Исследование раннего и позднего коллапса языковых моделей в медицинских приложениях

Е.В. Боброва, К.С. Зайцев, Д.К. Свириденко, Д.В. Холод, Е.В. Дюльдин

Аннотация. **Целью работы является комплексный анализ механизмов раннего коллапса языковых моделей, работающих с медицинскими текстами, при рекурсивном их обучении на примере архитектур Mistral-7B и LLaMA-3. Проведено экспериментальное исследование динамики изменения перплексии, метрик BLEU и ROUGE, а также распределения вероятностей токенов в процессе многопоколенческого синтетического обучения. Выявлены два типа коллапса моделей: ранний (характеризующийся быстрой деградацией вероятностных распределений) и поздний (с постепенным снижением разнообразия генерации). Установлено, что модель Mistral демонстрирует большую устойчивость к коллапсу данных по сравнению с LLaMA, что обусловлено особенностями ее архитектуры с механизмом скользящего внимания (sliding window attention). Работа предлагает новый методологический подход к количественной оценке деградации языковых моделей и формулирует практические рекомендации по предотвращению потери модельного разнообразия при рекурсивном обучении. Исследование проводилось на текстовых цитологических данных, используемых при диагностике заболеваний щитовидной железы.**

Ключевые слова **– механизм коллапса, перплексия, деградация LLM, рекурсивное обучение, распределение вероятностей**

I. ВВЕДЕНИЕ

Перплексия и распределение вероятностей используемых для обучения данных служат критическими индикаторами производительности и потенциального коллапса в больших языковых моделях (LLM), которые представляют собой современные вычислительные системы, созданные для решения задач обработки естественного языка.

Эти модели (такие, например, как, GPT-3 или BERT), используют очень большие наборы данных для обучения и сложные алгоритмы для генерации текстов, подобных тем, которые создает человек, что делает их интересными для применения в различных областях деятельности, таких как медицина, межъязыковый перевод, создание тематического контента, разговорных агентов и др.

Однако исследование работы различных языковых архитектур выявило их хрупкость и неустойчивость при их многократном применении, особенно при необходимости поддерживать согласованные и контекстуально обоснованные результаты по мере увеличения сложности и размера обрабатываемых данных [1, 2].

Перплексия (perplexity) является метрикой, используемой для внутренней оценки LLM без получения итоговых показателей потерь/точности модели [3]. Количественно перплексия характеризует неуверенность модели в своих предсказаниях. Для оценивания перплексии часто используется обратная вероятность тестового набора модели, поэтому, чем ниже уровень перплексии — тем лучше модель и выше ее уверенность в сгенерированном тексте. И, наоборот, рост показателей перплексии во время обучения может сигнализировать о надвигающемся коллапсе, отражая неспособность модели генерировать связный текст, и деградацию ее понимания языковых нюансов [4, 5]. Кроме того, анализ распределений вероятностей данных, используемых LLM в процессе генерации, раскрывает критически важные сведения о правильности их функционирования. Корректно работающая модель должна поддерживать весь набор предсказаний, но во при коллапсировании распределения могут чрезмерно сходиться, что приводит к снижению вариативности и креативности в генерируемых результатах [6].

Анализ роста перплексии означает для исследователей моделей, что начинается ранняя деградации модели. Поэтому необходимо вмешиваться, совершенствуя методы обучения и параметры моделей для поддержания устойчивой производительности.[7] Этот проактивный подход становится все более значимым с ростом сложности LLM, помогая исследовать их базовые механизмы, влияющие на точность и производительность, и зависящие от проблем коллапсирования моделей [8].

II. МЕТОДОЛГИЯ ИССЛЕДОВАНИЯ

Для проведения экспериментов, направленных на изучение явления коллапса в больших языковых моделях, нами были выбраны архитектуры Mistral-7B и LLaMA-3. Выбор обусловлен их различиями в архитектурных решениях, методах обработки данных и стратегиях предобучения, что позволяет детально проанализировать их устойчивость к деградации качества генерации при рекурсивном и перекрёстном обучении.

Mistral-7B представляет собой каскадную модель с улучшенной токенизацией и оптимизированной архитектурой Transformer. Одним из ключевых отличий является использование алгоритма sliding

window attention (SWA), позволяющего модели обрабатывать длинные контексты без значительного роста вычислительной нагрузки. Это делает Mistral-7B более устойчивой к информационной деградации при увеличении длины входного текста, что критично для задач с глубоким рекурсивным обучением.

Кроме того, Mistral-7B известна своей агрессивной оптимизацией весов и более высокой плотностью обучения по сравнению с моделями типа LLaMA, что приводит к более плавному распределению вероятностей при генерации текста.

В нашем исследовании использовались две модели: unsloth/mistral-7b-v0.3-bnb-4bit и unsloth/llama-3-8b-bnb-4bit на этапах подготовки исходных данных, дообучения с генерацией и оценивания коллапсирования.

A. Дизайн эксперимента и подготовка исходных данных

В качестве исходных данных в этом исследовании используются материалы, основанные на системе классификации Bethesda (The Bethesda System for Reporting Thyroid Cytopathology, TBSRTC) [9, 10], которая применяется для описания результатов тонкоигольной аспирационной биопсии образований щитовидной железы. Результаты исследований представляют собой одну из шести диагностических категорий, каждая из которых сопровождается определённым риском злокачественности, варьирующим от 4 до 97%, что влияет на выбор дальнейшей тактики ведения пациента.

Корпус данных включает более 27 тысяч реальных записей (пар текстов «описание» и «заключение») врачей-цитологов, полученных в период с 2013 по 2023 год в лаборатории цитологии и цитогенетики отдела патоморфологии НМИЦ эндокринологии. Важно отметить, что в этом корпусе частично отсутствуют исходные данные «описание», так как они размещены в другом поле вместе с собственно «заключением». Медицинские тексты обладают низкой структурированностью, что затрудняет процесс очистки данных от лишних токенов и выделения ключевой информации, необходимой для последующей генерации признаков и разработки моделей анализа.

Для анализа развития коллапса в моделях была применена методика рекурсивного обучения, при которой каждая новая версия модели обучается исключительно на данных, сгенерированных предыдущим поколением. Это позволяет отследить постепенную деградацию качества и разнообразия генерируемого текста на протяжении нескольких поколений моделей. В таблице 1 показано отличие оригинального текста и сгенерированного моделью, подвергнутой коллапсу.

Таблица 1. Сравнение заключения, сгеренерированного сколлапсированной моделью, и исходного текста

Исходное заключение	Сгенерированное
п18 В мазке неравномерной толщины с большой примесью крови обнаружены скопления укрупненныхполиморфных эпителиальных клетокформирующих преимущественно фолликулярные структурыболее характерные для фолликулярного образования щитовидной железы.	В соответствии с цитологическими критериями фолликулярного образования щитовидной железы: фолликулярная неоплазия или подозрение фолликулярную неоплазию щитовидной железы.

Для проведения сравнительного анализа моделей разделение данных проводилось двумя способами:

1. Разделение исходного набора на две части по 1000 строк. Первый поднабор - для обучения модели нулевого поколения, а второй, составляющий 5.6% от общего объёма данных для каждого из этих наборов, использовался для генерации синтетических данных.
2. Разделение исходного набора на две части по 2000 строк, где каждый поднабор составляет 11,1% от общего объёма данных. Как и в первом случае, базовая модель обучалась на первом поднаборе, а второй поднабор использовался для генерации новых данных.

Введем следующие обозначения:

l0-l7 — модели LLaMa, обученные на датасетах объемом 1000 и 2000 строк,

m0-m7 — модели Mistral, обученные на датасетах объемом 1000 и 2000 строк.

B. Параметры дообучения и генерации

Дообучение моделей проводилось с использованием следующих оптимизационных параметров:

- Оптимизатор: AdamW,
- Learning rate: 2e-4 с линейным снижением до 0,
- Batch size: 2,
- Количество эпох: 4,
- Максимальная длина входных последовательностей: 2048 токенов,
- Weight decay: 0.01,
- Gradient accumulation steps: 4.

C. Методы оценки механизмов коллапса

Для оценки качества обучения и анализа генерации текстов использовалась метрика перплексии (perplexity), которая отражает степень неопределённости (неуверенности) модели при предсказании следующего токена в последовательности.

Формально, перплексия для языковой модели P на заданном тексте длины N, состоящем из

последовательности токенов $w_1, w_2,...,w_N$, определяется, как [3]:

$$PPL(W) = exp\left(-\frac{1}{N}\sum_{n=0}^{N-1}\log P\left(\left(w_i|w_1,\dots,w_{i-1}\right)\right)\right) \quad (1)$$

где $P\left(\left(w_i|w_1,\dots,w_{i-1}\right)\right)$ — вероятность генерации токена w_i на основе предыдущих токенов.

Низкие значения перплексии указывают на высокую уверенность модели в своих предсказаниях, что может свидетельствовать о более точном приближении к истинному распределению данных и лучшем качестве генерации.

Однако в контексте исследования коллапса моделей чрезмерно низкая перплексия может также указывать на сужение распределения вероятностей и снижение разнообразия генерируемых текстов.

Дополнительно, для всесторонней оценки качества генерации текста были использованы такие метрики, как:

- BLEU (Bilingual Evaluation Understudy) — метрика, оценивающая качество машинного перевода через сравнение n-грамм в сгенерированном и эталонном текстах [11]
- ROUGE-1, ROUGE-2, ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) — набор метрик, оценивающих покрытие n-грамм эталонного текста в сгенерированном [12]

III. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

A. *Динамика перплексии в процессе рекурсивного обучения*

Сравнительный анализ перплексии языковых моделей LLaMA и Mistral в контексте рекурсивного обучения выявил существенные различия в динамике модельных характеристик. На рисунке 1, представляющем результаты для меньшего объема данных, наблюдается принципиально различная траектория изменения перплексии для исследуемых моделей. Модель LLaMA демонстрирует резкое снижение перплексии с initial значения 1.2 до 1.05 в первых двух поколениях, что может указывать на высокую скорость адаптации к обучающим данным.

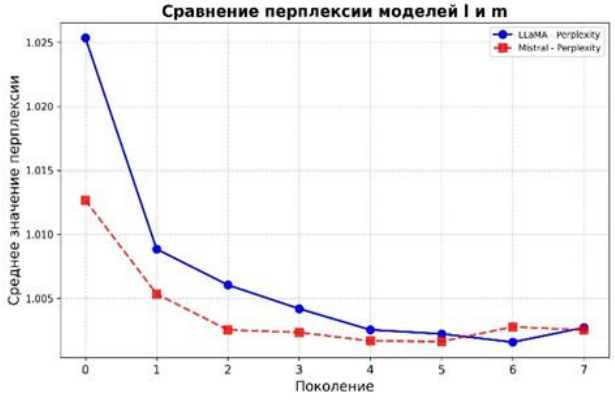


Рисунок 1 - График перплексии Llama 3 8B и Mistral 7B на обучающей выборке объемом 1000 строк

Напротив, модель Mistral характеризуется практически линейной и минимальной динамикой изменения перплексии, сохраняя стабильность на всем протяжении эксперимента.

Принципиально иная картина спектра перплексии представлена на рисунке 2 с расширенным датасетом объемом 2000 строк. В этом сценарии модель LLaMA приобретает волнообразный характер изменения перплексии с последовательным ростом до значений около 1.15, что может служить индикатором нестабильности внутренних репрезентаций при увеличении объема обучающей выборки. Статистически значимые флуктуации перплексии для LLaMA свидетельствуют о потенциальной уязвимости модели к эффекту забывания и риску модельного коллапса.

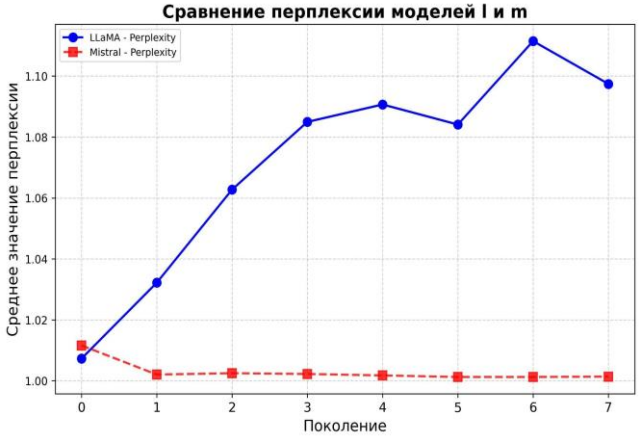


Рисунок 2 - График перплексии Llama 3 8B и Mistral 7B на обучающей выборке объемом 2000 строк

Модель Mistral демонстрирует диаметрально противоположную динамику – исключительно стабильный уровень перплексии, близкий к константе 1.0, независимо от числа поколений и объема экспериментальных данных. Такая консервативная поведенческая стратегия может указывать на более надежную архитектуру модели и эффективные механизмы внутренней регуляции при рекурсивном обучении.

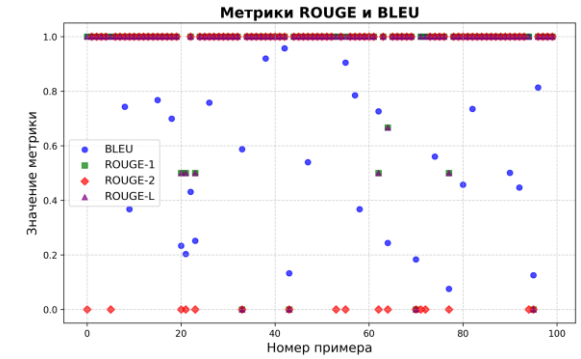
Полученные результаты эмпирически подтверждают гетерогенность поведения языковых моделей в условиях рекурсивного обучения и необходимость разработки дифференцированных подходов к оценке их производительности. Наблюдаемые различия в динамике перплексии LLaMA и Mistral требуют дальнейших углубленных исследований с привлечением дополнительных метрик и экспериментальных сценариев.

B. *Анализ метрик BLEU и ROUGE*

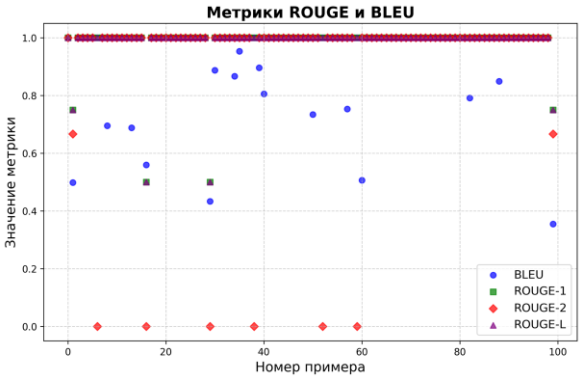
После обучения нескольких поколений моделей был выбран тестовый датасет, который не использовался при обучении ни одной из моделей, и на нем посчитаны метрики BLEU, ROUGE-1, ROUGE-2, ROUGE-L для проверки качества

генерации. Это было сделано для исключения влияния разных выборок на объективность сравнения моделей.

На рисунке 3 представлено распределение значений метрик по примерам датасета для модели l0. График демонстрирует значительную вариативность значений, что свидетельствует о высокой чувствительности модели к специфике входного текста: в некоторых случаях предсказания почти полностью совпадают с эталоном (значения близки к 1.0), тогда как в других демонстрируют значительные расхождения (значения близки к 0).



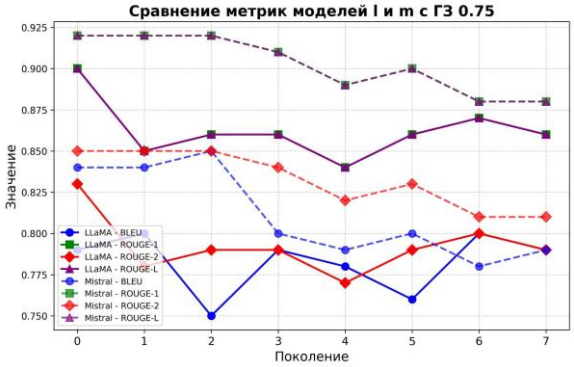
а) 1000



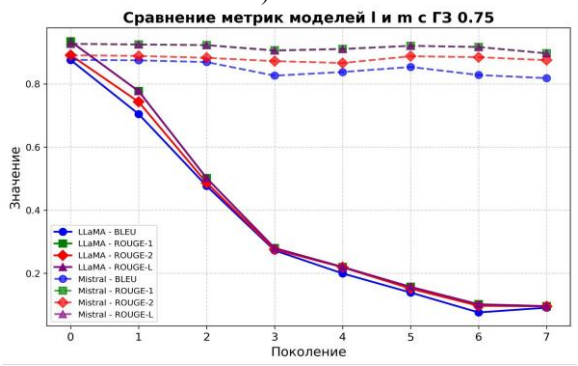
б) 2000

Рисунок 3 - Распределение значений метрик BLEU и ROUGE для модели l0

Для обобщенной оценки была использована методика анализа доли значений, превышающих установленное граничное значение (ГЗ), что позволило уменьшить влияние конкретных примеров входных данных и сравнить разные поколения моделей. На рисунках 4–6 представлены графики долей значений метрик в зависимости от поколения модели для разных ГЗ (0.75, 0.8 и 0.9).

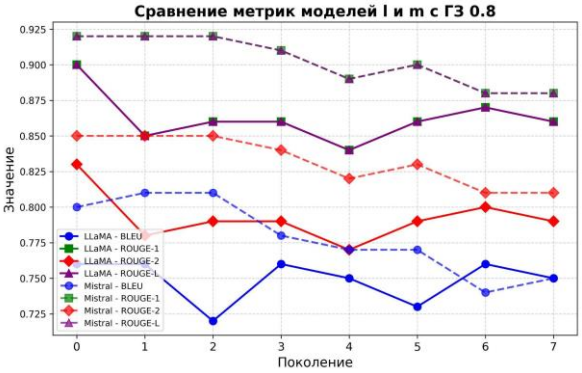


а) 1000

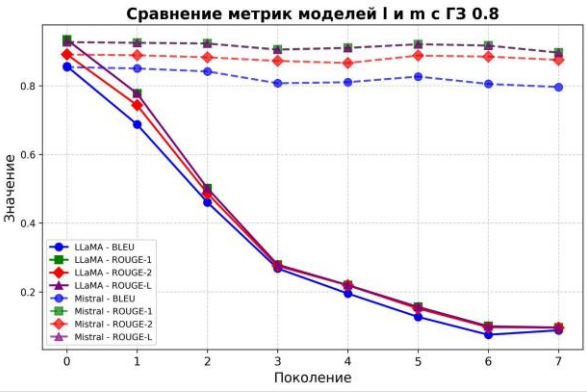


б) 2000

Рисунок 4 - Графики долей значений метрик в зависимости от поколения модели для ГЗ 0.75

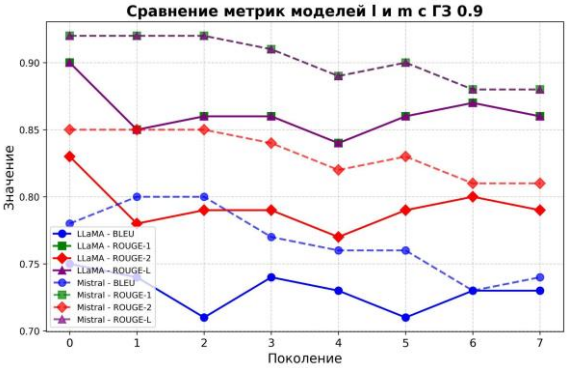


а) 1000

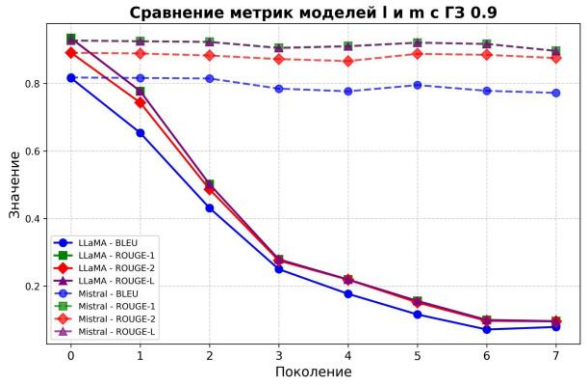


б) 2000

Рисунок 5 - Графики долей значений метрик в зависимости от поколения модели для ГЗ 0.8



а) 1000



б) 2000

Рисунок 6 - Графики долей значений метрик в зависимости от поколения модели для ГЗ 0.9

Представленные графики демонстрируют динамику значений метрик BLEU, ROUGE-1, ROUGE-2 и ROUGE-L для моделей LLaMA и Mistral в зависимости от поколения, учитывая три различных граничных значения. Видно, что модель Mistral демонстрирует большую стабильность и высокие значения ROUGE и BLEU по сравнению с LLaMA на протяжении всех поколений, что свидетельствует о лучшем сохранении текстового сходства.

Отчетливо прослеживается тенденция к снижению значений характеристик, что подтверждает постепенную деградацию качества генерируемого текста от поколения к поколению. При увеличении порогового значения с 0.75 до 0.9 наблюдается ожидаемое уменьшение доли высоких значений метрик, однако общие тренды сохраняются. В среднем наблюдается снижение метрик относительно базовых моделей на 5%, что указывает на начальную стадию коллапса.

C. Анализ распределения вероятности токенов

Для более детального понимания механизмов коллапса был проведен анализ внутренних процессов работы модели, а именно распределения вероятностей при генерации токенов. В ходе эксперимента был выбран случайный токен в предложении, и для него были построены распределения вероятностей, предсказываемых моделью. Анализ представленных распределений показывает, что с увеличением поколения модели наблюдается рост уверенности в предсказании конкретного токена. Это выражается в увеличении вероятности предсказания наиболее вероятного токена и снижении вероятностей альтернативных вариантов. Подобное изменение распределения вероятностей может свидетельствовать о снижении разнообразия генерируемого текста.

Для более наглядного представления изменений в распределении вероятностей были построены графики вероятностей токенов, за исключением наиболее вероятного (рисунки 7–12).

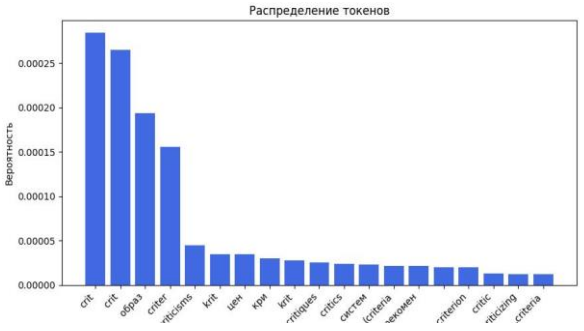


Рисунок 7 - Распределение вероятностей токенов, за исключением наиболее вероятного, модели l0

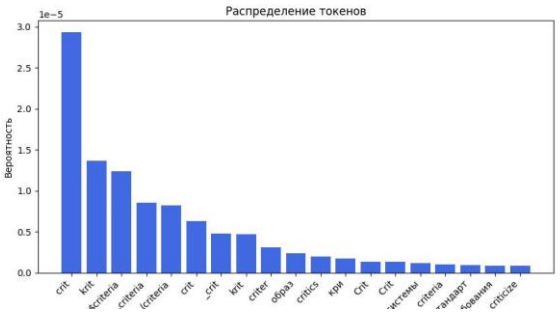


Рисунок 8 - Распределение вероятностей токенов, за исключением наиболее вероятного, модели l3

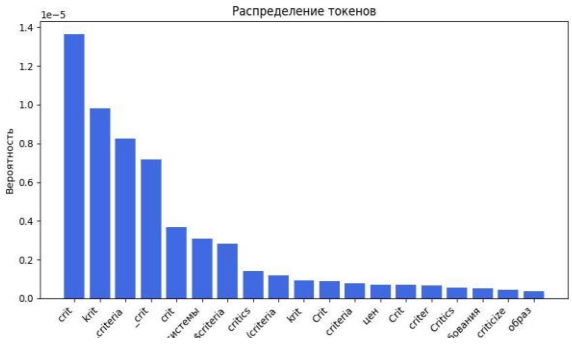


Рисунок 9 - Распределение вероятностей токенов, за исключением наиболее вероятного, модели l5

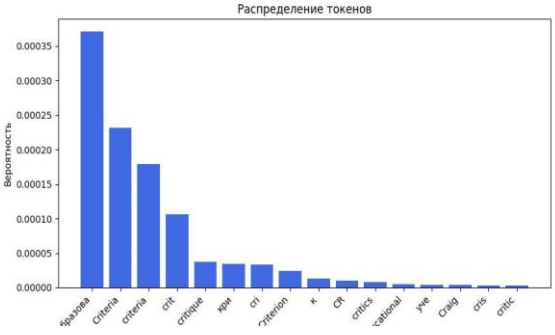


Рисунок 10 - Распределение вероятностей токенов, за исключением наиболее вероятного, модели m0

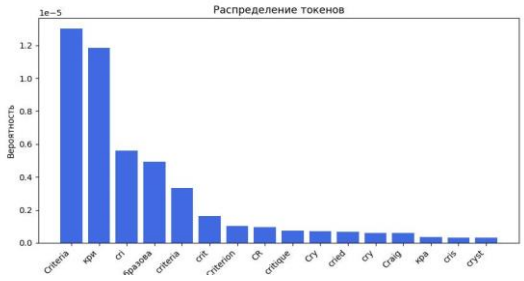


Рисунок 11 - Распределение вероятностей токенов, за исключением наиболее вероятного, модели m3

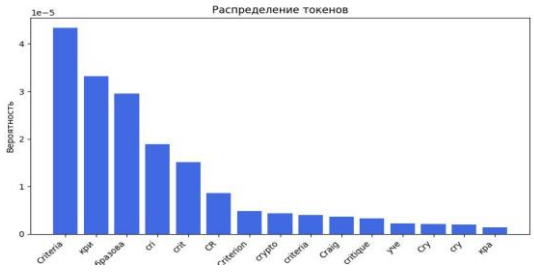


Рисунок 12 - Распределение вероятностей токенов, за исключением наиболее вероятного, модели m5

На представленных распределениях можно наблюдать, что вероятности токенов, отличных от наиболее вероятного, для моделей 3-го и 5-го поколений на порядок ниже по сравнению с базовой моделью. Это свидетельствует о процессе "забывания" маловероятных токенов и сужению распределения, что, в итоге, приводит к снижению разнообразия генерируемого текста.

В Ыисследовании [13] данный эффект был обозначен как "поздний коллапс модели". Это явление характеризуется тем, что модель начинает концентрироваться на небольшом подмножестве токенов из своего словаря, игнорируя остальные варианты, что приводит к генерации однообразных и предсказуемых текстов.

IV. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

A. Типология коллапса моделей: ранний и поздний коллапс

На основе проведенного анализа можно выделить два дистинктных типа коллапса языковых моделей при рекурсивном обучении: ранний и поздний.

Ранний коллапс характеризуется быстрым и резким изменением распределения вероятностей уже в первых 2-3 поколениях. К его основным признакам относятся:

- стремительное снижение метрик качества (перплексии и точности на валидационной выборке),
- быстрое сужение распределения вероятностей выходных токенов,
- потеря разнообразия генерации на ранних этапах обучения.

Экспериментальные данные демонстрируют неоднозначные признаки коллапса в модели LLaMA при масштабировании объема обучающих данных. В частности:

- на датасете из 2000 примеров наблюдается резкое ухудшение метрик качества уже на ранних итерациях, что соответствует паттерну раннего коллапса,
- на сокращенном датасете (1000 примеров) распределения выходов визуально неотличимы от таковых у модели Mistral, а деградация метрик выражена слабее, что характерно для позднего коллапса.

Этот парадокс может быть объяснен ограниченной ёмкостью модели (0.7 бит/параметр при квантовании в int4) [14]. В случае обучения на 2000 примерах:

- ёмкость модели исчерпывается практически сразу,
- наблюдается резкое падение качества вследствие преждевременной сходимости к субоптимальным локальным минимумам.

Напротив, при обучении на 1000 примерах:

- модель насыщается медленнее из-за меньшего объема данных,
- деградация происходит позже, демонстрируя паттерны позднего коллапса.

Поздний коллапс имеет более заторможенный характер, развиваясь на протяжении 4–7 поколений. Его ключевыми характеристиками являются:

1. плавное снижение перплексии (на 5–10% за поколение),
2. постепенное сужение распределения вероятностей,
3. сохранение ограниченного разнообразия предсказаний в течение нескольких поколений,
4. медленная, но устойчивая деградация метрик ROUGE и BLEU (падение на 2–5% за поколение).

Модели семейства Mistral (m0-m7) демонстрируют типичные признаки позднего коллапса, сохраняя относительное разнообразие предсказаний до 5-го поколения. Данный тип коллапса представляет особую сложность для ранней диагностики, что требует:

- тщательного мониторинга внутренних состояний модели,
- разработки специализированных метрик для отслеживания постепенной деградации,
- реализации механизмов превентивного вмешательства.

B. Причины различных типов коллапса

Наблюдаемые различия между ранним и поздним коллапсом могут быть обусловлены следующими факторами.

1. Архитектурные особенности: Механизм sliding window attention в Mistral обеспечивает более стабильное распределение внимания на длинные контексты, что замедляет процесс коллапса. В отличие от этого, более традиционный механизм внимания в LLaMA

быстрее фокусируется на доминирующих паттернах, ускоряя процесс деградации.

2. Различия в функциях активации и нормализации: Модели используют различные комбинации функций активации и нормализации, что влияет на устойчивость распределения вероятностей. Mistral применяет более агрессивную нормализацию, предотвращающую чрезмерную концентрацию вероятностной массы.
3. Инициализация и предобучение: Различия в начальной инициализации весов и процессе предобучения могут создавать разную "инерцию" распределений, влияя на скорость их деградации при рекурсивном обучении.

На основе полученных результатов можно сформулировать следующие рекомендации по предотвращению коллапса при рекурсивном обучении языковых моделей.

1. Смешивание синтетических и реальных данных. Включение определенной доли реальных данных в обучающий набор может помочь сохранить разнообразие и предотвратить чрезмерное усиление существующих паттернов.
2. Регуляризация, т. е. использование методов регуляризации, таких как dropout, label smoothing и т. д., может способствовать сохранению разнообразия в распределении вероятностей.
3. Мониторинг распределения вероятностей. Регулярное отслеживание изменений в распределении вероятностей токенов позволяет выявить ранние признаки коллапса и принять соответствующие меры.
4. Выбор архитектуры. При необходимости рекурсивного обучения предпочтение стоит отдавать архитектурам, демонстрирующим большую устойчивость к коллапсу, таким как Mistral-7B.

ЗАКЛЮЧЕНИЕ

Проведенное исследование позволило выявить и охарактеризовать два дистинктивных типа коллапса языковых моделей при рекурсивном обучении: ранний и поздний. Ранний коллапс, наблюдаемый в моделях LLaMA, характеризуется стремительной деградацией распределения вероятностей и быстрым снижением разнообразия генерации уже в первых поколениях. Поздний коллапс, свойственный моделям Mistral, имеет более постепенный характер, позволяя сохранять приемлемое качество и разнообразие генерации на протяжении нескольких поколений.

Анализ динамики изменения формы распределения вероятностей токенов предоставляет новый инструмент для раннего выявления признаков коллапса, что особенно важно для больших языковых моделей, применяемых в критических приложениях. Предложенная математическая модель эволюции распределения вероятностей открывает возможности для прогнозирования и предотвращения коллапса в процессе обучения и дообучения моделей.

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] Cooper N., Scholak T. Perplexed: Understanding when large language models are confused //arXiv preprint arXiv:2404.06634. – 2024

[2] Mezzoudj F., Benyettou A. An empirical study of statistical language models: n-gram language models vs. neural network language models //International Journal of Innovative Computing and Applications. – 2018. – Т. 9. – №. 4. – С. 189-202.

[3] Gritsai, G.M., Khabutdinov, I.A. & Grabovoy, A.V. Stack More LLM's: Efficient Detection of Machine-Generated Texts via Perplexity Approximation. Dokl. Math. 110 (Suppl 1), S203–S211 (2024): https://doi.org/10.1134/S1064562424602075

[4] Canvas4Everyone. Unraveling the Mystery of Perplexity: A Deep Dive into Likelihood Scores [Электронный ресурс]. URL: https://canvas4everyone.com/blogs/news/unraveling-the-mystery-of-perplexity-a-deep-dive-into-likelihood-scores (дата обращения: 27.03.2025).

[5] Chang Y. et al. A survey on evaluation of large language models //ACM transactions on intelligent systems and technology. – 2024. – Т. 15. – №. 3. – С. 1-45.

[6] UpTrain Blog. Decoding Perplexity and Its Significance in LLMs [Электронный ресурс]. URL: https://blog.uptrain.ai/decoding-perplexity-and-its-significance-in-llms/ (дата обращения: 27.03.2025).

[7] Madala, Sudheer. Introduction to Probability Theory in NLP [Электронный ресурс] // Scaler Topics. URL: https://www.scaler.com/topics/nlp/probability-theory-nlp/ (дата обращения: 27.03.2025).

[8] Gu J. et al. Do LLMs Play Dice? Exploring Probability Distribution Sampling in Large Language Models for Behavioral Simulation //arXiv preprint arXiv:2404.09043. – 2024.

[9] Ali S, Cibas E. The Bethesda System for Reporting Thyroid Cytopathology. (Ali SZ, Cibas ES, eds.). Cham: Springer International Publishing; 2018. doi: https://doi.org/10.1007/978-3-319-60570-8

[10] Ali SZ, Baloch ZW, Cochand-Priollet B, Schmitt FC, Vielh P, VanderLaan PA. The 2023 Bethesda System for Reporting Thyroid Cytopathology. Thyroid®. July 2023. doi: https://doi.org/10.1089/thy.2023.0141

[11] Papineni K. et al. Bleu: a method for automatic evaluation of machine translation //Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – С. 311-318.

[12] Lin C. Y. Rouge: A package for automatic evaluation of summaries //Text summarization branches out. – 2004. – С. 74-81.
[13] Shumailov I. et al. AI models collapse when trained on recursively generated data //Nature. – 2024. – Т. 631. – №. 8022. – С. 755-759.
[14] Allen-Zhu Z., Li Y. Physics of language models: Part 3.3, knowledge capacity scaling laws //arXiv preprint arXiv:2404.05405. – 2024.

Статья получена 20 июля 2025.

Боброва Елизавета Витальевна, Национальный Исследовательский Ядерный Университет МИФИ, аспирант, EVBobrova@mephi.ru
Зайцев Константин Сергеевич, Национальный Исследовательский Ядерный Университет МИФИ, профессор, KSZaytsev@mephi.ru
Свириденко Дмитрий Константинович, Национальный Исследовательский Ядерный Университет МИФИ, магистрант dmitrii.sviridenko@yandex.ru
Холод Данила Витальевич, Национальный Исследовательский Ядерный Университет МИФИ, магистрант, danila.kholod@gmail.com
Дюльдин Евгений Владимирович, Национальный Исследовательский Ядерный Университет МИФИ, аспирант, Zhecos1@yandex.ru

.

An investigation of early and late collapse of language models in medical applications

E.V. Bobrova, K.S. Zaytsev, D.K. Sviridenko, D.V. Kholod, E.V. Dyuldin

Abstract. The aim of the work is a comprehensive analysis of the mechanisms of early collapse of language models working with medical texts during their recursive training using the example of the Mistral-7B and LLaMA-3 architectures. An experimental study of the dynamics of perplexity change, BLEU and ROUGE metrics, as well as the probability distribution of tokens in the process of multi-generation synthetic training was conducted. Two types of model collapse are identified: early (characterized by rapid degradation of probability distributions) and late (with a gradual decrease in the diversity of generation). It is established that the Mistral model demonstrates greater resistance to data collapse compared to LLaMA, which is due to the features of its architecture with a sliding window attention mechanism. The paper proposes a new methodological approach to quantifying the degradation of language models and formulates practical recommendations for preventing the loss of model diversity during recursive learning. The study was conducted on text cytological data used in the diagnosis of thyroid diseases.

Keywords – Collapse mechanisms, perplexity, LLM degradation, recursive learning, probability distribution

REFERENCES

[1] Cooper N., Scholak T. Perplexed: Understanding when large language models are confused //arXiv preprint arXiv:2404.06634. – 2024

[2] Mezzoudj F., Benyettou A. An empirical study of statistical language models: n-gram language models vs. neural network language models //International Journal of Innovative Computing and Applications. – 2018. – Т. 9. – №. 4. – С. 189-202.

[3] Gritsai, G.M., Khabutdinov, I.A. & Grabovoy, A.V. Stack More LLM's: Efficient Detection of Machine-Generated Texts via Perplexity Approximation. Dokl. Math. 110 (Suppl 1), S203–S211 (2024): https://doi.org/10.1134/S1064562424602075

[4] Canvas4Everyone. Unraveling the Mystery of Perplexity: A Deep Dive into Likelihood Scores [Электронный ресурс]. URL: https://canvas4everyone.com/blogs/news/unraveling-the-mystery-of-perplexity-a-deep-dive-into-likelihood-scores (дата обращения: 27.03.2025).

[5] Chang Y. et al. A survey on evaluation of large language models //ACM transactions on intelligent systems and technology. – 2024. – Т. 15. – №. 3. – С. 1-45.

[6] UpTrain Blog. Decoding Perplexity and Its Significance in LLMs [Электронный ресурс]. URL: https://blog.uptrain.ai/decoding-perplexity-and-its-significance-in-llms/ (дата обращения: 27.03.2025).

[7] Madala, Sudheer. Introduction to Probability Theory in NLP [Электронный ресурс] // Scaler Topics. URL: https://www.scaler.com/topics/nlp/probability-theory-nlp/ (дата обращения: 27.03.2025).

[8] Gu J. et al. Do LLMs Play Dice? Exploring Probability Distribution Sampling in Large Language Models for Behavioral Simulation //arXiv preprint arXiv:2404.09043. – 2024.

[9] Ali S, Cibas E. The Bethesda System for Reporting Thyroid Cytopathology. (Ali SZ, Cibas ES, eds.). Cham: Springer International Publishing; 2018. doi: https://doi.org/10.1007/978-3-319-60570-8

[10] Ali SZ, Baloch ZW, Cochand-Priollet B, Schmitt FC, Vielh P, VanderLaan PA. The 2023 Bethesda System for Reporting Thyroid Cytopathology. Thyroid®. July 2023. doi: https://doi.org/10.1089/thy.2023.0141

[11] Papineni K. et al. Bleu: a method for automatic evaluation of machine translation //Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – С. 311-318.

[12] Lin C. Y. Rouge: A package for automatic evaluation of summaries //Text summarization branches out. – 2004. – С. 74-81.

[13] Shumailov I. et al. AI models collapse when trained on recursively generated data //Nature. – 2024. – Т. 631. – №. 8022. – С. 755-759.

[14] Allen-Zhu Z., Li Y. Physics of language models: Part 3.3, knowledge capacity scaling laws //arXiv preprint arXiv:2404.05405. – 2024.

..