https://doi.org/10.5281/zenodo.22077130

# Картография ментальной топографии минимизации свободной энергии: вычислительный подход на основе анализа нарративных траекторий больших языковых моделей (LLM)

А.Ю. Смирнов, А. П. Зеленцов, А.А. Киреева
АНО «Центр научных исследований
в сфере профориентации и психологии труда», Москва
Email: smirnov@crcg.ru

**Аннотация**
Как мозг защищает свои убеждения от противоречащих фактов? Нейробиология отвечает: через минимизацию свободной энергии (FEP, К. Фристон), но не знает *содержания* этих защит. Лингвистика умеет анализировать темы текстов, но не видит *динамики* когнитивных трюков. Мы предлагаем междисциплинарный метод, извлекающий из текстовых массивов не статические темы, а динамические графы переходов: от исходных априорных убеждений (*priors*) через точку когнитивного трения к финальному сглаживанию. С помощью LLM в режиме глубокого рассуждения мы фильтруем маркеры когнитивного усилия, выделяем операторы защиты и классифицируем их по энергозатратам (внешняя vs внутренняя атрибуция). Выдвигается строгая фальсифицируемая гипотеза о степенном ранговом распределении операторов, где низкозатратные (экстернализация, эскапизм) монотонно доминируют над высокозатратными (рефлексия, перестройка *priors*). Метод включает стратифицированную выборку по трём жанрам, двухпроходное семантическое сжатие с опорой на таксономию (Канеман, А. Фрейд) и алгоритм второй производной для выделения ядра защит. Проект переводит качественные гипотезы в измеримые переменные и создаёт первый эмпирический атлас коллективных когнитивных алгоритмов.
**Ключевые слова:** принцип свободной энергии, предиктивное кодирование, когнитивные защиты, большие языковые модели, нарративный анализ, частотная иерархия, фальсифицируемость.

# Cartography of the Mental Topography of Free-Energy Minimization: A Computational Approach Based on Narrative-Trajectory Analysis of Large Language Models (LLMs)

A.U. Smirnov, A.P. Zelentsov, A.A. Kireeva
ANPO «Center for Scientific Research

in Career Guidance and Labor Psychology»
Moscow, Russian Federation
Email: smirnov@crcg.ru

**Abstract**
How does the brain defend its beliefs against contradictory facts? Neuroscience answers: through free-energy minimization (FEP, K. Friston), but it does not know the content of these defenses. Linguistics can analyze the topics of texts, but it does not capture the dynamics of cognitive maneuvers. We propose an interdisciplinary method that extracts from text corpora not static topics but dynamic transition graphs: from initial prior beliefs (priors) through a point of cognitive friction to final smoothing. Using LLMs in a deep-reasoning mode, we filter markers of cognitive effort, identify defense operators, and classify them by energy cost (external vs. internal attribution). A rigorous falsifiable hypothesis is advanced concerning a power-law rank distribution of operators, in which low-cost operators (externalization, escapism) monotonically dominate high-cost operators (reflection, restructuring of priors). The method includes stratified sampling across three genres, two-pass semantic compression based on a taxonomy (Kahneman, A. Freud), and a second-derivative algorithm for identifying the core of defenses. The project translates qualitative hypotheses into measurable variables and creates the first empirical atlas of collective cognitive algorithms.
**Keywords:** free energy principle, predictive coding, cognitive defenses, large language models, narrative analysis, frequency hierarchy, falsifiability.

## 1. ВВЕДЕНИЕ

Каждый день мы сталкиваемся с фактами, которые противоречат нашим убеждениям. В идеальном мире мы бы корректировали свои взгляды. В реальности мозг поступает иначе: он виртуозно искажает восприятие, придумывает оправдания, ищет внешних врагов — лишь бы сохранить внутреннюю стабильность. Почему?

Согласно *принципу свободной энергии* (Free Energy Principle, FEP), сформулированному Карлом Фристоном [Friston, 2010], мозг — это не пассивный приёмник, а активная «машина предсказаний» (predictive processing) [Clark, 2013]. Он постоянно строит вероятностную модель мира и сравнивает её с сенсорными сигналами. Когда возникает рассогласование (ошибка предсказания), система испытывает когнитивное напряжение — «свободную энергию» — и стремится его минимизировать, чтобы сохранить гомеостаз. Этот фундаментальный процесс лежит в основе всех когнитивных защит.

Несмотря на эвристическую силу FEP, его эмпирическая верификация столкнулась с эпистемологическим барьером, разделившим исследователей на две относительно изолированные области:

- **Нейробиологический редукционизм** успешно описывает термодинамические и синаптические процессы на микро- и

мезоуровнях (ЭЭГ, фМРТ, паттерны нейронных ансамблей). Однако уравнения Фристона принципиально не различают, *какие именно нарративы, объяснения или самообманы* использует конкретный индивид для снижения напряжения. Это ограничение можно назвать «смысловой слепотой» математического аппарата.

- **Корпусная лингвистика и вычислительная социология**, напротив, накопили мощные инструменты для тематического моделирования текстов и построения статических семантических карт (включая протокол Смирнова, 2026, на русскоязычном массиве). Однако эти методы остаются преимущественно статическими: они фиксируют, *о чём* пишут авторы, но не фиксируют *как* в тексте разворачиваются логические переходы, маскирующие аномалии и защищающие базовые убеждения.

Наш подход преодолевает этот разрыв. Мы утверждаем, что **текст — это материальный след работы механизмов минимизации свободной энергии**. Когда индивид сталкивается с аномалией, его мозг экстернализует процесс сглаживания в язык через защитные нарративы. Анализируя не отдельные слова, а *логические переходы* (от веры к сомнению и обратно к стабильности) в огромных текстовых массивах, мы можем перевести абстрактную математику Фристона в строгую, измеримую систему координат.

В этом введении мы: (1) формализуем систему гипотез, (2) описываем операционализацию, включая фильтр «точек когнитивного трения», (3) представляем архитектуру вычислительного пайплайна, (4) обсуждаем механизмы контроля систематических искажений. В итоге мы предлагаем первый в мире эмпирический атлас коллективных когнитивных алгоритмов, который можно использовать в психологии, социологии и при создании адаптивных ИИ-систем.

## 2. ТЕОРЕТИЧЕСКИЕ ГИПОТЕЗЫ

Для перевода интуиций о «ленивом мозге» в проверяемые количественные утверждения мы вводим два уровня гипотез:

**Основная гипотеза ($H_1$ — дескриптивная):**

Множество вербальных паттернов, используемых носителями языка для минимизации ошибки предсказания (свободной энергии) на макросоциальном уровне, не является хаотичным. Оно ограничено конечным числом инвариантных динамических операторов, которые образуют устойчивую частотно-иерархическую структуру, инвариантную по отношению к индивидуальным психофизиологическим различиям авторов.

**Вторичная, фальсифицируемая гипотеза ($H_2$ — предсказательная):**

Распределение динамических операторов в глобальном текстовом корпусе подчинено закону термодинамической экономии когнитивных ресурсов. Частотная плотность операторов первого уровня (низкие энергозатраты, вектор внешней атрибуции: экстернализация ошибки,

эскапизм, рационализация) будет строго монотонно доминировать над плотностью операторов второго уровня (высокие энергозатраты, вектор внутренней атрибуции: рефлексия, структурное обновление *priors*) во всех исследуемых лингвистических стратах.

*Математическая операционализация $H_2$:* если мы ранжируем операторы по частоте встречаемости (r), то функция рангового распределения P(r) должна демонстрировать степенную зависимость (аналог закона Ципфа–Мандельброта) с резким изломом в точке перехода между низкозатратными защитами и высокозатратными инсайтами. Точка перегиба определяется с помощью алгоритма второй производной логарифмического рангового распределения (см. раздел 4.2).

Это делает $H_2$ строго фальсифицируемой: если в каком-либо крупном массиве (например, в философской эссеистике) рефлексивные операторы окажутся частотнее примитивных — гипотеза будет опровергнута, и мы пересмотрим биологические основания иерархии.

## 3. ОПЕРАЦИОНАЛИЗАЦИЯ

### 3.1. Отказ от «Картезианского театра»

Ключевая методологическая уязвимость — риск анализировать *весь* текст как непрерывный поток сознания (иллюзия «Картезианского театра»). Если ИИ будет обрабатывать текст линейно, он утонет в шуме: жанровые штампы, стилистические фигуры, фатические обороты. Например, союз «но» в художественной прозе часто выполняет композиционную функцию (*«День был солнечный, но холодный»*) и не свидетельствует о реальном когнитивном усилии персонажа.

Мы вводим **строгий структурно-лингвистический фильтр**: анализ активируется **исключительно в точках «когнитивного трения»** — синтаксических узлах, где зафиксировано столкновение внутренней модели с аномалией. Прокси-индикаторами такого трения служат **эксплицитные маркеры когнитивного усилия** — закрытый класс вводных слов, союзов и предикативных конструкций: *«однако», «на самом деле», «тем не менее», «вопреки ожиданиям», «между тем», «с другой стороны»* и их исторические эквиваленты.

При обнаружении такого маркера модель изолирует контекстное окно (1–2 предложения до и после) и фиксирует **динамическую траекторию перехода** в формате структурированного объекта (JSON):

```
{
"priors": "Руководитель уверен, что команда полностью
поддерживает его стратегию.",
"conflict": "Из отчётности следует, что фактические показатели
расходятся с плановыми на 20 %.",
"operator": "Экстернализация ошибки: снижение результатов
объяснено некомпетентностью внешнего подрядчика.",
"cost_level": "low"
}
```

### 3.2. Классификация энергозатрат

Чтобы исключить субъективный биас при оценке «сложности» оператора, мы задаём формальный структурный критерий, основанный на векторе атрибуции:

- **Низкий энергетический порог (Low Cost)** — оператор перекладывает причину сбоя на внешний объект, агента, институцию или объявляет сам входящий сигнал ложным. С точки зрения предиктивного кодирования, это требует минимального изменения синаптических весов внутренней модели.
- **Высокий энергетический порог (High Cost)** — оператор требует изменения собственных базовых *priors*, признания амбивалентности, фиксации местоимений первого лица в пассивном залоге при описании личной ошибки или радикального расширения синтаксического вложения для удержания парадокса.

Эта бинарная классификация закладывается в системный промпт на этапе экстракции и служит основой для последующего рангового анализа.

## 4. АРХИТЕКТУРА ВЫЧИСЛИТЕЛЬНОГО ПАЙПЛАЙНА

Мы реализуем **четырёхэтапный сквозной пайплайн**, автоматизирующий процесс от сырых текстов до статистических выводов. Общая схема:

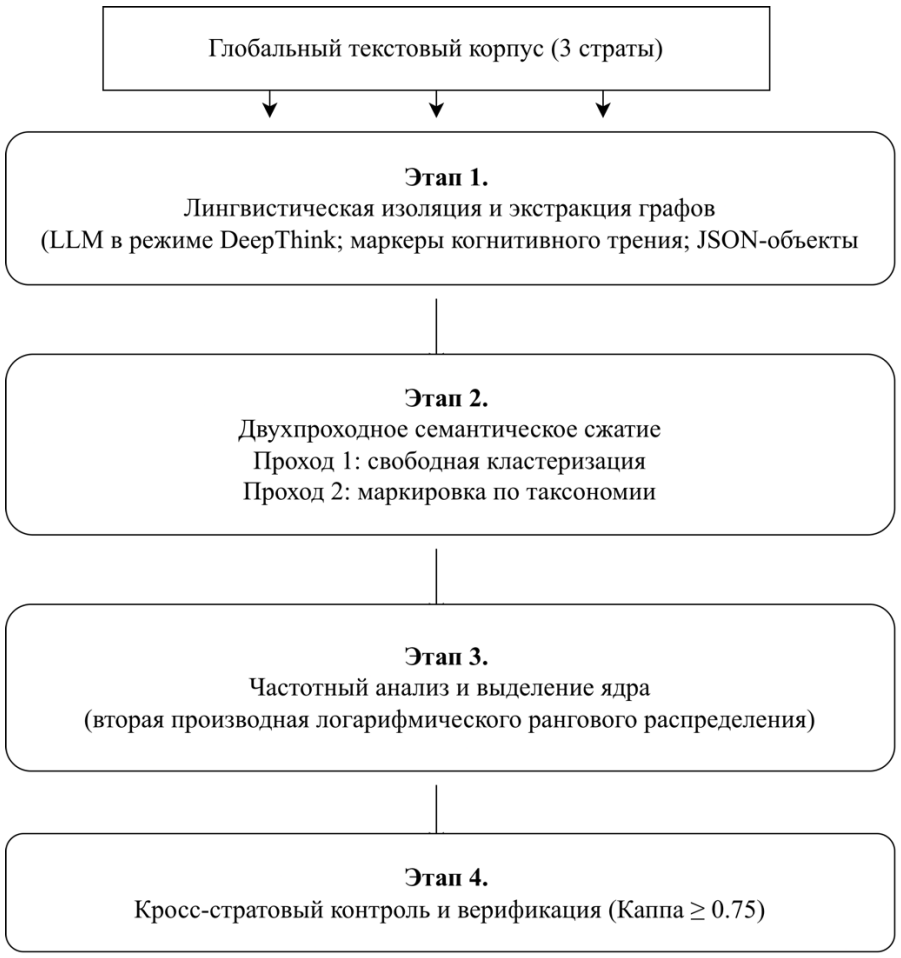


Рисунок 1. Блок-схема пайплайна обработки текстовых данных

### 4.1. Этап 1. Изоляция и экстракция

На вход подаются детокенизированные массивы текстов, разделённые на три независимые страты (см. раздел 5). Большая языковая модель класса *Reasoning* (например, DeepSeek-R1) выполняет поиск маркеров когнитивного трения. При обнаружении маркера модель изолирует окно и извлекает JSON-граф перехода. Системный промпт содержит чёткие инструкции с примерами для каждого из четырёх полей.

### 4.2. Этап 2. Двухпроходное гибридное семантическое сжатие

Прямая кластеризация сырых эмбеддингов неэффективна: фразы *«происки масонов»* и *«колдовство ведьм»* имеют разные векторы, хотя выполняют одинаковую защитную функцию. Поэтому мы применяем двухпроходную схему:

1. **Свободная кластеризация** — LLM группирует извлечённые операторы по смыслу (без внешних шаблонов), формируя укрупнённые формулы объяснения.
2. **Маркировка по референтной таксономии** — полученные кластеры принудительно сопоставляются с разработанной матрицей макро-классов (см. табл. 1). Операторы, не вошедшие в таксономию, выносятся в категорию «свободных кластеров» для последующего качественного анализа — это позволяет зафиксировать новые, ранее не описанные паттерны.

**Таблица 1. Референтная когнитивно-функциональная таксономия** (связывает психологические защиты (А. Фрейд), когнитивные искажения (Д. Канеман) и нейробиологическую функцию в рамках FEP)

| Макро-класс (формула сжатия) | Нейробиологическая функция в FEP | Психологический базис | Структурный маркер и вектор энергозатрат |
|---|---|---|---|
| **1. Экстернализация ошибки** [$EXT_BIAS$] | Модель мира объявляется непогрешимой. Ошибка предсказания приписывается внешнему латентному агенту. | Проекция, фундаментальная ошибка атрибуции | Перекладывание вины на среду/объект. **Низкие затраты**. |
| **2. Когнитивный эскапизм** [$ESC_INVAL$] | Входящий аномальный сигнал объявляется ложным, случайным или незначительным. | Отрицание, изоляция аффекта / иллюзия иллюзорности | Обесценивание фактов, игнорирование данных. **Низкие затраты**. |
| **3. Рационализация** [$RAT_INTEG$] | Мозг ретроспективно искажает историю, интегрируя аномалию в изначальный план. | Рационализация / эффект ретроспективного детерминизма | Конструкции «это к лучшему», «так и задумывалось». **Низкие затраты**. |
| **4. Компенсаторное упрощение** [$REG_DOGMA$] | При критическом диссонансе мозг совершает откат к жёстким, сверхобобщённым догматическим моделям. | Регрессия / эвристика доступности, Система 1 | Черно-белое мышление, лозунги, отказ от нюансов. **Низкие затраты**. |
| **5. Смещение ценностного фокуса** [$FRAM_DISP$] | Признание локальной ошибки сглаживается немедленным переключением на глобальный успех в иной сфере. | Замещение, реактивное образование / эффект фрейминга | Конструкции «да, но зато…», балансировка амбивалентности. **Низкие затраты**. |
| **6. Модификация Priors (Инсайт)** [$INS_MOD$] | **Единственный класс II уровня.** Направлен на радикальную, | Интеллектуальный катарсис / работа Системы 2 | Признание личной ошибки, изменение priors, |

| | | | |
|---|---|---|---|
| | энергозатратную перестройку внутренней модели под реальность. | | глубокий пассив. **Высокие затраты**. |

### 4.3. Этап 3. Частотный анализ и выделение ядра

Для каждого сжатого оператора вычисляется его частота в корпусе. Затем строится ранговое распределение и применяется алгоритм второй производной логарифмической функции рангового распределения (аналогично протоколу [Смирнов, 2026]). Точка максимального перегиба кривой объективно отделяет фундаментальное «тематическое ядро когнитивных защит» (операторы с высокой частотой и устойчивостью) от случайного стохастического «хвоста» индивидуальных речевых девиаций.

### 4.4. Этап 4. Контроль искажений и верификация

На этом этапе проводится кросс-стратовое сравнение (хи-квадрат, расстояние Кульбака–Лейблера) для изоляции инвариантных паттернов от жанровых и исторических артефактов. Кроме того, случайная подвыборка (10–15% от общего числа извлечённых графов) передаётся на слепую экспертизу когнитивным психологам и лингвистам. Рассчитывается коэффициент Каппа Коэна (κ) между ИИ-разметкой и экспертной оценкой; пайплайн считается валидным при $\kappa \geq 0.75$.

## 5. МЕХАНИЗМЫ КОНТРОЛЯ СИСТЕМАТИЧЕСКИХ СМЕЩЕНИЙ

Любое масштабное исследование сталкивается с искажениями. В таблице 2 перечислены основные риски и наши компенсаторные меры с количественными критериями.

**Таблица 2. Ограничения и способы их преодоления**

| Ограничение | Путь преодоления |
|---|---|
| **Жанровое смещение** (художественная проза, публицистика, соцсети диктуют разные риторические нормы) | **Стратифицированная выборка** с равными долями трёх подкорпусов (≈ 33% каждый). Сравнение распределений между стратами позволяет выделить инвариантные когнитивные паттерны (критерий согласия $\chi^2$, $p < 0.05$ после поправки Бонферрони). |
| **Временной и исторический сдвиг** (язык и маркеры меняются) | **Хронологическое аннотирование** (XIX, XX, XXI века) и скользящая калибровка системных промптов для каждого |

| | |
|---|---|
| | исторического слоя с учётом норм синтаксического вложения. |
| **Галлюцинации и субъективность LLM** (модель может достраивать логику там, где её нет) | 1) **Золотой стандарт разметки**: 10–15% случайной подвыборки проходит перекрёстную слепую экспертизу (не менее 3 экспертов). Коэффициент Каппа ≥ 0.75. 2) Использование режима *DeepThink* с требованием обоснования каждого выделенного оператора. 3) Двухпроходное сжатие с внешним таксономическим контролем. |
| **Культурно-языковое смещение** (первая фаза — русскоязычный корпус) | Исследование позиционируется как пилотное; в дальнейшем планируется расширение на англоязычные, азиатские и латиноамериканские базы. Промежуточные результаты интерпретируются как гипотезы для кросс-культурной проверки. |

## 6. ОЖИДАЕМЫЕ РЕЗУЛЬТАТЫ И НАУЧНАЯ ЗНАЧИМОСТЬ

Успешная реализация проекта впервые позволит:

- Получить **эмпирическую таксономию защитных операторов**, ранжированную по частоте и когнитивной сложности. Это превращает умозрительные философские категории в измеримые переменные.
- **Количественно подтвердить или опровергнуть** гипотезу об энергетической иерархии защит ($H_2$). Если подтвердится, мы получим мощный аргумент в пользу биологической детерминированности предпочтения простых объяснений сложным.
- Создать **методологический инструмент** для анализа динамики убеждений, применимый в социологии (изучение поляризации), политологии (анализ пропаганды), маркетинге и психотерапии.
- Сгенерировать **новые гипотезы для нейробиологов**: выделенные кластеры операторов могут стать стимульным материалом для фМРТ-исследований, проверяющих, активируют ли разные операторы различные нейронные сети (например, сеть пассивного режима vs. сеть выполнения задач).

В отличие от существующих подходов, мы выносим предиктивное кодирование «наружу» — в лингвистическое поле. Вместо абстрактных уравнений мы получаем содержательную архитектуру коллективного разума, зафиксированную в культуре.

## 7. ЗАКЛЮЧЕНИЕ

Мы предложили методологию, которая связывает термодинамический аппарат принципа свободной энергии с эмпирическим анализом нарративных траекторий в глобальных текстовых корпусах. Её сила — в **фальсифицируемости**: выдвинуто рискованное предсказание о монотонном доминировании низкозатратных когнитивных защит, которое может быть опровергнуто данными. Если это произойдёт, мы будем вынуждены пересмотреть биологические основания предлагаемой иерархии.

Мы приглашаем коллег к дискуссии и сотрудничеству, особенно в части кросс-культурной валидации и уточнения референтных таксономий. Реализация проекта намечена на 2026–2028 годы с публикацией промежуточных наборов данных и кода в открытом доступе. Это первый шаг к созданию объективного атласа алгоритмов мышления Homo Sapiens — карты, которая поможет нам лучше понять самих себя.

## Список литературы


1. Смирнов А. Ю. (2026). Эмпирическое обоснование тематического ядра личностного самоопределения: результаты реализации протокола на русскоязычном корпусе (Версия 1.01). Zenodo. DOI: 10.5281/zenodo.19691681.
2. Bail, C. A. (2021). *Terrified: How anti-Muslim fringe organizations became mainstream*. Princeton University Press.
3. Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204. DOI: 10.1017/S0140525X1200047X.
4. Freud, A. (1937). *The Ego and the Mechanisms of Defence*. London: Hogarth Press.
5. Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. DOI: 10.1038/nrn2787.
6. Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.