Почему ИИ продолжает «галлюцинировать» и можно ли это остановить

Дата публикации: 17.09.2025

Искусственный интеллект сегодня воспринимается как мощный инструмент, способный решать сложные задачи, генерировать тексты и помогать в исследовательской работе. Однако ключевая проблема, которая сопровождает языковые модели, — это так называемые «галлюцинации», то есть уверенное предоставление ложной информации. Новое исследование OpenAI представило одно из самых строгих математических объяснений этой проблемы, показав, что она не просто связана с качеством данных, а является фундаментальным следствием природы работы таких систем.

Языковые модели, включая ChatGPT, формируют ответы, предсказывая следующее слово в предложении на основе вероятностей. Этот процесс означает, что ошибка накапливается на каждом шаге генерации текста, а общая вероятность выдачи неверного результата становится выше, чем при ответах на простые бинарные вопросы. Даже при идеально чистых обучающих данных алгоритм будет допускать ошибки, поскольку задача классификации фактов сложна сама по себе. Чем реже встречается конкретная информация в обучающем наборе, тем выше риск ошибочного ответа. Пример с датами рождения известных личностей демонстрирует, что редкость упоминаний напрямую ведёт к высокой вероятности галлюцинаций.

Проблема усугубляется системой оценки современных моделей. Большинство бенчмарков, используемых в индустрии, основаны на бинарной системе, где ИИ получает ноль баллов как за честный ответ «я не знаю», так и за уверенно неправильный результат. В результате оптимальной стратегией становится угадывание, а не признание неопределённости. Математический анализ подтверждает, что при такой системе поощряется именно догадка, а не осторожность.

Решение, предложенное исследователями OpenAI, заключается в том, чтобы научить ИИ оценивать собственную уверенность в ответах. Система должна выдавать результат только в том случае, если вероятность правильности превышает заданный порог, например 75%. Такой подход уменьшает количество галлюцинаций, так как модель начинает отказываться от ответа в условиях высокой неопределённости. Однако это приведёт к радикальному изменению взаимодействия с пользователями. Представьте, что система начнёт отвечать «я не знаю» в 30% случаев — подобный опыт будет восприниматься многими как неудовлетворительный, особенно если они привыкли получать быстрые и

© HahgaPro, 2026

уверенные ответы.

Существует и экономический барьер. Для оценки вероятности правильности и анализа нескольких сценариев система должна проводить гораздо больше вычислений. Это увеличивает нагрузку на оборудование и расходы на обработку миллионов запросов ежедневно. В критически важных сферах, таких как медицина, логистика или финансовая торговля, затраты оправданы, потому что цена ошибки слишком высока. Но в массовых приложениях, ориентированных на потребителей, дорогие и медленные ответы проигрывают мгновенным, пусть и менее точным.

Основные выводы исследования можно сформулировать так: галлюцинации неизбежны из-за структуры языковых моделей, ошибки накапливаются при генерации текста, бинарные системы оценки поощряют догадки, а не честность, уменьшение ошибок возможно только при признании неопределённости, что ухудшает пользовательский опыт, снижение галлюцинаций требует больших вычислительных ресурсов.

В долгосрочной перспективе снижение стоимости электроэнергии на токен и развитие архитектуры микросхем могут сделать более реалистичным внедрение алгоритмов оценки уверенности в массовые ИИ-сервисы. Но фундаментальная проблема останется: пока бизнес-модель ориентирована на удобство и уверенные ответы, а не на строгую достоверность, галлюцинации будут сохраняться.

Это исследование не только демонстрирует ограничения нынешних технологий, но и подчёркивает необходимость переосмысления подхода к оценке работы ИИ. Возможно, для потребителей будущее искусственного интеллекта будет связано не столько с абсолютной точностью, сколько с новым балансом между скоростью, надёжностью и честностью системы в признании собственных ограничений.

© HahgaPro, 2026