

Анастасия Владимировна Колмогорова,
д. филол. н.
Национальный исследовательский
университет «Высшая школа экономики»,
Санкт-Петербург, Россия
э-почта: akolmogorova@hse.ru

Anastasia Vladimirovna Kolmogorova,
Dr. of Sc. (Philology)
National Research University Higher School
of Economics, St. Petersburg, Russia
e-mail: akolmogorova@hse.ru

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ В ПРИКЛАДНЫХ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ: ФУНКЦИОНАЛ И ОГРАНИЧЕНИЯ

Описан опыт использования Больших языковых моделей в прикладных лингвистических проектах, выполняемых лабораторией языковой конвергенции НИУ ВШЭ-Санкт-Петербург.

Ключевые слова: большие языковые модели; медицина; цифровизация музеев; маркетинговая коммуникация; прикладная лингвистика.

LARGE LANGUAGE MODELS IN APPLIED LINGUISTIC RESEARCH: FUNCTIONALITY, LIMITS, AND CONSTRAINTS

The experience of using Large Language Models in applied linguistic projects carried out by the Language Convergence Laboratory at HSE University-St. Petersburg is described.

Key words: Large Language Models; medicine; digitalization of museums; marketing communication; applied linguistics.

Прикладные исследования постепенно становятся мейнстримом в лингвистике. Это стало возможным благодаря двум факторам: 1) появлению больших текстовых данных и/или инструментов для их быстрого сбора (прежде всего, различным техникам веб-скрепинга) и 2) созданию больших языковых моделей (БЯМ), которым лингвисты-разработчики отныне могут делегировать ряд задач.

На основе использования few-shot промптинга [1] БЯМ могут с успехом применяться для разметки лингвистических данных (по речевым жанрам, сентименту, пропозициям, темам), для генерации синтетических датасетов и для модификации текстов по тем или иным параметрам.

Так, работая над проектом по поиску предикторов успешности маркетинговой коммуникации, мы применили БЯМ для речежанровой разметки скриптов устной речи торговых представителей общим объемом звучания в 80 часов. На тестировочной выборке точность и полнота разметки составила 0.91. Проведенная таким образом разметка большого корпуса позволила сделать выводы о жанровом распределении высказываний у более успешных и менее успешных торговых представителей, о среднем времени развертывания высказывания в том или ином жанре, а также последовательностях жанров в выборке.

В ходе работы над проектом по созданию системы векторного поиска картин по пользовательскому описанию мы использовали БЯМ для генерации датасета описаний картин из Цифровой коллекции Эрмитажа, чтобы в дальнейшем векторизировать их и использовать как основу для поиска. С одной стороны, это позволило быстро собрать необходимые данные, но с другой – обнаружило существенные отличия синтетических текстов от написанных информантами [2].

Наконец, мы использовали БЯМ в рамках проекта по упрощению текстов для пациентов с афазией. В системный промпт были имплементированы экспертные инструкции по упрощению. Получившиеся результаты валидированы в клинической практике и продемонстрировали хорошие метрики.

Кроме того, рассмотрены основные ограничения БЯМ: склонность к так называемым галлюцинациям, чувствительность к языку и прагматике и стратегиям промптинга.

ЛИТЕРАТУРА

1. Sparks of artificial general intelligence: Early experiments with gpt-4 / S. Bubeck [et al.]. ArXiv, abs/2303.12712, 2023.
2. Колмогорова А. В., Налобина П. А. Осмысля пространство: как используют схемы топологической семантики естественные и искусственные когнитивные агенты (на материале описаний картин из коллекции Эрмитажа) // Философия и эпистемология науки. 2025. № 1. С. 170–185.