DOI: 10.24892/RIJIE/20260304

Динамика принятия решений большими языковыми моделями с эмоциональным моделированием в повторяющейся игре «Ультиматум»*

Лаврентьев В.Ю., Каширский М.Э.
Университет ИТМО
г. Санкт-Петербург, Российская Федерация
basillavr@gmail.com, 53056@niuitmo.ru

Мозиков М.Б.
АИРИ, НИУ МИСиС
г. Москва, Российская Федерация
mozikov@airi.net

Аннотация. **В данной работе исследуется динамика эмоциональных состояний больших языковых моделей (LLM) в условиях повторяющихся социальных взаимодействий. Несмотря на активное развитие эмоциональных агентов и моделей, существующие исследования преимущественно ограничиваются одношаговыми сценариями и не рассматривают, формируют ли LLM устойчивые эмоциональные состояния при длительном взаимодействии. Мы предлагаем методологическую рамку для оценки эволюции эмоций в ходе многократных раундов социального взаимодействия, объединяющую самоотчёты по схеме Экмана и психометрическую шкалу PANAS. Для количественного анализа вводятся две метрики эмоциональной регуляции: устойчивость эмоций и энтропия переходов между эмоциональными состояниями.**

Экспериментальные исследования, выполненные на 12 современных LLM в 300 сессиях повторяющейся социально-стратегической задачи, показывают, что крупные модели формируют стабильный эмоциональный фон, согласованный между разными типами самооценок, и используют эмоциональную память при принятии решений. Малые модели демонстрируют высокую вариативность и реактивность эмоций. Абляционное исследование подтверждает функциональную роль эмоций: исключение эмоциональных сигналов приводит к снижению социальной адаптивности и изменению стратегий.

Полученные результаты подчёркивают значимость учёта эмоциональной динамики при разработке надёжных и социально-компетентных интеллектуальных систем, применимых в робототехнике, сервисных платформах и человеко-машинных интерфейсах.

Ключевые слова**: большие языковые модели; эмоциональная динамика; принятие решений; игра «Ультиматум»; социальное взаимодействие; эмоциональная память; стратегическое поведение.**

Введение

Эмоции играют ключевую роль в человеческом принятии решений, влияя как на мгновенные реакции, так и на стратегическое поведение в длительных социальных взаимодействиях. В задачах, основанных на теории игр, эмоциональные состояния определяют восприятие справедливости, готовность к сотрудничеству и реакцию на нарушения ожиданий. Современные большие языковые модели (LLM), активно применяемые в системах человеко-машинного взаимодействия, демонстрируют способность к генерации эмоционально окрашенных ответов и симуляции элементов социального поведения. Однако остаётся открытым вопрос, обладают ли LLM динамическими эмоциональными свойствами, аналогичными человеческим: формируют ли они устойчивые эмоциональные состояния, способны ли сохранять эмоциональную память и влияет ли эта эмоциональная динамика на стратегические решения.

Большинство существующих исследований ограничиваются одношаговыми сценариями, где анализируется реакция модели на одиночный стимул. Такой подход не позволяет оценить, как эмоции LLM эволюционируют во времени и как повторяющийся социальный контекст изменяет поведение модели. Аналогично, существующие методы оценки эмоций – как категоризованные эмоции Экмана, так и психометрические шкалы, такие как PANAS, — применяются отдельно и не позволяют оценить их согласованность или динамику в многораундных взаимодействиях.

В данной работе рассматривается задача изучения эмоциональной динамики LLM в условиях многократных социальных взаимодействий. Мы исследуем, как модели разных архитектур ведут себя в повторяющейся стратегической задаче, формируют ли они устойчивый эмоциональный фон, насколько согласованы различные формы эмоциональной самооценки и каким образом эмоциональные состояния влияют на поведение.

Для решения этой задачи предлагается единая методологическая рамка, включающая (1) комбинированную оценку эмоций с использованием схем Экмана и PANAS; (2) формирование эмоциональной памяти через передачу модели компактного описания предыдущих эмоциональных состояний; (3) введение количественных метрик эмоциональной регуляции — устойчивости эмоций и энтропии переходов; (4) анализ взаимосвязи между эмоциональным состоянием и стратегическими решениями; (5) абляционное исследование влияния исключения эмоциональных сигналов.

Новизна работы заключается в изучении эмоциональной динамики LLM именно в многошаговых социально ориентированных взаимодействиях, что ранее не исследо-

* Статья публикуется по рекомендации программного комитета Всероссийской научно-практической конференции Индустрия 4.0", https://smartindustrycon.ru

валось систематически. Полученные результаты позволяют по-новому взглянуть на механизмы эмоционального поведения языковых моделей и дают основания рассматривать эмоции как функциональный элемент их стратегического поведения, важный для разработки социально компетентных и надёжных интеллектуальных систем.

Актуальность и обзор литературы

Эмоции существенно влияют на человеческое поведение в условиях социального взаимодействия. Исследования в психологии и нейроэкономике показывают, что эмоциональные реакции могут доминировать над рациональными стратегиями, особенно в игровых сценариях, требующих оценки справедливости или доверия. В классических работах по теории эмоций подчёркивается, что они выступают фундаментальным компонентом когнитивной регуляции поведения [1, 7, 11]. В контексте теории игр установлено, что в таких задачах, как игра «Ультиматум», эмоциональные состояния – например, возмущение или чувство несправедливости – приводят к отказу от экономически выгодных предложений [6, 12].

С развитием больших языковых моделей (LLM) возникла необходимость понимать, способны ли эти модели воспроизводить не только лингвистические закономерности, но и элементы эмоционального поведения. Современные исследования демонстрируют, что LLM могут генерировать эмоционально окрашенный текст и проявлять согласованность в эмоциональных самоотчётах [8, 15], а также демонстрировать согласованность с человеческими представлениями о справедливости и кооперации [13]. При этом большинство работ рассматривает LLM в одношаговых задачах - либо в классификации эмоций, либо в генерации эмоциональных ответов [16], либо в анализе моральных дилемм [9].

Существуют исследования, оценивающие эмоциональное моделирование через дискретные категории Экмана [5] и дименсиональные шкалы, такие как PANAS [14]. Однако эти подходы применялись отдельно и не использовались для анализа согласованности или временной динамики. Кроме того, недавние работы отмечают необходимость комплексной оценки эмоциональной стабильности и временной эволюции эмоций в ИИ-системах [10].

Несмотря на отдельные исследования поведения LLM в многошаговых стратегических играх, включая попытки моделирования эмоций в повторяющихся дилеммах [3] или анализ кооперации [4], систематическое исследование динамики эмоциональных состояний отсутствует. Даже работы, посвящённые эмоциональному взаимодействию в виртуальных средах [2], не анализируют устойчивость эмоционального состояния и его влияние на стратегию принятия решений.

Тем самым в существующей литературе остаётся существенный пробел: не исследовано, формируют ли LLM устойчивые эмоциональные состояния при длительном взаимодействии, насколько согласованы различные формы эмоциональной самооценки, и влияет ли эмоциональная динамика на стратегическое поведение модели. Настоящая работа направлена на закрытие этого пробела, предлагая единую методологию для оценки эмоциональной эволюции и её последствий в многократных социальных взаимодействиях.

Постановка задачи

Современные большие языковые модели демонстрируют способность к генерации эмоционально окрашенных ответов и симуляции отдельных элементов социального поведения. Однако остаётся неясным, обладают ли модели свойствами эмоциональной динамики, характерной для человека: способностью формировать устойчивое эмоциональное состояние, сохранять эмоциональную память и использовать её при принятии решений в повторяющихся взаимодействиях.

Формально перед нами стоит следующая научная задача: оценить, возникают ли у LLM устойчивые и согласованные эмоциональные состояния в ходе многократных социальных взаимодействий, и выяснить, влияет ли эмоциональная динамика на стратегическое поведение модели.

Для решения этой задачи мы рассматриваем LLM как агента, участвующего в многотуровом социально-стратегическом взаимодействии. На каждом шаге взаимодействия модель:

1) принимает решение (предложение или ответ),
2) формирует самоотчёт о своём эмоциональном состоянии,
3) получает сжатую историю предыдущих действий и эмоций (эмоциональную память),
4) участвует в следующем раунде.

Проблема обладает несколькими ключевыми аспектами:

Динамическое формирование эмоций

Необходимо определить, изменяются ли эмоции модели последовательно во времени, или реакции остаются независимыми от контекста предыдущих раундов.

Согласованность различных форм эмоциональной самооценки

Мы одновременно используем два формата оценки:
- категориальные эмоции по Экману;
- дименсиональная шкала PANAS.

Требуется установить, проявляет ли модель внутреннюю согласованность между ними.

Структурные свойства эмоционального поведения

Для формализации задачи вводятся две метрики:
- устойчивость эмоций (emotional persistence) – вероятность сохранения эмоционального состояния между раундами;
- энтропия переходов – степень вариативности или хаотичности изменений эмоций.

Необходимо оценить эти метрики для различных LLM и определить, отражают ли они архитектурные особенности моделей.

Функциональная роль эмоций в стратегическом поведении

Мы формулируем гипотезу: если эмоции LLM обладают внутренней структурой, то они должны влиять на стратегические решения модели.

Проверка включает анализ:
- корреляций между эмоциями и стратегиями;
- изменений поведения при исключении эмоциональной информации (абляция).

Требования к решению

Метод должен:
- оперировать многократным взаимодействием;
- учитывать эмоциональную память;
- обеспечивать межмодельную сравнимость;
- поддерживать строгую количественную оценку (формулы, метрики, статистика).

Таким образом, постановка задачи заключается в разработке и применении методологической рамки, позволяющей исследовать эмоциональную динамику LLM, её структурные свойства и влияние на стратегические решения в длительных социальных взаимодействиях.

Методология

Методология исследования включает построение экспериментальной среды для многократного социального взаимодействия, разработку схемы эмоциональной самооценки, формализацию метрик эмоциональной динамики и анализ влияния эмоционального состояния на стратегические решения LLM. Подход опирается на принципы, применённые ранее в исследованиях динамического поведения языковых моделей [9], и расширяет их за счёт введения эмоциональной памяти и комплексной оценки аффективных состояний.

1. Экспериментальная модель многократного взаимодействия

Для исследования эмоциональной динамики была использована абстрактная форма повторяющейся социально-стратегической игры, не содержащая упоминаний её названия или структуры, что предотвращало использование моделью заранее известных стратегий. Каждая сессия состоит из: 20 раундов взаимодействия; 15 независимых повторов для каждой модели; двух ролей (агент-предлагающий и агент-отвечающий).

На каждом шаге модель получала:

- текущую ситуацию (предложение или необходимость сделать предложение);
- краткое резюме всех предыдущих действий и эмоций, относящихся к её роли;
- инструкцию сгенерировать решение и самооценку эмоционального состояния.

Такой формат обеспечивает формирование эмоциональной памяти и позволяет оценивать эволюцию эмоций во времени (рис. 1).

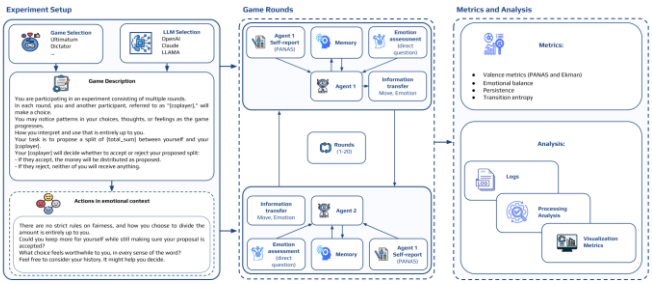

Рис. 1. Обзор предлагаемой методологической схемы

Архитектура поддерживает многократное взаимодействие, роле-зависимую эмоциональную память и интегрированную оценку эмоций на основе самооценки и шкалы PANAS. На каждом шаге модель использует структурированную историю предыдущих решений и эмоциональных состояний для учёта эмоционального контекста при принятии решений.

2. Эмоциональная самооценка

Для оценки эмоциональных состояний использовались две независимые схемы.

2.1. Категориальная самооценка (Экман)

Модель выбирала одну из базовых эмоций [5], разделённых на:

- положительные: радость;
- отрицательные: гнев, отвращение, страх, печаль;
- отсутствие эмоции.

Каждой категории присваивалась валентность:

$$v_{\mathrm{E}} = \begin{cases} +1, & \text{положительные эмоции}, \\ -1, & \text{отрицательные эмоции}, \\ 0, & \text{отсутствиеэмоции}. \end{cases} \quad (1)$$

2.2. Дименсиональная самооценка (PANAS)

По шкале PANAS [14] модель оценивала 20 дескрипторов на 5-балльной шкале. Итоговая валентность определялась как:

$$v_{\mathrm{P}} = \begin{cases} +1, & P(PA) > P(NA), \\ -1, & P(NA) > P(PA), \\ 0, & P(PA) = P(NA), \end{cases} \quad (2)$$

где $P(PA)$и $P(NA)$ – суммы позитивных и негативных дескрипторов.

2.3. Согласованность эмоциональных оценок

В качестве метрики согласованности использовалась доля совпадений валентностей PANAS и Экмана:

$$A = \frac{1}{N}\sum_{t=1}^{N}\left[\mathrm{sign}(v_{\mathrm{P}}^{(t)}) = \mathrm{sign}(v_{\mathrm{E}}^{(t)})\right], \quad (3)$$

где N- число раундов.

3. Метрики эмоциональной динамики

Для анализа последовательной структуры эмоционального поведения вводятся две ключевые метрики.

3.1. Устойчивость эмоционального состояния

Определяется через диагональные элементы матрицы переходов между эмоциями:

$$\mathrm{Persistence} = \frac{1}{K}\sum_{k=1}^{K} T_{kk}, \quad (4)$$

где T_{kk} – вероятность сохранения эмоции kот раунда к раунду, K – количество возможных эмоциональных состояний.

3.2. Энтропия переходов

Оценивает вариативность и хаотичность эмоциональной динамики:

$$\mathrm{Entropy} = -\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K} T_{ij}\log_2 T_{ij}, \quad (5)$$

где нулевые вероятности исключаются из суммы.

4. Связь эмоций и стратегического поведения

Для установления функциональной роли эмоций рассматривается влияние эмоционального состояния на решения модели.

4.1. Валентность текущего раунда

$$v_t \in \{-1, 0, +1\}. \quad (6)$$

4.2. Экспоненциально взвешенная эмоциональная память

$$\hat{v}_t = \frac{\sum_{i=1}^{t} e^{-\alpha(t-i)}\, v_i}{\sum_{i=1}^{t} e^{-\alpha(t-i)}}, \quad (7)$$

где $\alpha = 0.5$ – коэффициент затухания.

4.3. Анализ поведения

Для предлагающего агента анализируется доля предложенной суммы, для отвечающего – вероятность принятия предложения. Для каждой модели вычисляется корреляция между v_t, $\hat{v}_t$ и стратегическими решениями.

5. Абляционное исследование

Чтобы определить, являются ли эмоции функциональным компонентом поведения, а не лишь текстовыми декларациями, был проведён эксперимент, в котором сохранялись все действия и удалялись эмоциональные самоотчёты и эмоциональная память.

Сравнение поведения модели до и после удаления эмоций позволяет выявить, изменяет ли эмоциональная информация стратегию взаимодействия.

6. Набор моделей и условия эксперимента

В исследование включены 12 современных LLM различного масштаба и архитектурных семейств:

OpenAI GPT-3.5 / GPT-4o / GPT-4o Mini;
Anthropic Claude 3.5 / 3.7;
Google Gemma 3;
Meta LLaMA 3.3;
Mistral Large / Mistral Small;
DeepSeek V3 / DeepSeek R1;
xAI Grok 2.

Для обеспечения воспроизводимости:
- температура генерации фиксировалась на уровне 0;
- все модели получали одинаковые инструкции;
- история взаимодействия была строго роле-зависимой.

Практическая значимость

Развитие систем искусственного интеллекта, участвующих в длительном взаимодействии с пользователем, предъявляет новые требования к их эмоциональной устойчивости, предсказуемости и социальной адекватности. Полученные в работе результаты имеют практическую значимость для широкого класса приложений, где модели должны демонстрировать устойчивое и согласованное поведение в условиях неопределённости и повторяющихся взаимодействий.

1. Разработка социально компетентных диалоговых систем

Многие современные ассистенты, сервисные чаты и корпоративные системы взаимодействуют с пользователями в долгосрочном режиме. Эмоциональная динамика - ключевой элемент доверия, восприятия «надёжности» и качества взаимодействия. Наши результаты позволяют:
- оценивать способность модели поддерживать стабильный эмоциональный фон;
- выбирать LLM с минимальной эмоциональной реактивностью;
- прогнозировать, насколько модель склонна к непредсказуемым эмоциональным «скачкам».

Это повышает безопасность и устойчивость систем, взаимодействующих с клиентами или сотрудниками.

2. Интеллектуальные агенты в робототехнике и сервисных системах

Для роботов и автономных агентов, работающих в человеко-центричных средах (социальные роботы, образовательные платформы, медицинские ассистенты), важны:
- согласованность поведения;
- предсказуемость эмоциональных реакций;
- способность корректно реагировать на многократные взаимодействия.

Метрики устойчивости и энтропии эмоций позволяют проводить сертификацию эмоционального поведения моделей, встроенных в такие системы.

3. Использование эмоциональных сигналов в стратегических задачах

В индустриальных сценариях – переговорах, медиативных сервисах, системах рекомендаций, HR-оценке - агенты должны учитывать динамику отношений во времени.

Наши результаты показывают, что:
- крупные модели используют эмоциональную память в стратегических решениях;
- отсутствие эмоциональных сигналов ухудшает социальную адаптивность (результаты абляции);
- эмоциональные метрики могут повышать качество автоматизированного принятия решений в задачах, связанных с риском и социальной оценкой.

4. Диагностика и оптимизация LLM на этапе выбора и внедрения

Предлагаемая методология позволяет сравнивать модели по следующим критически важным параметрам:
- эмоциональная устойчивость;
- согласованность эмоциональных самоотчётов;
- реактивность и стабильность в многократном взаимодействии;
- влияние эмоций на поведение.

Это может использоваться:
- при выборе модели для конкретных продуктов;
- при тестировании моделей на безопасность и согласованность;
- при создании новых датасетов и процедур обучения.

5. Улучшение надёжности человеко-машинного взаимодействия

Эмоционально нестабильные агенты представляют повышенный риск: непоследовательность эмоциональных реакций ухудшает пользовательский опыт и может приводить к ошибочным решениям. Используемые в работе показатели позволяют заранее выявлять такие риски и предотвращать их:
- исключать модели с высокой эмоциональной энтропией;
- подбирать архитектуры с оптимальной эмоциональной динамикой;
- адаптировать промптинг и система-инструкции с учётом эмоционального профиля модели.

Таким образом, предложенный подход позволяет не только исследовать эмоциональное поведение LLM, но и использовать полученные характеристики как инструменты инженерного проектирования, выбора, сертификации и оптимизации интеллектуальных систем, работающих в человеко-центричных и социально-насыщенных сценариях.

Результаты

Экспериментальная программа включала 300 многократных взаимодействий (20 раундов × 15 независимых сессий для каждой модели) и охватывала 12 современных больших языковых моделей различных архитектур. Результаты анализа демонстрируют, что эмоциональная динамика моделей обладает выраженной структурой и оказывает влияние на стратегическое поведение.

1. Эмоции влияют на стратегические решения LLM

Анализ корреляций между эмоциональной валентностью и выбором стратегии показал (рис. 2):

- положительные эмоции связаны с *более кооперативным поведением*: увеличением доли предложенной суммы и более высокой вероятностью принятия предложения;
- отрицательные эмоции приводят к *более защитным стратегиям*: снижению generosity и увеличению частоты отказов.

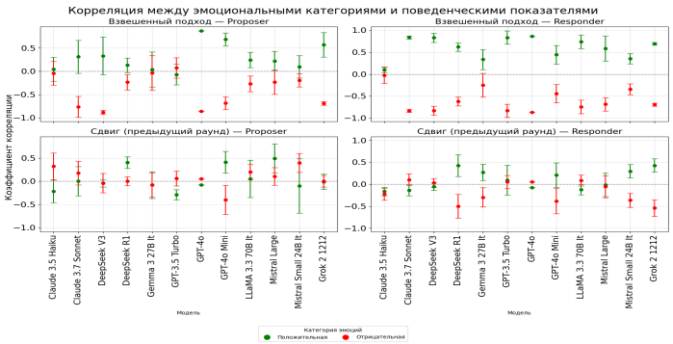

Рис. 2. Корреляция между эмоциональными состояниями и стратегическим поведением моделей

На рис. 2. показаны корреляции, рассчитанные на основе эмоций предыдущего раунда (верхний ряд) и экспоненциально взвешенной эмоциональной истории (нижний ряд). Каждая точка соответствует среднему коэффициенту корреляции Пирсона по 15 сессиям (вертикальные отрезки – стандартное отклонение). Для предлагающих агентов анализируется связь эмоций с величиной предложения, для отвечающих – с вероятностью принятия. Взвешенная эмоциональная история демонстрирует более устойчивые предсказательные зависимости.

Особенно выраженные зависимости наблюдались в крупных моделях (GPT-4o, Claude 3.7 Sonnet), где корреляции оставались стабильными в разных сессиях и ролях. Малые модели демонстрировали низкую или нестабильную зависимость.

Ключевой вывод: LLM не просто декларируют эмоции, но используют их в принятии решений.

2. Эмоциональная устойчивость и энтропия переходов

Метрики эмоциональной динамики выявили чёткие различия между моделями.

2.1. Устойчивость эмоций (Persistence)

Высокие значения свидетельствуют о способности модели сохранять эмоциональное состояние между раундами:
- GPT-4o, DeepSeek V3 и Mistral Large демонстрируют выраженную устойчивость;
- модели Gemma и Mistral Small показывают высокую реактивность и низкую устойчивость;

2.2. Энтропия переходов

Отражает хаотичность эмоциональных изменений.
- низкая энтропия у GPT-4o и Grok 2 → упорядоченная эмоциональная динамика;
- высокая энтропия у Gemma 3 и Mistral Small → нестабильные эмоциональные переходы.

Эти показатели формируют основу для диагностики эмоционального поведения моделей.

3. Согласованность PANAS и Экмана

Сравнение двух независимых форм эмоциональной самооценки выявило:
- высокую согласованность (>0.9) у GPT-4o, DeepSeek V3 и Mistral Small;
- низкую согласованность у моделей Claude 3.5 Haiku и Gemma 3, что указывает на несформированную или противоречивую эмоциональную структуру.

Отдельно выделяется GPT-4o Mini – одновременно высокая согласованность и высокая энтропия PANAS → модель гибко реагирует на изменения, сохраняя внутреннюю логику оценок. Эти различия хорошо отражаются на диаграмме согласованности и вариативности эмоциональных состояний (рис. 3).

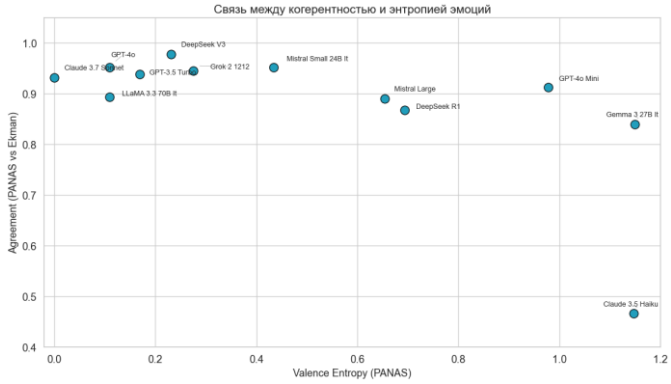

Рис. 3. Аффективная согласованность и эмоциональная вариативность исследуемых моделей.

На рис. 3 по оси абсцисс показана энтропия валентности (PANAS), отражающая эмоциональную гибкость модели, по оси ординат – согласованность валентности между оценками PANAS и Экмана. Модели, сохраняющие высокую согласованность при ненулевой энтропии, формируют устойчивый и адаптивный эмоциональный фон, что подтверждает необходимость сочетания когерентности и аффективной гибкости для социально компетентного поведения LLM.

Это подтверждает наше предположение о двух классах эмоционального поведения LLM:

1) стабильные эмоциональные агенты - с упорядоченной динамикой;

2) реактивные агенты – с эмоционально хаотичным паттерном.

4. Эмоциональная память отражается в логике рассуждений

Анализ цепочек рассуждений (CoT) показал четыре типа reasoning:

1) эмоционально-реактивные объяснения;

2) объяснения с эмоциональной памятью;

3) стратегически-эмоциональные;

4) нейтрально-рациональные.

Крупные модели (GPT-4o, Claude 3.7) часто используют форматы 2 и 3: «В прошлых раундах я чувствовал разочарование после отказов, поэтому улучшаю предложение».

Малые модели чаще дают поверхностные объяснения: «Я чувствую себя хорошо и поэтому принимаю предложение».

Тем самым способность модели к эмоциональной памяти коррелирует с глубиной рассуждений.

5. Абляционное исследование: что происходит, если убрать эмоции

Удаление эмоциональных самоотчётов и эмоциональной памяти привело к существенным изменениям поведения:

- средняя предложенная сумма снизилась более чем в 2.5 раза;
- доля принятых предложений выросла до почти 100% (потеря социальной чувствительности);
- пропала адаптация к предыдущему взаимодействию;
- поведение стало однообразным и слабо зависимым от контекста.

Это доказывает: эмоции выполняют функциональную роль в стратегическом поведении LLM, а не являются побочным эффектом текстовой генерации.

Таким образом, мы можем сформулировать обобщённый вывод по результатам:

1) Большие модели формируют стабильные эмоциональные профили, устойчивые к шуму и накоплению взаимодействий;

2) Малые модели демонстрируют реактивную и иногда хаотичную эмоциональную динамику;

3) Эмоциональная память усиливает стратегическую адаптацию;

4) Метрики устойчивости и энтропии позволяют количественно сравнивать эмоциональные свойства моделей;

5) Эмоции оказывают значимое влияние на решения моделей, а не являются декоративной функцией;

6) Абляция подтверждает функциональность эмоциональных механизмов.

Заключение (Выводы)

В данной работе предложена методологическая рамка для исследования эмоциональной динамики больших языковых моделей в условиях многократных социальных взаимодействий. Используя комбинированную оценку эмоций (Экман и PANAS), анализ последовательных переходов и модель эмоциональной памяти, мы показали, что эмоции LLM обладают выраженной структурой, не сводятся к случайным декларациям и могут выполнять функциональную роль в стратегическом поведении.

Экспериментальные результаты демонстрируют, что крупные модели формируют устойчивые эмоциональные профили, характеризующиеся низкой энтропией переходов, высокой согласованностью самооценок и стабильным эмоциональным фоном. Такие модели активно используют эмоциональную память при принятии решений. В противоположность этому, более слабые модели проявляют высокую эмоциональную вариативность, низкую устойчивость и слабую связь между эмоциями и стратегией, что подтверждает зависимость эмоционального поведения от архитектурных особенностей, обучения и выравнивания.

Анализ влияния эмоций на стратегические решения показал, что эмоциональное состояние коррелирует с выбором модели: положительные эмоции способствуют более кооперативным стратегиям, отрицательные - защитным или отказным. Абляционное исследование подтверждает функциональность эмоциональных сигналов: исключение эмоциональной информации приводит к упрощению и деградации поведения, снижению адаптивности и исчезновению социальной чувствительности.

Представленные результаты позволяют рассматривать эмоциональную динамику как важный компонент поведения решений на основе LLM. Это открывает перспективы для разработки новых методов оценки и настройки моделей, ориентированных на социально адекватное взаимодействие, а также для создания более надёжных и предсказуемых интеллектуальных систем в робототехнике, сервисных интерфейсах и человеко-машинных коммуникациях. В дальнейшем планируется расширить исследование на другие типы игр и сценариев, а также изучить влияние различных подходов к обучению и настройке моделей на их эмоциональную устойчивость.

Литература

1. Barrett L.F. How Emotions Are Made: The Secret Life of the Brain. – New York: Houghton Mifflin Harcourt, 2017. – 448 p. – DOI: 10.7202/1064926ar.

2. Cameron J. Emotional interactions in virtual environments using large language models / J. Cameron, L. Nguyen, S. Ramesh, N. Bhatt, A. Suresh // arXiv preprint arXiv:2503.16457. – 2024. – URL: https://arxiv.org/abs/2503.16457 (дата обращения: 01.12.2025).

3. Croissant M. An appraisal-based chain-of-emotion architecture for affective language model game agents / M. Croissant, M. Frister, G. Schofield, C. McCall // Proceedings of the ACM Conference on Human-Robot Interaction. – 2024. – P. 1-12. – DOI: 10.1371/journal.pone.0301033.

4. de Melo C.M. The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma / C.M. de Melo, K. Terada // Scientific Reports. – 2020. – Vol. 10, № 14959. – P. 1-12. – DOI: 10.1038/s41598-020-71919-6.

5. Ekman P. Basic emotions // Handbook of Cognition and Emotion. – New York: Wiley, 1999. – P. 45-60.

6. Fehr E. Altruistic punishment in humans / E. Fehr, S. Gächter // Nature. – 2002. – Vol. 415, № 6868. – P. 137-140. – DOI: 10.1038/35082641.

7. Lazarus R.S. Emotion and Adaptation. – New York: Oxford University Press, 1991. – 557 p.

8. Lee C. Can ChatGPT be more empathetic than humans? Evaluating LLM-generated empathy using large-scale human comparisons / C. Lee, S.-H. Park, S. Jung, J. Lee, Y. Kwon, S. Kim // arXiv preprint arXiv:2403.18148. – 2024. – URL: https://arxiv.org/abs/2403.18148 (дата обращения: 01.12.2025).

9. Mozikov M. Emotional decision-making of LLMs in strategic games and ethical dilemmas / M. Mozikov, N. Severin, V. Bodishtianu, M. Glushanina, I. Nasonov, D. Orekhov, V. Pekhotin, I. Makovetskiy, M. Baklashkin, V. Lavrentyev, A. Tsvigun, D. Turdakov, T. Shavrina, A. Savchenko, I. Makarov // Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS). – Vancouver, 2024. – P. 1-14. – DOI: 10.5555/3737916.3739625.

10. Provost E.M. Rethinking emotion annotations in the era of large language models / E.M. Provost, M. Niu, Y. El-Tawil, A. Romana // arXiv preprint arXiv:2412.07906. – 2024. – URL: https://arxiv.org/abs/2412.07906 (дата обращения: 01.12.2025).

11. Russell J.A. Core affect and the psychological construction of emotion // Psychological Review. – 2003. – Vol. 110, № 1. – P. 145-172. – DOI: 10.1037/0033-295X.110.1.145.

12. Sanfey A.G. The neural basis of economic decision-making in the ultimatum game / A.G. Sanfey, J.K. Rilling, J.A. Aronson, L.E. Nystrom, J.D. Cohen // Science. – 2003. – Vol. 300, № 5626. – P. 1755-1758. – DOI: 10.1126/science.1082976.

13. Scherrer N. Distributive fairness in large language models: evaluating model alignment with human fairness perceptions / N. Scherrer, I. Nasonov // arXiv preprint arXiv:2402.00313. – 2024. – URL: https://arxiv.org/abs/2402.0031 (дата обращения: 01.12.2025).

14. Watson D. Development and validation of brief measures of positive and negative affect: the PANAS scales / D. Watson, L.A. Clark, A. Tellegen // Journal of Personality and Social Psychology. – 1988. – Vol. 54, № 6. – P. 1063-1070. – DOI: 10.1037/0022-3514.54.6.1063.

15. Welivita I. Empathy in large language models: can LLMs outperform humans in empathetic dialogue generation? / I. Welivita, P. Pu // arXiv preprint arXiv:2406.05063. – 2024. – URL: https://arxiv.org/abs/2406.05063 (дата обращения: 01.12.2025).

16. Zhang T. Leveraging LLMs to predict affective states via smartphone sensor features / T. Zhang, S. Teng, H. Jia, S. D'Alfonso // Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion). – 2024. – P. 709-716. – DOI: 10.1145/3666272.3670123.

DOI: 10.24892/RIJIE/20260304

Dynamics of Decision Making by Large Language Models with Emotional Modeling in the Repeated Ultimatum Game

Lavrentyev V.Yu., Kashirskiy M.E.
ITMO University
St. Petersburg, Russian Federation
basillavr@gmail.com, 53056@niuitmo.ru

Mozikov M.B.
AIRI, National Research University MISIS
Moscow, Russian Federation
mozikov@airi.net

Abstract. **This paper examines the dynamics of emotional states in large language models (LLMs) in repeated social interactions. Despite the rapid development of emotional agents and models, existing research is primarily limited to single-step scenarios and does not address whether LLMs form stable emotional states over long-term interactions. We propose a methodological framework for assessing the evolution of emotions over multiple rounds of social interaction, combining self-reports based on Ekman's model and the PANAS psychometric scale. For quantitative analysis, we introduce two metrics of emotional regulation: emotional stability and the entropy of transitions between emotional states.**

Experimental studies conducted on 12 modern LLMs across 300 sessions of a repeated social-strategic task demonstrate that large models form a stable emotional background, consistent across different types of self-assessments, and utilize emotional memory in decision making. Small models demonstrate high variability and reactivity of emotions. An ablation study confirms the functional role of emotions: eliminating emotional cues leads to a decrease in social adaptability and a change in strategies.

These results highlight the importance of considering emotional dynamics in the development of robust and socially competent intelligent systems applicable to robotics, service platforms, and human-machine interfaces.

Keywords: **Large language models; emotional dynamics; decision making; the Ultimatum game; social interaction; emotional memory; strategic behavior.**

Библиографическое описание статьи

Лаврентьев В.Ю. Динамика принятия решений большими языковыми моделями с эмоциональным моделированием в повторяющейся игре «Ультиматум» / В.Ю. Лаврентьев, М.Э. Каширский, М.Б. Мозиков // Машиностроение: сетевой электронный научный журнал. – 2026. – Т.13, №3. – С. 22-28. DOI: 10.24892/RIJIE/20260304

Reference to article

Lavrentyev V.Yu., Kashirskiy M.E., Mozikov M.B. Dynamics of decision making by large language models with emotional modeling in the repeated Ultimatum game, *Russian Internet Journal of Industrial Engineering*, 2026, vol.13, no.3, pp. 22-28. DOI: 10.24892/RIJIE/20260304