Вектора представлений токенов в трансформерах лежат в высокоразмерном пространстве, что затрудняет визуализацию. Для оценки их взаимного расположения вводятся характеристики косинусной и сингулярной анизотропии. Анизотропия определяет степень неравномерности распределения векторов: при высокой анизотропии векторы концентрируются в узком конусе или вытягиваются вдоль одного направлении. Косинусная анизотропия вычисляется как матожидание косинусной близости векторов и оценивает углы между ними. Сингулярная анизотропия есть отношение квадрата максимального сингулярного числа к сумме квадратов всех сингулярных чисел и показывает концентрацию дисперсии вдоль главного направления. Исследования [1-3] подтверждают, что при прохождении через слои трансформера векторы сближаются, образуя узкий конус. Причины анизотропии дискуссионны: одни связывают её со слоями внимания [1], другие – со сжатием информации при абстрагировании [2, 3].

Эксперименты на моделях Qwen2.5-0.5/1.5/3B и GPT2-Medium/Large с текстами Википедии показали неравномерность распределений векторов. Косинусная анизотропия возрастает до 0.6–0.7 у выходных слоёв. Унимодальные гистограммы распределения косинусной близости с дисперсией 0.1–0.2 указывают на расположение векторов в узком конусе вдоль доминирующего направления. Наибольший рост косинусной анизотропии дают остаточные связи и механизм внимания. Сингулярная анизотропия немонотонна: быстро достигает максимума 0.8–0.9 на средних слоях, затем снижается до 0.2–0.5. Наибольшее влияние на неё оказывают остаточные связи и полносвязные слои. На средних слоях максимальное сингулярное число значительно превышает остальные, что свидетельствует о выравнивании векторов вдоль одного направления. Обнаружена сильная корреляция сингулярной анизотропии с нормой вектора первого токена [2]. Это указывает на выравнивание векторов именно вдоль вектора первого токена, вероятно, вследствие авторегрессионного внимания.

Дальнейшие исследования направлены на обоснование роли первого токена как аттрактора.

Список литературы

[1] Godey N., Clergérie É., Sagot B. Anisotropy is inherent to self-attention in transformers // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. 2024. P. 35–48.

[2] Queipo-de-Llano E. et al. Attention sinks and compression valleys in llms are two sides of the same coin // arXiv preprint arXiv:2510.06477. 2025.

[3] Skean O. et al. Layer by layer: Uncovering hidden representations in language models // arXiv preprint arXiv:2502.02013. 2025.