Специалисты Сбера разработали модель GigaEmbeddings, способную глубоко понимать и обрабатывать тексты на русском языке. Новинка была представлена на конференции ACL 2025 и нацелена на решение практических задач в бизнесе, таких как поиск, анализ обращений и построение чат-ботов.

Модель построена на базе GigaChat-3B и прошла трёхэтапное обучение: предварительное, тонкую настройку и мультизадачное. Архитектура была оптимизирована, что позволило сократить объём параметров на четверть без ухудшения качества. GigaEmbeddings уже доступна для использования на GitVerse и HuggingFace.

На рынке до сих пор не хватало эффективных решений для работы с русским языком. Большинство существующих инструментов требовали значительных ресурсов или давали слабые результаты при обработке текстов. Новая модель закрывает этот пробел, предлагая универсальный инструмент для различных отраслей.

GigaEmbeddings подходит для e-commerce, где особенно важен точный анализ пользовательских запросов, а также для создания интеллектуальных чат-ботов и RAG-систем, применяемых, например, в банках.