Оценка частичных доказательств в Lean 4 на основе анализа матриц внимания большой языковой модели

Студент: Линич А.А.
Научный руководитель: к.м.н. Баранников С.А.

Московский физико-технический институт

Июнь 2025 г.

Актуальность

- Большие языковые модели (БЯМ) достигли впечатляющих результатов в NLP, но формальное математическое рассуждение остаётся открытой задачей [1].
- Для управления генерацией (например, с откатами или в обучении с подкреплением) полезно использовать оценку промежуточных шагов доказательства [2] [3].

Цель

- Разработать интерпретируемый метод оценки корректности частичных доказательств, генерируемых БЯМ в Lean 4.

Задачи

- Построить датасет с частичными доказательствами на языке Lean 4 на основе датасета PropL.
- Разработать бинарный классификатор неоконченных доказательств, принимающий на вход матрицы внимания БЯМ.
- Сравнить полученный метод с базовыми моделями.

Формальная постановка задачи

Каждому примеру соответствует последовательность токенов $(P, \varphi, T = (t_1, \ldots, t_k))$, где P — промпт, φ — условие теоремы, T — частичное доказательство.

$$y = \begin{cases} 1, & \text{если } T \text{ может быть продолжено до доказательства} \\ 0, & \text{иначе.} \end{cases},$$

Формальная постановка задачи

Пусть БЯМ (DeepSeek-Prover-v2-7b) порождает матрицы внимания

$$A^{(\ell,q)} \in \mathbb{R}^{n\times n},$$

для каждого слоя ℓ и головы q, где n — количество токенов на входе БЯМ.

Необходимо разработать скалярный признак $s\left(A^{(\ell,q)}\right)$.

Классификация:

$$\hat{y} = \mathbb{I}\left[s\left(A^{(\ell,q)}\right) \geq \tau^*\right],$$

где (ℓ^*, q^*, τ^*) — слой, голова запроса и порог классификации соответственно.

Превосходство над базовыми моделями формально означает:

$$\text{Accuracy}(f) > \max_i \text{Accuracy}(f_i),$$

где f_i — базовые модели.

Базовые модели

- **Zero-shot классификация** с использованием предобученной БЯМ в режиме запроса без обучения [4].
- **CatBoost** по специальному вектору токена начала последовательности [5].

Методы: MTD_0 [6]

- Построение матрицы смежности неориентированного графа близости токенов:

$$D := 1 - \frac{1}{2}\left(A^{(\ell,q)} + \left(A^{(\ell,q)}\right)^{\top}\right), \quad D_{ii} := 0.$$

- Обнуление весов ребер внутри блоков $P, \varphi, t_1, \ldots, t_k$:

$$D_{ij} := 0, \quad \text{если } i, j \in B, B \in \{P, \varphi, t_1, \ldots, t_k\}.$$

- Вычисление веса минимального остовного дерева $T_{\min}$ с нормировкой на количество тактик в доказательстве:

$$MTD_0^{(\ell,q)} = \frac{1}{k} \sum_{(i,j) \in T_{\min}} D_{ij}.$$

Методы: самовнимание токенов (CBT) [9]

Пусть c_B индекс последнего токена блока B. Внимание токена, завершающего блок B на себя:

$$s_B^{(\ell,q)} := A_{c_B,c_B}^{(\ell,q)}.$$

Из нормировки матриц внимания:

$$\sum_{j=1}^{n} A_{i,j}^{(\ell,q)} = 1 \quad \Rightarrow \quad s_B^{(\ell,q)} = 1 - \sum_{j \neq c_B} A_{c_B,j}^{(\ell,q)}.$$

Введем CBT как среднее значение самовнимания токенов, заверщающих блоки:

$$\mathrm{CBT}^{(\ell,q)} = \frac{1}{k+2} \sum_{B \in \mathcal{B}} s_B^{(\ell,q)} = 1 - \frac{1}{k+2} \sum_{B \in \mathcal{B}} \sum_{j \neq c_B} A_{c_B,j}^{(\ell,q)}.$$

Построение классификаторов

- Для каждого метода (MTD_0 и CBT) и каждой пары слоя ℓ и головы запроса q вычисляется скалярный признак $s^{(\ell,q)}$.
- Далее для каждой пары определяется абсолютная разность между средними значениями признака на положительных и отрицательных примерах на валидационной выборке, и с ее помощью оценивается разность математических ожиданий:

$$(\ell^*, q^*) = \arg\max_{\ell,q} \left| \mathbb{E}_{y=1}[s^{(\ell,q)}] - \mathbb{E}_{y=0}[s^{(\ell,q)}] \right|.$$

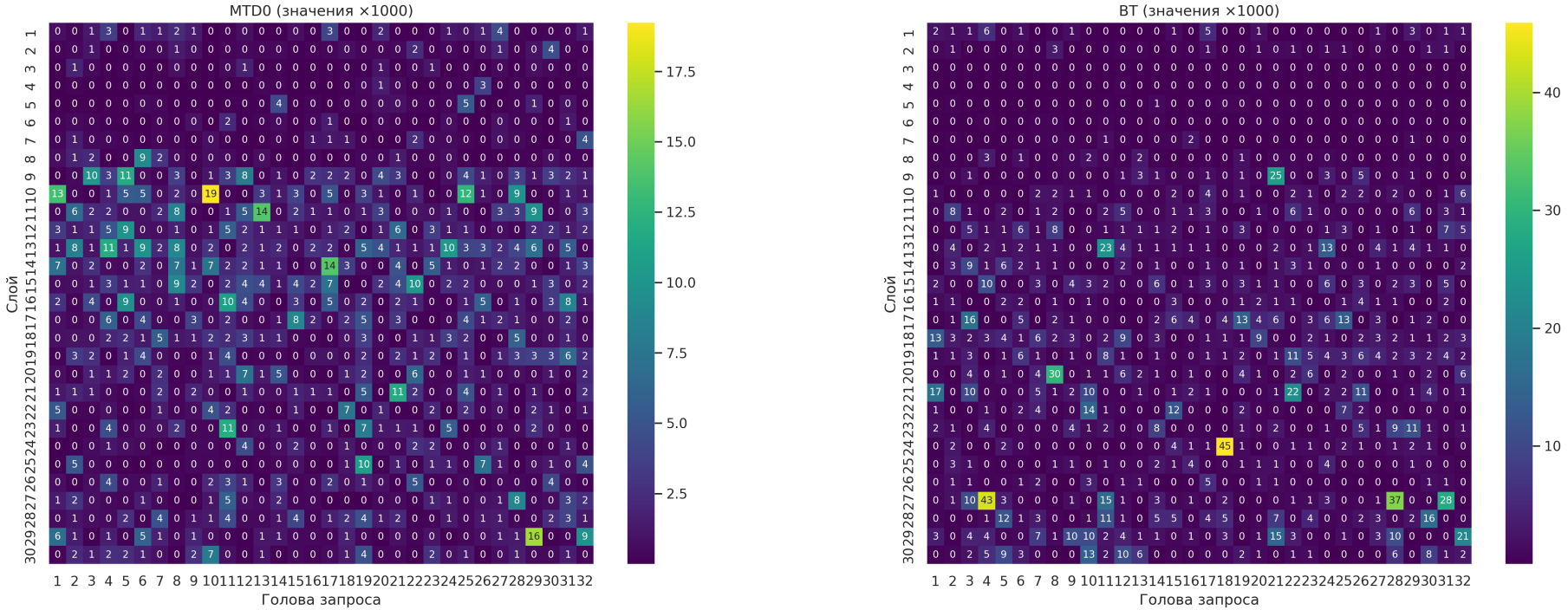
- По валидационной выборке для признака $s^{(\ell^*,q^*)}$ подбирается порог классификации:

$$\tau^* = \arg\max_{\tau} \text{Accuracy}(s^{(\ell^*,q^*)}, \tau).$$

Эксперимент

- Датасет с правильными и неправильными частичными доказательствами на основе PropL (частичные доказательства в Lean 4 из области пропозиционной логики) [7].
- 300 примеров на валидацию методов MTD_0, CBT, 1000 примеров на тест.
- БЯМ, фиксированная для всех методов: DeepSeek-Prover-v2-7b [8].

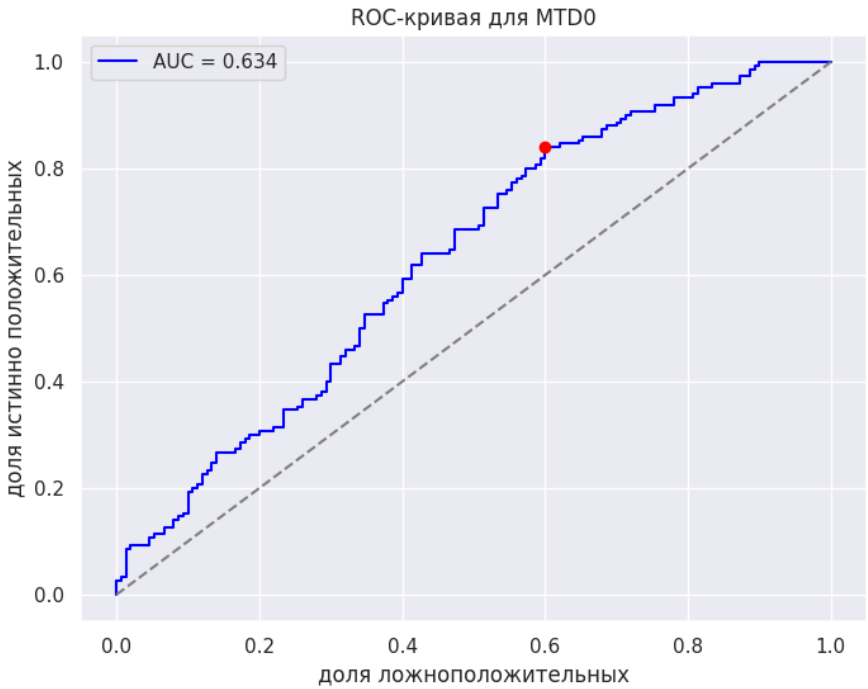
Выбор наилучших слоя и головы запроса

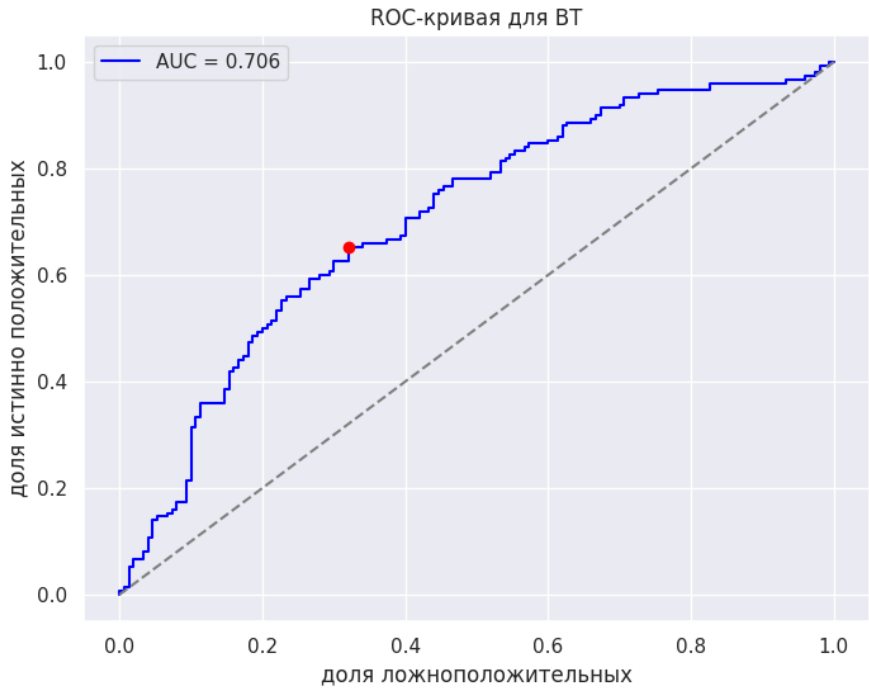


Значения $\hat{\Delta}_{\ell,q}$ для каждой пары (ℓ, q) на валидационной выборке. Лучшие слой и голова запроса — (10, 10) и (24, 18) для MTD_0 и CBT соответственно.

Выбор порога классификации

Для ℓ^*, q^* перебором выбирается порог классификации, максимизирующий точность.





MTD_0 CBT

Красной точкой обозначен выбранный порог τ^*.

Результаты

Метод	Accuracy	Precision	Recall	F1
MTD0	0.57	0.58	0.57	0.55
CBT	**0.62**	0.62	**0.62**	**0.62**
CatBoost	0.52	**0.70**	0.06	0.12
Zero-shot	0.58	0.60	0.58	0.56

CBT показывает наилучшее значение точности, превосходя оба бейзлайна и MTD_0.

Выводы

- Проведена обработка датасета PropL для получения правильных и неправильных частичных доказательств на Lean 4.
- Предложены два интерпретируемых признака, основанных на анализе матриц внимания большой языковой модели: MTD_0 и CBT.
- Построены бинарные классификаторы на основе предложенных признаков.
- Проведено сравнение с бейзлайнами. Оба метода показали сравнимую или лучшую точность, чем бейзлайны (CatBoost по специальному токену начала последовательности и zero-shot оценка самой модели).
- Метод CBT достиг максимальной точности на тестовой выборке (0.62), превзойдя оба бейзлайна и MTD_0.

Список литературы

1. Han, X. et al. “FormalMATH: Benchmarking Formal Mathematical Reasoning of Large Language Models.” *Advances in Neural Information Processing Systems*, vol. 34, pp. 7294–7305, 2021.

2. An, C. et al. “Learn from Failure: Fine-Tuning LLMs with Trial-and-Error Data for Intuitionistic Propositional Logic Proving.” In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 776–790.

3. Hu, Y. et al. “Learning to Utilize Shaping Rewards: A New Approach of Reward Shaping.” In: *Advances in Neural Information Processing Systems*, 2020.

4. Wang, X. et al. “Self-Consistency Improves Chain-of-Thought Reasoning in Language Models.” In: *International Conference on Learning Representations (ICLR)*, 2023.

Список литературы

5. Kowsari, K. et al. “Text Classification Algorithms: A Survey.” *Information*, vol. 10, no. 4, 2019.

6. Barannikov, S. et al. “Manifold Topology Divergence: A Framework for Comparing Data Manifolds.” In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

7. Zhang, W. et al. “PropLLM: A Dataset for Intuitionistic Propositional Logic Proofs.” arXiv preprint arXiv:2312.00552 (2023).

8. Ren, Z. et al. “DeepSeek-Prover-V2: Advancing Formal Mathematical Reasoning via Reinforcement Learning for Subgoal Decomposition.” preprint arXiv:2504.21801, 2025.

9. Škrlj, B. et al. “Exploring Neural Language Models via Analysis of Local and Global Self-Attention Spaces.” In: *EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 2021, pp. 76–83.