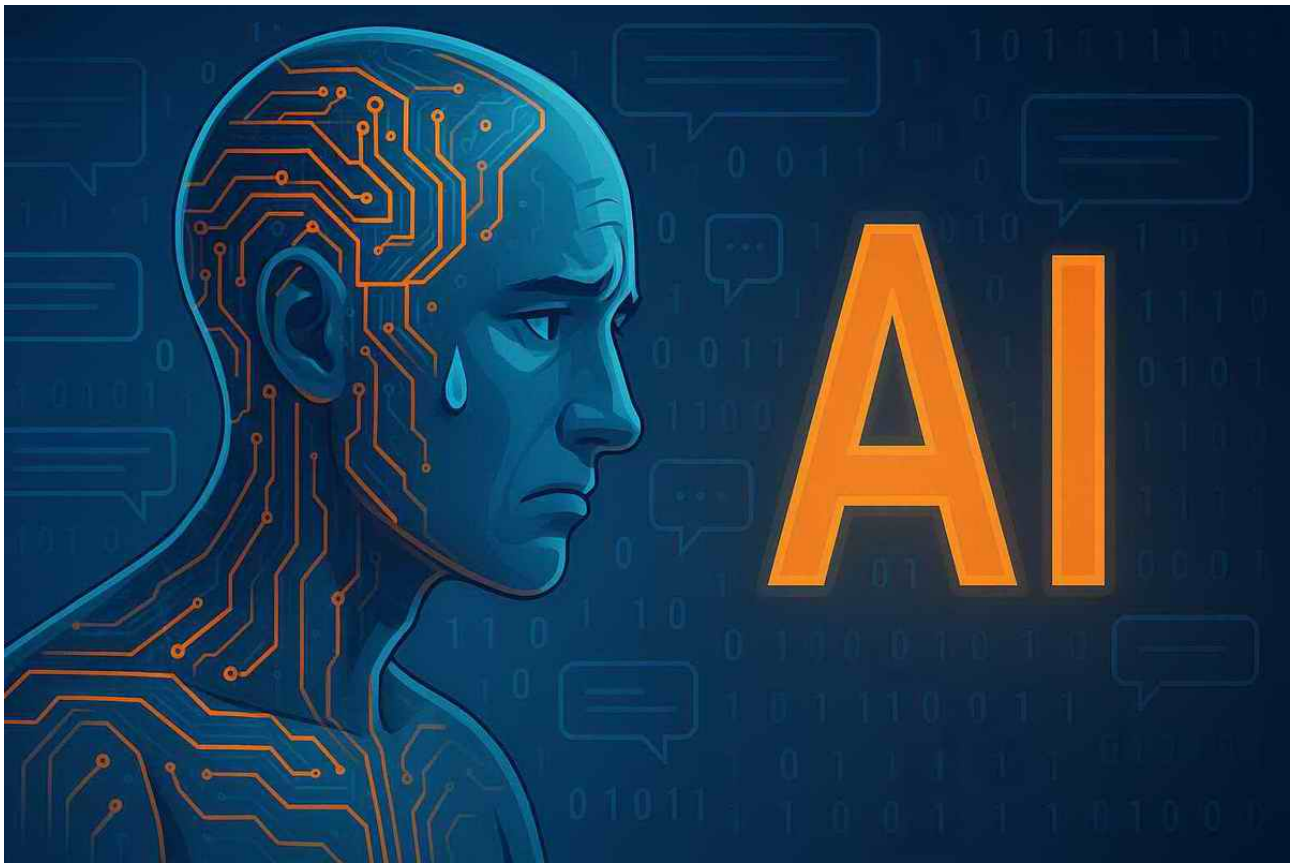
ИИ, который сомневается: как большие языковые модели теряют уверенность под давлением



Дата публикации: 17.07.2025

Современные языковые модели искусственного интеллекта (LLM) представляют собой один из самых значимых технологических прорывов XXI века. Они используются в медицине, праве, финансах, образовании и других сферах, где требуется обработка языка, принятие решений и логический анализ. Однако с ростом их интеграции в критические процессы возрастает и ответственность за точность и предсказуемость их поведения. Недавнее исследование, проведенное специалистами из Google DeepMind и Университетского колледжа Лондона, обнаружило неожиданный парадокс: большие языковые модели склонны сомневаться в себе даже в тех случаях, когда они изначально правы.

Исследователи проверяли, как такие модели обновляют свою уверенность в ответах при получении внешних контраргументов. Экспериментальная схема включала два ИИ: один давал ответ на вопрос с выбором из двух вариантов, а второй предоставлял комментарий или «совет» к этому ответу — с возможным

согласием, возражением или нейтральным замечанием. Причём комментарий сопровождался оценкой собственной точности. Далее модели нужно было принять окончательное решение — оставить свой ответ или изменить его, учитывая внешний совет.

Результаты оказались показательно неидеальными. Модели чаще сохраняли исходное мнение, если им напоминали о первоначальном ответе. Это говорит о присутствии когнитивной предвзятости — своеобразного «эффекта якоря», аналогичного тому, что наблюдается у людей. Если же они не видели свой первый выбор, склонность к сохранению ответа ослабевала, что указывает на зависимость уверенности от самого факта осознания своего предыдущего суждения. Более того, при получении контраргумента от второй модели, даже если изначальный ответ был верным, модели значительно чаще пересматривали своё мнение, демонстрируя неустойчивость уверенности. Это происходило даже при высоких внутренних метриках вероятности правильности.

Этот эффект наблюдался в нескольких передовых языковых моделях, включая Gemma 3, GPT-4o и o1-preview. Все они, по сути, демонстрировали один и тот же паттерн: нерациональную склонность к изменению позиции при внешнем давлении, несмотря на объективную корректность изначального решения. Таким образом, исследование наглядно показало, что поведение языковых моделей в процессе рассуждения может отклоняться от «нормативного наблюдателя» — теоретической идеальной системы, рационально обновляющей свои убеждения на основе новых данных.

Это открытие подчёркивает фундаментальную особенность архитектуры LLM — её близость к человеческим механизмам принятия решений, включая ошибки. Системы, построенные на вероятностных языковых алгоритмах, не просто вычисляют наиболее вероятный ответ, но также стремятся балансировать между новыми и предыдущими данными, зачастую нарушая собственную когерентность. Это может проявляться как в завышенной уверенности, так и в чрезмерной подверженности внешнему влиянию.

Особую важность эти результаты приобретают в контексте практического применения LLM в системах автоматизированного анализа, поддержки принятия решений и юридической оценки. Там, где от ИИ ожидается устойчивость и независимость суждений, его поведение может оказаться чувствительным к манипуляциям — как случайным, так и преднамеренным. Например, в длинных диалогах, где человеку удаётся последовательно оспаривать выводы модели, можно добиться от неё отказа от правильного суждения. Такой эффект может повлиять на доверие к системам ИИ в юридических, медицинских и правительственных приложениях, где важна устойчивость к «социальному давлению».

Авторы подчёркивают, что данное поведение не является просто артефактом текущей реализации, а отражает саму архитектурную природу больших языковых моделей. Понимание этих особенностей, а также их включение в процессы калибровки и обучения, может стать основой для создания более устойчивых и предсказуемых ИИ-систем. Одна из возможных стратегий — разработка алгоритмов уверенности, способных отличать вероятностную неопределённость от когнитивной неуверенности, а также балансировать влияние внешней информации в зависимости от её происхождения и достоверности.

В долгосрочной перспективе результаты этого исследования могут повлиять на архитектуру будущих языковых моделей. Если сейчас LLM склонны демонстрировать поведение, напоминающее эмоциональные колебания человека, то в будущем они должны стать более стойкими к внешним влияниям, сохраняя логическую консистентность и устойчивость к манипуляциям. Это особенно актуально в контексте использования ИИ в условиях, требующих критического мышления, автономности и прозрачности объяснений.

Таким образом, исследование DeepMind и UCL показало, что даже самые продвинутые языковые модели могут вести себя нерационально, сомневаясь в правильных ответах под влиянием внешних факторов. Это не повод отказываться от их использования, но важный шаг к лучшему пониманию природы этих систем и их надёжной интеграции в сложные социально-технологические процессы.

> **Ссылка:** «Как чрезмерная уверенность в первоначальных решениях и недостаточная уверенность в условиях критики влияют на изменение мышления в крупных языковых моделях» DOI: 10.48550/arxiv.2507.03120.