УДК 004.8

# ОЦЕНКА НЕОПРЕДЕЛЕННОСТИ В ТРАНСФОРМЕРНЫХ ЦЕПЯХ НА ОСНОВЕ ПРИНЦИПА СОГЛАСОВАННОСТИ ЭФФЕКТИВНОЙ ИНФОРМАЦИИ

**А. А. Красновский**[0000-0001-6842-7340]

*Университет Иннополис, г. Иннополис, Россия*

a.a.krasnovsky@gmail.com

***Аннотация***

Механистическая интерпретируемость позволяет выявлять функциональные подграфы в больших языковых моделях (LLM), известные как трансформерные цепи (Transformer Circuits, TC), которые реализуют конкретные алгоритмы. Однако отсутствует формальный способ, позволяющий за один проход количественно оценить, когда активная цепь ведет себя согласованно и, следовательно, ее состояние может быть признано корректным. Опираясь на ранее предложенную автором пучково-теоретическую формализацию причинной эмерджентности (Krasnovsky, 2025), мы специализируем ее для трансформерных цепей и вводим безразмерную однопроходную оценку согласованности эффективной информации (Effective Information Consistency Score, EICS). EICS сочетает нормализованную несогласованность пучка, вычисляемую из локальных якобианов и активаций, с гауссовским прокси EI для причинной эмерджентности на уровне цепи, полученным из того же состояния прямого прохода. Такая конструкция является прозрачной (white-box), однопроходной и делает единицы измерения явными, так что оценка безразмерна. Представлены практические рекомендации по интерпретации оценки, учету вычислительных затрат (с быстрыми и точными режимами) и анализ простейшего примера для проверки на адекватность.

***Ключевые слова****: механистическая интерпретируемость, трансформерные цепи, теория пучков, причинная эмерджентность, количественная оценка неопределенности, большие языковые модели (LLM).*

## ВВЕДЕНИЕ

Основополагающей целью механистической интерпретируемости является восстановление (или реконструкция) алгоритмических компонентов LLM на мезоуровне («трансформерных цепей») [1, 2]. Эти подграфы («головы внимания», MLP и их пути) связаны с такими задачами, как копирование или индукция [1] и извлечение фактов [3]. После идентификации цепи возникает естественный вопрос: *функционирует ли она согласованно при данном входном сигнале?* Одна и та же модель может ответить на фактический запрос правильно или «галлюцинировать». В рамках настоящей работы предположим, что *степень причинной согласованности* активной цепи различается между этими режимами.

Кроме того, для построения концептуального базиса будем рассматривать идеи теории пучков и причинной эмерджентности, адаптируя их для трансформерных цепей [4—9]. В качестве решения мы предлагаем метрику EICS, которая вычисляется за один прямой проход и дает количественную оценку. Концептуально неопределенность трактуется как потеря причинно-следственной связности: высокая активность при низкой несогласованности означает надежность системы, а обратная картина — ее рискованность. В отличие от подходов к оценке неопределенности (Uncertainty Quantification, UQ) типа «черный ящик» [10—13], EICS идентифицирует конкретные механизмы, ответственные за возникновение неопределенности.

## ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ

**Механистическая интерпретируемость.** Головы индукции и поведение обучения «в контексте» были описаны Олссоном и соавт. [1]. Недавние инструменты «графов атрибуций» (также известные как трассировка цепей) предоставляют процедуры обнаружения подграфов и их валидации [2]. Активно изучаются цепи, связанные со знаниями [3], взаимосвязь между локализацией и редактированием также является предметом всестороннего анализа [14].

**Количественная оценка неопределенности (UQ).** Методы UQ «черного ящика» включают распределенные свободные конформистские предсказания [10], пост-процессинг калибровки [11], глубокие ансамбли [12] и байесовские или

приближенно-байесовские методы тонкой настройки LLM, такие как Laplace-LoRA [13]. Наша задача состоит в обеспечении прозрачности за счет сигнала, генерируемого внутренними цепями.

**Клеточные пучки и причинная эмерджентность.** Клеточные пучки предлагают формализм для объединения локальных линейных отображений в глобально согласованные состояния, где степень несогласованности измеряется при помощи когомологий (кобоундариев) и операторов Лапласа — Ходжа [5, 6]. В свою очередь, причинная эмерджентность дает количественную оценку того, в каких случаях макромасштабные описания системы содержат больше эффективной информации, чем описания на уровне ее составных частей [7—9]. Мы адаптируем этот теоретический аппарат к анализу цепей трансформеров, вводя для этого явные и вычислимые показатели (прокси).

## ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ КОНЦЕПЦИИ

### 1. Трансформерные цепи

Рассмотрим нейросетевую архитектуру «трансформер» как ориентированный ациклический граф (DAG) $G = (V, E)$, где вершины соответствуют головам внимания или MLP, а ребра отражают поток информации. ТС — это подграф $G_M \subseteq G$, который предположительно реализует задачу. Для входа $x$ каждая вершина $v \in V_M$ имеет активацию $a_v \in \mathbb{R}^{d_v}$. Мы будем обозначать как $e = (u \to v) \in E$ ориентированное ребро и использовать локальные линеаризации в наблюдаемых активациях.

### 2. Клеточные пучки на графах и вычислимая несогласованность

Определим клеточный пучок $\mathcal{F}$ на *подлежащей неориентированной* версии $G_M$ (так что 1-коцепи живут на ребрах независимо от направления в DAG) со стеблями $\mathcal{F}(v) = \mathbb{R}^{d_v}$ и отображениями ограничений на ориентированных ребрах, задаваемыми якобианами, вычисленными в текущем состоянии:

$$\rho_e : \mathcal{F}(u) \to \mathcal{F}(v), \qquad \rho_{u\to v} := J_{u\to v} \equiv \left(\frac{\partial f_v}{\partial a_u}\right)_a.$$

Для 0-коцепи (назначения узлов) $s = \{s_v\}$ кобоундарий пучка $\delta^0 : C^0(G; f \to C^1(G; F)$ действует как

$$(\delta^0 s)_{(u\to v)} \;=\; \rho_{u\to v}\, s_u \;-\; s_v. \tag{1}$$

На графе (без 2-ячеек) $\delta^1 = 0$, следовательно, $H^1 \cong C^1/\operatorname{im}\delta^0$. Вместо того чтобы брать неканоническую «норму фактор-пространства», мы используем *нормализованную энергию несогласованности* из наблюдаемых активаций $a = \{a_v\}$:

$$C_{\mathrm{sh}}(G_M, a) \;=\; \frac{\left(\sum_{(u\to v)\in E_M} \|\rho_{u\to v} a_u - a_v\|_2^2\right)^{1/2}}{\varepsilon + \left(\sum_{(u\to v)\in E_M} \|a_u\|_2^2 + \|a_v\|_2^2\right)^{1/2}} \tag{2}$$

с малым $\varepsilon > 0$ для числовой устойчивости. Это безразмерная величина, которая равна $0$ тогда и только тогда, когда $a$ является (шумно) согласованным глобальным сечением. Можно опционально заменить $a$ на оценку наименьших квадратов $\hat{s} = \operatorname{argmin}_s \sum_e \| \rho_e s_u - s_v \|^2$; обе операции — это однопроходные вычисления (на основе произведения вектора на якобиан и якобиана на вектор: VJP/JVP).

**Вычисление JVP, инициированное узлами (эффективность).** Чтобы эффективно вычислить (2), мы используем схему JVP, инициированную узлами: для каждого исходного узла $u \in V_M$ выполняется один JVP с семенем $a_u$ для вычисления всех исходящих остаточных членов $\rho_{u\to v} a_u$ за один проход. Это уменьшает сложность с JVP для каждого ребра до $O(|V_M|)$ JVP (обычно 10—20 для средних цепей), сохраняя точное определение в (2).

### 3. Однопроходный гауссовский прокси для оценки эффективной информации

Истинная эффективная информация (EI) определяется интервенциями. Чтобы получить однопроходный прокси, предполагаем малыми изотропные локальные интервенции в текущем состоянии и аппроксимируем каждое отображение его якобианом. Для линейного отображения $y = Jx + \xi$ с изотропным $x$ единичной дисперсии и малым аддитивным шумом $\xi$ взаимная информация (в натах) пропорциональна $\frac{1}{2}\log\det(I + \alpha J^{\top} J)$ с масштабом $\alpha > 0$. Поэтому определим

$$\mathrm{EI}_G(J) := \tfrac{1}{2}\log\det(I + \alpha J^{\top} J),\;\; \Delta\mathrm{EI}_G(G_M) := \mathrm{EI}_G(J_M) \;-\; \textstyle\sum_{v\in V_M} \mathrm{EI}_G\,(J_v), \tag{3}$$

где $J_M$ — макро-якобиан от входов цепи к ее выходам (полученный путем линеаризации составного подграфа). Далее используем положительную часть $\Delta\mathrm{EI}_G^+ = \max(0, \Delta\mathrm{EI}_G)$ и необязательную нормализацию $\widetilde{\Delta\mathrm{EI}}_G := \Delta\mathrm{EI}_G^+ / \big(\varepsilon + \mathrm{EI}_G(J_M)\big)$, чтобы удерживать оценки в $[0,1)$.

**4. Гауссовская прокси-оценка эффективной информации – вывод и замечания по реализации**

**Определение модели.** Рассмотрим локальное, линейное описание цепи вокруг наблюдаемого состояния прямого прохода. Пусть $x \in \mathbb{R}^n$ обозначает малое стохастическое вмешательство на входах цепи, а $y \in \mathbb{R}^m$ — выходы цепи. Мы аппроксимируем

$$y \;=\; J\,x + \xi, \qquad x \sim \mathcal{N}(0, \sigma_x^2 I_n),\;\; \xi \sim \mathcal{N}\big(0, \sigma_\xi^2 I_m\big). \tag{4}$$

**Взаимная информация.** Для линейного гауссовского канала с независимыми гауссовскими входом и шумом

$$I(x;y) \;=\; \tfrac{1}{2}\log\det\big(I_m + \sigma_x^2/\sigma_\xi^2\, JJ^\top\big) \;=\; \tfrac{1}{2}\log\det\big(I_n + \sigma_x^2/\sigma_\xi^2\, J^\top J\big) \tag{5}$$

**Определение прокси.** Мы используем гауссовский прокси EI из (3); для цепи $G_M$ возникновение и его нормализованная положительная часть задаются, как в п. 3.3,

$$\widetilde{\Delta\mathrm{EI}}_G \;=\; \frac{\max(0, \Delta\mathrm{EI}_G)}{\varepsilon + \mathrm{EI}_G(J_M)}. \tag{6}$$

**Инвариантность, чувствительность и приближение при малых α.** $\mathrm{EI}_G(J)$ зависит только от сингулярных значений $J$. Его чувствительность к $\alpha$ равна

$$\frac{\partial}{\partial\alpha}\left(\tfrac{1}{2}\log\det(I + \alpha J^\top J)\right) \;=\; \tfrac{1}{2}\,\mathrm{tr}[(I + \alpha J^\top J)^{-1} J^\top J],$$

и для $\alpha\,\sigma_{\max}^2 \ll 1$

$$\frac{1}{2}\log\det(I + \alpha J^\top J) \;\approx\; \frac{\alpha}{2}\,\|\,J\,\|_F^2.$$

См. также (5) для связи со взаимной информацией линейного гауссовского канала и (4) для постановки линейной модели.

**Вычисление.** Используем разложение Холецкого или собственное разложение для малых $n$; для больших $n$ используем оцениватели лог-детерминанта Hutch++ или Ланцоша только с JVP/VJP-произведениями. Остаточные связи обработаем путем построения линейного блочного оператора или вычисления лог-детерминанта с помощью методов Крылова.

**МЕТОД: ОЦЕНКА СОГЛАСОВАННОСТИ ЭФФЕКТИВНОЙ ИНФОРМАЦИИ**

**1. Определение**

Даны $G_M$, активации $a$ и якобианы ребер $\{\rho_{u\to v}\}$ из одного прямого прохода. Определим

$$\mathrm{EICS}(G_M; a) \;=\; \frac{\widetilde{\Delta\mathrm{EI}}_G(G_M)}{1+C_{\mathrm{sh}}(G_M,a)}. \tag{7}$$

Высокие значения EICS означают (i) сильную макроуровневую интеграцию информации относительно частей и (ii) низкое внутреннее несогласие на ребрах.

**Почему это устраняет предыдущие проблемы.** 1) Мы никогда не берем норму фактор-пространства $H^1$; а измеряем *энергию несогласия* (2) напрямую и безразмерно. 2) Термин EI – это четко сформулированный гауссовский прокси лог-определителя; единицы измерения – наты, которые становятся безразмерными через нормализацию. 3) DAG не содержат направленных циклов, но несогласованность пучка остается осмысленной на неориентированном 1-скелете.

**2. Практическое использование и интерпретация**

По конструкции $C_{\mathrm{sh}} \geq 0$ и $\widetilde{\Delta\mathrm{EI}}_G \in [0,1)$, следовательно, $\mathrm{EICS} \in [0,1)$. Используем тренировочное подмножество, чтобы выбрать порог $\tau$ (AUROC/F1), а также сообщаем компоненты $1/(1+C_{\mathrm{sh}})$ и $\widetilde{\Delta\mathrm{EI}}_G$, чтобы диагностировать факторы. Для $\alpha$ либо устанавливаем $\alpha = 1$ (априорное отношение сигнал — шум), либо выбираем $\alpha$ так, чтобы значение $\frac{1}{2}\log\det(I + \alpha J_M^{\top} J_M)$ оставалось в целевом межквартильном диапазоне и избегало насыщения.

**ТЕОРЕТИЧЕСКИЕ СВОЙСТВА**

**Допущение 1 (Локальная линейность и ограниченность).** *Вдоль $G_M$ отображения локально линейны, якобианы $\{\rho_e\}$ липшицевы в окрестности $a$ и нормы*

*операторов ограничены. Лапласиан Ходжа пучка* $L = \delta^{0\dagger}\delta^{0}$ *(с внутренним произведением, индуцированным весами ребер) имеет спектральный зазор* $\lambda_2(L) > 0$ [5].

**Утверждение 1 (Вычислимость за один проход).** *При выполнении Допущения 1 как* $C_{sh}(G_M, a)$*, так и* $\widetilde{\Delta EI}_G(G_M)$ *являются детерминированными функциями одного прямого прохода и его произведений якобиана на вектор. Следовательно, для вычисления* $EICS$ *требуется константное (*$O(1)$*) число прямых проходов.*

*Базис доказательства.* $\delta^0$ строится из $\{\rho_e\}$, вычисленных в $a$. Как остаточный член (2), так и термины лог-детерминанта (3) являются функциями этих объектов. Для вычисления не требуется стохастическое моделирование по входным данным.

**Утверждение 2 (Устойчивость к малым возмущениям вне цепи (оценка)).** *Пусть* $y$ *обозначает выходы цепи. Рассмотрим аддитивное внешнее возмущение* $\eta$*, которое входит в* $G_M$ *с усилением не более* $\gamma$ *в норме оператора. При утверждении 1*

$$\| \hat{s} - s^{\star} \| \leq \frac{\gamma}{\lambda_2(L)} \| \eta \|, \qquad \| \Delta y \| \leq \kappa \| \hat{s} - s^{\star} \|$$

*для некоторой локальной константы Липшица* $\kappa$*. В частности, малые значения* $C_{sh}$ *(означающие высокие значения* $\lambda_2(L)$*) приводят к сужению интервалов оценок.*

*О роли* $\lambda_2(L)$. Оценка масштабируется как $1/\lambda_2(L)$. На практике $\lambda_2(L)$ зависит от i) связности и ii) весов ребер, индуцируемых локальными якобианами. Мы рекомендуем: a) *сообщать* $\lambda_2(L)$ для каждой цепи; b) *нормировать веса ребер* по операторным нормам $\rho_{u\to v}$ и c) при необходимости *регуляризовать* $L \leftarrow L + \beta I$, когда эмпирическое $\lambda_2(L)$ близко к нулю, — это ужесточает практическую оценку, не изменяя $C_{\mathrm{sh}}$ или EICS.

## АЛГОРИТМ

### 1. Однопроходный EICS для трансформерной цепи

**Алгоритм 1.** Однопроходный EICS для трансформерной цепи

1: **Вход:** модель $\mathcal{M}$, вход $x$, цепь $G_M = (V_M, E_M)$, масштаб $\alpha > 0$.
2: **Выход:** $\mathrm{EICS}(G_M; a)$.
3: **Прямой проход и активации:** выполняем $\mathcal{M}(x)$ и записываем $\{a_v\}_{v \in V_M}$.
4: **Якобианы ребер:** для каждого $(u \to v) \in E_M$ вычисляем $\rho_{u \to v} = (\partial f_v / \partial a_u)_a$ с помощью VJP/JVP.
5: **Несогласованность пучка:** вычисляем $C_{\mathrm{sh}}(G_M, a)$ по формуле (2). *Реализация:* используем *JVP, инициированные узлами* (одна JVP на исходный узел $u$), чтобы вычислить все $\rho_{u \to v} a_u$ для исходящих ребер.
6: **Гауссовский прокси EI:** строим макро-якобиан $J_M$ и якобианы узлов $\{J_v\}$. Вычисляем $\widetilde{\Delta \mathrm{EI}}_G$ по формулам (3), (6).
**Быстрый режим (ранжирование):** приближение для малых $\alpha$; используем методы Hutch++ или Ланцоша с 4–8 зондами для каждого $J_v$ и 8–12 для $J_M$.
**Точный режим (малые блоки):** вычисляем $\mathrm{logdet}(I + \alpha J^\top J)$ через разложение Холецкого или SVD.
7: **Оценка:** возвращаем $\mathrm{EICS} = \widetilde{\Delta \mathrm{EI}}_G / (1 + C_{\mathrm{sh}})$.

### 2. Вычислительные затраты и масштабирование

**Порядок величины.** С использованием JVP, инициированных узлами, вычисление $C_{\mathrm{sh}}$ требует $O(|V_M|)$ JVP (примерно $10$–$20$ для средних цепей). В *быстром* режиме EI (приближение Фробениуса для малых $\alpha$ + зондирование Hutchinson) общее количество работ автоград обычно составляет $\sim 50$–$200$ JVP/VJP-произведений на ограниченном подграфе, или около $\sim 2$–$6$ эквивалентов прямого прохода. В *точном* режиме EI (большие $\alpha$ или явные факторизации) ожидайте $\sim 5$–$15$ эквивалентов прямого прохода. Расчет выполняется без применения методов Монте-Карло по входным данным: все величины детерминированно вычисляются из одного прямого прохода.

**Параметры масштабирования.** a) Пакетно инициируйте JVP для всех узлов; b) ограничьте EI топ-$k$ сингулярными направлениями через Ланцоша; c) кешируйте промежуточные линейные отображения вдоль TC; d) предпочитайте малые $\alpha$ для задач ранжирования.

## ПРЕДЛАГАЕМАЯ ВАЛИДАЦИЯ

### 1. Протокол оценки

Мы описываем протокол оценки для задачи фактических вопроса–ответа, используя рабочий процесс графа атрибуции [2] для идентификации цепи извлечения фактов $G_{\mathrm{fact}}$. Предлагаем два набора: (A) вопросы с проверяемыми ответами; (B) адверсариальные или провоцирующие галлюцинации подсказки [15]. Ожидается, что у множества $A$ будет более высокий $\mathrm{EICS}$ (низкий $C_{\mathrm{sh}}$, положительное $\widetilde{\Delta\mathrm{EI}}_G$), а у множества $B$ — пониженные оценки. Сравниваемые подходы включают лог-вероятность, энтропию, дисперсию глубоких ансамблей [12] и размеры конформных наборов [10].

### 2. Базовые методы и абляционный анализ

**1) Корреляция активаций на ребрах (EAC).** Средняя корреляция Пирсона между $a_u$ и $a_v$ по $(u \to v) \in E_M$ (по измерениям).

**2) Остаток выравнивания на ребрах (EAR).**

$$\frac{1}{|E_M|} \sum_{(u \to v)} \| \hat{\rho}_{u \to v} a_u - a_v \|_2$$

с оценкой наименьших квадратов по каждому ребру $\hat{\rho}_{u \to v}$ (без связи пучка).

**Абляции.** (A1) $1/(1 + C_{\mathrm{sh}})$ только; (A2) $\widetilde{\Delta\mathrm{EI}}_G$ только.

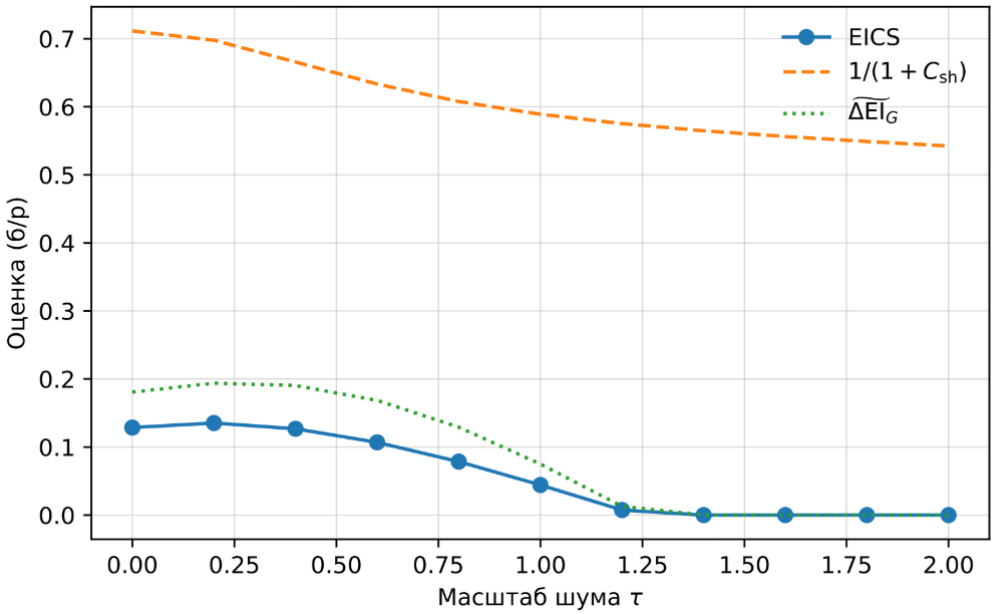


Рис. 1. Простейшая проверка адекватности результатов для цепи из 6 узлов с двумя параллельными ветвями. По мере увеличения шума узлов $\tau$ несогласованность пучка $C_{\mathrm{sh}}$ возрастает (а $1/(1+C_{\mathrm{sh}})$ падает). Мы также уменьшаем выравнивание между ветвями при увеличении $\tau$ (декогерентность ребер), что вызывает уменьшение прокси возникновения $\widetilde{\Delta \mathrm{EI}}_G$ и общей EICS. Кривые показывают средние значения по начальным генераторам. Определения следуют (2), (3) и (7).

**3. Простейшая проверка на адекватность (аналитический и симуляционный протоколы)**

**Настройка.** 6-узловая прямоугольная ТС с линейными блоками и аддитивным гауссовским шумом на ребрах; изменяйте шум на подмножестве ребер. Вычисляйте $C_{\mathrm{sh}}$, $\widetilde{\Delta \mathrm{EI}}_G$ EICS и базовые методы для разных начальных генераторов.

**Аналитическая проверка (малые α).** При $\alpha = \sigma_x^2/\sigma_\xi^2$ увеличение аддитивного шума уменьшает термины EI; шум на ребрах увеличивает остатки в (2). Следовательно, EICS уменьшается с шумом, что соответствует интуиции (см. рис. 1 для простейшей проверки на адекватность, иллюстрирующей эти тенденции).

**ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ОГРАНИЧЕНИЙ ИССЛЕДОВАНИЯ**

**Зависимость от цепи.** EICS имеет смысл лишь настолько, насколько корректно задан $G_M$.

**Линеаризация**. Основанный на якобианах пучок и прокси EI предполагают локальную линейность.

**Стоимость.** JVP, инициированные узлами, быстрые лог-детерминанты и топ-$k$ направления смягчают затраты.

**Место EICS среди методов UQ типа «черный ящик».** EICS предоставляет *механистические* доказательства, дополняя калибровочные и конформные инструменты.

## ЗАКЛЮЧЕНИЕ

Представлена реализация концепции применения пучковой и причинно-эмерджентной перспективы к трансформерным цепям как практическая, однопроходная оценка (EICS). Заменив плохо определенные нормы когомологий на нормализованную энергию несогласованности и определив гауссовский лог-детерминантный прокси EI, получим, что EICS является как вычислимой, так и безразмерной оценкой. Кроме того, детально описаны практические рекомендации, анализ вычислительных затрат для быстрых и точных режимов и простейшая проверка на адекватность.

## ПРИЛОЖЕНИЯ

### Код простейшей проверки на адекватность для рис. 1

В приведенном ниже скрипте воспроизводится рис. 1. Он реализует примитивную двухветвевую модельную цепь, вычисляет $\boldsymbol{C}_{\mathbf{sh}}$ (2), нормализованную прокси-оценку эмерджентности $\boldsymbol{\Delta\widetilde{\mathbf{EI}}_G}$ (3) и $\mathbf{EICS}$ (7) как функции масштаба шума $\boldsymbol{\tau}$.

```
import numpy as np, matplotlib.pyplot as plt
EPS, D, N_SEEDS = 1e-8, 32, 100
TAUS = np.linspace(0.0, 2.0, 11)
alpha, align = 1.0, 0.9  # фиксированное отношение сигнал-шум; высокая начальная согласованность ветвей

def rand_matrix(d, scale=0.8, rng=None):
    rng = np.random.default_rng() if rng is None else rng
    return scale * rng.normal(size=(d, d)) / np.sqrt(d)

def ei_proxy(J, alpha):  # 0.5 * сумма log(1 + alpha * sigma^2)
    s = np.linalg.svd(J, compute_uv=False)
    return 0.5 * np.sum(np.log1p(alpha * (s**2)))
```

```
def build_branch_mats(D=32, align=0.9, rng=None):
    rng = np.random.default_rng(123) if rng is None else rng
    U = rand_matrix(D, 0.8, rng); A = rand_matrix(D, 0.9, rng); W =
rand_matrix(D, 0.9, rng)
    W13 = U;  W23 = (1-align)*rand_matrix(D,0.8,rng) + align*U
    W34 = A;  W35 = (1-align)*rand_matrix(D,0.9,rng) + align*A
    W46 = W;  W56 = (1-align)*rand_matrix(D,0.9,rng) + align*W
    return W13, W23, W34, W35, W46, W56

def metrics_at_tau(tau, rng):
    W13,W23,W34,W35,W46,W56 = build_branch_mats(D, align, rng)
    # Декогерентность ребер: уменьшаем согласованность между ветвями с ростом
tau
    h = min(1.0, tau/2.0)
    nrg = np.random.default_rng(rng.integers(10**9))
    W56 = (1-h)*W56 + h*rand_matrix(D, 0.9, nrg)
    W35 = (1-h)*W35 + h*rand_matrix(D, 0.9, nrg)

    # Две параллельные подцепи (части), макро-оператор — их сумма
    Jb1 = W46 @ W34 @ W13
    Jb2 = W56 @ W35 @ W23
    JM  = Jb1 + Jb2

    EI_macro = ei_proxy(JM, alpha)
    EI_parts = ei_proxy(Jb1, alpha) + ei_proxy(Jb2, alpha)
    dEI_g = max(0.0, EI_macro - EI_parts) / (EPS + EI_macro)

    a1 = rng.normal(size=D); a2 = rng.normal(size=D)
    a3 = W13@a1 + W23@a2; a4 = W34@a3; a5 = W35@a3; a6 = W46@a4 + W56@a5
    a1o = a1 + tau*rng.normal(size=D); a2o = a2 + tau*rng.normal(size=D)
    a3o = a3 + tau*rng.normal(size=D); a4o = a4 + tau*rng.normal(size=D)
    a5o = a5 + tau*rng.normal(size=D); a6o = a6 + tau*rng.normal(size=D)

    edges = [(a1o,a3o,W13),(a2o,a3o,W23),(a3o,a4o,W34),
             (a3o,a5o,W35),(a4o,a6o,W46),(a5o,a6o,W56)]
    num = den = 0.0
    for au,av,W in edges:
        r = W@au - av; num += r@r; den += au@au + av@av
    Csh = np.sqrt(num) / (EPS + np.sqrt(den))
    EICS = dEI_g / (1.0 + Csh)
    return Csh, dEI_g, EICS

C_m, d_m, S_m = [], [], []
C_e, d_e, S_e = [], [], []
for tau in TAUS:
    Cs, ds, Ss = [], [], []
    for k in range(N_SEEDS):
```

```
        rng = np.random.default_rng(1000 + k)
        Csh, dEIg, EICS = metrics_at_tau(tau, rng)
        Cs.append(Csh); ds.append(dEIg); Ss.append(EICS)
    Cs, ds, Ss = map(np.array, (Cs,ds,Ss))
    C_m.append(Cs.mean()); d_m.append(ds.mean()); S_m.append(Ss.mean())
    C_e.append(Cs.std(ddof=1)/np.sqrt(N_SEEDS))
    d_e.append(ds.std(ddof=1)/np.sqrt(N_SEEDS))
    S_e.append(Ss.std(ddof=1)/np.sqrt(N_SEEDS))

TAUS = np.array(TAUS); C_m=np.array(C_m); d_m=np.array(d_m); S_m=np.array(S_m)
C_e=np.array(C_e); d_e=np.array(d_e); S_e=np.array(S_e)
plt.figure(figsize=(6.2,4.2))
plt.plot(TAUS, S_m, '-o', label='EICS')
invC = 1.0/(1.0 + C_m)
plt.plot(TAUS, invC, '--', label=r'$1/(1+C_{\mathrm{sh}})$')
plt.plot(TAUS, d_m, ':', label=r'$\widetilde{\Delta \mathrm{EI}}_G$')
plt.xlabel(r'Масштаб шума $\tau$'); plt.ylabel('Оценка (безразмерно)')
plt.title('Простейшая проверка на адекватность: влияние шума на EICS и компо-
ненты')
plt.grid(True, linewidth=0.5, alpha=0.5); plt.legend(frameon=False)
plt.tight_layout(); plt.savefig('fig_toy_noise_curve.pdf',
bbox_inches='tight')
```

**СПИСОК ЛИТЕРАТУРЫ**

1. *Olsson C., Elhage N., Nanda N., et al.* In-context Learning and Induction Heads. 2022. arXiv : 2209.11895.

2. *Anthropic*. Circuit Tracing / Attribution Graphs: Methods & Applications: Transformer Circuits Team. 2025. Access mode: https://transformer-circuits.pub/2025/attribution-graphs/ (дата обращения: 2025-08-20).

3. *Yao Y., Zhang N., Xi Z., Wang M., Xu Z., Deng S., and Chen H.* Knowledge Circuits in Pretrained Transformers // Advances in Neural Information Processing Systems (NeurIPS). 2024. Vol. 37. P. 118571–118602.

4. *Krasnovsky A.A.* Sheaf-Theoretic Causal Emergence for Resilience Analysis in Distributed Systems. 2025. arXiv: 2503.14104.

5. *Hansen J., Ghrist R.* Toward a Spectral Theory of Cellular Sheaves // Journal of Applied and Computational Topology. 2019. Vol. 3, No. 4. P. 315–358.

6. *Robinson M.* Topological Signal Processing. Springer, 2014.

7. *Rosas F.E., Mediano P.A.M., Jensen H.J., Seth A.K., Barrett A.B., Carhart-Harris R.L., and Bor D.* Reconciling Emergences: An Information-Theoretic Approach to Identify Causal Emergence in Multivariate Data // PLOS Computational Biology. 2020. Vol. 16, No. 12. P. e1008289.

8. *Tononi G., Sporns O.* Measuring Information Integration // BMC Neuroscience. 2003. Vol. 4. P. 31.

9. *Oizumi M., Albantakis L., Tononi G.* From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 // PLOS Computational Biology. 2014. Vol. 10, No. 5. P. e1003588.

10. *Angelopoulos A.N., Bates S.* A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. 2021. arXiv: 2107.07511.

11. *Guo C., Pleiss G., Sun Y., and Weinberger K.Q.* On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR. 2017. P. 1321–1330.

12. *Lakshminarayanan B., Pritzel A., Blundell C.* Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30.

13. Bayesian Low-rank Adaptation for Large Language Models (Laplace-LoRA). 2023. ICLR 2024 version. arXiv: 2308.13111.

14. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models / Hase P., Bansal M., Kim B., and Ghandeharioun A. // Advances in Neural Information Processing Systems (NeurIPS). 2023. Vol. 36. P. 17643–17668.

15. *Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2025. Vol. 43, No. 2. P. 1–55.

# MEASURING UNCERTAINTY IN TRANSFORMER CIRCUITS WITH EFFECTIVE INFORMATION CONSISTENCY

**A. A. Krasnovsky**[0000-0001-6842-7340]

*Innopolis University, Innopolis, Russia*

a.a.krasnovsky@gmail.com

***Abstract***

Mechanistic interpretability has identified functional subgraphs within large language models (LLMs), known as Transformer Circuits (TCs), that appear to implement specific algorithms. Yet we lack a formal, single-pass way to quantify when an active circuit is behaving coherently and thus likely trustworthy. Building on the author's prior sheaf-theoretic formulation of causal emergence (Krasnovsky, 2025), we specialize it to transformer circuits and introduce the single-pass, dimensionless Effective-Information Consistency Score (EICS). EICS combines (i) a *normalized sheaf inconsistency* computed from local Jacobians and activations, with (ii) a *Gaussian EI proxy* for circuit-level causal emergence derived from the same forward state. The construction is white-box, single-pass, and makes units explicit so that the score is dimensionless. We further provide practical guidance on score interpretation, computational overhead (with fast and exact modes), and a toy sanity-check analysis.

***Keywords****:* mechanistic interpretability, ransformer circuits, sheaf theory, causal emergence, uncertainty quantification, large language models (LLMs).

**REFERENCES**

1. *Olsson C., Elhage N., Nanda N., et al.* In-context Learning and Induction Heads. 2022. arXiv: 2209.11895.

2. *Anthropic*. Circuit Tracing / Attribution Graphs: Methods & Applications: Transformer Circuits Team. 2025. Access mode: https://transformer-circuits.pub/2025/attribution-graphs/ (accessed: 2025-08-20).

3. *Yao Y., Zhang N., Xi Z., Wang M., Xu Z., Deng S., and Chen H.* Knowledge Circuits in Pretrained Transformers // Advances in Neural Information Processing Systems (NeurIPS). 2024. Vol. 37. P. 118571–118602.

4. *Krasnovsky A.A.* Sheaf-Theoretic Causal Emergence for Resilience Analysis in Distributed Systems. 2025. arXiv : 2503.14104.

5. *Hansen J., Ghrist R.* Toward a Spectral Theory of Cellular Sheaves // Journal of Applied and Computational Topology. 2019. Vol. 3, No. 4. P. 315–358.

6. *Robinson M.* Topological Signal Processing. Springer, 2014.

7. *Rosas F.E., Mediano P.A.M., Jensen H.J., Seth A.K., Barrett A.B., Carhart-Harris R.L., and Bor D.* Reconciling Emergences: An Information-Theoretic Approach to Identify Causal Emergence in Multivariate Data // PLOS Computational Biology. 2020. Vol. 16, No. 12. P. e1008289.

8. *Tononi G., Sporns O.* Measuring Information Integration // BMC Neuroscience. 2003. Vol. 4. P. 31.

9. *Oizumi M., Albantakis L., Tononi G.* From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 // PLOS Computational Biology. 2014. Vol. 10, No. 5. P. e1003588.

10. *Angelopoulos A.N., Bates S.* A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. 2021. arXiv : 2107.07511.

11. *Guo C., Pleiss G., Sun Y., and Weinberger K.Q.* On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR. 2017. P. 1321–1330.

12. *Lakshminarayanan B., Pritzel A., Blundell C.* Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30.

13. Bayesian Low-rank Adaptation for Large Language Models (Laplace-LoRA). 2023. ICLR 2024 version. arXiv : 2308.13111.

14. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models / Hase P., Bansal M., Kim B., and Ghandeharioun A. // Advances in Neural Information Processing Systems (NeurIPS). 2023. Vol. 36. P. 17643–17668.

15. *Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2025. Vol. 43, No. 2. P. 1–55.

## СВЕДЕНИЯ ОБ АВТОРЕ

***КРАСНОВСКИЙ Анатолий Анатольевич*** – аспирант Университета Иннополис. Исследовательская работа сфокусирована на фундаментальных вопросах интерпретируемости сложных систем, а также на разработке и применении методов математического моделирования для их анализа. Академические изыскания подкреплены более чем десятилетним опытом работы в IT-индустрии, где занимался проектированием и разработкой высоконагруженных распределенных систем. Имеет степень магистра с отличием в области прикладной математики и компьютерных наук.

***Anatoly Anatolievich KRASNOVSKY*** is a Ph.D. student at Innopolis University. His research focuses on fundamental questions in the interpretability of complex systems, as well as the development and application of mathematical modeling methods for their analysis. His academic pursuits are grounded in over a decade of experience in the IT industry, where he designed and developed high-load distributed systems. He holds an M.S. in Applied Mathematics and Computer Science with honors.

email: a.a.krasnovsky@gmail.com

ORCID: 0000-0001-6842-7340

*Материал поступил в редакцию 13 октября 2025 года*