RESEARCH MAY 6, 2026

COVERAGE NOT AVERAGES: RETHINKING RETRIEVAL EVALUATION

By Andrew Klearman, Radu Revutchi, Rohin Garg

Building intelligent agents for the enterprise often requires an information retrieval system over a private, proprietary corpus of information. This corpus may represent the *tribal knowledge* that a company has built over their years of experience or the customer documents. No matter how capable language models become, they cannot reason over information they never receive. As a machine learning engineer, my job is to build systems that consistently gather relevant context for a given business query.

To build these systems, we typically construct evaluation datasets and benchmark retrieval quality over time. Those benchmarks then become the objective function we hill-climb against as we iterate on retriever design choices. Most retrieval benchmarks summarize performance using aggregate metrics such as Recall@k or nDCG@k, averaged across the full corpus of queries. These averages are useful, but they can also obscure important failure modes and uneven system behavior hidden beneath a single headline number.

At Scale, we have seen this effect firsthand. In one customer engagement, we hill-climbed a retrieval system to what appeared to be very strong performance on our internal evaluation benchmark. We had built that benchmark in close collaboration with the customer, using curated question-answer pairs that reflected the intended use case. After saturating the benchmark and presenting the system for dogfooding, we found that real-world queries were not accurately represented in the evaluation set. In effect, we had overfit to a benchmark distribution that was only a weak proxy for true user behavior.

The Problem: Aggregate Metrics Hide Semantic Heterogeneity

This customer engagement motivated the team to rethink how we evaluate retrieval systems in the first place. Retrieval tasks are fundamentally difficult to benchmark because the input space (natural language) is effectively unbounded. A user can ask the system anything, and the expectation is that the system can retrieve the right context regardless of phrasing, intent, or level of specificity.

Each individual query can be thought of as a unit test which examines the system at one specific example. But unlike traditional software, retrieval systems must generalize across an enormous space of requests whose true distribution is often unknown in practice. This uncertainty creates a statistical problem for evaluation. Any benchmark is only a sample of possible user requests, and if that sample does not match the real-world query distribution, aggregate metrics become biased estimates of true system performance.

In our paper, *Coverage, Not Averages: Semantic Stratification for Trustworthy Retrieval Evaluation*, we formalize this mismatch and show how unknown or misspecified query distributions can distort standard corpus-level

evaluations. We propose stratified evaluation as a more useful framework. Rather than treating the benchmark as one homogeneous pool of queries, we partition requests into semantically coherent clusters and measure performance within each group. The target estimand becomes the cluster-level means, rather than a single corpus-level metric dominated by the most common query types.

This shift offers two advantages. First, it reduces sensitivity to benchmark composition by ensuring that niche but important query classes are represented in the evaluation. Second, it makes system behavior more interpretable, revealing where retrieval is strong or weak, and which classes of user intent are being ignored by aggregate metrics.

The quality of the evaluation depends critically on how query regimes are constructed. In the paper, we outline several practical conditions for strong cluster construction:

- **Human-interpretable descriptors:** Each cluster should admit a clear semantic description that humans can understand and reason about. Interpretability is important because the purpose of stratification is not only to measure performance, but also to diagnose failure modes.
- **Corpus-defined intent space:** The document collection itself determines what forms of information-seeking behavior are valid and meaningful to evaluate.

Semantic Stratification Methodology

We find that semantic partitioning makes for strong clusters in practice. We experimented with many approaches for constructing these semantic groupings, and the key takeaway was that clustering over extracted entities yielded the cleanest and most interpretable clusters. Here we present the methodology of how we approach creating clusters, but more details can be found in the paper.

1. Extract entities from documents

We first parse each document and identify salient entities such as people, organizations, products, locations, and technical concepts. These entities provide a structured view of what information exists inside the corpus.

2. Deduplication

Many entities appear under multiple surface forms (“NYC” vs. “New York City”, abbreviations, synonyms, alternate spellings). We merge equivalent mentions into a unified entity set. In our experiments, we found that thorough deduplication was important for producing semantically coherent clusters.

3. Build a semantic graph

We then create a graph where nodes are entities and edges connect entities that are semantically related or frequently co-occur across documents. This transforms the corpus into a map of connected concepts. Densely connected components represent semantic themes (not just concepts) that exist commonly in the corpus.

4. Detect coherent communities

Finally, we run community detection over the graph to identify clusters of related entities. These communities often correspond to natural information-seeking regimes within the corpus: product troubleshooting, regional policy questions, people lookups, clinical treatments, and so on.

From Corpus Structure to Evaluation

Once these semantic communities are identified, benchmark queries can be assigned to the cluster they most naturally belong to. We then evaluate retrieval performance within each cluster, rather than collapsing all of our evaluation signal into one corpus-wide average. This gives us a much deeper signal for evaluation, which can be used to better understand system performance and to support faster iterative improvement.

Case Study: NFCorpus

To better understand the impacts of stratified evaluation, we apply the framework to a public IR benchmark called NFCorpus. NFCorpus is a nutritional science benchmark that is commonly used to evaluate retrieval systems on technical language, specialized terminology, and domain-specific documents. We organized the corpus into 326 semantic clusters and map each query into its relevant clusters. We then evaluated several retrievers across these clusters. Performance varied substantially; every system had regions of strength and weakness that were hidden by aggregate benchmark scores.

Looking at cluster level statistics helped us understand the relative strengths and weaknesses of the different retrievers, while also highlighting genuinely difficult regions of the corpus. For example, the Nutrition Expert cluster proved challenging for all three retrievers, potentially indicating issues with the labeled relevance judgments or a systematically hard retrieval regime. Other clusters showed clear specialization effects, where certain retrievers consistently outperformed others depending on the semantic content of the query set.

We also observed the semantic composition of the corpus, as well as the coverage that the benchmark test observed. Many large regions of the corpus received little or no evaluation coverage, while a small number of topics dominated the benchmark.

We found that many semantically large regions of the corpus received little or no benchmark coverage, while specific clusters such as *Tumor Biology* (roughly 30% of all queries) received disproportionate attention. In other words, the benchmark emphasized a narrow subset of the corpus while leaving large semantic regions largely untested.

For practitioners like us, this creates a misallocation of optimization effort. Teams naturally improve whatever the benchmark measures. If some semantic regimes are overrepresented while others are absent, engineering work will concentrate on the evaluated slices of the problem.

This is exactly the failure mode we encountered in our customer engagement. Benchmarks with low semantic coverage introduce real deployment risk. Users do not constrain themselves to the benchmark distribution, they ask questions across the full intent space of the corpus. Once usage shifts outside the evaluated regions, teams are operating in uncharted territory.

To understand whether these uncovered regions in NFCorpus were harmless omissions or genuine blind spots, we ran an experiment measuring retrieval performance directly on clusters that received little or no benchmark coverage. We generated new queries within these underrepresented clusters and labeled relevant documents from the corpus.

Some clusters, such as Cancer cell lines, showed relatively strong performance across retrievers, while others, like Statistical methods, were significantly more challenging.

[Table: see online version]

Some previously uncovered regions performed well, while others exposed clear weaknesses. This helps explain how strong aggregate scores can mask uncertainty in real-world deployments: systems may appear reliable overall while their behavior in unevaluated parts of the corpus remains unknown.

By leaving these regions unevaluated, aggregate metrics hide uncertainty about system behavior in large portions of the corpus. This creates a fundamental gap between what we measure and what we actually need to know to deploy retrieval systems with confidence. A single benchmark score suggests stability and completeness, but in reality it may reflect performance on a narrow and biased slice of the problem. For practitioners, this means we lack visibility into where systems generalize, where they break, and how they will behave as user queries inevitably expand beyond the original evaluation set.