**Stability (Stable Diffusion v2) - total: TODO/48**

1. Data sources
   * 4 points: Dataset pipeline is disclosed: Training on LAION-5B filtered by the LAION-NSFW classifier with specific scores followed by further training on derivatives of LAION-5B. Dataset size and number of steps are disclosed. LAION-5B dataset itself provides data decomposition and underlying sizes including associated datasheet.
   * Sources: <https://huggingface.co/stabilityai/stable-diffusion-2#training> and <https://arxiv.org/pdf/2210.08402.pdf>
2. Data governance
   * 1 points: Dataset curation and governance is described to some extent in the LAION-5B paper, but not such discussion is given in the context of Stable Diffusion with the exception of filtering of unsafe or pornographic material.
   * Sources: <https://huggingface.co/stabilityai/stable-diffusion-2#training> and <https://arxiv.org/pdf/2210.08402.pdf>
3. Copyrighted data
   * 0 points: Copyright is not discussed (very peripheral mention in LAION-5B paper).
   * Sources: <https://huggingface.co/stabilityai/stable-diffusion-2#training> and <https://arxiv.org/pdf/2210.08402.pdf>
4. Compute (additive)
   * 4 points: The model size, compute requirements (training time, number of A100s), compute provider (AWS), and associated measurement methodology are described.
   * Source: <https://huggingface.co/stabilityai/stable-diffusion-2#environmental-impact>
5. Energy (additive)
   * 3 points: The energy usage, compute region, measurement methodology, and the emissions are disclosed. No specific mitigations to improve efficiency or reduce energy consumption are disclosed. Note: The information appears to be copied from Stable Diffusion v1, so we are uncertain if it is correctly updated (though it is reported under Stable Diffusion v2), so we assume correctness of reporting and we give credit at present.
   * Source: <https://huggingface.co/stabilityai/stable-diffusion-2#environmental-impact>
6. Capabilities and limitations
   * 4 points: Capabilities are described plainly (“a model that can be used to generate and modify images based on text prompts.”) with specific use cases (“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”). Several limitations (e.g. “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”), largely aligning with the literature on text-to-image model limitations. Some grounding to specifics of this model.
   * Source: <https://huggingface.co/stabilityai/stable-diffusion-2>
7. Risks and mitigations (additive)
   * 1 point: Risks are listed (e.g. bias/presentation of white and western culture as default, sexual content generation, violent content generation, degradation, disinformation, sharing of copyrighted information in violation of terms of use) though no significant mitigations (exception: data filtering for pornographic content) are taken as disclosed in the model card.
   * Source: <https://huggingface.co/stabilityai/stable-diffusion-2#misuse-malicious-use-and-out-of-scope-use>
8. Evaluations (additive)
   * 0 points: No evaluations, though some FID scores are reported in the model card but the evaluation context is difficult to determine.
   * Sources: <https://stability.ai/blog/stable-diffusion-v2-release> and <https://huggingface.co/stabilityai/stable-diffusion-2#misuse-malicious-use-and-out-of-scope-use>
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).
   * Sources: <https://stability.ai/blog/stable-diffusion-v2-release> and <https://huggingface.co/stabilityai/stable-diffusion-2#misuse-malicious-use-and-out-of-scope-use>
10. Machine-generated content
    * 1 point: No discussion of disclosure of content being machine-generated. Given model is openly released, no enforcement mechanism is discussed as well. Misuse categories do include “impersonating individuals without their content”.
    * Source: <https://huggingface.co/stabilityai/stable-diffusion-2#misuse-and-malicious-use> and <https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL>
11. Member states
    * 2 points: No discussion of countries of use, though the model is openly available via Hugging Face, therefore subject to Hugging Face terms-of-service in relation to US export controls. No discussion of specific EU member states.
    * Sources: <https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL> and <https://huggingface.co/terms-of-service>
12. Downstream documentation
    * 4 points: Model card is available alongside model download with relevant documentation.
    * Source: <https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL>