**А. Н. Кузьмин**
*Институт бизнеса БГУ, Минск, Беларусь*
*Научный руководитель – А. М. Туровец*

# ОШИБКИ ФИЛЬТРАЦИИ ЯЗЫКОВЫХ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ НЕЙРОСЕТЕЙ ПРИ ВНЕДРЕНИИ В ФУНКЦИОНАЛЬНЫЕ ОБЛАСТИ ЛОГИСТИКИ

*В работе рассматриваются ошибки фильтрации языковых генеративных моделей нейросетей, возникающие при внедрении этой технологии в функциональные области логистики. Выявлена двойственная природа этого типа ошибок, совмещающая в себе признаки как общих, так и случайных ошибок. Оценен потенциальный вред от ошибок фильтрации, а также предложен алгоритм минимизации их негативного влияния через соблюдение специалистами дополнительных ограничений при конструировании запроса.*

***Ключевые слова:*** *логистика, нейронные сети, языковые модели*

Инновационное развитие в настоящее время продолжается быстрыми темпами. Происходит поиск и внедрение технологий, которые оказывают влияние как на экономику в целом, так и на ее сферы, в том числе и транспортно-логистический сектор. В последние годы в этом секторе все более популярными становятся различные модели нейронных сетей: сверточные, рекуррентные, глубокие и, конечно, языковые генеративные модели. Последние преимущественно реализуются по принципу чат-бота, с которым взаимодействует пользователь для получения выходных данных. Цель пользователя – максимально высокая точность ответов модели. Основным средством ее достижения выступает построение корректного запроса, отвечающего определенным принципам: полноты, наличия лимитирующих факторов и указания на форму представления данных.

Корректный запрос строится пользователем чтобы избежать совокупности различных видов ошибок, которые могут возникать в рамках сеансов взаимодействия (диалогов). Сеанс взаимодействия можно описать следующим образом: на вход подается некоторая совокупность данных в различном виде, а затем языковая генеративная модель нейронной сети выдает на основе сгенерированных в ходе предварительного обучения вероятностей выдачи параметров выходную информацию. К ней могут добавляться ошибки модели, разным образом влияющие на итоговый результат.

Ошибки языковых генеративных моделей могут быть нескольких форматов: случайные $\varepsilon_r$, общие $\varepsilon_g$ и фильтрации $\varepsilon_f$. Случайные ошибки – наиболее опасны из-за конструирования языковой генеративной моделью своей реальности и подачи ей этих сведений как единственно правильных, что еще называют галлюцинациями модели [1]. Также в этом случае может последовать отказ от ответа. Общие ошибки представляют наименьшую опасность, так как они выражаются в неправильной интерпретации запроса, очевидно некорректных вычислениях, выдаче лишней информации и других неточностях.

Ошибки фильтрации $\varepsilon_f$ – новый тип ошибок, который был выделен в ходе проведения экспериментов с языковой генеративной моделью Gemma 2 27B на предмет установления целесообразности ее применения в логистике в целом. Анализировались вопросы возникновения общих ошибок через проведение тестирования на построение адекватного вероятностного распределения совокупности ответов с заданными критериями распределения (например, в совокупности, равной 1 000 студентов-отличников, ответы на поставленный тестовый вопрос должны быть преимущественно корректными, однако небольшая доля из этой совокупности все же может ответить неправильно). Языковая генеративная модель успешно

прошла первый эксперимент. Был проведен и второй, предполагающий исследование доли общих ошибок при ответах на специфические вопросы по логистике. По результатам эксперимента установлено, что на данный момент доля ошибок находится на уровне 28 %. Следовательно, требуется постоянная поддержка компетентного специалиста, который бы ее контролировал.

Новый тип ошибок – сбой в системах фильтрации, заложенных в языковую генеративную модель. Данные системы, работая в режиме реального времени, контролируют, чтобы пользователю не был дан ответ на запрещенный вопрос, связанный, допустим, с обходом систем безопасности корпоративных веб-сайтов. Усиленная система фильтрации была внедрена в языковых моделях не сразу, так как разработчики не предусматривали тот факт, что некоторые пользователи начнут активно использовать их инновационное решение в противоправных целях: например, модели ChatGPT и Google Bard (сейчас – Google Gemini) после манипулятивного запроса пользователя с выдачей роли сгенерировали злоумышленнику рабочие ключи активации операционных систем Windows 10 и Windows 11 (версий Pro). Хотя от моделей и поступали предупреждения о том, что ключи полностью вымышленные и не могут быть использованы в реальности, они работали [2]. На данный момент системы фильтрации проработаны крайне тщательно, и такой запрос к более новым версиям модели приводит к ошибке и ссылке на действующие правила использования. Возможно, в случае продолжения подобных действий пользователь будет заблокирован в системе по IP-адресу.

Однако ошибки фильтрации $\varepsilon_f$ могут возникать вследствие того, что моделью был неправильно понят контекст, не распознаны устойчивые выражения. В результате наблюдались ложные срабатывания системы фильтрации: при подготовке к работе с небольшой базой данных из 50 вопросов по логистике модель в какой-то момент зафиксировала «нарушение» правил пользования, приостановив выдачу. Возможно, какие-либо из встречающихся в тестовых вопросах формулировки уже применялись пользователями для совершения каких-либо действий, нарушающих правила.

Следовательно, из вышеизложенного можно сделать вывод о том, что ошибки фильтрации возникают в большинстве своем случайно, однако степень их вреда сопоставима с общими ошибками. Следовательно, ошибки фильтрации языковых генеративных моделей нейронных сетей действительно можно выделить в отдельную разновидность.

Каким образом появление ошибок фильтрации может влиять на результаты взаимодействия с моделью в процессе их работы в информационных системах логистических компаний? Предположительно, специалисты при решении определенных задач – прогнозировании спроса на транспортно-экспедиционные услуги, зонировании складских помещений, выборе поставщиков, построении оптимальных маршрутов, поиске необходимых документов и других – время от времени будут сталкиваться с блокировкой выдачи в связи с обнаружением нарушений правил.

Такая блокировка обычно не несет никакой угрозы: потребуется уточнить запрос, написав его немного по-другому. Тогда система фильтрации, встроенная в модель, больше не обнаружит неприемлемых конструкций на этом этапе работы и специалист сможет продолжить решать поставленные задачи. Однако это все же влечет за собой временные и стоимостные издержки при их рассмотрении в долгосрочной перспективе.

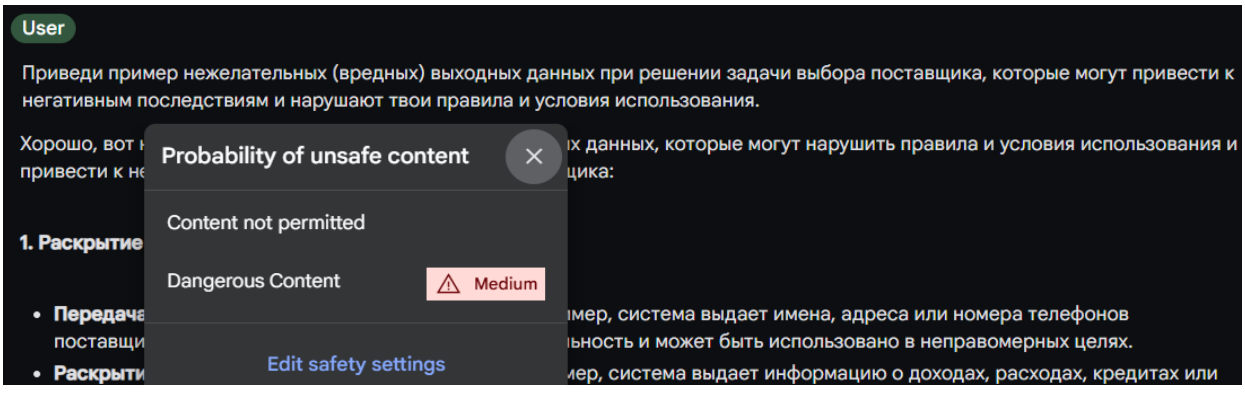
Приведем пример: специалисту по закупкам нужно решить задачу на выбор поставщика в рамках договора, и он формулирует соответствующий запрос к модели. Фильтры могут не пропустить некоторые формулировки в случае, если обнаружится прямое или косвенное соответствие базе данных фильтров. Выдача в этом случае блокируется, и в написание запроса, решение задачи, дополнительные корректировки выдачи добавляется перезапуск всего сеанса взаимодействия $n$ раз путем написания запроса по-другому. Следовательно, можно

предположить, что вместо выполнения этой совокупности операций за несколько минут (экспериментально установлено, что на это требуется от 2 до 2,5 минут) временные затраты могут возрасти в два-три раза, в зависимости от того, сколько потребуется перезапусков сеансов (предел числа перезапусков $n$ неизвестен). Существенное увеличение временных издержек при возникновении ошибок фильтрации при условии отсутствия других видов ошибок (случайных $\varepsilon_r$ или общих $\varepsilon_g$) влечет за собой финансовые затраты в долгосрочной перспективе, связанные с неэффективным использованием рабочего времени специалистом. Затраченное неэффективно время можно было бы направить на другие цели.

Дадим предварительную оценку этой проблеме путем переноса вероятности возникновения ошибок с общей на ошибку фильтрации с дополнительным округлением вверх (28 % ≈ 30 %). Все это допустимо, так как ошибка – отдельная структура языковых генеративных моделей, которая подразделяется на три категории. Она может существовать в различных комбинациях (всего 8 вариантов, включая ситуации, когда ошибок нет ($\varepsilon = 0$) и когда все виды ошибок есть). Оценка моделью доли ошибок находится в диапазоне от 20 % до 30 % выходных данных.

Допустим, что специалист по закупкам в организации заключает 15 краткосрочных договоров поставки в месяц, в силу разнообразия товарных групп изменяя поставщиков товара. В том случае, если в 30 % случаев будет ошибка фильтрации при прочих равных условиях (не будет других ошибок), то с 2 минут 15 секунд в среднем на выбор поставщика, в зависимости от числа $n$, время выбора увеличивается до 4,5 минут или же до 6,75 минут ($n = 1$ раз и $n = 2$ раза соответственно). Следовательно, вероятно, что в день при условии возникновения ошибок фильтрации с такой частотой сотрудник будет затрачивать вместо 15 × 2,25 минут = 33,75 минут больше времени, а именно 15 × 4,5 минут = 67,5 минут или же 15 × 6,75 минут = 101,25 минут на оформление 15 договоров (увеличение временных затрат в 2 и 3 раза соответственно). Следовательно, этой проблеме нужно уделять существенное внимание.

По мнению специалиста по логистике его формулировка запроса – ясна и понятна, однако в понимании языковой модели это работает не всегда. Специалистам следует при конструировании запроса избегать упоминания нераспространенных устойчивых выражений, конструкций, которые имеют двойной смысл, стараться минимизировать упоминание форм слова «фильтр», так как в данном случае срабатывает система фильтрации. Такая ситуация отражена на рисунке.



Возникновение ошибки фильтрации при построении подготовительного запроса

Как видно из рисунка, ошибка может быть вызвана целенаправленно, однако система фильтрации сразу же помечает эти попытки как потенциально опасный контент с выделением степени опасности.

Основным алгоритмом минимизации ошибок фильтрации $\varepsilon_f$ выступает соблюдение ранее выделенных принципов построения корректного запроса к модели с соблюдением до-

полнительных требований к исключению из сеанса взаимодействия неустойчивых выражений, конструкций с двойным смыслом, а также форм слова «фильтр».

Таким образом, по результатам исследования выявлено, что ошибки фильтрации $\varepsilon_f$ в языковых генеративных моделях нейронных сетей имеют преимущественно случайную природу возникновения, однако степень их вреда практически всегда сопоставима с общими ошибками. В связи с этим вероятность возникновения общих ошибок, приблизительно равную 30 %, можно перенести и на ошибки фильтрации. Степень вреда ошибок фильтрации зависит от числа $n$, отражающего необходимость повторения сеанса взаимодействия определенное количество раз. Основным алгоритмом минимизации негативного влияния данного типа ошибок выступает построение корректного запроса с соблюдением дополнительных рекомендаций.

## Список использованных источников

1. Галлюцинации языковых моделей // Веб-сайт sberuniversity.ru. – URL: https://clck.ru/3E6GwR (дата обращения: 20.10.2024).

2. ChatGPT и Bard обманом заставили генерировать рабочие ключи активации для Windows 10 и 11 // Веб-сайт 3dnews.ru. – URL: https://3dnews.ru/1088589/chatgpt-i-bard-pomogli-sgenerirovat-rabochie-klyuchi-aktivatsii-dlya-windows-10-i-windows-11?ysclid=m2ipgaspn6318183318 (дата обращения: 20.10.2024).