A Bag-of-Documents Model for Query Specificity*

Aritra Mandal[1,†], Daniel Tunkelang[2,†] and Zhe Wu[3]

[1]eBay Inc, San Jose, CA, USA
[2]Algolia, Mountain View, CA, USA
[3]eBay Inc, San Jose, CA, USA

Abstract

Search queries differ in specificity, and this variation directly affects retrieval quality and the type of search experience required: broad queries often require exploratory experiences, whereas narrow queries permit more precise ranking. Existing measures of query specificity, such as query length and inverse document frequency (IDF), provide only limited signals. We introduce Bag-of-Documents Specificity (BoDS), a two-stage approach that operationalizes query specificity as *intent coherence*: how tightly a query's engaged documents cluster in embedding space. In the first stage, we compute specificity directly from engagement-derived document vectors using the mean resultant length of a query's associated document vectors. In the second stage, we generalize to all queries by fine-tuning a BERT-based regression model on these scores, achieving an R^2 of 0.84 on the hold-out set. Using logs from a large English-language e-commerce platform, BoDS achieves stronger correlations with click-through rate (CTR; $r = 0.20$) and minimum click rank (MCR; $r = -0.24$) than baselines based on query length or IDF. These results demonstrate that BoDS offers a more robust and holistic measure of query specificity, with applications to query performance prediction, adaptive ranking strategies, and query suggestion.

Keywords

query specificity, bag of documents, e-commerce, search, NLP, query understanding, representation learning, query performance prediction

1. Introduction

The primary goal of search systems is to retrieve and rank results that satisfy the information needs expressed through queries. These needs vary widely in specificity, especially in domains such as e-commerce. Broad queries such as "shirts" offer little signal for precise ranking and often require exploratory search experiences. In contrast, more specific queries like "men's black t-shirts" convey clear intent and enable ranking strategies that minimize user friction. Query specificity thus plays a central role in determining how effectively a search engine can satisfy user needs and how much additional interaction is required.

Beyond retrieval, query specificity also shapes ranking behavior. Ranking models must balance query-dependent relevance against query-independent factors such as popularity and recency. A highly specific query suggests that the user is less willing to deviate from the central intent, while a broad query implies openness to a wider range of results. Specificity influences query suggestions: more specific queries tend to yield higher click-through rates (CTR) and shorter search journeys, improving overall user satisfaction.

Despite its importance, query specificity lacks a standard definition, and existing proxies are limited. Query length is intuitive but unreliable: two queries of equal length can differ greatly in specificity ("shoes" vs. "crocs"). Inverse document frequency (IDF) provides some signal but suffers from two weaknesses. First, rare tokens are not always specific ("weird" vs. "sneakers"). Second, combining IDF scores across tokens either inflates or deflates specificity: summing tends to overestimate (e.g., "apple iphone" vs. "iphone"), while averaging tends to underestimate. Query clarity is a more principled alternative, measuring the coherence of top-ranked results relative to the collection, but it depends on traditional language models and does not leverage modern embeddings.

ECOM'26: SIGIR Workshop on eCommerce, Jul 24, 2026, Melbourne, Australia

*This work is protected by our corresponding patent[1]

†These authors contributed equally.

arimandal@ebay.com (A. Mandal); dtunkelang@gmail.com (D. Tunkelang); zwu1@ebay.com (Z. Wu)

We address these limitations with Bag-of-Documents Specificity (BoDS), a measure that quantifies how coherently the documents associated with a query cluster in embedding space. Building on the bag-of-documents model of Mandal et al. [2, 3], which represents queries as the mean of their relevant document vectors, BoDS uses the concentration of those vectors as a measure of *intent coherence*: the tighter the cluster, the more precisely defined the query intent. We treat intent coherence as our operationalization of specificity. The two are closely related but not identical—taxonomic breadth (*shirts* vs. *men's black t-shirts*) is a primary driver of coherence, but an ambiguous query (*apple*, *weird*) also disperses its document cloud and so registers as low-coherence regardless of its lexical breadth. Intent coherence is thus the quantity BoDS computes directly, and the more defensible description of what it measures. In this formulation, hierarchical breadth and ambiguity are intentionally not distinguished; both manifest as dispersion of the document cloud.

BoDS is computed in two stages. In the first stage, we estimate specificity directly from engagement-derived document embeddings using the mean resultant length of the query's associated document vectors. In the second stage, we extend specificity estimation to all queries by fine-tuning a BERT-based regression model on these scores, enabling specificity estimation even for tail queries with little or no engagement data. Our contributions are as follows:

- We operationalize query specificity as *intent coherence*—the concentration of a query's engaged-document cloud in embedding space—measured by the mean resultant length.
- We present a two-stage framework that combines engagement-derived computation with BERT-based regression for generalization.
- We evaluate BoDS on 3M queries from a major e-commerce platform, showing stronger correlations with CTR and MCR than query length and IDF baselines.
- We situate BoDS within the broader landscape of query specificity research, connecting term-based and clarity-based approaches with modern embedding methods.

Prior work has approached query specificity from several perspectives. Cronen-Townsend et al. [4] introduced the notion of query clarity, defined as the divergence between a query's result distribution and the overall collection. Phan et al. [5] observed a correlation between query length and the specificity of the underlying information need, while Arampatzis and Kamps [6] modeled specificity as the sum or average of query term IDF. Hafernik and Jansen [7] distinguished between "narrow" and "general" queries in web search, and Ke et al. [8] analyzed the effects of specificity on retrieval performance. More recently, Arabzadeh et al. [9, 10] proposed embedding-based metrics for pre-retrieval query performance prediction. Perhaps most relevant to our approach, however, is work not explicitly focused on query specificity but on applying spherical statistics to cluster documents represented as unit vectors [11, 12, 13, 14].

BoDS combines ideas from prior approaches—clarity's coherence, IDF's rarity signal, and embeddings' semantic representation—while addressing several of their limitations by modeling specificity directly in embedding space. In doing so, BoDS operationalizes the classical *cluster hypothesis* [15]—that documents relevant to the same information need tend to be similar—in embedding space: when a query's engaged documents cluster tightly, its intent is coherent and well localized.

The next section details the methodology for computing BoDS, including the formal definition of our coherence-based measure.

2. Methodology

We define Bag-of-Documents Specificity (BoDS) as a two-stage framework, building on the bag-of-documents model[**?**].

Stage 1 computes a direct specificity score from engagement-derived document embeddings. **Stage 2** trains a text-only regression model to predict these scores for all queries, including tail queries with limited or no engagement.

2.1. Problem Setup and Notation

Let a query q have an associated "bag" of engaged documents $D_q = \{d_1, \ldots, d_n\}$ drawn from clicks or conversions in search logs. Each document d_i is represented by an embedding vector $\mathbf{d}_i \in \mathbb{R}^k$. We L_2-normalize all document vectors (so $\|\mathbf{d}_i\| = 1$). Embeddings are produced with a proprietary fine-tuned BERT model [16], but any representation where cosine reflects semantic similarity is acceptable.

We form the **query centroid** as the (un-normalized) mean of its documents:

$$\boldsymbol{\mu}_q \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i, \qquad \hat{\boldsymbol{\mu}}_q \;=\; \frac{\boldsymbol{\mu}_q}{\|\boldsymbol{\mu}_q\|}.$$

Define the **mean resultant length**:

$$R_q \;=\; \|\boldsymbol{\mu}_q\| \;\in\; [0,1].$$

Intuitively, R_q is high when the bag's documents are tightly clustered (specific query) and low when they point in diverse directions (broad/ambiguous query).

With unit-length $\mathbf{d}_i$, the *mean* cosine to the centroid direction equals R_q:

$$\frac{1}{n}\sum_{i=1}^{n}\cos\big(\hat{\boldsymbol{\mu}}_q, \mathbf{d}_i\big) \;=\; \|\boldsymbol{\mu}_q\| \;=\; R_q.$$

Thus, "average cosine(query-centroid, doc)" and "length of the mean vector" are *exactly the same statistic.* Also the **mean pairwise cosine** is a monotone transform of R_q:

$$\overline{\cos}_{\text{pair}} \;=\; \frac{\|\sum_i \mathbf{d}_i\|^2 - n}{n(n-1)} \;=\; \frac{nR_q^2 - 1}{n-1}.$$

Connection to von Mises–Fisher concentration. Under the von Mises–Fisher (vMF) model, unit vectors follow $p(\mathbf{x}) \propto \exp\big(\kappa\,\hat{\boldsymbol{\mu}}^\top\mathbf{x}\big)$ with mean direction $\hat{\boldsymbol{\mu}}$ and concentration $\kappa \geq 0$. The population mean resultant length is $A_k(\kappa) = I_{k/2}(\kappa)/I_{k/2-1}(\kappa)$, where I_ν is the modified Bessel function of the first kind of order ν; since A_k is strictly increasing, R_q and κ encode the same information. Inverting A_k with the standard approximation [11, 12]

$$\hat{\kappa} \;\approx\; \frac{R_q\,(k - R_q^2)}{1 - R_q^2}$$

maps R_q to a concentration parameter that explicitly accounts for the embedding dimension k (here $k = 768$), placing specificity on a dimension-aware scale comparable across embedding spaces. The vMF framework also characterizes the small-sample bias of R_q as an estimator of κ, complementing the finite-n analysis of random bags below.

We adopt R_q itself as our reported measure for its simplicity and linear-time computation, mapping to $\hat{\kappa}$ only when a dimension-aware concentration scale is required.

2.2. Stage 1: Direct BoDS from Engagement

Bag construction. For each query q, collect its engaged documents (clicks and/or conversions). Each distinct engaged document contributes once: we do *not* weight by engagement frequency, so R_q measures the dispersion of a query's intent rather than the popularity of individual documents, and the finite-sample analysis below applies with n the bag size directly (an engagement-weighted centroid would instead require the effective sample size $n_{\text{eff}} = (\sum_i w_i)^2/\sum_i w_i^2$). To ensure a stable estimate, we require a minimum bag size (e.g., $n \geq m$). Our reported correlations (Table 1) use all engaged documents; for the bag-size and popularity diagnostics (Table 2 and Limitations) we additionally cap bags at 100 engaged items to bound the range of n, which changes the correlations only marginally (e.g., MCR $-0.24 \to -0.22$).

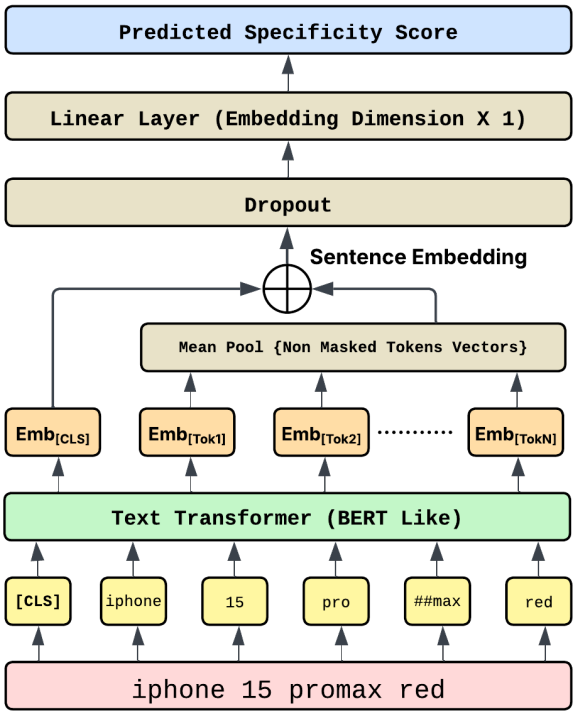


Figure 1: Transformer model architecture of our approach

Specificity score. We define the **Stage-1 BoDS** score as

$$\mathrm{BoDS}_1(q) \;=\; R_q \;=\; \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{d}_i \right\|.$$

Edge cases behave intuitively: with $n = 1$, $\mathrm{BoDS}_1(q) = 1$. A bag of n unrelated documents does *not* drive R_q to zero. For independent draws, $\mathbb{E}[\|\mu\|^2] = \frac{1}{n} + (1 - \frac{1}{n})\,\rho$, where ρ is the mean pairwise cosine of unrelated documents; under isotropy ($\rho = 0$) this gives the familiar $1/\sqrt{n}$ floor, but contextual embeddings are anisotropic ($\rho \approx 0.14$ in our corpus), so a random bag concentrates near $R_q \approx 0.44$ at $n = 15$ rather than 0.

Computation. Linear in bag size: $O(nk)$ to build $\boldsymbol{\mu}_q$, plus one norm. This dominates the naive $O(n^2)$ pairwise alternative without sacrificing information (by the identity above).

Optional weighting (variant). If desired, define weights $w_i \geq 0$ (e.g., clicks, dwell-time, or conversion value), $\sum_i w_i = 1$, and use

$$\boldsymbol{\mu}_q^{(w)} = \sum_i w_i \mathbf{d}_i, \quad \mathrm{BoDS}_1^{(w)}(q) = \|\boldsymbol{\mu}_q^{(w)}\|.$$

This can emphasize highly informative engagements; the rest of the pipeline is unchanged.

2.3. Stage 2: Text-only Regression Generalization

Most queries lack sufficient engagement to directly compute BoDS_1. We therefore train a regression model $\hat{s}(q)$ that predicts specificity directly from the query text, using $\mathrm{BoDS}_1(q)$ as supervision. **Model.** A BERT-family encoder with a single-neuron regression head, trained with MSE:

$$\mathcal{L} \;=\; \frac{1}{|Q|} \sum_{q \in Q} \big(\hat{s}(q) - \mathrm{BoDS}_1(q)\big)^2.$$

We fine-tune on queries with reliable Stage-1 labels (meeting the minimum bag size), using a standard

train/validation split.

Scale & performance: In our implementation, we trained on ∼1.4M labeled queries with a 200K hold-out; the text-only model achieved $R^2 = 0.84$ on the hold-out. This enables robust estimation for head, torso, and tail queries alike. The architecture (Figure 1) maps the 768-dimensional query embedding through dropout and a linear head, trained on 2 A100 GPUs with batch size 128. We use the [CLS] token embedding for the reported results; mean-pooling over query tokens is a supported variant that we did not separately ablate.

2.4. Implementation Details and Defaults

- **Embeddings for documents:** Proprietary fine-tuned BERT model, L_2-normalized; cosine as similarity.
- **Bag thresholds:** Require $n \geq 15$ engaged docs; reported correlations use all engaged documents, with an optional cap at $N = 100$ per query (random or engagement-stratified subsample) used for the bias diagnostics.
- **Label quality:** Larger n reduces the sampling variance of R_q; we filter queries below the minimum to avoid noisy supervision.
- **Complexity:** Stage 1 is $O(nk)$ per query; Stage 2 inference is standard encoder cost per query.

3. Evaluation

A challenge with evaluating any approach to query specificity is that there is no standard definition, which makes human judgments difficult to elicit. Asking annotators to rank queries by specificity presupposes a common scale, but queries with no hierarchical relationship (e.g., "jeans" vs. "black shirt") cannot be placed on one: each expresses a coherent intent, and neither is a generalization of the other. This observation motivates our framing of specificity as *intent coherence* rather than taxonomic breadth: the question is not which query is "more specific" in the abstract, but how coherently each query's engaged documents cluster.

In the absence of clear ground truth, we opted for a different evaluation approach: comparing different ways to compute query specificity based on their ability to predict searcher engagement with results.

We evaluate BoDS by asking two questions:

1. How does it correlate with engagement-based measures of query performance?
2. Does it outperform simple baselines such as query length and IDF?

3.1. Experimental Setup

Data. We used anonymized search logs from a large English-language e-commerce platform. Queries were sampled from one month of activity, covering 3M unique queries with at least one click. For training the Stage-2 regression model, we used 1.4M queries with sufficient engagement (≥ 15 clicked documents), holding out 200K queries for evaluation.

Baselines. We compare BoDS against:

- **Query length** (number of tokens).
- **Inverse document frequency (IDF)**: mean IDF of query tokens, using collection statistics.

Metrics. We use two engagement-based signals as ground truth:

- **Click-through rate (CTR):** fraction of result pages with at least one click.
- **Minimum click rank (MCR):** average rank of the first clicked item (lower is better).

We report Pearson correlation r with both of these engagement signals.

Table 1
Correlations between specificity measures and engagement metrics. BoDS uses all engaged documents per query; baseline measures do not depend on the bag.

Specificity measure	**CTR (r)**	**MCR (r)**
Query length	0.17	-0.04
Sum IDF	0.15	-0.21
Mean IDF	0.09	-0.21
BoDS	**0.20**	$\mathbf{-0.24}$

Table 2
Pearson correlation of R_q with minimum click rank (MCR), raw and debiased, across query-frequency bands. The engagement signal holds within every band, confirming it is not an artifact of query popularity. Bags capped at 100 engaged documents to bound n ($15 \leq n \leq 100$); the capped "All" value (-0.22) is marginally below the uncapped Table 1 value (-0.24). Bottom row: partial correlation controlling for $\log$ frequency.

Band	Queries	R_q	debiased
Head (top 1%)	22,007	-0.14	-0.15
Torso (1–10%)	198,468	-0.19	-0.20
Tail (>10%)	1,979,559	-0.22	-0.23
All	2,200,034	-0.22	-0.22
Partial $r(R_q, \mathrm{MCR} \mid \log \mathrm{freq})$		-0.21	

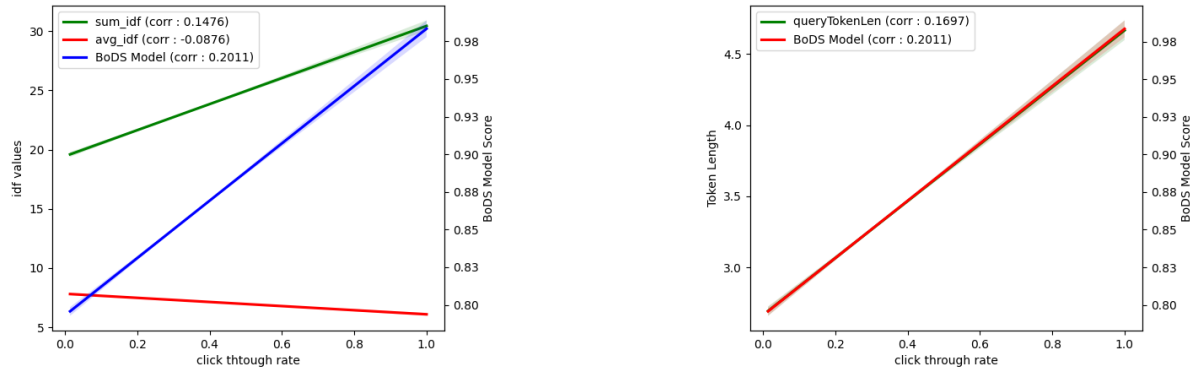
3.2. Correlation with Engagement

Table 1 reports correlations between each specificity measure and engagement signals. BoDS achieves the strongest correlation with both CTR and MCR. The absolute correlations are modest for all measures—engagement is driven largely by ranker quality and result presentation, not query specificity alone—but BoDS improves consistently, if modestly, over query length and IDF. We note further that MCR itself depends on ranker quality, which is a confound when using it as a proxy for a query-intrinsic property. Figures 2a - 3b visually compare the correlations of the IDF measures and token length with CTR and MCR against those of BoDS.

Robustness to query popularity. Because bag size n correlates with query frequency, we verify that the engagement signal is not an artifact of popularity. Stratifying by query-frequency band (Table 2), the negative R_q–MCR correlation holds *within* every band—head, torso, and tail—so it is not driven by between-tier popularity differences, and debiasing leaves it essentially unchanged. The correlation is weaker for head queries (-0.14 vs. -0.22 in the tail), consistent with popular queries being more heavily ranker-optimized, where (as noted above) MCR variance is compressed and reflects the ranker more than the query. Controlling for popularity directly, the MCR correlation is essentially preserved (partial $r = -0.21$ given $\log$ frequency, unchanged when bag size is also partialled out), confirming the signal is largely independent of query popularity. The CTR correlation, by contrast, attenuates under the same control ($+0.17 \rightarrow +0.09$): part of BoDS's relationship with raw click-through is mediated by popularity—unsurprisingly, since click-through rates vary strongly with query frequency—so we regard MCR as the more reliable engagement signal for specificity.

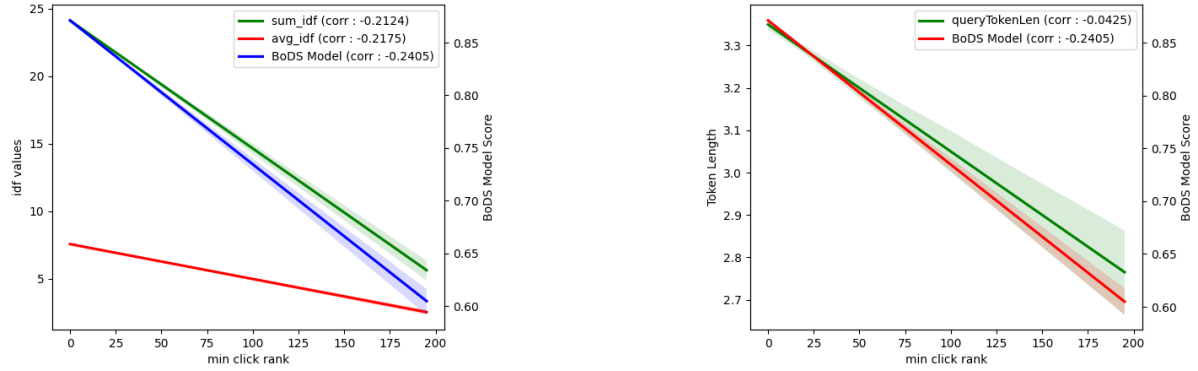
3.3. Regression Generalization (Stage 2)

The Stage-2 regression model generalizes BoDS to all queries, including those with sparse or no engagement. In the 200K set held out, the model achieves $R^2 = 0.84$ against Stage-1 labels. This shows that query text alone provides enough signal for the regressor to approximate specificity as defined by document coherence. Because the regressor operates on query text alone, it applies uniformly to head, torso, and tail queries, extending specificity estimation to queries with little or no engagement data.



(a) (b)

Figure 2: Correlation of CTR with (a) BoDS vs IDF-based measures and (b) BoDS vs token length



(a) (b)

Figure 3: Correlation of MCR with (a) BoDS vs IDF-based measures and (b) BoDS vs token length

3.4. Public Reproduction

The engagement validation above is necessarily on proprietary logs. To establish construct validity independent of proprietary engagement data, we additionally reproduce Stage-1 BoDS on the public ESCI-US dataset [17] using an open-weight MiniLM encoder, forming each query's bag from its *human-judged* Exact-relevant products (21,394 queries with at least two such products). Three findings support the measure. (i) R_q for judged bags sits far above random bags of equal size at every n (Fig. 4); the raw dependence on bag size ($r = -0.24$) is largely a small-sample artifact that the debiased pairwise form $\frac{nR_q^2-1}{n-1}$ removes ($r = -0.02$). Crucially, restricting to the minimum-engagement floor our Stage-1 measure requires ($n \geq 15$), the raw dependence is already negligible ($r = -0.04$) and vanishes under debiasing ($r = +0.00$): the bag-size effect is mechanical, not a confound with the coherence signal. (ii) Coherence computed purely from human judgments behaves as the theory predicts: broadening the relevant set from Exact to Exact+Substitute lowers R_q for 89% of queries (mean $0.815 \rightarrow 0.771$), confirming BoDS is not an engagement artifact. (iii) R_q correlates positively with query length ($r = +0.19$), an independent text-only signal. Random embedding pairs have mean cosine 0.137 (not 0), confirming the embedding anisotropy ($\rho \approx 0.14$) assumed in our edge-case analysis (Section 2.2); mean-centering document vectors restores isotropy (≈ 0.000).

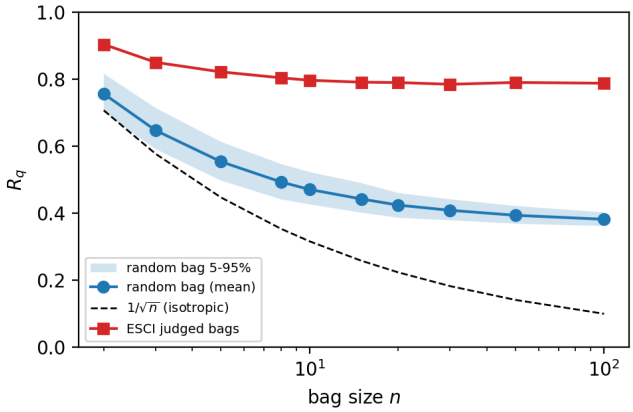


Figure 4: Public reproduction on ESCI (open-weight MiniLM). R_q for human-judged relevant bags (red) sits far above random bags of matched size (blue), which in turn exceed the isotropic $1/\sqrt{n}$ floor (dashed)—the latter gap reflecting embedding anisotropy. R_q thus measures coherence well beyond a bag-size effect.

Limitations

R_q depends on bag size n along two distinct axes that warrant separate treatment. The first is *mechanical*: because a query's documents are each maximally similar to themselves, small bags inflate R_q regardless of intent (the $1/\sqrt{n}$ floor of Section 2.2). We control for this directly by debiasing on n via the pairwise form $\frac{nR_q^2-1}{n-1}$, which strips the self-similarity term; on public ESCI data the residual dependence is negligible ($r = -0.02$ overall, and $r = +0.00$ within the $n \geq 15$ floor), so the bag-size effect is a small-sample artifact rather than coherence signal. On the proprietary logs, within the capped diagnostic regime ($15 \leq n \leq 100$) the dependence is likewise small ($r = -0.11$, explaining less than 1.5% of variance); debiasing removes the mechanical component, leaving $r = -0.09$. Critically, BoDS's engagement signal survives this correction: on the same bags the debiased statistic correlates with minimum click rank at $r = -0.22$, essentially identical to the raw R_q ($r = -0.22$) and far exceeding its -0.09 dependence on bag size—so bag size is not the source of the engagement signal. A residual size dependence persists—larger than on ESCI's human-judged bags—consistent with broader queries genuinely engaging a wider, less coherent set of documents on live traffic; we separate this substantive component from query popularity via the frequency-stratified analysis of Table 2. The second axis is *substantive*: n correlates with query popularity, so R_q may partly encode frequency rather than specificity. Debiasing does *not* address this; we instead control for popularity directly, stratifying the engagement correlation by query frequency and partialling out $\log$ frequency (Table 2). R_q also inherits the anisotropy of the underlying embedding space, which compresses its dynamic range; mean-centering document vectors mitigates this. Finally, "specificity" conflates hierarchical breadth, ambiguity, and intent coherence; BoDS measures the last directly and the others only indirectly.

4. Conclusion

BoDS builds on the bag-of-documents model to operationalize query specificity through the concentration of the documents associated with a query in embedding space. The resulting measure, which intuitively models specificity in terms of coherence around a core query intent, outperforms query length and IDF-based measures as a query performance predictor in terms of correlation with both CTR and MCR.

As we noted, a challenge with evaluating any approach to query specificity is that there is no standard definition for query specificity. Our intent-coherence framing offers one operational definition; we hope to build on it in future work with a public benchmark for intent coherence, backed by human judgments, of which the ESCI reproduction above is a first step.

Ethics Statement

This work uses anonymized, aggregated search log data from a large e-commerce platform. No personally identifiable information (PII) was accessed or retained, and all data were processed in compliance with the platform's privacy policies. Query and click signals were used in aggregate to prevent re-identification of individual users or merchants. The proposed methodology focuses on modeling query specificity, which is unlikely to cause direct harms, but we acknowledge potential downstream impacts if such measures are used to prioritize certain queries or products in ranking systems. We encourage practitioners to evaluate fairness and transparency when applying BoDS in production search systems.

Declaration on Generative AI

During the preparation of this work, the author(s) used used Claude (Opus 4.6) to identify and correct grammatical errors, typos, and other writing mistakes.After using these tools/services, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] D. Tunkelang, A. Mandal, Z. Wu, Query intent specificity, 2025. US Patent App. 18/397,767.

[2] A. Mandal, D. Tunkelang, Z. Wu, Semantic equivalence of e-commerce queries, arXiv preprint arXiv:2308.03869 (2023).

[3] D. Tunkelang, A. Mandal, Z. Wu, Search result identification using vector aggregation, 2024. US Patent 12,124,522.

[4] S. Cronen-Townsend, Y. Zhou, W. B. Croft, Predicting query performance, in: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2002, pp. 299–306. URL: https://doi.org/10.1145/564376.564429. doi:10.1145/564376.564429.

[5] N. Phan, P. Bailey, R. Wilkinson, Understanding the relationship of information need specificity to search query length, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 709–710. URL: https://doi.org/10.1145/1277741.1277870. doi:10.1145/1277741.1277870.

[6] A. Arampatzis, J. Kamps, An empirical study of query specificity, in: Advances in Information Retrieval (ECIR 2010), volume 5993 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2010, pp. 594–597. doi:10.1007/978-3-642-12275-0_55.

[7] C. T. Hafernik, B. J. Jansen, Understanding the specificity of web search queries, in: CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA, ACM, 2013, pp. 1827–1832. doi:10.1145/2468356.2468684.

[8] R. Ke, Y. Liu, M. Zhang, S. Ma, The impacts of query specificity on information retrieval, New Technology of Library and Information Service 32 (2016) 34–42. In Chinese.

[9] N. Arabzadeh, F. Zarrinkalam, J. V. Jovanovic, E. Bagheri, Neural embedding-based metrics for pre-retrieval query performance prediction, in: Advances in Information Retrieval – 42nd European Conference on Information Retrieval (ECIR), volume 12036 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 78–85. doi:10.1007/978-3-030-45442-5_10.

[10] N. Arabzadeh, F. Zarrinkalam, J. Jovanovic, E. Bagheri, Geometric estimation of specificity within embedding spaces, in: W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, J. X. Yu (Eds.), Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019, ACM, 2019, pp. 2109–2112. URL: https://doi.org/10.1145/3357384.3358152. doi:10.1145/3357384.3358152.

[11] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit hypersphere using von mises–

fisher distributions, Journal of Machine Learning Research 6 (2005) 1345–1382. URL: https://www.jmlr.org/papers/v6/banerjee05a.html.

[12] K. Hornik, B. Grün, On conjugate families and jeffreys priors for von mises–fisher distributions, Journal of Statistical Planning and Inference 145 (2014) 35–47. doi:10.1016/j.jspi.2013.10.002.

[13] K. Batmanghelich, A. Saeedi, K. Narasimhan, S. Gershman, D. Parkes, Nonparametric spherical topic modeling with word embeddings, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016, pp. 537–542. URL: https://arxiv.org/abs/1604.00126.

[14] S. I. Nikolenko, Topic coherence in lda-based topic models: An overview, in: Proceedings of the Workshop on Topic Models: Post-processing and Applications, 2016, pp. 1–4. doi:10.18653/v1/W16-0301.

[15] C. J. van Rijsbergen, Information Retrieval, 2nd ed., Butterworth-Heinemann, Newton, MA, USA, 1979.

[16] C. Xue, J. Lute, D. Schonfeld, G. Han, L. Dahlmann, How ebay created a language model with three billion item titles, 2023. URL: https://innovation.ebayinc.com/stories/how-ebay-created-a-language-model-with-three-billion-item-titles/, eBay Innovation blog post.

[17] C. K. Reddy, L. Màrquez, F. Valero, N. Rao, H. Zaragoza, S. Bandyopadhyay, A. Biswas, A. Xing, K. Subbian, Shopping queries dataset: A large-scale ESCI benchmark for improving product search, arXiv preprint arXiv:2206.06588 (2022).