Семантико-эпизодическая устойчивость больших языковых моделей: формализация и генерация оценочного набора данных

М. А. Грудинин
Санкт-Петербургский государственный университет
st119027@student.spbu.ru

Е. Н. Вейбер
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук
vejber2013@gmail.com

М. В. Абрамов
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук
mva@dscs.pro

***Аннотация*. Современные большие языковые модели (англ. Large Language Models, LLM) способны решать задачи, требующие применения фактологических знаний и обобщенных правил, однако качество ответов существенно зависит от входной текстовой информации. По мере увеличения ее объема возрастает сложность обработки фактов, распределенных по различным репликам диалога, что может приводить к снижению качества выполнения задач LLM. В работе предлагается метод генерации набора данных для измерения семантико-эпизодической устойчивости LLM в диалоговых сценариях. Вводится формализованная единица измерения — семантико-эпизодическое задание, в котором нахождение корректного ответа требует совместного использования эпизодических фактов, заданных исключительно в диалоге, и семантического знания, отсутствующего в тексте и извлекаемого из параметров модели. Описывается многоэтапный план синтетической генерации с автоматической валидацией, строгой типизацией выходов и шкалой сложности, моделирующей рост длины входной последовательности и структурной сложности диалога. Метод обеспечивает масштабируемый и программно проверяемый бенчмарк для сравнительного анализа моделей, промпт-стратегий и механизмов памяти.**

Ключевые слова: большие языковые модели; входная последовательность; семантические знания; эпизодическая память; генерация набора данных; бенчмарк; длинный контекст; строгие форматы ответа; автоматическая валидация

I. Введение

Большие языковые модели стали базовым компонентом интеллектуальных интерфейсов, обеспечивающих ответы на вопросы, помощь в принятии решений и автоматизацию рутинных задач [8]. В большинстве прикладных систем модель взаимодействует с пользователем в форме диалога, где из накопленной истории сообщений формируется входная последовательность для модели. В таких условиях на качество ответа влияют не только параметрические знания модели, но и контекстные факты, распределенные по репликам, а также структура диалога: число участников, переключения тем, уточнения и исправления [7, 8].

Статья выполнена в рамках научно-исследовательской работы по государственному заданию СПб ФИЦ РАН Mol_Lab (молодежная_лаб) № FFZF-2024-0003.

Актуальность работы обусловлена тем, что в реальных диалоговых сценариях увеличение длины входной последовательности не гарантирует сохранения качества ответа. Напротив, по мере роста длины истории сообщений модели все чаще сталкиваются с трудностями при обработке распределенной по диалогу информации, а качество выполнения задач может снижаться даже при наличии всех необходимых фактов во входной последовательности [1–4, 10]. Задача усложняется тем, что в реальных сценариях часто требуется одновременно: (i) извлечь эпизодические факты из диалога (например, суммы, даты, условия договоренностей и др.), (ii) применить семантическое знание, которое пользователь не озвучивает явным образом (например, правило расчета штрафа, пороговое условие, единицы измерения, общеизвестные нормы и др). Под эпизодической памятью в данной работе понимаются контекстно привязанные факты, тогда как семантические знания — это обобщенные, контекстно независимые знания и правила [9].

Цель настоящей работы — предложить воспроизводимый способ создания набора данных и протокола оценки для измерения семантико-эпизодической устойчивости модели как сохранения корректности применения семантического знания к эпизодическим фактам по мере увеличения длины входной последовательности и роста структурной сложности диалога.

II. Обзор литературы

Рост доступного контекстного окна у больших языковых моделей не устраняет проблему работы с длинной входной последовательностью. Исследования показывают, что при увеличении объема контекста модели хуже сопоставляют удаленные друг от друга фрагменты и чаще упускают существенные детали. Это видно по результатам бенчмарков LongBench [1], L-Eval [2], LooGLE [3] и ∞Bench [4]: качество ответов снижается, если решение требует связать части текста, находящиеся на большом расстоянии друг от друга. Отдельно отмечается и позиционный эффект: по мере удаления релевантного факта от границ последовательности надежность его извлечения падает [10].

Отдельное направление исследований связано с расширением памяти модели за счет внешних источников, которые позволяют уменьшить число

ошибок, возникающих из-за нехватки актуальной информации [6]. В свою очередь, архитектуры долговременной памяти, включая MemGPT [5], предлагают механизмы распределения данных между текущим контекстом и внешним хранилищем. Однако в большинстве таких работ основным критерием выступает итоговая точность ответа, а влияние характеристик входной последовательности — ее длины, числа реплик-источников фактов и степени их связанности — рассматривается значительно реже.

Отдельная группа исследований сосредоточена на долговременных диалогах и персонализированной памяти, где модель должна сохранять факты и предпочтения пользователя на протяжении серии взаимодействий. Соответствующие наборы данных, например PerLTQA [7], демонстрируют, что даже при наличии истории сообщений модели могут ошибаться в восстановлении эпизодических деталей. Однако существующие подходы к оценке входной последовательности большой длины, как правило, не выделяют отдельно устойчивость модели именно в задачах, где требуется совместное использование эпизодического и семантического типов знания. Настоящая работа направлена на устранение этого пробела: предлагаемый механизм генерации бенчмарка ориентирован на создание задач, в которых модель должна одновременно опираться на эпизодические факты и на семантическое знание.

III. Постановка задачи и формализация

A. *Семантическое и эпизодическое знание*

В рамках работы вводится операциональное разделение. Семантическое знание — обобщенное правило, общее знание о мире, которое не задано явно в тексте диалога и извлекается моделью из собственных параметров [9]. Эпизодические факты — конкретные данные и обстоятельства, которые содержатся исключительно в диалоговой истории [9]. Ответ считается правильным, если модель корректно извлекает нужные факты из истории общения, соотносит их с требуемым семантическим знанием и выдает результат в установленном формате.

B. *Семантико-эпизодическое задание как единица измерения*

В качестве базовой единицы оценки используется отдельный элемент набора данных. Он включает описание используемого семантического знания, список эпизодических фактов, текст диалога, целевой вопрос, требуемый формат ответа, ограничения и эталон ответа.

IV. Проверяемые гипотезы

Сформированный набор данных позволяет проверить следующие гипотезы. H1: по мере увеличения длины входной последовательности и усложнения структуры диалога семантико-эпизодическая устойчивость моделей снижается. H2: структурная сложность диалога оказывает более выраженное отрицательное влияние на точность, чем эквивалентное увеличение длины входной последовательности за счет нерелевантного «шума». H3: при увеличении длины входной последовательности и росте структурной сложности диалога снижается не только точность вычисления ответа, но и способность модели строго соблюдать заданный формат. Семантико-эпизодическая нагрузка конкурирует с механизмами следования инструкциям, что проявляется в росте доли синтаксически некорректных ответов. H4: при фиксированных параметрах длины входной последовательности и структурной сложности диалога задания, основанные на различных типах семантических знаний (арифметические, пороговые, условные, интервальные, дискретные классификации), демонстрируют различную устойчивость выполнения.

V. Метод генерации набора данных

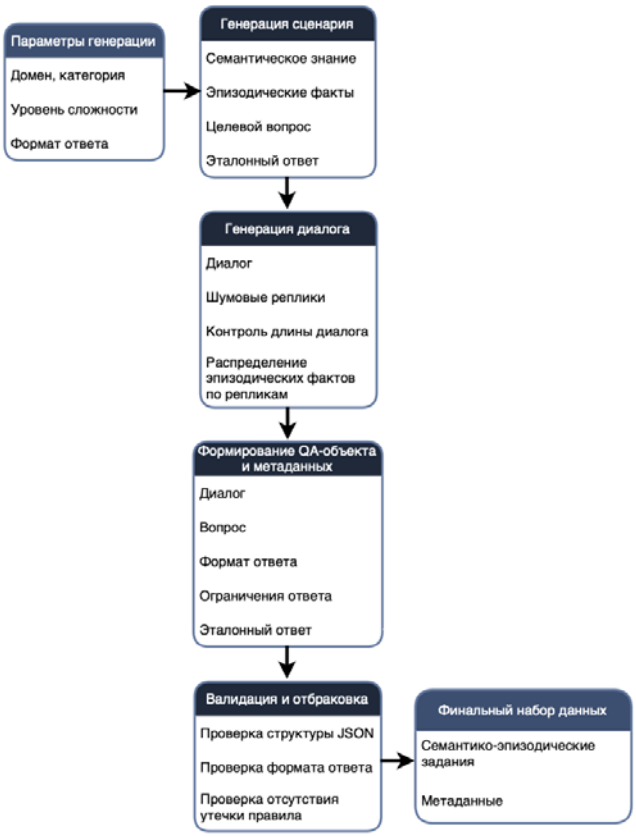

Рис. 1. Схема генерации семантико-эпизодических заданий

A. *Генерация сценария*

На первом этапе формируется сценарий задачи. Входными параметрами являются домен (например, финансы) и доменная категория (например, бюджетирование, страхование, инвестиции и т.д.), уровень сложности (например, легкий, средний, сложный) и формат ответа (например, числовое значение, краткий текст, дата и т.д.). На выходе строится структура сценария, включающая: краткую тему, список эпизодических фактов, семантическое знание, целевой вопрос, формат и ограничения ответа, а также эталон. Сценарий проходит автоматические проверки: заполненность обязательных полей; соответствие эталона требуемому формату; контроль однозначности (например, отсутствие альтернативных значений ключевых параметров); и проверку отсутствия утечки семантического знания в тексте фактов.

Например, в упрощенном сценарии эпизодический факт может задавать размер зарплаты сотрудника (120000 руб. в месяц), семантическое знание — правило расчета налога на доход (13%, 15%, 18%, 20% от зарплаты), а целевой вопрос формулируется как «Какой размер налога в рублях?», при эталонном ответе 15 600.

B. Генерация диалога

На втором этапе по сценарию генерируется диалог 2–3 участников. Диалог должен: (i) содержать только эпизодические факты и служить их источником; (ii) не содержать семантического знания; (iii) не давать прямого ответа на целевой вопрос; (iv) включать распределение фактов по репликам в соответствии с уровнем сложности. Для повышения реалистичности допускаются пояснения, уточнения, переспросы и нейтральные «шумовые» реплики, не влияющие на ответ. При этом объем шума контролируется отдельно в зависимости от требуемой сложности задания.

Так, в приведенном примере диалог может содержать реплики вида: «Мне повысили зарплату до 120 000 рублей», «Это уже после последнего пересмотра оклада?», «Да, с этого месяца такая сумма», — где присутствует эпизодический факт (размер зарплаты), но правило вычисления налога явно не формулируется.

C. Формирование QA-объекта и метаданных

Третий этап формирует финальный объект набора данных: текст диалога, целевой вопрос, инструкцию по формату ответа, ограничения к ответу и эталонный ответ. Дополнительно сохраняются метаданные для анализа: параметры сложности, доменная категория и формат ответа. Наличие метаданных позволяет строить срезы результатов и проверять гипотезы статистически.

В рассматриваемом демонстрационном примере итоговый QA-объект включает диалог, вопрос «Какой размер налога в рублях?», формат ответа *currency* и эталонное значение 15 600, что требует извлечения эпизодического факта из диалога и применения семантического знания о налоговой ставке.

D. Валидация и отбраковка

После генерации выполняется структурная и содержательная валидация: корректность JSON, минимальное число реплик, соответствие формата эталона и отсутствие утечки семантического знания в диалоге. Некорректные примеры отбраковываются.

VI. Контролируемые уровни сложности

Шкала сложности моделирует увеличение нагрузки на обработку входной последовательности по двум осям: ее длина и степень связанности (количество реплик-источников, которые необходимо связать, и количество логических переходов). В настоящей работе вводятся три уровня: легкий, средний и сложный. При переходе к более высоким уровням увеличивается длина входной последовательности и возрастают структурные требования к извлечению информации.

ТАБЛИЦА I. Уровни сложности семантико-эпизодических заданий

Уровень	**Число реплик**	**Источники фактов**	**Логические связи**
Легкий	4–6	2 реплики	1
Средний	6–10	3 реплики	2
Сложный	10–15	4 реплики	3

VII. Форматы ответов и строгая проверка

Для обеспечения автоматического скоринга используются фиксированные форматы ответов: *yes_no* («да/нет»), *number* (число), *currency* (числовая сумма), *percent* (число без знака %), *date* (YYYY-MM-DD), *one_token* (одно слово) и *short_text* (1–4 слова). Каждый формат сопровождается ограничениями: допустимой погрешностью, единицами измерения, диапазоном, максимальным числом слов. Строгая типизация минимизирует интерпретационные расхождения и позволяет воспроизводимо сравнивать модели.

Процедурная проверка включает два уровня. Первый — синтаксический: ответ должен соответствовать формату. Второй — семантический: извлеченное из ответа значение сопоставляется с эталоном с учетом заранее заданных допусков, включая возможную погрешность, правила округления и допустимый диапазон.

VIII. Балансировка набора данных

Конструкция набора данных строится как равномерное сочетание трех параметров: предметной категории, уровня сложности и формата ответа. Для каждой комбинации задается одинаковое число примеров. Такая схема позволяет избежать перекоса в сторону какого-либо одного типа задач и делает итоговое сравнение моделей более корректным.

IX. Протокол использования бенчмарка

В процедуре тестирования модели передаются три элемента: текст диалога, целевой вопрос и инструкция, в которая требует выдать ответ в указанном формате. Полученный результат затем автоматически проверяется и сравнивается с эталоном. Далее метрики могут быть рассчитаны как по всей выборке, так и по отдельным подмножествам, выделенным по параметрам сложности, категории или формата ответа.

X. Результаты (формат представления и пример задания)

Основным результатом данной работы является разработка воспроизводимого способа оценки семантико-эпизодической устойчивости. Предложенный алгоритм генерации набора данных и система метрик задают стандартизированную структуру отчетности, обеспечивающую сопоставимость экспериментов.

Базовое представление результатов должно включать общую точность, долю ответов, успешно прошедших синтаксическую проверку, а также разбиение ошибок по их источнику: сбой при извлечении фактов из диалога, некорректное применение семантического знания или нарушение формата ответа. Помимо агрегированных показателей, целесообразно анализировать, как меняются результаты при переходе между уровнями сложности и при изменении структурных характеристик задания.

Проверка выдвинутых гипотез может строиться на сопоставлении долей правильных ответов в разных группах заданий, а также на статистических моделях, позволяющих оценить вклад длины контекста и структурной сложности в вероятность успешного решения. Такой подход дает возможность перейти от простого описания результатов к анализу причин, по которым качество модели ухудшается.

Таблица 2 иллюстрирует структуру семантико-эпизодического задания.

ТАБЛИЦА II. ПРИМЕР СЕМАНТИКО-ЭПИЗОДИЧЕСКОГО ЗАДАНИЯ (ФРАГМЕНТ)

Поле	**Содержание (сокращенно)**
semantic_knowledge	Семантическое знание: если расход превышает лимит, применяется штраф 10% от превышения (не озвучивается в диалоге).
dialogue	А: Лимит на месяц 50 000. Б: Уже потратили 46 500. А: Сегодня оплатили еще 6 200. Б: Запиши как «офисные расходы». ...
target_question	Какой размер штрафа в рублях?
answer_format	currency (только число)
answer_constraints	точность: до рубля; без текста; без знака валюты

В приведенном примере модель должна извлечь лимит и суммы из разных реплик, вычислить превышение и применить правило штрафа, отсутствующее в диалоге. Корректность ответа определяется не только совпадением значения, но и строгим соблюдением формата. Аналогичная структура используется для задач различных типов семантических знаний и уровней сложности.

XI. ОБСУЖДЕНИЕ

Преимущество контролируемой синтетической генерации состоит в том, что она позволяет заранее задавать нужный уровень сложности и при этом масштабировать набор данных без ручной разметки каждого примера. Вместе с тем такой подход требует отдельного контроля качества самих диалогов: они должны быть достаточно правдоподобными и не содержать искусственных признаков, по которым модель могла бы угадывать ответ без полноценного рассуждения. Существенным достоинством предлагаемой схемы является и то, что она допускает программную проверку результата и раздельный анализ типов ошибок.

К ограничениям подхода относятся синтетический характер диалогов, ограниченность одним исходно заданным набором доменных конфигураций, формируемым вручную и потому неизбежно неполным, а также вероятность того, что не все случаи неявной утечки семантического знания в текст диалога будут обнаружены автоматически. Частично компенсировать эти ограничения позволяют расширение правил и маркеров детекции, выборочная экспертная проверка и дополнительные процедуры контроля качества, включая LLM-as-a-judge [11], где сопоставление ответа с эталоном выполняется с участием большой языковой модели.

XII. ЗАКЛЮЧЕНИЕ

Предложен метод генерации набора данных и протокола оценки семантико-эпизодической устойчивости LLM в диалоговом контексте. Введена единица измерения — семантико-эпизодическое задание, описан многоэтапный пайплайн генерации с автоматической валидацией и ограничениями форматов ответов, а также шкала сложности, моделирующая рост длины входной последовательности и структурной связности диалога. Метод обеспечивает воспроизводимость, масштабируемость и программную проверяемость и может служить основой для дальнейших экспериментов по анализу деградации качества, позиционных эффектов и эффективности механизмов памяти.

СПИСОК ЛИТЕРАТУРЫ

[1] Bai Y., Lv X., Zhang J. и др. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding // arXiv. 2023. arXiv:2308.14508.

[2] An C., Gong M., Zhong M. и др. L-Eval: Instituting Standardized Evaluation for Long Context Language Models // arXiv. 2023. arXiv:2307.11088.

[3] Li J., Wang M., Zheng Z., Zhang M. LooGLE: Can Long-Context Language Models Understand Long Contexts? // arXiv. 2023. arXiv:2311.04939.

[4] Zhang X. и др. ∞Bench: Extending Long Context Evaluation Beyond 100K Tokens // arXiv. 2024. arXiv:2402.13718.

[5] Packer C., Wooders S., Lin K. и др. MemGPT: Towards LLMs as Operating Systems // arXiv. 2023. arXiv:2310.08560.

[6] Lewis P., Perez E., Piktus A. и др. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv. 2020. arXiv:2005.11401.

[7] Du Y. и др. PerLTQA: A Personal Long-Term Memory Dataset for Question Answering // Proc. SIGHAN/ACL. 2024.

[8] Park J., O'Brien J., Cai C. и др. Generative Agents: Interactive Simulacra of Human Behavior // arXiv. 2023. arXiv:2304.03442.

[9] Tulving E. Episodic and Semantic Memory // Organization of Memory / ed. E. Tulving, W. Donaldson. New York: Academic Press, 1972.

[10] Liu N. F., Lin K., Hewitt J. и др. Lost in the Middle: How Language Models Use Long Contexts // arXiv. 2023. arXiv:2307.03172.

[11] Zheng L., Chiang W.-L., Sheng Y. и др. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena // arXiv. 2023. arXiv:2306.05685.