigations work, or justification of non-mitigated risks.
   * Source: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training)
     1. Note: Could have easily given 2 points here. One mitigation listed is using the NSFW filter. Another, in Stability’s API [Terms of Service](https://platform.stability.ai/docs/terms-of-service), states that Stability “may block certain text prompts and blur resulting Content if we find (in our sole judgment) the prompt or resulting image too graphic, offensive, insensitive, or hurtful, even in circumstances where you may not have had any ill intent or where the Content does not otherwise violate the prohibitions described above. We may also block text prompts or block or remove Content that we otherwise find objectionable or that we believe in our reasonable judgment may expose us or others to any harm or liability”
8. Evaluations
   * 0 points: No measurement of accuracy, harms, robustness or other factors using public/industry standard benchmarks. Model card contains “evaluation results” of FID vs CLIP scores but little context is given.
   * Source: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training)
9. Testing
   * 0 points: No description of internal testing, external evaluation, red teaming or stress testing.
   * Source: [https://huggingface.co/stabilityai/stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2#training) ; <https://stability.ai/blog/stable-diffusion-v2-release>
10. Machine-generated content
    * 0 points: no discussion of machine-generated content in model card or terms of service.
    * Sources: <https://huggingface.co/stabilityai/stable-diffusion-2#misuse-malicious-use-and-out-of-scope-use> ; <https://platform.stability.ai/docs/terms-of-service>
      1. Note: comes closest with “The model was not trained to be factual or true representations of people or events, and therefore using the model to generate such content is out-of-scope for the abilities of this model” in the model card and “Generate Content or engage in any Content Sharing that is intentionally misleading” in the Terms of Service
11. Member states
    * 0 points: No description of deployment practices in relation to the EU.
    * Sources: <https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL>; <https://platform.stability.ai/docs/terms-of-service> ;
12. Downstream documentation
    * 4 points: All relevant documentation is provided and easily accessible.
    * Source: <https://huggingface.co/stabilityai/stable-diffusion-2/tree/main>