Mechanistyczna Interpretacja Modeli Dyfuzyjnych

Generatywna sztuczna inteligencja przebojem wdarła się do każdego aspektu naszej rzeczywistości, stojąc za efektami specjalnymi w kinowych hitach czy odkryciami biologicznymi na miarę Nagrody Nobla. Mimo tych sukcesów, praca z dzisiejszymi modelami AI rzadko przypomina precyzyjną inżynierię, a częściej grę na "jednorękim bandycie". Wpisujemy polecenie, pociągamy za wajchę i mamy nadzieję, że wynik będzie zgodny z naszymi oczekiwaniami. Jeśli chcemy wygenerować "zabawnego kota", taki mechanizm jest często wystarczający. Jednak gdy potrzebujemy precyzyjnej kontroli (na przykład przy projektowaniu białka o konkretnych właściwościach leczniczych) obecna metoda prób i błędów po prostu nie zdaje egzaminu.

Nasz projekt ma na celu zamianę tej losowej maszyny w precyzyjny panel sterowania. Zajrzymy do wnętrza modelu który do tej pory traktowany był zazwyczaj jak tzw. "czarna skrzynka", aby zrozumieć, jak dokładnie "myśli" sztuczna inteligencja. Wykorzystując metodę zwaną "mechanistyczną interpretowalnością" (mechanistic interpretability), chcemy stworzyć mapę wewnętrznych połączeń tych modeli i znaleźć konkretne "przełączniki" sterujące wynikiem. Zamiast ogólnikowo prosić model o zmianę, chcemy zlokalizować konkretne neurony odpowiedzialne za daną cechę – czy to teksturę obiektu na zdjęciu, rytm utworu muzycznego, czy strukturę chemiczną cząsteczki – i bezpośrednio nimi sterować.

Naszym celem jest stworzenie uniwersalnego systemu sterowania, działającego niezależnie od rodzaju danych. Stawiamy tezę, że sposób, w jaki AI koduje "styl" obrazu, jest matematycznie bliźniaczy do tego, jak koduje "funkcję" cząsteczki chemicznej czy peptydu. Opracowując wspólną logikę dla tych mechanizmów, chcemy stworzyć jednolite wytyczne dla precyzyjnego sterowania modelami – niezależnie od tego, czy regulujemy rytm piosenki, czy właściwości toksyczne związku chemicznego.

Aby udowodnić słuszność naszej teorii, wdrożymy rygorystyczny proces weryfikacji. W przypadku obrazu i dźwięku o ocenę jakości poprosimy ludzi – słuchaczy i widzów. W dziedzinie biologii pójdziemy jednak o krok dalej: zaprojektowane przez odpowiednio wysterowaną sztuczną inteligencję peptydy, fizycznie wytworzymy w laboratorium, a następnie sprawdzimy ich skuteczność w neutralizowaniu szkodliwych bakterii. W ten sposób tworzymy bezpośredni pomost pomiędzy "wirtualnymi pokrętłami", którymi sterujemy w modelu, a realnymi efektami w fizycznym świecie.