Автономная некоммерческая организация
«Институт Искусственного интеллекта»

121170, Москва, Кутузовский пр-т, 32, к. 1, 4.В.08
E-mail: info@airi.net, Web: airi.net
ОГРН 1207700493978 ИНН 7730261209 КПП 773001001

ОТЗЫВ НАУЧНОГО РУКОВОДИТЕЛЯ

на дипломную работу Беликовой Юлии Андреевны

на тему: Применение white-box подходов для улучшения вопросно-ответных навыков в больших языковых моделях

Выпускная квалификационная работа Беликовой Ю.А. посвящена решению актуальной проблемы борьбы с контекстуальными галлюцинациями в больших языковых моделях (БЯМ). Эта проблема является ключевым барьером для надёжного применения вопросно-ответных систем, особенно в таких критически важных областях, как медицина и финансы. В работе предложены и апробированы перспективные white-box подходы, которые повышают фактологическую точность моделей за счёт анализа их внутренних состояний.

Предложенная методология носит комплексный характер и отличается как теоретической, так и практической новизной. В теоретическом плане впервые предложена методология необучаемой детекции галлюцинаций, основанная на ансамблировании топологического анализа графов внимания (метод MTopDiv) и вероятностных метрик неопределенности модели. Кроме того, разработана оригинальная методология минимизации галлюцинаций путем дообучения модели методом прямой оптимизации предпочтений (DPO) на автоматически сгенерированных парах. Практическая значимость работы заключается в разработке алгоритмов, адаптированных для применения в условиях ограниченных данных и вычислительных ресурсов. Предложенные подходы продемонстрировали высокую эффективность на широком спектре стандартных и индустриальных наборов данных. Достоверность полученных результатов подтверждается широкой серией вычислительных экспериментов с использованием общепринятых метрик и современных языковых моделей.

К сильным сторонам работы относятся глубокий обзор литературы, логичная структура и апробация ключевых подходов на ведущих международных конференциях, что подтверждает их научную значимость. В качестве ограничений и направлений для будущей работы следует отметить необходимость проверки масштабируемости методов на более крупных и разнообразных архитектурах, а также их адаптации для эффективной работы с современными длинными контекстами.

В процессе работы над диссертацией Беликова Ю.А. проявила себя как самостоятельный, инициативный и высокомотивированный исследователь. Она неоднократно успешно представляла результаты своих исследований на научных семинарах, демонстрируя отличные навыки публичных выступлений и ведения научной дискуссии.

Выпускная квалификационная работа Беликовой Ю.А. является завершенным научным исследованием, соответствует всем требованиям, установленным к ВКР студентов МФТИ,

и я рекомендую присвоить ей квалификацию магистра по направлению 09.04.01 «Информатика и вычислительная техника».

Научный руководитель, Коновалов Василий Павлович
к.т.н., старший научный сотрудник
Вычислительная семантика
АНО «Институт искусственного интеллекта»

Подпись научного руководителя, сотрудника АНО «Институт искусственного интеллекта»,

удостоверяю
HR Директор – начальник отдела
АНО «Институт искусственного интеллекта» Звонарева М.И.