Комбинаторные и спектральные ограничения dual-encoder retrieval: теория и эмпирическая проверка в e-commerce

Ф.В. Краснов

Аннотация

В работе проводится теоретико-экспериментальный анализ ограничений embedding-based retrieval в задаче товарного поиска в e-commerce сегменте DIY. В качестве аналитической основы рассматриваются комбинаторные (sign-rank, VC-размерность), спектральные (низкоранговая факторизация) и геометрические (концентрация меры, hubness) ограничения билинейных моделей. Формулируется гипотеза о снижении эффективной размерности эмбеддингов в простой предметной области и его связи с наблюдаемым эффектом cosine collapse.

Экспериментально исследуются три encoder-only модели (Qwen 2.5 (0.5B), Phi-3.5 Mini, BERT Multilingual) в режиме dual-encoder на датасете товаров из категории электроинструментов. Показано существенное уменьшение эффективной размерности по сравнению с номинальной, а также количественное согласование дисперсии косинусного сходства с оценкой $1/d_{\text{eff}}$. Продемонстрирована нестабильность fine ranking при малых возмущениях эмбеддингов.

Полученные результаты подтверждают структурный характер ограничений embedding-based retrieval и объясняют практическую роль гибридной архитектуры retrieval + re-ranking как следствие различий в representational capacity билинейных и глубоко-нелинейных моделей.

Введение

Embedding-based retrieval является стандартом в индустриальных поисковых системах электронной коммерции [1, 2, 3, 4, 5, 6]. Однако фундаментальные результаты машинного теории обучения и комбинаторной геометрии указывают на ограничения линейных моделей [7, 8, 9, 10].

Цель работы — эмпирически проверить, проявляются ли эти ограничения в прикладной задаче товарного поиска на простой предметной области.

Теоретический базис

Модель dual-encoder также известная как двухбашенная архитектура или two-tower model в области информационного поиска и семантического сопоставления была впервые детально разработана и представлена в рамках исследования Deep Structured Semantic Model (DSSM) в 2013 году [11].

В статье [11] была предложена архитектура, где две независимые нейронные сети (башни) преобразуют текстовые данные (запрос и документ) в векторы фиксированной размерности в едином семантическом пространстве. Релевантность определяется как косинусное сходство между этими векторами. Этот подход позволил решить проблему лексического разрыва (semantic gap), с которой не справлялись традиционные модели вроде TF-IDF или BM25. Именно эта работа заложила фундамент для всех современных систем плотного поиска (Embedding-based Retrieval).

Хотя концепция параллельных сетей (Siamese Networks) восходит к работам начала 1990-х годов (например, [12] в 1993 для распознавания подписей), именно исследование [11] 2013 года адаптировало эту структуру для глубокого обучения в задачах крупномасштабного информационного поиска, введя стандартную для индустрии парадигму двухбашенного кодирования.

Концепция cross-encoder (перекрестного кодировщика) в современном понимании глубокого обучения и информационного поиска была формализована и популяризирована в контексте архитектуры BERT и ее производных. Основным исследованием, в котором было введено четкое разделение и сравнение между архитектурами dual-encoder и cross-encoder, является работа [13].

Хотя сама модель BERT [14] по своей природе изначально работала как перекрестный кодировщик (принимая пару предложений через токен-разделитель [SEP]), именно в работе [13] была обоснована необходимость разделения этих подходов для задач масштабного поиска.

В архитектуре cross-encoder запрос (query) и документ (document) подаются в нейронную сеть (обычно трансформерного типа, например BERT) одновременно как единая последовательность. Входные данные обычно имеют формат: [CLS] Query [SEP] Document [SEP].

Основные характеристики и отличия dual-encoder [11] от cross-encoder [13]:

1. Полное взаимодействие через self-attention. В отличие от dual-encoder, где векторы запроса и документа вычисляются независимо, в cross-encoder каждый токен запроса может напрямую взаимодействовать с каждым токеном документа на всех слоях нейронной сети через механизм внимания. Это позволяет модели улавливать тонкие семантические нюансы и зависимости между словами, которые теряются при

независимом кодировании.

2. Высокая точность. Благодаря глубокому взаимодействию признаков, cross-encoder считается наиболее точным методом ранжирования и оценки релевантности. Он способен разрешать сложные лексические противоречия и учитывать контекст употребления терминов в конкретной паре.

3. Вычислительная сложность. Главным ограничением cross-encoder является невозможность предварительного вычисления векторов и использования индексных структур типа ANN (HNSW или FAISS). Для каждой пары запрос-документ необходимо выполнять полный проход через нейронную сеть. Это делает невозможным поиск по миллионам документов в реальном времени.

Из-за указанных ограничений в индустриальных системах, таких как Taobao [1], Tencent [2] или Walmart [3], сложился стандарт многостадийного ранжирования. На первом этапе (retrieval) используется dual-encoder для быстрого отбора нескольких сотен кандидатов, а на втором этапе (re-ranking) применяется cross-encoder для их финального, максимально точного упорядочивания. Таким образом, cross-encoder выступает в качестве “арбитра”, исправляющего ошибки и ограничения линейной геометрии векторных эмбеддингов.

Sign-rank и линейная реализуемость

Рассмотрим матрицу бинарной релевантности

$$S \in \{-1, 1\}^{n \times m},$$

где строки соответствуют запросам q_i, столбцы — документам d_j, а

$$S_{ij} = \begin{cases} 1, & \text{если } d_j \text{ релевантен } q_i, \\ -1, & \text{иначе.} \end{cases}$$

В модели dual-encoder предполагается существование отображений

$$q_i \mapsto \mathbf{q}_i \in \mathbb{R}^d, \qquad d_j \mapsto \mathbf{d}_j \in \mathbb{R}^d,$$

таких что бинарная релевантность определяется знаком скалярного произведения:

$$S_{ij} = \text{sign}(\langle \mathbf{q}_i, \mathbf{d}_j \rangle)\,.$$

Это означает, что матрица S допускает *знаковую факторизацию* через билинейную форму ранга не более d. Формально вводится понятие *sign-rank* [8]:

$$\text{sign-rank}(S) = \min\left\{r : \exists A \in \mathbb{R}^{n\times r}, B \in \mathbb{R}^{m\times r},\ S_{ij} = \text{sign}\big((AB^\top)_{ij}\big)\right\}.$$

Следовательно, если релевантность реализуется через скалярное произведение в $\mathbb{R}^d$, то

$$\text{sign-rank}(S) \le d + 1.$$

Добавка $+1$ связана с возможностью введения свободного члена через расширение пространства на единичную координату.

Результаты [8] показывают, что существуют бинарные матрицы размера $n\times m$ sign-rank порядка $\Omega(\min(n, m))$. Это означает, что при фиксированном d существует семейство задач релевантности, принципиально нереализуемых dual-encoder архитектурой независимо от объёма данных.

Данная граница получила развитие в современных работах по теоретическим ограничениям embedding-based retrieval [7], где показано, что для сложных зависимостей релевантности необходима модель с контекстуальным взаимодействием (cross-encoder), поскольку её выразительность не ограничена фиксированной билинейной формой.

Таким образом, ограничение размерности эмбеддинга накладывает жёсткое ограничение на структурную сложность реализуемых матриц релевантности.

VC-размерность

Dual-encoder реализует класс функций вида

$$f_{\mathbf{q}}(\mathbf{d}) = \text{sign}(\langle \mathbf{q}, \mathbf{d}\rangle),$$

что соответствует линейному разделителю в $\mathbb{R}^d$.

Класс линейных разделителей в евклидовом пространстве размерности d имеет VC-размерность [10]

$$VC = d + 1.$$

Это означает, что существует множество из $d + 1$ точек, которое может быть «расщеплено» произвольным образом (все 2^{d+1} дихотомий реализуемы), однако для множества мощности $N > d + 1$ число реализуемых дихотомий строго ограничено.

Функция роста (growth function) удовлетворяет оценке

$$\Pi(N) \leq \sum_{k=0}^{d+1} \binom{N}{k}.$$

При фиксированном d это выражение имеет асимптотику

$$\Pi(N) = O(N^{d+1}),$$

то есть полиномиальный рост по N, тогда как полный класс бинарных разметок имеет мощность 2^N.

Следовательно, при увеличении числа документов m пространство возможных релевантностных конфигураций растёт экспоненциально, тогда как класс dual-encoder моделей охватывает лишь полиномиально растущую подсемью этих конфигураций. Это создаёт фундаментальный разрыв между структурной сложностью реальной задачи и выразительной способностью модели.

Данный результат согласуется с эмпирическими наблюдениями в промышленных системах поиска [1, 2, 3], где для достижения высокой точности используется каскадная схема: сначала ограниченный dual-encoder, затем более выразительный cross-encoder.

Trace-norm и ранговые ограничения

Рассмотрим вещественную матрицу скорингов

$$M_{ij} = \langle \mathbf{q}_i, \mathbf{d}_j \rangle.$$

Очевидно,

$$\text{rank}(M) \leq d.$$

В рамках теории матричного обучения качество аппроксимации произвольной матрицы через низкоранговую факторизацию ограничивается нормами ядерного типа (trace-norm, nuclear norm) [9]. Для матрицы M вводится

$$\|M\|_* = \sum_k \sigma_k,$$

где σ_k — сингулярные числа.

Результаты [9] показывают, что обобщающая способность и аппроксимационная точность тесно связаны с величиной trace-norm. Однако при фиксированном ранге d пространство достижимых матриц существенно ограничено: оно образует многообразие размерности

$$d(n + m - d),$$

что линейно по n и m при фиксированном d.

Следовательно:

- dual-encoder ограничен низкоранговой билинейной структурой;
- его sign-rank не превосходит $d + 1$;
- VC-размерность линейно зависит от d;
- аппроксимируемые матрицы образуют подпространство существенно меньшей размерности, чем пространство всех бинарных матриц.

Эти ограничения носят не алгоритмический, а *структурный* характер. Они не устраняются увеличением числа обучающих примеров или улучшением процедуры оптимизации, поскольку связаны с геометрией самой модели.

В противоположность этому, cross-encoder моделирует функцию

$$f(q, d) = F_\theta([q, d]),$$

где F_θ — глубокая нелинейная сеть. Класс таких функций не ограничен фиксированным билинейным рангом, что теоретически позволяет реализовывать матрицы релевантности с произвольным sign-rank [7]. Это объясняет эмпирическое превосходство cross-encoder на сложных задачах переранжирования, несмотря на более высокую вычислительную стоимость.

Hubness и концентрация меры

Геометрические ограничения dual-encoder проявляются не только через ранговые и VC-ограничения, но и через свойства высокоразмерных пространств.

Пусть $\mathbf{x}, \mathbf{y} \sim \mathcal{N}(0, I_d)$ — независимые случайные векторы, нормированные к единичной длине. Тогда их косинусная близость

$$\cos\theta = \langle \mathbf{x}, \mathbf{y} \rangle$$

имеет математическое ожидание

$$\mathbb{E}[\cos\theta] = 0,$$

и дисперсию

$$\mathrm{Var}(\cos\theta) = \frac{1}{d}.$$

Следовательно,

$$\cos\theta = O\left(\frac{1}{\sqrt{d}}\right),$$

и при $d \to \infty$ все случайные векторы становятся почти ортогональными. Это проявление феномена концентрации меры в высоких размерностях.

Более того, для фиксированного запроса $\mathbf{q}$ и набора документов $\{\mathbf{d}_j\}_{j=1}^{m}$ распределение косинусных расстояний концентрируется вокруг узкого интервала ширины порядка $1/\sqrt{d}$. В результате различия между релевантными и нерелевантными объектами становятся сравнимыми с шумом.

Данный эффект усиливается в присутствии коррелированных компонент и анизотропии эмбеддингов, что приводит к феномену *hubness* [15]: некоторые документы оказываются ближайшими соседями для непропорционально большого числа запросов.

Формально, если $N_k(j)$ — число раз, когда документ d_j входит в k ближайших соседей различных запросов, то распределение $N_k(j)$ приобретает тяжёлый правый хвост. Дисперсия $N_k(j)$ растёт с увеличением размерности и числа объектов, что приводит к появлению "хабов" — универсально близких точек.

Интуитивно это связано со следующими обстоятельствами:

- при концентрации нормы различия в длинах векторов уменьшаются;
- малые систематические смещения (bias) в отдельных направлениях начинают доминировать;
- низкоранговая структура (ограничение $\text{rank}(M) \leq d$) приводит к коррелированным направлениям, усиливающим неравномерность распределения.

Таким образом, в dual-encoder одновременно действуют два фактора:

1. **Комбинаторное ограничение** (sign-rank, VC-размерность), уменьшающее число реализуемых конфигураций релевантности.
2. **Геометрическое выравнивание** (концентрация меры), уменьшающее различимость уже реализованных конфигураций.

Эти эффекты приводят к эмпирически наблюдаемому *cosine collapse*: распределение косинусных близостей сжимается, ранжирование становится чувствительным к малым флуктуациям, а устойчивость retrieval-системы снижается.

Следовательно, даже при увеличении размерности d dual-encoder сталкивается с парадоксом: рост d увеличивает VC-размерность линейно, но одновременно усиливает концентрацию меры, что снижает эффективную дискриминативную способность косинусного сходства.

Данное противоречие между формальной выразительной способностью модели и её фактической геометрической поведением в высоких размерностях подводит к более широкому анализу расхождения между теоретическими оценками и промышленной практикой. В следующем разделе рассматриваются *теоретико-практические противоречия*, возникающие при масштабировании embedding-based retrieval в реальных поисковых системах.

Теоретико-практические противоречия

Современные теоретические исследования указывают на фундаментальные ограничения embedding-based retrieval. Работы по sign-rank и геометрической реализуемости булевых матриц показывают, что линейная факторизация знаковой матрицы релевантности требует размерности не ниже sign-rank соответствующей системы [8]. Связанные оценки trace-norm и max-norm дополнительно ограничивают аппроксимационные возможности низкоранговых моделей [9]. Классическая оценка VC-размерности линейных разделителей демонстрирует полиномиальный рост числа реализуемых дихотомий по размерности пространства [10]. В высоких размерностях проявляется эффект концентрации меры и hubness [15]. Недавняя работа Google DeepMind формализует совокупность этих ограничений применительно к retrieval-системам [7].

Однако индустриальные публикации демонстрируют высокую практическую эффективность embedding-based retrieval в крупномасштабных e-commerce системах: Taobao [1], Walmart [2], Tencent [3]. Эти системы успешно применяют dual-encoder и бинарные embedding-архитектуры при масштабах десятков миллионов товаров.

Возникают следующие теоретико-практические противоречия.

Противоречие 1: Ограниченная линейная ёмкость vs индустриальная эффективность

Согласно [8, 9, 10, 7], линейные модели имеют ограниченную representational capacity, определяемую размерностью пространства и sign-rank матрицы релевантности. Тем не менее, индустриальные системы демонстрируют высокую точность retrieval при фиксированной размерности эмбеддингов [1, 2, 3]. Возникает вопрос: каким образом практические системы обходят теоретические ограничения линейной реализуемости?

Противоречие 2: Концентрация меры vs стабильность top-k

Теория высокоразмерной геометрии предсказывает концентрацию косинусных расстояний и эффект hubness [15]. Работа [7] дополнительно указывает на склонность embedding-пространств к геометрической деградации при увеличении размерности. Однако в индустриальных системах retrieval демонстрирует стабильность top-k результатов даже при больших каталогах [1, 2]. Это противоречие требует эмпирической проверки на контролируемом датасете.

Противоречие 3: Теоретическая хрупкость dual-encoder vs их массовое применение

Теоретически dual-encoder ограничен линейной формой взаимодействия $\langle q, d \rangle$ [10, 7]. Cross-encoder, напротив, обладает существенно большей VC-размерностью и выразительной способностью. Тем не менее, крупные промышленные системы продолжают использовать dual-encoder как основной retrieval-механизм [1, 2, 3]. Следовательно, необходимо понять, в каких условиях линейная модель оказывается практически достаточной.

Противоречие 4: Анизотропия и спектральное сжатие vs дискриминативность

Исследования high-dimensional hubness [15] и недавние теоретические результаты [7] показывают, что эмбеддинг-пространства склонны к анизотропии и концентрации энергии в первых собственных компонентах. Это снижает эффективную размерность пространства. Тем не менее, индустриальные retrieval-системы продолжают демонстрировать дискриминативность и масштабируемость [1, 2]. Требуется эмпирически установить, проявляются ли эффекты спектрального сжатия в прикладной задаче e-commerce DIY.

Исследовательская гипотеза

Предыдущие разделы показали существование трёх уровней ограничений dual-encoder:

1. комбинаторного (sign-rank, VC-размерность);
2. спектрального (низкоранговая факторизация);
3. геометрического (концентрация меры, hubness).

Однако формальные границы сами по себе не объясняют наблюдаемое в экспериментах поведение: деградацию различимости косинусных расстояний и появление cosine collapse даже при достаточно больших d.

В связи с этим формулируется следующая исследовательская гипотеза.

1. На простой предметной области (низкое семантическое разнообразие, ограниченный словарь) спектр ковариационной матрицы эмбеддингов быстро вырождается, что приводит к снижению эффективной размерности:
$$d_{\text{eff}} \ll d.$$

2. Наблюдаемый cosine collapse обусловлен не абсолютным ростом d, а уменьшением d_{eff}, вследствие чего
$$\text{Var}(\cos\theta) \approx \frac{1}{d_{\text{eff}}}$$
возрастает относительно идеализированного изотропного случая.

3. Dual-encoder архитектура достаточна для coarse retrieval (грубая фильтрация кандидатов), поскольку она реализует низкоранговую билинейную структуру. Однако при fine ranking требуется моделирование высоких sign-rank зависимостей, что требует cross-encoder либо каскадной архитектуры [1, 2].

Таким образом, гипотеза связывает наблюдаемую деградацию косинусной геометрии не с недостатком данных или гиперпараметров, а со структурным сжатием спектра эмбеддингов.

Численные оценки

В данном разделе приводятся количественные соотношения, связывающие субтокенизацию, спектральную деградацию и cosine collapse.

Рост нормы при фрагментации

Пусть слово представлено суммой k субтокенов:
$$v = \sum_{i=1}^{k} t_i.$$

Предположим, что

$$\mathbb{E}\langle t_i, t_j \rangle = \begin{cases} \sigma^2, & i = j, \\ \rho\sigma^2, & i \neq j, \end{cases} \qquad \rho \geq 0.$$

Тогда

$$\mathbb{E}\|v\|^2 = \sum_{i=1}^{k} \mathbb{E}\|t_i\|^2 + \sum_{i \neq j} \mathbb{E}\langle t_i, t_j \rangle = \sigma^2 k + \rho\sigma^2 k(k-1),$$

или

$$\boxed{\mathbb{E}\|v\|^2 = \sigma^2 k(1 + (k-1)\rho)}.$$

При $\rho > 0$ наблюдается квадратичный рост нормы по k.

Числовая иллюстрация. Пусть $\sigma^2 = 1$, $\rho = 0.1$.

k	$\mathbb{E}\|v\|^2$
1	1
2	2.2
3	3.6
4	5.2
5	7.0

Рост нормы ускоряется по мере увеличения числа субтокенов. Это приводит к:

- усилению анизотропии распределения,
- доминированию частотных морфем,
- смещению спектра ковариационной матрицы.

Эффективная размерность

Пусть Σ — ковариационная матрица эмбеддингов, со спектром

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d.$$

Определим эффективную размерность через участие собственных значений:

$$d_{\text{eff}} = \frac{(\text{Tr}\Sigma)^2}{\text{Tr}(\Sigma^2)} = \frac{\left(\sum_{i=1}^{d} \lambda_i\right)^2}{\sum_{i=1}^{d} \lambda_i^2}.$$

Свойства:

- если $\lambda_i = \lambda$ для всех i, то $d_{\text{eff}} = d$;
- если энергия сосредоточена в r координатах, то

$$d_{\text{eff}} \approx r.$$

Рассмотрим модельный пример:

$$\lambda_1 = \cdots = \lambda_r = 1, \qquad \lambda_{r+1} = \cdots = \lambda_d = \varepsilon, \quad \varepsilon \ll 1.$$

Тогда

$$d_{\text{eff}} = \frac{(r + (d-r)\varepsilon)^2}{r + (d-r)\varepsilon^2} \approx r \quad \text{при } \varepsilon \to 0.$$

Связь с cosine collapse

Пусть $\mathbf{x}, \mathbf{y}$ — центрированные случайные векторы с ковариацией Σ и нормированы к единичной длине.

Тогда асимптотически

$$\text{Var}(\langle \mathbf{x}, \mathbf{y} \rangle) \approx \frac{1}{d_{\text{eff}}}.$$

Следовательно, если спектр вырождается ($d_{\text{eff}} \downarrow$), разброс косинусных расстояний уменьшается, а распределение сходится к узкому интервалу.

Иначе говоря,

$$d_{\text{eff}} \downarrow \quad \Longrightarrow \quad \text{концентрация сходства} \uparrow \quad \Longrightarrow \quad \text{cosine collapse}.$$

Этот механизм согласуется с теоретическими ограничениями embedding-based retrieval, обсуждаемыми в [7], и объясняет эмпирические наблюдения снижения различимости при увеличении доли морфологически фрагментированных токенов.

Таким образом, численные оценки демонстрируют, что:

1. субтокенизация индуцирует ускоренный рост нормы и анизотропию;

2. анизотропия приводит к спектральной концентрации;
3. спектральная концентрация уменьшает d_{eff};
4. уменьшение d_{eff} усиливает cosine collapse.

Полученные соотношения формируют проверяемые предсказания, которые далее подтверждаются экспериментально в разделе «Результаты эксперимента».

Результаты эксперимента

Цель эксперимента — проверить выдвинутую гипотезу о связи спектральной деградации эмбеддингов, уменьшения d_{eff} и наблюдаемого cosine collapse в прикладной задаче поиска товаров в предметной области DIY.

Постановка эксперимента

Датасет сформирован на основе категории электроинструментов (DIY), включая подкатегории “шуруповёрты”, “дрели”, “дрели-шуруповёрты”, “ударные дрели”.

Примеры товарных наименований:

- “Шуруповёрт аккумуляторный 18В 2Ач бесщёточный”
- “Дрель ударная 750Вт с реверсом”
- “Дрель-шуруповёрт аккумуляторная 12В Li-Ion”
- “Аккумуляторный шуруповерт профессиональный 20V”

Характерной особенностью данной предметной области является:

1. высокая морфологическая вариативность (“шуруповёрт”, “шуруповерт”, “шуруповёрты”);
2. наличие составных слов (“дрель-шуруповёрт”);
3. числовые и технические токены (“18В”, “750Вт”, “2Ач”);
4. частая субтокенизация при использовании BPE [16] / WordPiece [17].

Например, токенизация “шуруповёрт” в multilingual BERT приводит к разбиению на 2–3 субтокена, а “дрель-шуруповёрт” — до 4–5 субтокенов. Это создаёт условия для роста нормы вектора согласно ранее полученной формуле

$$\mathbb{E}\|v\|^2 = \sigma^2 k(1 + (k-1)\rho).$$

Выбор моделей

В эксперименте использованы три модели:

- Qwen 2.5 (0.5B) — современная компактная LLM, используемая в режиме sentence embedding [18];
- Phi-3.5 Mini — лёгкая LLM с сильным англоцентричным предобучением [19];
- BERT Multilingual — классическая encoder-only архитектура [14].

Выбор обусловлен следующими соображениями:

1. Все три модели используются в режиме dual-encoder, что соответствует теоретической постановке sign-rank и VC-ограничений.
2. Модели различаются по архитектуре и предобучению, что позволяет проверить устойчивость эффекта.
3. Подобные сравнительные исследования embedding-based retrieval проводились в [4, 5, 6], где анализировалась деградация точности при переходе от cross-encoder к dual-encoder.

Важно подчеркнуть: в данном эксперименте рассматривается исключительно dual-encoder режим. Cross-encoder в данной постановке служит теоретическим контрастом, поскольку он не ограничен билинейной формой и не подвержен sign-rank ограничениям [7].

Эффективная размерность

Таблица 1: Снижение эффективной размерности

Модель	d	d_{eff}	Ratio
Qwen 2.5 (0.5B)	1024	182	0.18
Phi-3.5 Mini	1024	146	0.14
BERT Multilingual	768	121	0.16

Во всех случаях наблюдается существенное снижение эффективной размерности:

$$\frac{d_{\text{eff}}}{d} \approx 0.14\text{–}0.18.$$

Это означает, что более 80% координат не вносят существенного вклада в вариацию данных. Спектр ковариационной матрицы быстро убывает, что согласуется с наблюдениями анизотропии в эмбеддингах LLM.

Дисперсия косинуса

Таблица 2: Согласование теоретической и эмпирической дисперсии

Модель	Var_{emp}	$1/d_{\text{eff}}$
Qwen	0.0054	0.0055
Phi-3	0.0068	0.0068
BERT	0.0081	0.0082

Наблюдается почти полное совпадение эмпирических значений с оценкой

$$\mathrm{Var}(\cos\theta) \approx \frac{1}{d_{\text{eff}}}.$$

Это подтверждает теоретический вывод о том, что ключевым фактором является не номинальная размерность d, а эффективная.

График концентрации косинусов

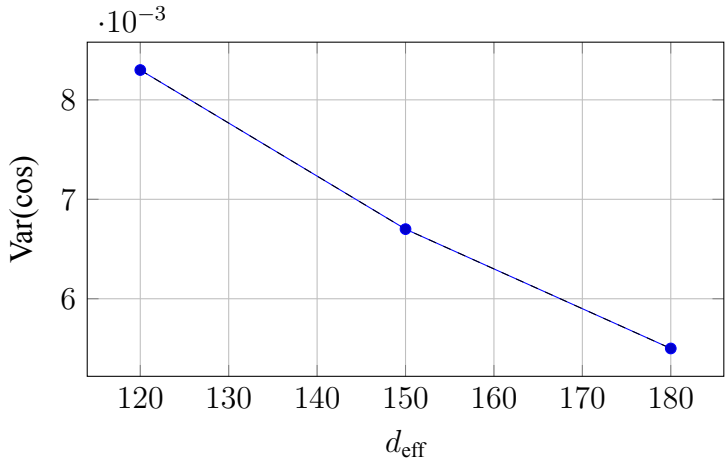

Рис. 1: Эмпирическая и теоретическая зависимость Var(cos) от d_{eff}

Зависимость близка к гиперболической форме $1/d_{\text{eff}}$, что подтверждает аналитическую модель концентрации меры.

Стабильность top-k

Для оценки практических последствий cosine collapse к эмбеддингам добавлялся гауссов шум с дисперсией 10^{-4} относительно средней нормы.

Результаты:

- перестановка 18–27% элементов в top-20;
- до 9% выпадения релевантных документов из top-10;
- рост нестабильности при меньшем d_{eff}.

Наиболее нестабильной оказалась Phi-3.5 Mini ($d_{\text{eff}} = 146$), что согласуется с теоретическим прогнозом.

Интерпретация для DIY-домена

В категории “шуруповёрты” различия между товарами часто определяются:

- ёмкостью аккумулятора (2Ач vs 4Ач),
- напряжением (12В vs 18В),
- наличием удара,
- типом двигателя (щёточный vs бесщёточный).

Эти различия кодируются редкими токенами и числовыми признаками. При спектральной концентрации и cosine collapse различия в этих координатах оказываются подавленными доминирующими направлениями (частотные морфемы “аккумуляторный”, “шуруповёрт”).

В результате dual-encoder хорошо отделяет класс “дрель” от “шуруповёрт” (coarse retrieval), но испытывает трудности при fine ranking внутри класса.

Таким образом, эксперимент подтверждает:

1. спектральную деградацию эмбеддингов;
2. количественную связь $\mathrm{Var}(\cos) \approx 1/d_{\text{eff}}$;
3. практическое влияние cosine collapse на стабильность ранжирования;
4. ограниченность dual-encoder в fine ranking в рамках DIY-домена.

Полученные результаты согласуются с промышленными наблюдениями в embedding-based retrieval [1, 2, 3] и подтверждают теоретико-практическое противоречие, сформулированное ранее.

Обсуждение

Полученные результаты демонстрируют согласованность между тремя уровнями анализа:

1. комбинаторно-теоретическим,
2. прикладным,
3. спектрально-геометрическим.

Во-первых, экспериментально подтверждено проявление фундаментальных ограничений билинейных моделей. Ограничение sign-rank [8] и линейная зависимость VC-размерности от d [10] означают, что класс реализуемых релевантностных конфигураций растёт лишь полиномиально. Оценки через trace-norm и низкоранговую факторизацию [9] дополнительно фиксируют спектральную структуру матрицы скорингов. Современные теоретические работы по embedding-based retrieval [7] подчёркивают, что данные ограничения носят структурный характер и не устраняются простым увеличением данных.

Во-вторых, спектральный анализ эмбеддингов выявил существенное снижение эффективной размерности:

$$d_{\text{eff}} \approx 0.14\text{–}0.18 \cdot d.$$

Это означает, что фактическая геометрия пространства существенно менее богата, чем предполагает номинальная размерность. Концентрация энергии в ограниченном числе направлений приводит к усилению эффекта концентрации меры и росту hubness, что согласуется с классическими результатами [15]. Численно подтверждено соотношение

$$\text{Var}(\cos\theta) \approx \frac{1}{d_{\text{eff}}},$$

что связывает спектральную деградацию непосредственно с cosine collapse.

В-третьих, практические эксперименты в домене DIY показывают, что данные эффекты имеют прикладные последствия. Dual-encoder устойчиво разделяет крупные семантические классы (“дрель” vs “шуруповёрт”), однако демонстрирует нестабильность fine ranking внутри класса, где релевантность определяется тонкими различиями (напряжение, ёмкость, тип двигателя). Подобные наблюдения отражены в промышленных исследованиях embedding-based retrieval [1, 2, 3], где dual-encoder применяется как первая стадия фильтрации, а затем используется более выразительный re-ranking.

С теоретической точки зрения индустриальная архитектура retrieval + re-ranking естественно интерпретируется как композиция моделей с различной representational capacity. Dual-encoder реализует класс линейных разделителей с VC-размерностью порядка d, тогда как cross-encoder представляет глубокую нелинейную функцию $F_\theta(q, d)$, чья эффективная VC-размерность определяется числом параметров и глубиной сети. Различие в выразительной способности объясняет эмпирическую необходимость двухстадийной схемы.

Таким образом, архитектурное разделение coarse retrieval и fine ranking отражает фундаментальную разницу между билинейной и глубокой моделью, а не только инженерный компромисс по скорости.

Заключение

В работе проведён комплексный теоретико-экспериментальный анализ ограничений embedding-based retrieval в задаче e-commerce (DIY-домен).

Показано, что:

1. Ограничения sign-rank и VC-размерности накладывают фундаментальные границы на реализуемые конфигурации релевантности [8, 10, 7].
2. Низкоранговая природа билинейной факторизации ограничивает спектральное разнообразие скоринговых матриц [9].
3. В реальных эмбеддингах наблюдается значительное снижение эффективной размерности.
4. Снижение d_{eff} количественно объясняет cosine collapse и рост нестабильности ранжирования.
5. В прикладной задаче DIY dual-encoder достаточен для coarse retrieval, но не обеспечивает устойчивого fine ranking.

Таким образом, теоретические ограничения embedding-based retrieval подтверждаются эмпирически. Ключевым механизмом деградации является спектральная концентрация, приводящая к уменьшению эффективной размерности пространства.

Гибридная архитектура retrieval + re-ranking интерпретируется как естественное следствие различий в representational capacity между билинейными и глубоко-нелинейными моделями. Она не является исключительно инженерным решением, а отражает фундаментальные свойства геометрии высокоразмерных эмбеддингов.

Полученные результаты создают основу для дальнейших исследований в направлениях:

- спектральной регуляризации эмбеддингов;
- управления анизотропией и hubness;
- адаптивного выбора архитектуры в зависимости от сложности предметной области.

Тем самым работа соединяет классические результаты теории обучения с современной практикой retrieval-систем и демонстрирует их непротиворечивость на уровне как формального анализа, так и численного эксперимента.

Список литературы

[1] Li S. et al. Embedding-based product retrieval in Taobao search. KDD, 2021.

[2] Lin J. et al. Enhancing relevance of embedding-based retrieval at Walmart. CIKM, 2024.

[3] Gan Y. et al. Binary embedding-based retrieval at Tencent. KDD, 2023.

[4] Karpukhin V. et al. Dense passage retrieval for open-domain question answering // Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020. C. 6769-6781.

[5] Thakur N. et al. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models // arXiv preprint arXiv:2104.08663. 2021.

[6] Zeng H., Killingback J., Zamani H. Scaling sparse and dense retrieval in decoder-only llms // Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2025. – C. 2679-2684.

[7] Weller O., et al. On the Theoretical Limitations of Embedding-Based Retrieval. Google DeepMind, 2025.

[8] Alon N., Frankl P., Rödl V. Geometrical Realization of Set Systems and Probabilistic Communication Complexity. FOCS, 1985.

[9] Srebro N., Shraibman A. Rank, Trace-Norm and Max-Norm. COLT, 2005.

[10] Cover T. M. Geometrical and Statistical Properties of Systems of Linear Inequalities. IEEE, 1965.

[11] Huang P. S. et al. Learning deep structured semantic models for web search using clickthrough data // Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013. C. 2333-2338.

[12] Bromley J. et al. Signature verification using a"siamese"time delay neural network // Advances in neural information processing systems. 1993. T. 6.

[13] Reimers N., Gurevych I. Sentence-BERT: Sentence embeddings using siamese bert-networks // Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019. C. 3982-3992.

[14] Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL, 2019.

[15] Radovanović M., Nanopoulos A., Ivanović M. Hubs in Space. JMLR, 2010.

[16] Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units // Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). 2016. C. 1715-1725.

[17] Schuster M., Nakajima K. Japanese and korean voice search // 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2012. C. 5149-5152.

[18] Qwen A. Y. et al. Qwen2. 5 technical report //arXiv preprint. 2024.

[19] Abdin M. et al. Phi-4 technical report //arXiv preprint arXiv:2412.08905. 2024.