Ученые разработали программу для комплексной оценки эффективности работы ИИ-агентов

07.10.2026

Исследователи Санкт-Петербургского Федерального исследовательского центра РАН (СПб ФИЦ РАН) и Санкт-Петербургского отделения Математического института им. В. А. Стеклова РАН (ПОМИ РАН) разработали программное обеспечение, которое позволит оценить как цифровые помощники, использующие в своей работе технологии искусственного интеллекта (ИИ-агенты), справляются с задачами разной сложности. Разработка позволит проводить тестирование различных моделей на предмет ошибок или угроз, а также может применяться пользователями для индивидуального выбора ИИ-агентов под свои конкретные задачи. Результаты исследования опубликованы в научном журнале IEEE Access (Q1).

Сегодня в ведущих технологических странах мира в экономику и повседневную жизнь каждого человека активно внедряются ИИ-агенты — это программные системы, которые способны воспринимать окружение, формулировать цели, планировать действия и выполнять их с помощью искусственного интеллекта (больших языковых моделей) и подключённых инструментов (API, базы данных, устройства и т.д.). Они работают как “умные исполнители задач” как в реальных, так и в цифровых процессах. По данным из открытых источников, уже к 2030 году до 20% всех цифровых покупок будут совершаться автономными ИИ-агентами, а мировой рынок таких технологий вырастет в 30 раз.

Однако, как и любые цифровые системы ИИ-агенты могут допускать ошибки и совершать некорректные действия с информацией, что чревато не только утечкой конфиденциальных данных, но и серьезными экономическими рисками. Поэтому для их безопасного использования ученые СПб ФИЦ РАН ведут разработки цифровых инструментов для мониторинга того насколько корректно работают подобные решения.

«Мы разработали MultiCAT-Bench – это программное обеспечение, которое позволяет пользователям ИИ-агентов понять, какие именно факторы делают работу ИИ-агентов с инструментами сложной. Существующие аналоги не позволяют детально оценить, насколько сложные задачи способен решать тот или иной ИИ-агент и где проходят границы его надежности. Наша разработка помогает выявить потенциально уязвимые сценарии и понять, в каких условиях ИИ-агенты начинают чаще ошибаться, это важно как для оценки существующих моделей, так и для проектирования более надежных агентных систем с учетом выявленных ограничений», – отмечает руководитель Лаборатории прикладного искусственного интеллекта СПб ФИЦ РАН Максим Абрамов.

Программа для оценки ИИ-агентов позволяет проводить комплексную и детализированную оценку решения задач, которые делятся на 10 категорий по уровню сложности (число вызываемых и доступных функций агента, структура и тип параметров, сложность формулировок и «лингвистический шум» и проч.). Всего MultiCAT-Bench содержит 3789 тестовых задач.

Эксперименты по сравнению качества работы различных ИИ-агентов показали, что модели особенно сильно теряют качество при длинных цепочках действий и большом числе параметров, причем чаще пропускают необходимые действия, чем совершают лишние.

«С помощью разработанного теста мы сравнили десять моделей и выявили несколько важных закономерностей. Самый распространенный тип ошибки — пропуск необходимого действия. Кроме того, моделям сложнее дается скорее «ширина», чем «глубина» задачи: сильнее всего точность снижается при росте числа необходимых вызовов инструментов — примерно в 1,6 раза. Неожиданно сложным оказался и мультипользовательский сценарий, когда один запрос содержит поручения от разных людей. Наконец, мы обнаружили еще одну важную особенность: рассуждающие модели значительно лучше распознают ситуации, когда для выполнения задачи не хватает данных и необходимо запросить дополнительную информацию у пользователя», – поясняет старший научный сотрудник лаборатории математической логики и дискретной математики ПОМИ РАН **Сергей Николенко**.

Работа помогает понять, в каких условиях ИИ-агенты становятся менее надежными, и проектировать системы с учетом этих ограничений. Например, сложные задачи можно разбивать на более короткие последовательности действий, дополнительно контролировать передачу важных параметров и предусматривать механизмы уточнения информации в случаях, когда исходных данных недостаточно.

«Наше программное обеспечение могут использовать разработчики и тестировщики при проектировании ИИ-агентов, а пользователи, которые применяют такие модели, могут подобрать себе наиболее эффективного ИИ-агента для той или иной сферы, будь то, например, работа с клиентами, финансы, электронная коммерция, юриспруденция, или подбор кадров», – дополняет стажер-исследователь Лаборатории прикладного искусственного интеллекта СПб ФИЦ РАН **Александр Попцов**.

Основная часть исследования выполнена сотрудниками СПб ФИЦ РАН, в работе также участвовал ученый Санкт-Петербургского отделения Математического института им. В. А. Стеклова РАН. Проект поддержан грантом Минобрнауки по направлению “Молодежные лаборатории”. Разработка соответствует Концепции научно-технологического развития Санкт-Петербурга в области создания передовых методов оценки и тестирования систем искусственного интеллекта для критических применений.