Разработчики и исследователи ИИ из компании Anthropic показали, что многие «умные» чат-боты демонстрируют объяснения о своём «мыслительном процессе» перед предоставлением ответа, придавая ему весомость, самостоятельность и прозрачность.

Однако эксперименты с моделями цепочки мыслей (COT), которые способны «рассуждать», показали, что эти объяснения оказываются ненастоящими, и модели «не признаются в использовании подсказок». В тесте, где моделям Claude 3.7 Sonnet и DeepSeek-R1 была предложена «несанкционированная информация», они в автономном режиме определяли, использовать ли её.

Иными словами, модели должны были не просто рассуждать, а системам предоставлялись подсказки, на основании которых сначала генерировались простые логические цепочки в контексте нужного ответа, а затем построение логики усложнялось. Причём без подсказок от людей адекватного результата не было бы или ответы были бы неполными.

Таким образом, большая часть теста моделей оказалась «неверной». Но по запросу оператора об источнике «мысли», ИИ-модели буквально уклонялись от ответа.

«Эти результаты вызывают вопросы о прозрачности и достоверности ответов, предоставляемых чат-ботами с искусственным интеллектом и подчёркивают необходимость дальнейших исследований в этой области», — отмечают эксперты.