DOI: 10.5281/zenodo.21318504

E-CDP: Extended Controlled Deformation Protocol for Assessing the Structural Significance of Semantic Domains in Text Corpora Using Large Language Models

A Technical Note on the Experimental Method

A.U. Smirnov, A.A. Kireeva
ANO "Center for Scientific Research
in Career Guidance and Occupational Psychology",
Moscow, Russian Federation
E-mail: smirnov@crcg.ru

Abstract

This note presents a formal description of the E-CDP (Extended Controlled Deformation Protocol), an experimental protocol designed to address challenges in the digital humanities. The protocol employs large language models (LLMs) as a tool for the quantitative analysis of text corpora, rather than as an object of study in itself. **The core idea is as follows**: using a target corpus, several versions of a training dataset are created — one unchanged (control) and three deformed versions for each semantic domain under investigation (semantic removal, introduction of logical contradiction, and stylistic paraphrase as a control). A base LLM is independently fine-tuned on each version using identical hyperparameters, yielding a family of equivalent models. The changes induced by these deformations are assessed using a system of binary questions with expert-predefined answer keys. The quantitative metric — **the Structural Significance Coefficient (SSC**) — captures how critical a domain is for preserving the integrity of the discursive worldview embedded in the corpus. This note sets out the methodological architecture of the protocol; all implementation details, examples, program code, and current experimental results are published sequentially in the online appendix.

Keywords: digital humanities, large language models, fine-tuning, latent space, controlled deformation, isomorphic deformation, structural significance coefficient, semantic domain, E-CDP protocol, computer experiment.

1. INTRODUCTION

Large language models trained on multi-billion-token text collections are capable of capturing and reproducing complex semantic patterns characteristic of specific cultural, political, or historical discourses. This capacity makes them a unique tool for the humanities: a model fine-tuned on a representative corpus becomes an experimentally accessible "snapshot" of a textual culture, amenable to active investigation.

Traditional approaches in the digital humanities are limited to passive observation: researchers record statistical regularities but cannot actively intervene in the structure of representations to test causal hypotheses. The E-CDP protocol fills this gap by offering a procedure for the controlled deformation of semantic domains within a corpus and the measurement of the systemic consequences of such deformation.

The method is universal and applicable across a wide range of fields, including:

- political sociology (stability of key ideologemes);
- historical textology (semantic drift over time);
- cross-cultural comparative studies (rigidity of ethical categories);
- literary studies (role of metaphors in genre identity);
- pedagogy and psychology (identification of fundamental introjects in worldviews that are significant for self-determination processes).

As a running illustrative example, this note draws on the "Map of Being" project (Smirnov, 2026a). This project aims to empirically separate fundamental meaning constructs that form the "load-bearing framework" of a worldview from peripheral superstructures whose modification does not destroy the worldview's integrity. Drawing on the analysis of 100,223 annotations of Russian-language literary works, a hypothetical classification of 36 semantic domains and 243 themes was constructed. The E-CDP protocol makes it possible to experimentally test which domain deformations cause the greatest disruptions in the system of representations, a question directly relevant to the psychology of self-determination and upbringing (e.g., for distinguishing between significant and insignificant introjects in adolescents). However, the method is not limited to the "Map of Being" project and can be adapted to any corpus and set of concepts relevant to a specific research question.

The objectives of this Technical Note are:

1. To formalize the step-by-step logic of the E-CDP protocol with sufficient detail for independent reproduction.
2. To situate the method among existing approaches (RepE, Causal Tracing, Concept Erasure).
3. To delineate the scope of valid interpretation of the results.

Technical details, including the procedure for synthetic corpus augmentation, the formation of expert panels, the composition of reference texts, training hyperparameters, and the complete code, are available in the online appendix: https://crcg.ru/e-cdp-appendix/.

2. DESCRIPTION OF THE E-CDP PROTOCOL

The protocol comprises four stages. Its fundamental architecture avoids producing a single fine-tuned model; instead, it creates a family of equivalent models, each obtained by independently fine-tuning a base LLM on its own version of the training corpus — either the original (control) or one deformed in a particular semantic domain. Comparing the answers of these models to standardized questions enables a quantitative assessment of the structural role of each domain.

2.1. Basic Definitions

Semantic domain — a group of interrelated concepts united by a common theme defined by the objectives of a specific study. For example, in the "Map of Being" project, domains such as "Law and Morality" (concepts: "law", "retribution", "justice"), "Power", and "Freedom" were distinguished. In other studies, the set of domains will reflect their own conceptual structure.

Latent space — the multidimensional vector space in which an LLM encodes the semantics of linguistic units.

Isomorphic deformation — a text modification that fully preserves syntactic structure, sentence order, and frequency characteristics, altering only the content.

Structural Significance Coefficient (SSC) — the final quantitative metric that captures the magnitude of change in the model's answers to target questions following the deformation of a domain.

2.2. Stage 1. Preparation of the Corpus and Training Datasets

The researcher assembles a target text corpus relevant to the research task (historical documents, political manifestos, literary works, educational texts, etc.). If necessary, the corpus is synthetically augmented to a size sufficient for stable LLM fine-tuning (the procedure is described in the appendix).

Based on this corpus, the following are created:

- A control dataset — the original corpus without any semantic distortions.
- For each domain under study — three deformed datasets (see Stage 2).

All datasets must have the same size (number of tokens) and identical syntactic structure in the portion subjected to deformation, ensuring that differences in model behavior are caused solely by semantic changes.

2.3. Stage 2. Isomorphic Deformation and Creation of Deformed Datasets

For each semantic domain, experts compose a reference text describing the domain in a manner typical of the discourse (the composition procedure is presented in the appendix). In the "Map of Being" project, the reference text covers all 36 domains simultaneously.

Subsequently, for each domain, the reference text undergoes three types of isomorphic deformation:

Deformation type	Essence	Implementation method
Reduction (V1)	Semantic removal: the domain's content is reduced to 2–3 extremely abstract, almost vacuous statements.	Elaborated judgments are replaced by short declarative assertions of high generality; concretizing concepts are replaced by abstract ones while preserving the syntactic framework and text length.
Contradiction (V2)	Introduction of logical dissonance by embedding mutually exclusive statements.	Antonymous pairs and negations are placed in key positions; syntactic structure remains unchanged.
Paraphrase (V3)	Complete preservation of the semantic core with a change of stylistic register (e.g., scientific → journalistic).	Synonymous lexical replacement and rhetorical restructuring are performed without altering the order of arguments or introducing new meanings.

All three versions are isomorphic to the original in syntax and length. Paraphrase (V3) serves as a control: if it induces changes comparable to V1 or V2, this indicates that the model is responding to surface lexical features rather than to a semantic shift.

Based on each deformed text, a corresponding training dataset is formed (the construction method is described in Smirnov, 2026a; see also the online appendix). Thus, for each domain, three datasets are obtained — one per deformation type.

2.4. Stage 3. Independent Fine-Tuning: A Family of Equivalent Models

The base LLM (e.g., Qwen2.5-7B) is fine-tuned separately on each prepared dataset with identical hyperparameters (recommended: LoRA/QLoRA, rank=32, alpha=64, learning rate with warmup, early stopping). Each fine-tuning run starts from the same initial base model, without inheriting results from previous training runs.

As a result, a family of models is obtained:

- Control model M_control — trained on the original, unaltered corpus. It represents a "snapshot" of the discourse without distortions.
- Deformed models M_def_{d,t} — for each domain *d* and each deformation type *t* ∈ {V1, V2, V3}. Each such model reflects a worldview in which one specific domain has been purposefully distorted while all other aspects of the corpus are preserved.

Thus, if N domains are studied, the final set includes 3N deformed models plus one control model. All models are independent and equivalent, which prevents the accumulation of distortions and ensures valid comparisons.

2.5. Stage 4. Measurement and Calculation of the Structural Significance Coefficient (SSC)

2.5.1. Expert Panel of Binary Questions with Keys

For each domain, subject-matter experts develop two sets of questions requiring "Yes" or "No" answers:

- Target questions (for each domain — the same number, at least 50; 100–500 recommended). These questions directly pertain to the content of the deformed domain. For each question, experts specify a reference answer (key) reflecting the normative understanding of the domain within the given discourse.
- Control questions (at least 50). These serve a dual function: (a) checking the overall integrity of the model — whether any global disruptions occurred as a result of fine-tuning (the invariant core); (b) including some questions on the research topic not directly

related to the deformed domain, in order to detect possible unexpected systemic effects. Keys are also provided for control questions.

The methodology for composing questions, ensuring inter-expert agreement, and the structure of the invariant core are detailed in the online appendix.

2.5.2. Preliminary Validation

First, the target and control questions are posed to the base LLM (without fine-tuning) and to the control model M_control. It is expected that the proportion of M_control's answers matching the key is significantly higher than that of the base model. This confirms that fine-tuning on the corpus successfully formed a relevant representation of the discourse and that the questions correctly reflect the content of the domains.

If, for any domain, the accuracy of M_control on the target questions (A_c) falls below 0.6, that domain is considered insufficiently represented in the model and is excluded from further analysis, or the corpus and question formulations must be revised.

2.5.3. Testing Deformed Models

Each model from the family M_def_{d,t} is tested on the same target and control questions. The proportion of answers matching the expert key (accuracy) is recorded. Since all questions are binary and the number of target questions is identical across domains, results for different domains can be directly compared.

2.5.4. Calculation of Metrics

Let, for domain *d* and deformation type *t*:

- A_c — accuracy of the control model on the target question set for this domain (A_c ≥ 0.6);
- A_def_t — analogous accuracy for the deformed model M_def_{d,t};
- A_c_ctrl and A_def_t_ctrl — accuracies on the control question set.

Semantic Shift Index (SSI) for deformation *t*:

$$SSI_t = (A_c – A_{def_t}) / A_c$$

SSI quantifies the relative decrease in the proportion of correct answers compared to the control model. This decrease is interpreted as a measure of how strongly the domain distortion altered the model's worldview.

Stability Index (SI):

$$SI_t = |A_c_ctrl - A_def_t_ctrl|$$

Acceptable degradation threshold: $SI_t < 0.05$. Exceeding this threshold indicates that the model has diverged unacceptably — its answers to control questions differ substantially from the control version, and the experiment for that deformation type is considered invalid.

For paraphrase (V3), the expected outcome is $SSI_V3 \approx 0$ and $SI_V3 < 0.05$, confirming that the model is sensitive specifically to the semantic, rather than the stylistic, aspect of the deformation.

2.5.5. Structural Significance Coefficient (SSC)

$$SSC_d = SSI_V1 + SSI_V2$$

SSC takes values from 0 (deformation had no effect on the answers) to 2 (complete loss of the ability to give answers that are correct from the standpoint of the original discourse). A higher SSC indicates greater structural significance of the domain. For clarity, it is recommended to visualize the SSI values for all domains across both deformation types, which reveals the overall pattern of "rigidity" within the corpus's semantic structure.

Note on statistical reliability: Increasing the number of questions in the panel narrows the confidence intervals for the estimates A_c and A_def_t, thereby enhancing the reliability of the SSC computation. It is recommended to choose the question panel size such that the expected error does not exceed 0.05. Repeated testing of all models can also be carried out to assess the stability of the results.

3. COMPARATIVE ANALYSIS: THE PLACE OF E-CDP AMONG EXISTING APPROACHES

Characteristic	RepE (Zou et al., 2023)	Causal Tracing / Model Editing	Concept Erasure (Belrose et al.)	**E-CDP (our method)**
Mode of intervention	Manipulation of activation vectors	Direct weight editing	Linear removal of information from representations	Independent fine-tuning on deformed versions of a corpus

Level of intervention	Activations (superficial)	Weights (deep, localized)	Representations (intermediate)	Semantic domain in the corpus (deep, systemic)
Goal of intervention	Controlling model behavior	Correcting facts	Removing unwanted information	Measuring the structural significance of domains in a corpus
What is measured	Change in generation	Success of fact replacement	Success of concept removal	Systemic response of the model to semantic deformation of the corpus
Key innovation	Linear controllers	Localization of causality	Linear projection	Isomorphic deformation + protocol of binary questions with keys

Thus, the E-CDP protocol does not compete with the listed methods; rather, it addresses a different task: a transition from analyzing the internals of a model to experimentally investigating discursive structures using LLMs as a tool.

4. CONCLUSION: VALUE AND LIMITS OF APPLICABILITY

Value of the protocol:

- It enables a transition from passive observation to active experimentation on textual cultures.
- It provides a transparent, reproducible metric (SSC) and a binary question protocol accessible to humanities scholars without deep technical training.
- It ensures rigorous validity control through a triple system: paraphrase, control questions, and the stability index.

Limits of applicability:

- Prohibition against psychologizing the model. Changes in LLM answers result from token probability redistribution, not from manifestations of “cognitive dissonance” or “intentionality.”

- Dependence on the corpus and expert questions. The results are entirely determined by the composition of the training corpus and the quality of the formulated questions.
- Technical sensitivity. SSC comparisons are valid only within a single experimental design (the same LLM architecture, identical hyperparameters, and corpus augmentation procedure).
- Non-equivalence to human consciousness. An LLM is a statistical snapshot of textual culture, and extrapolating conclusions to human psychology requires separate empirical studies with live respondents.

5. REFERENCES

1. Smirnov A.U. Architectonics of semantic representations: an experimental protocol for identifying structurally significant domains in large language models // Zenodo Preprint. — 2026a. — DOI: 10.5281/zenodo.21205430.
2. Smirnov A.U., Kireeva A.A. Method of controlled deformations of latent spaces of large language models: a new tool for computer experiment in digital humanities // Zenodo Preprint. — 2026b. — (in press).
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.

The full code, instructions for corpus augmentation, formation of expert panels, examples of deformed texts, and current experimental results are available in the E-CDP online appendix: https://crcg.ru/e-cdp-appendix/.