

Предисловие от редактора

Уважаемые коллеги,

Рад представить вашему вниманию очередной том серии «Записок научных семинаров ПОМИ», продолжающий подсерию «Исследования по прикладной математике и информатике» и посвящённый прежде всего искусственному интеллекту, машинному обучению и обработке естественного языка. Далее я кратко представлю содержание работ.

И. Алексеевская и К. Архипенко предлагают атаку Soft-Prompt, обходящую защитные механизмы больших языковых моделей с помощью «мягкого» перефразирования вредоносных запросов ансамблем моделей. Г. Э. Ангени и В. А. Ершов вводят клиентообусловленное спекулятивное декодирование, ускоряющее генерацию языковых моделей в мультиарендных сервисах за счёт учёта статистики запросов отдельных клиентов. О. Барышников и соавторы представляют систему Query2Diagram, автоматически строящую UML-диаграммы в ответ на запросы разработчиков. Н. Бутаков с соавторами предлагают эффективный алгоритм планирования распределённых AutoML-конвейеров с ограничениями по времени выполнения. Ю. Дементьев и соавторы получают новые результаты о полудуплексной коммуникационной сложности. С. Елистратов и соавторы исследуют направленный шум и неявную регуляризацию в оптимизаторе Lion.

Д. Галимзянова и соавторы описывают CleanComedy — методы порождения «дружелюбного», свободного от токсичности юмора. Р. Гарипов, А. Забашта и А. Фильченков предлагают архитектуру глубокого обучения для извлечения мета-признаков из табличных данных. И. Иов и соавторы исследуют применение больших языковых моделей для оптимизации конструирования признаков в AutoML. А. Лаэтин, Б. Пулфер и С. Смирнов предлагают иерархическую инициализацию временных масштабов в архитектуре Mamba-2 для многомасштабного моделирования последовательностей. И. А. Молодецких и соавторы разрабатывают методологию сбора и анализа субъективной заметности артефактов нейросетевых моделей суперразрешения.

У меня в этом томе выходят два обзора, точнее, большой обзор в двух частях: сначала обзор методов фильтрации и оценки обучающих

данных для больших языковых моделей, а потом, во второй работе, обзор работ о функциях влияния данных и их масштабировании.

Р. Обловатный и соавторы предлагают метод обнаружения галлюцинаций в языковых моделях с дополненной поиском генерацией (RAG), основанный на вероятностных расстояниях. Е. Орлов и В. Малых изучают ранжирование ответов на Stack Overflow в текстовой постановке. Завершает том работа Е. А. Шумицкой, Д. С. Ватолина и А. В. Анциферовой об уязвимостях рандомизированного сглаживания в регрессионных моделях и адаптивных атаках на метрики качества изображений.

Мы будем рады принять работы российских и зарубежных учёных по искусственному интеллекту и другим областям прикладной математики и информатики к рассмотрению для публикации в следующих томах серии.

Сергей Николенко,
д.ф.-м.н., ПОМИ РАН