УДК 004.89

КЛАССИФИКАЦИЯ ОСНОВНЫХ ОГРАНИЧЕНИЙ ЯЗЫКОВЫХ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ НЕЙРОННЫХ СЕТЕЙ

А. Н. Кузьмин

магистрант, Белорусский государственный университет, г. Минск, Беларусь, artem160104k@gmail.com

Научный руководитель **М. А. Седнина**

старший преподаватель, Международный институт дистанционного образования Белорусского национального технического университета, г. Минск, Беларусь, sednina@bntu.by

В статье рассмотрены и классифицированы основные ограничения языковых генеративных моделей нейронных сетей в случае их функционирования на предприятиях и в организациях. Оценено влияние выявляемых ограничений данных моделей на их инновационный потенциал. Обоснована необходимость дальнейшего изучения ограничений для описания пределов эффективности данных моделей и разработки новых метрик оценки эффективности их работы.

Ключевые слова: цифровизация; нейронные сети; информационные технологии.

CLASSIFICATION OF THE MAIN LIMITATIONS OF GENERATIVE LANGUAGE MODELS OF NEURAL NETWORKS

A. N. Kuzmin

master's student, Belarusian State University, Minsk, Belarus, artem160104k@gmail.com

Supervisor **M. A. Sednina**

senior lecturer, International Institute for Distance Education of the Belarusian National Technical University, Minsk, Belarus, sednina@bntu.by

The paper examines and classifies the main limitations of generative language models of neural networks in the case of their functioning in enterprises and organizations. The impact of the revealed limitations of these models on their innovation potential is estimated. The necessity of studying the limitations for describing the limits of the effectiveness of these models and developing new metrics for evaluating their effectiveness is substantiated.

Keywords: digitalisation; neural networks; information technologies.

В настоящее время популярность языковых генеративных моделей нейронных сетей растет (в августе 2025 года только ChatGPT использовало около 800 млн пользователей в неделю) [1]. Расширяются диапазоны их применения на предприятиях и в организациях.

Однако при внедрении языковых генеративных моделей нейронных сетей в различных предприятиях и организациях вопрос эффективности часто не находится в приоритете: основными критериями качества выступают стоимость внедрения и, собственно, подписки на мо-

дель. Неполное представление о других критериях в среднесрочной перспективе может привести к снижению общей эффективности работы данного инновационного решения, интегрированного в информационную систему и, возможно, к отказу от его массового использования.

Организации IT-сектора являются наиболее прогрессивными в плане внедрения данного инновационного решения по сравнению с другими, так как одно из назначений этого вспомогательного инструмента – изменение подхода к работе специалистов – проявляется наиболее сильно.

Следовательно, необходимо, основываясь на результатах предыдущих исследований о принципах работы данных моделей [2] выявить основные ограничения языковых генеративных моделей нейронных сетей локального характера (тех, которые наблюдаются при работе с данными инновационными решениями в IT-компаниях).

Известно, что нейронные сети рассматриваемого типа (архитектуры GPT) обучаются на больших массивах текстовых данных. В процессе обучения происходит формирование параметров с вероятностями выдачи, на основе которых формируются выходные данные.

Однако сам набор данных, состоящий из буквенных, цифровых и иных символов, можно разбить на отдельные сегменты, которые могут существовать независимо друг от друга и быть базой обучения для локальных языковых генеративных моделей нейронных сетей.

Следовательно, общий или же отдельные наборы данных будут иметь некоторые ограничения, из которых, в свою очередь, сформируются пределы эффективности – те точки, в которой весь потенциал модели исчерпается в рамках всей компании или отдела, так как языковые генеративные модели несовершенны, и все еще существует определенная доля ошибок.

При анализе основных ограничений учитывался критерий оптимальности в рамках набора данных вида X_n – минимальное число ошибок ε, стремящееся к 0 (для максимального полезного эффекта от модели необходимо, чтобы она совершала как можно меньшее число ошибок).

В ходе исследования было выявлено, что имеются ограничения, которые заданы разработчиками – размер контекстного окна, измеряемый в токенах; возможности самой модели, который могут быть директивно ограничены (пример – модель DeepSeek, к которой не подключен никакой из графических модулей, поэтому выходная информация – только текстовая, а также ограничения выдачи самой информации через базу обучения, что не дает развернуть использование данной модели более масштабно).

Ограничения можно задавать и самостоятельно через промпт-инжиниринг [3]. Вышеперечисленное можно объединить под названием прямых ограничений.

Следующая группа представляет собой косвенные ограничения, которые существенно лимитируют эффективность работы моделей. Они не закладываются директивно разработчиками или пользователем, а возникают в рамках самой языковой генеративной модели в процессе поступления входных данных в модель (взаимодействий с ней). К одному из них можно отнести, например, выходные данные после возникновения ошибок модели в том случае, если промпт-инжинирингом и другими методами эти ошибки не удалось исправить – тогда практический вопрос или их группа так и останутся неразрешенными без помощи человека. В IT-компаниях рассматриваемые модели нейронных сетей часто используются для написания программного кода, но в сложных случаях все еще требуется усилие оператора (то есть, специалиста), так как возникают ошибки. Следовательно, образуется нереализованный потенциал инновационного решения, который пока нельзя исправить на уровне самой модели, что характерно и для других ограничений.

Тональность – второе ограничение данной группы, ведь от нее зависит направленность выдачи. Этот фактор, в свою очередь, связан с параметром температуры t, влияющим на «креативность» той или иной модели.

Третий ограничение второй группы связано с ограниченным вниманием – модель не способна запомнить все контекстное окно сразу: оно разбивается ей на внутренние составляющие, что может привести изменению итоговых результатов.

Четвертое ограничение – невозможность быстрого и точного изменения формы представления данных. Он представлен отдельно, так как в данном случае модель направлена на преобразование уже полученных данных.

Необходимо понимать, что выявленные ограничения являются основными (не были выявлены те, которые гипотетически могут существовать для отдельных областей). Следовательно, дальнейшее изучение ограничений в разрезе отдельных отраслей целесообразно для разработки новых метрик эффективности работы языковых генеративных моделей нейронных сетей для отдельных отраслей, а также для описания пределов эффективности работы данных моделей.

Таким образом, выявлено 6 основных ограничений языковых генеративных моделей нейронных сетей. Первая группа – прямые – устанавливаются разработчиками модели или пользователями через промпт-инжиниринг. Вторая группа – косвенные – образуются в модели в процессе взаимодействия с ней. Отметим, что для отдельных отраслей могут существовать и специфические ограничения, лимитирующие общий инновационный потенциал языковых генеративных моделей нейронных сетей. Следовательно, для полного его раскрытия целесообразно продолжить выявление дополнительных ограничений и описание пределов эффективности.

Библиографические ссылки

1. ChatGPT Statistics (2025) – daily & monthly active users // Demandsage.com : сайт. URL: https://www./chatgpt-statistics/ (date of access: 25.09.2025).

2. *Кузьмин А. Н., Туровец А. М.* Проблема минимизации общих ошибок языковых генеративных моделей нейронных сетей в логистических системах // Бизнес. Инновации. Экономика : сб. науч. ст. / Ин-т бизнеса БГУ. Минск, 2024. Вып. 10. С. 155–160.

3. What is prompt engineering? what are the benefits of using prompt Engineering // Medium.com : сайт. URL: https://clck.ru/3PRDjy (date of access: 25.09.2025).