УДК 004.832

Топологический анализ математических доказательств, генерируемых большими языковыми моделями на языке доказательства теорем Lean 4

А. А. Линич[1], С. А. Баранников[2,3]

[1]Московский физико-технический институт (национальный исследовательский университет)
[2]Сколковский институт науки и технологий
[3]Национальный центр научных исследований (CNRS)

Современные математические доказательства становятся все сложнее, и проверка их корректности требует значительных усилий даже от экспертов [1]. В связи с этим активно развивается направление формализации математических доказательств с использованием инструментов автоматической проверки. Одним из таких инструментов является язык Lean 4, основанный на исчислении конструкций [2]. Доказательство в Lean 4 состоит из условия теоремы и последовательности тактик. В Lean 4 некорректное или неполное доказательство не подлежит компиляции, что делает успешно скомпилированный код эквивалентом формальной верификации доказательства [2].

Формализация доказательств на формальных языках остается трудоемкой задачей. Для упрощения формализации доказательств и возможной автоматизации вывода новых теорем исследуется использование больших языковых моделей (БЯМ) [3]. Один из подходов — генерация полных доказательств на основе формулировки теоремы. Далее речь пойдет именно о таких моделях.

При генерации доказательств с помощью БЯМ важно уметь определять качество доказательств. Это необходимо как для контроля качества, так и для улучшения самой модели. Стандартные подходы к оценке включают сравнение с эталонным доказательством на уровне токенов и проверку с помощью компиляции. Однако оба метода имеют ограничения. В первом случае доказательство может быть верным, но отличаться от эталонного порядком шагов или обозначениями переменных, что затрудняет оценку сходства. Во втором случае проверка через компиляцию не позволяет оценить частично правильные доказательства: если код не скомпилировался, невозможно определить, насколько он близок к корректному.

Работа сосредоточена на применении Manifold Topology Divergence (MTopDiv) — топологического метода сравнения двух многообразий — для оценки качества доказательств, сгенерированных БЯМ в Lean 4. В данной работе он адаптирован для анализа графов, построенных на матрицах внимания БЯМ, вместо стандартного подхода с облаком точек [4]. Далее описаны шаги его вычисления.

Для графа, построенного на матрице внимания трансформерной модели для фиксированного слоя и головы, рассматриваются ребра с соответствующими весами. Веса внутри одной тактики зануляются, что эквивалентно стягиванию всех токенов этой тактики в одну точку. Аналогично зануляются веса ребер, соединяющих токены в промпте и в условии теоремы. Далее на получившемся графе строится минимальное остовное дерево, и подсчитывается сумма весов его ребер. В данной работе предлагается нормировать эту сумму не на длину ответа модели, а на количество ребер в графе.

Цель эксперимента — с помощью mtd0 выявить голову внимания, наиболее чувствительную к различию между корректными и некорректными доказательствами, а затем оценить ее разделяющую способность на тестовой выборке. Для этого использовался датасет с примерами доказательств, сгенерированных методом проб и ошибок [5]. Для каждой теоремы была доступна корректная последовательность тактик, а некорректная моделировалась путем перемешивания всех релевантных тактик из доказательства с пробами и ошибками. Затем последовательность обрезалась до длины правильного доказательства (шесть тактик), чтобы избежать влияния разного масштаба данных. В качестве модели для получения матриц внимания использовалась Qwen2.5-coder-1.5B [6]. Рассматривались все 28 слоев и 12 голов внимания. Для каждого слоя и головы вычислялась нормированная сумма весов ребер минимального остовного дерева. Затем эти значения усреднялись отдельно по множеству корректных и некорректных доказательств, после чего оценивалась абсолютная разница между ними.

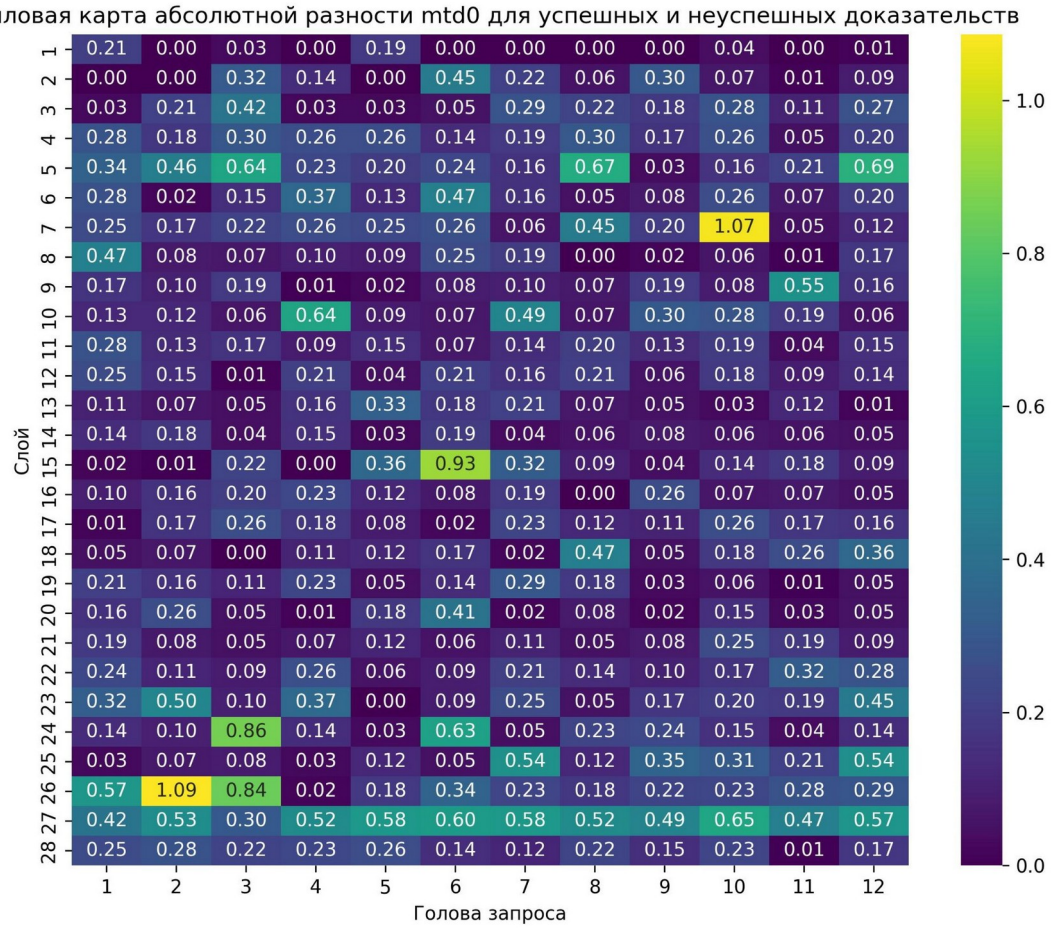

Рис. 1. Тепловая карта абсолютной разности среднего mtd0 на успешных и неуспешных доказательствах. Значения домножены на 10^3

Как показано на рис. 1, наибольшую чувствительность к различию между правильными и ошибочными доказательствами показала вторая голова 26-го слоя. Для нее разность средних значений нормированной суммы ребер составила более 10^{-3} что свидетельствует о наличии различий в топологических характеристиках графов, построенных на основе соответствующих матриц внимания модели. Интересно, что более половины голов на предпоследнем слое показали разницу не менее 0.5×10^{-3}, что может свидетельствовать о более осмысленной обработке признаков на этом уровне.

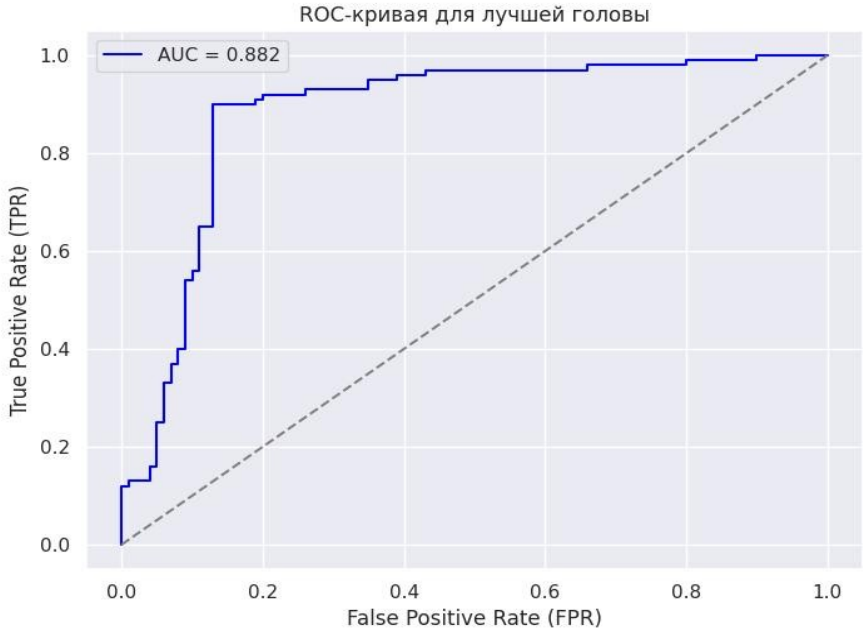

Рис. 2. Оценка качества лучшей головы

Как видно на рис. 2, AUC составляет 0.882, что свидетельствует о высокой разделяющей способности значения mtd0 для лучшей головы.

В ходе эксперимента выявлена голова с высокой разделяющей способностью, позволяющей различать правильные и ошибочные доказательства на основе топологических характеристик графов матриц внимания. Полученные результаты подтверждают перспективность предложенного подхода для анализа качества сгенерированных доказательств.

Литература

1. Hales T. C. Formal Proof // Notices of the American Mathematical Society. 2008. V. 55(11). P. 1370–1380.
2. de Moura L., Ullrich S. The Lean 4 Theorem Prover and Programming Language // In A. Platzer, G. Sutcliffe (Eds.), Automated Deduction – CADE 28. Springer International Publishing, 2021. P. 625–635.
3. Polu S., Sutskever I. Generative Language Modeling for Automated Theorem Proving // arXiv preprint. 2020. arXiv:2009.03393.
4. Barannikov S., Trofimov I., Sotnikov G., Trimbach E., Korotin A., Filippov A., Burnaev E. Manifold Topology Divergence: a Framework for Comparing Data Manifolds // In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, J. Wortman Vaughan (Eds.), Advances in Neural Information Processing Systems. Vol. 34. Curran Associates, Inc., 2021. P. 7294–7305.
5. KomeijiForce. PropL [Электронный ресурс]. URL: https://huggingface.co/datasets/KomeijiForce/PropL (дата обращения: 07.03.2025).
6. Hugging Face. Qwen2.5-Coder-1.5B [Электронный ресурс]. URL: https://huggingface.co/Qwen/Qwen2.5-Coder-1.5B?utm_source=chatgpt.com (дата обращения: 07.03.2025).