Специалисты Центра практического искусственного интеллекта Сбербанка создали метод, который значительно уменьшает количество недостоверных ответов больших языковых моделей. Разработка позволяет точнее определять случаи, когда искусственный интеллект выдаёт правдоподобную, но ложную информацию.

Новый подход демонстрирует повышение точности обнаружения некорректных ответов почти на 30% по сравнению с существующими методами. Особенность решения заключается в возможности эффективной работы с малым объёмом данных — для обучения достаточно всего 250 примеров.

Метод предназначен для использования в RAG-системах, которые являются ключевым компонентом современных мультиагентных решений искусственного интеллекта. Эти системы работают с контекстно-зависимыми вопросами и ответами.

Разработка позволяет компаниям экономить ресурсы, которые ранее требовались для масштабной разметки данных. Улучшение качества детекции ошибок ИИ способствует повышению надёжности и доверия к искусственному интеллекту в промышленных решениях.

Результаты исследования были представлены на международной конференции SIGIR 2025. Работа учёных Сбера способствует к решению одной из наиболее актуальных проблем в области искусственного интеллекта — минимизации рисков распространения недостоверной информации.