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 *Савкина Максима Константиновича*
на тему: *Генерация синтетического датасета для детекции именованных сущностей с применением больших языковых моделей*

В выпускной работе исследуются методы генерации синтетических текстовых данных для задачи распознавания именованных сущностей (NER) с использованием больших языковых моделей. В рамках исследования разработаны две отдельные методологии: первая предназначена для генерации синтетических данных в задаче детекции персональной информации (PII); вторая - для генерации синтетического датасета NER в открытом домене с широким покрытием тематик и типов сущностей. Полученные датасеты были использованы для обучения и тестирования моделей, показавших высокое качество решения соответствующих задач.

Актуальность работы обусловлена несколькими ключевыми факторами. Во-первых, ручная разметка данных для задач извлечения именованных сущностей (NER) трудоемкая и дорогостоящая. Во-вторых, генерация синтетических текстов предоставляет значительно больший контроль над содержанием и стилем текстов - это позволяет создавать корпуса, адаптированные под конкретные прикладные задачи. Наконец, особенно остро стоит проблема обнаружения персональных данных (PII), где число открытых датасетов крайне ограничено из-за невозможности публикации персональных сведений. Генерация искусственных текстов и персональных данных позволяет создавать реалистичные примеры текстов с PII, не раскрывая при этом реальную пользовательскую информацию.

Достоверность полученных результатов подтверждается тем, что в открытый доступ выложен код для воспроизведения экспериментов. Проведен анализ результатов, позволяющий оценить эффективность предложенного метода. Однако стоит отметить, что в работе отсутствует усреднение метрик по нескольким запускам и не приведены значения стандартного отклонения, что затрудняет оценку устойчивости модели и снижает воспроизводимость результатов.

Предложенные в работе методологии позволяют генерировать датасеты с произвольной структурой, стилистикой и любыми наборами именованных сущностей. Это дает возможность быстро и с минимальными затратами создавать корпуса, адаптированные под конкретные задачи, существенно снижая необходимость ручной разметки. Обученная на этих синтетических данных модель опубликована в открытом доступе и регулярно скачивается, что подтверждает её востребованность на практике. Дополнительно, наличие публикаций на конференциях уровня A/A* показывает признание значимости результатов со стороны научного сообщества.

К сильным сторонам работы можно отнести 1) разработку востребованных датасетов, что подтверждается их практическим использованием и загрузками; 2) публикации на международных конференциях, свидетельствующие о научной значимости полученных результатов. Вместе с тем, одним из недостатков является незавершенность работы по второму датасету, для него требуется более глубокий и обширный анализ.

Характеристика студента:

В ходе работы студент проявил инициативность, способность к постановке исследовательских задач, а также умение последовательно решать их с использованием актуальных инструментов.

Студент изучил значительный объем профильной научной литературы, включая новейшие публикации, релевантные тематике научной работы. Он провёл обзор существующих подходов, выявил их ограничения и предложил собственные решения, продемонстрировав глубокое понимание предметной области.

Студент выступил с постерами на двух международных NLP-конференциях EMNLP'2024 и NAACL'2025 и четырех московских постерных сессиях (Sber RnD Day'2024, Fall into ML'2024, IT Purple'2025 и DataFest'2025), представляя результаты исследований, выполненных в рамках выпускной квалификационной работы. Кроме того, он провел открытый семинар по теме диплома. Всё это подтверждает высокий уровень сформированных навыков публичных выступлений и научной коммуникации.

Заключение:
Выпускная квалификационная работа соответствует всем требованиям, установленным к ВКР студентов МФТИ. Работа отличается научной новизной, высокой степенью проработки, практической значимостью и качеством оформления. Студент заслуживает присвоения квалификации магистра.

Оценка выпускной работы: отлично.

Научный руководитель, *Коновалов Василий Павлович*
к.т.н., *старший научный сотрудник,*
подразделение вычислительной семантики,
АНО «Институт искусственного интеллекта»

Подпись рецензента, сотрудника АНО «Институт искусственного интеллекта»,

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