**Введение**

Добрый день! Меня зовут Голиков Алексей Александрович, я являюсь аспирантом Елабужского института Казанского Федерального Университета. Сегодня я представлю свой доклад «Анализ эффективности ансамблирования больших языковых моделей».

**Наведение на левую верхнюю область «Улучшение качества ответов больших языковых моделей с течением времени»**

В последние годы большие языковые модели вроде ChatGPT начинают обладать всё большими возможностями. На графике отмечено улучшение качества прохождения моделями теста на общие знания (бенчмарк MMLU), из которого следует, что мы постепенно приближаемся к достижению моделями уровня эксперта во всех областях сразу. И данный процесс лишь ускоряется с течением времени – за последний год вышло множество больших языковых моделей, которые по интеллекту оставляли позади прошлые модели. В данном исследовании в качестве языковых моделей были взяты передовые на момент проведения исследования модели ChatGPT-4, Nemotron-4-340B, Gemini-1.5-Pro, Qwen-2-72B.

**Наведение на правую верхнюю область «Тесты предметного модуля «Современный русский язык» ЕИ КФУ»**

Основными способами оценки качества работы больших языковых моделей являются результаты прохождения ими различных тестов на общие и специализированные знания. В данной работе большим языковым моделям предлагалось ответить на вопросы тестов предметного модуля «Современный русский язык», использующихся в Елабужском институте Казанского Федерального Университета – всего 185 вопросов из 6 блоков. На слайде приведен пример одного из вопросов.

**Наведение на левую среднюю область «Точность различных языковых моделей»**

Вышеупомянутые модели на предложенном тесте демонстрируют точность, представленную на диаграмме – лучшие результаты показали ChatGPT-4 и Nemotron-4-340B – 64,3%.

**Наведение на правую среднюю область «Корреляция ответов различных языковых моделей»**

При этом видим, что ответы моделей имеют умеренную корреляцию – от 0,48 до 0,66. Это означает, что данные модели, хотя дают и близкие по точности результаты, но отвечают правильно на разные вопросы. Это дает шанс попробовать объединить способности различных моделей, чтобы повысить итоговую точность. В классическом машинном обучении это называется ансамблирование – техника, с помощью которой зачастую побеждают на соревнованиях по анализу данных. Новизна данного исследования заключается в применении ансамблирования больших языковых моделей.

**Наведение на левую нижнюю область «Метод случайного леса для формирования ансамбля LLM»**

Для ансамблирования больших языковых моделей были реализованы алгоритмы логистической регрессии, градиентного бустинга и случайного леса с разделением множества вопросов и ответов на часть для обучения моделей (80% случайно отобранных вопросов и ответов моделей на них) и часть для тестирования (остальные 20%). Признаками для обучения алгоритмов в данном случае являлись ответы отдельных моделей, а целевой переменной – заранее известные правильные ответы. При обучении алгоритмов выполнялась оптимизация их гиперпараметров методом случайного поиска. Наилучший результат продемонстрировал алгоритм случайного леса, который заключается в голосовании отдельных деревьев решений, обученных на случайных подмножествах признаков и данных.

**Наведение на правую нижнюю область «Метод случайного леса для формирования ансамбля LLM»**

Точность ансамбля LLM в реализации случайного леса составила 67,6%, что на 4,3% лучше, чем отдельной самой точной модели.

**Заключение**

Таким образом, в работе было продемонстрировано, что ансамблирование больших языковых моделей способно улучшить итоговое качество ответов, однако важно выбирать достаточно близкие по качеству модели, чьи ответы при этом не имеют существенной корреляции.