Вики

Галлюцинации искусственного интеллекта (Hallucinations of artificial intelligence)

> Синонимы: AI hallucinations, Artificial hallucination, Confabulation, Искусственная галлюцинация, Конфабуляция

Галлюцинациями искусственного интеллекта (ИИ) называют феномен, когда генеративные модели ИИ, такие как LLM или генераторы изображений, формируют контент с ошибками, искаженными или несуществующими паттернами и артефактами, а также ложной или бессмысленной информацией.

Например, чат-боты LLM могут вставлять правдоподобно звучащие, но ложные утверждения в ответы. Нередко указываются неверные даты, события или статистические данные, неправильное приписывание цитат или выдумка ссылок, описание несуществующих технологий или концепций как реальных.

В начале 2020-х аналитики обозначили галлюцинации как одну из наиболее значимых проблем генеративного ИИ, и даже прозвучали предложения прекратить новые разработки в этом направлении. В настоящее время выявление и устранение галлюцинаций и их последствий представляет собой важнейшую задачу в области практического внедрения и обеспечения надежности генеративного ИИ.

Термин «галлюцинация», заимствованный из описания ложных восприятий человека, в контексте обработки информации появился в начале 1980-х в области компьютерного зрения. В 2000-х галлюцинациями называли сбои в работе систем обработки естественного языка. И только в 2010-х этим термином стали обозначать генерацию фактически неверных или вводящих в заблуждение результатов системами ИИ в таких задачах, как машинный перевод и идентификация объектов.

Широкое распространение термин получил в начале 2020-х с началом бума систем ИИ. Разработчики генеративных моделей предупреждали о возможных галлюцинациях, определяя их как «уверенные утверждения, которые не соответствуют действительности». Хотя некоторые исследователи избегают термина «галлюцинация», считая его потенциально вводящим в заблуждение из-за антропоморфности.

Везде, где системы ИИ используются при принятии решений, их галлюцинации могут представлять определенные риски. Некоторые из них могут быть незначительными — например, когда чат-бот дает неверный ответ на простой вопрос, пользователь может оказаться недостаточно информированным. Но есть гораздо более критические области, где галлюцинации будут иметь фатальный характер.

Например, галлюцинации в системах ИИ автономных транспортных средств могут приводить к авариям. В страховании, где ИИ используется для определения права клиента на страховое покрытие, галлюцинации могут иметь последствия, меняющие жизнь человека. В проектировании с использованием ИИ они могут привести к неверным техническим решениям, которые впоследствии могут вызвать катастрофу.

Выделяют две группы причин возникновения галлюцинаций:

1. Первая связана с тем, как работают генеративные модели. Они предсказывают слова на основе вероятности — то есть выбирают те, которые чаще всего встречаются рядом в обучающих данных (например, «пылесос» и «уборка»). Но если такие слова не подходят по смыслу в конкретном контексте, модель может создавать бессмысленные фразы или выдумывать факты.
2. Вторая — с качеством данных для обучения. Если данные неполные, неточные или содержат ошибки, модель будет воспроизводить эти ошибки в своих ответах.

На рисунке ниже показаны наборы изображений, на которых моделям сложно обучаться. В них похожие объекты встречаются в разных контекстах: например, в одном случае рядом оказываются собака и кекс с изюмом, а в другом — собака и швабра.

В результате можно ожидать, что при генерации контента о кулинарии и выпечке модель не к месту вставит туда изображение с собакой. А в материале о собаках читатель может обнаружить картинку со шваброй.

Считается, что полностью устранить галлюцинации возможно лишь при 100% точности моделей. Поскольку это недостижимо, галлюцинации остаются неизбежными. Поэтому для борьбы с их последствиями были выработаны различные стратегии:

- критическая оценка результатов работы генеративного ИИ;
- повышение качества обучающих данных с использованием алгоритмов расширенного поиска (RAG);
- применение технологий и инструментов, которые повышают эффективность доступа к разнообразным источникам качественных, проверенных данных, таких как

MCP-серверы;

- использование специальных методик проектирования подсказок: например, явное указание модели отвечать «я не знаю», если есть неопределенность, ограничение области ответов или требование указывать источники;
- использование программных инструментов для проверки соответствия сгенерированного контента априори заданным требованиям;
- многократное выполнение одного и того же задания с последующим выбором наиболее точного ответа (самосогласование).

Однако в творческих областях (например, в искусстве) нестандартные и даже «странные» результаты генерации могут представлять интерес.