Hypotheticals Document Embedding (HyDE) - Analysis and evaluation

Matthieu Olekhnovitch 3rd Year Student
ENSAE
molekhnovitch@ensae.fr

Abstract

Hypothetical Document Embeddings (HyDE) demonstrated a powerful approach for effective zero-shot dense retrieval without relevance labels, pivoting through LLM-generated documents and unsupervised contrastive encoders. While HyDE has shown strong performance comparable to supervised methods on tasks often emphasizing top-k relevance (e.g., web search, single-answer QA), its efficacy in multi-retrieval settings—where a query requires identifying a set of distinct relevant documents—remains less explored. This project re-evaluates the HyDE paradigm specifically for multi-retrieval tasks, such as evidence gathering for complex question answering or retrieving diverse perspectives. We investigate whether the single hypothetical document generated by the LLM, combined with the lossy compression of the contrastive encoder, adequately captures the necessary signal diversity to identify multiple, distinct relevant documents within the corpus embedding space. We conduct experiments on established multi-retrieval benchmarks, analyzing HyDE's performance using set-based metrics (e.g., Recall@k, nDCG@k across a wider k, F1@k). Our analysis aims to quantify HyDE's effectiveness and potential limitations in scenarios demanding comprehensive information retrieval, comparing its performance against unsupervised baselines and providing insights into its applicability for tasks **beyond single-best-answer retrieval**.

1 State-of-the-Art in Zero-Shot Dense Retrieval

1.1 Shift Towards Dense Retrieval and the Supervision Challenge

Information retrieval has seen a significant shift from traditional sparse, lexical methods (e.g., BM25) towards dense retrieval techniques, driven by advances in deep learning and transformer architectures. Dense retrieval maps queries and documents into a shared low-dimensional semantic space, allowing for relevance matching based on meaning rather than just keyword overlap. However, the effectiveness of these models typically hinges on large-scale supervised training data, often involving query-document pairs with relevance labels, like the MS-MARCO dataset. This reliance on supervision poses challenges, as such datasets are expensive to create, may not exist for specific domains or languages, and can carry licensing restrictions limiting practical use (for some inaccessible).

1.2 Approaches Mitigating Supervision Dependence

To address this, several research directions have emerged. Transfer learning is a common paradigm, where models are pre-trained or fine-tuned on a large, labeled dataset (like MS-MARCO) and then applied to target tasks, often evaluated using benchmarks like BEIR. While often effective, this still presupposes the availability and suitability of a large source dataset.

Unsupervised dense retrieval aims to remove the need for relevance labels entirely during the encoder training phase. A prominent approach involves contrastive learning, exemplified by Contriever.

Contriever learns document representations by training an encoder to pull embeddings of augmented versions of the same document closer together while pushing apart embeddings of different documents. While purely unsupervised, Contriever often underperforms supervised models and sometimes struggles against strong lexical baselines like BM25, especially in zero-shot settings.

1.3 Hypothetical Document Embeddings (HyDE)

Recently, Hypothetical Document Embeddings (HyDE) introduced a novel paradigm for precise zero-shot dense retrieval. HyDE ingeniously sidesteps the difficulty of learning query-document relevance matching without labels. Instead, it leverages two key components:

1. An instruction-following Large Language Model (LLM) (we used TinyLlama-1.1B-Chat-v1.0): Given a user query, the LLM is instructed to generate a hypothetical document that answers or addresses the query. This generated document captures the relevance patterns expected in a real answer, though it may contain factual inaccuracies.
2. An unsupervised contrastive encoder (we used BAAI/bge-small-en-v1.5): This encoder, trained only on document similarities, maps the hypothetical document into an embedding vector. This vector is then used to search against the embeddings of the real document corpus (encoded using the same unsupervised encoder).

HyDE effectively offloads the task of understanding relevance to the powerful generative capabilities of the LLM and uses the dense encoder as a "lossy compressor" and grounding mechanism to find similar real documents in the embedding space. Therefore we then rely a lot more on the LLM itself to find good 'hypothetical' answers.

1.4 Original HyDE Performance and Limitations

The original work demonstrated that HyDE significantly outperforms unsupervised Contriever and achieves performance competitive with fully supervised, fine-tuned models (like ContrieverFT, ANCE, DPR) across web search, QA, fact verification, and multilingual tasks, all without requiring any relevance labels for the retrieval system itself.

However, the original evaluation of HyDE, like much work in dense retrieval, primarily focused on tasks where retrieving the single best document or a small set of top-ranked documents is key (measured by metrics like nDCG@10, Recall@1k). The efficacy of HyDE in **multi-retrieval** scenarios—where a query necessitates retrieving a set of distinct, relevant documents to provide a comprehensive answer or cover multiple facets of a topic—has not been explicitly studied. It remains an open question whether the single hypothetical document generated by the LLM provides sufficient signal diversity to effectively identify multiple, varied relevant documents within the corpus, which is the primary focus of this evaluation.

1.5 Post-HyDE Developments in Zero-Shot Retrieval

Following the introduction of HyDE, subsequent research has further explored and expanded the role of LLMs while also refining unsupervised techniques. The period from late 2022 onwards has seen several key trends emerge.

One major direction involves using LLMs more deeply for *representation enhancement*, often in an offline manner distinct from HyDE's online document generation. Instead of generating a hypothetical document at query time, methods focus on using LLMs during pre-processing or training. For instance, LLMs can generate synthetic queries for each document in the corpus; these synthetic queries augment the document's representation or are used in contrastive training frameworks, aiming to better align document embeddings with potential user intents without explicit relevance labels . Similarly, LLMs are employed for sophisticated query rewriting or expansion techniques, transforming the user's initial query into variations more likely to match relevant documents in the embedding space.

A paradigm shift is represented by *generative retrieval*, where LLMs are trained or prompted to directly generate document identifiers (e.g., titles, unique IDs, or defining prefixes) corresponding to relevant documents, bypassing the conventional retrieve-then-rank pipeline based on embedding similarity. This approach treats retrieval as a sequence generation task, leveraging the LLM's parametric knowledge to directly map queries to document pointers, potentially offering a different mechanism for capturing relevance.

Alongside LLM-centric methods, advancements continue in *unsupervised and self-supervised contrastive learning*. Building on foundations like Contriever, newer approaches focus on improving robustness to domain shifts or incorporate sophisticated data augmentation strategies. Some methods explore using LLMs not to generate hypothetical documents, but to provide weaker supervisory signals, such as estimated relevance scores between queries and passages, which can then guide the training of a dense encoder. Techniques like alternating distillation, where retriever and reranker models iteratively teach each other in an unsupervised loop, have also shown promise for improving zero-shot effectiveness.

Furthermore, hybrid methods combining these advanced zero-shot dense retrieval techniques (including HyDE itself or its successors) with traditional sparse methods like BM25 continue to be relevant, often using rank fusion strategies to achieve robust performance across diverse information needs.

In essence, the post-HyDE era is characterized by deeper integration of LLMs for both representation learning and direct retrieval, alongside refinements in unsupervised learning, all aimed at pushing the boundaries of retrieval effectiveness without reliance on costly human-annotated relevance data.

2 Experiments & Re-evaluation of HyDE for Multi-Retrieval

Building upon the motivation to assess HyDE's suitability for multi-retrieval scenarios and acknowledging the inherent challenges, this section details the experimental methodology, presents the core findings, and discusses their implications. Our primary objective is to evaluate whether HyDE's single hypothetical document generation mechanism effectively supports the retrieval of multiple, distinct relevant passages required for complex information needs.

2.1 Experimental Setup

2.1.1 Datasets and Tasks

Given the scarcity of well-documented and diverse retrieval datasets in NLP, we relied on the HuggingFace Hub as a platform for identifying suitable benchmarks. For this study, we selected a dataset with explicit multi-retrieval characteristics:

- **RAG-Mini-BioASQ:** Derived from the BIOASQ competition, this dataset focuses on biomedical semantic indexing and question answering. It is particularly well-suited for retrieval tasks, as it provides pre-split passages and, for the question-answering test set, includes a 'relevantpassageids' column that identifies multiple passages relevant to each question (up to 157).
- **Corpus Details:** The dataset comprises a corpus of 40,221 passages, each annotated with a unique identifier, alongside 4,719 question-answer pairs. Each pair is associated with up to 157 relevant passage identifiers, supporting rich multi-retrieval evaluation.

The primary task investigated is set-based passage retrieval. To ensure deterministic evaluation, we developed custom metrics and did not use the free-text 'Answer' column.

While an informative metric could be the number of questions judged well-answered by a third-party language model given the retrieved context and reference answer, we did not adopt this approach due to its non-deterministic nature and high sensitivity to the capabilities and variability of the answering model.

2.1.2 Evaluation Metrics

Evaluating performance in a multi-retrieval scenario requires metrics sensitive not only to whether the required set of 'k' relevant documents is retrieved, but also to the quality of their ranking. Finding all necessary documents buried deep in the list is less useful than finding them at the top ranks. While simple set metrics like Recall@k confirm retrieval within the top 'k' slots, they disregard the internal ranking quality which is really important in RAG systems for example.

To effectively capture both relevance and ranking quality, we adopt **Normalized Discounted Cumulative Gain (nDCG)** as our primary evaluation metric, following its use in the HyDE paper. nDCG is well-suited for this task as it rewards the retrieval of relevant documents—i.e., those included in the query's ground-truth set—while also emphasizing their rank by assigning exponentially higher weights to documents appearing earlier in the list.

Unlike the original HyDE implementation, however, we extend the cutoff rank $topk$ to a sufficiently large value to ensure that all relevant passages are included among the retrieved results. This allows the nDCG formulation to more fully reflect retrieval quality without being artificially constrained by a fixed or limited context size.

For a given query with a ground-truth set containing 'k' relevant documents, and a retrieval system returning a ranked list of size $N = topk$ (e.g., $N = 1000$ in our experiments), we compute nDCG@N that we note nDCG as follows:

1. **Relevance Assignment:** A retrieved document at rank i (where $1 \leq i \leq N$) is assigned a binary relevance score rel_i: $rel_i = 1$ if the document is in the ground-truth set for the query, and $rel_i = 0$ otherwise.
2. **Discounted Cumulative Gain (DCG@N):** This score aggregates the relevance of documents, discounting by rank:

$$DCG@N = \sum_{i=1}^{N} \frac{rel_i}{\log_2(i+1)}$$

3. **Ideal Discounted Cumulative Gain (IDCG@N):** This represents the maximum possible DCG@N score, achieved by ranking the k relevant documents perfectly at the top positions:

$$IDCG@N = \sum_{i=1}^{\min(k,N)} \frac{1}{\log_2(i+1)}$$

(This assumes the ideal ranking places all k relevant items first, up to the list limit N).

4. **Normalized DCG (nDCG@N):** The final score normalizes the actual DCG by the ideal DCG, yielding a value between 0 and 1:

$$nDCG@N = \frac{DCG@N}{IDCG@N}$$

(If IDCG@N is 0, typically nDCG@N is defined as 0).

An nDCG@N score of 1.0 indicates perfect ranking: all relevant documents found within the top N ranks are placed at the very top of the list, in the best possible order up to rank $\min(k, N)$.

As the number of required relevant documents (k) typically varies per query, we compute these metrics on a per-query basis and report the average scores across the entire test set to represent overall system performance.

2.1.3 Models and Baselines

We compare the following retrieval approaches:

- **Unsupervised Encoder Alone:** The base contrastive encoder used in HyDE (here we used BAAI/bge-small-en-v1.5), applied directly to the query embedding. This isolates the contribution of the HyDE generation step.
- **HyDE:** Our primary model of interest. :
  - We used TinyLlama/TinyLlama-1.1B-Chat-v1.0 for hypothetical document generation. The prompt is the same used in HyDE paper for COVID19 dataset.
  - Unsupervised Encoder: BAAI/bge-small-en-v1.5
- **Sparse Baseline:** BM25

2.1.4 Implementation Details

All experiments were conducted on a local machine equipped with an NVIDIA RTX 3060 GPU (12GB VRAM). The software environment was built around Python 3.12 with CUDA support enabled to accelerate neural model inference where applicable. To maintain full control over the experimental setup and avoid external dependencies or rate limits, all models were executed locally.

For vector-based retrieval, we utilized the FAISS library with a flat index, running on CPU. While neural inference was GPU-accelerated, we deliberately chose to keep the FAISS store on CPU to ensure compatibility with most hardware environments.

All source code, including preprocessing scripts, and evaluation metrics, is publicly available on GitHub[1].

2.2 Results

The main retrieval results on the RAG-Mini-BioASQ dataset (evaluated over the first 500 questions) are presented in Table 1. We report both the mean and median of the normalized Discounted Cumulative Gain (nDCG) to provide a robust assessment of retrieval performance across varying question difficulties.

Due to the computational cost associated with the generative component of the HyDE method — requiring approximately one hour to process 500 queries — we limited our evaluation to this subset for consistency across methods. In contrast, both BM25 and Encoder Only methods retrieved results for the full 4,719-question set in under a minute. Nevertheless, we found that this representative subset of 500 questions provided a sufficiently diverse and challenging benchmark to assess comparative performance.

Table 1: Multi-retrieval performance comparison on RAG-Mini-BioASQ. Higher nDCG values indicate better alignment with the ground-truth passage set.

Method	Mean nDCG	Median nDCG
BM25	0.561	0.595
Encoder Only	**0.617**	**0.647**
HyDE	0.570	0.606

The results show that the **Encoder Only** approach outperforms both BM25 and HyDE in terms of both mean and median nDCG. Although HyDE demonstrates competitive performance, its gains are less pronounced in this biomedical multi-retrieval setting, potentially due to the domain-specific nature of the passages and the formulation of questions, which may limit the effectiveness of hypothetical document generation across different semantically large sets of relevant passages. Interestingly, the HyDE method shows the highest number of 1-valued nDCG scores, indicating that while it may not consistently outperform other methods on average, it occasionally achieves perfect retrieval on certain queries.

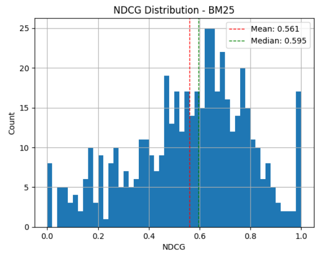
To further illustrate the distribution and spread of performance across individual queries, we include the per-query nDCG distributions for each method in Figure 1.

2.3 Discussion and Future Directions.

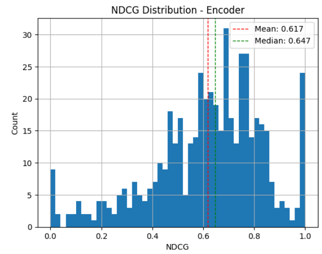
While HyDE was originally proposed as a promising retrieval enhancement via hypothetical document generation, our results suggest that its advantage does not necessarily extend to multi-retrieval contexts such as RAG-Mini-BioASQ. In this biomedical setting—characterized by multiple relevant passages per query—HyDE's single-hypothesis generation may inadequately capture the semantic diversity needed to retrieve a full set of relevant documents. This limitation is further compounded by its high computational overhead, which poses scalability concerns for large retrieval sets.

To improve HyDE's efficiency and retrieval coverage in such scenarios, several extensions could be explored. First, generating multiple hypothetical answers per query could better span the space of

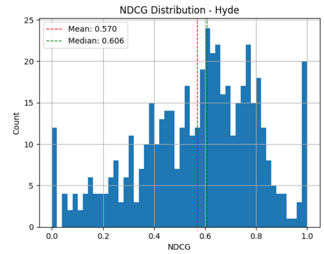
[1] `https://github.com/Matt-Olek/ML4NLP-Class-Project`



(a) nDCG distribution for BM25



(b) nDCG distribution for Encoder Only



(c) nDCG distribution for HyDE

Figure 1: nDCG distribution comparisons across BM25, Encoder Only, and HyDE methods.

relevant contexts, improving recall across semantically dispersed relevant passages. Second, filtering or diversifying generations using prompt engineering or contrastive reranking could make the output more effective for dense retrieval. Finally, a hybrid pipeline combining Encoder Only retrieval with a lightweight generative refinement stage may offer a practical balance between effectiveness and efficiency.

As part of ongoing work toward production-ready Retrieval-Augmented Generation (RAG) systems, we are actively exploring multi-hypothesis HyDE strategies coupled with dense retrieval, LLM-based reranking, and hybrid scoring. Preliminary results indicate that this enriched pipeline — blending efficient encoding with the semantic depth of generation and reranking — can offer a powerful and scalable solution for complex information retrieval tasks.