УДК 004.891.2

РАЗРАБОТКА МЕТОДОВ ДЕТЕКТИРОВАНИЯ СГЕНЕРИРОВАННЫХ ЯЗЫКОВЫМИ МОДЕЛЯМИ ТЕКСТОВ

Ярулин Э.С. (ИТМО)
Научный руководитель – кандидат технических наук Шаламов В.В. (ИТМО)

Введение. В последние годы проблема автоматической детекции текстов, генерируемых искусственным интеллектом, стала особенно актуальной из-за широкого распространения больших языковых моделей (LLM). Тексты, создаваемые такими моделями, могут быть практически неотличимы от человеческого письма, что порождает вопросы о достоверности публикуемой информации и угрозу плагиата. Для обеспечения безопасности информационного пространства и сохранения доверия к цифровому контенту необходимо разработать эффективные методы выявления машинного происхождения текста. Наиболее перспективным направлением является анализ вероятностных характеристик токенов и синтаксических паттернов, что позволяет классифицировать текст без прямого сравнения с исходным корпусом [1].

Основная часть. С помощью современных математических моделей и алгоритмов машинного обучения решаются следующие ключевые задачи детекции сгенерированного текста:

Задача оценки вероятностных характеристик. Речь идёт о вычислении перплексии, кросс-перплексии и сопутствующих метрик, которые позволяют судить о том, насколько «ожидаемы» языковые конструкции для той или иной модели. Сравнивая поведение нескольких языковых моделей, можно выявлять скрытые паттерны, указывающие на машинную генерацию [1].

Задача анализа стилистических и лингвистических особенностей. При генерации текста искусственный интеллект может допускать специфические ошибки согласования, предпочитать определённые синтаксические структуры или «слишком ровно» распределять редкие слова. Выделяют ряд методов, связанных с синтаксическим парсингом, морфемным анализом (особенно важным в русском языке) и сравнением распределения лексем. Избыточная регулярность или, напротив, искусственно внедрённая хаотичность может приводить к некорректному стилю, который детекторы машинного текста выявляют [2].

В реальных условиях детектор должен обладать широким охватом: уметь распознавать тексты разной тематики (технические, публицистические, научные обзоры и т.д.), а также адаптироваться к изменениям в языковых моделях.

Выводы. Проведён анализ актуальных подходов к задаче детекции машинно-сгенерированных текстов. Разработаны принципы сегментации задач на вероятностные методы и лингвистический анализ. Выделены ключевые трудности, такие как использование обфускации и многостилевой генерации, что требует комплексных алгоритмов, способных сочетать статистические и синтаксические признаки.

Список использованных источников:

1. Brian Tufts, Xuandong Zhao, Lei Li. A Practical Examination of AI-Generated Text Detectors for Large Language Models // arXiv preprint arXiv:2412.05139 – 2024
2. Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, Nicholas Andrews. Few-Shot Detection of Machine-Generated Text using Style Representations // arXiv preprint arXiv:2401.06712. – 2024