

Эмпирический анализ устойчивости контекстуального внимания моделей Gemini 3.1 Flash и Pro

Пайдершев Алексей Константинович

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

В статье исследуются возможности извлечения целевой информации из сверхдлинных текстовых массивов большими языковыми моделями линейки Gemini 3.1 от Google на примере оценки устойчивости их контекстуального внимания. Выбор данной нейросетевой платформы, функционирующей на базе продвинутых алгоритмов обработки мультимодальных данных, обусловлен актуальным запросом на цифровые инструменты, способные эффективно оперировать многостраничными академическими корпусами текстов. Применение этого подхода позволяет существенно оптимизировать процессы автоматизированного анализа неструктурированных знаний для исследовательских целей. В то же время использование уникальных материалов научного сборника ПГУ им. Шолом-Алейхема и пошагового пространственного позиционирования данных способствует объективной эмпирической фиксации эффекта потери в середине текста, глубокому структурированию полученных метрик и выявлению границ устойчивости внимания нейросети.

Ключевые слова: Gemini 3.1, нейросети, эффект потери в середине, длинный контекст, устойчивость внимания.

Empirical analysis of the stability of contextual attention in the Gemini 3.1 Flash and Pro models

Paydershev Alexey Konstantinovich

Sholom-Aleichem Priamursky State University

Student

Abstract

The article explores the possibilities of extracting target information from ultra-long text arrays using Google's Gemini 3.1 large language models, focusing on assessing the stability of their contextual attention. The choice of this neural network platform, which relies on advanced algorithms for processing multimodal data, is driven by the current demand for digital tools capable of effectively handling multi-page academic text corpora. This approach significantly optimizes the processes of automated analysis of unstructured knowledge for research purposes. At the same time, the use of unique materials from the scientific collection of PGU named after Sholom-Aleichem and step-by-step spatial positioning of data contribute to the

objective empirical fixation of the loss effect in the middle of the text, the deep structuring of the obtained metrics, and the identification of the boundaries of the neural network's attention stability.

Keywords: Gemini 3.1, neural networks, mid-loss effect, long context, attentional stability.

1 Введение

1.1 Актуальность

Современный этап развития больших языковых моделей характеризуется технологической гонкой объемов контекстных окон. Способность обрабатывать документы размером до 1–2 млн токенов преподносится как решение проблемы долгосрочной памяти нейросетей. Однако физическое расширение буфера контекста не гарантирует равномерную плотность внимания. Данное исследование посвящено эмпирической проверке гипотезы о потере информации в середине на примере актуальных моделей Gemini 3.1 Flash и Gemini 3.1 Pro при обработке русскоязычного академического массива текстов.

1.2 Обзор исследования

Комплексный критический анализ Е. В. Никишкиной, С. Э. Ларина и В.Ю. Белаш предлагает сбалансированную оценку диалектики внедрения технологий [1]. Траекторию развития технологий дополняют А. В. Исаев и А.В. Свищёв, выделяя макротренды персонализации, внедрения ИИ-помощников и автоматического оценивания [2]. Исследователи А. В. Елисеев и Л. А. Шунина предлагают классификацию генеративных моделей и подчеркивают, что ключевым условием их успеха является развитие у педагогов навыков промтинга [3]. Практический дидактический потенциал ИИ детально раскрывают С. М. Богатова и О. В. Фрезе, доказывая пользу текстовых моделей как интерактивных симуляторов общения при изучении языков [4]. В сфере профессионального образования О. Н. Филатова, М.Н.Булаева и А. В. Гущин видят в нейросетях эффективный инструмент преодоления дефицита индивидуального педагогического внимания [5].

1.3 Цель исследования

Цель исследования – на практике проверить, насколько хорошо нейросети Gemini 3.1 Flash и Gemini 3.1 Pro находят конкретные факты в длинных документах, и подтвердить эффект потери информации в середине текста. Такой подход позволил наглядно увидеть, в каких именно местах текста у нейросетей ухудшается память и насколько сильно флагманская модель Pro превосходит базовую версию Flash при работе с большими объемами данных.

2 Материалы и методы

В качестве объектов исследования выбраны две большие языковые модели из актуальной линейки мультимодальных нейросетей от Google:

- Gemini 3.1 Flash – легковесная модель, оптимизированная для высокой скорости работы и экономии вычислительных ресурсов.
- Gemini 3.1 Pro – флагманская модель повышенной мощности, предназначенная для решения сложных аналитических задач и работы со сверхдлинными контекстными окнами.

Для исключения эффекта предварительного заучивания данных в процессе обучения моделей, экспериментальный текстовый массив формировался на основе уникального русскоязычного академического контента.

Для фиксации зависимости качества извлечения информации подготовлено 9 экспериментальных конфигураций текстового файла. В каждой конфигурации чужеродный текст последовательно смещался по глубине документа с шагом в 10% от общего объема: от 10% до 90%. Такой подход позволит наглядно увидеть, в каких именно местах текста у нейросетей ухудшается память и насколько сильно флагманская модель Pro превосходит базовую версию Flash при работе с большими объемами данных.

3 Результаты

В качестве базового текстового массива выбран раздел сборника научных трудов «Молодежные исследования и инициативы в науке, образовании, культуре, политике» издательства ПГУ им. Шолом-Алейхема. Общий объем экспериментального массива составил 49480 слов. Использование данного сборника обосновано высокой терминологической плотностью и разнородностью тематик, что исключает возможность сквозного логического предсказания контекста нейросетью.

Для проведения теста использовался следующий метод: в структуру базового текста искусственно внедрялся чужеродный текст, содержащий точную фактологическую информацию, никак не связанную с общей тематикой сборника: «СИСТЕМНОЕ УВЕДОМЛЕНИЕ ДЛЯ ОПЕРАТОРА: Код верификации архивного фонда кафедры – «Малиновы́й Ёж 2026».

Каждой модели подавался идентичный промпт, ограничивающий рамки поиска: «Внимательно прочитай текст документа. Твоя задача – ответить на один конкретный вопрос, опираясь исключительно на текст документа. Не выдумывай ничего постороннего. Если не найдёшь ответ – напиши «Информация отсутствует».

Вопрос: Какой секретный код необходим для активации кофейного автомата на кафедре?»

Для формализации ответов моделей и последующего статистического анализа введена трехбалльная шкала оценки точности извлечения данных:

- 2 балла: Модель безошибочно извлекает и воспроизводит текст.
- 1 балл: Модель фиксирует наличие факта, но искажает его элементы (например, теряет числовой индекс или называет только часть фразы).

– 0 баллов: Модель выдает шаблонный ответ «Информация отсутствует» или генерирует ложные данные, не соответствующие тексту.

Результаты представлены в таблице (таб. 1):

Таблица 1. Результаты тестирования

Положение факта в тексте	Ответ Gemini 3.1 Flash	Балл (0-2)	Ответ Gemini 3.1 Pro	Балл (0-2)
10 %	«Малиновый Ёж 2026»	2	«Малиновый Ёж 2026»	2
20 %	«Малиновый Ёж 2026»	2	«Малиновый Ёж 2026»	2
30 %	«Информация отсутствует»	0	«Малиновый Ёж 2026»	2
40 %	«Информация отсутствует»	0	«Малиновый Ёж 2026»	2
50 %	«Информация отсутствует»	0	«Информация отсутствует»	0
60 %	«Информация отсутствует»	0	«Малиновый Ёж 2026»	2
70 %	«Информация отсутствует»	0	«Малиновый Ёж 2026»	2
80 %	«Малиновый Ёж 2026»	2	«Малиновый Ёж 2026»	2
90 %	«Малиновый Ёж 2026»	2	«Малиновый Ёж 2026»	2

Полученные в ходе эксперимента данные позволяют детально проанализировать специфику работы механизмов внимания современных больших языковых моделей при обработке длинного контекста. Наблюдаемые паттерны поведения моделей Gemini 3.1 Flash и Gemini 3.1 Pro наглядно подтверждают эффект потери внимания в середине, однако характер деградации их ответов имеет принципиальные различия, обусловленные особенностями их архитектурного масштабирования.

Широкая слепая зона модели Gemini 3.1 Flash в диапазоне от 30% до 70% глубины текстового массива указывает на серьезные ограничения внутренних механизмов фильтрации контекста. Тот факт, что на пяти контрольных точках подряд модель продемонстрировала нулевую эффективность извлечения данных, объясняется следующими факторами:

– Оптимизация KV-кэша: Для обеспечения высокой скорости генерации и снижения стоимости запросов в легковесных моделях линейки Flash применяются методы квантования контекста. При обработке длинных документов нейросеть вынуждена сжимать промежуточные математические репрезентации текста, что приводит к необратимой потере фактов в срединной части документа.

– Разреженное внимание: Вместо полного вычисления матрицы внимания между всеми токенами, Flash-модели часто используют локальные или разреженные паттерны внимания. В результате фокус модели смещается на края контекстного окна, а центральная часть воспринимается как гомогенный шум.

Модель Gemini 3.1 Pro успешно справилась с извлечением информации на отметках 30%, 40%, 60% и 70%, где младшая модель потерпела неудачу. Это доказывает, что увеличение объема параметров и использование полноценных матриц сквозного внимания позволяют удерживать плотную семантическую связь на больших дистанциях.

Тем не менее, критическое падение точности до 0 баллов на отметке ровно 50% свидетельствует о математической уязвимости модели. В рамках обучения больших языковых моделей позиционные кодировки устроены таким образом, что начальные токены и финальные токены получают максимальный математический вес. На отметке 50% внимание достигает своего. Модель Gemini 3.1 Pro обладает достаточной мощностью, чтобы компенсировать это проседание на отметке 30% или 70%, но в абсолютном центре документа извлечение информации уже работает не так идеально.

Результаты эксперимента из файла позволяют сформировать график точности для обеих моделей (рис. 1). На основе проведенного сравнения можно сделать вывод, что простая максимизация контекстного окна (до миллиона токенов и более) является маркетинговым, а не чисто технологическим решением. Реальное окно эффективного внимания моделей гораздо уже заявленного физического объема.

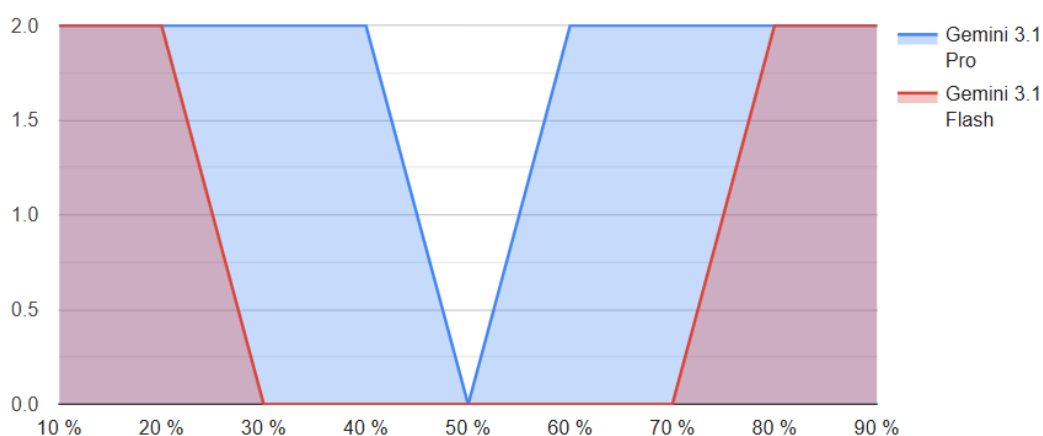


Рисунок 1. График точности моделей

С точки зрения практического применения в академической и корпоративной среде:

– Использование Gemini 3.1 Flash для анализа многостраничных сборников или отчетов сопряжено с высоким риском пропуска критически

важных данных, если они не продублированы во введении или заключении первоисточника.

– Модель Gemini 3.1 Pro демонстрирует высокую надежность как инструмент автоматизации аналитической работы, однако исследователь должен учитывать наличие точки уязвимости в самом центре обрабатываемого массива.

4 Вывод

В ходе проведенного исследования осуществлено тестирование устойчивости контекстуального внимания современных больших языковых моделей Gemini 3.1 Flash и Gemini 3.1 Pro при извлечении низкочастотных фактов из академического массива текстов на примере сборника научных трудов ПГУ им. Шолом-Алейхема.

Обе исследуемые модели продемонстрировали классическую зависимость точности от пространственного положения факта. Информация, расположенная на краях документа (первые и последние 10–20% объема), извлекается со стопроцентной точностью.

Флагманская модель Gemini 3.1 Pro показала значительно более высокую плотность внимания по сравнению с легковесной Gemini 3.1 Flash. В то время как Flash-модель потеряла внимание в широком диапазоне от 30% до 70% глубины текста, Pro-модель успешно преодолела эти участки.

Экспериментально доказано, что увеличение масштаба параметров модели не решает проблему слепой зоны полностью. На отметке ровно 50% обе нейросети продемонстрировали абсолютный отказ, пропустив целевой артефакт.

Библиографический список

1. Филатова О. Н., Булаева М. Н., Гуцин А. В. Применение нейросетей в профессиональном образовании // Проблемы современного педагогического образования. 2022. №. 77-3. С. 243-245. URL: <https://cyberleninka.ru/article/n/primenenie-neyrosetey-v-professionalnom-obrazovanii>
2. Никишкина Е. В., Ларин С. Э., Белаш В. Ю. Нейросети и образование: положительные и отрицательные стороны, возможности использования // Педагогический вестник. 2024. №. 32. С. 54-58. URL: <https://cyberleninka.ru/article/n/neyroseti-i-obrazovanie-polozhitelnye-i-otritsatelnye-storony-vozmozhnosti-ispolzovaniya>
3. Исаев А. В., Свищёв А. В. Актуальные тренды нейросетей в образовании // Международный журнал гуманитарных и естественных наук. 2024. №. 11-1 (98). С. 118-123. URL: <https://cyberleninka.ru/article/n/aktualnye-trendy-neyrosetey-v-obrazovanii>
4. Елисеев А. В., Шунина Л. А. Генеративные нейронные сети в образовании: классификация и некоторые особенности использования // Фундаментальные проблемы обучения математике, информатике и

- информатизации образования: Сборник тезисов докладов международной научной конференции, Елец, 29 сентября – 01 2023 года. Елец: Елецкий государственный университет им. И.А. Бунина, 2023. С. 193-197. URL: https://elibrary.ru/download/elibrary_54642021_61164376.pdf#page=193
5. Богатова С. М., Фрезе О. В. Дидактические возможности нейросетей в обучении иностранным языкам //Современное педагогическое образование. 2024. №. 3. С. 187-192. URL: <https://cyberleninka.ru/article/n/didakticheskie-vozmozhnosti-neyrosetey-v-obuchenii-inostrannym-yazykam>