

УДК 81 '32

Липатова Дарья Андреевна
студент департамента филологии
Научный исследовательский
университет
«Высшая школа экономики»
г. Санкт-Петербург, Россия

Daria Lipatova
Student of the Department of Philology
National Research University «Higher
School of Economics»
Saint Petersburg, Russia
dalipatova@edu.hse.ru

СРАВНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ И МАРКОВСКИХ ЦЕПЕЙ ДЛЯ ГЕНЕРАЦИИ АНЕКДОТОВ НА РУССКОМ ЯЗЫКЕ

В статье рассматриваются возможности автоматической генерации анекдотов на русском языке с использованием марковских цепей и больших языковых моделей (large language models, LLM). Целью исследования является сравнительный анализ этих методов в задаче генерации юмористических текстов, учитывающих формальные признаки анекдота. Для обучения марковских моделей был собран корпус из более чем 300 000 анекдотов, очищенный от дубликатов и выбросов. Результаты показали, что увеличение параметра *state_size* улучшает связность и осмысленность текстов, но снижает их разнообразие, при этом многие генерации представляют собой комбинации уже существующих анекдотов. В работе также тестировались LLM: GPT-4, YandexGPT (3.0, 5.0 Lite, 5.0 Pro), Gemini (2.0 Flash, 2.0 Flash Thinking, 2.5 Pro) и Grok 3. Выяснилось, что некоторые из них (особенно GPT-4 и Gemini 2.0 Flash) лучше сохраняют ключевые признаки жанра анекдота, создавая тексты с «речью от автора» и характерными персонажами, тогда как другие отклоняются от этих признаков или выдают заготовленные варианты. Результаты исследования подчеркивают необходимость дальнейших экспериментов с промптами и обучающими данными для повышения оригинальности и качества генерации.

Ключевые слова: обработка естественного языка; автоматическая генерация текста; анекдоты; большие языковые модели; марковские цепи.

COMPARISON OF LARGE LANGUAGE MODELS AND MARKOV CHAINS FOR GENERATING JOKES IN RUSSIAN

The article deals with the possibilities of automatic generation of anecdotes in Russian using Markov chains and large language models (LLM). The aim of the study is a comparative analysis of these methods in the task of generating humorous texts that take into account the formal features of the anecdote. A corpus of more than 300,000 anecdotes, cleaned of duplicates and outliers, was collected to train the Markov models. Results showed that increasing the *state_size* parameter improves the coherence and meaningfulness of the texts but reduces their diversity, with many generations being combinations of pre-existing anecdotes. The LLMs tested in this work are GPT-4, YandexGPT (3.0, 5.0 Lite, 5.0 Pro), Gemini (2.0 Flash, 2.0 Flash Thinking, 2.5 Pro) and Grok 3. It turned out that some of them (especially GPT-4 and Gemini 2.0 Flash) were better at preserving key features of the anecdote genre, producing texts with “speech from the author” and distinctive characters, while others deviated from these features or produced pre-prepared variants. The results of the study emphasize the need for further experimentation with prompts and training data to improve originality and generation quality.

Key words: natural language processing; automatic text generation; anecdotes; large language models; Markov chains.

В условиях автоматизации различных областей человеческой деятельности, включая сферу творчества [1], актуальным становится исследование возможностей генерации юмористических текстов средствами искусственного интеллекта [2; 3]. Анекдоты как лингвокультурный феномен представляют собой сложный объект для генерации, так как их создание требует не только владения языковыми шаблонами, но и учета контекстуальных и когнитивных факторов, лежащих в основе комического эффекта. Однако на текущий момент не было обнаружено примеров систематического анализа генерации анекдотических текстов на русском языке.

Целью этой работы является сравнительный анализ больших языковых моделей (Large Language Models, LLM) и марковских цепей в задаче генерации анекдотов на русском языке. Выбор данных методов обусловлен их принципиальным различием: марковские цепи являются классическим статистическим подходом к генерации текста, в то время как LLM представляют собой современный нейросетевой метод. Марковские цепи основаны на предположении, что вероятность появления следующего элемента текста зависит только от текущего состояния, а не от всей предшествующей последовательности. Большие языковые модели работают на архитектуре трансформеров, которая анализирует текст с помощью механизма внимания, позволяющего находить взаимосвязи и зависимости токенов, тем самым учитывая больший контекст, чем другие модели. Такое сопоставление позволяет провести наглядное сравнение простых и продвинутых алгоритмов генерации.

Для сравнения результатов генерации были выбраны признаки анекдота, выделенные в работе Е. Я. Шмелевой и А. Д. Шмелева [4]. Авторы обращают внимание на наличие постоянного набора персонажей с устойчивыми речевыми и поведенческими характеристиками, известными носителям языка и не требующими представления [4, с. 23]. Кроме того, они акцентируют внимание на структуре текста: разделении на часть «от автора» и речь персонажей [4, с. 31]. Текст «от автора» подчиняется специфическим синтаксическим требованиям: начальные предложения часто имеют не прямой порядок слов, а глаголы преимущественно употребляются в настоящем времени или в форме совершенного вида прошедшего времени с результативным значением [4, с. 32–33].

Материалом исследования стали данные, собранные с сайтов, где пользователи размещают анекдоты (Таблица 1), общим объемом 358408 текстов. После удаления дубликатов, пустых записей и выбросов по длине размер итогового датасета составил 309723 анекдота.

Т а б л и ц а 1

Источники анекдотов для датасета

Адрес ресурса	Подраздел	Кол-во анекдотов
https://anekdotov.net/	«Популярно сегодня» с 1 января 2006 года по 15 февраля 2025 года	36868
https://anekdoty.ru/	Страницы категорий (напр. «про Вовочку»)	6295
https://www.anekdot.ru/	Страница самых популярных анекдотов дня с 8 ноября 1995 года по 15 февраля 2025 года	246598
https://anekdotovstreet.com/	Анекдоты дня с 1 января 2019 года по 15 февраля 2025 года	68647

Сначала рассмотрим результаты, полученные при использовании марковских цепей, для обучения которых применялась библиотека `markovify` [5]. Для генерации был выбран метод `markovify.NewLineText` с параметрами по умолчанию, так как он учитывает переносы строк как границы анекдотов в отличие от `markovify.Text`, который полагается на точки как маркеры конца предложения. Пример генерации:

(1) Так или иначе, но всё равно по ней ботинком, чтобы кому-то понравиться.– А тебе какого размера это вырастет, а уже ссоримся... – Таки я подожду.

Основная проблема результатов – отсутствие смысловой связности: хотя предложения грамматически корректны, их части логически не связаны. Для улучшения результатов был увеличен параметр `state_size` (количество предыдущих слов, учитываемых при генерации следующего) с 2 до 3. Некоторые результаты:

(2) У нас дома все решает папа. А кто у тебя? – Кот. – Какой? – Разжиреловый.

(3) Бредет грустный Жириновский в одиночестве по Виа Монтенаполеоне мимо шикарных миланских бутиков и думает: – Главное, чтобы по утрам меня одевали и причесывали птички, а то разнесу всю деревню к такой-то матери. Ну, в смысле, справа налево. Много думал.

Качество генерации улучшилось: пример (2) демонстрирует полностью связный и смешной текст. В случае (3) осмысленные фрагменты стали длиннее, хотя полной связности достичь не удалось.

При дальнейшем увеличении `state_size` до 4 модель начала иногда возвращать `None` – это происходит, когда невозможно сгенерировать текст, удовлетворяющий условию совпадения с оригиналом менее чем на 70 % (`max_overlap_ratio = 0.7`). Тем не менее, большинство генераций остаются успешными:

(4) Где-то в России. Доктор – пациенту: – Поздравляю. – С чем?! – Я болел

(5) – Потерпевшая! Назовите наиболее яркую примету преступника? – Он был гонщиком, Нурислам. – Автогонщиком? – Самогонщиком.

Хотя некоторые сгенерированные предложения остаются нелогичными, доля удачных примеров возросла. Наблюдается закономерность, отмеченная авторами исследования [6]: успешные генерации часто представляют собой комбинации нескольких анекдотов из датасета [6, с. 3]. Например, пример (5) объединяет два анекдота:

1) – Потерпевшая! Назовите наиболее яркую примету преступника? – Он был без маски.

2) – Бабуль, а почему ты в деда влюбилась? – Он был гонщиком, Нурислам. – Автогонщиком? – Самогонщиком.

Дальнейшее увеличение `state_size` не приводит к существенному улучшению качества, а лишь повышает частоту появления `None` и текстов, практически дословно повторяющих анекдоты из датасета.

Таким образом, анекдотам, сгенерированным марковскими цепями, удастся сохранить особенности данных, на которых они обучались, в частности, диалогическую форму и речь автора в некоторых случаях. Однако часто их сложно назвать оригинальными, так как они часто представляют собой комбинацию из имеющихся в датасете текстов. Это подтверждает ограниченность классических марковских цепей.

Далее обратимся к генерациям, полученным в ходе применения больших языковых моделей. Для того, чтобы иметь возможность выделить общие особенности генераций языковых моделей, мы использовали несколько: GPT-4 [7], YandexGPT [8] (версии 3.0, 5.0 Lite, 5.0 Pro), Gemini [9] (2.0 Flash, 2.0 FlashThinking, 2.5 Pro) и Grok 3 [10]. Во всех случаях генерация производилась по промпту «сгенерируй анекдот», выбранному из-за своей короткости и нейтральности. В некоторых случаях были попытки генерации для других коротких промптов, о них указано отдельно.

Для первых генераций используется модель GPT-4 в интерфейсе ChatGPT от OpenAI. Рассмотрим результаты 10 генераций без ограничений по промпту «сгенерируй анекдот». Все сгенерированные тексты имеют форму диалога, но только в одном из результатов перед диалогом присутствует «речь от автора»:

(6) Мужчина приходит к психиатру:

– Доктор, у меня проблема – я постоянно разговариваю сам с собой.

– И вам это мешает?

– Нет, но я такой зануда, что даже себя раздражаю!

Можно заметить, что эта «речь от автора» тоже не совсем отвечает признакам анекдота: хотя в ней используется глагол в настоящем времени, порядок слов остается прямым, а не меняется на обратный.

Получившиеся анекдоты можно разделить на несколько групп по действующим героям. 3 из них можно отнести к категории «у врача»: все из

них представляют собой диалог доктора и пациента, и диалогическая часть начинается с обращения «доктор» (– *Доктор, у меня плохая память...*; – *Доктор, у меня постоянно такое чувство...*). 5 представляют собой диалог мужа и жены, диалог начинается с обращений «дорогая», «дорогой» или «милый». Оставшиеся два – звонок в службу поддержки и разговор с начальником. Можно сделать вывод, что в получившихся анекдотах сохраняется наличие постоянных для анекдотов героев.

Анекдоты получились разными по качеству, но для наиболее удачных из них часто можно найти очень похожие в интернете. Примеры представлены в Таблице 2.

Т а б л и ц а 2

Сравнение некоторых результатов генерации ChatGPT
с анекдотами в открытом доступе

Генерация ChatGPT	Анекдот из интернета
– Дорогой, ты где? – На рыбалке! – А кто это там хихикает? – Рыбы!	Звонок на мобильник: – Милый, ты где? – Я на рыбалке. – А кто это там так громко дышит? – Это щука.
– Доктор, у меня постоянно такое чувство, что меня все игнорируют! – Следующий!	– Доктор, мне кажется, меня все игнорируют! – Следующий!

Эти сходства затрудняют возможность говорить об этих анекдотах как о сгенерированных нейросетью, так как их юмористическая часть была создана человеком.

Перейдем к результатам генерации при помощи YandexGPT от компании «Яндекс». При генерации в базовом режиме (модель YandexGPT версии 3.0) анекдоты часто повторяются и, по своей сути, не являются генерацией – модель просто выдает один из заложенных заранее вариантов. При запросе «придумай анекдот сама» модель отвечает «Я не могу придумывать анекдоты, но могу рассказать вам один», и далее выдает один из имеющихся анекдотов.

При генерации в продвинутом режиме (модель YandexGPT версии 5.0) при запросе «сгенерируй анекдот» модель продолжает выдавать заготовленные анекдоты. Однако при запросе «придумай анекдот сама» начинают появляться уникальные тексты. Несколько примеров:

- (7) *Дорогой, я тебе борщ сварила!*
А можно сначала посмотреть его резюме?

(8) Почему программисты всё путают?

Потому что они начинают считать с нуля.

(9) Почему программисты не любят ходить на похороны?

Потому что там нельзя сказать: “Давайте оптимизируем этот процесс!”

По примерам видно, что в сгенерированных текстах есть типичные персонажи, но по другим признакам от анекдота они отличаются. В примере (7) есть речь героев, но нет речи от автора (хотя ее можно несложным образом воссоздать из контекста). Примеры (8) и (9) от анекдота еще дальше, так как в них нет даже прямой речи героев (эти тексты скорее можно назвать короткими шутками). Также можно заметить, что все результаты недлинные и ограничены двумя репликами, что может говорить о некоторой однотипности результатов при генерации без дополнительных условий.

Кроме использованных ранее моделей в интерфейсе Алисы, можно использовать модели YandexGPT 5 Lite и YandexGPT 5 Pro – модели без поиска и других дополнительных функций Алисы. При генерации с Yandex GPT 5 Lite и запросе «придумай анекдот сама» получились тексты, более удачные по выделенным признакам, чем сгенерированные ранее:

(10) Встречаются два математика:

– Ты слышал, у меня кошка научилась считать до трёх!

– Удивительно! А моя собака освоила интегралы. Теперь она считает, что кость длиной в два интеграла лучше, чем кость длиной в один интеграл.

(11) Встретились на улице два филолога. Один говорит:

– Знаете, я сегодня увидел такое редкое и красивое слово, что решил запомнить его навсегда!

– И какое же?

– «Досвидос»! Теперь буду использовать его вместо прощания.

В получившихся анекдотах прослеживается наличие типичных персонажей, есть «речь от автора» и прямая речь героев; «речь от автора» имеет обратный порядок слов, и в ней используются глаголы в настоящем или прошедшем времени. Эти анекдоты уже длиннее, чем генерации предыдущей модели, но у них есть другая частотная особенность: тексты часто начинаются с «встретились/встречаются».

При генерации с YandexGPT 5 Pro получились следующие тексты:

(12) Почему кошка всегда падает на четыре лапы? Потому что у неё есть девять жизней на тренировке!

(13) Почему кролики любят играть в прятки? Потому что у них всегда есть запасное укрытие – собственная нора!

(14) Почему у фотографа всегда хорошее настроение? Потому что он знает: даже самый хмурый день можно превратить в яркий снимок!

Похожие результаты мы получали ранее: такие тексты скорее относятся к шуткам, чем к анекдотам. Здесь также можно обратить внимание на однотипность результатов.

Среди моделей Gemini от компании Google для генерации мы будем использовать 2.0 Flash, 2.0 FlashThinking и 2.5 Pro.

Некоторые примеры генерации с Gemini 2.0 Flash:

(15) Сидят два программиста в баре. Один говорит другому:

– Слушай, я вчера такое обновление установил, у меня теперь кофе сам варится!

Второй отвечает:

– Да ну? А у меня, знаешь, какая программа? Она мне носки стирает и гладит!

Первый с недоверием:

– Не может быть!

Второй:

– Ну, как тебе сказать... Иногда вылетает ошибка «Носки не найдены»...

(16) Встречаются два нейросети. Одна говорит другой:

– Я тут научилась писать стихи, как Пушкин!

Вторая отвечает:

– О, круто! А я могу генерировать картины, как Ван Гог!

Первая:

– Да ладно! А можешь нарисовать мне кота, играющего на пианино?

Вторая:

– Легко! Но... боюсь, у него будет девять пальцев и три хвоста.

Первая:

– Ну, это же нейросеть! Главное, чтобы было креативно!

Можно заметить, что получившиеся тексты длиннее, чем были во всех предыдущих моделях. Также интересно, что кроме наличия «речи от автора» (соответствующей критериям, выделенным ранее) в начале текста, она появляется в течение всего анекдота.

Генерация с «думающей» моделью Gemini 2.0 FlashThinking показала немного другие результаты:

(17) Два друга разговаривают. Один жалуется:

– Мой новый телефон такой умный, что постоянно меня перебивает!

Второй отвечает:

– А мой такой старый, что вообще не слушает!

(18) Приходит пациент к врачу и говорит:

– Доктор, у меня такое чувство, будто я курица!

Врач отвечает:

– Садитесь, больной, сейчас посмотрим ваши... кхм... перья.

Здесь анекдоты получились короче, но все характеристики, как и у результатов работы предыдущей модели, остались прежними.

Наконец, посмотрим на работу модели Gemini 2.5 Pro.

(19) – Внучек, у меня что-то с компьютером случилось! На экране какая-то картинка висит и ничего не нажимается!

– Бабуль, а какая картинка?

– Ну, такой цветочек красивый... в горшочке...

– Бабушка, это же заставка! Ты просто мышкой подвигай!

– Ах, вот оно что! Спасибо, родной! А я уже лейку приготовила, думала, полить его надо!

(20) Учительница на уроке математики спрашивает:

– Вовочка, сколько будет дважды два?

Вовочка, не задумываясь:

– Марь Иванна, а мы покупаем или продаем?

Здесь появляются примеры разной длины и разных форм: в первом примере нет «речи от автора», но во втором она появляется в нескольких местах.

Последняя модель, которую мы попробуем использовать для генерации, – это Grok 3, разработанный компанией xAI. Ниже представлены несколько результатов генерации:

(21) Мама спрашивает сына:

– Почему ты так долго сидишь за компьютером?

Сын:

– Мам, я не сижу, я спасаю мир от пиксельных монстров, это моя миссия!

(22) Парень приходит к гадалке и спрашивает:

– Скажите, когда я разбогатею?

Гадалка смотрит в шар и говорит:

– Через 10 лет, если перестанешь спрашивать глупости и начнешь работать.

(23) Учитель спрашивает ученика:

– Почему ты не сделал домашку?

Ученик:

– Да я вчера весь день пытался понять, почему у кота девять жизней, а у меня ни одной нормальной.

В результате работы этой модели получились небольшие анекдоты, в которых есть типичные персонажи, «речь от автора», в которой, однако, прямой порядок слов, и прямая речь героев. Стоит отметить сходство с генерациями Gemini, в которых речь автора появляется не только в начале анекдота, а перед каждой репликой.

В этой работе были проанализированы возможности генерации анекдотов с использованием марковских цепей и больших языковых моделей и были получены следующие выводы:

1. Марковские модели сохраняют те признаки анекдота, которые сохранены в примерах обучающей выборки, поэтому количество сохраненных

признаков варьируется. Их генерации часто представляют собой комбинацию больших фрагментов анекдотов из датасета, что может приводить к повышенному комическому эффекту, но снижает их оригинальность.

2. Некоторые LLM (особенно Gemini 2.0 Flash, GPT-4) более последовательно воспроизводят формальные признаки анекдота. Однако другие модели могут выдавать готовые шаблоны или шутки вместо анекдотов. Для некоторых LLM также характерна генерация анекдотов схожей структуры и длины.

В дальнейших исследованиях возможно сравнение генераций LLM при использовании различных промптов, дообучение моделей на более разнообразных и качественных датасетах.

ЛИТЕРАТУРА

1. Ivcevic Z., Grandinetti M. Artificial intelligence as a tool for creativity [Electronic resource] // Journal of Creativity. 2024. V. 34. № 2. DOI: <https://doi.org/10.1016/j.yjoc.2024.100079>
2. Gorenz D., Schwarz N. How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. [Electronic resource] // PLOS ONE 19(7). 2024. DOI: <https://doi.org/10.1371/journal.pone.0305364>
3. Ritchie G. Can Computers Create Humor? // AI Magazine. №30(3). 2009. P. 71–81.
4. Шмелева Е. Я., Шмелев А. Д. Русский анекдот. Текст и речевой жанр. М.: Языки славянской культуры, 2002. 144 с.
5. GitHub – jsvine/markovify: A simple, extensible Markov chain generator. [Electronic resource] URL: <https://github.com/jsvine/markovify> (date of access: 03.02.2025).
6. Chaudhary T., Goel M., Mamidi R. Towards Conversational Humour Analysis and Design. 2021. [Electronic resource] (date of access: 16.01.2025). DOI: <https://doi.org/10.48550/arXiv.2103.00536>
7. GPT-4 | OpenAI [Electronic resource] URL: <https://openai.com/index/gpt-4/> (date of access: 26.04.2025).
8. Нейросеть Алиса для решения любых задач. URL: <https://alice.yandex.ru/> (дата обращения: 28.04.2025).
9. Google Gemini [Electronic resource] URL: <https://gemini.google.com/> (date of access: 29.04.2025).
10. Grok [Electronic resource] URL: <https://grok.com/> (date of access: 29.04.2025).