РЕЦЕНЗИЯ
на выпускную квалификационную работу

Беликовой Юлии Андреевны
на тему «Применение white-box подходов для улучшения вопросно-ответных навыков в больших языковых моделях»

Работа посвящена актуальной проблеме контекстуальных галлюцинаций в вопросно-ответных системах, использующих большие языковые модели. С учетом широкого применения языковых моделей в критически важных областях, проблема генерации фактически некорректных ответов представляет серьезную угрозу для надежности таких систем. Поэтому исследование методов детекции и минимизации галлюцинаций является важной задачей направления обработки естественного языка.

Теоретическая значимость работы заключается в разработке новых подходов к детекции контекстуальных галлюцинаций через ансамблирование топологического анализа карт внимания с метриками неопределенности и модификацию стандартных методов пробинга для обучения на малом количестве данных, а также в создании методологии автоматического формирования пар предпочтений для дообучения языковых моделей минимизировать контекстуальные галлюцинации. Практическая ценность определяется адаптацией методов для работы в условиях ограниченных ресурсов и их применимостью в прикладных RAG (Retrieval-Augmented Generation) задачах. Достоверность результатов подтверждается экспериментальной валидацией на академических и индустриальных датасетах с использованием общепринятых метрик оценки. Эксперименты проведены на разных архитектурах моделей со статистически значимыми результатами.

К положительным сторонам работы относятся качественный обзор литературы, отличная структура изложения, инновационная методология и практическая ориентированность решения. Беликова Ю.А. демонстрирует глубокое понимание предметной области и умение систематизировать существующие подходы. Среди ограничений работы следует отметить фокус на идеальном контексте, тогда как в реальных RAG-системах контекст может быть нерелевантным, что требует дополнительного исследования.

Магистерская диссертация заслуживает оценки «отлично». Рекомендую присвоить Беликовой Ю.А. квалификацию магистра по направлению **09.04.01 «Информатика и вычислительная техника»**.

Рецензент:
Миле Митрович, PhD,
Исполнительный директор по исследованию данных ПАО Сбербанк

20.06.2025