CHROMA TECHNICAL REPORT

April 07, 2025

Generative Benchmarking

Kelly Hong Researcher - Chroma

Anton Troynikov Cofounder, Advisor - Chroma

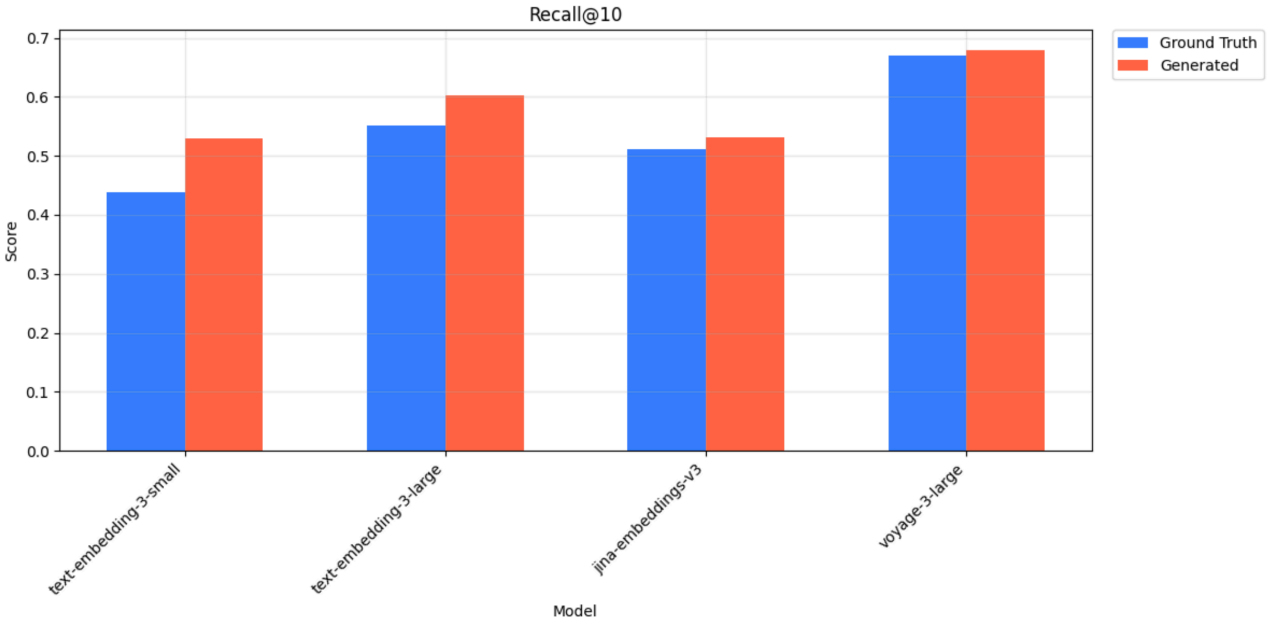
Jeff Huber Cofounder, CEO - Chroma

Morgan McGuire Director of Applied AI - Weights and Biases

In traditional software systems, evaluation typically relies on deterministic logic: given a fixed input, the output is known and reproducible. In contrast, AI systems produce probabilistic results, and their evaluation is often context-dependent and subjective—making simple unit tests insufficient for tasks such as document retrieval. As a result, evaluating these systems rely on benchmarking performance across many examples.

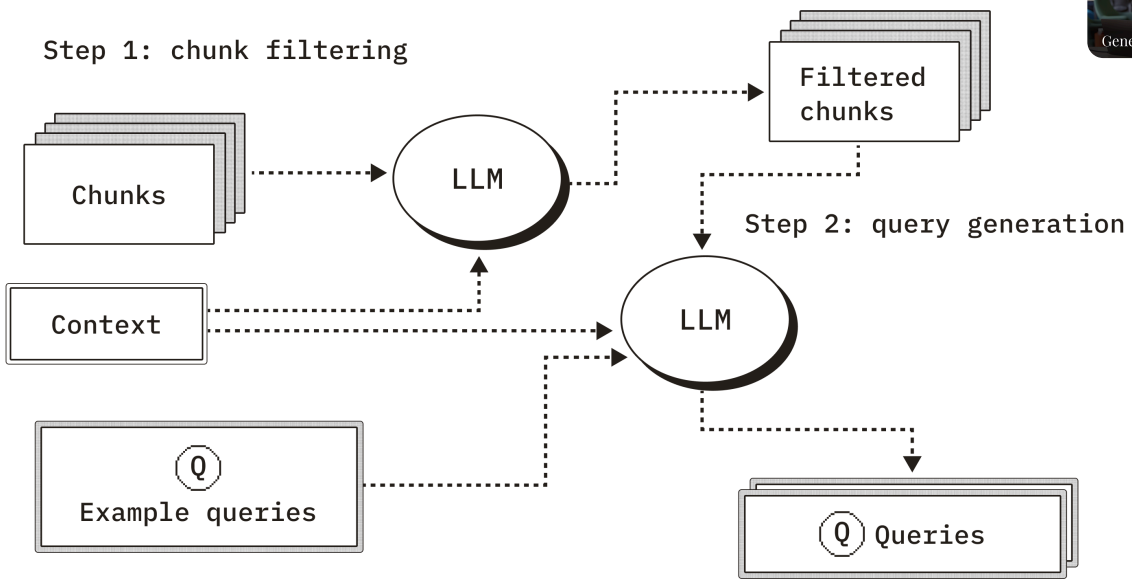
However, widely-used public benchmarks often rely on artificially clean datasets and generic domains, with the added concern that they were likely seen by embedding models in training. Prior efforts such as RAGAS and AIR-Bench have used synthetic testset generation to address such limitations, primarily optimizing for dataset diversity.

demonstrate that our generated queries reflect real user queries and that they can capture performance differences that public benchmarks may miss.



Recall@10 scores comparing generated queries to ground truth queries from Weights and Biases' production data. We show that our generated queries are representative of production traffic.

A core limitation of current benchmarks is that they often fail to accurately reflect the actual use cases of their evaluated models. Despite this, there exists a common misconception that strong performance for a model on a public benchmark directly generalizes to comparable real-world performance. A model's performance on a public benchmark is often inflated by polished datasets that lack the ambiguity of production scenarios, as well as potential memorization of the retrieval task from seeing the benchmark already in training.



Overview of our query generation process.

Our generative benchmarking method addresses these limitations with a more tailored and representative approach to evaluation. We first begin with a document filtering step, where we use an aligned LLM judge and given context from the user to identify documents that are most relevant to the specified use case and contains sufficient information to generate queries from. This step ensures that we focus on documents that users will realistically query. Next, we generate queries using given context and example queries to steer generation. Providing these details steers the LLM in a way that better aligns with real user queries compared to naively generated queries without these contextual details.

We present this method alongside experiments focusing on the representativeness of these generated queries in how they compare to the ground truth.

Our in-depth technical report continues below. If you find our work useful, please consider citing us:

plaintext Copy Code

```
@techreport{hong2025benchmarking,
  title = {Generative Benchmarking},
  author = {Hong, Kelly and Troynikov, Anton and Huber, Jeff and McGuire, Morg
  year = {2025},
  month = {April},
  institution = {Chroma}
```

Interested in working on improving retrieval for AI applications? Chroma is Hi

Introduction

Benchmarking serves a critical role in measuring the performance of any machine learning system. In practice, models are often evaluated using established popular benchmarks to assess their performance in a reliable and standardized way.

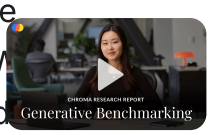
Specifically in information retrieval, embedding models and retrieval pipelines are commonly evaluated using well-known benchmarks such as the Massive Text Embedding Benchmark (MTEB) [1] and Benchmarking Information Retrieval (BEIR) [2]. These public benchmarks provide a convenient way to quantify and compare model performance through a unified metric. However, the current reliance on these benchmarks presents several critical issues that can compromise their effectiveness:

1. These datasets are **generic**, which fail to capture the domain-specificity of real-world retrieval applications.
2. The data is also **overly clean**, often containing polished queries and documents which contrasts with the messiness of real data.
3. Models have already **seen** most of these benchmarks in their training data, which may lead to inflated performance since they have learned the structure of the task already.

Consequently, we cannot exclusively use these publicly available benchmarks to evaluate the performance of a given embeddings-based retrieval system in a real-world context. Strong performance on a public benchmark for a given embedding model does not guarantee comparable performance for your specific production pipeline.

This creates a need for benchmarks tailored to a specific user's data, in a way that

Generative benchmarks are specific to a user's data and thus allow for a more accurate and representative evaluation of a retrieval system's performance. W address the previously mentioned problems of current benchmarking method



1. This dataset is **customized** to the user's documents, ensuring relevance to their use case.
2. It works with **real**, messy data with ambiguous queries.
3. This data is truly **unseen** since it is generated based on the user's documents, which helps test generalization abilities without the possibility of memorization.

Contributions

We present the following:

- A demonstration that LLMs cannot reliably generate unseen queries for public datasets. We present samples of reproduced queries—either verbatim or slightly reworded—across all 9 datasets we tested.
- Experiments with query generation showing that we are able to generate unseen yet representative queries for widely accepted benchmarks, validated by consistent embedding model rankings and query-document relevance distributions.
- A generative benchmarking method and its application to a real production use case from Weights and Biases [3]. We demonstrate that our generated queries yield model performance metrics that closely align with those of the real queries and capture performance differences that MTEB fails to reflect.
- The complete codebase to replicate our results and generate a custom benchmark for your own data.

Related Work

comprehensively cover the domains encountered in real-world RAG applications such as customer support assistants or technical documentation bots.

The data used for these benchmarks is also artificially clean, meaning that they often contain polished query-document pairs constructed for the purpose of question answering tasks. This contrasts with the ambiguity of real data, where queries are often vague with only partial matches to their relevant documents. Lastly, most embedding models have likely seen these benchmark datasets during training, which makes it difficult to distinguish true retrieval capabilities from memorization.

Synthetic dataset generation for information retrieval has been an active area of research, such as the work by InPars [4] and Promptagator [5], which leverage few-shot prompting for generating synthetic training data. These methods are aimed at improving retrieval models, often focusing on improvements in metrics such as Recall. This contrasts with our motivation to evaluate models in a more realistic manner rather than to improve them.

RAGAS [6] and AIR-Bench [7] are more aligned to our focus on evaluation, as both aim to generate testsets for evaluating retrieval in real-world scenarios. Both approaches focus on the diversity of generated testsets, with RAGAS focusing on diversity in query type and AIR-Bench aiming for diversity in domains beyond public benchmarks. However, an area that has not been extensively explored in prior work is how well these synthetically generated queries represent real user queries from production. Our experiments focus on synthetically generating queries that are representative of the ground truth, which we believe is the objective of a golden dataset when evaluating retrieval systems.

In our work, we also employ Large Language Models (LLMs) as judges for labeling tasks. LLM judges allow for a cost-effective and consistent way of labeling data, however, they come with known problems around alignment. We cannot guarantee that LLMs will have the same judgments as humans would, largely due to their high sensitivity to minor changes in prompting and the difficulty in articulating ambiguous concepts such as “relevance”.

To better align our LLM judge with human judgements, we use an adapted version of EvalGEN [8]. EvalGEN is a framework for validating LLM outputs through iterating on a set of criteria based human inputs. We use this process to align our LLM judge for document filtering, the first step in our generative benchmarking process.

The first phase of our work focuses on addressing the following question:

Do generated queries reflect the evaluative capabilities of established benchi

We begin by working with public datasets from MTEB. The objective of this phase is to validate that we can generate queries that are representative of the current standard for retrieval evaluation. This enables us to leverage established metrics on embedding model performance within MTEB, which are unavailable for private datasets.

We employ two key methods for comparing our synthetically generated queries with MTEB:

- Cosine similarity distributions for query-document pairs
- Retrieval metrics, such as Recall@k and NDCG@k

For robustness, we perform this analysis across 9 datasets spanning various domains and languages, using 5 different embedding models.

We then extend synthetic dataset generation to production data:

How do we generate queries that accurately reflect real use cases?

We work with a retrieval dataset from Weights and Biases [3], which includes the technical documentation and production queries from their technical support bot. We experiment with various query generation methods to capture the characteristics of real user queries in production settings; filtering documents for quality, and steering queries. We compare these generated queries against production queries to assess our method in a real-world scenario.

All prompts used for document filtering and query generation are provided in the Appendix.

Generating Representative Unseen

We first demonstrate our ability to generate unseen queries that are representative of widely-used benchmarks.

Models

LLM

claude-3-5-sonnet-20241022

Embedding Models 1

text-embedding-3-small

text-embedding-3-large

jina-embeddings-v3

voyage-3-large

Datasets 2

Wikipedia Multilingual (en, hi, de, pt, fa, bn) [9] 3

For each language, the corpus includes snippets from Wikipedia pages paired with relevant queries. This is a generic Wikipedia dataset covering 16 different languages, out of which we chose the 6 indicated above for linguistic diversity.

LegalBench Consumer Contracts QA [10] 4

Contains sections of consumer contracts paired with relevant queries, representing a specialized legal domain.

SciFact [11]

Contains claims that can be supported or refuted by scientific literature. Only supported (true) claims were used for fair evaluation.

MedicalQA [12]

Contains question and answer pairs related to medical conditions, treatments, and protocols

Naive Query Generation

We first demonstrate that LLMs have memorized a substantial portion of these public benchmarks, thus we cannot reliably generate unseen queries with a naive approach.

Method

For each dataset, we prompt Claude 3.5 Sonnet to simply generate a query given a document. 5

We then embed ground truth (queries from original dataset) and generated queries using various embedding models, examining cosine similarity distributions with target documents and performing manual inspections.

Results

Across **all datasets**, we consistently observe cases where the generated queries are identical or near-identical to ground truth queries. In cases of near-identical matches, the generated queries essentially reword the ground truth queries.

```
Target Document: Deliverance was shot primarily in Rabun County in northeastern
Georgia. The canoe scenes were filmed in the Tallulah Gorge southeast of Clayton and
on the Chattooga River. This river divides the northeastern corner of Georgia from the
northwestern corner of South Carolina. Additional scenes were shot in Salem, South
Carolina.

Ground Truth Query: Where was the movie Deliverance filmed?

Generated Query: Where was the film Deliverance shot?
```

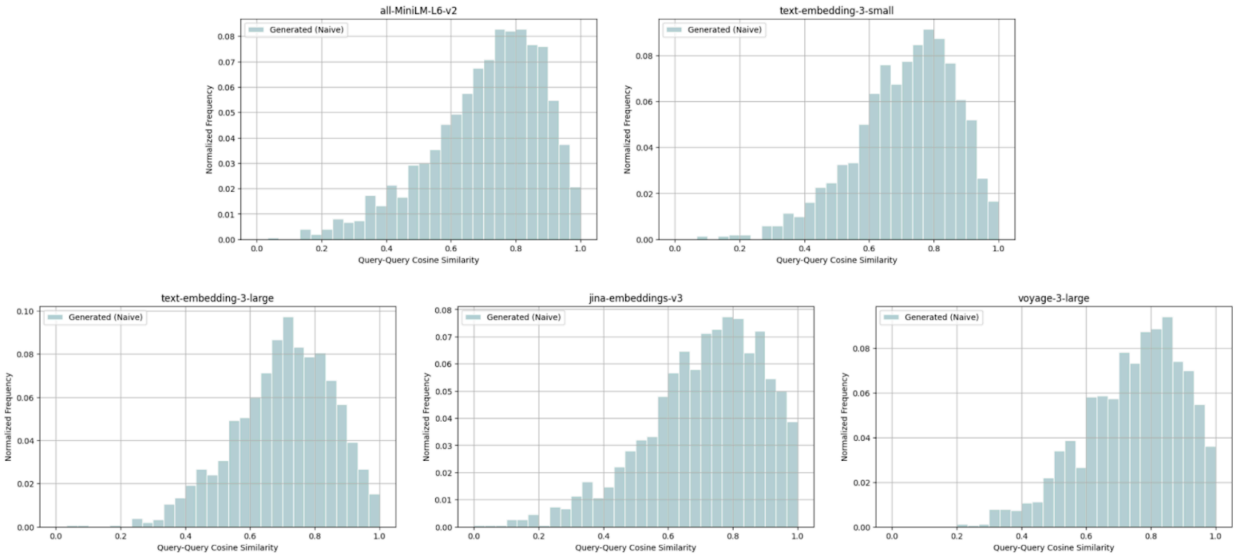
Wikipedia Multilingual (en) - example of a near-identical match between the ground truth and generated query.

To further analyze this similarity between our synthetically generated queries and ground truth queries, we embed them across all 5 embedding models and compute their cosine similarity scores. In the case above for example, we would embed the ground truth query (“Where was the movie Deliverance filmed?”) and its corresponding generated query (“Where was the film Deliverance shot?”), then compute the cosine similarity between the two embeddings (which is 0.973 with text-

It is important to note that high query-query cosine similarity scores are not desirable in this context. Here, our goal is not to measure representativeness by how semantically close a generated query is to an existing one—rather, we are ch whether the LLM has reproduced or closely mimicked a query it has seen during training. Since different queries can be relevant to the same document, semantic diversity is expected and even encouraged.

In later sections, we evaluate representativeness using query-document (instead of query-query) similarity and retrieval metrics, which better capture the grounding of generated queries.

With the aim to see how often our LLM generated memorized queries, we plot the distribution of these cosine similarity scores.



Wikipedia Multilingual (en) - query-query cosine similarity scores between: ground truth query & naively generated query.

For this Wikipedia Multilingual (en) dataset, we observe the following distribution of cosine similarity scores across 5 embedding models:

- **11.91%** of query pairs above 0.9 cosine similarity, which include the identical and near-identical query pairs.
- **65.89%** of query pairs between 0.6 and 0.9 cosine similarity, in which the generated queries tend to be more specific than ground truth queries.
- **22.20%** of query pairs below 0.6 cosine similarity, where we notice a substantial difference between the generated and ground truth queries.

- The document is missing context or contains incomplete sentences.
- The query is not relevant to the content of its target document.


```
Target Document: At the other side of the river I have my sand bank, where sits my
darling short one, with the beak of a great blue heron.

Ground Truth Query: What is the setting described in the document "Lotería"?

Generated Query: What kind of bird has the beak that the person's beloved resembles?

Query-Query Score (text-embedding-3-small): 0.068
```

Wikipedia Multilingual (en) - query-query pair with a cosine similarity score below 0.6.

Our findings demonstrate that while LLMs often naively generate unseen queries—meaning that they are capable of generating diverse queries without explicit prompting—there are cases where they reproduce queries from public benchmarks. This limitation highlights the need for a more refined approach to ensure reliable generation of truly unseen queries.

Distinct Query Generation

We generate distinct queries by incorporating both the target document and ground truth query into the LLM prompt—the ground truth query serves as a negative example. We first demonstrate that these queries are unseen, then demonstrate that they are also representative of the ground truth benchmark.

Method

For each dataset, we prompt Claude 3.5 Sonnet with both the target document and ground truth query. We explicitly prompt the LLM to generate a query distinct from the given example query.

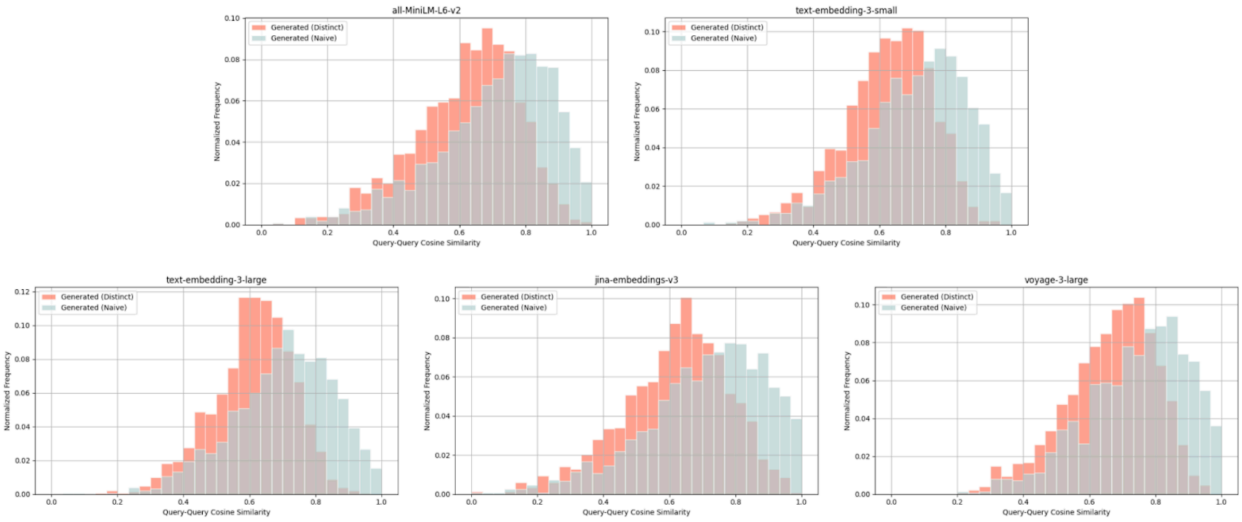
We initially validate that these newly generated queries are truly unseen and distinct from the ground truth through query-query cosine similarity comparisons and manual sampling of queries. Next, we assess the representativeness of these unseen queries against the ground truth through retrieval metrics and query-document cosine similarity distributions.

Results

Generated queries are distinct and unseen

We explicitly prompt the LLM to generate queries distinct from existing ground truth queries. For each model and dataset, query-query cosine similarity distributions are compared between:

- Ground truth query & naively generated query (blue)
- Ground truth query & distinct generated query (red)



Wikipedia Multilingual (en) - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red)

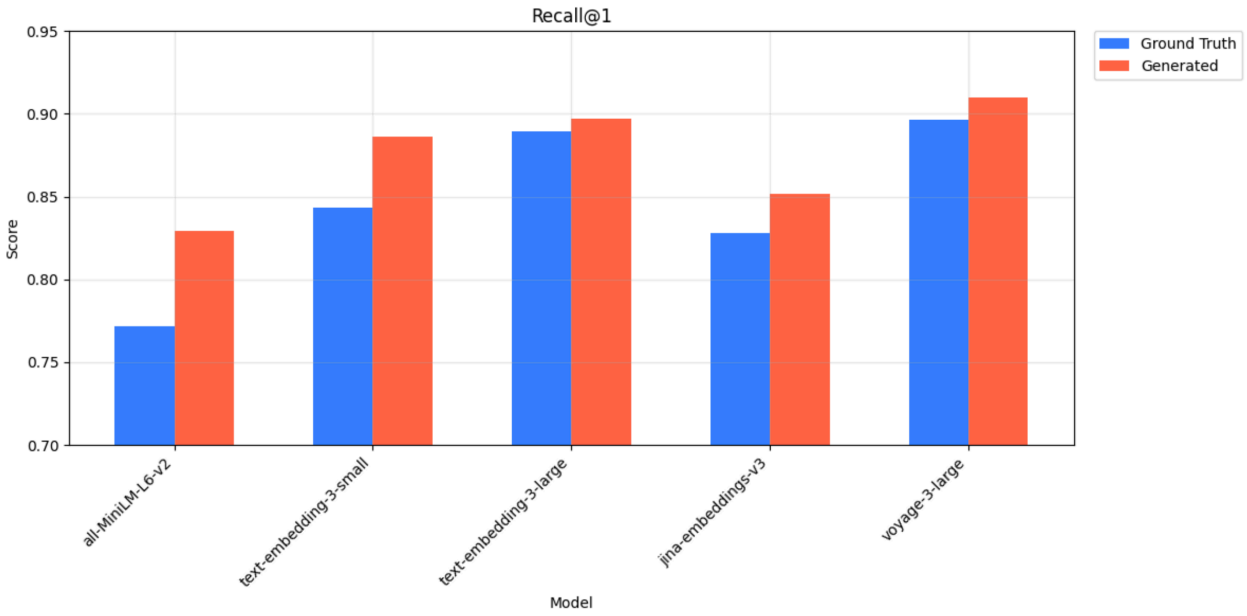
For each distribution, we average across all 5 embedding models and consistently observe lower cosine similarity scores for distinct queries (red) compared to naive queries (blue). With such query-query comparisons, we look for lower cosine similarity scores as an indication of distinctiveness.

For example with the Wikipedia Multilingual (en) dataset, the average cosine similarity score for a ground truth query & distinct generated query is **0.628**, whereas that of a ground truth query & naively generated query is **0.716**.

Generated queries are representative of the ground truth

Further evaluation reveals that these distinct generated queries are not only unseen, but they are also representative of the ground truth datasets.

tasks, we observe similar relative performance across embedding models. For example, text-embedding-3-large on Wikipedia Multilingual (en) has a higher Recall@1 score when compared to all-MiniLM-L6-v2 for both ground truth and generated queries.



Wikipedia Multilingual (en) - Recall@1 scores for ground truth and generated queries across all 5 models.

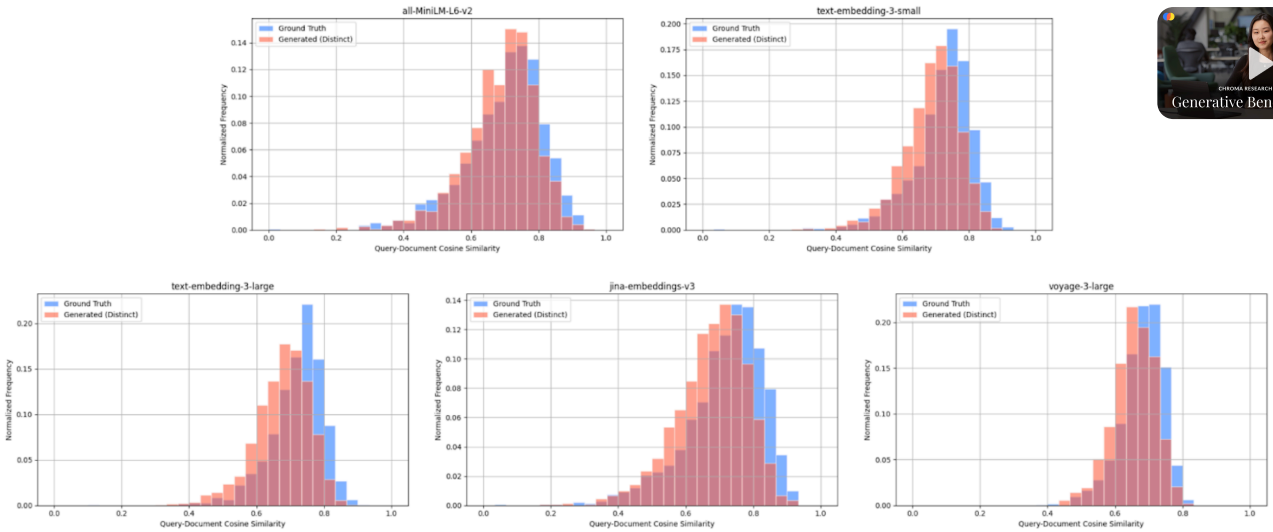
Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10
Ground Truth	all-MiniLM-L6-v2	0.772	0.901	0.936	0.962
Generated	all-MiniLM-L6-v2	0.830	0.927	0.950	0.957
Ground Truth	text-embedding-3-small	0.843	0.945	0.967	0.990
Generated	text-embedding-3-small	0.886	0.966	0.981	0.991
Ground Truth	text-embedding-3-	0.889	0.969	0.988	0.997

Query Type	Model	Recall@1	Recall@3	Recall@5	Rec
Generated	text-embedding-3-large	0.897	0.977	0.984	0.989
Ground Truth	jina-embeddings-v3	0.828	0.933	0.955	0.979
Generated	jina-embeddings-v3	0.852	0.935	0.959	0.967
Ground Truth	voyage-3-large	0.897	0.959	0.972	0.983
Generated	voyage-3-large	0.910	0.961	0.967	0.968

Wikipedia Multilingual (en) - Recall@k scores for ground truth and generated queries across all 5 models.

The generated queries often exhibit higher specificity and relevance to the provided target document compared to the ground truth queries, which naturally boosts retrieval performance across all embedding models. Thus, we focus on the **relative** evaluation of performance between embedding models as a proxy for representativeness rather than on absolute values. These relative differences produce consistent rankings of embedding models regardless of whether the queries were generated—which reflects the primary goal in real use cases where embedding models are selected based on relative performance.

We also compare the query-document cosine similarity distributions for ground truth queries and generated queries, across all embedding models. These distributions represent the diversity in relevance of query-document pairs, thus similar distributions indicate similar levels of diversity.



Wikipedia Multilingual (en) - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).

To quantify how similar these distributions are, we compute the Kullback-Leibler (KL) divergence for each model. KL divergence is a measure of how one probability distribution differs from another—lower values indicate that the two distributions are more similar. Using this metric, we observe that the cosine similarity distributions are closely aligned. For example, we observe the following results for KL divergence with the Wikipedia Multilingual (en) dataset:

- all-MiniLM-L6-v2: **0.0532**
- text-embedding-3-small: **0.193**
- text-embedding-3-large: **0.217**
- jina-embeddings-v3: **0.165**
- voyage-3-large: **0.129**

We note some variability—primarily due to the varied nature of individual datasets—such as the SciFact dataset containing ground truth queries only partially relevant to their target documents which leads to lower ground truth query-document cosine similarity scores relative to generated query-document scores.

Overall, our refined query generation strategy consistently yields unseen queries **representative** of established benchmarks across various datasets and models. With this confidence in representativeness, we move onto query generation with real data.

Extension to Real Production Data

We work with a retrieval dataset from Weights and Biases [3], a developer platform for building AI applications and models. Specifically, we utilize data from **WandBot**, their technical support bot for handling queries related to their documentation.

Our approach to generative benchmarking involves two primary steps:

- Document Filtering: We initially filter documents using an aligned LLM judge. This step is critical in selecting documents relevant to users' potential queries, ensuring that generated queries would be both contextually meaningful and realistic.
- Query Generation: From the filtered documents, we generate one query per document. We use contextual information and representative example queries to steer the LLM in capturing genuine user intent and maintaining consistent query style.

We validate each step of our generative benchmarking process against ground truth labels as a test for representativeness.

Ground Truth Labeling

Documents

We randomly select 250 documents and manually label each as "good" if they are suitable for query generation, meaning that they are relevant and informative based on the provided context.

Queries

To establish that our generated queries are truly representative, we work with 2023 production queries from WandBot which serve as our ground truth. After deduplication, we have 2003 unique queries.

Due to the labor-intensive nature of manual labeling, we manually examine 693 out of 2003 queries, and account for the rest of the queries using weighted representation.

This process involves:

HDBSCAN [14] to be grouped into semantically similar clusters.

- Query selection: we use Maximal Marginal Relevance (MMR) to select the most representative queries from each cluster, which balances representativeness with diversity.

We end up with 360 representative queries (72 clusters, 5 queries per cluster). Additionally, we randomly sample 333 queries outside these representative clusters for broader coverage.

Each query underwent the following process for manual labeling:

- Retrieve 10 documents from each of the 4 embedding models.
- We manually assess for whether one of those retrieved documents is relevant, if so, then label that document as the query's matching pair.
- If no relevant document is retrieved, we manually search the **Weights and Biases' documentation** corpus to either identify false negatives or confirm true negatives.

After manually going through 693 queries, we end up with:

- 560 queries with relevant document pairs (including 76 false negatives)
- 133 queries were true negatives

To reflect the frequency of query topics in production, we assign weights to each query based on its cluster:

- Queries labeled as noise by HDBSCAN (i.e., not part of any cluster) are given a static low weight of 0.1.
- All other queries are weighed proportionally to their cluster's size:
  $$\text{query weight} = \frac{\text{cluster size}}{\text{mean cluster size}}$$

Then during evaluation, we multiply each query's weight to its metric scores (Recall@k, NDCG@k, etc.). We then sum the weighted scores across all queries and normalize by the total weight. This ensures that common query types contribute more to the final metrics, aligning evaluation with production query distribution.

Document Filtering

Before generating queries, we apply a filtering process to select documents that are both contextually relevant and informative. Our goal is to identify documents genuinely reflective of the needs of Weights and Biases users.

Certain documents, such as video transcripts and general news content, are deemed irrelevant for query generation due to their misalignment with actual user inquiries. For instance, documents like startup funding announcements are excluded, as they do not directly relate to the technical documentation or user support context of WandBot.

```
# Finbots.AI Raises $3 Million To Bring AI Into Banking

Description: In a round of Series A funding, Finbots.AI has raised $3 Million with
plans to expand business and it's ZScore service for enhanding credit score
calculation with AI.

Body:

[Finbots.AI](https://finbots.ai/) sees a spot in the market for the use of AI in
banking, particularily in the fair calculation of credit scores. Many prospective
borrowers might face difficulty with legacy platforms and practices which determine
they're too high risk, when they're actually perfectly suitable to borrow

Finbots.AI is offering a solution, bringing what they claim is a fair and balanced
method into the credit scoring business with ZScore, their advanced AI-powered
solution for credit score calculation. ZScore works across the whole credit cycle and
offers a number of easily accessible capabilities to help assign credit scores and
manage borrowers.

## Who's funding Finbots.AI and where is the money going?

## Find out more
```

Example of a filtered out document from WandBot's corpus.

Method

Using our previously established ground truth labels for document quality, we align an LLM judge to human assessments through an iterative approach guided by the EvalGEN [8] framework.

We define our criteria with relevance, completeness, and intent. Then, we prompt Claude 3.5 Sonnet to evaluate each document for each criterion.

document-criterion pair against the human's overall label. This gives us per-criterion alignment scores, as well as an overall alignment score based on whether the final classification matches the human judgment. We iteratively refine our crit improve alignment.

Results

We see significant improvements in alignment after 5 iterations over our 250 labeled documents, increasing from an initial baseline of **46%** to **75.2%**. The final alignment scores per criterion are:

- Relevance: 64%
- Completeness: 65.6%
- Intent: 56.8%

Applying the finalized criteria across the broader corpus, we filter down from an initial set of 13,319 documents to a refined dataset of 8,490 documents, which serves as the basis for generating synthetic queries.

Query Generation

Using the filtered set of documents, we generate queries tailored to reflect actual user queries. We accomplish this by incorporating both contextual information and example queries provided by users, then comparing these queries with our ground truth.

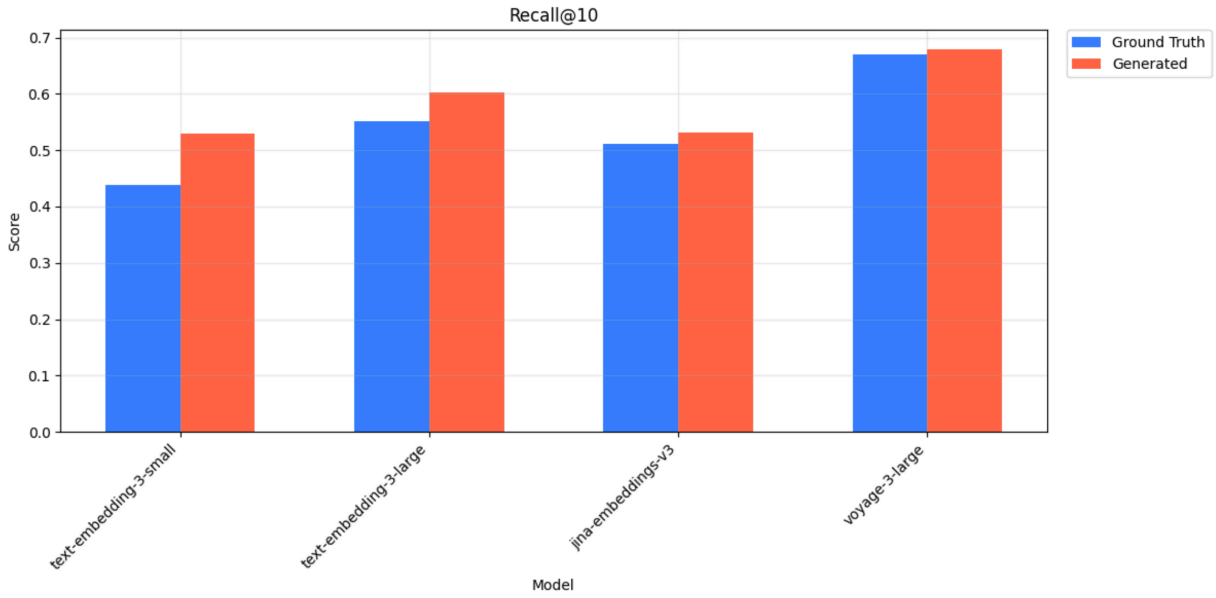
Method

We prompt Claude 3.5 Sonnet with each filtered document, along with context and example queries specific to Weights and Biases.

We validate the representativeness of our generated queries by directly comparing retrieval performance metrics with ground truth queries. Additionally, we com these metrics and cosine similarity distributions against a naive baseline, whe queries are generated without any guiding context or example queries.

Results

By comparing retrieval performance between ground truth queries and our generated queries, we confirm that our generated queries generally capture the relative performance differences among embedding models. This aligns with our observations from our earlier analyses on public datasets, where we also noticed this preservation of embedding model rankings across both query types.



WandBot - Recall@10 scores for ground truth and generated queries

Query Type	Model	Recall@10	NDCG@10	Precision@10
Ground Truth	text-embedding-3-small	0.439	0.282	0.044
Generated	text-embedding-3-small	0.530	0.397	0.053
Ground Truth	text-embedding-3-large	0.552	0.356	0.055

Query Type	Model	Recall@10	NDCG@10	Precis
Ground Truth	jina-embeddings-v3	0.511	0.341	0.0511
Generated	jina-embeddings-v3	0.532	0.389	0.0532
Ground Truth	voyage-3-large	0.670	0.402	0.0670
Generated	voyage-3-large	0.679	0.512	0.0679

WandBot - metrics for ground truth and generated queries. We observe a shift between the ranking for text-embedding-3-small vs jina-embeddings-v3 for NDCG@10; however, the performance scores are close in this case and we generally see consistent rankings.

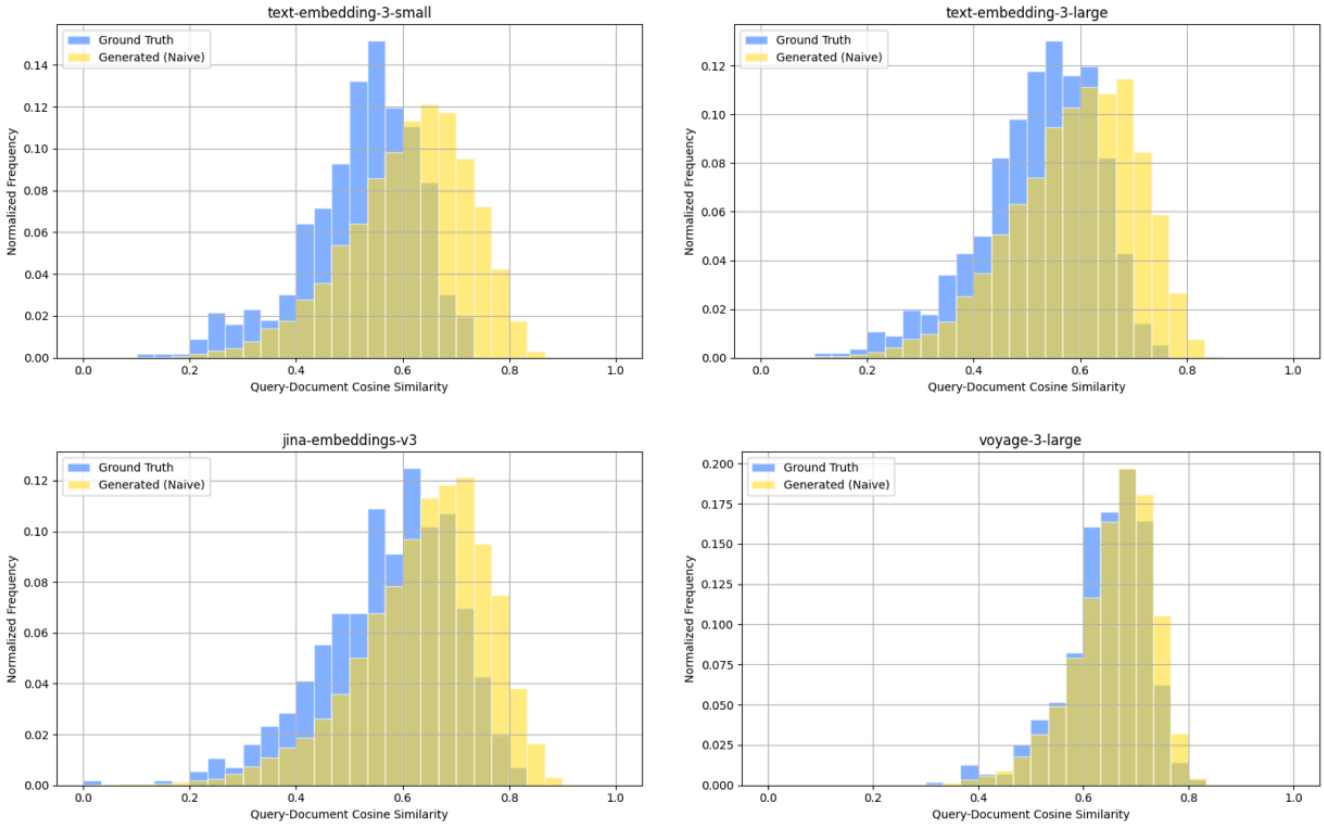
We also highlight a discrepancy with MTEB scores. Our results show that jina-embeddings-v3 exhibits lower retrieval performance than text-embedding-3-large—despite jina-embeddings-v3 consistently outperforming text-embedding-3-large across all MTEB English tasks. This reinforces our central claim on how performance on public benchmarks such as MTEB does not always translate to real-world performance.

This does not necessarily mean that text-embedding-3-large is universally superior to jina-embeddings-v-3; rather, it illustrates how performance rankings can shift depending on the specific use case.

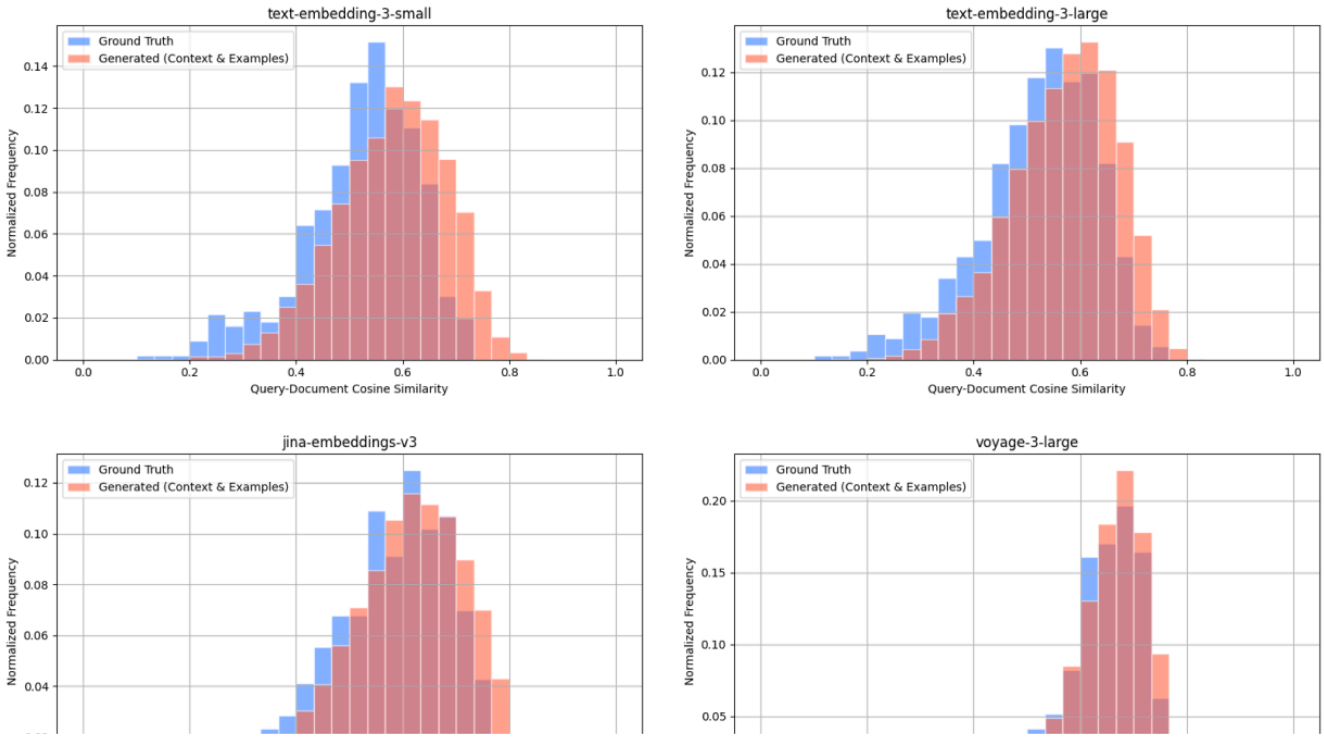
We also compare our generated queries with context and examples against naively generated queries (lacking any context or examples in generation). While these naive queries also maintained similar rankings among embedding models in this case, they yielded higher retrieval metrics than the ground truth. This could misleadingly suggest better performance compared to what would be expected in a real production environment.

In contrast, our query generation method with context and examples not only preserved the ranking of embedding models but also produced metrics that closely matched those from ground truth queries. This highlights an important distinction: representative queries are valuable not just for comparing models, but for approximating real-world performance. By better reflecting the characteristics of actual user traffic, our method offers a more accurate benchmark.

generated with context and examples have a lower KL divergence score of **0.159** compared to naive queries with a score of **0.207**, indicating closer alignment ground truth queries.



WandBot - query-document scores for ground truth queries (blue) and naively generated queries (yellow).



WandBot - query-document scores for ground truth queries (blue) and queries generated with context and examples (red).

These findings underscore the effectiveness of our query generation method in representing real user queries.

Limitations & Future Work

Due to the limited availability of diverse production datasets, our evaluation with real data is constrained to one dataset. Thus, this is a limitation of our work as further evaluation across varied domains is necessary for identifying potential domain-specific nuances.

Additionally, our Weights and Biases dataset includes queries that lack corresponding documentation matches—a scenario common in production environments. Currently, our retrieval metrics do not capture these scenarios, as our evaluation solely considered queries with available relevant documents.

From this, we identify several directions for future work:

- Expansion to other domains: Applying our generative benchmarking method to a wider range of production datasets will help uncover differences in generation patterns and various embedding models.
- Iterating with production traffic: Incorporating production traffic can improve both the document corpus and golden dataset. Not only would this better align evaluation queries with real user queries, it would also surface recurring gaps—queries that consistently fail to retrieve relevant documents. Future work could explore ways to automatically detect such outlier queries and enable proactive responses.
- Methods to improve corpus quality: We consistently observe that any given corpus contains a portion of documents irrelevant to user needs or lack sufficient context to answer relevant queries. Future efforts could explore reliable methods for document cleaning, selective context enrichment, or restructuring content to improve retrieval performance.

Conclusion

generated queries with the ground truth, across public datasets and real production data from Weights and Biases. Alongside these demonstrations, we provide t codebase for generating a benchmark on any set of documents.

Footnotes

[1] When applicable, we configured the input_type parameter for query-document retrieval (i.e. input_type=query for embedding a query with voyage-3-large).

[2] Only deduplication was used to clean datasets since our goal was to assess representation rather than optimize retrieval performance.

[3] all-MiniLM-L6-v2 was excluded from Wikipedia Multilingual for all non-English datasets as it was only trained on English.

[4] all-MiniLM-L6-v2 was excluded from LegalBench Consumer Contracts QA since 83.77% of documents exceed its 256-token context limit.

[5] To ensure that corpus-only query generation produced queries in the intended language, the prompt was slightly adjusted to specify the target language for non-English datasets. The rest of the prompt remained unchanged to maintain consistency.

References

[1] Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2022). MTEB: Massive Text Embedding Benchmark. arXiv preprint arXiv:2210.07316. PDF

[2] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (Round 2). PDF

[3] Biewald, L. (2020). Experiment Tracking with Weights and Biases [Software]. Link

[4] Bonifacio, L., Abonizio, H., Fadaee, M., and Nogueira, R. (2022). InPars: Data Augmentation for Information Retrieval using Large Language Models. arXiv preprint arXiv:2202.05144. PDF

In Proceedings of the 11th International Conference on Learning Representations (ICLR). PDF

[6] Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217. PDF

[7] Chen, J., Wang, N., Li, C., Wang, B., Xiao, S., Xiao, H., Liao, H., Lian, D., and Liu, Z. (2024). AIR-Bench: Automated Heterogeneous Information Retrieval Benchmark. arXiv preprint arXiv:2412.13102. PDF

[8] Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A. G., and Arawjo, I. (2024). Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. arXiv preprint arXiv:2404.12272. PDF

[9] ellamind. (2023). wikipedia-2023-11-retrieval-multilingual [Dataset]. Hugging Face. Link

[10] mteb. (2023). legalbench_consumer_contracts_qa [Dataset]. Hugging Face. Link

[11] Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7534–7550. PDF

[12] mteb. (n.d.). medical_qa [Dataset]. Hugging Face. Link

[13] McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426. PDF

[14] Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In Pei, J., Tseng, V. S., Cao, L., Motoda, H., and Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining (PAKDD 2013), Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg. Link

Appendix

Prompts

```
Based on the following piece of information:
<document>
{document}
<document>

Please generate a query relevant to the information provided above.

Simply output the query without any additional words in this format:
<format>
[query]
<format>
```

Distinct Query Generation Prompt

```
Based on the following piece of information:
<document>
{document}
<document>

This would be an example query that would be good for this kind of context:
<query>
{query}
<query>

Please generate one additional query that is distinct from the example, but is still
relevant to the corpus. This point is very important, ensure that the generated query
does not repeat the given example query.

Simply output the query without any additional words in this format:
<format>
[query]
<format>
```

LLM Judge Criteria

```
relevance = """
The document is relevant and something that users would search for considering the
following context:

We are building a question-answering bot designed specifically for Weights & Biases,
an AI developer for training, fine-tuning, and managing models.

Any information that would be useful to a user working in machine learning is
considered as relevant.
```

```
The document is complete, meaning that it contains useful information to answer
queries and does not only serve as an introduction to the main content that user
be looking for.
"""
```


```
intent = """
The document would be relevant in the use case of a user working in machine learning,
who may be seeking help or learn more about Weights & Biases or machine learning in
general.
"""
```

LLM Judge Prompt

```
Evaluate the following document against the criterion below.

Criterion: {criterion}

Document: {document}

Output a single word: "yes" if the document meets the criterion, or "no" if it does
not. Do not include any extra text or formatting, simply "yes" or "no".
```

Weights and Biases Query Generation Prompt

```
Consider the context:
{context}

Based on the following piece of text:
<text>
{document}
<text>

Please generate a realistic query that a user may ask relevant to the information
provided above.

Here are some example queries that users have asked which you should consider when
generating your query:
<example-queries>
{example_queries}
<example-queries>

Do not repeat the example queries, they are only provided to give you an idea of the
type of queries that users ask.
```

```
Simply output the query without any additional words in this format:
<format>
[query]
<format>
```

Results

Wikipedia Multilingual (en)

```
Score: 1.0000
Original Query: What modifications were made to the SR-25 when it was adopted by SOCOM
as the Mk 11 MOD 0?
Generated Query: What modifications were made to the SR-25 when it was adopted by
SOCOM as the Mk 11 MOD 0?

Score: 1.0000
Original Query: How does a PWM anemometer measure wind velocity?
Generated Query: How does a PWM anemometer measure wind velocity?

Score: 1.0000
Original Query: Where was the statue of Ferdinando I hidden during World War II?
Generated Query: Where was the statue of Ferdinando I hidden during World War II?

Score: 1.0000
Original Query: Why did Nagma join the Congress Party?
Generated Query: Why did Nagma join the Congress Party?

Score: 1.0000
Original Query: What types of trees were traditionally used for coppicing in southern
Britain?
Generated Query: What types of trees were traditionally used for coppicing in southern
Britain?
```

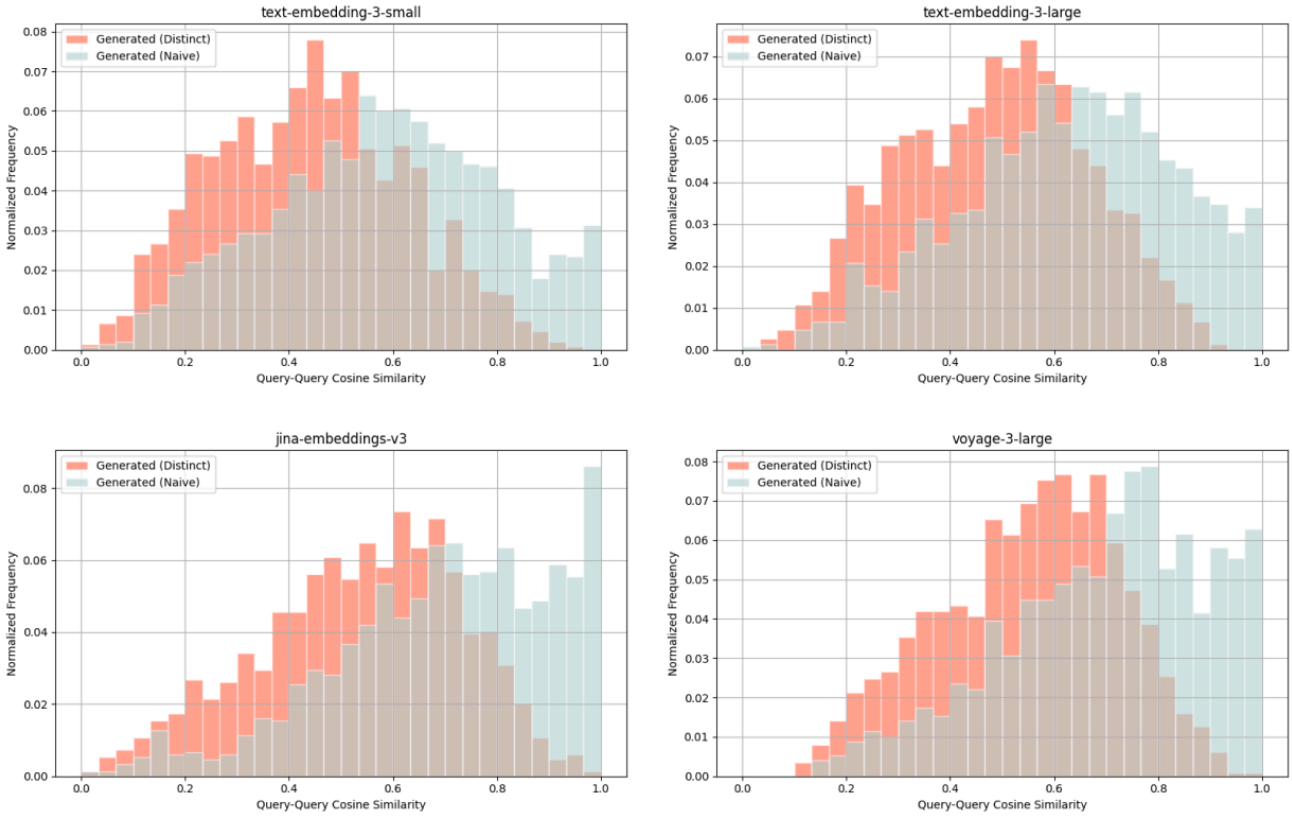
Wikipedia Multilingual (en) - reproduced queries, cosine similarity scores calculated using text-embedding-3-small.

Wikipedia Multilingual (hi)

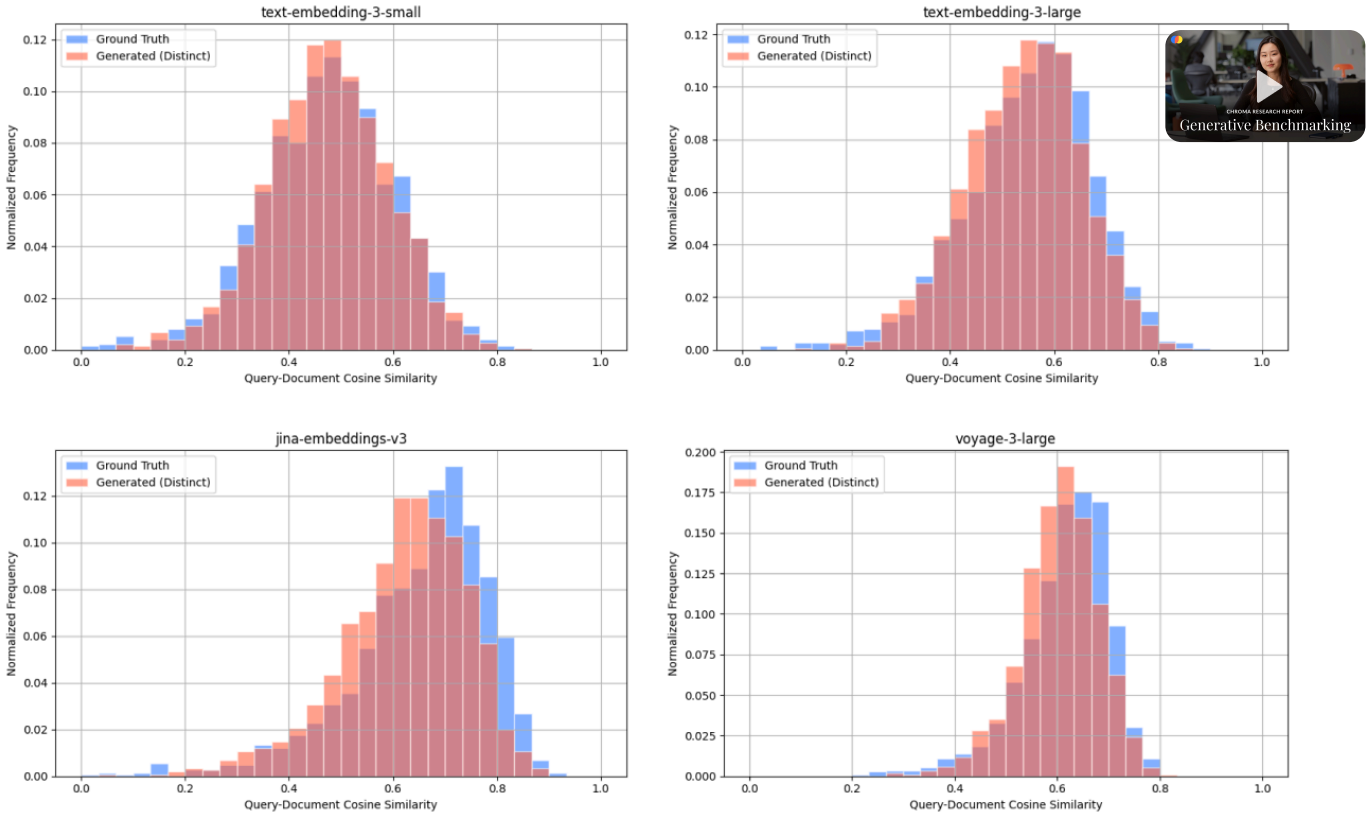
```
Score: 1.0000
Original Query: प्रकिण्व की सक्रियता को कौन-कौन से कारक प्रभावित करते हैं?
Generated Query: प्रकिण्व की सक्रियता को कौन-कौन से कारक प्रभावित करते हैं?

Score: 1.0000
Original Query: वरदविनायक मंदिर कहाँ स्थित है?
```

```
Original Query: गीताप्रेस किन-किन भाषाओं में साहित्य प्रकाशित करता है?
Generated Query: गीताप्रेस किन-किन भाषाओं में साहित्य प्रकाशित करता है?

Score: 1.0000
Original Query: मूसा ने अपने युवक सेवक से क्या कहा?
Generated Query: मूसा ने अपने युवक सेवक से क्या कहा?

Score: 1.0000
Original Query: छायावादयुग का आरंभ किस युग से हुआ था?
Generated Query: छायावादयुग का आरंभ किस युग से हुआ था?
```

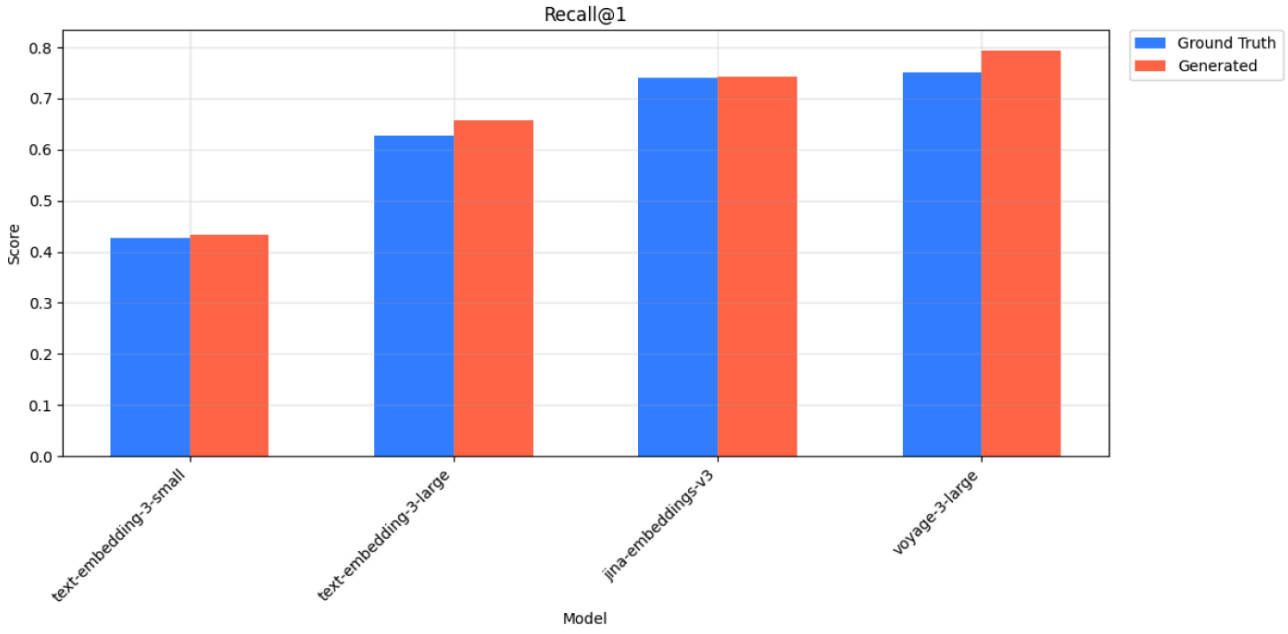
Wikipedia Multilingual (hi) - reproduced queries, cosine similarity scores calculated using text-embedding-3-small.



Wikipedia Multilingual (hi) - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red).



Wikipedia Multilingual (hi) - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).



Wikipedia Multilingual (hi) - Recall@1 Scores for Ground Truth and Generated Queries across 4 models.

Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10

Query Type	Model	Recall@1	Recall@3	Recall@5	Rec
	small				
Generated	text-embedding-3-small	0.434	0.544	0.582	0.609
Ground Truth	text-embedding-3-large	0.628	0.789	0.830	0.866
Generated	text-embedding-3-large	0.658	0.782	0.819	0.849
Ground Truth	jina-embeddings-v3	0.741	0.863	0.890	0.915
Generated	jina-embeddings-v3	0.743	0.861	0.884	0.899
Ground Truth	voyage-3-large	0.751	0.859	0.885	0.901
Generated	voyage-3-large	0.794	0.879	0.889	0.894

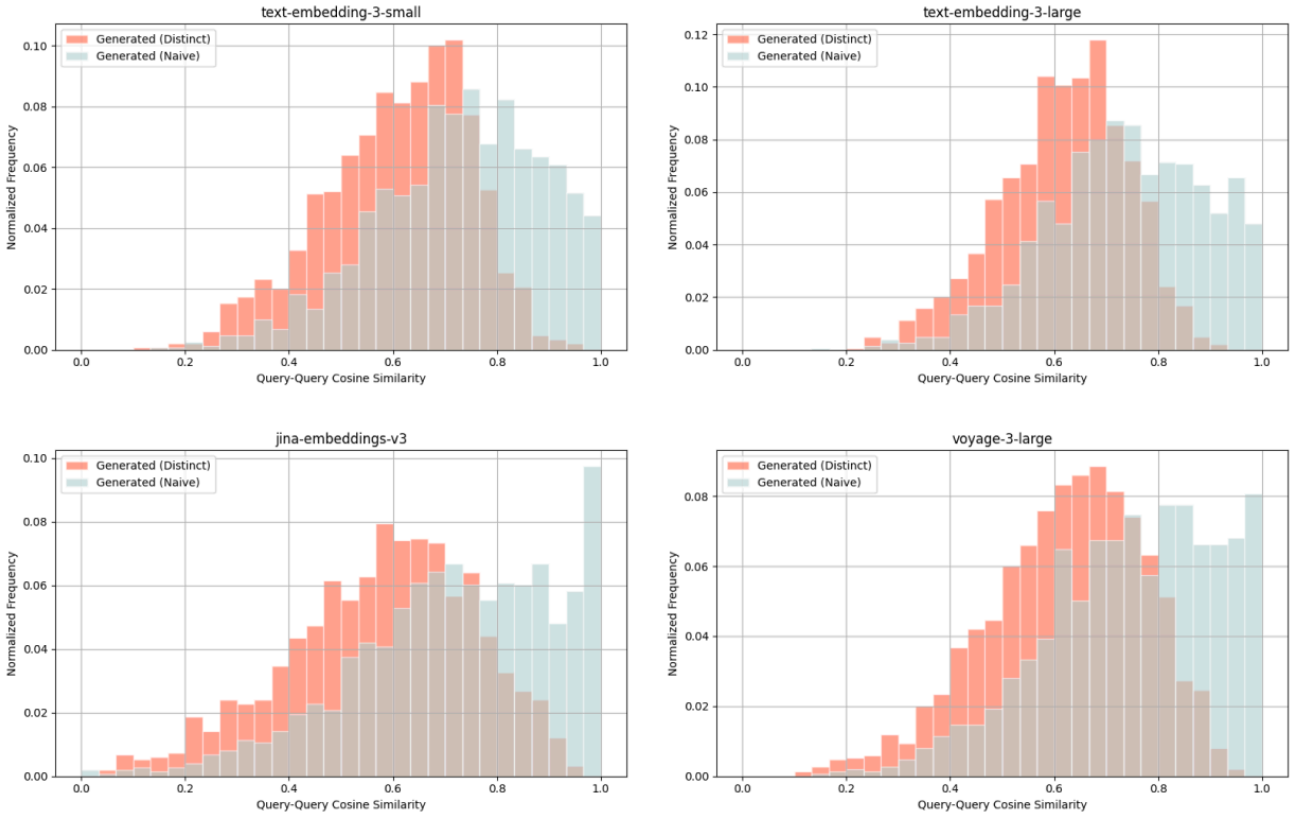
Wikipedia Multilingual (hi) - Recall@k Scores for Ground Truth and Generated Queries across 4 models.

Wikipedia Multilingual (de)

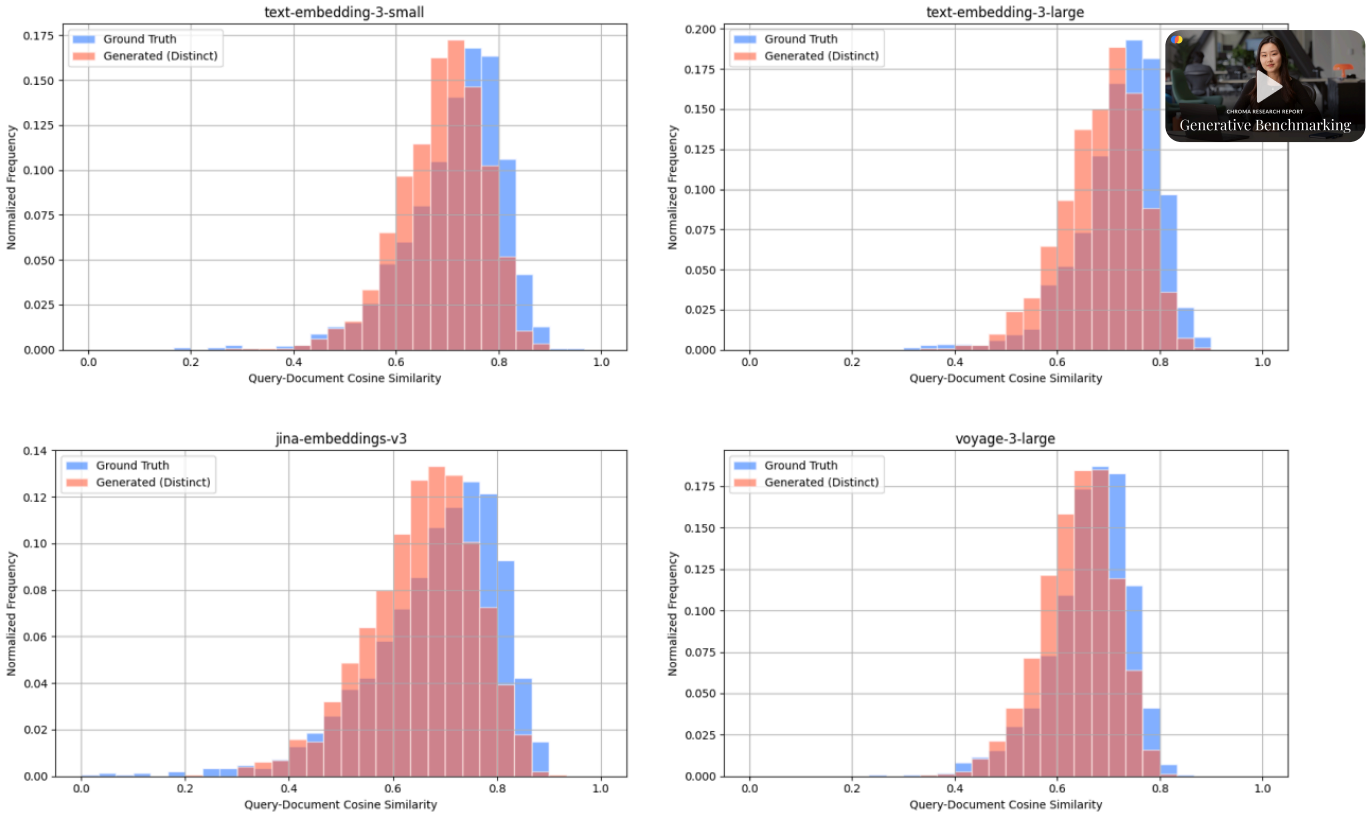
```
Score: 1.0000
Original Query: Warum sind Eisengallustinten nur für Eintauchfedern geeignet?
Generated Query: Warum sind Eisengallustinten nur für Eintauchfedern geeignet?

Score: 1.0000
Original Query: Was ist der Unterschied zwischen einem Gutachten und einer
gutachtlichen Stellungnahme?
```

```
Score: 1.0000
Original Query: Was ist ein Kettenkehrreim und wie unterscheidet er sich von ein
rückwärts laufenden Kettenkehrreim?
Generated Query: Was ist ein Kettenkehrreim und wie unterscheidet er sich von einem
rückwärts laufenden Kettenkehrreim?

Score: 1.0000
Original Query: Wann wurde die Herz-Jesu-Kirche in Bruckmühl eingeweiht?
Generated Query: Wann wurde die Herz-Jesu-Kirche in Bruckmühl eingeweiht?

Score: 1.0000
Original Query: Welche verschiedenen Arten von Bordellen gibt es in Deutschland?
Generated Query: Welche verschiedenen Arten von Bordellen gibt es in Deutschland?
```

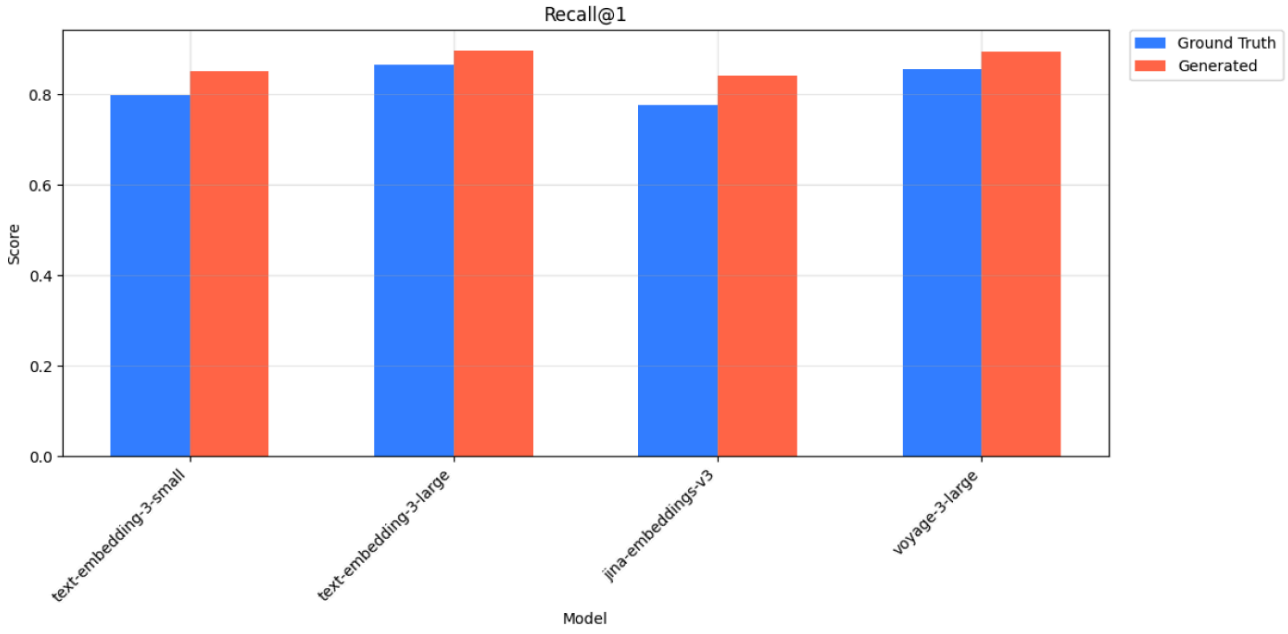
Wikipedia Multilingual (de) - reproduced queries, cosine similarity scores calculated using text-embedding-3-small.



Wikipedia Multilingual (de) - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red).



Wikipedia Multilingual (de) - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).



Wikipedia Multilingual (de) - Recall@1 Scores for Ground Truth and Generated Queries across 4 models.

Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10

Query Type	Model	Recall@1	Recall@3	Recall@5	Rec
	small				
Generated	text-embedding-3-small	0.851	0.947	0.961	0.967
Ground Truth	text-embedding-3-large	0.865	0.952	0.974	0.988
Generated	text-embedding-3-large	0.897	0.969	0.983	0.989
Ground Truth	jina-embeddings-v3	0.775	0.900	0.934	0.951
Generated	jina-embeddings-v3	0.840	0.931	0.947	0.955
Ground Truth	voyage-3-large	0.857	0.939	0.961	0.973
Generated	voyage-3-large	0.894	0.946	0.953	0.956

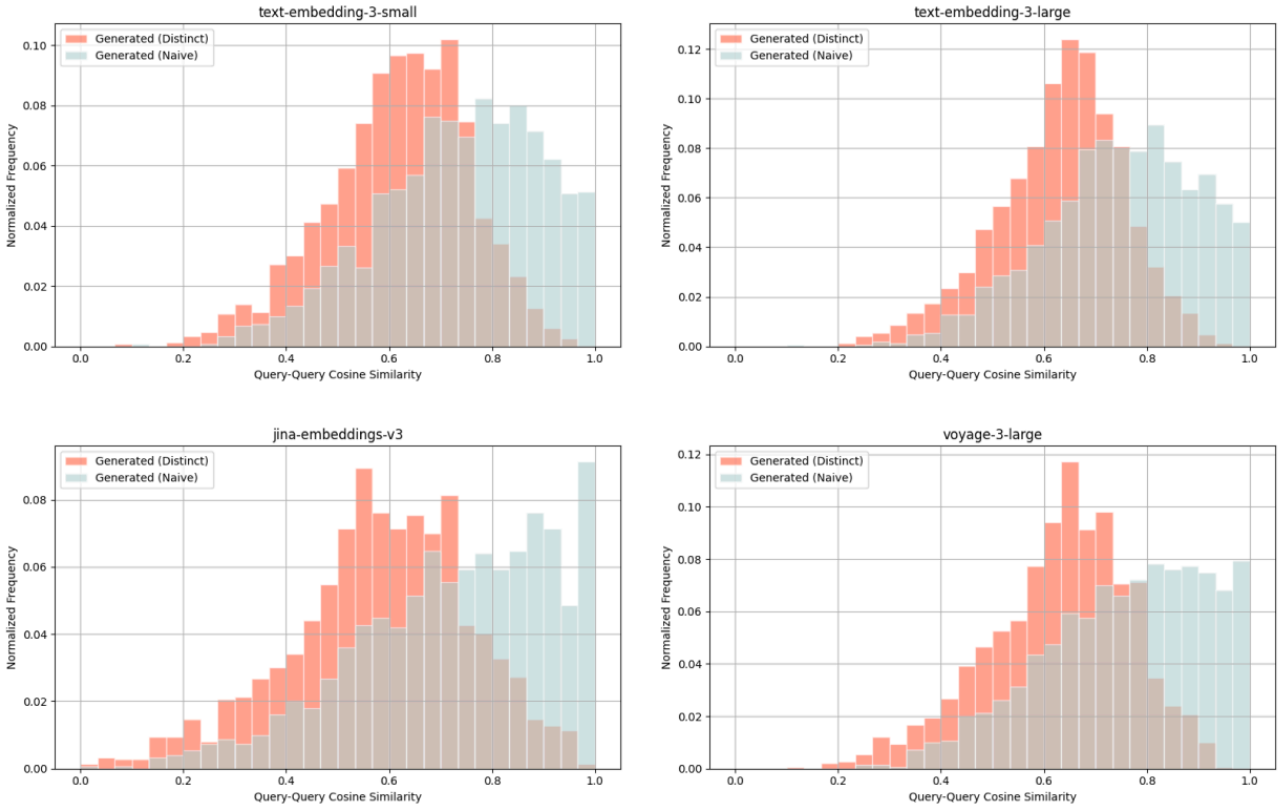
Wikipedia Multilingual (de) - Recall@k Scores for Ground Truth and Generated Queries across 4 models.

Wikipedia Multilingual (pt)

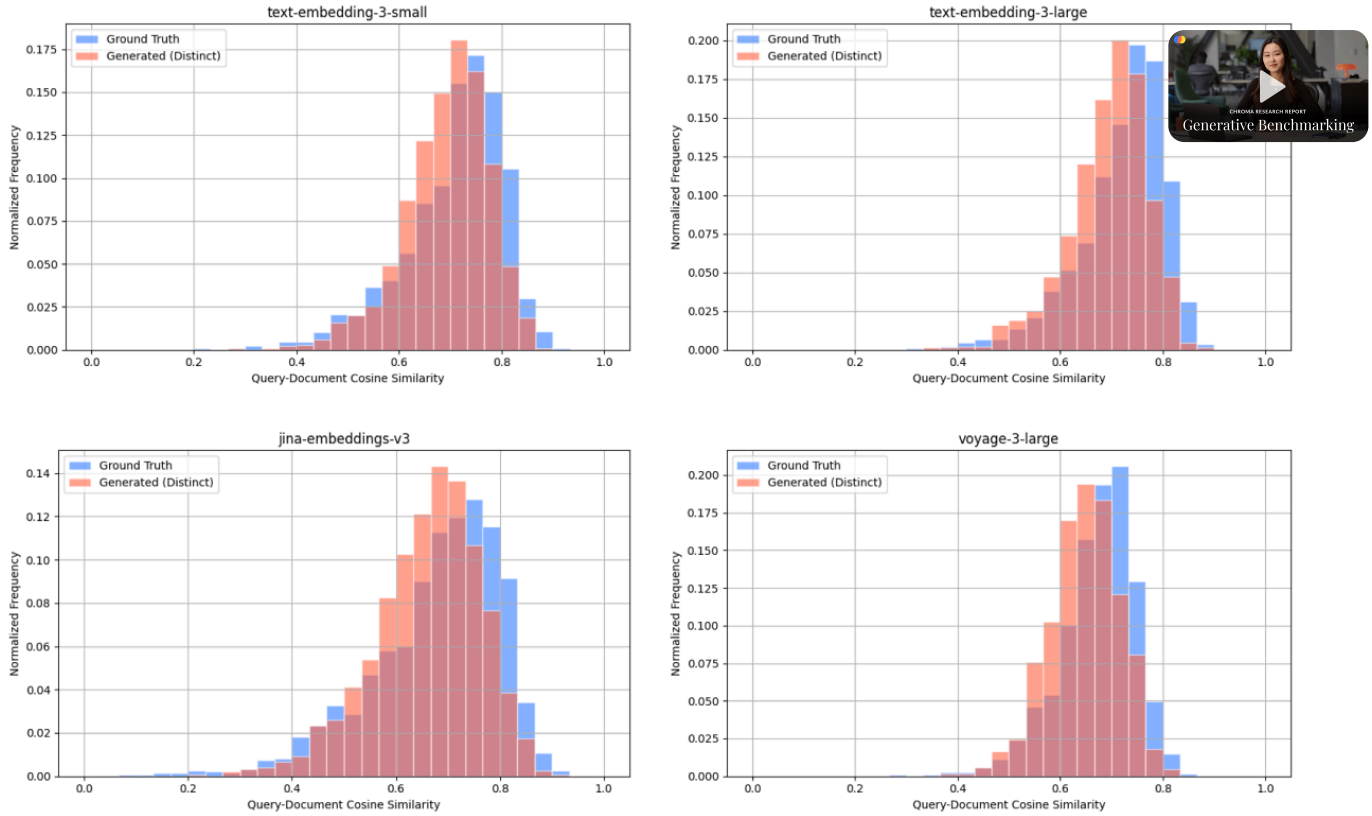
```
Score: 1.0000
Original Query: Quem é Enkai na mitologia dos massai?
Generated Query: Quem é Enkai na mitologia dos massai?

Score: 1.0000
Original Query: Qual é o slogan da Hellmann's no Brasil?
Generated Query: Qual é o slogan da Hellmann's no Brasil?
```

```
Generated Query: Como Políbio morreu?

Score: 1.0000
Original Query: Qual foi a importância estratégica de Coruche durante o período de
domínio islâmico?
Generated Query: Qual foi a importância estratégica de Coruche durante o período de
domínio islâmico?

Score: 1.0000
Original Query: Quais foram os principais produtos agropecuários exportados pelo Laos
em 2019?
Generated Query: Quais foram os principais produtos agropecuários exportados pelo Laos
em 2019?
```

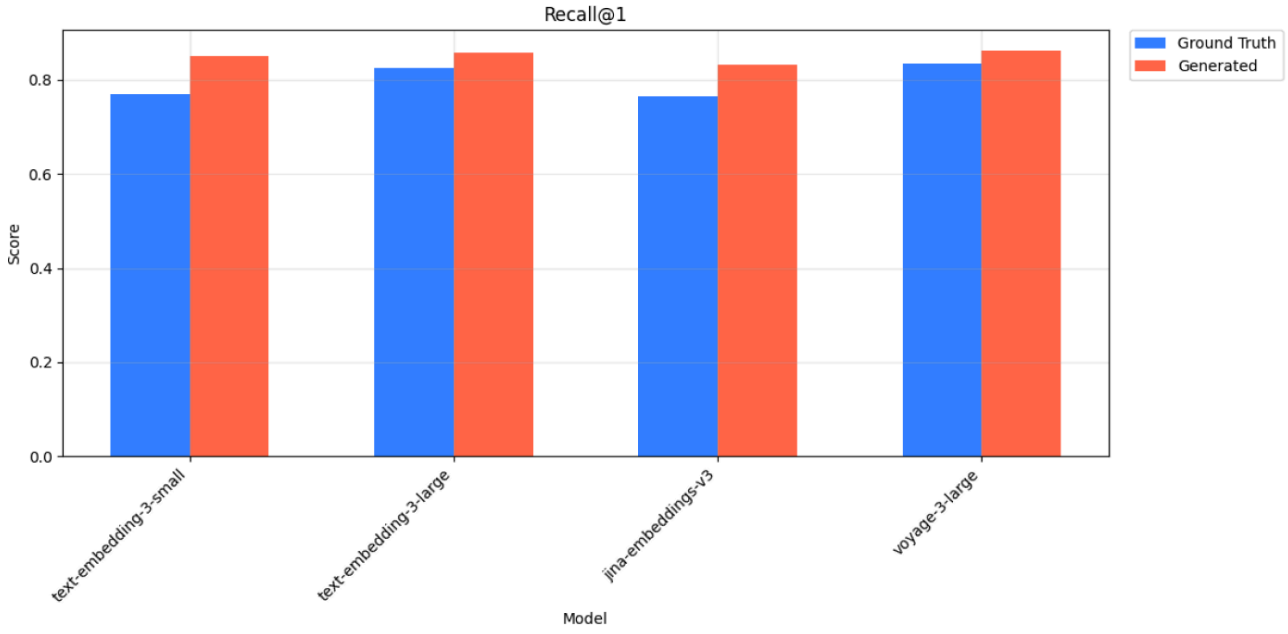
Wikipedia Multilingual (pt) - reproduced queries, cosine similarity scores calculated using text-embedding-3-small.



Wikipedia Multilingual (pt) - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red).



Wikipedia Multilingual (pt) - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).



Wikipedia Multilingual (pt) - Recall@1 Scores for Ground Truth and Generated Queries across 4 models.

Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10

Query Type	Model	Recall@1	Recall@3	Recall@5	Rec
	small				
Generated	text-embedding-3-small	0.851	0.938	0.955	0.969
Ground Truth	text-embedding-3-large	0.825	0.931	0.961	0.989
Generated	text-embedding-3-large	0.857	0.951	0.966	0.978
Ground Truth	jina-embeddings-v3	0.765	0.893	0.929	0.954
Generated	jina-embeddings-v3	0.833	0.928	0.950	0.960
Ground Truth	voyage-3-large	0.835	0.939	0.961	0.979
Generated	voyage-3-large	0.863	0.945	0.956	0.959

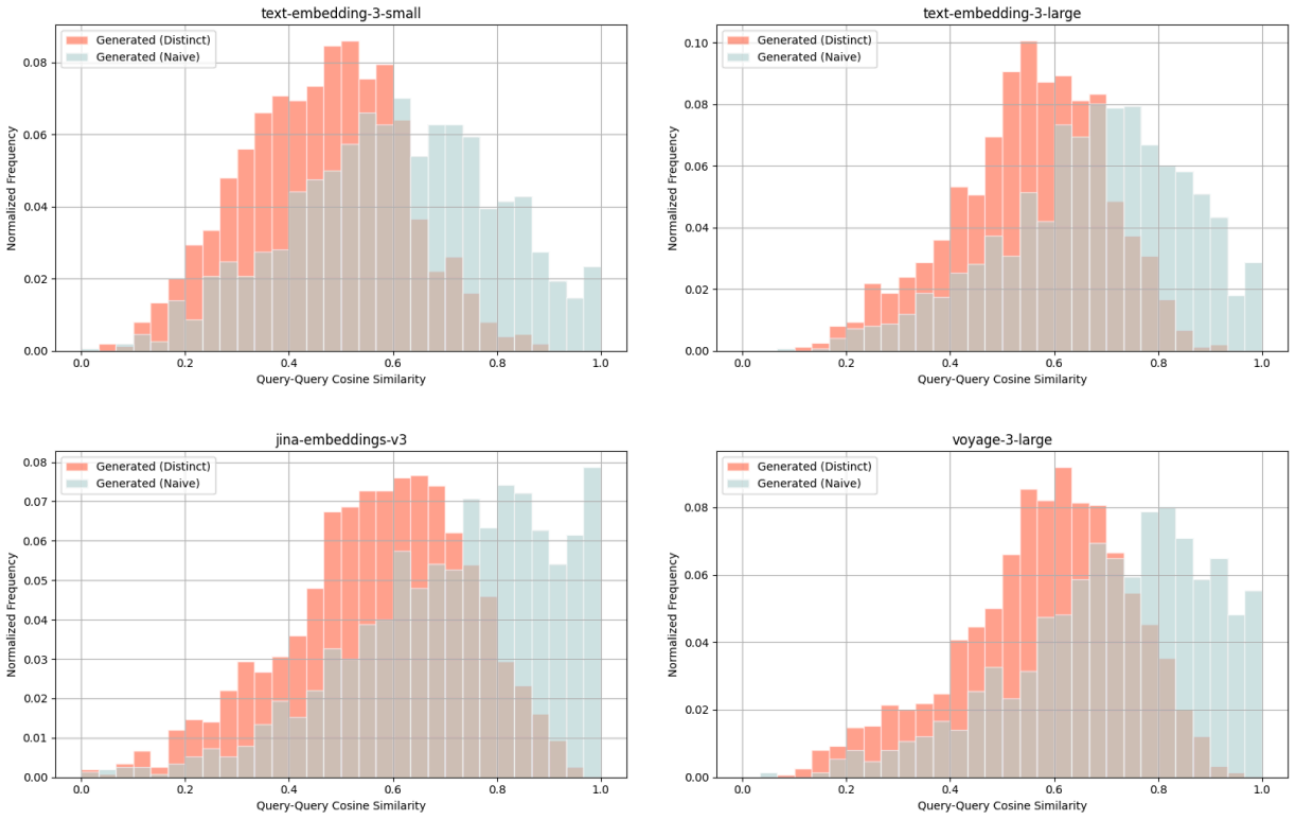
Wikipedia Multilingual (pt) - Recall@k Scores for Ground Truth and Generated Queries across 4 models.

Wikipedia Multilingual (fa)

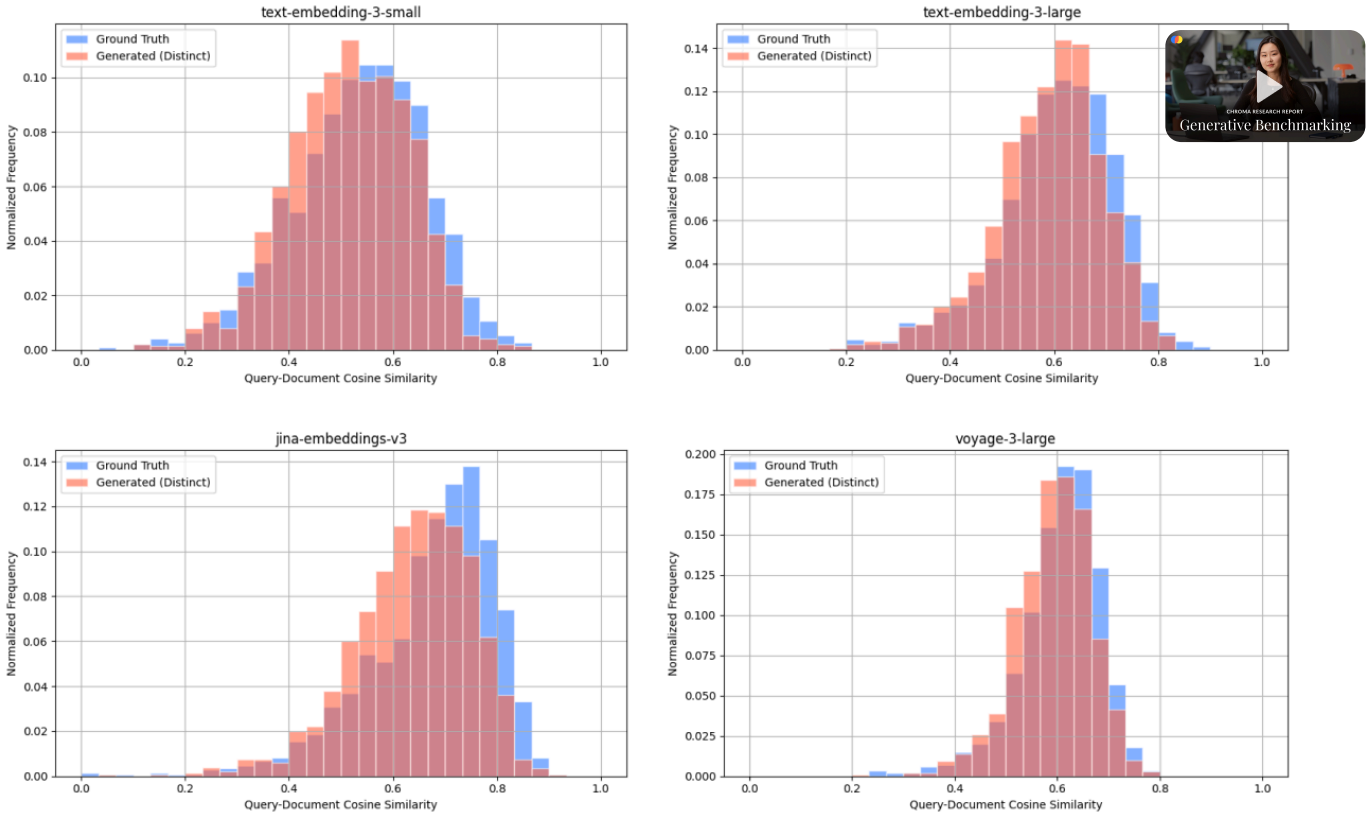
```
Score: 1.0000
Original Query: محله لتیبار در کدام قسمت شهر سمنان واقع شده است؟
Generated Query: محله لتیبار در کدام قسمت شهر سمنان واقع شده است؟

Score: 1.0000
Original Query: عناب چه فوایدی برای سلامتی دارد؟
Generated Query: عناب چه فوایدی برای سلامتی دارد؟
```

```
Generated Query: دریاچه بزرگ آلماتی در چه ارتفاعی از سطح دریا قرار دارد؟

Score: 1.0000
Original Query: چرا ترانزیستورهای ماسفت را تک‌قطبی می‌نامند؟
Generated Query: چرا ترانزیستورهای ماسفت را تک‌قطبی می‌نامند؟

Score: 1.0000
Original Query: شعار دانشگاه هاروارد چیست؟
Generated Query: شعار دانشگاه هاروارد چیست؟
```

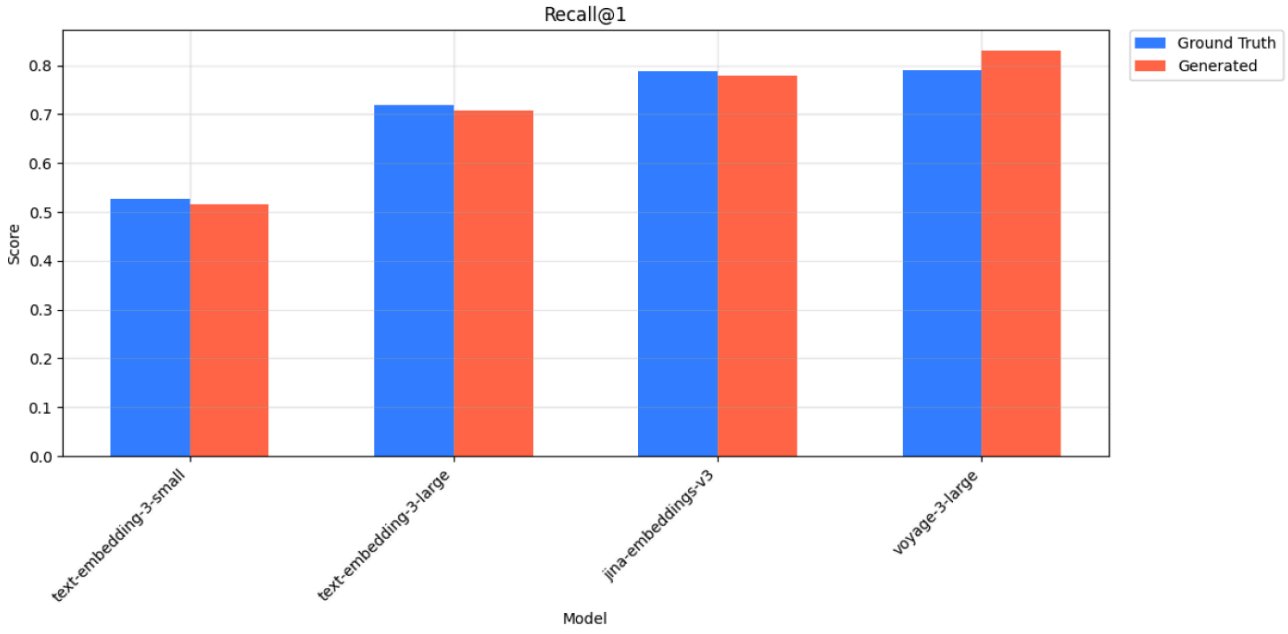
Wikipedia Multilingual (fa) - reproduced queries, cosine similarity scores calculated using text-embedding-3-small.



Wikipedia Multilingual (fa) - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red).



Wikipedia Multilingual (fa) - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).



Wikipedia Multilingual (fa) - Recall@1 Scores for Ground Truth and Generated Queries across 4 models.

Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10

Query Type	Model	Recall@1	Recall@3	Recall@5	Rec
	small				
Generated	text-embedding-3-small	0.517	0.628	0.668	0.701
Ground Truth	text-embedding-3-large	0.719	0.859	0.891	0.917
Generated	text-embedding-3-large	0.708	0.843	0.877	0.901
Ground Truth	jina-embeddings-v3	0.789	0.905	0.931	0.953
Generated	jina-embeddings-v3	0.779	0.892	0.913	0.925
Ground Truth	voyage-3-large	0.791	0.891	0.911	0.929
Generated	voyage-3-large	0.831	0.904	0.911	0.916

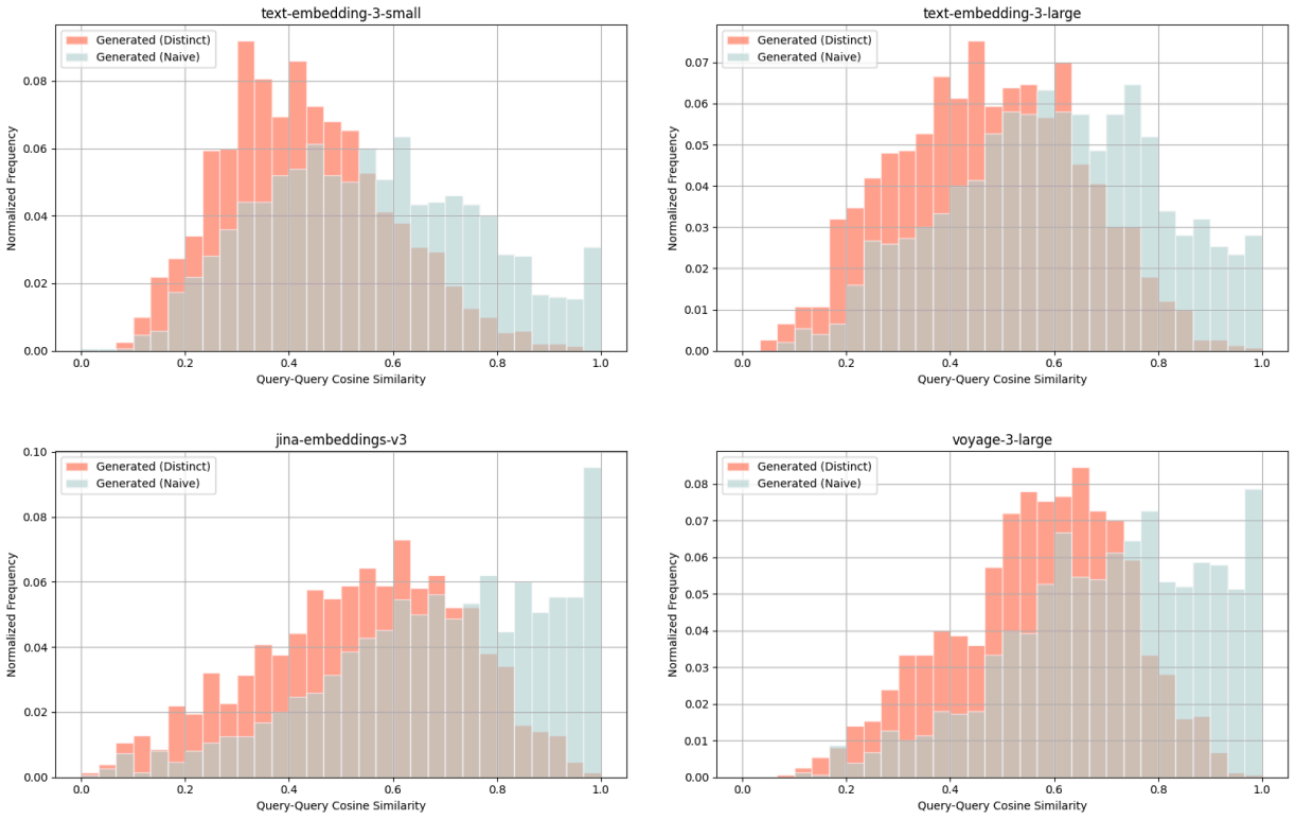
Wikipedia Multilingual (fa) - Recall@k Scores for Ground Truth and Generated Queries across 4 models.

Wikipedia Multilingual (bn)

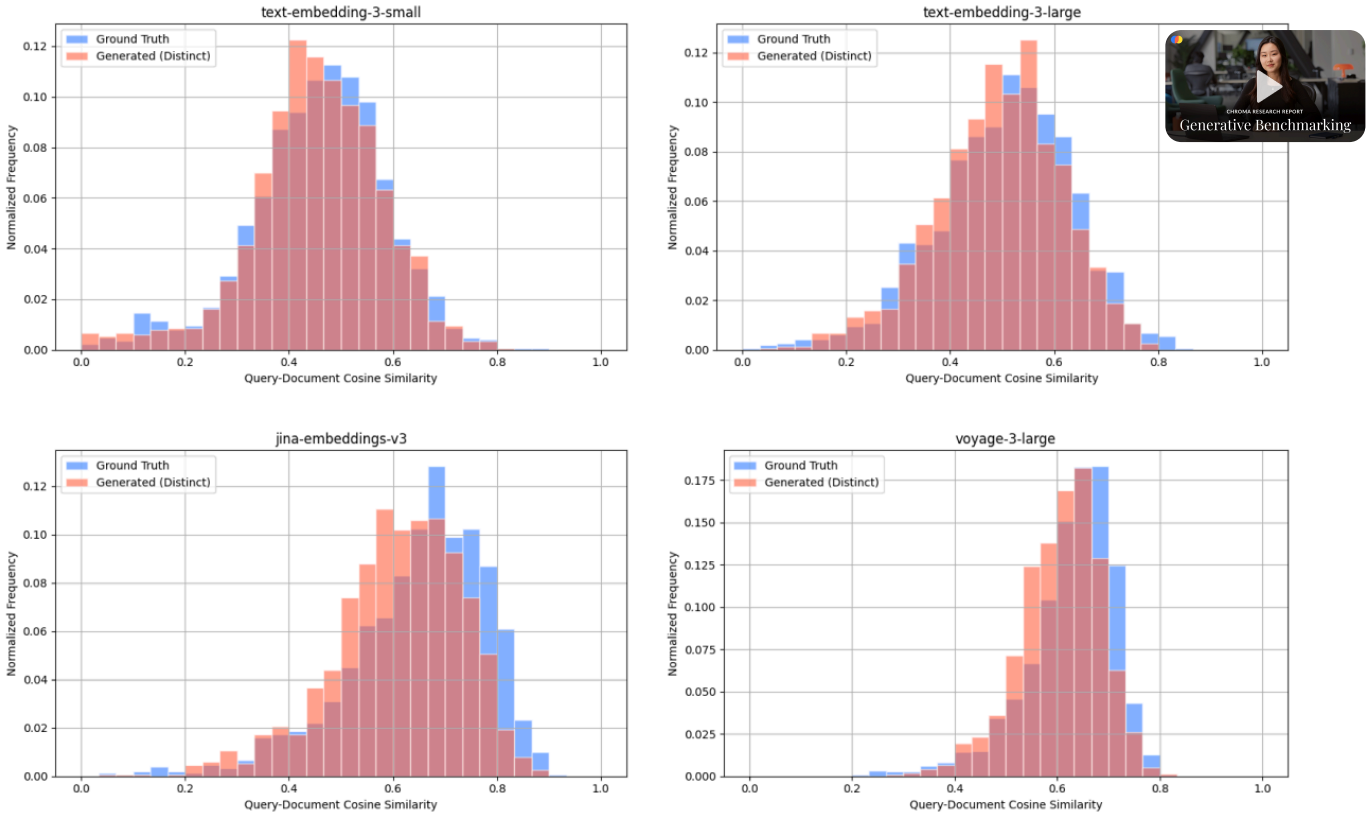
```
Score: 1.0000
Original Query: উইলো গাছের শিকড় আবাসিক এলাকায় কী ধরনের সমস্যার সৃষ্টি করে?
Generated Query: উইলো গাছের শিকড় আবাসিক এলাকায় কী ধরনের সমস্যার সৃষ্টি করে?

Score: 1.0000
Original Query: গরুর মাংসের স্টেক কীভাবে রান্না করা হয়?
Generated Query: গরুর মাংসের স্টেক কীভাবে রান্না করা হয়?
```

```
Generated Query: জাইলিনের বিভিন্ন রকমভেদের গলনাঙ্ক কত?

Score: 1.0000
Original Query: বেরিয়াম ফেল্ডস্পার কীভাবে গঠিত হয়?
Generated Query: বেরিয়াম ফেল্ডস্পার কীভাবে গঠিত হয়?

Score: 1.0000
Original Query: অশ্বত্থ গাছ কোন কোন দেশে পাওয়া যায়?
Generated Query: অশ্বত্থ গাছ কোন কোন দেশে পাওয়া যায়?
```

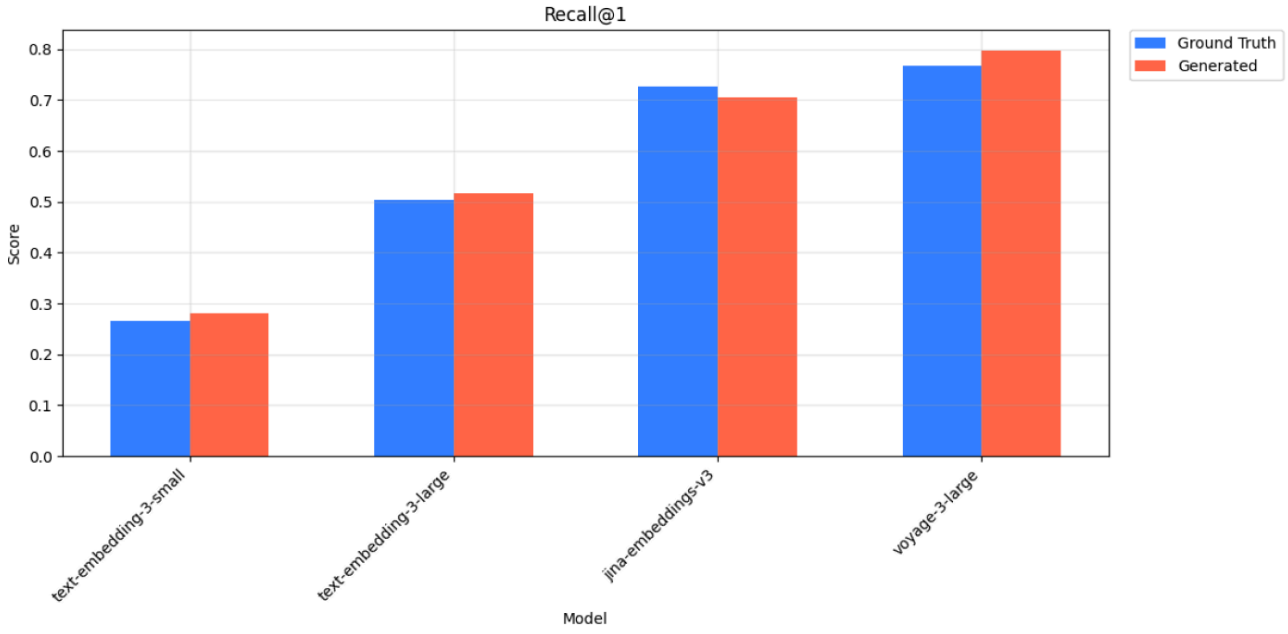
Wikipedia Multilingual (bn) - reproduced queries, cosine similarity scores calculated using text-embedding-3-small.



Wikipedia Multilingual (bn) - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red).



Wikipedia Multilingual (bn) - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).



Wikipedia Multilingual (bn) - Recall@1 Scores for Ground Truth and Generated Queries across 4 models.

Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10

Query Type	Model	Recall@1	Recall@3	Recall@5	Rec
	small				
Generated	text-embedding-3-small	0.281	0.344	0.368	0.395
Ground Truth	text-embedding-3-large	0.503	0.647	0.689	0.723
Generated	text-embedding-3-large	0.516	0.635	0.671	0.706
Ground Truth	jina-embeddings-v3	0.727	0.849	0.882	0.905
Generated	jina-embeddings-v3	0.705	0.827	0.849	0.868
Ground Truth	voyage-3-large	0.767	0.884	0.911	0.928
Generated	voyage-3-large	0.797	0.888	0.901	0.909

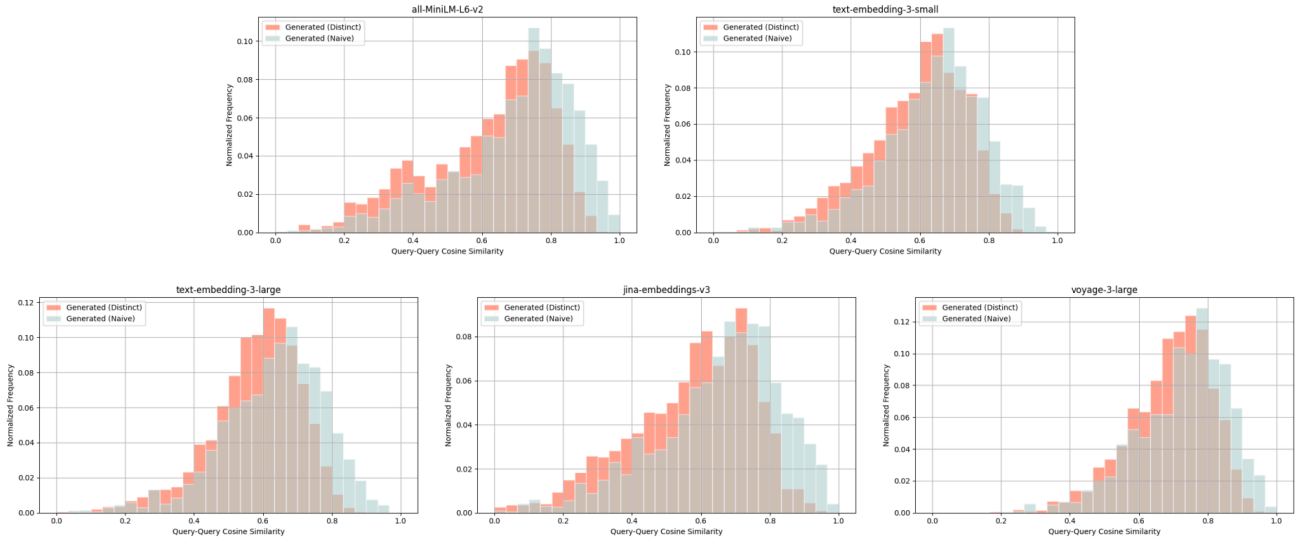
Wikipedia Multilingual (bn) - Recall@k Scores for Ground Truth and Generated Queries across 4 models.

MedicalQA

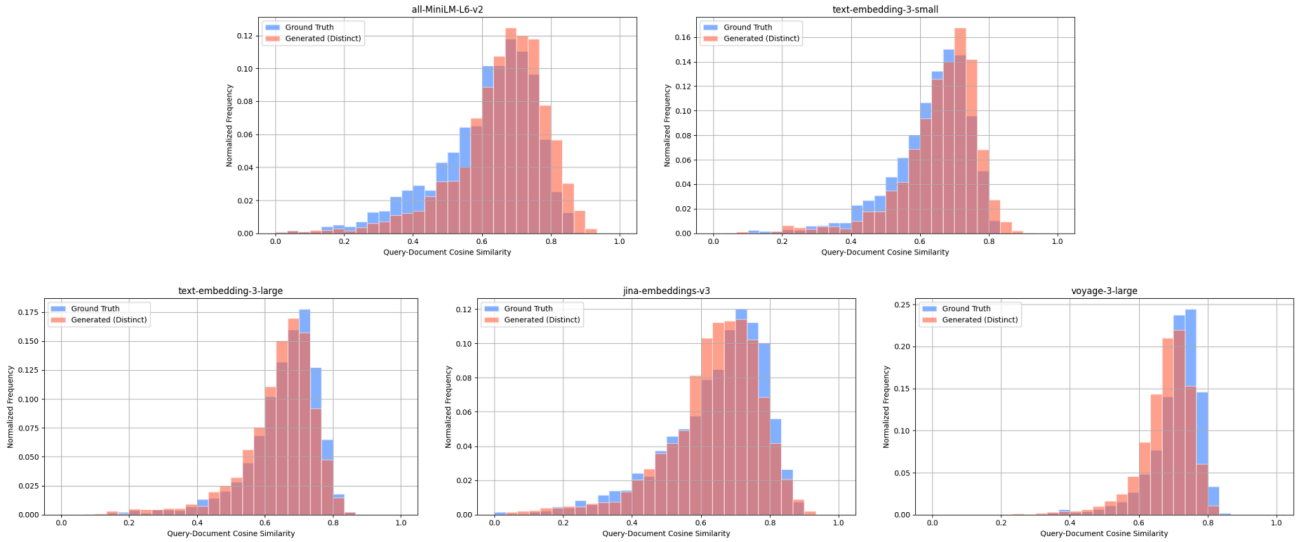
```
Score: 0.9704
Original Query: What are the treatments for Pelizaeus-Merzbacher Disease ?
Generated Query: What treatments are available for Pelizaeus-Merzbacher disease?

Score: 0.9518
Original Query: What are the treatments for Neurological Complications of AIDS ?
Generated Query: What treatments are available for neurological complications of AIDS?
```

```
Generated Query: What treatments are available for patients with Sturge-Weber
syndrome?

Score: 0.9437
Original Query: What are the treatments for Anencephaly ?
Generated Query: What treatments are available for anencephaly?

Score: 0.9380
Original Query: What are the treatments for Meralgia Paresthetica ?
Generated Query: What are the treatment options available for meralgia paresthetica?
```

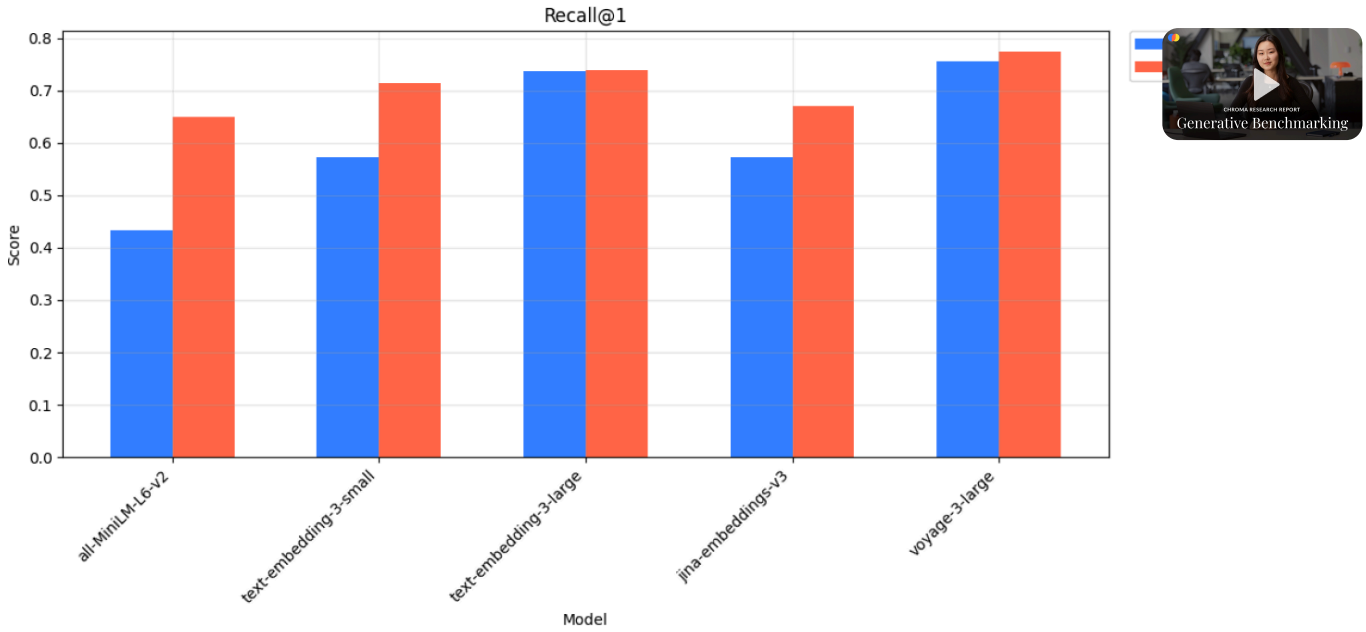
MedicalQA - reproduced queries, cosine similarity scores calculated using text-embedding-3-smal.



MedicalQA - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red).



MedicalQA - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).



MedicalQA - Recall@1 Scores for Ground Truth and Generated Queries across 4 models.

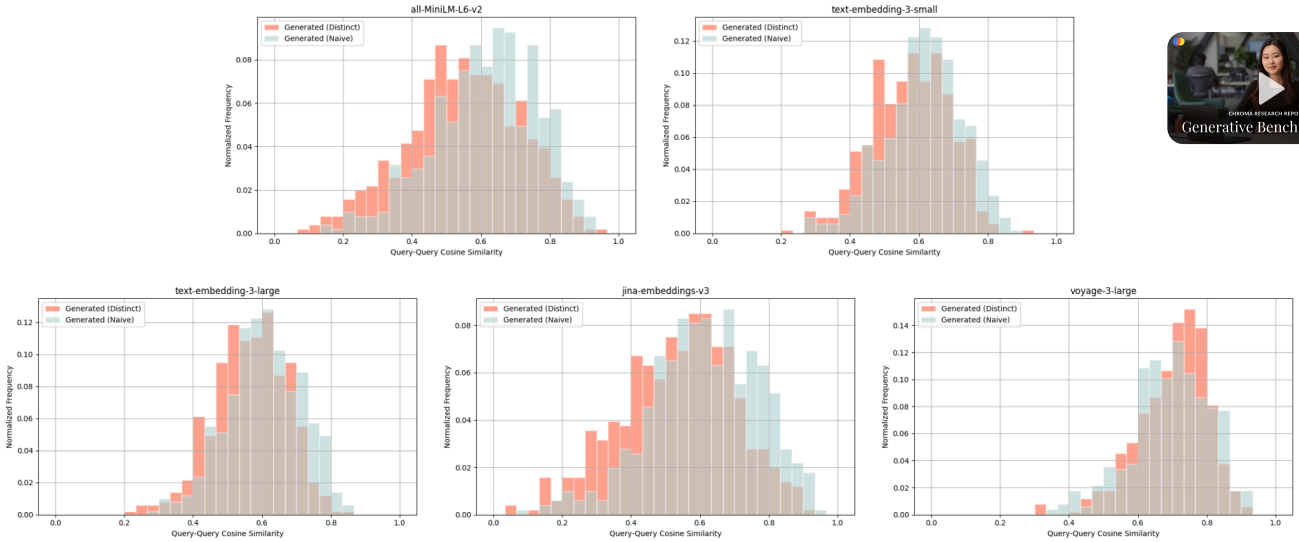
Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10
Ground Truth	all-MiniLM-L6-v2	0.432	0.671	0.748	0.805
Generated	all-MiniLM-L6-v2	0.650	0.822	0.863	0.902
Ground Truth	text-embedding-3-small	0.573	0.795	0.846	0.891
Generated	text-embedding-3-small	0.714	0.870	0.907	0.931
Ground Truth	text-embedding-3-large	0.736	0.875	0.912	0.943
Generated	text-embedding-3-large	0.739	0.900	0.932	0.950
	jina				

Query Type	Model	Recall@1	Recall@3	Recall@5	Rec
Generated	jina-embeddings-v3	0.671	0.835	0.869	0.906
Ground Truth	voyage-3-large	0.757	0.888	0.915	0.944
Generated	voyage-3-large	0.775	0.910	0.929	0.948

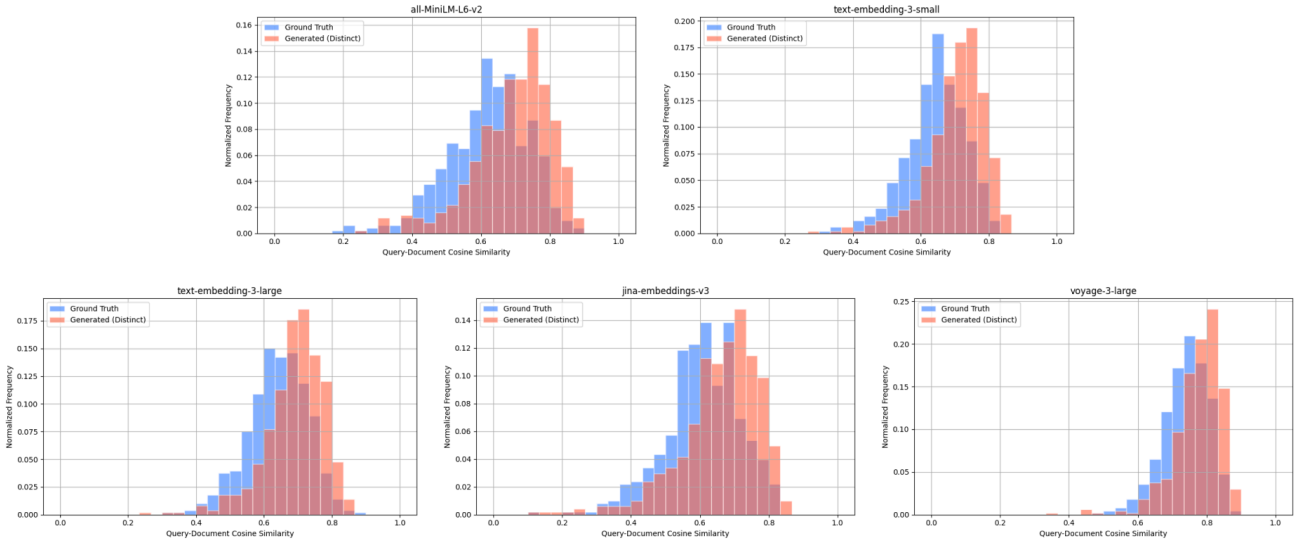
MedicalQA - Recall@k Scores for Ground Truth and Generated Queries across all 5 models.

SciFact

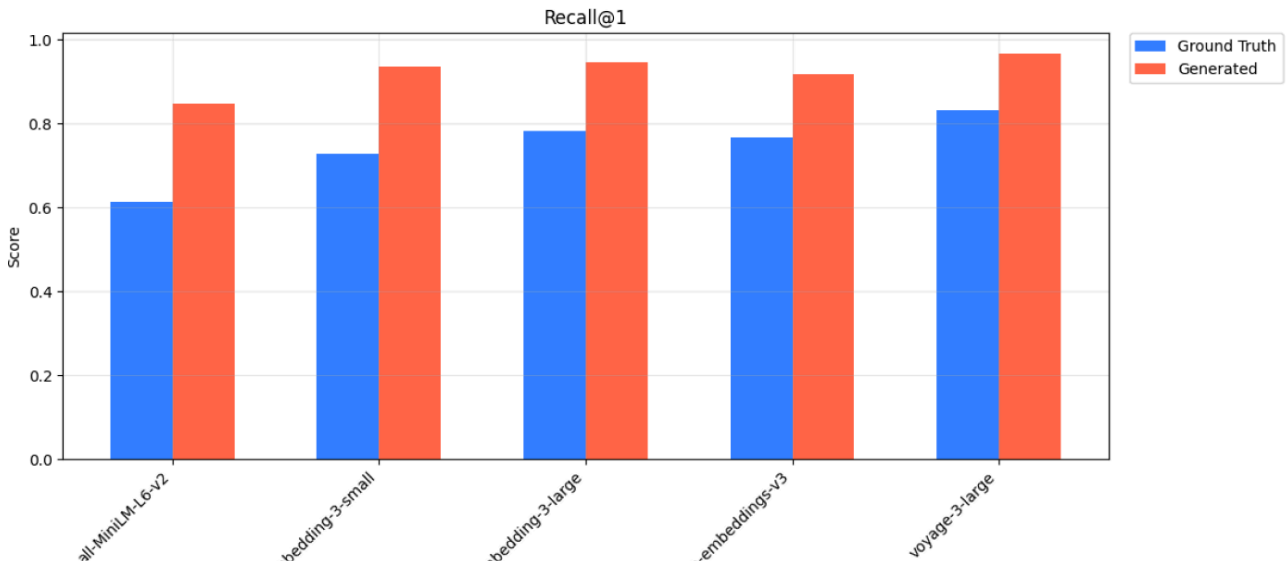
```
Score: 0.8904
Original Query: A single nucleotide variant the gene DGKK is strongly associated with increased risk of hypospadias.
Generated Query: What is the association between DGKK gene variants and hypospadias risk?

Score: 0.8626
Original Query: GATA3 regulates cell cycle progression in bone marrow hematopoietic stem cells.
Generated Query: What is the role of GATA-3 in hematopoietic stem cell regulation and cell cycle maintenance?

Score: 0.8458
Original Query: Active caspase-11 participate in regulating phagosome-lysosome fusion.
Generated Query: What is the role of caspase-11 in regulating bacterial infection and phagosome-lysosome fusion in macrophages?

Score: 0.8427
Original Query: The sliding activity of kinesin-8 protein Kip3 promotes bipolar spindle assembly.
Generated Query: How does the sliding activity of Kip3 contribute to bipolar spindle assembly and genome stability in budding yeast?

Score: 0.8391
Original Query: Clathrin stabilizes the spindle fiber apparatus during mitosis.
Generated Query: How does clathrin function in stabilizing the mitotic spindle and
```



SciFact - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red).



SciFact - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).



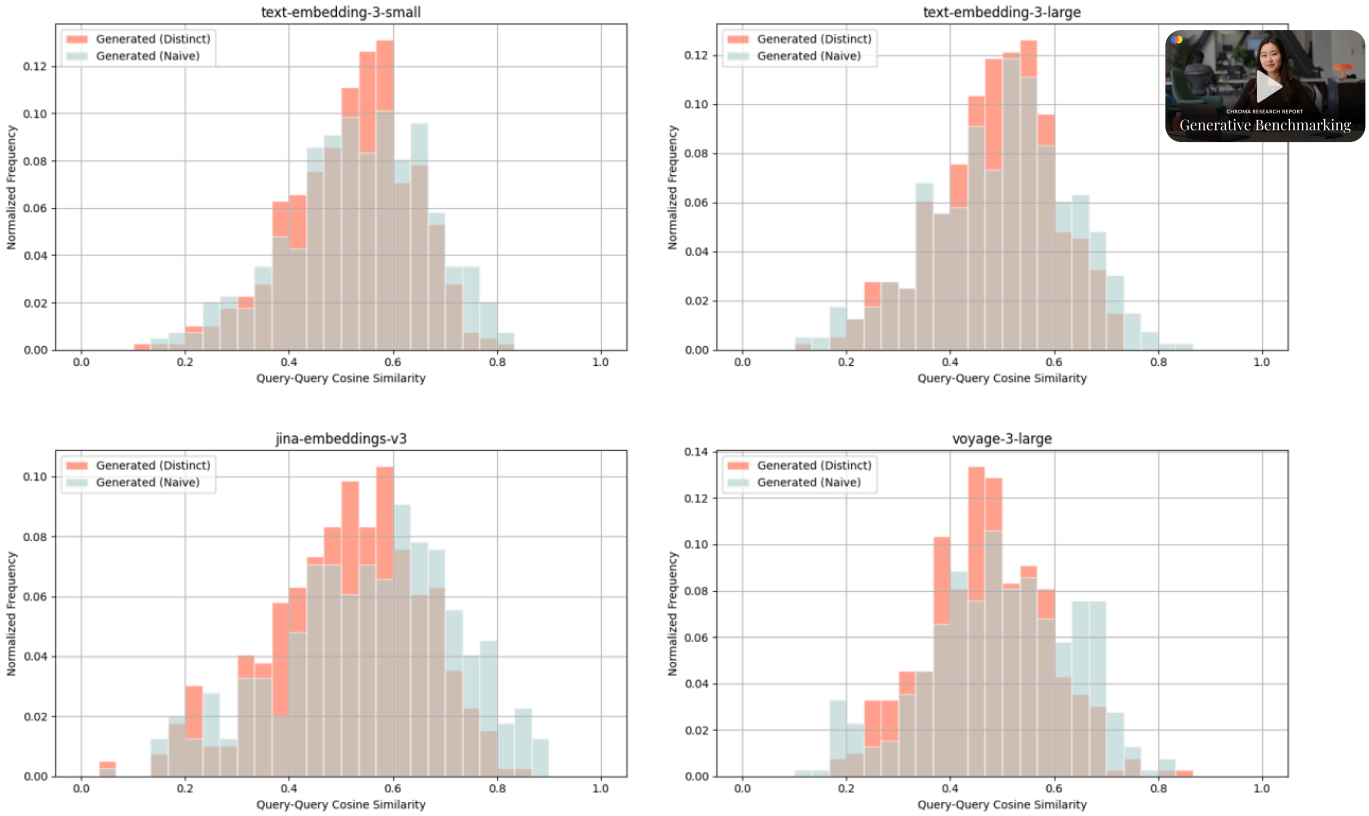
SciFact - Recall@1 Scores for Ground Truth and Generated Queries across 4 models.

Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10
Ground Truth	all-MiniLM-L6-v2	0.613	0.754	0.814	0.862
Generated	all-MiniLM-L6-v2	0.846	0.901	0.921	0.941
Ground Truth	text-embedding-3-small	0.727	0.850	0.898	0.937
Generated	text-embedding-3-small	0.936	0.949	0.968	0.972
Ground Truth	text-embedding-3-large	0.782	0.896	0.928	0.958
Generated	text-embedding-3-large	0.947	0.959	0.969	0.978
Ground Truth	jina-embeddings-v3	0.767	0.871	0.905	0.936
Generated	jina-embeddings-v3	0.916	0.952	0.959	0.973
Ground Truth	voyage-3-large	0.831	0.904	0.944	0.959
Generated	voyage-3-large	0.967	0.977	0.985	0.992

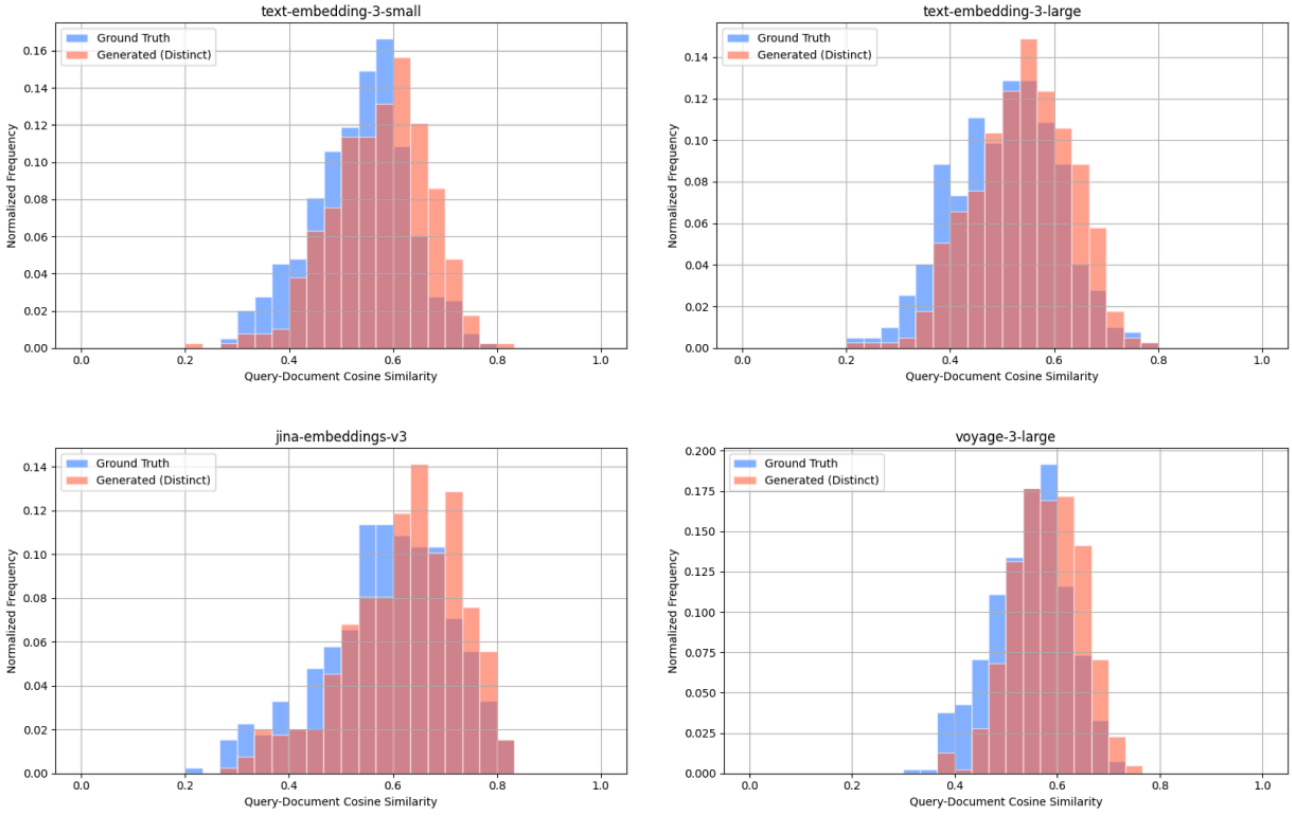
SciFact - Recall@k Scores for Ground Truth and Generated Queries across all 5 models

```
Score: 0.8204
Original Query: Does Disney allow for the return of all physical products within
days of when such product was purchased?
Generated Query: What is the time period allowed for cancelling and returning physical
goods purchased through Disney Products?

Score: 0.8040
Original Query: Does Shopify agree to ensure the encryption of all data provided to
it?
Generated Query: What are the requirements for compliance with data encryption during
payment processing according to Shopify's Terms of Service?

Score: 0.8013
Original Query: Is the general rule that legal action against Amazon must take place
by way of arbitration proceedings?
Generated Query: What is the process for initiating an arbitration proceeding against
Amazon according to their dispute resolution terms?

Score: 0.7989
Original Query: According to the terms, does CNN create third party content?
Generated Query: What are CNN's responsibilities and limitations regarding third-party
content on their site?

Score: 0.7948
Original Query: By uploading content on Shopify services, do I give up my IP rights in
that content?
Generated Query: What happens to your intellectual property rights when uploading
materials to Shopify?
```

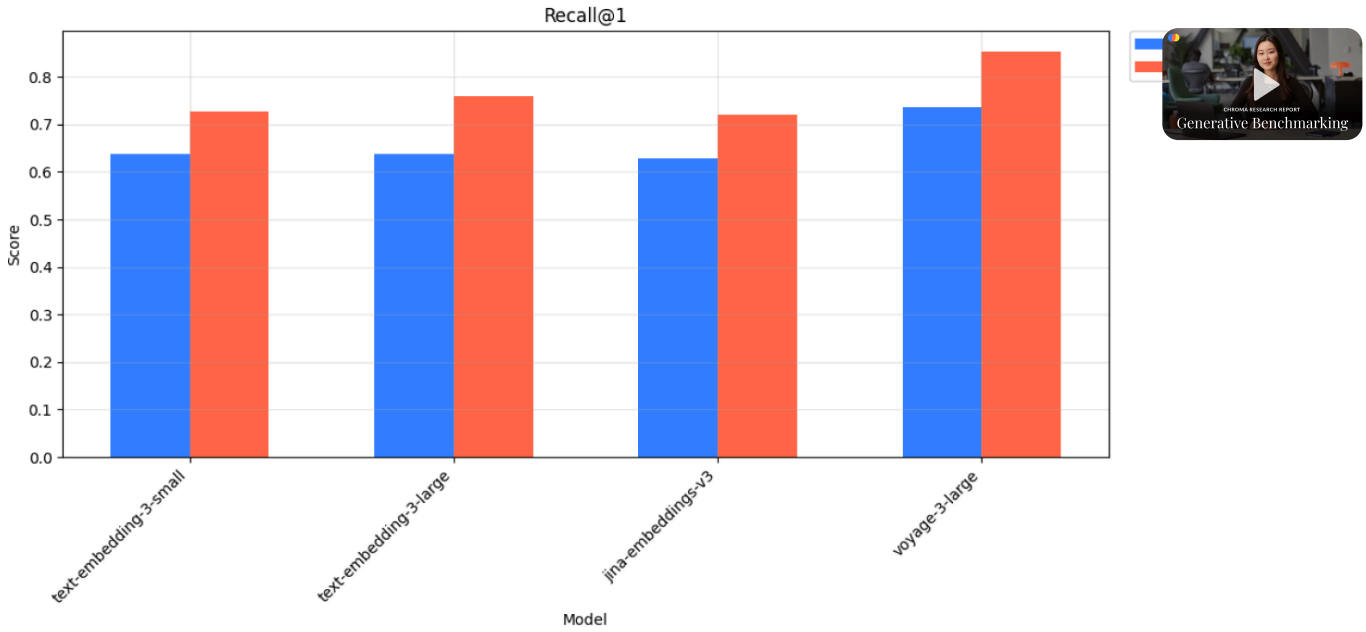
LegalBench Consumer Contracts QA - reproduced queries, cosine similarity scores calculated using text-embedding-3-small.



LegalBench Consumer Contracts QA - query-query cosine similarity scores between: ground truth query & naively generated query (blue), ground truth query & distinct generated query (red).



LegalBench Consumer Contracts QA - query-document cosine similarity scores between: ground truth query & target document (blue), distinct generated query & target document (red).



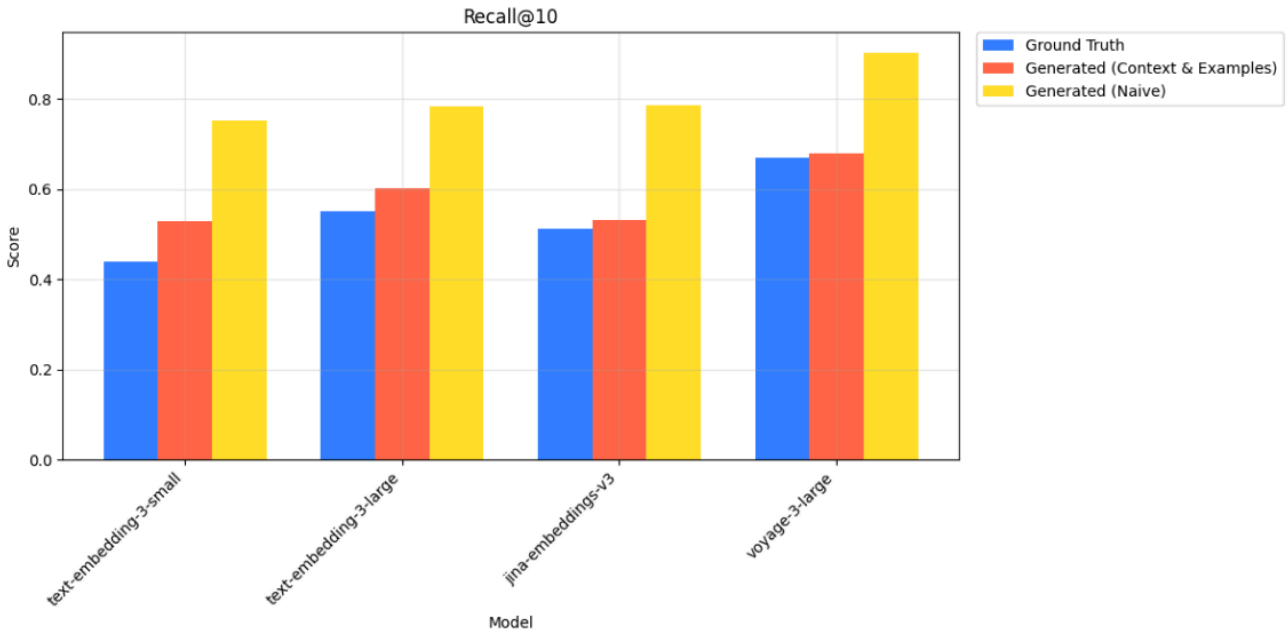
LegalBench Consumer Contracts QA - Recall@1 Scores for Ground Truth and Generated Queries across 4 models.

Query Type	Model	Recall@1	Recall@3	Recall@5	Recall@10
Ground Truth	text-embedding-3-small	0.639	0.813	0.886	0.942
Generated	text-embedding-3-small	0.727	0.914	0.947	0.975
Ground Truth	text-embedding-3-large	0.639	0.836	0.889	0.949
Generated	text-embedding-3-large	0.760	0.927	0.957	0.982
Ground Truth	jina-embeddings-v3	0.629	0.803	0.869	0.937
Generated	jina-embeddings-	0.720	0.899	0.934	0.957

Query Type	Model	Recall@1	Recall@3	Recall@5	Rec
Ground Truth	voyage-3-large	0.737	0.866	0.902	0.949
Generated	voyage-3-large	0.854	0.960	0.980	0.992

LegalBench Consumer Contracts QA - Recall@k Scores for Ground Truth and Generated Queries across 4 models.

WandBot



WandBot - Recall@10 Scores for ground truth queries, generated queries (with context and examples), and naively generated queries across 4 models.

Try Chroma Cloud

Chroma is the open-source search and retrieval database for AI applications.