

ROUTIR: Fast Serving of Retrieval Pipelines for Retrieval-Augmented Generation

Eugene Yang¹, Andrew Yates¹, Dawn Lawrie¹,
James Mayfield¹, and Trevor Adriaanse²

¹ Human Language Technology Center of Excellence, Johns Hopkins University,
Baltimore, MD 21211, USA

{eugene.yang, andrew.yates, lawrie, mayfield}@jhu.edu

² Johns Hopkins University, Baltimore, MD 21211, USA
tadriaa1@jhu.edu

Abstract. Retrieval models are key components of Retrieval-Augmented Generation (RAG) systems, which generate search queries, process the documents returned, and generate a response. RAG systems are often dynamic and may involve multiple rounds of retrieval. While many state-of-the-art retrieval methods are available through academic IR platforms, these platforms are typically designed for the Cranfield paradigm in which all queries are known up front and can be batch processed offline. This simplification accelerates research but leaves state-of-the-art retrieval models unable to support downstream applications that require online services, such as arbitrary dynamic RAG pipelines that involve looping, feedback, or even self-organizing agents. In this work, we introduce ROUTIR, a Python package that provides a simple and efficient HTTP API that wraps arbitrary retrieval methods, including first stage retrieval, reranking, query expansion, and result fusion. By providing a minimal JSON configuration file specifying the retrieval models to serve, ROUTIR can be used to construct and query retrieval pipelines on-the-fly using any permutation of available models (e.g., fusing the results of several first-stage retrieval methods followed by reranking). The API automatically performs asynchronous query batching and caches results by default. While many state-of-the-art retrieval methods are already supported by the package, ROUTIR is also easily expandable by implementing the *Engine* abstract class. The package is open-sourced and publicly available on GitHub: <http://github.com/hltcoe/routir>.

Keywords: search service · retrieval-augmented generation · asynchronous query batching · multi-stage retrieval · online evaluation

1 Introduction

Information retrieval research typically requires system comparison using a fixed set of queries and documents, following the Cranfield paradigm [7]. Such experiments can be conducted through either batched or sequential query processing, depending on whether measuring query latency is critical to the study. Regardless, the queries are predefined, so the experimental environment can be static

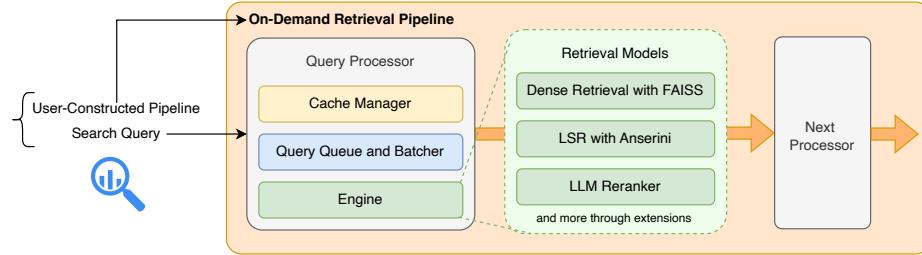


Fig. 1. Service architecture of ROUTIR. The user’s HTTP request specifies a retrieval pipeline and search query, illustrated on the left. ROUTIR orchestrates the retrieval pipeline on the fly and processes the query with each query processor. Each processor manages queuing, caching, and query processing. The “Next Processor” enables, for example, reranking or result fusion. See Figure 2 for a full JSON example of a request.

and read directly from a file. However, embedding retrieval models in a larger dynamic system pipeline creates challenges and overhead when conducting experiments.

Particularly in Retrieval-Augmented Generation (RAG), the primary pipeline involves one or more large language models (LLMs) [9, 46] that generate search queries, digest retrieved documents, and draft the response. Instead of a linear process such as Fusion-in-Decoder [17] or GINGER [20], RAG systems are increasingly complex and dynamic, with multiple rounds of retrieval [1, 9, 13, 39, 47]. A static experimental environment is not sufficient for these systems, because the queries are not known up front. Tools for offline query processing need to be modified to include the generative components or wrapped to provide an API that can accommodate online querying. However, wrapping offline retrieval pipelines with reasonable query and resource efficiency is non-trivial, and this task is orthogonal to implementing a retrieval model. We introduce an open-source package, ROUTIR, that allows users to provide pipelines composed of the latest retrieval models as an HTTP API for use with downstream applications like dynamic RAG systems.

ROUTIR supports (1) simple service configuration for common model architectures without additional coding; (2) minimal wrappers for incorporating new models; (3) dynamic batching and queuing for concurrent and asynchronous requests; (4) fast and robust caching in memory and Redis;³ and (5) reliable and easy-to-use HTTP API endpoints that do not require additional client-side packages. Such capabilities provide fast and reproducible experimental environments for RAG research, while enabling new retrieval models to be incorporated easily with minimal engineering overhead. While ROUTIR is designed primarily for academic research, which does not implement various security measures for an API service, it can be used for internal prototyping in industry to spin up a proof-of-concept application.

³ <https://redis.io/>

ROUTIR has been deployed in various research settings where it has been demonstrated to be robust and reliable. During JHU SCALE 2025,⁴ a ten-week research workshop at Johns Hopkins University attended by more than 50 researchers working on long-form RAG, ROUTIR was able to provide the retrieval API for PLAID-X [48], SPLADE-v3 [21], and Qwen3 Embedding [53] on the TREC NeuCLIR [23], TREC RAGTIME [24], TREC BioGen [16], and TREC RAG (i.e., MS-MARCO v2 Passage) [42] document collections using only three NVIDIA 24GB TITAN RTXs. In this use case, all three retrieval systems were able to provide search results with a reasonable latency without caching, and to provide results nearly instantaneously when results were cached. With asynchronous HTTP requests, which is the typical use case in RAG research, since multiple queries are usually generated and searched at the same time, ROUTIR provides a throughput of 3 to 10 queries per second depending on the underlying model. ROUTIR also powered the search service for the TREC RAGTIME track⁵ to provide the same PLAID-X model for all track participants. With only CPU resources, the endpoint provides a latency of around 600 ms on an AWS virtual machine when serving both TREC NeuCLIR and RAGTIME collections.

In this paper, we document the design decisions, architecture, and several use cases for ROUTIR. The package is publicly available on PyPI with implementation available on GitHub.

2 Related Work

Academic IR platforms have a long history going back to at least the mid-1980s [4]. Platforms like Galago [5], Indri [40], Patapsco [8], and Terrier [35] support offline experiments with traditional statistical methods, while platforms like Anserini [25, 50], Capreolus [51, 52], Experimaestro [36], FlexNeuART [2], LLM4Ranking [27], LLM-Rankers [54], OpenNIR [30], PyTerrier [32], RankLLM [54], and Tevatron [14, 29] provide support for modern neural first-stage retrieval and reranking methods [26]. These platforms are typically designed for offline IR experiments following the Cranfield paradigm. They provide reproducible indexing and searching capabilities to support experimentation where test queries are predefined and can be processed sequentially or in batch. Rather than providing general-purpose toolkits, there has been a recent trend towards streamlining experimentation using tools designed to run specific benchmarks with a predefined set of collections, such as MTEB [33] and BEIR [41]. While these academic tools are useful for offline retrieval evaluation and have greatly advanced the field, their focus on offline usage limits their ability to be embedded in larger systems. They typically do not provide Web APIs to interface with downstream applications.

This limitation particularly affects RAG, where retrieval is usually an upstream or interleaving component providing retrieved information to generative

⁴ <https://hltcoe.jhu.edu/research/scale/scale-2025/>

⁵ https://trec-ragtime.github.io/search_api.html

models [1, 9, 13, 39, 43, 46, 47]. This is a different setting from offline experimentation where the queries are fixed. RAG pipelines like Open Deep Research⁶ can dynamically generate queries to incrementally gather information for generation. There are two ways to support such iterative RAG pipelines: (re-)implement the generative component within the IR platform or provide an API that can be dynamically queried by an external generative library. PyTerrier-RAG [31] takes the former approach, whereas ROUTIR takes the latter by dynamically accepting and serving queries using state-of-the-art IR methods, including those implemented by platforms designed for offline use. Rather than reimplementing the underlying retrieval methods and RAG workflows, ROUTIR provides a streamlined approach to wrap existing methods, compose them into retrieval pipelines, and query them online through a HTTP API. This allows ROUTIR to serve the fast-growing RAG community by seamlessly embedding retrieval models into RAG platforms instead of requiring that RAG methods be reimplemented within an academic IR platform.

Production search platforms like ElasticSearch⁷, OpenSearch⁸, and Vespa⁹ provide HTTP APIs for first-stage retrieval and reranking over collections that have been indexed by those platforms. These tools are complex, closely integrated with their underlying inverted index and vector database data structures, and optimized for production use. These design choices make extending the platforms with new methods non-trivial. In contrast, ROUTIR provides a simple API and flexible method classes that make adding new retrieval models straightforward.

3 ROUTIR Architecture

ROUTIR is a thin, robust wrapper around retrieval models that provides online service capabilities that are orthogonal to the retrieval models themselves. The end user submits an HTTP request to a ROUTIR endpoint for each search query, e.g., `{"service": "qwen3-neuclir", "query": "where is machu picchu", "limit": 15}`, and receives the retrieval results of the requested 15 documents in a dictionary of document IDs and scores. A separate request delivers the document data associated with the document ID.

The core design principle of ROUTIR is to be as lightweight as possible while providing a flexible service layer to mitigate the overhead of serving state-of-the-art models when they are released. In ROUTIR, we limit dependencies to only essential packages and leave model-specific ones, such as Huggingface Transformers [45] and PyJNIs¹⁰, as extras for users to install when they are needed.

⁶ https://github.com/langchain-ai/open_deep_research

⁷ <https://www.elastic.co/elasticsearch>

⁸ <https://opensearch.org/>

⁹ <https://vespa.ai/>

¹⁰ <https://github.com/kivy/pyjnius>

In this section, we describe the ROUTIR service architecture, which is illustrated in Figure 1. ROUTIR has three main components: Engines, Processors, and Pipelines.

- *Engines* provide one or more core retrieval capabilities: first-stage retrieval using an index, reranking, query rewriting, and result fusion. To add new methods to ROUTIR, the user writes an Engine subclass that may simply wrap an existing implementation. A special *Relay* Engine can be used to access Engines provided by another node running ROUTIR.
- *Processors* receive search queries as input and perform actions before passing the queries to an Engine. By default, they are used to cache results and to batch queries that arrive in quick succession.
- *Pipelines* describe how engines are composed to produce a ranking. For example, a pipeline could indicate that results from two first-stage retrieval engines should be fused and then reranked. Pipelines can be composed of available Engines on the fly.

3.1 Retrieval Engines

Each retrieval model or system should be wrapped as an Engine, which implements the interface for serving queries. Each engine can provide one or more of the four core retrieval capabilities: (a) index searching, (b) query-passage scoring (for reranking), (c) query rewriting (or generation), and (d) result fusion.

Allowing each engine instance to provide multiple capabilities can minimize the memory footprint when different tasks share the same underlying models. For instance, most bi-encoder models can provide both first-stage retrieval that results in a ranked list of documents, and query-passage scoring (reranking) that enables fast passage selection when a RAG pipeline [6, 15, 18] needs to compress the document input to the most relevant passages to feed to a downstream generation step. To improve retrieval effectiveness, an Engine can provide a reranker, such as monoT5 [37], Qwen3 Reranker [53], or Rank1 [44]. Rerankers take a query and a list of strings (e.g. documents) as input and output a ranking over the input list. In ROUTIR, modules can be initiated in Python as a standalone model instance to provide a unified interface; this is similar to what PyTerrier extensions such as `pyterrier-colbert` provide. However, the primary benefit of this thin wrapper is that it provides a robust and flexible online search service.

Queries as inputs to the Engine are batched (as detailed in Section 3.2) to provide better service throughput; this is helpful because most bi-encoders and cross-encoders can score multiple queries and documents in the same matrix multiplication, leading to better GPU utilization. The common measurement for query-time efficiency has been query latency measured by sequentially issuing queries to the model during offline evaluation [38]. However, when serving multiple users or queries as a service, ROUTIR optimizes for throughput (i.e., the number of queries served in a fixed period of time) since queries can often be served asynchronously. Especially for a RAG pipeline that issues multiple

queries simultaneously [11, 49], all retrieval results must arrive before generation begins; this suggests the need for high-throughput rather than low-latency (although the two qualities are usually correlated).

Multi-Server Request Routing. A special type of Engine is a *Relay* – an Engine that relays requests to another ROUTIR endpoint. This capability is particularly useful when computing resources are divided among multiple machines or if some compute nodes are not exposed to a public IP (a common setup in academic research clusters). This is similar to the proxy service in LiteLLM¹¹ that relays LLM requests to compute endpoints without exposing multiple machines to the end users. While ROUTIR does not offer load-balancing at the request level (which may be included in future versions), it offers triage at the model level to direct requests to models with different resource requirements to different machines. ROUTIR also supports importing services from a list of endpoints to simplify configuration (more on this in Section 4.2). This feature provides the backbone for collectively serving multiple retrieval models with one endpoint in a distributed computing environment, which is crucial to facilitate complex retrieval pipelines (more on this later in this section).

3.2 Query Processor

Each Engine is further wrapped in a Processor class, which handles caching and queuing of input search queries. When the processor receives a query, it is added to the service queue for batching. The queue dispatches a batch of queries to the engine whenever the batch is full (size configurable) or the maximum wait time is reached (typically 50 to 100 ms; also configurable). When a set of subqueries is generated by a RAG system and sent to the endpoint individually through asynchronous HTTP requests, they are usually batched on the server side to allow them to be processed together by the Engine. The end user can also simultaneously process multiple top-level queries in a RAG pipeline and use the batching capability of the retrieval server. This exploits the asynchronous nature of HTTP requests. With the native support of asynchronous operations in Python, firing multiple retrieval and LLM requests to an external server without blocking the program from advancing to other operations until the results are actually needed is a key ingredient to accelerate the speed of RAG toolkits such as LangGraph,¹² AutoGen,¹³ DSPy [19], and GPT Researcher [12].

While batching adds some overhead in gathering queries, it prevents queries from being processed sequentially. This results in greater throughput when handling multiple queries. This is generally not handled by offline IR toolkits such as PyTerrier [32] and Anserini [50], since online serving is not the primary use case for those tools. ROUTIR provides the essential wrappers to serve retrieval mod-

¹¹ <https://github.com/BerriAI/litellm>

¹² <https://www.langchain.com/langgraph>

¹³ <https://microsoft.github.io/autogen>

```

1  {
2    "pipeline": "{qwen3-neuclir,plaidx-neuclir}RRF%
3    "collection": "neuclir",
4    "query": "where is Taiwan"
5 }
```

Fig. 2. Example Pipeline Request. The pipeline issues the query to `qwen3-neuclir` and `plaidx-neuclir` engines, fuses the results with reciprocal rank fusion, takes the top 50 documents from the fused result, and finally reranks using Rank1 [44] reranker.

Table 1. Operators for pipeline construction string. Please refer to <https://github.com/hltcoe/routir/blob/main/src/routir/pipeline/parser.py#L7> for the full context-free grammar.

Operator	Operation	Description
<code>e1 >> e2</code>	Pipe	Pass the retrieval results of <code>e1</code> to a downstream engine <code>e2</code> , such as a reranker.
<code>e1%k</code>	Limit	Only retain the top k retrieved documents from <code>e1</code> .
<code>{e1, e2}</code>	Parallel Pipelines	Pass the upstream results or query (if at the beginning of a pipeline) to a list of parallel pipelines (<code>e1</code> and <code>e2</code>).
<code>xx{e1, e2}</code>	Query Generation	Generate multiple sub-queries with method <code>xx</code> and issue them to all parallel pipelines.
<code>}xx</code>	Result Fusion	Fuse retrieval results from parallel pipelines with method <code>xx</code> .

els, including those supported by PyTerrier and Anserini, to efficiently embed them in a RAG system pipeline.

Furthermore, processors also cache retrieval results to prevent duplicate requests to the Engine instance. ROUTIR supports both in-memory and Redis caches, which provides flexibility to support different cache integrity needs.

3.3 On-demand Pipeline Construction

In addition to serving the query with a single retrieval model, ROUTIR supports on-demand pipeline construction from the end user request. This is illustrated in Figure 2, where the pipeline combines the results of two first-stage retrievers using reciprocal rank fusion and reranks the fused results. ROUTIR parses the pipeline string provided by the user. It understands that the Engines corresponding to the two first-stage retrievers can be run in parallel and that fusion and reranking are sequential steps with the `>>` pipe operator. In addition, asynchronous requests are issued to prioritize throughput when producing the final retrieval results.

The pipeline string is defined using a context-free grammar that supports the construction of linear pipelines. Table 1 summarizes the operators available

for the pipeline construction string. Dynamic pipeline construction allows a user or RAG system to control the pipeline as needed on the fly to accommodate runtime constraints such as latency, coverage, and query difficulty. The context-free grammar can even be part of the input to the language model, enabling it to generate the pipeline as part of an agentic workflow.

4 Serving Models using ROUTIR

In this section, we describe how ROUTIR can be configured and extended to serve different retrieval models. This serves as an introduction to all the features provided by ROUTIR; readers are encouraged to explore further in the documentation and the source code on GitHub.

4.1 Resource and Setup

Computing resources needed to run the barebone ROUTIR are very minimal, for example, a single processor with 200 MB memory can host a ROUTIR instance with only Relay Engine. However, resources for hosting each retrieval model depend on its own requirements. For example, it is more efficient to host a dense retrieval model with a GPU for encoding the queries; Faiss indexes usually require a larger system memory to hold the in-memory index for better performance.

ROUTIR can be installed through `pip` or `uv`. Please refer to the documentation for more details. It can also be hosted with a command as simple as `uvx routir config.json` without explicit package installation.

4.2 Server Configuration

ROUTIR uses a JSON file to express the configuration for the services with two primary blocks: `services` and `collections`. The `services` block is a list of dictionaries that specifies all the Engines to initialize on the endpoint. The `collections` block specifies a list of document collections that the endpoint will serve based on the document IDs requested. The additional top-level `server_imports` and `file_imports` fields allow the user to specify external ROUTIR endpoints and custom Python scripts to include during the initialization. All available services on each endpoint in `server_imports` are automatically relayed. This allows offloading computationally expensive models, such as LLM reranking, to another machine, while still enabling their integration with other services in the end user’s custom retrieval pipelines. Figure 3 demonstrates an example configuration JSON object.

Each dictionary in the `services` list defines a processor and its underlying Engine. The field `engine` specifies the Engine class to initiate, which can be any of those included in ROUTIR as built-in engines such as `PLAIDX`, or ones that are implemented by the user in a separate Python script. This is accomplished via

```

1  {
2    "server_imports": [ "http://localhost:5000" ],
3    "file_imports": [ "./examples/rank1_extension.py" ],
4    "services": [
5      {
6        "name": "rank1",
7        "engine": "Rank1Engine",
8        "config": {}
9      },
10     {
11       "name": "qwen3-neuclir",
12       "engine": "Qwen3",
13       "batch_size": 16,
14       "config": {
15         "index_path": "hfds:routir/neuclir-qwen3-8b-faiss-
16             PQ2048x4fs",
17         "embedding_model_name": "Qwen/Qwen3-Embedding-8B",
18       }
19     },
20   ],
21   "collections": [
22     {
23       "name": "neuclir",
24       "doc_path": "./neuclir.collection.jsonl"
25     }
26   ]
}

```

Fig. 3. Example configuration file. Refer to the ROUTIR documentation for more fields. In the example, we load two engines: Rank1 reranker from a custom script (see Section 4.3), and qwen3-neuclir using a FAISS index loaded from Huggingface Datasets containing Qwen3 8B embeddings. The specified document collection can be used to retrieve document text via an API or to pass documents to reranking models.

the `file_imports` field that allows users to specify scripts to load on-demand without modifying the package or implementing another custom entry script.

The `collections` field lists each collection serviced by an engine in a dictionary containing the name of the collection and a path to a JSONL file. Each JSON object is a document containing an ID field with arbitrary fields carrying its content. ROUTIR loads each JSONL collection file and builds a memory offset lookup table to efficiently look up the document when serving the content. Such lookup tables allow ROUTIR random access the collection file based on document IDs without reading the file sequentially or loading the entire collection into memory.

ROUTIR provides several built-in Engine types that can be used to serve models with common architectures such as dense bi-encoders with FAISS indices, multi-vector dense retrieval with PLAID-X [48], and learned-sparse retrieval

```

1 class PyseriniBM25(Engine):
2     def __init__(self, name: str = None, config=None, **kwargs):
3         super().__init__(name, config, **kwargs)
4         self.searcher = LuceneSearcher(self.index_path)
5         self.searcher.set_bm25(0.9, 0.4)
6
7     @async
8     def search_batch(self, queries, limit=20):
9         return [
10             {docobj.docid: docobj.score
11              for docobj in self.searcher.search(query, k=lm)}
12             for query, lm in zip(queries, limit)
13         ]

```

Fig. 4. Example code snippet for integrating Pyserini with ROUTIR. This example can be extended with more flexible parameter configuration or even allowing the endpoint users to specify the retrieval model. Full example can be found at https://github.com/hltcoe/routir/blob/main/examples/pyserini_extension.py.

models with Anserini [50]. These Engines are implemented based on an `Engine` Python abstract class that ensures a common interface.

These Engine types cover most use cases when using neural models. However, since there is not yet a common reranker architecture (besides the general pointwise, pairwise, listwise, and setwise paradigms), we have selectively implemented a few rerankers as built-in Engines with an example script for loading a more complex model through `file imports`.¹⁴ In the next section, we describe how to incorporate an external IR toolkit into ROUTIR.

4.3 Integration with Existing IR Toolkits

Figure 4 demonstrates a simple example that integrates Pyserini [25] into ROUTIR. For index retrieval, one only needs to implement the initialization of the Engine, which contains index loading and hyperparameter settings if applicable, and the `search_batch` method, which takes a batch of queries and returns a list of results in the same order as the input queries.

Since all search methods are asynchronous, ROUTIR does not wait for a module to finish searching before attending to the next API request. However, since Python asynchronous functions are still single-threaded on a single processor, unless the Engine spawns another search process or calls out to another process, processes can be blocked. For example, a standalone Lucene instance searching a batch of queries may block the Python process from accepting an API request. We provide some implementation guidance on how these concurrency issues can be overcome in the documentation. However, such engineering issues are model

¹⁴ Rank1 [44], a pointwise reasoning reranker, integration script can be found in https://github.com/hltcoe/routir/blob/main/examples/rank1_extension.py.

Table 2. Effectiveness and efficiency on the NeuCLIR 2023 MLIR Task.

Model	Effectiveness	Batched Throughput	Seq. Latency
	nDCG@20	(query/sec) \uparrow	(sec/query) \downarrow
Multi-Vector: PLAID-X [48]	0.402	7.05	0.24
LSR: MILCO [34]	0.413	3.27	2.46
Dense: Qwen3 Embedding [53]	0.430	9.60	1.23

and toolkit-dependent. They can generally be solved by hosting different models in separate ROUTIR instances and combining them via `server_imports` to a joint endpoint.

5 Experiment and Analysis

To demonstrate the adaptability of ROUTIR, we report effectiveness and efficiency using the TREC 2023 NeuCLIR MLIR task [22], which has 76 queries and about 10 million web documents in Chinese, Persian, and Russian extracted from CommonCrawl News. We experiment with the following three multilingual models with distinct architectures and stacks:

- Multi-vector dense using PLAID-X [48]. The PLAID-X model was reported as the state of the art in 2023 during TREC. The model is based on XLM-RoBERTa-Large [28] and is served using an NVIDIA TITAN RTX 24G with the PLAID-X implementation.
- Learned-sparse retrieval using MILCO [34] with Anserini [50]. The MILCO model is also based on XLM-RoBERTa-Large and is served with the same GPU using the Huggingface Transformer. The index is served with Anserini via PyJNIs.
- Dense retrieval using Qwen3 Embedding [53] with a FAISS index [10] using vLLM.¹⁵ The Qwen3 Embedding model is served with vLLM with parameters cast to FP16 to fit on the TITAN RTX GPU. The document embeddings are indexed with FAISS using product quantization of 2048 dimensions and 4-bit fast scan (`PQ2048x4fs`).

We experimented with two request modes: batched and sequential. When queries are batched, we issue all 76 queries with asynchronous HTTP requests to the ROUTIR endpoints and report the throughput, i.e., the number of queries processed per second. In the sequential mode, we issued the next query after receiving results from the previous one and report the latency, i.e., the number of seconds to process each query.

As shown in Table 2, all three models demonstrate strong throughput, although LSR with Anserini is the slowest. However, it can be greatly accelerated with tools such as Seismic [3] that are tailored for LSR searching. All three models are able to take advantage of batched queries to provide faster overall speed.

¹⁵ <https://github.com/vllm-project/vllm>

```

1  class LiveRAG_PLAIDX_Search():
2      def __init__(self, query: str, **kwargs):
3          self.url = os.getenv("RETRIEVER_ENDPOINT")
4          self.query = query
5
6      def get_content(self, collection, doc_id):
7          return requests.post(self.url+"/content", json={
8              "collection": collection, "id": doc_id
9          }).json()
10
11     def search(self, max_results: int = 5):
12         response = requests.post(f"{self.url}/query", json={
13             "service": "plaidx-liverag",
14             "query": str(self.query), "limit": max_results
15         }).json()["result"]
16
17     return [
18         {"score": score, "href": str(doc_id),
19          **self.get_content("liverag", doc_id)}
20         for doc_id, score in results.items()
21     ]

```

Fig. 5. Example search module for GPT Researcher.

Particularly in FAISS, ROUTIR uses the batched search capability to search the index for all queries in the batch at once, which has the highest throughput despite being five times slower than PLAID-X in sequential latency.

These results demonstrate the adaptivity of ROUTIR and robustness to the input types. In the next section, we demonstrate how to integrate ROUTIR with a RAG system.

6 Example of Using ROUTIR in a RAG System

GPT Researcher is an open-source RAG toolkit that integrates various search engines such as Google and DuckDuckGo into a RAG system that supports planning, self-reflections, and multi-agent pipelines. Figure 5 depicts a retriever in GPT Researcher interfacing with ROUTIR, requiring only 21 lines of code. Since ROUTIR uses the HTTP API as the interface, it does not require any other dependencies to interface with GPT Researcher. Figure 5 is an abbreviated version of the implementation (excluding some sanity checks and try-except blocks) used in HLTCOE’s system for the LiveRAG challenge at SIGIR 2025 [11]. While it would also be possible to implement GPT Researcher inside an academic IR platform, doing so would require substantially more engineering effort.

7 Summary

In this work, we introduced ROUTIR, a simple and robust toolkit for wrapping and serving retrieval models to serve online queries. We described the design principles and architecture of ROUTIR and presented several use cases both in configuring the service and integration with RAG pipelines. This highlighted its ability to increase query throughput, a particularly desirable feature for RAG systems. ROUTIR has demonstrated its effectiveness and reliability as the backbone for the 2025 JHU SCALE Workshop for more than 50 researchers and also as the search service for 2025 TREC RAGTIME Track.

ROUTIR is still under development with more features in the near term planned, including integration of LLM rerankers, Model Context Protocol (MCP) interface, and better resource management. ROUTIR is completely open-sourced on GitHub and welcomes community feedback, feature requests, and pull requests.

References

1. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-rag: Learning to retrieve, generate, and critique through self-reflection (2024)
2. Boytsov, L., Nyberg, E.: Flexible retrieval with nmslib and flexneuart. arXiv preprint arXiv:2010.14848 (2020)
3. Bruch, S., Nardini, F.M., Rulli, C., Venturini, R.: Efficient inverted indexes for approximate retrieval over learned sparse representations. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 152–162, ACM (2024), <https://doi.org/10.1145/3626772.3657769>, URL <https://doi.org/10.1145/3626772.3657769>
4. Buckley, C.: Implementation of the smart information retrieval system. Tech. rep., Cornell University (1985)
5. Cartright, M.A., Huston, S.J., Feild, H.: Galago: A modular distributed processing and retrieval system. In: OSIR@SIGIR, pp. 25–31 (2012)
6. Cheng, X., Wang, X., Zhang, X., Ge, T., Chen, S.Q., Wei, F., Zhang, H., Zhao, D.: xrag: Extreme context compression for retrieval-augmented generation with one token. Advances in Neural Information Processing Systems **37**, 109487–109516 (2024)
7. Cleverdon, C.: The aslib cranfield research project on the comparative efficiency of indexing systems. Aslib Proceedings **12**(12), 421–431 (12 1960), ISSN 0001-253X, <https://doi.org/10.1108/eb049778>, URL <https://doi.org/10.1108/eb049778>
8. Costello, C., Yang, E., Lawrie, D., Mayfield, J.: Patapsco: A python framework for cross-language information retrieval experiments. In: Proceedings of the 44th European Conference on Information Retrieval (ECIR) (2022)
9. Dong, G., Jin, J., Li, X., Zhu, Y., Dou, Z., Wen, J.R.: Rag-critic: Leveraging automated critic-guided agentic workflow for retrieval augmented genera-

- tion. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3551–3578 (2025)
10. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvassy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024)
 11. Duh, K., Yang, E., Weller, O., Yates, A., Lawrie, D.: HLTCOE at LiveRAG: GPT-Researcher using ColBERT retrieval. arXiv preprint arXiv:2506.22356 (2025)
 12. Elovic, A.: gpt-researcher (Jul 2023), URL <https://github.com/assafelovic/gpt-researcher>
 13. Fang, J., Meng, Z., MacDonald, C.: KiRAG: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 18969–18985, Association for Computational Linguistics, Vienna, Austria (Jul 2025), ISBN 979-8-89176-251-0, <https://doi.org/10.18653/v1/2025.acl-long.929>, URL <https://aclanthology.org/2025.acl-long.929/>
 14. Gao, L., Ma, X., Lin, J.J., Callan, J.: Tevatron: An efficient and flexible toolkit for dense retrieval. ArXiv **abs/2203.05765** (2022)
 15. Guo, S., Ren, Z.: Dynamic context compression for efficient rag. arXiv preprint arXiv:2507.22931 (2025)
 16. Gupta, D., Demner-Fushman, D., Hersh, W., Bedrick, S., Roberts, K.: TREC biogen track 2025 (2025), URL <https://trec-biogen.github.io/>
 17. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering. In: EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics, pp. 874–880, Association for Computational Linguistics (2021)
 18. Jeong, Y., Kim, J., Lee, D., Hwang, S.w.: ECORAG: Evidentiality-guided compression for long context RAG. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025, pp. 26607–26628, Association for Computational Linguistics, Vienna, Austria (Jul 2025), ISBN 979-8-89176-256-5, <https://doi.org/10.18653/v1/2025.findings-acl.1365>, URL <https://aclanthology.org/2025.findings-acl.1365/>
 19. Khattab, O., Singhi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T.T., Moazam, H., et al.: Dspy: Compiling declarative language model calls into self-improving pipelines. arXiv preprint arXiv:2310.03714 (2023)
 20. Lajewska, W., Balog, K.: Ginger: Grounded information nugget-based generation of responses. In: Proceedings of the 48th International ACM SIGIR Conference (SIGIR '25) (2025), URL <https://krisztianbalog.com/files/sigir2025-ginger.pdf>, SIGIR 2025 paper
 21. Lassance, C., Déjean, H., Formal, T., Clinchant, S.: Splade-v3: New baselines for splade (2024), URL <https://arxiv.org/abs/2403.06789>
 22. Lawrie, D., MacAvaney, S., Mayfield, J., McNamee, P., Oard, D.W., Soldaini, L., Yang, E.: Overview of the TREC 2023 NeuCLIR track. In: Proceedings of The Thirty-Second Text REtrieval Conference, NIST (2023)

23. Lawrie, D., MacAvaney, S., Mayfield, J., McNamee, P., Oard, D.W., Soldaini, L., Yang, E.: Overview of the TREC 2024 neuclir track. arXiv preprint arXiv:2509.14355 (2025)
24. Lawrie, D., MacAvaney, S., Mayfield, J., Soldaini, L., Yang, E., Yates, A.: TREC RAGTIME track 2025 (2025), URL <https://trec-ragtime.github.io/>
25. Lin, J., Ma, X., Lin, S.C., Yang, J.H., Pradeep, R., Nogueira, R.: Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 2356–2362, SIGIR ’21, Association for Computing Machinery, New York, NY, USA (2021), ISBN 9781450380379, <https://doi.org/10.1145/3404835.3463238>, URL <https://doi.org/10.1145/3404835.3463238>
26. Lin, J., Nogueira, R., Yates, A.: Pretrained Transformers for Text Ranking: BERT and Beyond. Springer Nature (2022)
27. Liu, Q., Duan, H., Chen, Y., Lu, Q., Sun, W., Mao, J.: Llm4ranking: An easy-to-use framework of utilizing large language models for document reranking (2025), URL <https://arxiv.org/abs/2504.07439>
28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach (2019)
29. Ma, X., Gao, L., Zhuang, S., Zhan, J.S., Callan, J., Lin, J.: Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. arXiv preprint arXiv:2505.02466 (2025)
30. MacAvaney, S.: Opennir: A complete neural ad-hoc ranking pipeline. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 845–848 (2020)
31. Macdonald, C., Fang, J., Parry, A., Meng, Z.: Constructing and evaluating declarative rag pipelines in pyterrier. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 4035–4040 (2025)
32. Macdonald, C., Tonello, N., MacAvaney, S., Ounis, I.: Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, p. 4526–4533, CIKM ’21, Association for Computing Machinery, New York, NY, USA (2021), ISBN 9781450384469, <https://doi.org/10.1145/3459637.3482013>, URL <https://doi.org/10.1145/3459637.3482013>
33. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: MTEB: Massive text embedding benchmark. In: Vlachos, A., Augenstein, I. (eds.) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2014–2037, Association for Computational Linguistics, Dubrovnik, Croatia (May 2023), <https://doi.org/10.18653/v1/2023.eacl-main.148>, URL <https://aclanthology.org/2023.eacl-main.148/>

34. Nguyen, T., Lei, Y., Ju, J.H., Yang, E., Yates, A.: Milco: Learned sparse retrieval across languages via a multilingual connector. arXiv [cs.IR] (2025)
35. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier : A high performance and scalable information retrieval platform. In: Proceedings of the 2006 SIGIR Open Source Workshop (2006), URL <https://api.semanticscholar.org/CorpusID:16510983>
36. Piwowarski, B.: Experimaestro and datamaestro: Experiment and dataset managers (for ir). In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2173–2176 (2020)
37. Pradeep, R., Nogueira, R., Lin, J.: The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models (2021), URL <https://arxiv.org/abs/2101.05667>
38. Schurman, E., Brutlag, J.: Performance related changes and their user impact. In: velocity web performance and operations conference (2009)
39. Shao, Y., Jiang, Y., Kanell, T.A., Xu, P., Khattab, O., Lam, M.S.: Assisting in writing wikipedia-like articles from scratch with large language models. arXiv preprint arXiv:2402.14207 (2024)
40. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis, vol. 2, pp. 2–6, Amherst, MA, USA (2005)
41. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models (2021), URL <https://arxiv.org/abs/2104.08663>
42. Upadhyay, S., Pradeep, R., Thakur, N., Lin, J., Craswell, N.: TREC rag track 2025 (2025), URL <https://trec-rag.github.io/>
43. Wang, H., Prasad, A., Stengel-Eskin, E., Bansal, M.: Retrieval-augmented generation with conflicting evidence. In: Proceddings of the 2nd Conference on Language Modeling (2025)
44. Weller, O., Ricci, K., Yang, E., Yates, A., Lawrie, D., Durme, B.V.: Rank1: Test-time compute for reranking in information retrieval. In: Second Conference on Language Modeling (2025)
45. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Huggingface’s transformers: State-of-the-art natural language processing (2020), URL <https://arxiv.org/abs/1910.03771>
46. Xu, R., Shi, W., Zhuang, Y., Yu, Y., Ho, J.C., Wang, H., Yang, C.: Collab-RAG: Boosting retrieval-augmented generation for complex question answering via white-box and black-box LLM collaboration. In: Second Conference on Language Modeling (2025), URL <https://openreview.net/forum?id=C0Ds4jSGhN>
47. Yan, S.Q., Gu, J.C., Zhu, Y., Ling, Z.H.: Corrective retrieval augmented generation. arXiv preprint arXiv:2401.15884 (2024)

48. Yang, E., Lawrie, D., Mayfield, J., Oard, D.W., Miller, S.: Translate-distill: learning cross-language dense retrieval by translation and distillation. In: European Conference on Information Retrieval, pp. 50–65, Springer (2024)
49. Yang, E., Lawrie, D., Weller, O., Mayfield, J.: HLT COE at TREC 2024 NeuCLIR track (2025), URL <https://arxiv.org/abs/2510.00143>
50. Yang, P., Fang, H., Lin, J.: Anserini: Enabling the use of lucene for information retrieval research. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 1253–1256, SIGIR ’17, Association for Computing Machinery, New York, NY, USA (2017), ISBN 9781450350228, <https://doi.org/10.1145/3077136.3080721>, URL <https://doi.org/10.1145/3077136.3080721>
51. Yates, A., Arora, S., Zhang, X., Yang, W., Jose, K.M., Lin, J.: Capreolus: A toolkit for end-to-end neural ad hoc retrieval. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 861–864 (2020)
52. Yates, A., Jose, K.M., Zhang, X., Lin, J.: Flexible ir pipelines with capreolus. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 3181–3188 (2020)
53. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025)
54. Zhuang, S., Zhuang, H., Koopman, B., Zucccon, G.: A setwise approach for effective and highly efficient zero-shot ranking with large language models. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24 (2024)