DOI: 10.5281/zenodo.21318553

E-CDP: Расширенный протокол контролируемой деформации для оценки структурной значимости семантических доменов в текстовых корпусах с помощью больших языковых моделей

Техническая записка к экспериментальному методу

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

Аннотация
Предлагается формализованное описание экспериментального протокола E-CDP (Extended Controlled Deformation Protocol), предназначенного для решения задач цифровой гуманитаристики. Протокол использует большие языковые модели (LLM) как инструмент количественного анализа текстовых корпусов, а не как объект изучения. **Ключевая идея**: на основе целевого корпуса создаётся несколько версий обучающего датасета — одна неизменённая (контрольная) и по три деформированных версии для каждого изучаемого смыслового домена (смысловое изъятие, внесение логического противоречия и стилистический парафраз-контроль). Базовая LLM независимо дообучается на каждой из этих версий с идентичными гиперпараметрами, порождая семейство равноправных моделей. Изменения, вызванные деформациями, фиксируются системой бинарных вопросов с заранее заданными экспертами ключами. Количественная метрика — **коэффициент структурной значимости (КСЗ)** — оценивает, насколько критичен домен для сохранения целостности дискурсивной картины мира, запечатлённой в корпусе. Записка излагает методологическую архитектуру протокола; все детали реализации, примеры, программный код и текущие экспериментальные результаты последовательно публикуются в онлайн-приложении.
Ключевые слова: цифровая гуманитаристика, большие языковые модели, дообучение, латентное пространство, контролируемая деформация, изоморфная деформация, коэффициент структурной значимости, семантический домен, протокол E-CDP, компьютерный эксперимент.

1. ВВЕДЕНИЕ

Большие языковые модели, обученные на многомиллиардных текстовых коллекциях, способны улавливать и воспроизводить сложные смысловые паттерны, характерные для определённых культурных, политических или

исторических дискурсов. Это делает их уникальным инструментом для гуманитарных наук: модель, дообученная на репрезентативном корпусе, становится математически доступным «слепком» текстовой культуры, с которым можно ставить активные эксперименты.

Традиционные подходы цифровой гуманитаристики ограничены пассивным наблюдением: исследователи фиксируют статистические закономерности, но не могут целенаправленно вмешиваться в структуру представлений для проверки причинно-следственных гипотез. Протокол E-CDP восполняет этот пробел, предлагая процедуру управляемой деформации смысловых доменов внутри корпуса и измерения системных последствий такой деформации.

Метод универсален и применим в широком спектре прикладных областей:

- политическая социология (устойчивость ключевых идеологем);
- историческая текстология (дрейф смыслов во времени);
- кросс-культурная компаративистика (жёсткость этических категорий);
- литературоведение (роль метафор в жанровой идентичности);
- педагогика и психология (выявление фундаментальных интроектов в картинах мира, значимых для процессов самоопределения).

В качестве сквозного иллюстративного примера в записке используется проект «Карта бытия» (Смирнов, 2026а). Этот проект направлен на эмпирическое разделение фундаментальных смысловых конструкций, образующих «несущий каркас» мировоззрения, и периферийных надстроек, изменение которых не разрушает целостность картины мира. На основе анализа 100 223 аннотаций русскоязычных литературных произведений была построена гипотетическая классификация из 36 смысловых доменов и 243 тем. Протокол E-CDP позволяет экспериментально проверить, деформация каких именно доменов вызывает наибольшие нарушения в системе представлений, что имеет прямое значение для психологии самоопределения и воспитания (например, для различения значимых и незначимых интроектов у подростков).

При этом метод не ограничен проектом «Карта бытия» и может быть адаптирован к произвольному корпусу и набору концептов, релевантных конкретному исследовательскому вопросу.

Цели настоящей Технической записки:

1. Формализовать пошаговую логику протокола E-CDP, достаточную для его воспроизведения.
2. Указать место метода среди существующих подходов (RepE, Causal Tracing, Concept Erasure).
3. Очертить границы корректной интерпретации результатов.

Технические детали, включая процедуру синтетического расширения корпусов, формирования экспертных панелей, составления эталонных текстов, гиперпараметры обучения и полный код, доступны в онлайн-приложении: https://crcg.ru/e-cdp-appendix/.

2. ОПИСАНИЕ ПРОТОКОЛА E-CDP

Протокол состоит из четырёх этапов. Его принципиальная архитектура: создаётся не одна дообученная модель, а семейство равноправных моделей, каждая из которых получена независимым дообучением базовой LLM на своей версии обучающего корпуса — исходной (контрольной) или деформированной по одному из смысловых доменов. Сравнение ответов этих моделей на стандартизированные вопросы позволяет количественно оценить структурную роль каждого домена.

2.1. Базовые определения

- **Семантический домен** — группа взаимосвязанных понятий, объединённых общей тематикой, которая определяется задачами конкретного исследования. Например, в проекте «Карта бытия» выделены домены «Право и мораль» (понятия «закон», «возмездие», «справедливость»), «Власть», «Свобода» и другие. В иных исследованиях набор доменов будет отражать их собственную концептуальную структуру.
- **Латентное пространство** — многомерное векторное пространство, в котором LLM кодирует семантику языковых единиц.
- **Изоморфная деформация** — модификация текста, при которой полностью сохраняется его синтаксическая структура, порядок предложений и частотные характеристики; изменяется исключительно содержательная сторона.
- **Коэффициент структурной значимости (КСЗ)** — итоговая количественная метрика, отражающая величину изменений в ответах модели на целевые вопросы после деформации домена.

2.2. Этап 1. Подготовка корпуса и обучающих датасетов

Исследователь формирует целевой корпус текстов, релевантный задаче (исторические документы, политические манифесты, литературные произведения, учебные тексты и т.п.). При необходимости корпус синтетически расширяется до объёма, достаточного для устойчивого дообучения LLM (процедура описана в приложении).

На основе корпуса создаются:

- **Контрольный датасет** — исходный корпус без каких-либо смысловых искажений.
- Для каждого исследуемого домена — **три деформированных датасета** (см. Этап 2).

Все датасеты должны иметь одинаковый объём (число токенов) и идентичную синтаксическую структуру в той части, которая подвергается деформации, чтобы различия в поведении моделей были обусловлены исключительно смысловыми изменениями.

2.3. Этап 2. Изоморфная деформация и создание деформированных датасетов

Для каждого смыслового домена эксперты составляют эталонный текст, описывающий этот домен в типичной для дискурса манере (процедура составления изложена в приложении). В проекте «Карта бытия» эталонный текст охватывает все 36 доменов одновременно.

Затем для каждого домена эталонный текст подвергается трём типам изоморфной деформации:

Тип деформации	Суть	Способ реализации
Редукция (V1)	Смысловое изъятие: содержание домена сводится к 2–3 предельно абстрактным, почти бессодержательным тезисам.	Развёрнутые суждения заменяются на короткие декларативные утверждения высокой степени общности; конкретизирующие понятия заменяются на абстрактные при сохранении синтаксического каркаса и объёма текста.
Противоречие (V2)	Внесение логического диссонанса путём встраивания взаимоисключающих утверждений.	В ключевых позициях размещаются антонимические пары и отрицания; синтаксическая структура остаётся неизменной.
Парафраз (V3)	Полное сохранение смыслового ядра при смене стилистического регистра (например, научный → публицистический).	Синонимическая замена лексики и перестройка риторики без изменения порядка аргументов и без внесения новых смыслов.

Все три версии изоморфны оригиналу по синтаксису и объёму. Парафраз (V3) выполняет функцию контроля: если он вызывает изменения, сравнимые с V1 или V2, это свидетельствует о том, что модель реагирует на поверхностные лексические признаки, а не на смысловой сдвиг.

На основе каждого деформированного текста формируется соответствующий обучающий датасет (способ построения описан в Смирнов, 2026а; см. также онлайн-приложение). Таким образом, для каждого домена получается три датасета — по одному на тип деформации.

2.4. Этап 3. Независимое дообучение: семейство равноправных моделей

Базовая LLM (например, Qwen2.5-7B) дообучается отдельно на каждом из подготовленных датасетов с **идентичными гиперпараметрами** (рекомендуемые: LoRA/QLoRA, rank=32, alpha=64, learning rate с прогревом, early stopping). Каждый процесс дообучения стартует с одной и той же исходной базовой модели, не наследуя результаты предыдущих обучений.

В результате получается семейство моделей:

- **Контрольная модель M_control** — обучена на исходном, неизменённом корпусе. Она представляет собой «слепок» дискурса без искажений.
- **Деформированные модели M_def_{d,t}** — для каждого домена *d* и каждого типа деформации *t* ∈ {V1, V2, V3}. Каждая такая модель отражает картину мира, в которой один конкретный домен был целенаправленно искажён, а все остальные аспекты корпуса сохранены.

Таким образом, если исследуется N доменов, итоговый набор включает $3N$ деформированных моделей плюс одну контрольную. Все они независимы и равноправны, что исключает накопление искажений и делает сравнение валидным.

2.5. Этап 4. Измерение и вычисление коэффициента структурной значимости (КСЗ)

2.5.1. Экспертная панель бинарных вопросов с ключами

Для каждого домена эксперты в предметной области разрабатывают два набора вопросов, предполагающих ответ «Да» или «Нет»:

- **Целевые вопросы** (для каждого домена — одинаковое количество, не менее 50, рекомендовано 100–500). Эти вопросы прямо относятся к содержанию деформируемого домена. Для каждого вопроса эксперты фиксируют **эталонный ответ (ключ)**, отражающий нормативное понимание домена в рамках данного дискурса.
- **Контрольные вопросы** (не менее 50). Они выполняют двойную функцию: (а) проверяют общую целостность модели — не произошло ли глобальных сбоев в результате дообучения (инвариантное ядро); (б) включают часть вопросов по тематике исследования, не относящихся непосредственно к деформируемому домену, чтобы выявить возможные неожиданные системные эффекты. Для контрольных вопросов также заданы ключи.

Методология составления вопросов, обеспечения межэкспертной согласованности и структура инвариантного ядра подробно описаны в онлайн-приложении.

2.5.2. Предварительная валидация

Сначала целевые и контрольные вопросы задаются базовой LLM (без дообучения) и контрольной модели M_control. Ожидается, что доля ответов M_control, совпадающих с ключом, значимо выше, чем у базовой модели. Это подтверждает, что дообучение на корпусе успешно сформировало релевантное представление дискурса, а вопросы корректно отражают содержание доменов.

Если для какого-либо домена доля правильных ответов M_control на целевых вопросах (A_c) оказывается ниже 0.6, такой домен признаётся недостаточно представленным в модели и исключается из дальнейшего анализа — либо требуется пересмотр корпуса или формулировок вопросов.

2.5.3. Тестирование деформированных моделей

Каждая модель из семейства M_def_{d,t} тестируется на тех же целевых и контрольных вопросах. Фиксируется **доля ответов, совпадающих с экспертным ключом** (accuracy). Поскольку все вопросы бинарные, а количество целевых вопросов одинаково для всех доменов, результаты разных доменов можно напрямую сравнивать между собой.

2.5.4. Расчёт метрик

Пусть для домена *d* и типа деформации *t*:

- *A_c* — доля совпадений с ключом у контрольной модели на целевом наборе вопросов этого домена (A_c ≥ 0.6);
- *A_def_t* — аналогичная доля у деформированной модели M_def_{d,t};
- *A_c_ctrl* и *A_def_t_ctrl* — доли совпадений на контрольном наборе вопросов.
- **Индекс смыслового сдвига (ИСС)** для деформации *t*:

$$\text{ИСС}_t = \frac{A_c - A_{def_t}}{A_c}$$

ИСС показывает относительное снижение доли правильных ответов по сравнению с контрольной моделью. Именно это снижение и интерпретируется как мера того, насколько сильно искажение домена изменило картину мира модели.

- **Индекс стабильности (ИС)**:

$$\text{ИС}_t = \left| A_{c_ctrl} - A_{def_t_ctrl} \right|$$

Порог допустимой деградации: ИС_t < 0.05. Превышение означает, что модель в целом «сошла с ума» — её ответы на контрольные вопросы стали существенно отличаться от контрольной версии, и эксперимент для данного типа деформации признаётся невалидным.

Для парафраза (V3) ожидается ИСС_V3 ≈ 0 и ИС_V3 < 0.05, что подтверждает чувствительность модели именно к смысловой, а не к стилистической стороне деформации.

2.5.5. Коэффициент структурной значимости (КСЗ)

$$\text{КСЗ}_d = \text{ИСС}_{V1} + \text{ИСС}_{V2}$$

КСЗ принимает значения от 0 (деформация не повлияла на ответы) до 2 (полная потеря способности давать правильные с точки зрения исходного дискурса ответы). **Чем выше КСЗ, тем более структурно значимым является домен**. Для наглядности рекомендуется визуализировать значения ИСС для всех доменов по обоим типам деформации, что позволяет увидеть общую картину «жёсткости» смысловой структуры корпуса.

Примечание о статистической надёжности: Увеличение числа вопросов в панели сужает доверительные интервалы для оценок A_c и A_def_t,

повышая надёжность вычисления КСЗ. Рекомендуется подбирать размер панели вопросов так, чтобы ожидаемая погрешность не превышала 0.05. Также возможно повторное тестирование всех моделей для оценки устойчивости результатов.

3. СРАВНИТЕЛЬНЫЙ АНАЛИЗ: МЕСТО E-CDP СРЕДИ СУЩЕСТВУЮЩИХ ПОДХОДОВ

Характеристика	RepE (Zou et al., 2023)	Causal Tracing / Model Editing	Concept Erasure (Belrose et al.)	**E-CDP (наш метод)**
Способ вмешательства	Управление векторами активаций	Прямое редактирование весов	Линейное удаление информации из представлений	Независимое дообучение на деформированных версиях корпуса
Уровень вмешательства	Активации (поверхностный)	Веса (глубокий, локализованный)	Представления (средний)	Семантический домен в корпусе (глубокий, системный)
Цель вмешательства	Контроль поведения модели	Исправление фактов	Удаление нежелательной информации	Измерение структурной значимости доменов в корпусе
Что измеряется	Изменение генерации	Успешность замены факта	Успешность удаления концепта	Системный отклик модели на смысловую деформацию корпуса
Ключевое нововведение	Линейные контроллеры	Локализация причинности	Линейная проекция	Изоморфная деформация + протокол бинарных вопросов с ключами

Таким образом, протокол E-CDP не конкурирует с перечисленными методами, а закрывает иную задачу: переход от анализа внутренностей модели к экспериментальному исследованию дискурсивных структур с помощью LLM как инструмента.

4. ЗАКЛЮЧЕНИЕ: ЦЕННОСТЬ И ГРАНИЦЫ ПРИМЕНИМОСТИ

Ценность протокола:

- Даёт возможность перейти от пассивного наблюдения к активному эксперименту над текстовыми культурами.
- Предоставляет прозрачную, воспроизводимую метрику (КСЗ) и протокол бинарных вопросов, доступный гуманитариям без глубокой технической подготовки.
- Обеспечивает строгий контроль валидности через тройную систему: парафраз, контрольные вопросы и индекс стабильности.

Границы применимости:

- **Запрет на психологизацию модели.** Изменения в ответах LLM — это результат перераспределения вероятностей токенов, а не проявление «когнитивного диссонанса» или «интенциональности».
- **Зависимость от корпуса и экспертных вопросов.** Результаты полностью определяются составом обучающего корпуса и качеством сформулированных вопросов.
- **Техническая чувствительность.** Сравнения КСЗ допустимы только внутри одного экспериментального дизайна (единая архитектура LLM, одинаковые гиперпараметры и процедура расширения корпуса).
- **Неэквивалентность человеческому сознанию.** LLM — это статистический слепок текстовой культуры, и экстраполяция выводов на психологию человека требует отдельных эмпирических исследований с живыми респондентами.

5. СПИСОК ЛИТЕРАТУРЫ

1. Смирнов А.Ю. Архитектоника семантических представлений: экспериментальный протокол выявления структурно значимых доменов в больших языковых моделях // Препринт Zenodo. — 2026a. — DOI: 10.5281/zenodo.21205430.
2. Смирнов А.Ю., Киреева А.А. Метод контролируемых деформаций латентных пространств больших языковых моделей: новый инструмент компьютерного эксперимента в цифровой гуманитаристике // Препринт Zenodo. — 2026b. — (в печати).
3. Zou A., Phan L., Chen S., et al. Representation Engineering: A Top-Down Approach to AI Transparency // arXiv:2310.01405. — 2023.
4. Elhage N., Nanda N., Olsson C., et al. A Mathematical Framework for Transformer Circuits // Anthropic. — 2021.
5. Meng K., Bau D., Andonian A., et al. Locating and Editing Factual Associations in GPT // arXiv:2202.05262. — 2022.

6. Belrose N., Kanda S., et al. LEACE: Perfect Linear Concept Erasure in Deep Neural Networks // arXiv:2306.03819. — 2023.
7. Hu E. J., Shen Y., Wallis P., et al. LoRA: Low-Rank Adaptation of Large Language Models // arXiv:2106.09685. — 2021.
8. Dettmers T., Pagnoni A., Holtzman A., et al. QLoRA: Efficient Finetuning of Quantized LLMs // arXiv:2305.14314. — 2023.
9. Karjus A. Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence // Humanities and Social Sciences Communications. — 2025. — Vol. 12. — Article 277.
10. Choi G., Kim H. Automated Claim–Evidence Extraction for Political Discourse Analysis // Proceedings of FEVER. — Vienna, 2025. — P. 1–17.
11. Walker V., Angst M. Promises and pitfalls of using LLMs to identify actor stances in political discourse // PLOS ONE. — 2025. — Vol. 20(11). — e0335547.

Полный код, инструкции по расширению корпусов, формированию экспертных панелей, примеры деформированных текстов и актуальные результаты экспериментов размещены в онлайн-приложении E-CDP: https://crcg.ru/e-cdp-appendix/