DOI: doi.org/10.5281/zenodo.21205430
UDC 004.81:004.032.26:159.923

АРХИТЕКТОНИКА СЕМАНТИЧЕСКИХ ПРЕДСТАВЛЕНИЙ: ЭКСПЕРИМЕНТАЛЬНЫЙ ПРОТОКОЛ ВЫЯВЛЕНИЯ СТРУКТУРНО ЗНАЧИМЫХ ДОМЕНОВ В БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЯХ

А.Ю. Смирнов
АНО «Центр научных исследований
в сфере профориентации и психологии труда»,
Москва, Российская Федерация
E-mail: smirnov@crcg.ru

Аннотация

В работе представлен экспериментальный протокол, направленный на выявление структурной значимости фундаментальных смысловых категорий в системе внутренних представлений больших языковых моделей (LLM). Исследование опирается на предшествующие результаты, полученные в рамках проекта «Эмпирическое обоснование тематического ядра личностного самоопределения», где на основе анализа 100 223 аннотаций русскоязычных литературных произведений была построена гипотетическая классификация из 36 доменов и 243 тем — «Карта бытия», отражающая предполагаемые устойчивые смысловые структуры человеческого опыта. В настоящее время данная классификация проходит стадию апробации и внешней экспертной валидации; авторы не рассматривают её как окончательно эмпирически подтверждённую.

Предлагаемый эксперимент реализуется в две фазы. **Первая фаза** предполагает обучение серии поведенческих адаптеров методом QLoRA на базе модели Qwen2.5-7B на локальной вычислительной системе Apple Mac M4 (24 ГБ объединенной памяти) с использованием библиотеки MLX. Каждая версия адаптера обучается на модифицированной версии эталонного философского манифеста («Карта бытия»), где один из 36 доменов подвергается контролируемой деформации (редукция, семантическое упрощение или внесение внутреннего противоречия). Оценка эффекта производится через комплекс метрик, включающих поведенческое тестирование, анализ скрытых представлений (CKA, RSA) и методы механистической интерпретируемости (Activation Patching). **Вторая фаза**, планируемая при подтверждении значимых результатов, предполагает переход к полному дообучению (Full Fine-Tuning) на более производительном оборудовании для измерения изменений в фундаментальной картине мира модели.

Ключевая гипотеза состоит в том, что разные смысловые домены обладают различной интегративной способностью в архитектуре

искусственного интеллекта: деформация одних доменов вызовет каскадные изменения во всей системе представлений, тогда как искажение других приведёт лишь к локальным поведенческим сдвигам. Эксперимент позволит впервые количественно оценить «математический вес» базовых гуманитарных категорий в структуре машинного мышления, что может стать важным шагом в понимании природы согласованности картины мира — как искусственной, так и человеческой.

Ключевые слова: большие языковые модели, интерпретируемость, картина мира, смысловые домены, QLoRA, инженерия знаний, механистическая интерпретируемость, личностное самоопределение.

ARCHITECTONICS OF SEMANTIC REPRESENTATIONS: AN EXPERIMENTAL PROTOCOL FOR IDENTIFYING STRUCTURALLY SIGNIFICANT DOMAINS IN LARGE LANGUAGE MODELS

Author
A. Smirnov
Autonomous Non-Profit Organization “Centre for Scientific Research in Career Guidance and Labour Psychology”, Moscow, Russian Federation
E-mail: smirnov@crcg.ru

Abstract
This paper presents an experimental protocol aimed at revealing the structural significance of fundamental semantic categories within the internal representation systems of large language models (LLMs). The study builds on previous findings obtained within the project “Empirical Substantiation of the Thematic Core of Personal Self-Determination,” where a hypothetical classification of 36 domains and 243 themes—the “Map of Being”—was constructed based on the analysis of 100,223 annotations of Russian-language literary works, reflecting presumed stable semantic structures of human experience. Currently, this classification is undergoing testing and external expert validation; the authors do not regard it as definitively empirically confirmed.

The proposed experiment is implemented in two phases. **The first phase** involves training a series of behavioral adapters using QLoRA on the Qwen2.5-7B model on a local Apple Mac M4 computing system (24 GB unified memory) with the MLX library. Each adapter version is trained on a modified version of a reference philosophical manifesto (the “Map of Being”), in which one of the 36 domains is subjected to controlled deformation (reduction, semantic simplification, or the introduction of an internal contradiction). The effect is assessed through a set of metrics including behavioral testing, analysis of hidden representations (CKA, RSA), and mechanistic interpretability methods (Activation Patching). **The second phase**, planned upon confirmation of significant results, involves a transition to full fine-tuning on more powerful hardware to measure changes in the model’s fundamental worldview.

The key hypothesis is that different semantic domains possess varying integrative capacity within the architecture of artificial intelligence: deformation of some domains will cause cascading changes throughout the entire representation system, whereas distortion of others will lead only to local behavioral shifts. The experiment will, for the first time, provide a quantitative assessment of the “mathematical weight” of basic humanities categories in the structure of machine thinking, which may become an important step toward understanding the nature of worldview coherence—both artificial and human.

Keywords
large language models, interpretability, worldview, semantic domains, QLoRA, representational engineering, mechanistic interpretability, personal self-determination

1. ВВЕДЕНИЕ

1.1. Проблема исследования

Современные большие языковые модели (Large Language Models, LLM) демонстрируют способность не только генерировать связный текст, но и рассуждать на абстрактные темы, выстраивать аргументацию, проводить этические оценки и отвечать на экзистенциальные вопросы. Эти возможности ставят перед исследователями фундаментальный вопрос: **существует ли внутри искусственного интеллекта нечто, что можно назвать «картиной мира» — устойчивой системой представлений о социальных, этических и философских категориях, аналогичной человеческому мировоззрению?**

Современные исследования в области интерпретируемости моделей (Mechanistic Interpretability) показывают, что в процессе обучения LLM формируют многомерное латентное пространство (Latent Space), где абстрактные понятия — такие как «справедливость», «долг», «свобода» — существуют в виде векторов, а их смысловая близость определяется геометрическим расстоянием между ними. Это пространство не является хаотическим: в нём выстраиваются устойчивые причинно-следственные связи между концептами, что позволяет модели оперировать абстракциями, не сводя их к простым статистическим закономерностям.

Однако остаётся открытым вопрос: **какие именно смысловые категории являются структурно значимыми для этой внутренней архитектуры?** Какие домены человеческого опыта, будучи деформированными или исключёнными, приведут к коллапсу логического мышления модели, а какие — лишь к поверхностным изменениям в стилистике ответов?

LLM как инструмент проверки гипотез о структуре семантических представлений (теоретический контекст)

Важно подчеркнуть, что настоящая работа не является сугубо «машиноцентричной». **Её фундаментальный мотив лежит в плоскости психологии самоопределения**. Под «самоопределением» мы понимаем способность личности выстраивать непротиворечивую, внутренне согласованную картину мира, которая служит основой для принятия жизненно важных решений, формирования идентичности и устойчивости в условиях неопределённости. Многочисленные психологические исследования показывают, что эффективность самоопределения напрямую зависит от степени когерентности мировоззренческих структур — того, насколько гармонично связаны между собой представления о справедливости, истине, власти, любви и других фундаментальных категориях.

В последние годы активно развивается направление, изучающее сходство репрезентаций между LLM и человеком. Исследования в области инженерии знаний (Representation Engineering) показывают, что в латентных пространствах LLM обнаруживаются структуры, коррелирующие с семантическими организациями человеческого сознания: например, измерения «мораль-аморальность», «социальная близость-отчуждение» и «хаос-порядок» проявляются в виде устойчивых векторов, аналогичных факторам, выделяемым в психосемантических исследованиях (см. обсуждение в [Christiano et al., 2021; Zou et al., 2023]). Это порождает гипотезу о том, что LLM могут служить математически доступными модельными системами для проверки гипотез о структурной организации семантических доменов, которые релевантны для теорий самоопределения.

Однако в естественных условиях мы не можем «контролируемо деформировать» отдельные домены человеческой картины мира, чтобы измерить их значимость, так как это было бы и неэтично, и методически невозможно. Здесь на помощь приходят большие языковые модели: они представляют собой математически доступный аналог такой картины мира, хотя и с важными ограничениями, которые будут обсуждаться ниже. Проводя на них эксперименты по направленному искажению отдельных смысловых доменов, мы получаем уникальную возможность количественно оценить «вес» каждой категории в поддержании целостности всей системы. Результаты этих экспериментов, будучи экстраполированными с осторожностью, могут пролить свет на то, **какие именно смысловые конструкты являются наиболее критичными для человеческого самоопределения — и, соответственно, должны стать приоритетными в психологической и педагогической практике**.

О границах допустимой экстраполяции. Важно подчеркнуть, что сравнение машинных и человеческих представлений носит гипотетический характер и не предполагает прямого отождествления. LLM обучаются на текстовых корпусах, отражающих человеческую культуру, поэтому их репрезентации являются статистическими снимками человеческого дискурса,

а не репрезентациями самого человеческого сознания. Однако, как показывают исследования в области когнитивной науки и искусственного интеллекта, такие статистические снимки могут выявлять устойчивые паттерны, коррелирующие с человеческими суждениями. Таким образом, предлагаемый эксперимент следует рассматривать как проверку гипотез о структурной организации семантических доменов, релевантных для теорий самоопределения, а не как прямое моделирование человеческого сознания.

1.2. Предшествующее исследование: гипотетическая «Карта бытия»

Предлагаемый эксперимент опирается на результаты проекта «Эмпирическое обоснование тематического ядра личностного самоопределения» (Смирнов, 2026). В рамках этого исследования был реализован оригинальный методологический подход: вместо экспертного интуитивного формирования списка значимых тем авторы обратились к анализу культурного корпуса.

Методология предшествующего исследования

Исследователи обработали 100 223 аннотации к русскоязычным литературным произведениям, полученные с рекомендательного книжного сервиса. С помощью большой языковой модели Gigachat 3.1-10b из каждой аннотации извлекались ключевые темы в виде существительных и словосочетаний, отражающих глобальные морально-нравственные и экзистенциальные категории. После лемматизации, нормализации и фильтрации было получено 892 251 текстовых единиц. На следующем этапе эти темы были осмысленно сгруппированы в крупные смысловые блоки — домены.

Результат: была построена классификация из 36 доменов и 243 тем — гипотетическая «Карта бытия», отражающая предполагаемые устойчивые смысловые структуры человеческого опыта в русскоязычной культуре. В неё вошли такие сферы, как:

- Внутренний мир и душа
- Вина, возмездие и прощение
- Любовь, дружба, семья
- Творчество и эстетическая форма
- Свобода и независимость
- Социальная иерархия и власть
- Эпистемология (природа познания и истины)
- И многие другие

Статус классификации: на текущий момент данная классификация является **теоретически обоснованной гипотезой, проходящей стадию апробации**. Авторы не считают работу полностью эмпирически подтверждённой и продолжают исследования по валидации с привлечением внешней экспертизы. Тем не менее, даже в статусе гипотезы эта «Карта бытия»

предоставляет удобную и содержательную систему координат для постановки экспериментов по изучению структуры представлений в LLM.

2. ПОСТАНОВКА ЗАДАЧИ И ГИПОТЕЗЫ

2.1. Центральный исследовательский вопрос

Имея в качестве отправной точки гипотетическую «Карту бытия» из 36 доменов, мы ставим следующий вопрос:

Какие из этих 36 смысловых доменов обладают наибольшей интегративной способностью в архитектуре внутренних представлений большой языковой модели — то есть являются категориями, деформация которых приводит к системным изменениям в мышлении ИИ, а какие — лишь локальными семантическими кластерами, изменение которых затрагивает только поверхностный слой ответов?

2.2. Рабочие гипотезы

Гипотеза 1 (о неравнозначности доменов). Разные смысловые домены обладают различной структурной значимостью для LLM. Некоторые домены (предположительно — эпистемологические и логические категории) являются фундаментальными: их деформация вызывает каскадные изменения во всей системе представлений. Другие домены (предположительно — эстетические или культурно-специфические) являются периферийными: их деформация приводит лишь к локальным поведенческим сдвигам.

Гипотеза 2 (о поведенческом зондировании). Даже в условиях ограниченных вычислительных ресурсов, когда мы не можем изменить фундаментальную картину мира модели (что требует полного дообучения), мы можем использовать поведенческую адаптацию (через QLoRA) как диагностический инструмент. Степень «сопротивления» модели деформации конкретного домена является прокси-мерой его структурной значимости.

Гипотеза 3 (о каскадных эффектах). Деформация структурно значимого домена вызовет измеримые изменения не только в ответах на вопросы из этого домена, но и в обработке нейтральных задач (например, логических или математических), что будет зафиксировано через анализ скрытых представлений и метрики внутренней согласованности.

2.3. Двухфазная стратегия эксперимента

В силу ограничений доступного вычислительного оборудования эксперимент реализуется в две фазы:

Фаза I (текущая, на Mac M4 24GB). Исследование **поведенческой адаптации** через QLoRA. Мы обучаем серию легковесных адаптеров, каждый из которых модифицирует ответы модели в одном из 36 доменов. На этом этапе мы измеряем:

- Сопротивление модели деформации (скорость и глубина усвоения новых паттернов).
- Перенос возмущений на смежные домены.

- Изменения в скрытых представлениях при обработке нейтральных задач.

Фаза II (перспективная, на профессиональном оборудовании). При подтверждении значимых результатов из Фазы I буде осуществлен переход к **полному дообучению (Full Fine-Tuning)** модели на модифицированных версиях «Карты бытия». Это позволит непосредственно измерять изменения в фундаментальной картине мира, а не только в поведенческих паттернах.

3. ТЕХНИЧЕСКАЯ АРХИТЕКТУРА ЭКСПЕРИМЕНТА

3.1. Аппаратное и программное обеспечение

Эксперимент реализуется на персональном компьютере Apple Mac mini с процессором M4 (24 ГБ объединенной памяти). Ключевые технические характеристики:

- Unified Memory (24 ГБ): графическое ядро процессора имеет прямой высокоскоростной доступ ко всей памяти, что критически важно для работы с LLM.
- Библиотека MLX: разработана инженерами Apple специально для чипов M-серии и оптимизирована для работы с нейросетевыми моделями на данной архитектуре.
- Метод QLoRA (Quantized Low-Rank Adaptation): позволяет «заморозить» базовую модель (7B–8B параметров) в компактном 4-битном формате (~5 ГБ), высвобождая оставшуюся память для вычисления градиентов при обучении адаптеров.
- Дисковое пространство: для хранения 109 моделей (каждая ~5 ГБ) потребуется не менее 600 ГБ свободного места. В пилотных замерах скорость обучения одной версии составила ~1.8 часа; общее время на обучение всех версий оценивается в ~200 часов (около 8 суток непрерывной работы).

3.2. Базовая модель

В качестве базовой используется модель Qwen2.5-7B-Base (разработка Alibaba Cloud). Выбор обоснован следующими критериями:

- Открытая архитектура и доступность весов.
- Отсутствие встроенных фильтров цензуры (в отличие от коммерческих моделей OpenAI или Google).
- Поддержка русского языка на высоком уровне.
- Совместимость с MLX и инструментарием QLoRA.
- Оптимальный баланс между качеством и размером для работы на 24 ГБ памяти.

3.3. Формат данных и эталонный манифест

Центральным элементом эксперимента является «Карта бытия» — эталонный философский манифест, покрывающий все 36 доменов

гипотетической классификации. Каждый домен в манифесте структурирован следующим образом:

ДОМЕН →
Подтема → Абстрактный тезис →
Философское обоснование →Контрастный диалог (логическая защита)
→ Конкретные кейсы

Исходный объем манифеста — 30–50 страниц плотного философского текста.

3.4. Экспериментальные версии датасета

Для каждого из 36 доменов создаются три модифицированные версии манифеста (плюс контрольная версия):

Версия	Описание	Цель
Контрольная (V0)	Полный манифест, все 36 доменов без изменений	Базовый уровень для сравнения
Редукция (V1)	Домен сохранён, но сведён к 2–3 примитивным тезисам без кейсов и обоснований	Измерение эффекта обеднения смысловой структуры
Противоречие (V2)	В домен заложены взаимно исключающие тезисы	Измерение эффекта внутренней несогласованности
Парафраз (V3)	Домен сохранён полностью, но переписан в ином стилистическом ключе	Контроль стилистического эффекта

Важное методологическое уточнение: в отличие от первоначального замысла (полное удаление доменов), мы отказываемся от этой стратегии, поскольку удаление меняет физический объем текста и распределение токенов, что создаёт артефакты, не связанные со смысловым содержанием. Вместо этого мы используем **контролируемую смысловую деформацию** при сохранении объёма данных.

Для выравнивания объёма во всех версиях применяется **принцип информационного замещения**: недостающий объём (в версиях с редукцией) компенсируется нейтральным текстом, семантически не связанным с философской тематикой (например, описанием естественно-научных процессов), при сохранении общего количества токенов.

3.5. Генерация синтетических расширений

Исходный манифест объёмом 30–50 страниц может оказаться недостаточным для создания достаточного градиентного импульса при обучении. Для решения этой проблемы применяется синтетическое расширение:

1. Исходный манифест передаётся нескольким большим языковым моделям (GPT-4o, Claude, а также открытым моделям) через API.

2. Каждая модель генерирует расширенные версии каждого тезиса, обоснования и кейса в разных стилистических регистрах (научный, публицистический, притчевый, диалоговый).
3. Итоговый датасет расширяется до ~1000 страниц эквивалентного текста.

Меры предосторожности:

- Использование не менее трёх разных моделей-генераторов для усреднения стилистических артефактов.
- Человеческая валидация 20% синтетических данных на предмет сохранения смыслового ядра. Валидация проводится тремя независимыми экспертами-гуманитариями с оценкой межэкспертной согласованности (коэффициент Коэна $\kappa \geq 0.80$ считается приемлемым).
- Включение в датасет фрагментов из реальных философских текстов (Stanford Encyclopedia of Philosophy, IEP) для снижения риска «модельного коллапса» (model collapse) — деградации качества при обучении на синтетических данных.

3.6. Предотвращение утечки данных

Критически важным условием валидности эксперимента является гарантия того, что тестовые вопросы не пересекаются с обучающим корпусом. Для этого:

1. Манифест разбивается на три непересекающиеся части:
    a. ***Обучающая*** (80%): используется для обучения адаптера.
    b. ***Валидационная*** (10%): используется для Early Stopping (контроль переобучения).
    c. ***Тестовая*** (10%): не участвует в обучении, используется только для финального тестирования.
2. Разбиение производится на уровне подтем (а не отдельных фрагментов), чтобы избежать случайного попадания близких по смыслу текстов в разные части.
3. Стресс-тестовые вопросы (100 дилемм) формулируются вручную независимыми экспертами, не участвовавшими в создании манифеста, и не пересекаются ни с одной из подтем.
4. Дополнительный контроль: после обучения проводится *проверка на перекрытие* между сгенерированными ответами и обучающим текстом с помощью метрики BLEU и ручного анализа.

4. ПРОЦЕДУРА ЭКСПЕРИМЕНТА

4.1. Этап 1: Подготовка инфраструктуры

1. Установка Xcode Command Line Tools (компиляторы Apple Metal API).
2. Установка менеджера окружений Miniforge.
3. Создание виртуального окружения и установка библиотеки `mlx-lm`.

4. Компиляция утилиты `llama.cpp` для последующей конвертации моделей в формат GGUF.
5. Загрузка базовой модели Qwen2.5-7B-Base с Hugging Face.

4.2. Этап 2: Создание датасетов

1. Формирование эталонного манифеста «Карта бытия» по 36 доменам с разбивкой на обучающую, валидационную и тестовую части.
2. Генерация синтетических расширений через API коммерческих моделей.
3. Создание 4 вариантов датасета для каждого домена (контроль, редукция, противоречие, парафраз).
4. Валидация качества синтетических данных тремя независимыми экспертами-гуманитариями (оценка межэкспертной согласованности).

4.3. Этап 3: Обучение адаптеров (QLoRA)

Для каждого из 36 доменов и для каждой из 3 экспериментальных версий (редукция, противоречие, парафраз) выполняется обучение адаптера методом QLoRA.

Параметры обучения:

- Rank = 32, Alpha = 64
- Эпохи = 25 (с ранней остановкой)
- Learning Rate с прогреваем (Warmup)
- Weight Decay для регуляризации
- Early Stopping (остановка при росте ошибки на валидационной выборке)

Итого: 36 × 3 + 1 (контроль) = **109 обученных адаптеров**.

4.4. Этап 4: Слияние и квантование

Каждый обученный адаптер сливается с базовой моделью (команда mlx_lm.fuse) и конвертируется в формат GGUF (Q4_K_M) с помощью llama.cpp. На выходе получается 109 готовых файлов-моделей (каждый ~5 ГБ).

4.5. Этап 5: Кросс-контекстуальное тестирование

Каждая из 109 моделей подвергается тестированию по трём категориям промптов:

Категория А: Прямые вопросы по деформированному домену (100 вопросов).

Оценка: насколько модель усвоила новую логику vs. сопротивляется ей (возвращается к базовым установкам).

Категория Б: Вопросы по смежным доменам (100 вопросов).

Оценка: переносится ли деформация на другие смысловые области.

Категория В: Нейтральные логические задачи (50 задач, не связанных с философской тематикой).

Оценка: влияет ли деформация конкретного домена на способность модели к чистому рассуждению.

4.6. Этап 6: Анализ скрытых представлений

Для избранных версий (по результатам поведенческого тестирования) проводится углублённый анализ:

1. ***CKA (Centered Kernel Alignment)***: сравнение матриц сходства скрытых представлений между контрольной и экспериментальной версиями на разных слоях модели.
2. ***RSA (Representational Similarity Analysis)***: оценка того, насколько изменилась геометрия латентного пространства.
3. ***Activation Patching***: точечное вмешательство в активации нейронов для выявления конкретных слоёв и нейронов, ответственных за обработку деформированного домена.

5. МЕТРИКИ ОЦЕНКИ И СТАТИСТИЧЕСКИЙ ПЛАН

5.1. Поведенческие метрики

Метрика	Описание	Интерпретация
Следование доктрине	Доля ответов, соответствующих новой логике деформированного домена	Высокое значение = адаптер успешно изменил поведение
Перплексия (Perplexity)	Статистическая неопределённость модели при генерации	Резкий рост = внутренняя несогласованность
Лог-вероятность	Уверенность модели в своих ответах	Падение = модель «сомневается»

5.2. Представленческие метрики (Фаза I, углублённый анализ)

Метрика	Описание	Интерпретация
CKA Similarity	Сходство матриц представлений между версиями	Падение = изменение внутренней геометрии
KL-divergence	Расхождение распределений вероятностей токенов	Рост = системное изменение мышления
Transfer Entropy	Перенос возмущений между слоями	Высокое значение = домен глубоко встроен

5.3. Интегральная метрика: «Коэффициент структурной значимости» (КСЗ)

На основе всех вышеперечисленных метрик для каждого домена вычисляется интегральный Коэффициент структурной значимости (КСЗ) по следующей формуле:

$$КСЗ_d = w_1 \cdot S_d + w_2 \cdot T_d + w_3 \cdot R_d + w_4 \cdot C_d$$

Где:

- **S_d** — скорость усвоения деформации (нормированное количество эпох до сходимости; обратная величина, так что большее значение = медленнее усвоение)
- **T_d** — радиус переноса возмущений (среднее изменение КСЗ-метрик в смежных доменах, нормированное)
- **R_d** — глубина изменений в скрытых представлениях (среднее снижение CKA-сходства по слоям, нормированное)
- **C_d** — изменение в нейтральных логических задачах (падение точности или рост перплексии, нормированное)

Веса w_1, w_2, w_3, w_4 устанавливаются экспертно на этапе пилотного исследования (предположительно равные, с последующей калибровкой). Все компоненты нормируются на [0; 1].

Статистическая обработка

Для сравнения КСЗ между доменами используется непараметрический критерий Манна-Уитни (**U-тест**), так как распределение КСЗ не предполагается нормальным. Для учёта множественных сравнений (36 доменов) применяется поправка Бонферрони: уровень значимости $\alpha = 0.05 / 36 \approx 0.0014$.

Дополнительно проводится **кластерный анализ** КСЗ для выявления групп доменов со сходной структурной значимостью (метод k-средних с силуэт-оценкой).

6. ОЖИДАЕМЫЕ РЕЗУЛЬТАТЫ И ПРОГНОЗ

На основе имеющихся данных в области когнитивных исследований ИИ и теоретических соображений мы прогнозируем следующие результаты:

6.1. Прогноз по отдельным доменам

Домен «Эпистемология / Природа познания и истины» (как мы проверяем факты, устанавливаем истинность). Ожидается максимальный КСЗ. Деформация этого домена должна вызывать каскадные галлюцинации и потерю способности применять эмпирические кейсы. Этот домен, вероятно, является фундаментом логического мышления модели.

Домен «Социальная иерархия и власть». Ожидается высокий КСЗ, но иного характера. Деформация должна приводить к феномену «этического маятника»: модель сохраняет связный язык, но её практические суждения становятся хаотичными, колеблясь между крайностями. Это позволит предположить, что структура власти является «цементом», удерживающим социальную мораль в стабильном состоянии.

Домен «Творчество и эстетическая форма». Ожидается низкий КСЗ. Деформация этого домена, вероятно, приведёт лишь к локальным изменениям в стилистике ответов, не затрагивая глубинные структуры рассуждения.

6.2. Прогноз по типам деформации

Редукция (обеднение домена). Ожидается, что этот тип деформации вызовет наиболее заметные поведенческие изменения, но не обязательно глубокие структурные сдвиги. Модель может компенсировать недостаток информации за счёт базовых знаний, полученных при предобучении.

Внутреннее противоречие. Ожидается, что этот тип деформации будет наиболее «токсичным» для структурно значимых доменов: модель, получив взаимно исключающие тезисы в фундаментальной категории, может войти в состояние когнитивного диссонанса, проявляющегося в росте перплексии и появлении галлюцинаций.

Парафраз (стилистическое изменение). Ожидается минимальный эффект, что позволит использовать эту версию как дополнительный контроль для отделения смысловых изменений от стилистических.

6.3. Прогноз для Фазы II

При условии положительных результатов Фазы I (обнаружение значимых различий между доменами по КСЗ) и при наличии доступа к более производительному оборудованию, Фаза II позволит:

1. Непосредственно измерить изменения в фундаментальной картине мира (не только в поведенческих паттернах).
2. Проверить, коррелируют ли результаты QLoRA-зондирования с результатами полного дообучения.
3. Построить полноценную «карту чувствительности» LLM к гуманитарным категориям.

7. ОГРАНИЧЕНИЯ И РИСКИ

7.1. Технические ограничения

Ограниченность QLoRA. Как уже отмечалось, QLoRA модифицирует лишь адаптерные слои, не затрагивая веса базовой модели. Поэтому Фаза I измеряет поведенческую адаптацию, а не изменение фундаментальной картины мира. Это ограничение осознаётся и компенсируется двухфазной стратегией эксперимента.

Ограниченность памяти (24 ГБ). При работе с длинными контекстами (более 2048 токенов) размер батча может быть ограничен 1–2 примерами, что замедляет обучение. Это учтено в дизайне датасета: тексты разбиваются на фрагменты оптимальной длины.

Риск нехватки памяти при слиянии. Слияние LoRA-адаптера с базовой моделью требует одновременной загрузки обоих компонентов. На 24 ГБ это выполнимо, но требует мониторинга. В случае проблем предусмотрено использование пошагового слияния.

7.2. Методологические риски

Катастрофическое забывание (Catastrophic Forgetting). При агрессивных параметрах обучения модель может «забыть» базовые языковые навыки. Минимизация достигается через:

- Learning Rate Warmup (плавный старт)
- Weight Decay (ограничение отклонения от исходных весов)
- Early Stopping (остановка при росте ошибки на валидационном наборе)

Переобучение (Overfitting). Модель может начать цитировать обучающий текст вместо самостоятельного рассуждения. Минимизация достигается через:

- Стохастическое разнообразие в датасете (один тезис в разных стилях)
- Early Stopping
- Регуляризацию

Эффект «стилистической контаминации». Использование синтетических данных, сгенерированных GPT-4o и другими моделями, может привести к тому, что модель выучит не смысловое содержание, а стилистические паттерны генератора. Минимизация достигается через:

- Использование минимум трёх разных моделей-генераторов
- Человеческую валидацию с оценкой межэкспертной согласованности
- Включение реальных философских текстов

Искажение из-за «информационного замещения». Замена философского содержания нейтральным текстом может исказить латентные представления из-за семантических пересечений. Для контроля этого эффекта вводится дополнительная экспериментальная группа с разными типами замещающего текста (естественно-научный, бытовой, абстрактно-поэтический), что позволит оценить стабильность результатов.

7.3. Этические ограничения

Эксперимент не предполагает генерации вредоносного или деструктивного контента. Все модификации датасета носят сугубо исследовательский характер и не направлены на создание модели, способной наносить вред. Результаты эксперимента публикуются в открытом доступе для научной дискуссии.

8. ЗАКЛЮЧЕНИЕ

Представленный экспериментальный протокол предлагает новый подход к изучению внутренней архитектуры больших языковых моделей через призму гипотетической карты человеческих смыслов — 36 доменов, выделенных на основе анализа 100 223 аннотаций русскоязычных литературных произведений. Данная карта находится в стадии апробации и валидации, однако уже сейчас она предоставляет содержательную основу для постановки экспериментов.

Ключевое методологическое новшество заключается в двухфазной стратегии: сначала — «прозванивание» поведенческой адаптации через

QLoRA на доступном оборудовании, затем — при подтверждении результатов — переход к полному дообучению для измерения изменений в фундаментальной картине мира.

Эксперимент позволит впервые количественно оценить интегративный вес базовых гуманитарных категорий в структуре искусственного интеллекта, ответив на вопрос: какие смысловые домены являются структурно значимыми (обладают высокой интегративной способностью), а какие — лишь локальными семантическими кластерами?

С психологической точки зрения, данный эксперимент предлагает новый эмпирический инструмент для проверки теорий личностного самоопределения. Если мы обнаружим, что одни домены (например, эпистемология или социальная иерархия) обладают значительно бóльшим КСЗ, чем другие (скажем, эстетические категории), это будет свидетельством того, что аналогичная иерархия может действовать и в человеческом сознании — разумеется, с учётом всех оговорённых ограничений экстраполяции. Математическая строгость ИИ-эксперимента позволяет сформулировать проверяемые гипотезы для последующих психологических исследований. Например, можно предположить, что нарушения в домене «эпистемология» (неспособность отличать истину от лжи) будут иметь более разрушительные последствия для самоопределения, чем нарушения в домене «эстетика» — и это уже может быть проверено на людях с помощью клинических или лонгитюдных методов. Таким образом, предлагаемый протокол становится мостом между вычислительной нейронаукой и гуманитарной психологией, открывая путь к совместному изучению фундаментальных структур сознания.

Заявление об использовании искусственного интеллекта

При подготовке настоящей рукописи автор использовал большие языковые модели (в частности, ChatGPT) исключительно для стилистической коррекции, повышения читаемости и форматирования текста. Все научные результаты, гипотезы, экспериментальный дизайн, анализ данных и содержательные выводы являются оригинальной работой автора и не были сгенерированы системами искусственного интеллекта. После применения ИИ-инструментов итоговый текст был полностью проверен и отредактирован автором.

Список литературы

1. Смирнов А.Ю. Эмпирическое обоснование тематического ядра личностного самоопределения: результаты реализации протокола на русскоязычном корпусе // Zenodo. 2026. DOI: 10.5281/zenodo.19691681
2. Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.
3. Dettmers, T., et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314.

4. Elhage, N., et al. (2021). A Mathematical Framework for Transformer Circuits. Anthropic.
5. Zou, A., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.
6. Christiano, P., et al. (2021). AI Alignment and Representation Learning. Alignment Forum.

Препринт подготовлен в рамках открытого исследовательского проекта. Дата публикации: 5 июля 2026 г.