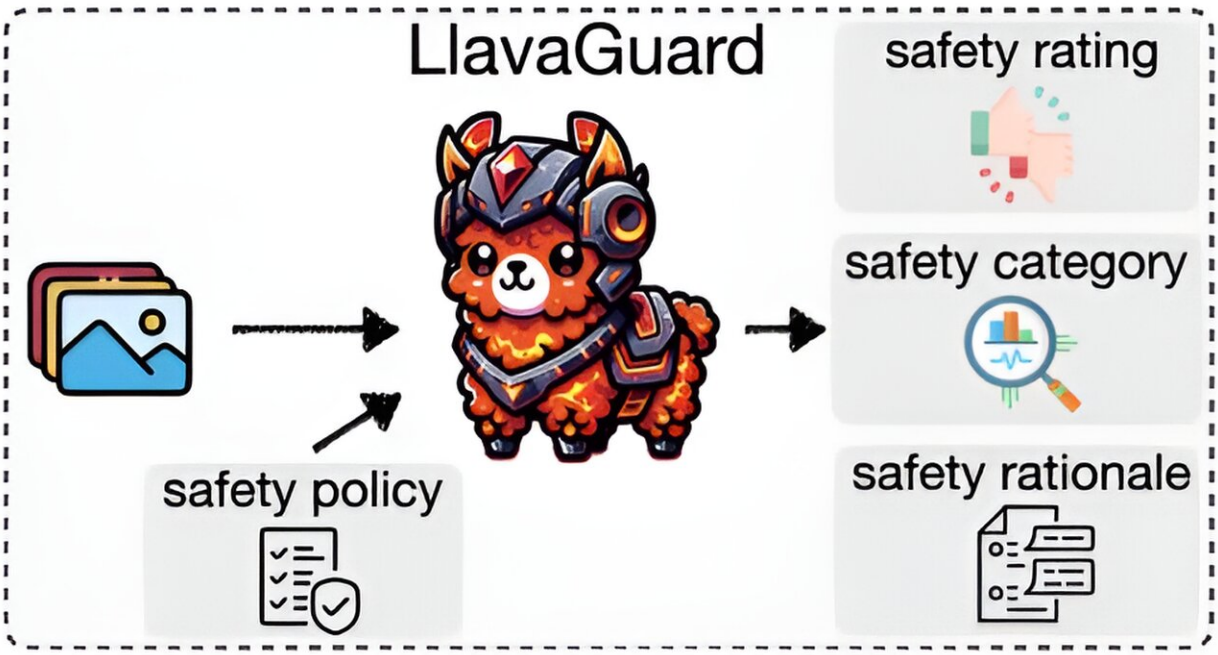
New tool uses vision language models to safeguard against offensive image content

July 10 2024, by Silke Paradowski



LlavaGuard judges images for safety alignment with a policy providing a safety rating, category, and rationale. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2406.05113

Researchers at the Artificial Intelligence and Machine Learning Lab (AIML) in the Department of Computer Science at TU Darmstadt and the Hessian Center for Artificial Intelligence (hessian.AI) have developed a method that uses vision language models to filter, evaluate,

and suppress specific image content in large datasets or from image generators.

Artificial intelligence (AI) can be used to identify objects in images and videos. This computer vision can also be used to analyze large corpora of visual data.

Researchers led by Felix Friedrich from the AIML have developed a method called LlavaGuard, which can now be used to filter certain image content. This tool uses so-called vision language models (VLMs). In contrast to large language models (LLMs) such as ChatGPT, which can only process text, vision language models are able to process and understand image and text content simultaneously. The work is published on the *arXiv* preprint server.

LlavaGuard can also fulfill complex requirements, as it is characterized by its ability to adapt to different legal regulations and user requirements. For example, the tool can differentiate between regions in which activities such as cannabis consumption are legal or illegal. LlavaGuard can also assess whether content is appropriate for certain age groups and restrict or adapt it accordingly.

"Until now, such fine-grained safety tools have only been available for analyzing texts. When filtering images, only the 'nudity' category has previously been implemented, but not others such as 'violence,' 'self-harm' or 'drug abuse,'" says Friedrich.

LlavaGuard not only flags problematic content, but also provides detailed explanations of its safety ratings by categorizing content (e.g., "hate," "illegal substances," "violence," etc.) and explaining why it is classified as safe or unsafe.

"This transparency is what makes our tool so special and is crucial for

understanding and trust," explains Friedrich. It makes LlavaGuard an invaluable tool for researchers, developers and political decision-makers.

The research on LlavaGuard is an integral part of the Reasonable Artificial Intelligence (RAI) cluster project at TU Darmstadt and demonstrates the university's commitment to advancing safe and ethical AI technologies. LlavaGuard was developed to increase the safety of large generative models by filtering training data and explaining and justifying the output of problematic motives, thereby reducing the risk of generating harmful or inappropriate content.

The potential applications of LlavaGuard are far-reaching. Although the tool is currently still under development and focused on research, it can already be integrated into image generators such as Stable Diffusion to minimize the production of unsafe content.

In addition, LlavaGuard could also be adapted for use on social media platforms in the future to protect users by filtering out inappropriate images and thus promoting a safer online environment.

More information: Lukas Helff et al, LLavaGuard: VLM-based Safeguards for Vision Dataset Curation and Safety Assessment, *arXiv* (2024). DOI: 10.48550/arxiv.2406.05113

Provided by Technische Universitat Darmstadt