УДК 004.8

ДЕТЕКЦИЯ ИСПОЛЬЗОВАНИЯ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ В СТУДЕНЧЕСКИХ РАБОТАХ: МЕТОДЫ, ОГРАНИЧЕНИЯ И ЭТИЧЕСКИЕ АСПЕКТЫ

Кацубин В.А.[1]

студент,

Поволжский государственный университет телекоммуникаций и информатики,

Самара, Россия

Аннотация

Статья посвящена анализу методов автоматической детекции текстов, сгенерированных большими языковыми моделями (LLM), в академическом контексте. Рассматриваются три класса методов - статистические, стилометрические и гибридные, - а также их практические ограничения: высокий уровень ложноположительных срабатываний, уязвимость к постпроцессинговым атакам и зависимость от языка и домена. Особое внимание уделяется этическим аспектам: рискам для академической справедливости, дифференцированному воздействию на отдельные группы студентов и вопросам конфиденциальности. Предлагаются принципы ответственного применения инструментов детекции в образовательных учреждениях.

Ключевые слова: большие языковые модели, детекция ИИ-текстов, академическая честность, этика ИИ, генеративный ИИ, образование.

[1] Научный руководитель: **Куваева Е.Н.**, ст. преподаватель, Поволжский государственный университет телекоммуникаций и информатики, Самара, Россия.
Kuvaeva E.N., Senior Lecturer, Povolzhskiy State University of Telecommunications and Informatics, Samara, Russia.

DETECTION OF GENERATIVE MODEL USAGE IN STUDENT WORKS: METHODS, LIMITATIONS AND ETHICAL ASPECTS

Kacubin V.A.

Student,

Povolzhskiy State University of Telecommunication and Informatics,

Russia, Samara

Abstract

The article analyzes methods for automatic detection of texts generated by large language models (LLMs) in academic contexts. Three classes of methods are examined - statistical, stylometric, and hybrid - along with their practical limitations: high false-positive rates, vulnerability to post-processing attacks, and language-domain dependency. Special attention is paid to ethical aspects: risks to academic fairness, differential impact on student groups, and privacy concerns. Principles for the responsible use of detection tools in educational institutions are proposed.

Keywords: large language models, AI text detection, academic integrity, AI ethics, generative AI, education.

Широкое распространение больших языковых моделей (LLM) - в первую очередь ChatGPT, Gemini и Claude - радикально изменило образовательную среду. По данным опроса Европейского студенческого союза (2023), более 60% студентов используют генеративный ИИ при подготовке учебных работ [1]. Это порождает серьёзные вопросы об академической честности и достоверности оценивания знаний. Академические учреждения реагируют внедрением автоматических детекторов, однако их надёжность вызывает споры [2; 3]. Цель настоящей статьи - систематизировать современные методы детекции ИИ-

текстов, выявить их технические ограничения и проанализировать сопутствующие этические риски.

LLM - это нейронные сети, обученные на масштабных корпусах текстов и способные генерировать связный, стилистически разнообразный текст по заданному запросу [4]. В академическом контексте их использование варьируется от вспомогательного (исправление грамматических ошибок, структурирование черновика) до полного замещения самостоятельной работы студента. Perkins et al. [5] выделяют четыре уровня вовлечённости: использование ИИ в качестве справочника, инструмента редактирования, помощника в написании и, наконец, замены автора. Именно последний уровень представляет наибольшую угрозу для академической честности и является основным объектом детекции.

Статистические детекторы основаны на анализе вероятностных свойств текста. Метод DetectGPT [6] опирается на гипотезу о том, что LLM-тексты занимают локальные максимумы логарифмической вероятности в пространстве модели: незначительные перефразировки снижают вероятность, тогда как для человеческих текстов это нехарактерно. Техника «водяных знаков» [7] предусматривает встраивание скрытых статистических паттернов на этапе генерации. Инструмент GLTR [8] визуализирует распределение токенов по вероятности, помогая эксперту выявлять характерно предсказуемые участки. Однако статистические методы демонстрируют высокую чувствительность к стилю конкретной модели и ухудшают качество при работе с короткими текстами (менее 150 слов).

Стилометрические подходы анализируют лексические и синтаксические паттерны, типичные для LLM-текстов: избыточное использование клише, равномерную длину предложений, ограниченное лексическое разнообразие и гиперкоррекцию пунктуации. На практике коммерческие детекторы (Turnitin AI, GPTZero) комбинируют стилометрические признаки с нейросетевой

классификацией, что повышает точность, но не устраняет принципиальных ограничений метода.

Гибридные системы объединяют несколько методов детекции, дополняя их контекстуальным анализом (поведение студента в LMS, история редактирования документа). Перспективным направлением является мультиагентная верификация, при которой несколько независимых классификаторов голосуют за итоговое решение, снижая дисперсию ошибок.

Фундаментальной проблемой остаётся высокий уровень ложноположительных срабатываний (False Positive Rate, FPR). В исследовании Liang et al. было показано, что при анализе эссе студентов-неносителей английского языка FPR достигает 61,3%. Это означает, что честная работа может быть ошибочно квалифицирована как сгенерированная ИИ, что влечёт серьёзные академические последствия для студента.

Несложные трансформации текста - перефразирование, синонимическая замена, изменение порядка слов - существенно снижают вероятность обнаружения ИИ-генерации большинством детекторов. Показано, что даже однократное прохождение текста через инструмент перефразирования снижает точность детектора GPTZero с 94% до 42%.

Большинство детекторов обучены преимущественно на англоязычных корпусах и показывают значительно худшие результаты для русскоязычных, арабских и китайских текстов. Аналогичная проблема существует на уровне академических дисциплин: детектор, настроенный на гуманитарные тексты, теряет эффективность при анализе технических или медицинских работ.

По мере выхода новых поколений LLM детекторы устаревают: статистические свойства генерируемых текстов изменяются, а обучение детектора на текстах GPT-3 делает его менее эффективным против GPT-4 и последующих моделей. Это создаёт гонку вооружений между разработчиками моделей и детекторов.

Автоматическая детекция противоречит принципу презумпции невиновности, если вероятностный вывод алгоритма рассматривается как доказательство нарушения. Использование детектора как единственного доказательства в дисциплинарном расследовании недопустимо с точки зрения процессуальной справедливости.

Высокий FPR для отдельных групп студентов (не носители языка, студенты с дислексией, представители незападных образовательных традиций) означает, что детекторы де-факто создают дискриминирующую среду. Это нарушает принципы равенства образовательных возможностей.

Акцент на контроле, а не на обучении смещает педагогический фокус. Ряд исследователей указывает, что развитие у студентов навыков критического мышления, академической рефлексии и цифровой грамотности более эффективно противодействует злоупотреблениям ИИ, чем технический мониторинг [3; 5].

Передача студенческих работ в облачные детекторы порождает вопросы соответствия требованиям GDPR (в ЕС) и аналогичного российского законодательства о персональных данных. Провайдеры детекторов нередко не раскрывают политику хранения и использования загруженных текстов.

На основании проведённого анализа можно сформулировать четыре принципа ответственного применения инструментов детекции. Принцип 1 - многофакторная оценка: вывод детектора должен рассматриваться как один из факторов в совокупности с очной проверкой знаний, анализом истории редактирования и устным собеседованием [6]. Принцип 2 - прозрачность: студенты должны заранее знать, что их работы проходят автоматическую проверку и какие последствия влечёт обнаружение. Принцип 3 - право на апелляцию: должна существовать независимая процедура обжалования результатов детекции с участием компетентного эксперта. Принцип 4 - педагогический приоритет: внедрение детекторов должно сопровождаться

разработкой альтернативных форм контроля - проектных работ, устных защит, портфолио, - снижающих мотивацию к использованию ИИ.

Детекция ИИ-сгенерированных текстов представляет собой технически сложную и этически неоднозначную задачу. Ни один из существующих методов не обеспечивает приемлемого сочетания точности и справедливости. Перспективным направлением является не совершенствование детекторов само по себе, а реформирование форматов академических заданий, ориентированных на демонстрацию мышления, а не воспроизведение текста. Ответственный подход требует институционального регулирования, прозрачности и гарантий академической справедливости для всех категорий студентов.

Библиографический список

1. Бхатт У., Гассеми М., Веллер А. Кто несёт ответственность за вред, причинённый ИИ в здравоохранении // The Lancet Digital Health. - 2022. - Т. 4, № 3. - С. e147–e148.
2. Коттон Д. Р. Э., Коттон П. А., Шипвей Дж. Р. Чат-боты и списывание: обеспечение академической добросовестности в эпоху ChatGPT // Innovations in Education and Teaching International. - 2024. - Т. 61, № 2. - С. 228–239.
3. Европейский студенческий союз. Опрос об использовании ИИ в высшем образовании / Политический документ ESU. - Брюссель, 2023. - 34 с.
4. Флориди Л., Кириатти М. GPT-3: природа, охват, ограничения и последствия // Minds and Machines. - 2020. - Т. 30, № 4. - С. 681–694.
5. Германн С., Штробельт Х., Раш А. М. GLTR: статистическое обнаружение и визуализация сгенерированного текста // Материалы конференции ACL 2019. - 2019. - С. 111–116.
6. Ипполито Д., Дакворт Д., Каллисон-Берч К., Экк Д. Автоматическое обнаружение сгенерированного текста проще всего, когда оно

обманывает и людей // Материалы конференции ACL 2020. - 2020. - С. 1808–1822.

7. Каснеци Э., Сесслер К., Кюхеманн С. и др. ChatGPT во благо? О возможностях и вызовах больших языковых моделей для образования // Learning and Individual Differences. - 2023. - Т. 103. - С. 102274.
8. Кирхенбауэр Дж., Гейпинг Дж., Вэнь Й. и др. Водяной знак для больших языковых моделей // Материалы конференции ICML 2023. - 2023. - С. 17061–17084.