SMART Закупка

Только что

Хотите, чтобы ИИ не делился секретами? Тренируйтесь сами

11 марта 2023 года подразделение Samsung Device Solutions разрешило сотрудникам использовать ChatGPT . Пошли проблемы . Отчет The Economist Korea , опубликованный менее чем через три недели, выявил три случая «утечки данных». Два инженера использовали ChatGPT для устранения неполадок в конфиденциальном коде, а руководитель использовал его для расшифровки протокола совещания. Samsung изменил курс, запретив использование сотрудниками не только ChatGPT, но и всего внешнего генеративного ИИ.

Ситуация с Samsung иллюстрирует проблему, с которой сталкивается любой, кто использует сторонние инструменты генеративного ИИ, основанные на большой языковой модели (LLM). Самые мощные инструменты искусственного интеллекта могут поглощать большие фрагменты текста и быстро выдавать полезные результаты, но эта функция может легко привести к утечке данных.

«Это может быть хорошо для личного использования, но как насчет корпоративного использования? […] Вы не можете просто отправить все свои данные в OpenAI , на их серверы», — говорит Талеб Алашкар , главный технический директор компании по компьютерному зрению AlgoFace и MIT Research Affiliate.

Наивные пользователи ИИ передают личные данные
Проблемы конфиденциальности данных генеративного ИИ сводятся к двум ключевым проблемам.

ИИ связан теми же правилами конфиденциальности, что и другие технологии. Временный запрет ChatGPT в Италии произошел после инцидента с безопасностью в марте 2023 года , который позволил пользователям просматривать истории чатов других пользователей. Эта проблема может затронуть любую технологию, хранящую пользовательские данные. Италия сняла запрет после того, как OpenAI добавила функции , дающие пользователям больший контроль над тем, как их данные хранятся и используются.

Но ИИ сталкивается с другими уникальными проблемами. Генеративные модели ИИ не предназначены для воспроизведения обучающих данных и, как правило, не способны на это в любом конкретном случае, но это не невозможно. В документе под названием «Извлечение обучающих данных из диффузионных моделей», опубликованном в январе 2023 года, описывается, как стабильная диффузия может генерировать изображения, похожие на изображения в обучающих данных. Иск Doe против GitHub включает в себя примеры кода, сгенерированного Github Copilot, инструмента на базе LLM от OpenAI, который соответствует коду, найденному в обучающих данных. Это приводит к опасениям, что генеративный ИИ, контролируемый третьей стороной, может непреднамеренно привести к утечке конфиденциальных данных частично или полностью. Некоторые генеративные инструменты искусственного интеллекта, в том числе ChatGPT, усугубляют этот страх, включая

пользовательские данные в свой обучающий набор. У организаций, обеспокоенных конфиденциальностью данных, не остается иного выбора, кроме как запретить их использование.

«Подумайте о страховой компании, или крупных банках, или [министерстве обороны] или клинике Мэйо», — говорит Алашкар, добавляя, что «каждый ИТ-директор, технический директор, директор по безопасности или менеджер в компании занят изучением этих политик и поиском лучших практики. Я думаю, что большинство ответственных компаний сейчас очень заняты, пытаясь найти то, что нужно».

Эффективность — это ответ на частный ИИ
У проблем конфиденциальности данных ИИ есть очевидное решение. Организация может обучаться, используя свои собственные данные (или данные, которые она получила с помощью средств, отвечающих требованиям конфиденциальности данных) и развернуть модель на оборудовании, которым она владеет и управляет. Но очевидное решение сопряжено с очевидной проблемой: оно неэффективно . Процесс обучения и развертывания генеративной модели ИИ дорог и сложен в управлении для всех, кроме самых опытных и хорошо финансируемых организаций.

«Когда вы начинаете тренироваться на 500 графических процессорах, все идет не так. Вы действительно должны знать, что делаете, и это то, что мы сделали, и мы объединили это в интерфейсе», — говорит Навин Рао , соучредитель и генеральный директор MosaicML . Компания Рао предлагает третий вариант: размещенную модель ИИ, работающую в защищенной среде MosaicML. Моделью можно управлять через веб-клиент, интерфейс командной строки или Python.

«Когда вы начинаете тренироваться на 500 графических процессорах, все идет не так. Вы действительно должны знать, что делаете». — Навин Рао, соучредитель и генеральный директор MosaicML.

«Вот платформа, вот модель, а вы сохраняете свои данные. Обучите свою модель и сохраните вес модели. Данные остаются в вашей сети», — объясняет Джули Чой, директор MosaicML по маркетингу и связям с общественностью. Чой говорит, что компания работает с клиентами в финансовой сфере и другими, которые «действительно инвестируют в свою собственную интеллектуальную собственность».

Хостинговый подход является растущей тенденцией. Intel сотрудничает с Boston Consulting Group над частной моделью ИИ , IBM планирует выйти на арену с ИИ Watsonx , а существующие сервисы, такие как Sagemaker от Amazon и Microsoft Azure ML, развиваются в соответствии со спросом. 11 марта 2023 года подразделение Samsung Device Solutions разрешило сотрудникам использовать ChatGPT . Пошли проблемы . Отчет The Economist Korea , опубликованный менее чем через три недели, выявил три случая «утечки данных». Два инженера использовали ChatGPT для устранения неполадок в конфиденциальном коде, а руководитель использовал его для расшифровки протокола совещания. Samsung изменил курс, запретив использование сотрудниками не только ChatGPT, но и всего внешнего генеративного ИИ.

Ситуация с Samsung иллюстрирует проблему, с которой сталкивается любой, кто использует сторонние инструменты генеративного ИИ, основанные на большой языковой модели (LLM). Самые мощные инструменты искусственного интеллекта могут поглощать большие фрагменты текста и быстро выдавать полезные результаты, но эта функция может легко привести к утечке данных.

«Это может быть хорошо для личного использования, но как насчет корпоративного

использования? […] Вы не можете просто отправить все свои данные в OpenAI , на их серверы», — говорит Талеб Алашкар , главный технический директор компании по компьютерному зрению AlgoFace и MIT Research Affiliate.

Наивные пользователи ИИ передают личные данные
Проблемы конфиденциальности данных генеративного ИИ сводятся к двум ключевым проблемам.

ИИ связан теми же правилами конфиденциальности, что и другие технологии. Временный запрет ChatGPT в Италии произошел после инцидента с безопасностью в марте 2023 года , который позволил пользователям просматривать истории чатов других пользователей. Эта проблема может затронуть любую технологию, хранящую пользовательские данные. Италия сняла запрет после того, как OpenAI добавила функции , дающие пользователям больший контроль над тем, как их данные хранятся и используются.

Но ИИ сталкивается с другими уникальными проблемами. Генеративные модели ИИ не предназначены для воспроизведения обучающих данных и, как правило, не способны на это в любом конкретном случае, но это не невозможно. В документе под названием «Извлечение обучающих данных из диффузионных моделей», опубликованном в январе 2023 года, описывается, как стабильная диффузия может генерировать изображения, похожие на изображения в обучающих данных. Иск Doe против GitHub включает в себя примеры кода, сгенерированного Github Copilot, инструмента на базе LLM от OpenAI, который соответствует коду, найденному в обучающих данных.

Фотография женщины по имени Энн Грэм Лотц рядом с созданным искусственным интеллектом изображением Энн Грэм Лотц, созданным с помощью Stable Diffusion. Сравнение показывает, что изображение генератора ИИ значительно похоже на исходное изображение, которое было включено в обучающие данные модели ИИ.Исследователи обнаружили, что Stable Diffusion иногда может создавать изображения, похожие на обучающие данные.
ИЗВЛЕЧЕНИЕ ОБУЧАЮЩИХ ДАННЫХ ИЗ ДИФФУЗИОННЫХ МОДЕЛЕЙ
Это приводит к опасениям, что генеративный ИИ, контролируемый третьей стороной, может непреднамеренно привести к утечке конфиденциальных данных частично или полностью. Некоторые генеративные инструменты искусственного интеллекта, в том числе ChatGPT, усугубляют этот страх, включая пользовательские данные в свой обучающий набор. У организаций, обеспокоенных конфиденциальностью данных, не остается иного выбора, кроме как запретить их использование.

«Подумайте о страховой компании, или крупных банках, или [министерстве обороны] или клинике Мэйо», — говорит Алашкар, добавляя, что «каждый ИТ-директор, технический директор, директор по безопасности или менеджер в компании занят изучением этих политик и поиском лучших практики. Я думаю, что большинство ответственных компаний сейчас очень заняты, пытаясь найти то, что нужно».

Эффективность — это ответ на частный ИИ
У проблем конфиденциальности данных ИИ есть очевидное решение. Организация может обучаться, используя свои собственные данные (или данные, которые она получила с помощью средств, отвечающих требованиям конфиденциальности данных) и развернуть модель на оборудовании, которым она владеет и управляет. Но очевидное решение сопряжено с очевидной проблемой: оно неэффективно . Процесс обучения и развертывания генеративной модели ИИ дорог и сложен в управлении для всех, кроме самых опытных и хорошо финансируемых организаций.

«Когда вы начинаете тренироваться на 500 графических процессорах, все идет не так. Вы

действительно должны знать, что делаете, и это то, что мы сделали, и мы объединили это в интерфейсе», — говорит Навин Рао , соучредитель и генеральный директор MosaicML . Компания Рао предлагает третий вариант: размещенную модель ИИ, работающую в защищенной среде MosaicML. Моделью можно управлять через веб-клиент, интерфейс командной строки или Python.

«Когда вы начинаете тренироваться на 500 графических процессорах, все идет не так. Вы действительно должны знать, что делаете». — Навин Рао, соучредитель и генеральный директор MosaicML.

«Вот платформа, вот модель, а вы сохраняете свои данные. Обучите свою модель и сохраните вес модели. Данные остаются в вашей сети», — объясняет Джули Чой, директор MosaicML по маркетингу и связям с общественностью. Чой говорит, что компания работает с клиентами в финансовой сфере и другими, которые «действительно инвестируют в свою собственную интеллектуальную собственность».

Хостинговый подход является растущей тенденцией. Intel сотрудничает с Boston Consulting Group над частной моделью ИИ , IBM планирует выйти на арену с ИИ Watsonx , а существующие сервисы, такие как Sagemaker от Amazon и Microsoft Azure ML, развиваются в соответствии со спросом.

График, показывающий обучение модели ИИ, размещенной на MosaicML. На графике отмечены несколько точек, в которых произошли аппаратные сбои. Обучение возобновлялось автоматически после каждого отказа оборудования.MosaicML может обучить хост LLM менее чем за 10 дней и автоматически компенсирует аппаратные сбои, возникающие во время обучения. МОЗАИКАML
Обучение размещенной модели ИИ остается дорогим, сложным и трудоемким, но значительно меньшим, чем обучение в одиночку. 5 мая 2023 года MosaicML объявила, что обучила модель LLM под названием MPT-7B менее чем за 200 000 долларов США за девять с половиной дней и без вмешательства человека. OpenAI не раскрывает стоимость обучения своих моделей, но оценивает стоимость обучения GPT-3 как минимум в 4,6 миллиона долларов .

Развертывание размещенной модели искусственного интеллекта также дает организациям контроль над вопросами, граничащими с конфиденциальностью, такими как доверие и безопасность. Чой говорит, что приложение для чата по питанию обратилось к MosaicML после того, как обнаружило, что его предложения искусственного интеллекта вызвали реакцию «постыдить жир». Приложение, которое в то время использовало конкурирующий LLM, не могло предотвратить нежелательные ответы, потому что оно не контролировало обучающие данные или веса, используемые для точной настройки выходных данных.

«Мы действительно считаем, что безопасность и конфиденциальность данных имеют первостепенное значение при создании систем искусственного интеллекта. Потому что, в конце концов, ИИ — это ускоритель, и он будет обучаться на ваших данных, чтобы помочь вам принимать решения», — говорит Чой.

Ссылка на статью: Хотите, чтобы ИИ не делился секретами? Тренируйтесь сами