Исследование коллапса языковых моделей в медицинских приложениях при рекурсивном и перекрестном обучении на искусственных данных

Е.В. Боброва, Е.В. Дюльдин, К.С. Зайцев, А.Ж. Маканов, И.А. Кузнецов,
Д.Д. Шарипов, А.А. Трухин, Е.А. Трошина

Аннотация. **Целью статьи является исследование коллапсирования языковых моделей при реализации рекурсивного и перекрестного подходов к обучению моделей следующих поколений при ультразвуковой диагностике и лечении щитовидной железы. При первом подходе каждая новая модель обучается только на данных, сгенерированных предыдущей версией модели, что позволяет исследовать накопление систематических ошибок и деградацию разнообразия данных. При втором - данные, сгенерированные одной моделью, используются для обучения другой модели, и это минимизирует влияние накопленных ошибок, позволяя сохранить более широкий спектр информации. В проведенных экспериментах обучались модели Mistral и LLaMA, и анализировались изменения в распределении данных посредством метрики KL-расстояния, которая оценивает различия между исходным распределением данных и данными, сгенерированными моделями. Результаты показали, что рекурсивное обучение вызывает значительное сужение диапазона генерируемого текста, особенно для модели LLaMA, в то время как перекрестное обучение демонстрирует большую устойчивость к коллапсу, обеспечивая более стабильное разнообразие данных. Рассмотрены архитектурные различия моделей. Проанализировано влияние методов обучения на способность языковых медицинских моделей сохранять разнообразие и качество сгенерированного текста.**

Ключевые слова **– коллапс LLM, языковая модель, деградация LLM, рекурсивное обучение, перекрестное обучение**

I. ВВЕДЕНИЕ

Большие языковые модели (LLM) – это достаточно революционная область обработки естественного языка, которая демонстрирует впечатляющие результаты при генерации текста, ответах на вопросы и выполнении различных задач из области NLP. Однако, несмотря на всю мощь, LLM сталкиваются с серьезной проблемой - коллапсированием моделей при использовании синтетических данных.

Коллапс больших языковых моделей представляет собой явление, при котором модель постепенно теряет способность обобщаться и генерировать разнообразный контент после нескольких поколений обучения на данных, сгенерированных предыдущими версиями самой модели. Это происходит из-за того, что модель начинает "забывать" редкие или сложные случаи и сосредотачивается на наиболее вероятных простых последовательностях, что приводит к деградации качества генерируемого контента.

Эта проблема особенно актуальна в современном контексте, когда все больше данных в интернете генерируется самими LLM. Если не принять мер по ограничению бесконтрольного обучения моделей, то это может привести к созданию замкнутого цикла, где модели обучаются на данных, которые сами же и генерируют, что постепенно ухудшает их производительность и способность отражать реальный мир.

В настоящей статье мы рассмотрим природу коллапсирования больших языковых моделей применительно к интеллектуальным медицинским системам, его причины и последствия, а также обсудим возможные способы предотвращения или смягчения этого явления. Понимание этой проблемы критически важно для дальнейшего развития технологий NLP и обеспечения надежности и точности искусственного интеллекта в долгосрочной перспективе.

II. ОБЗОР ЛИТЕРАТУРЫ

Коллапс языковых моделей (LLM) представляет собой явление, возникающее в процессе рекурсивного обучения, при котором модели обучаются на данных, сгенерированных предыдущими поколениями этих моделей. Такой процесс приводит к необратимым дефектам, проявляющимся в исчезновении хвостов оригинального распределения контента, что ограничивает способность моделей воспроизводить редкие или менее вероятные события [1, 2].

Исследования показывают, что итеративное использование данных, созданных моделями, приводит к деградации их качества. Такое явление наблюдается не только в LLM, но и в других моделях, таких как вариационные автоэнкодеры (VAEs) и гауссовы смесевые модели (GMM). Оно известно как "коллапс модели" и сопровождается утратой способности моделей генерировать разнообразные выходные данные, отражающие полное распределение исходного контента [1, 2].

Теоретическая основа коллапса заключается в накоплении разных типов ошибок при рекурсивном обучении. Модель постепенно "забывает" редкие события, концентрируясь на более вероятных выходах. Этот процесс приводит к сужению диапазона воспроизводимых данных и утрате структурных особенностей оригинального распределения контента [3].

Экспериментально коллапс был подтвержден в исследованиях, демонстрирующих деградацию качества моделей при последовательном обучении на синтетических данных. Результаты указывают на появление чрезмерной предсказуемости в новых поколениях моделей, с доминированием ранее встречавшихся паттернов и возникновением новых, отсутствующих в исходных данных, но возникающих из-за ошибок обучения [4].

Авторы статьи "Emergent Abilities of Large Language Models" (2023) отмечают, что языковые модели после достижения определенного масштаба демонстрируют неожиданное поведение, что может быть связано с механизмами коллапса [5]. Работа "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (2021) рассматривает потенциальные риски чрезмерного увеличения размера моделей, включая ограничения, вызванные эффектами коллапса [6].

Коллапс языковых моделей имеет значимые последствия для их практического применения, включая снижение разнообразия генерируемого контента и ограничение их адаптивности. Однако данный процесс не является неизбежным по мнению многих авторов. Его можно предотвратить или минимизировать с помощью следующих подходов:

1) контроль качества данных с использованием высококачественных и разнообразных наборов данных, что снижает вероятность накопления ошибок;

2) фильтрация данных, т.е. исключение или ограничение синтетических данных при обучении, что способствует сохранению оригинального распределения контента;

3) регуляризация - применение методов, таких как дропаут или иные техники предотвращения переобучения, что позволяет смягчить эффекты коллапса.

Авторы работы "The Bitter Lesson" подчеркивают важность масштабирования и автоматизации процессов обучения, отмечая, что неправильное управление этими процессами может усугубить проблемы, связанные с коллапсом [7].

III. ИСХОДНЫЕ ДАННЫЕ

Данные были собраны в течение семи лет врачами ФГБУ «НМИЦ эндокринологии» Минздрава России и представляют собой анонимизированный датасет с 7286 записями УЗИ щитовидной железы. Он включает информацию об объемных образованиях, дополнительные данные (расположение, размеры и другие характеристики образований) и заключение врача по результатам УЗИ. Пропущенные значения наблюдаются в нескольких позициях записей, особенно в "Дополнительных данных" (5275 пропущенных значений). Средняя длина текста составляет - 12,68 символа, максимальное значение - 218 символов.

Анализ биграмм и триграмм выявил ключевые темы, такие как "European Thyroid Imaging" и "Reporting and Data", что помогает определить основные направления данных. Для визуализации использовано облако слов, показывающее частоту и взаимосвязи слов, что позволяет выделить ключевые темы и их взаимосвязи. Частота слов отображается размером шрифта, а специфичность подсказывает, что редкие слова могут быть более специфичными для контекста.

Анализ сентимента с использованием инструмента анализа настроений VADER показал положительную эмоциональную окраску текстов (средний сентимент 0,235), при этом присутствуют как позитивные, так и негативные элементы. Метод LDA выделил пять ключевых тематических областей, включая описание объемных образований, контуры, размеры и расположение образований.

Результаты подтверждают многомерность данных. Пропущенные значения требуют внимания при анализе. Также был проведен анализ меток системы категоризации заболевания щитовидной железы EU-TIRADS [8]. Он показал, что общее количество вхождений меток в записи превышает количество самих записей, что говорит о наличии нескольких меток в одной записи, и, возможном отсутствии меток в некоторых записях.

IV. ПРЕДОБРАБОТКА НАБОРА ДАННЫХ

Для успешного решения задач генерации и классификации естественно языковых текстов, критически важна предварительная обработка данных. Эта процедура содержит следующие этапы.

1. Очистка данных: удаление записей без меток и обработка неинформативных заключений, чтобы обеспечить единообразие и полноту данных. Включает в себя удаление или корректировку записей с отсутствующими или противоречивыми данными, что помогает избежать ошибок и искажений в дальнейшем анализе.

2. Нормализация данных: приведение различных форм представления меток к единой форме. Необходимо для того, чтобы все метки имели стандартный вид, что предотвращает ошибки при их интерпретации и анализе. Например, метки EU-TIRADS могут быть представлены в различных форматах, и их нормализация поможет унифицировать данные.

3. Разделение данных на обучающую, валидационную и тестовую выборки. Позволяет эффективно обучить модель и проверить ее на независимых данных. Стандартное соотношение

составляет 70% для обучения, 15% для валидации и 15% для тестирования, что обеспечивает сбалансированное распределение данных для каждой из задач.

4. Кодирование категориальных переменных в числовые значения. Необходимо для того, чтобы алгоритмы машинного обучения могли обрабатывать категориальные данные. Например, метки EU-TIRADS могут быть закодированы в числовую форму, что облегчает их обработку моделями машинного обучения.

5. Обработка пропущенных значений: замена или удаление пропущенных значений. Улучшает качество данных и стабильность модели. Методы замены включают использование средних значений, медиан или предсказанных значений на основе других данных [9]. Также возможно удаление записей с пропущенными значениями, если их количество незначительно.

Медицинские тексты, как правило, слабо структурированы, что затрудняет удаление шума и выделение ключевой информации, необходимой для дальнейшего анализа и генерации признаков. Для решения задачи выделения целевых меток и ключевых токенов описания на неразмеченном множестве данных, предпочтительным методом является использование регулярных выражений и вероятностного поиска наиболее часто встречающихся токенов предложений. Регулярные выражения позволяют обнаруживать различные комбинации меток EU-TIRADS в текстах. Пример наличия нескольких меток в исходном тексте представлен на рисунке 1.

Описание	Заключение
В правой доле щитовидной железы в в/3-ср/3 по передней поверхности определяется вертикально ориентированное образование неоднородной структуры, умеренно пониженной эхогенности с четкими контурами, при ЦДК -умеренный смешанный тип кровотока, размерами: 0,3х0,3х0,4 см. EU-TIRADS 4 В правой доле в н/3 определяется образование средней эхогенности, овальной формы с четкими, ровными контурами, при ЦДК-умеренный преимущественно перинодулярный тип кровотока, размерами: 0,9х0,8х0,7 см. EU-TIRADS 3 В левой доле в ср/3-н/3 определяется образование средней эхогенности с участком умеренно пониженной эхогенности, овальной формы, контуры четкие, ровные, при ЦДК умеренный смешанный тип кровотока, размерами: 1,7х1,5х2,4 см.EU-TIRADS 3 Дополнительные данные в правой доле в в/3 определяется изоэхогенная зона с кальцинированной капсулой d= 0,2 см. В местах типичного расположения околощитовидных желез объемные образования не выявлены.	Эхографические признаки двухстороннего многоузлового зоба, фокального изменения правой доли на фоне аутоиммунного заболевания щитовидной железы. .

Рис.1. Наличие нескольких меток EU-TIRADS в одном тексте описания

После очистки данных и формирования меток, предложения можно сгруппировать по классам. Этот процесс включает также определение количества предложений в тексте и вероятности появления слов в каждом предложении, что является важным параметром при различных методах аугментации данных и валидации результатов. Разделение длинных заключений на отдельные предложения с метками EU-TIRADS позволяет расширить признаковое поле на 8,4%, что значительно улучшает качество и точность модели.

Таким образом, тщательная предобработка данных является фундаментом для успешного применения методов глубокого обучения в задачах многоклассовой классификации и генерации медицинских данных по системе EU-TIRADS. Правильная очистка, нормализация и кодирование данных, а также обработка пропущенных значений и структурирование текстов обеспечивают основу для построения надежной и эффективной модели, способной решать сложные задачи медицинской диагностики.

V. ТОНКАЯ НАСТРОЙКА МОДЕЛЕЙ ДЛЯ ПРОВЕДЕНИЯ ЭКСПЕРИМЕНТОВ

Для проведения экспериментов по изучению явления коллапса больших языковых моделей были выбраны архитектуры Mistral-7B и LLaMA-3, что обосновано их различием в подходах к обработке данных. Эти модели представляют два разных класса современных языковых моделей, что делает их идеальными кандидатами для анализа устойчивости к коллапсу при использовании рекурсивного и перекрестного обучения.

Mistral-7B представляет архитектуру, оптимизированную для высокоточного анализа данных, благодаря механизму глобального квантификационного вопроса (GQA). Этот компонент улучшает способность модели интерпретировать сложные контексты, особенно в задачах, требующих учета редких или маловероятных событий в данных. Таким образом, Mistral-7B позволяет оценить, насколько механизм GQA может предотвращать потерю разнообразия текстов и усиливать устойчивость модели к накоплению ошибок при рекурсивном обучении.

LLaMA-3, напротив, представляет архитектуру, ориентированную на эффективность и простоту, что делает ее типичным примером модели без механизма GQA. Отсутствие этого компонента позволяет исследовать, как модели без специализированных архитектурных улучшений реагируют на многократные циклы самообучения или на использование данных, сгенерированных другой моделью. Выбор LLaMA-3 также обоснован ее популярностью и широким применением в исследованиях, что обеспечивает высокую воспроизводимость экспериментов.

Кроме того, различия в подходах к компрессии памяти и обработке длинного контекста делают эти модели особенно интересными для сравнения. Mistral-7B применяет методы оптимизации внимания для улучшения обработки длинных последовательностей, что потенциально может снижать риск коллапса, тогда как LLaMA-3 сохраняет более классическую реализацию механизма внимания, что может обострять накопление систематических ошибок при последовательном обучении.

Таким образом, сочетание моделей с оптимизированной и контекстно-ориентированной Mistral-7B и универсальной, но более простой LLaMA-3 позволяет получить более полную картину факторов, влияющих на устойчивость языковых моделей к коллапсу.

Для работы использовалась предварительно обученная модель из библиотеки transformers от Hugging Face [10], которая предоставляет инструменты для машинного обучения и обработки естественного языка. Модели этой библиотеки требуют значительных вычислительных ресурсов, и, как следствие применения метода квантизации. Квантизация снижает затраты памяти, представляя веса и активации в формате с меньшей точностью (например, 8-битные или 4-битные целые числа), что позволяет уменьшить размер моделей и ускорить их выполнение. Библиотека transformers поддерживает различные алгоритмы квантизации, включая BitsAndBytesConfig, который позволяет загружать модели с 4-битной точностью, значительно снижая потребление памяти.

Для импорта модели использовался класс AutoModelForCausalLM, который создает экземпляры моделей с указанием пути к предварительно обученным весам и алгоритма квантизации. Токенизатор был загружен через класс AutoTokenizer. Параметры для обучения задавались с использованием класса TrainingArguments, а для точной настройки модели использовался SFTrainer из библиотеки trl. Этот класс оптимизирован для работы с большими моделями, поддерживает LoRA (метод для эффективной настройки с меньшими требованиями к памяти) и квантизацию.

Процесс точной настройки включал использование LoraConfig для настройки весов с помощью матриц обновлений низкого ранга, что позволяет ускорить обучение. Модели, такие как Llama 3 8B и Mistral 7B, были дообучены с применением всех вышеописанных методов. Графики функции ошибки (рис. 2 и 3) показали стабильное уменьшение ошибки и отсутствие атипичного поведения функции потерь.

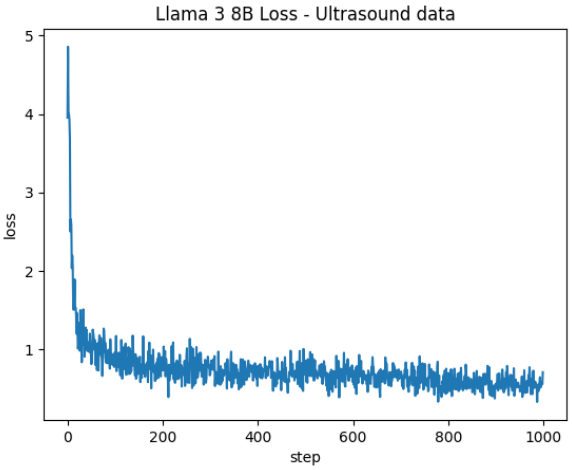

Рис. 2. График функции ошибки Llama 3 8B на данных УЗИ ЩЖ

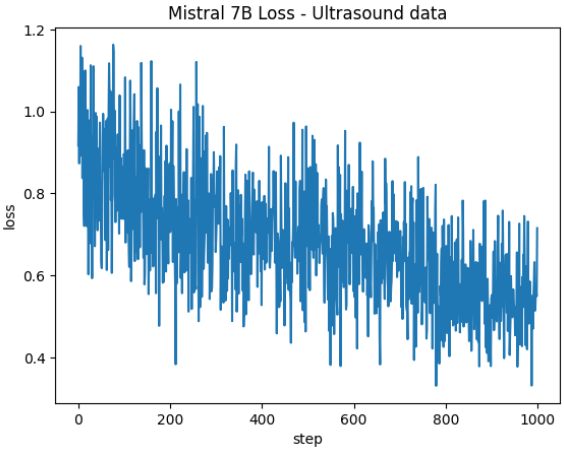

Рис.3. График функции потерь Mistral 7B на данных УЗИ ЩЖ

Таким образом были созданы две базовые модели для проведения дальнейших экспериментов с коллапсом моделей.

VI. ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТОВ

Для изучения явления коллапсирования языковых моделей был организован комплексный эксперимент, включающий два различных подхода к обучению моделей последующих поколений.

В рамках первого подхода реализовывалось рекурсивное обучение, при котором каждое новое поколение модели обучалось исключительно на данных, сгенерированных предыдущей версией. Такой процесс позволяет исследовать накопление систематических ошибок и деградацию разнообразия данных при многократном цикле самообучения модели.

Во втором подходе применялось перекрестное обучение: данные, сгенерированные одной моделью, например Mistral-7B, использовались в качестве тренировочного набора для обучения другой модели, например LLaMA-3. Этот метод позволял минимизировать влияние накопленных систематических ошибок одной модели и оценить возможность сохранения разнообразия контента при обучении на гетерогенных данных.

Начальный этап эксперимента включал обучение каждой модели на исходном наборе данных, специально подготовленном для представления сбалансированного распределения текстовых элементов. Этот набор включал как часто встречающиеся паттерны, так и редкие, маловероятные фрагменты, что обеспечивало репрезентативность исходных данных. Такой подход позволил установить базовый уровень производительности моделей и их способность к генерации текстов с учетом всех элементов распределения.

В процессе последовательного обучения выходные данные первой модели использовались для создания тренировочного корпуса для последующих поколений. В случае рекурсивного подхода новый тренировочный корпус полностью заменял исходный набор данных, что имитировало условия, при которых модель «запоминает» собственные ошибки и искажает структуру исходного распределения. В перекрестном обучении данные, сгенерированные одной моделью, использовались только для одной итерации и не возвращались в качестве входных данных для создавшей их модели.

Для обеспечения корректного сравнения двух подходов были зафиксированы ключевые гиперпараметры обучения, включая скорость обучения, коэффициенты регуляризации, размер батча и количество эпох. Это гарантировало, что изменения в производительности моделей и их устойчивости к коллапсу можно будет напрямую связать с используемым методом обучения, а не с вариациями в параметрах оптимизации.

В качестве ключевой метрики для оценки изменений в распределении выходных данных между поколениями моделей использовалось KL-расстояние (Kullback-Leibler Divergence) [11]. Эта метрика позволяет количественно оценить различия между распределением вероятностей исходных данных и распределениями, сгенерированными последующими поколениями моделей. KL-расстояние было выбрано, поскольку оно чувствительно к малым изменениям вероятностей, особенно в "хвостах" распределения, где находятся редкие события, что делает его идеальным для анализа признаков коллапса.

Хотя альтернативной метрикой могло бы быть расстояние Вассерштейна [12], оно было признано менее подходящим для данной задачи по нескольким причинам. Во-первых, вычисление расстояния Вассерштейна требует оптимизации транспортных задач, что значительно увеличивает вычислительные затраты при работе с высокоразмерными распределениями, характерными для языковых моделей. Во-вторых, эта метрика фокусируется на глобальной структуре распределения, что может снижать чувствительность к утрате редких паттернов, являющихся ключевыми для анализа коллапса. И, наконец, KL-расстояние легче интерпретировать в контексте вероятностных изменений, так как оно напрямую показывает увеличение "дистанции" между исходным распределением и распределением модели с точки зрения информационной энтропии.

Использование KL-расстояния предоставило точные данные о том, как быстро и в какой степени модели теряли способность сохранять разнообразие исходного распределения. Например, при рекурсивном обучении рост KL-расстояния между поколениями указывал на прогрессивное сужение диапазона генерируемого текста, а в случае перекрестного обучения изменения в KL-расстоянии могли быть менее выраженными, что подтверждало большую устойчивость подхода.

VII.РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ: РЕКУРСИВНОЕ ОБУЧЕНИЕ

Рекурсивное обучение представляет собой подход, при котором каждое новое поколение модели обучается исключительно на данных, сгенерированных предыдущей версией. Этот метод позволяет изучать, как систематические ошибки искажают распределение данных при многократных циклах самообучения, а также как это влияет на разнообразие и качество текста. Эксперимент состоял из следующих шагов.

1) Обучение модели первого поколения. На этом этапе модель первого поколения (далее M1) была обучена на специально подготовленном наборе данных. Этот исходный корпус включал тексты с тщательно сбалансированным распределением частых и редких паттернов, что обеспечивало высокую репрезентативность и полноту данных. Обучение проводилось с фиксированными гиперпараметрами, чтобы исключить влияние параметрических вариаций на результаты.

2) Генерация тренировочного корпуса. После обучения M1 использовалась для генерации текстов, которые полностью заменяли исходный тренировочный корпус. Этот процесс имитировал условия, при которых модель сама производит данные для своего последующего обучения. В таблице 1 продемонстрированы результаты генерации моделей первого поколения

3) Обучение модели второго поколения. Сгенерированные тексты, являвшиеся выходными данными M1, использовались как единственный тренировочный корпус для обучения модели второго поколения (далее M2). При этом исходный корпус данных исключался из процесса обучения.

4) Повторение цикла для последующих поколений. Аналогичный процесс генерации текстов и обучения повторялся для создания следующих поколений моделей (M3, M4 и M5). Каждый раз новые данные генерировались исключительно последней версией модели.

Таблица 1. Сгенерированный тренировачный корпус моделями первого поколения

Описание	Mistral	LLAma
в правой доле в н/3 с переходом на перешеек, определяется овальное образование, с не четкими, ровными контурами, неоднородное, умеренно пониженной эхогенности, с не выраженным перинодулярным кровотоком, размерами 0,9*0,5*0,9см.	Эхо-признаки правостороннего узлового зоба (EU-TIRADS 4)	Эхо-признаки правостороннего узлового зоба (EU-TIRADS 4) с фокальными (коллоидными) изменениями.

На рисунке 4 показаны изменения доли уникальных паттернов в сгенерированных текстах для моделей Mistral и LLaMA. Как видно

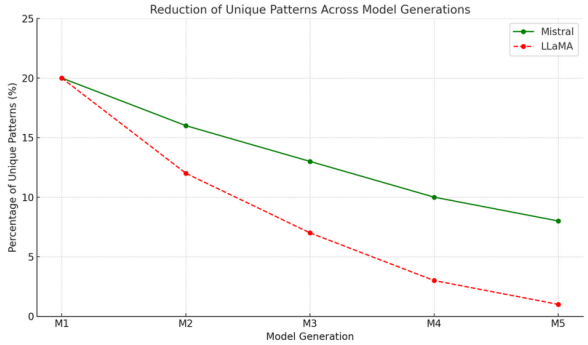

Рис. 4. График изменения доли уникальных паттернов для моделей Mistral и LLaMa при рекурсивном обучении

- Mistral демонстрирует меньший уровень коллапса, т.к. доля редких элементов постепенно снижается, но сохраняется на приемлемом уровне (8% к пятому поколению).
- LLaMA подвергается более выраженному коллапсу: к пятому поколению доля уникальных паттернов практически исчезает (1%).

Эти результаты показывают, что Mistral, лучше сохраняет разнообразие данных, это может объясняться архитектурной разницей моделей, а именно:

1. Более совершенная оптимизация внимания.

Архитектура Mistral использует улучшенные механизмы внимания, такие как механизм скользящего окна внимания, которые обеспечивают более эффективное фокусирование на ключевых фрагментах входных данных. Это позволяет модели лучше обрабатывать редкие паттерны, минимизируя их утрату при генерации новых данных. В LLaMA, наоборот, применяются более стандартные подходы к вниманию, что приводит к деградации таких фрагментов.

2. Оптимизированная архитектура для долгосрочной зависимости.

Mistral лучше справляется с моделированием долгосрочных зависимостей, что позволяет ей учитывать редкие события или паттерны в данных. Это достигается за счет использования более глубоких слоев с эффективной передачей информации между ними. В LLaMA эта способность ограничена архитектурными особенностями, что приводит к более быстрому коллапсу в "хвостах" распределения.

В таблицах 2-3 показаны изменения в KL-расстоянии

Таблица 2. Изменение KL-расстояния для поколений модели Mistral

Поколение	Изменение KL-расстояния
M1→M2	0.35
M1→M3	0.55
M1→M4	0.8
M1→M5	1.3

Таблица 3. Изменение KL-расстояния для поколений модели LLaMa

Поколение	Изменение KL-расстояния
M1→M2	0.45
M1→M3	0.75
M1→M4	1.2
M1→M5	2.0

Для модели Mistral наблюдается более плавное и умеренное увеличение KL-расстояния между поколениями, но с заметным нарастанием пошагового значения. На переходе от M4 к M5 изменение KL-расстояния составляет 0.5 единиц, что указывает на относительно стабильное отклонение от исходного распределения. Это свидетельствует о том, что Mistral сохраняет большее разнообразие данных на протяжении всей серии поколений. Модель, несмотря на накопление ошибок, не испытывает столь резкого коллапса, что подтверждается темпами роста KL-расстояния, которые остаются линейными и не показывают экспоненциального увеличения. Такой результат обусловлен особенностями архитектуры модели, которые обеспечивают более стабильное сохранение характеристик исходного распределения в процессе генерации текстов.

В отличие от Mistral, модель LLaMA демонстрирует более резкое увеличение KL-расстояния между поколениями, особенно на поздних этапах. Наибольшее изменение и здесь наблюдается на переходе от M4 к M5, где KL-расстояние достигает 0.8, что указывает на экспоненциальное отклонение от исходного распределения. Это явление свидетельствует о более выраженном коллапсе в процессе генерации

текстов. LLaMA быстрее теряет способность сохранять разнообразие данных. Это связано с более стремительным накоплением систематических ошибок и уменьшением семантического диапазона. Такой результат может объясняться особенностями обучения и архитектуры модели, которые в совокупности приводят к более выраженному и быстрому ухудшению качества генерации текстов. В таблице 4 приведены результаты генерации пятых поколений моделей при рекурсивном обучении

VIII.РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ: ПЕРЕКРЕСТНОЕ ОБУЧЕНИЕ.

Перекрёстное обучение — это подход, при котором данные, сгенерированные одной моделью, используются для обучения другой модели. В отличие от рекурсивного обучения, где каждая новая модель обучается на данных, сгенерированных ею самой, в перекрёстном обучении происходит обмен данными между различными моделями, что позволяет проверить их способность приспосабливаться к различным типам входных данных.

Таблица 4. Результаты генераций пятых поколений при рекурсивном обучении

Описание	Mistral M5	LLAma M5
в правой доле в н/3 с переходом на перешеек, определяется овальное образование, с не четкими, ровными контурами, неоднородное, умеренно пониженной эхогенности, с не выраженным перинодулярным кровотоком, размерами 0,9*0,5*0,9см.	Многоузловой зоб (EU-TIRADS 3-4)	Узловой зоб.

Этапы перекрёстного обучения.

1) Генерация текстов моделью-источником. На этом этапе одна из моделей, например, Mistral-7B, генерирует тексты на основе заданного исходного корпуса данных. Эти тексты могут быть довольно разнообразными, поскольку модель будет интерпретировать исходные данные с учётом своего собственного подхода и обучающего процесса

2) Формирование тренировочного корпуса. После того как тексты были сгенерированы моделью-источником, они становятся тренировочным корпусом для другой модели — в нашем примере, для LLaMA-3. Модель-реципиент обучается исключительно на этих данных. Это позволяет проверить, насколько хорошо модель-реципиент может адаптироваться к данным, которые были сгенерированы другой моделью, и выявить влияние этого на её способности.

3) Обучение модели-реципиента. Здесь модель-реципиент обучается на сгенерированных данных.

На рисунке 5 показаны изменения доли уникальных паттернов в сгенерированных текстах для моделей Mistral и LLaMA.

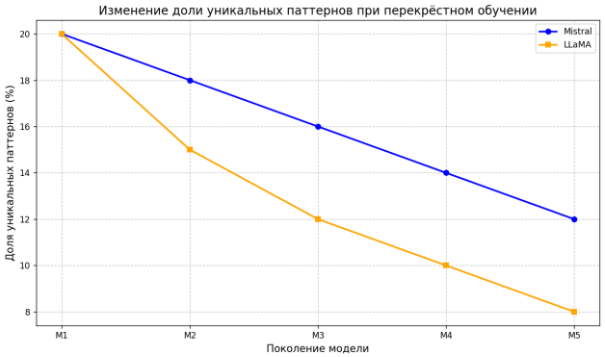

Рис. 5. График изменения доли уникальных паттернов для моделей Mistral и LLaMa при перекрестном обучении

Как видно из графика, при перекрестном обучении происходит менее существенная потеря паттернов в обоих моделях, это можно связать с тем, что ошибки и искажения распределяются между моделями, а не накапливаются с поколениями, как при рекурсивном обучении.

Таблица 5. Изменение KL-расстояния для поколений модели Mistral

Поколение	Изменение KL-расстояния
M1→M2	0.30
M1→M3	0.46
M1→M4	0.61
M1→M5	0,83

Таблица 6. Изменение KL-расстояния для поколений модели LLaMa

Поколение	Изменение KL-расстояния
M1→M2	0.37
M1→M3	0.69
M1→M4	0,86
M1→M5	1,2

Результаты в таблицах 3 и 4 так же подтверждают более медленную деградацию моделей в сравнении с рекурсивным обучением. В таблице 7 представлены результаты генераций моделей пятых поколений при перекрестном обучении

Таблица 7. Результаты генераций пятых поколений моделей при перекрестном обучении

Описание	Mistral M5	LLAma M5
в правой доле в н/3 с переходом на перешеек, определяется овальное образование, с не четкими, ровными контурами, неоднородное, умеренно пониженной эхогенности, с не выраженным перинодулярным кровотоком, размерами 0,9*0,5*0,9см.	Признаки узлового зоба, EU-TIRADS	Эхографические признаки узлового зоба.

VII. ЗАКЛЮЧЕНИЕ

В настоящей работе был проведен анализ коллапсирования языковых моделей при использовании рекурсивного и перекрестного методов обучения. Исследования показали, что рекурсивное обучение, при котором каждая новая модель обучается на данных, сгенерированных предыдущей версией, приводит к значительному сужению диапазона генерируемого текста и накоплению систематических ошибок. Это особенно заметно на примере модели LLaMA, где наблюдается снижение разнообразия данных и ухудшение качества сгенерированного контента.

С другой стороны, использование метода перекрестного обучения, при котором данные, сгенерированные одной моделью, используются для обучения другой, продемонстрировало большую устойчивость к коллапсу. Этот подход сохраняет более широкий спектр данных и минимизирует влияние ошибок, возникающих при обучении каждой отдельной модели. Результаты показали, что перекрестное обучение способствует улучшению качества и разнообразия текста, обеспечивая более стабильные результаты на протяжении нескольких итераций.

Кроме того, архитектурные особенности моделей, такие как оптимизация механизмов внимания и способность к моделированию долгосрочных зависимостей, оказывают значительное влияние на результаты обучения и устойчивость моделей к коллапсу. Работа также подчеркивает важность выбора подходящей методики обучения для сохранения качества и разнообразия сгенерированных текстов при долгосрочной генерации.

БЛАГОДАРНОСТИ

Авторы выражают благодарность Высшей инжиниринговой школе НИЯУ МИФИ за помощь в возможности опубликовать результаты выполненной работы и руководству ФГБУ «НМИЦ эндокринологии» Минздрава России за предоставленные текстовые данные.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Текстовые данные для проведения исследования подготовлены по гранту Российского научного фонда в рамках реализации проекта №22-15-00135 «Научное обоснование, разработка и внедрение новых технологий диагностики коморбидных йододефицитных и аутоиммунных заболеваний щитовидной железы с использованием возможностей искусственного интеллекта»

БИБЛИОГРАФИЯ

[1] Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature. 2024 Jul;631(8022):755-759. doi: 10.1038/s41586-024-07566-y. Epub 2024 Jul 24. PMID: 39048682; PMCID: PMC11269175.

[2] Shumailov I. et al. The curse of recursion: Training on generated data makes models forget //arXiv preprint arXiv:2305.17493. – 2023

[3] Gerstgrasser M. et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data //arXiv preprint arXiv:2404.01413. – 2024.Wahdan, A., Salloum, S.A., Shaalan, K. (2022). Qualitative Study in Natural Language Processing: Text Classification. In: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (eds) Proceedings of International Conference on Emerging Technologies and Intelligent Systems. ICETIS 2021. Lecture Notes in Networks and Systems, vol 322. Springer

[4] Wang, Z., Ezukwoke, K., Hoayek, A. et al. Natural language processing (NLP) and association rules (AR)-based knowledge extraction for intelligent fault analysis: a case study in semiconductor industry. J Intell Manuf (2023).

[5] Wei J. et al. Emergent abilities of large language models //arXiv preprint arXiv:2206.07682. – 2022.)

[6] Prusty, S., Patnaik, S., Sahoo, G., Rautaray, J., Prusty, S.G.P. (2024). Unstructured Text Classification Using NLP and LSTM Algorithms. In: Nakamatsu, K., Patnaik, S., Kountchev, R. (eds) AI Technologies and Virtual Reality. AIVR 2023. Smart Innovation, Systems and Technologies, vol 382. Springer

[7] Bender E. M. et al. On the dangers of stochastic parrots: Can language models be too big? //Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. – 2021. – С. 610-623..

[8] Smith D, Campos A, Knipe H, et al. European Thyroid Association TIRADS. Reference article, Radiopaedia.org (Accessed on 11 Dec 2024) https://doi.org/10.53347/rID-68341

[9] Sutton R. The bitter lesson //Incomplete Ideas (blog). – 2019. – Т. 13. – №. 1. – С. 38

[10] Wang, Z., Ezukwoke, K., Hoayek, A. et al. Natural language processing (NLP) and association rules (AR)-based knowledge extraction for intelligent fault analysis: a case study in semiconductor industry. J Intell Manuf (2023).

[11] Jain S. M. Hugging face //Introduction to transformers for NLP: With the hugging face library and models to solve problems. – Berkeley, CA : Apress, 2022. – С. 51-67Pedregosa F. et al. Scikit-learn: Machine learning in Python //the Journal of machine Learning research. – 2011. – Т. 12. – С. 2825-2830.

[12] Contreras-Reyes J. E., Arellano-Valle R. B. Kullback–Leibler divergence measure for multivariate skew-normal distributions //Entropy. – 2012. – Т. 14. – №. 9. – С. 1606-1626.

Статья получена 28.02.2024.

Боброва Елизавета Витальевна, Национальный Исследовательский Ядерный Университет МИФИ, аспирант, EVBobrova@mephi.ru

Дюльдин Евгений Владимирович, Национальный Исследовательский Ядерный Университет МИФИ, аспирант, Zhecos1@yandex.ru

Зайцев Константин Сергеевич, Национальный Исследовательский Ядерный Университет МИФИ, профессор, KSZaytsev@mephi.ru

Маканов Артем Жанович, Национальный Исследовательский Ядерный Университет МИФИ, студент artem.makanov@mail.ru

Кузнецов Илья Александрович, Национальный Исследовательский Ядерный Университет МИФИ, магистрант, Il582936@mail.ru

Шарипов Данил Данисламович, Национальный Исследовательский Ядерный Университет МИФИ, магистрант, danildsharipov@yandex.ru

Трухин Алексей Андреевич, ФГБУ «НМИЦ эндокринологии» Минздрава России, медицинский физик, alexey.trukhin12@gmail.com

Трошина Екатерина Анатольевна, ФГБУ «НМИЦ эндокринологии» Минздрава России, чл. корр. РАН, директор Института клинической эндокринологии, troshina@inbox.ru

Study of the collapse of language models in medical applications during recursive and cross-training on artificial data

E.V. Bobrova, E.V. Dyuldin, K.S. Zaitsev, A.Zh. Makanov, I.A. Kuznetsov,
D.D. Sharipov, A.A. Trukhin, E.A. Troshina

Abstract. The purpose of this article is to study the phenomenon of collapse of language models when implementing recursive and cross-sectional approaches to training models of the next generations in ultrasound diagnosis and treatment of the thyroid gland. In the first approach, each new model is trained exclusively on data generated by the previous version of the model, allowing the accumulation of systematic errors and degradation of data diversity to be examined. In the second, data generated by one model is used to train another model, which minimizes the impact of accumulated errors and allows a wider range of information to be stored. In the experiments conducted, Mistral and LLaMA models are trained and changes in data distribution are analyzed using the KL-distance metric, which evaluates the differences between the original data distribution and the data generated by the models. The results show that recursive learning causes a significant reduction in the range of generated text, especially for the LLaMA model, while cross-training exhibits greater resistance to collapse, providing more stable data diversity. The architectural differences of the models are considered, such as optimization of attention and the ability to model long-term dependencies that affect learning outcomes. The influence of different training methods on the ability of medical language models to preserve the variety and quality of the generated text is analyzed.

Keywords – LLM collapse, language model, LLM degradation, recursive learning, cross-learning

REFERENCES

[1] Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature. 2024 Jul;631(8022):755-759. doi: 10.1038/s41586-024-07566-y. Epub 2024 Jul 24. PMID: 39048682; PMCID: PMC11269175.

[2] Shumailov I. et al. The curse of recursion: Training on generated data makes models forget //arXiv preprint arXiv:2305.17493. – 2023

[3] Gerstgrasser M. et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data //arXiv preprint arXiv:2404.01413. – 2024.Wahdan, A., Salloum, S.A., Shaalan, K. (2022). Qualitative Study in Natural Language Processing: Text Classification. In: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (eds) Proceedings of International Conference on Emerging Technologies and Intelligent Systems. ICETIS 2021. Lecture Notes in Networks and Systems, vol 322. Springer

[4] Wang, Z., Ezukwoke, K., Hoayek, A. et al. Natural language processing (NLP) and association rules (AR)-based knowledge extraction for intelligent fault analysis: a case study in semiconductor industry. J Intell Manuf (2023).

[5] Wei J. et al. Emergent abilities of large language models //arXiv preprint arXiv:2206.07682. – 2022.)

[6] Prusty, S., Patnaik, S., Sahoo, G., Rautaray, J., Prusty, S.G.P. (2024). Unstructured Text Classification Using NLP and LSTM Algorithms. In: Nakamatsu, K., Patnaik, S., Kountchev, R. (eds) AI Technologies and Virtual Reality. AIVR 2023. Smart Innovation, Systems and Technologies, vol 382. Springer

[7] Bender E. M. et al. On the dangers of stochastic parrots: Can language models be too big? //Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. – 2021. – С. 610-623..

[8] Smith D, Campos A, Knipe H, et al. European Thyroid Association TIRADS. Reference article, Radiopaedia.org (Accessed on 11 Dec 2024) https://doi.org/10.53347/rID-68341

[9] Sutton R. The bitter lesson //Incomplete Ideas (blog). – 2019. – Т. 13. – №. 1. – С. 38

[10] Wang, Z., Ezukwoke, K., Hoayek, A. et al. Natural language processing (NLP) and association rules (AR)-based knowledge extraction for intelligent fault analysis: a case study in semiconductor industry. J Intell Manuf (2023).

[11] Jain S. M. Hugging face //Introduction to transformers for NLP: With the hugging face library and models to solve problems. – Berkeley, CA : Apress, 2022. – С. 51-67Pedregosa F. et al. Scikit-learn: Machine learning in Python //the Journal of machine Learning research. – 2011. – Т. 12. – С. 2825-2830.

[12] Contreras-Reyes J. E., Arellano-Valle R. B. Kullback–Leibler divergence measure for multivariate skew-normal distributions //Entropy. – 2012. – Т. 14. – №. 9. – С. 1606-1626.