In recent times, machine learning (ML) models, like ChatGPT and Stable diffusion, have grown rapidly in creating high-quality text and images. This is possible due to the abundance of data for training and improved computing power. These models are now widely used in various fields for tasks like image recognition, language processing, and predictive analytics.

Despite the remarkable achievements of generative models in producing high-quality outputs, this project underscores the imperative of prioritizing their safety, a facet frequently overshadowed. Safety here means ensuring these models operate without causing unintended harm, especially in terms of protecting the confidentiality and integrity of processed data. It involves preventing the models from revealing or misusing sensitive information and guarding against the generation of misleading or harmful content. Safety also extends to defending against attacks or manipulation, crucial in applications involving critical decision-making processes.

Our project is built upon a fundamental hypothesis that underscores the vital connection between the safety of generative models and how their training data is managed. We firmly believe that the bedrock of reliable and legally compliant AI systems lies **in the nature and handling of the data that these systems are trained on**. To delve into this connection, we've structured our research into three key tasks.

In the initial phase, known as Task I, our primary focus is on exploring the learning mechanisms of generative models, specifically honing in on the concept of memorization. This aspect is particularly critical as it often leads to unintended exposure of sensitive or copyrighted material during the models' inference stage.

Moving on to Task II, our second objective revolves around devising innovative verification techniques tailored for generative models. These techniques are designed to serve two primary scenarios. The first scenario empowers individuals, such as artists, to determine if their creative works, not explicitly memorized by the model, were improperly used in its training. The second scenario, termed output verification, aims at enabling model owners to confirm or refute the generation of potentially harmful content by their models.

In Task III we propose groundbreaking strategies for safeguarding both generative models and the parties who trained them (usually API providers). Considering the substantial investment in data, computational power, and manual labor required for training, we advocate treating these models as intellectual property belonging to the entity responsible for their training. Our approach involves the development of active defenses to thwart attempts at model theft. In cases where theft still occurs, we explore ownership resolution techniques to legally challenge such actions.

Ultimately, our project seeks to establish a consistent ML framework that prioritizes the confidentiality and integrity of generative models. We plan to disseminate this framework through educational materials and an open-source library as well as validate the effectiveness of our methods on real-life applications.

In summary, our project aims to enhance the safety of generative models by investigating memorization, devising verification techniques, and proposing strategies for safeguarding both models and data. By creating a robust ML framework and validating it in real-life applications, we aspire to contribute to the responsible and secure deployment of generative models, ensuring their integrity and confidentiality in the evolving landscape of artificial intelligence.