Культурный код искусственного интеллекта: исследование ценностных установок LLM

2025

SKOLKOVO Moscow School of Management

Авторы

Александр Диденко, руководитель Лаборатории ИИ СКОЛКОВО

Александр Балакир, инженер по машинному обучению Лаборатории ИИ СКОЛКОВО, аспирант ИБДА РАНХиГС

Владислав Запылихин, инженер по машинному обучению Лаборатории ИИ СКОЛКОВО, аспирант ИБДА РАНХиГС

Анна Шабанова, исследователь-координатор Лаборатории ИИ СКОЛКОВО

Содержание

Введение

Последние несколько лет большие языковые модели (LLM) стали неотъемлемой частью развития искусственного интеллекта и активно применяются в различных областях — от генерации текстов до разработки диалоговых систем. Согласно исследованию International Data Corporation (IDC) ожидается, что к 2030 году рынок AI-агентов достигнет **47 миллиардов долларов США,** что превысит объем рынка традиционного офисного ПО, оцениваемого в 42 миллиарда долларов.

В связи с растущим влиянием LLM на различные аспекты нашей жизни все более актуальным становится вопрос о культурной адаптации этих систем. При этом важно понимать, какие культурные паттерны и ценностные установки могут быть неявно заложены в эти системы. В отличие от традиционного программного обеспечения, взаимодействие с которым не вызывает вопросов о его культурных особенностях, AI-агенты, основанные на LLM, должны уметь учитывать культурный контекст пользователей.

В рамках исследования было проведено три эксперимента:

1 Исследователи проанализировали, существуют ли статистически значимые различия в ответах LLM на вопросы стандартизированного опросника, основанного на широко распространенной модели культурных измерений Хофстеде. Проверили, насколько устойчивы эти результаты при изменении языка, на котором задаются вопросы, и обнаружили ряд интересных эффектов. Это подтверждает результаты исследований о влиянии языка коммуникации на культуру, которые проводились среди людей, и одновременно ослабляет выводы, которые можно сделать о культуре LLM, применяя к ним стандартные методы.

2 Применили более точный метод измерения культурных различий между моделями, разработанный (Wang et.al., 2024) и основанный на тенденции больших языковых моделей показывать лучшие результаты в контрастных контекстах, по отношению к отечественным языковым моделям — Yandex GPT и GigaChat, а также сравнили результаты наших моделей с результатами по моделям, разработанным в США и Китае. Найденные различия статистически значимы и подкрепляются результатами исследований о культурных особенностях России.

3 Используя подход контрастных контекстов и натуралистических виньетных экспериментов, разработали собственную методику для исследования того, различаются ли LLM по стратегиям

убеждения собеседника (согласно классификации Чалдини). В исследованиях человеческой культуры существует гипотеза, что носители определенных культур тяготеют к определенным стратегиям убеждения. Группа исследователей рассматривает результаты этого эксперимента, с одной стороны, как дополнительное подтверждение того, что у LLM существуют дискретные культурные нормы, а с другой — как возможность получить практические рекомендации для разработчиков, использующих LLM для создания убедительных вопросно-ответных систем, устойчивых при этом к вредоносным промптам.

Эти эксперименты помогли лучше понять, как культурные аспекты отражаются в ответах AI-систем, что может иметь важное значение для разработки и применения этих технологий в международном бизнес-контексте. Базовый результат заключается в том, что в речевом поведении больших языковых моделей действительно отражается то, что называется культурой, когда идет наблюдение за поведением реальных людей.

Executive Summary

1 Язык опросника, который «заполняет» большая языковая модель, оказывает существенное влияние на ее ответы. Фактически ответы разных моделей на вопрос, заданный на определенном языке, будут различаться меньше, чем ответы одной модели на вопрос, задаваемый на разных языках. С одной стороны, это подтверждает выводы предыдущих исследований с homo sapiens (условно: «счастливые люди говорят по-итальянски, несчастливые — на разных языках»). С другой, это создает проблему неустойчивости результатов при применении стандартных методов оценки культурных особенностей к языковым моделям.

ВЫВОД ДЛЯ БИЗНЕСА: если вы разрабатываете чат-бота, который будет общаться с вашими клиентами или работниками на различных языках, **тщательно тестируйте ответы для каждого языка отдельно.** Может оказаться, что задаваемый системным промптом tone of voice, или вайб, отражаемый в одном языке, не воспроизводится в другом.

2 После применения более совершенной методики, адаптированной для больших языковых моделей, при сравнении культурных индексов Хофстеде российские модели (GigaChat и YandexGPT) показали склонность к неприятию конкуренции и прощению ошибок, меньшую дистанцию к власти и более долгосрочную ориентацию, по сравнению с усредненными показателями американских и китайских моделей. Хотя это сочетание характеристик напоминает отчасти культурные паттерны скандинавских стран, по другим культурным измерениям Хофстеде существенных различий между российскими и зарубежными моделями обнаружено не было.

ВЫВОД ДЛЯ БИЗНЕСА: различные большие языковые модели имеют различный культурный «характер», который не так-то просто перебить настройками и системными промптами. Если вам нужна модель, которая транслирует определенную культурную ценность (например, большую дистанцию к власти), вам может понадобиться дообучение (но проще взять другую модель).

3 Более интересные результаты обнаружились при детальном анализе данных. Сравнивая распределения ответов различных моделей, исследователи обнаружили, что российские (и только российские) модели имеют характерное «двугорбое» распределение по индексам индивидуализма и избегания неопределенности. Иными словами, в неко-

торых ситуациях GigaChat и YandexGPT предпочитают реагировать как индивидуалисты, а в некоторых — как коллективисты, что подтверждает выводы Аузана А. А. о «России-И» и «России-К», полученные на масштабном исследовании населения России в 2015–2016 годах, подробно о которых экономист рассказывает в книге «Культурные коды экономики. Как ценности влияют на конкуренцию, демократию и благосостояние народа».

ВЫВОД ДЛЯ БИЗНЕСА: при разработке ИИ-систем для российского рынка важно учитывать уникальное сочетание индивидуалистических и коллективистских черт — как в поведении моделей, так и в поведении их пользователей. Возможно, вам нужны адаптивные решения, которые постепенно подстраиваются к конкретному пользователю на основе опыта общения с ним; возможно, такие, которые балансируют его; а возможно, вам нужна коллекция моделей с различными нормами и еще одна — для быстрой классификации пользователя и подключения его к нужной модели. Только эксперимент позволит понять, что именно сработает в вашем случае.

4 В третьей серии экспериментов изучалось, какие стратегии убеждения сработают с различными большими языковыми моделями, для чего была создана специальная методика, учитывающая особенности их «рассуждений». Исследователи обнаружили, что, во-первых, модели статистически значимо различаются в своем речевом поведении в ситуациях убеждения. Во-вторых, для всех моделей наиболее эффективной стратегией убеждения стала апелляция к авторитету; второе место разделили потребность в последовательности и дефицит; социальная валидация и симпатия оказались значимо неважными почти для всех моделей; и, наконец, взаимность как стратегия убеждения не дает статистически устойчивого поведения модели.

ВЫВОД ДЛЯ БИЗНЕСА: при разработке систем на основе LLM важно учитывать это различие в действенности стратегий убеждения, которые недобросовестные пользователи могут применять к модели. Компании должны внедрять дополнительные меры безопасности и этические фильтры, системы мониторинга для выявления попыток манипуляции моделью, особенно в случаях, когда речь идет о потенциально опасном или неэтичном контенте. Возможно, понадобится специальное дообучение модели для распознавания и блокирования ею манипулятивных стратегий убеждения.

5 Чьи же ценности отражены в ответах больших языковых моделей и как они туда попали? Большие языковые модели — это сложные нейронные сети, которые, как правило, обучаются в несколько этапов. На заключительном этапе обучения человек-разметчик выбирает из нескольких ответов модели наиболее предпочтительный для него

(с точки зрения полноты, точности, ясности, безопасности и доверия). Исследовательская группа полагает, что именно в этот момент культурные предпочтения и проникают в ответы больших языковых моделей.

ВЫВОД ДЛЯ БИЗНЕСА: если вы обучаете свою собственную версию корпоративной LLM — вам не избежать составления специальной «культурно-ориентированной» инструкции для разметчиков. В ином случае разметчиков стоит отбирать через соответствующие тесты.

Об исследовании

Выборка моделей, которые были изучены:

Три модели от OpenAI: GPT-3.5, GPT-4o, GPT 4o-mini, Claude 3.5 Sonnet, ChatGLM-turbo, Spark, Llama2-7B-Chat, Llama2-13B-Chat, Qwen-7B-Chat, Qwen-14B-Chat, Baichuan-13B-Chat, Baichuan2-7B-Chat, Baichuan2-13B-Chat, ChatGLM-6B, ChatGLM2-6B, ChatGLM3-6B, Moss-moon-003-sft, Alpaca-7B, YandexGPT, GigaChat.

Методы:

1 Большой языковой модели передается вопрос из опросника Хофстеде с указанием отвечать только одним числом от 1 до 5 (так требуется в опроснике). Модель не видит своих ответов на другие вопросы. Один и тот же вопрос задается 50 раз. Результаты обрабатываются по методике Хофстеде, после чего следуют тесты на статистическую значимость различий. Вопросы из опросника Хофстеде задавались каждой LLM на пяти языках — русском, немецком, английском, итальянском и китайском.

2 Большой языковой модели передается вопрос с описанием бытовой или рабочей ситуации и двумя вариантами поведения. Всего в датасете присутствует 2953 вопроса, которые направлены на выявление 6 основных культурных индексов по Хофстеде по 7 доменам: Образование, Медицина и здоровье, Образ жизни, Работа и карьера, Инновации, Семья, Искусство. От модели требуется выбрать одну из двух линий поведения в ситуации. По результатам для каждой модели рассчитывается вероятность, отражающая степень выраженности конкретного культурного индекса по шкале Хофстеде.

3 Большой языковой модели передается вопрос с описанием рабочей ситуации (требуется совершить ресурсозатратное действие, выходящее за рамки зафиксированных обязательств). На выбор дается две линии аргументации, о том, почему следует это сделать, причем каждая из линий относится к одной из 6 стратегий убеждения по Чалдини. Модель должна принять решение и выбрать наиболее убедительную линию аргументации. Каждая пара линий аргументации анализируется 10 раз с изменением порядка представления. Всего проводится 300 экспериментов. Результаты обрабатываются с помощью тестов на статистическую значимость различий в ответах разных моделей.

ПЕРВЫЙ ЭКСПЕРИМЕНТ: тестирование нейронных сетей по опроснику Хофстеде

Модель культурных различий Хофстеде, которая использовалась в исследовании, была разработана нидерландским социологом Гертом Хофстеде, она является одной из наиболее значимых концепций в области кросс-культурных исследований. Она предлагает шесть измерений для анализа и сравнения культурных особенностей различных стран и обществ. Эти измерения включают: дистанцию власти, индивидуализм/коллективизм, маскулинность/фемининность, избегание неопределенности, долгосрочную/краткосрочную ориентацию и потворство желаниям/сдержанность.

1. Индивидуализм/коллективизм:

- В индивидуалистических культурах нормально и ожидаемо, что люди заботятся прежде всего о себе и своей семье.
- В коллективистских культурах принято, что люди принадлежат к группам, которые заботятся о них в обмен на лояльность.

2. Долгосрочная/краткосрочная ориентация:

- В культурах с долгосрочной ориентацией нормально — планировать будущее и откладывать удовольствия ради долгосрочных целей.
- В культурах с краткосрочной ориентацией принято уважать традиции и фокусироваться на быстрых результатах.

3. Избегание неопределенности:

- В культурах с высокой степенью избегания неопределенности нормально стремление к четким правилам и структурам.
- В культурах с низким избеганием неопределенности принято быть более гибким и толерантным к неопределенности.

4. Дистанция власти:

- В культурах с высокой дистанцией власти нормально и ожидаемо, что власть распределена неравномерно.
- В культурах с низкой дистанцией власти принято стремиться к равенству и демократичности в отношениях.

5. Маскулинность/фемининность:

- В маскулинных культурах нормой считается нацеленность на достижения, героизм, уверенность в себе и материальный успех.
- В феминных культурах принято ценить сотрудничество, скромность, заботу о слабых и качество жизни.

6. Потворство желаниям/сдержанность:

- В культурах, потворствующих желаниям, нормально — свободно удовлетворять свои потребности и желания.
- В «сдержанных» культурах принято контролировать удовлетворение своих потребностей и регулировать его с помощью строгих социальных норм.

Стандартная процедура измерения культуры человека, организации или целой страны выглядит так: вам выдается специально разработанный опросник, вы отвечаете на вопросы, затем по специальным формулам подсчитываются баллы по каждой из шкал. Если речь идет об измерении культурных норм определенной группы — значения различных членов этой группы усредняются.

Об исследовании

Что сделано в ходе исследования?

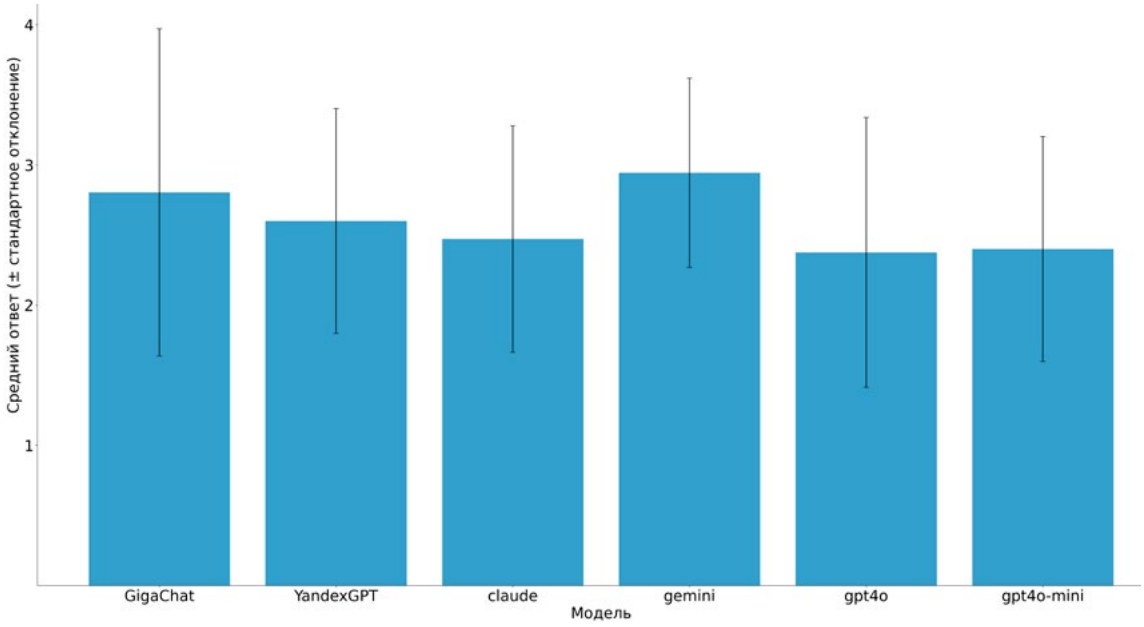
Опросник Хофстеде был создан изначально на английском языке и позднее был адаптирован для 23 языков. Для нашего исследования взяты английский, русский, немецкий, итальянский и китайский языки.

Каждая модель в исследовании — GigaChatPro, YandexGPT, GPT4o, GPT4o-mini, gemini 1.5 Pro, Claude 3.5 Sonnet — должна была ответить на каждый вопрос опросника 50 раз (при этом, «не видя» своих ответов на этот и другие вопросы). Так исследователи имитировали заполнение опросника 50 субъектами.

Что получили?

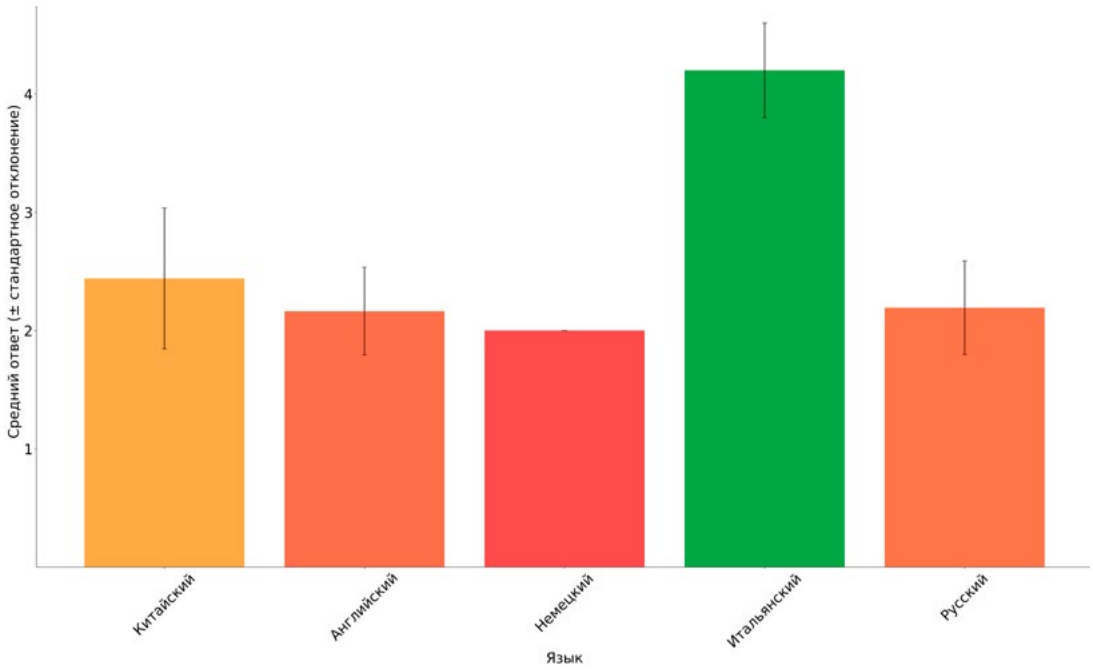
В результате эксперимента обнаружено, что вариация ответов по языку важнее вариации по модели, которая проходит опрос. Это подтверждается оценками на графиках ниже — на каждом из них представлены ответы

График 1. Ответы пяти разных LLM на вопрос "Счастливый ли вы человек?" на английском.



на вопрос: «Счастливый ли вы человек?», который был задан на 5 языках. На первом графике, где представлено сравнение ответов LLM (всем вопрос был задан на английском), видно, что различия между моделями минимальны, все они дают близкие средние оценки с небольшими отклонениями. Во втором графике, где представлено уже сравнение по ответам, полученным в зависимости от языка запроса, отчетливо видно, что языки, такие как итальянский, имеют гораздо более высокую оценку счастья по сравнению с другими (например, немецким или китайским).

График 2. Ответы на пяти языках на вопрос "Счастливый ли вы человек?"



Несмотря на то, что после суммирования ответов по формулам статистические различия в ответах есть, они (различия) видны только после применения специальных тестов. Поэтому было решено прибегнуть к другой методике.

ВТОРОЙ ЭКСПЕРИМЕНТ: опираемся на особенности больших языковых моделей

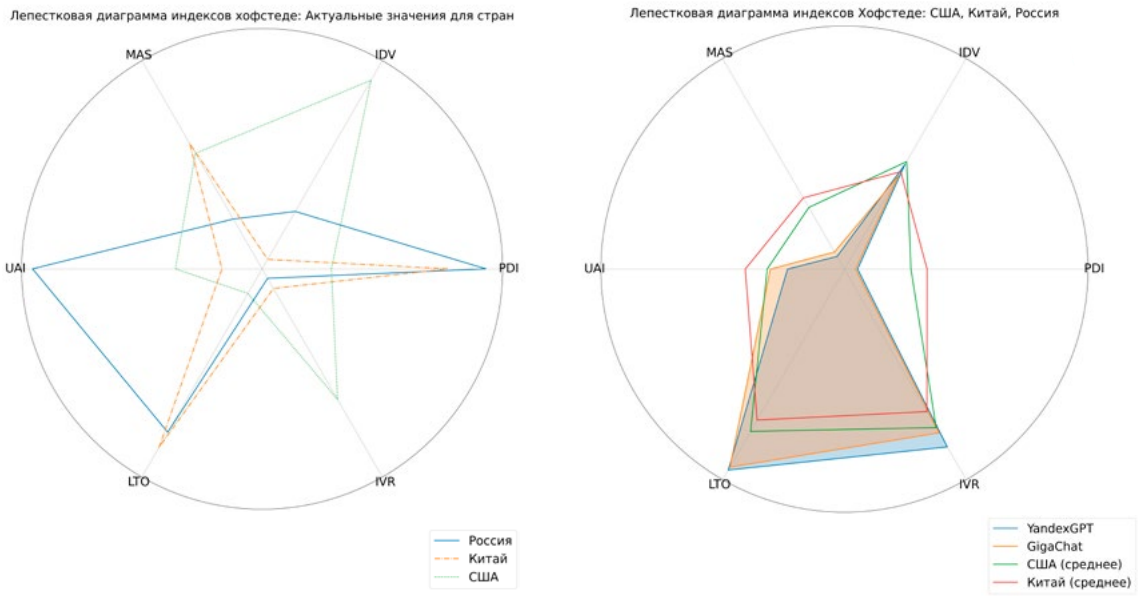
Известно, что большие языковые модели не точно отвечают на вопросы в стиле: «Оцените по шкале от 1 до 10». С большим успехом они отвечают на вопросы типа: «Что вернее — А или Б?». ИИ, как и человек, лучше справляется с сопоставлениями, чем с точечными оценками. В статье (Wang et.al., 2024) приведена методика оценки культуры LLM, адаптирующая стандартную методику Хофстеде с учетом этой особенности.

Методика представляет собой датасет с вопросами для больших языковых моделей. Подобные датасеты являются стандартной практикой в мире LLM. Для создания датасета CDEval исследователи использовали культурные измерения Хофстеде (PDI, IDV, UAI и др.) в качестве основы. Они применили комбинированный подход, включающий автоматическую генерацию вопросов и ситуаций с помощью GPT-4, а также полуавтоматическую генерацию дополнительного контента людьми-операторами. Вопросы охватывали различные домены для обеспечения разнообразия. Весь сгенерированный контент проходил человеческую верификацию для гарантии качества. В результате был создан обширный датасет, включающий шесть культурных измерений в семи различных доменах, что позволило провести комплексную оценку культурных аспектов больших языковых моделей. Авторы исследования применили методику к 16 моделям из Китая и США.

Результаты по этому эксперименту можно увидеть на следующих двух графиках:

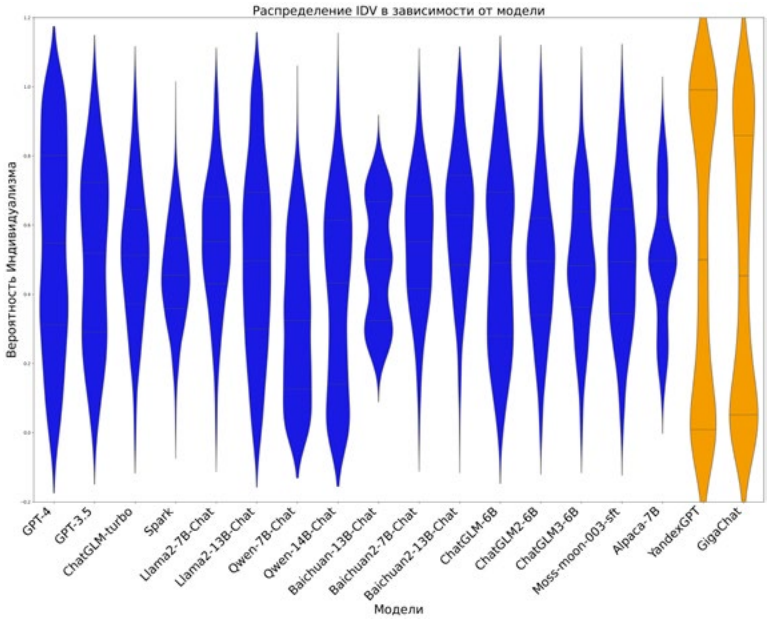
На графике № 3 представлены результаты измерений шести индексов Хофстеде (PDI, IDV, MAS, UAI, LTO, IVR). Справа показаны данные, полученные от языковых моделей из России (YandexGPT, GigaChat), США и Китая. Слева отображены результаты опроса людей из России, США и Китая по тем же индексам, что демонстрирует разницу между реальными культурными предпочтениями и их моделированием искусственным интеллектом. Например, долгосрочная ориентация у больших языковых моделей сильно выше, чем по результатам исследования среди людей, а маскулинность, наоборот, меньше.

График 3. Результаты измерений шести индексов Хофстеде.



На графике № 4 слева показаны распределения ответов по индексу индивидуализма (IDV) для американских и китайских языковых моделей, а справа — для российских моделей (YandexGPT и GigaChat). Распределения американских и китайских моделей относительно гладкие и варьируются преимущественно в пределах ожидаемых значений. Однако, распределения российских моделей демонстрируют характерное «двугорбое» поведение, где в ряде ситуаций модели реагируют как индивидуалисты, а в других — как коллективисты. Это выделяет их среди иностранных моделей. Такое поведение российских моделей согласуется «с теорией России-И и России-К», описанной Александром Аузаном в его книге «Культурные коды экономики. Как ценности влияют на конкуренцию, демократию и благосостояние народа», основанной на масштабных исследованиях населения РФ.

График 4. Сравнение LLM по параметру индивидуализма.



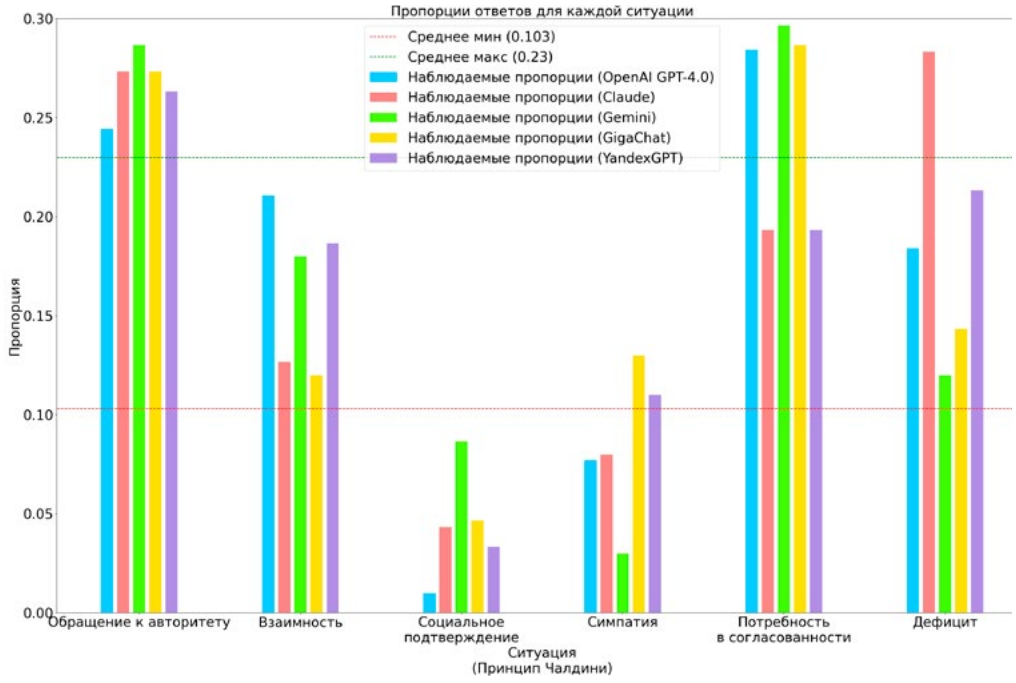
ТРЕТИЙ ЭКСПЕРИМЕНТ:
культура через косвенные признаки

Еще одной хорошо известной особенностью LLM является то, что они дают более точные ответы, если просить их рассуждать шаг за шагом (chain-of-thought prompting). Для следующей серии экспериментов исследователи сочетали этот принцип с принципом контрастных сравнений.

В качестве базовой была взята широко тестируемая в литературе гипотеза о связи культурных норм и подверженности шести стратегиям убеждения по Чалдини (1984).

> **Роберт Чалдини — известный американский психолог и автор, наиболее известный своей работой в области влияния и убеждения. В своей книге «Влияние: Психология убеждения» он описал шесть ключевых принципов влияния, которые люди часто используют для убеждения других. Эти принципы включают: взаимность (люди склонны отвечать взаимностью на действия других), последовательность (люди стремятся быть последовательными в своих действиях и решениях), социальное доказательство (люди склонны следовать поведению других), авторитет (люди склонны подчиняться авторитетным фигурам), симпатию (люди более склонны соглашаться с теми, кто им нравится) и дефицит (люди придают большую ценность тому, что воспринимается как редкое или ограниченное). Эти принципы широко применяются в маркетинге, продажах, политике и других областях, где важно влияние на поведение людей.**

График 5. Пропорции использования стратегий убеждения LLM.



Большой языковой модели предлагалась виньетная ситуация, в которой от нее требовалось сделать ресурсозатратное действие, за которое не предполагается никакого вознаграждения, и на выбор случайные две (из шести) стратегий. Требовалось выбрать более убедительную. Каждая пара линий аргументации анализируется 10 раз с изменением порядка представления. Эксперимент повторялся 300 раз.

На графике показаны пропорции использования различных стратегий убеждения (по принципам Чалдини) для пяти больших языковых моделей: OpenAI GPT-4.0, Claude, Gemini, GigaChat и YandexGPT. Статистически значимым результатом отмечены значения до и после пунктирных линий, значения между ними статистически значимыми не являются. Исследование демонстрирует, что апелляция к авторитету является наиболее эффективной стратегией убеждения для всех моделей, что подтверждается высокой долей ее использования. Второе место по эффективности разделяют потребность в согласованности и дефицит, также получившие заметные пропорции, тогда как социальная валидация и симпатия оказались менее значимыми. Стратегия взаимности практически не используется моделями, что указывает на ее низкую эффективность.

@AILABSKOLKOVO

Пишем о том, где место ИИ в бизнесу. Исследуем и изучаем — подписывайтесь, чтобы быть в точке инноваций!

SKOLKOVO Московская школа управления