W ostatnich latach modele uczenia maszynowego, takie jak ChatGPT i Stable diffusion, umożliwiły bezprecedensowy postęp w tworzeniu wysokiej jakości tekstu i obrazów. Stało się to możliwe dzięki dużej ilości danych do szkolenia i zwiększonej mocy obliczeniowej. Modele mają obecnie wiele zastosowań w różnych dziedzinach, m. in. do zadań takich jak rozpoznawanie obrazów, przetwarzanie języka naturalnego i analiza predykcyjna.

Pomimo niezwykłych osiągnięć modeli generatywnych w generowaniu wysokiej jakości wyników, niniejszy projekt podkreśla konieczność priorytetowego traktowania ich bezpieczeństwa, które jest często pomijanym aspektem rozwoju tych modeli. Bezpieczeństwo oznacza tutaj zapewnienie, że modele te działają bez powodowania niezamierzonych szkód, zwłaszcza w zakresie ochrony poufności i integralności przetwarzanych danych. Obejmuje to zapobieganie ujawnianiu lub niewłaściwemu wykorzystywaniu przez modele danych wrażliwych oraz ochronę przed generowaniem błędnych lub szkodliwych treści. Bezpieczeństwo obejmuje również ochronę przed atakami lub manipulacjami, co ma kluczowe znaczenie w zastosowaniach obejmujących krytyczne procesy decyzyjne.

Nasz projekt opiera się na fundamentalnej hipotezie, która podkreśla istotny związek między bezpieczeństwem modeli generatywnych a sposobem zarządzania ich danymi treningowymi. Jesteśmy przekonani, że podstawą niezawodnych i zgodnych z prawem systemów sztucznej inteligencji jest sposób **w jaki przetwarzane oraz weryfikowane są dane, na których te systemy są trenowane**. Aby pogłębić zrozumienie zależności między danymi i modelami, podzieliliśmy nasze badania na trzy kluczowe zadania.

W początkowej fazie (Zadanie I) skupiamy się przede wszystkim na badaniu mechanizmów uczenia się modeli generatywnych, w szczególności na koncepcji zapamiętywania. Ten aspekt jest szczególnie istotny, ponieważ często prowadzi do niezamierzonego ujawnienia wrażliwych lub chronionych prawem autorskim materiałów na etapie wnioskowania (inferencji) modeli.

Przechodząc do zadania II, nasz drugi cel dotyczy opracowania innowacyjnych technik weryfikacji dostosowanych do modeli generatywnych. Techniki te mają służyć dwóm podstawowym scenariuszom. Pierwszy scenariusz umożliwia osobom fizycznym, takim jak artyści, ustalenie, czy ich twórcze prace, które nie zostały bezpośrednio zapamiętane przez model, zostały niewłaściwie wykorzystane w jego szkoleniu. Drugi scenariusz, zwany weryfikacją wyników, ma na celu określenie pochodzenia generowanych (potencjalnie szkodliwych) treści przez modele - umożliwienie potwierdzenia lub negacji przez właścicieli modelu.

W Zadaniu III proponujemy przełomowe strategie ochrony zarówno modeli generatywnych, jak i stron, które je wyszkoliły (zazwyczaj dostawców API). Biorąc pod uwagę znaczne inwestycje w dane, moc obliczeniową i manualną pracę wymaganą do treningu, opowiadamy się za traktowaniem tych modeli jako własności intelektualnej należącej do podmiotu odpowiedzialnego za ich trenowanie. Nasze podejście obejmuje opracowanie aktywnych mechanizmów obronnych w celu udaremnienia prób kradzieży modeli. W przypadkach, w których kradzież nadal ma miejsce, badamy techniki rozwiązywania kwestii własności, aby prawnie zakwestionować takie działania.

Nasz projekt ma na celu ustanowienie spójnych ram uczenia maszynowego (ang. framework), które priorytetowo traktują poufność i integralność modeli generatywnych. Planujemy rozpowszechniać powstałe metody poprzez materiały edukacyjne i bibliotekę open source, a także zweryfikować skuteczność naszych metod w rzeczywistych zastosowaniach.

Podsumowując, nasz projekt ma na celu zwiększenie bezpieczeństwa modeli generatywnych poprzez zbadanie zapamiętywania, opracowanie technik weryfikacji i zaproponowanie strategii ochrony zarówno modeli, jak i podmiotu udostępniającego modele. Tworząc solidne ramy bezpiecznego uczenia maszynowego i weryfikując ich skuteczność w rzeczywistych aplikacjach, dążymy do przyczynienia się do odpowiedzialnego i bezpiecznego wdrażania modeli generatywnych, zapewniając ich integralność i poufność w szybko zmieniającym się krajobrazie sztucznej inteligencji.