Novelty Assessment Report

Paper: Revela: Dense Retriever Learning via Language Modeling

PDF URL: https://openreview.net/pdf?id=e7pAjJZJWb

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-04

Abstract

Dense retrievers play a vital role in accessing external and specialized knowledge to augment language models (LMs). Training dense retrievers typically requires annotated query-document pairs, which are costly to create and scarce in specialized domains (e.g., code) or in complex settings (e.g., requiring reasoning). These practical challenges have sparked growing interest in self-supervised retriever learning. Since LMs are trained to capture token-level dependencies through a self-supervised learning objective (i.e., next token prediction), we can analogously cast retrieval as learning dependencies among chunks of tokens. This analogy naturally leads to the question: How can we adapt self-supervised learning objectives in the spirit of language modeling to train retrievers? .

To answer this question, we introduce Revela, a unified and scalable training framework for self-supervised retriever learning via language modeling. Revela models semantic dependencies among documents by conditioning next token prediction on local and cross-document context through an in-batch attention mechanism. This attention is weighted by retriever-computed similarity scores, enabling the retriever to be optimized as part of language modeling. We evaluate Revela on domain-specific (CoIR), reasoning-intensive (BRIGHT), and general-domain (BEIR) benchmarks across various retriever backbones. Without annotated or synthetic query-document pairs, Revela surpasses larger supervised models and proprietary APIs on CoIR and matches them on BRIGHT. It achieves BEIR's unsupervised SoTA with ~ 1000x less training data and 10x less compute. Performance increases with batch size and model size, highlighting Revela's scalability and its promise for self-supervised retriever learning.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Self-Supervised Dense Retriever Learning via Language Modeling**

A total of **50 papers** were analyzed and organized into a taxonomy with **26 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Self-Supervised Pre-training Architectures**
- **Contrastive and Self-Supervised Representation Learning**
- **Synthetic Data Generation and Query Augmentation**
- **Large Language Model Adaptation for Dense Retrieval**
- **Domain Adaptation and Transfer Learning**
- **Hybrid and Enhanced Retrieval Strategies**
- **Specialized Retrieval Applications and Objectives**
- **Self-Supervised Dense Retriever Learning via Language Modeling**
- **Surveys, Frameworks, and Comparative Studies**
- **Non-Retrieval Applications**

Complete Taxonomy Tree

- Self-Supervised Dense Retriever Learning via Language Modeling Survey Taxonomy
- Self-Supervised Pre-training Architectures
    - Masked Auto-Encoder Based Pre-training (4 papers)
    - [7] RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder (Shitao Xiao, 2022) View paper
    - [25] Challenging Decoder helps in Masked Auto-Encoder Pre-training for Dense Passage Retrieval (Li, 2023) View paper
    - [43] ConTextual Masked Auto-Encoder for Dense Passage Retrieval (Xing Wu, 2023) View paper
    - [44] ConTextual Mask AutoEncoder for Dense Passage Retrieval (Wu Xing, 2023) View paper
    - Condenser and Corpus-Aware Pre-training (2 papers)
    - [6] Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval (Gao, 2021) View paper
    - [35] Long: Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval (Callan, 2022) View paper
    - Bag-of-Word and Simplified Decoder Approaches (1 papers)
    - [37] Drop your Decoder: Pre-training with Bag-of-Word Prediction for Dense Passage Retrieval. (Guangyuan Ma, 2024) View paper
    - Structure-Aware and Multimodal Pre-training (3 papers)
    - [5] SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features (Tschannen, 2025) View paper
    - [22] Structure-aware language model pretraining improves dense retrieval on structured data (Gu Yu, 2023) View paper
    - [26] Conditioned Masked Language and Image Modeling for Image-Text Dense Retrieval (Ziyang Luo, 2022) View paper
- Contrastive and Self-Supervised Representation Learning
    - Iterative Contrastive Learning (2 papers)
    - [10] Laprador: Unsupervised pretrained dense retriever for zero-shot text retrieval (Canwen Xu, 2022) View paper

- ◦ [39] Unsupervised Context Aware Sentence Representation Pretraining for Multi-lingual Dense Retrieval (Wu Ning, 2022) View paper
- ◦ Counterfactual and Robustness-Oriented Contrastive Learning (1 papers)
- ◦ [4] Unsupervised dense retrieval with conterfactual contrastive learning (Chen Hai-tian, 2024) View paper
- ◦ Self-Supervised Fine-Tuning with Masked Language Modeling (2 papers)
- ◦ [11] Self-supervised fine-tuning for efficient passage re-ranking (Meoungjun Kim, 2021) View paper
- ◦ [12] Self-supervised Contrastive BERT Fine-tuning for Fusion-Based Reviewed-Item Retrieval (Pour, 2023) View paper
- ◦ Discriminative Representation via Self-Teaching (1 papers)
- ◦ [49] Discriminative Language Model via Self-Teaching for Dense Retrieval (Lu Chen, 2022) View paper

- Synthetic Data Generation and Query Augmentation
  - ◦ Autoencoding with Synthetic Query Generation (2 papers)
  - ◦ [3] Designing accurate retrieval systems using language models (Sachan, 2024) View paper
  - ◦ [18] Questions Are All You Need to Train a Dense Passage Retriever (Devendra Singh Sachan, 2022) View paper
  - ◦ Hypothetical Document Generation for Zero-Shot Retrieval (1 papers)
  - ◦ [2] Precise zero-shot dense retrieval without relevance labels (Callan, 2023) View paper
  - ◦ Noisy Self-Training with Synthetic Queries (1 papers)
  - ◦ [20] Noisy Self-Training with Synthetic Queries for Dense Retrieval (Fan Jiang, 2023) View paper
  - ◦ Instruction-Tuning and Unsupervised Query Augmentation (2 papers)
  - ◦ [13] ReGen: Zero-Shot Text Classification via Training Data Generation with Progressive Dense Retrieval (Meng Yu, 2023) View paper
  - ◦ [14] Unsupervised Text Representation Learning via Instruction-Tuning for Zero-Shot Dense Retrieval (Zimeng Qiu, 2024) View paper
- Large Language Model Adaptation for Dense Retrieval
  - ◦ Unsupervised LLM-to-Retriever Adaptation (2 papers)
  - ◦ [17] Llama2Vec: Unsupervised Adaptation of Large Language Models for Dense Retrieval (Chaofan Li, 2023) View paper
  - ◦ [46] Making Large Language Models A Better Foundation For Dense Retrieval (Li Chaofan, 2023) View paper
  - ◦ LLM-Based Document Expansion for Pre-training (1 papers)
  - ◦ [42] Pre-training with Large Language Model-based Document Expansion for Dense Passage Retrieval (Ma Guangyuan, 2023) View paper
  - ◦ End-to-End LLM-Driven Retrieval Systems (1 papers)
  - ◦ [8] Self-retrieval: End-to-end information retrieval with one large language model (Jiawei Chen, 2024) View paper
- Domain Adaptation and Transfer Learning
  - ◦ Pseudo-Relevance Labeling for Domain Adaptation (2 papers)
  - ◦ [29] Domain Adaptation for Dense Retrieval through Self-Supervision by Pseudo-Relevance Labeling (Li, 2022) View paper
  - ◦ [36] Domain Adaptation for Dense Retrieval and Conversational Dense Retrieval through Self-Supervision by Meticulous Pseudo-Relevance Labeling (Li, 2024) View paper
  - ◦ Multilingual and Cross-Lingual Adaptation (2 papers)
  - ◦ [19] Knowledge Enhanced Pre-training for Cross-lingual Dense Retrieval (H Zhang, 2024) View paper
  - ◦ [23] Lexicon-enhanced self-supervised training for multilingual dense retrieval (Houxing Ren, 2022) View paper
- Hybrid and Enhanced Retrieval Strategies
  - ◦ Retrieval-Augmented Pre-training (1 papers)
  - ◦ [1] Retrieval augmented language model pre-training (Kelvin Guu, 2020) View paper
  - ◦ Aggregation and Fusion for Retrieval (2 papers)
  - ◦ [38] Enhancing Fusion Techniques and Language Model Fine-tuning for Natural Language-based Reviewed-Item Retrieval (Farinneya, 2023) View paper
  - ◦ [48] Aggretriever: A Simple Approach to Aggregate Textual Representations for Robust Dense Passage Retrieval (Sheng-Chieh Lin, 2022) View paper
  - ◦ Causal Retrieval and Long-Range Modeling (1 papers)
  - ◦ [40] Efficient Long-range Language Modeling with Self-supervised Causal Retrieval (X Hu, 2024) View paper
- Specialized Retrieval Applications and Objectives
  - ◦ Retrieval-Oriented Masking Strategies (1 papers)
  - ◦ [50] Retrieval Oriented Masking Pre-training Language Model for Dense Passage Retrieval (Long, 2022) View paper
  - ◦ Domain-Specific Retrieval Systems (3 papers)
  - ◦ [9] Iterative self-supervised learning for legal similar case retrieval (Yao Liu, 2024) View paper
  - ◦ [15] Easyrec: Simple yet effective language models for recommendation (Xubin Ren, 2025) View paper
  - ◦ [27] Improving first-stage retrieval of point-of-interest search by pre-training models (Lang Mei, 2023) View paper
  - ◦ Reasoning-Aware and Interactive Retrieval (2 papers)
  - ◦ [28] RaDeR: Reasoning-aware Dense Retrieval Models (Debrup Das, 2025) View paper
  - ◦ [31] Boosting search engines with interactive agents (Adolphs, 2021) View paper
- Self-Supervised Dense Retriever Learning via Language Modeling ★ (1 papers)
  - ◦ [0] Revela: Dense Retriever Learning via Language Modeling (Anon et al., 2026) View paper
- Surveys, Frameworks, and Comparative Studies (7 papers)
  - ◦ [16] Self-supervised information retrieval trained from self-generated sets of queries and relevant documents (G. Moro, 2022) View paper
  - ◦ [24] A Self-supervised Learning Algorithm for Unsupervised Information Retrieval in Big Data Corpora (Virendra Tank, 2025) View paper
  - ◦ [30] Advancements in Self-Supervised Learning: A Review of Unsupervised Information Retrieval Techniques for Big Data (Virendra Tank, 2026) View paper
  - ◦ [32] Self-supervised information retrieval: a novel approach based on Deep Metric Learning and Neural Language Models (Rossi, 2021)
  - ◦ [33] Strategies for Efficient Text Clustering and Retrieval based on Language Models (è¨□èª□ã□¢ã□□ã□«ã□«å□ºã□¥ã□□å□¹ç□□ç□□ã□ªã□□ã□ã□¹ã□□ã□¯ã□©ã□¹ã□¿ã□ªã□³ã□°ã□¨æ¤□ç´¢, 2024) View paper

  - [34] Novel language models and methods for semantic representation learning, self-supervised retrieval, and summarization (Valgimigli, 2024) View paper
  - [41] Adapter des modÃ¨les de recherche d'information basÃ©s sur les rÃ©seaux neuronauxprofonds pour les documents longs et les nouveaux domaines (Li, 2023) View paper
- Non-Retrieval Applications (3 papers)
  - [21] A study of situational reasoning for traffic understanding (Zhang, 2023) View paper
  - [45] Embedding Earth: Self-supervised contrastive pre-training for dense land cover classification (Tarasiou, 2022) View paper
  - [47] High-Order Correlation-Guided Slide-Level Histology Retrieval With Self-Supervised Hashing. (Shengrui Li, 2023) View paper

Narrative

Core task: self-supervised dense retriever learning via language modeling. This field centers on training neural retrievers without manually labeled query-document pairs by leveraging language model objectives and self-generated signals. The taxonomy reveals a rich landscape organized around several complementary directions. Self-Supervised Pre-training Architectures explore masked language modeling variants and encoder-decoder designs (e.g., RetroMAE[7], Corpus Aware Pretraining[6]) that shape retrieval-oriented representations. Contrastive and Self-Supervised Representation Learning emphasizes contrastive objectives and metric learning frameworks (e.g., Contrastive BERT[12], Deep Metric Learning[32]) to pull relevant items closer in embedding space. Synthetic Data Generation and Query Augmentation tackle the scarcity of labeled data by creating pseudo-queries or augmented examples (e.g., Self-Generated Queries[16], Questions Are All[18]). Large Language Model Adaptation branches adapt decoder-only or instruction-tuned LLMs for dense retrieval (e.g., Llama2Vec[17], Instruction-Tuning[14]), while Domain Adaptation and Transfer Learning address cross-lingual and specialized corpus challenges (e.g., Cross-lingual Dense[19], Legal Case Retrieval[9]). Hybrid and Enhanced Retrieval Strategies combine lexical and neural signals (e.g., Fusion Techniques[38]), and Specialized Retrieval Applications target domains like geospatial data or histology (e.g., Embedding Earth[45], Slide-Level Histology[47]).

A particularly active line of work focuses on designing effective self-supervised objectives that align language model pre-training with retrieval goals, balancing reconstruction fidelity and discriminative power. Some studies emphasize retrieval-oriented masking strategies (e.g., Retrieval Oriented Masking[50], ConTextual MAE[43]) to guide the model toward salient query-document features, while others explore pseudo-relevance feedback loops (e.g., Pseudo-Relevance Labeling[29], Meticulous Pseudo-Relevance[36]) to iteratively refine retrieval quality. Revela[0] sits within the core Self-Supervised Dense Retriever Learning via Language Modeling branch, directly addressing how language modeling can bootstrap retrieval without external supervision. Its emphasis on leveraging LM signals aligns it closely with works like REALM[1] and Designing Accurate Retrieval[3], which also integrate language model objectives into retrieval pipelines, though Revela[0] appears to focus more explicitly on self-supervised mechanisms rather than hybrid or instruction-based adaptations seen in neighboring efforts like Instruction-Tuning[14] or LLM Dense Foundation[46].

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on a specific technical approach: using language modeling objectives (next token prediction conditioned on cross-document context) to train dense retrievers in a self-supervised manner. The sibling subtopics represent orthogonal categorizations: one covers non-text domains (land cover, histology images) where self-supervised learning is applied to retrieval, while the other encompasses meta-level contributions (surveys, frameworks, comparative studies) rather than novel methods. The original leaf is narrowly scoped to a particular training paradigm for text retrieval, while siblings either change the domain or the contribution type.

Similarities: - All subtopics fall under the broader umbrella of self-supervised learning applied to retrieval tasks - All involve learning representations or retrieval mechanisms without explicit human annotations - The original leaf and Non-Retrieval Applications both focus on specific technical methods (though in different domains)

Differences: - The original leaf is specific to text-based dense retrieval using language modeling objectives, while Non-Retrieval Applications covers image/spatial domains - The original leaf proposes a novel training method, while Surveys/Frameworks/Comparative Studies provide meta-analyses rather than new techniques - The original leaf emphasizes cross-document semantic dependencies through language modeling, while siblings either lack this text-specific mechanism (Non-Retrieval) or aggregate multiple approaches (Surveys) - Non-Retrieval Applications explicitly excludes text-based methods, creating a clear domain boundary with the original leaf

Suggested Search Directions: - Investigate whether other language modeling objectives (e.g., masked language modeling, contrastive language-image pretraining adaptations) are used for dense retriever training - Explore if there are hybrid methods combining language modeling with other self-supervised signals (contrastive learning, clustering) for retrieval - Check for papers on self-supervised dense retrieval that use different conditioning mechanisms beyond cross-document context weighting

Sibling Subtopics

- **Non-Retrieval Applications** (leaves: 1, papers: 3)
- Scope: Papers applying self-supervised learning to non-retrieval tasks such as land cover classification or histology image retrieval.
- Exclude: Excludes text-based dense retrieval methods; see other categories.
- **Surveys, Frameworks, and Comparative Studies** (leaves: 1, papers: 7)
- Scope: Survey papers, comprehensive frameworks, or studies comparing multiple self-supervised retrieval techniques across benchmarks.
- Exclude: Excludes papers proposing single novel methods; see other categories based on technique.

Contributions Analysis

Overall novelty summary. ```json { "paragraphs": ["The paper introduces Revela, a framework that trains dense retrievers through language modeling objectives by conditioning next token prediction on cross-document context weighted by retriever similarity scores. According to the taxonomy tree, this work occupies a singleton leaf node titled 'Self-Supervised Dense Retriever Learning via Language Modeling' with no sibling papers, suggesting it addresses a relatively sparse and specialized research direction. The taxonomy contains 50 papers across 36 topic areas, yet this particular leaf stands alone, indicating that explicitly framing retriever training as chunk-level language modeling with in-batch attention mechanisms represents a less crowded conceptual space within the broader self-supervised retrieval literature.",

"The taxonomy reveals that Revela's closest neighbors lie in adjacent branches: Self-Supervised Pre-training Architectures explores masked auto-encoder designs like RetroMAE and Condenser models that compress semantics into dense vectors, while Contrastive and Self-Supervised Representation Learning emphasizes metric learning frameworks without architectural novelty. Synthetic Data Generation branches generate pseudo-queries or hypothetical documents to enable training, and Large Language Model Adaptation adapts autoregressive LLMs into dual encoders. Revela diverges from these directions by avoiding synthetic query generation, eschewing explicit contrastive losses, and focusing instead on modeling semantic dependencies through language modeling objectives applied directly to document chunks, positioning it at a conceptual intersection that the taxonomy captures as a distinct leaf.", "Among 26 candidates examined across three contributions, the literature search found limited direct overlap. The core Revela framework contribution examined 10 candidates with zero refutable matches, suggesting novelty within the examined scope. The in-batch attention mechanism contribution reviewed 6 candidates, again with no refutations. However, the performance claim contribution examined 10 candidates and identified 1 refutable match, indicating that at least one prior work within this limited sample demonstrates comparable effectiveness without query-document pairs. The search scope—26 papers from semantic search and citation expansion—provides a snapshot rather than exhaustive coverage, meaning these statistics reflect top-K similarity matches rather than comprehensive field-wide analysis.",

"Based on the limited search scope of 26 candidates, Revela appears to occupy a relatively novel position by explicitly casting retrieval as chunk-level language modeling with similarity-weighted cross-document attention. The singleton taxonomy leaf and low refutation rates across most contributions suggest conceptual distinctiveness, though the performance contribution shows at least one overlapping prior result. The analysis does not cover the full breadth of self-supervised retrieval literature, particularly work published after the search cutoff or in non-indexed venues, leaving open questions about broader field coverage."] } ```

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Revela framework for self-supervised retriever learning via language modeling

Description: Revela is a novel framework that trains dense retrievers through language modeling by conditioning next token prediction on both local context and cross-document context via an in-batch attention mechanism. The retriever is optimized jointly with the language model without requiring annotated or synthetic query-document pairs.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction

URL: View paper

Brief Assessment

Contrastive Span[55] focuses on contrastive span prediction for pre-training encoders in dense retrieval, using autoencoder-based approaches with group-wise contrastive learning. This differs fundamentally from Revela's approach of training retrievers through next token prediction conditioned on cross-document context via in-batch attention mechanisms.

2. Structure-aware language model pretraining improves dense retrieval on structured data

URL: View paper

Brief Assessment

Structure-aware Pretraining[22] focuses on retrieval for structured data (code, products) using alignment between structured/ unstructured data pairs and masked entity prediction. Revela addresses general dense retrieval via in-batch attention and next-token prediction without requiring paired data.

3. Condenser: a pre-training architecture for dense retrieval

URL: View paper

Brief Assessment

Condenser[51] focuses on pre-training architecture modifications (adding a condenser head with skip connections) to improve dense encoder readiness, rather than self-supervised retriever training via language modeling objectives. The candidate does not demonstrate prior work on conditioning next token prediction on cross-document context through in-batch attention mechanisms.

4. Unleashing the Power of LLMs in Dense Retrieval with Query Likelihood Modeling

URL: View paper

Brief Assessment

Query Likelihood[54] focuses on adapting LLMs for dense retrieval through query likelihood maximization as an auxiliary task before contrastive learning, rather than proposing a self-supervised framework that jointly trains retrievers and language models without query-document pairs through in-batch attention mechanisms.

5. Unsupervised dense retrieval with relevance-aware contrastive pre-training

URL: View paper

Brief Assessment

Relevance-aware Contrastive[53] focuses on contrastive pre-training with pseudo-positive pairs from data augmentation, not on language modeling objectives for retriever training. The candidate uses contrastive loss with relevance weighting, while the original conditions next token prediction on cross-document context via in-batch attention.

6. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval

URL: View paper

Brief Assessment

Corpus Aware Pretraining[6] focuses on corpus-level contrastive pre-training using span pairs from documents, not on conditioning next token prediction on cross-document context via in-batch attention as in Revela. The technical approaches differ fundamentally in their training mechanisms.

7. Lexicon-enhanced self-supervised training for multilingual dense retrieval

URL: View paper

Brief Assessment

Lexicon-enhanced[23] focuses on multilingual dense retrieval using sparse-dense hybrid mining and query generation, not on self-supervised retriever training via language modeling objectives. The candidate does not address conditioning next token prediction on cross-document context through in-batch attention mechanisms.

8. Unsupervised dense information retrieval with contrastive learning

URL: View paper

Brief Assessment

Contrastive Learning[52] uses contrastive learning with data augmentation (cropping, inverse cloze task) to train retrievers, not language modeling with in-batch attention as in Revela.

9. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features

URL: View paper

Brief Assessment

SigLIP 2[5] focuses on vision-language encoders for image-text tasks using contrastive learning and self-supervised techniques. It does not address dense retrieval or language modeling objectives for text-based retrieval systems.

10. Precise zero-shot dense retrieval without relevance labels

URL: View paper

Brief Assessment

Precise Zero-Shot[2] focuses on zero-shot dense retrieval using hypothetical document generation via instruction-following LMs (e.g., InstructGPT) combined with unsupervised contrastive encoders. In contrast, Revela trains retrievers jointly with language models through next-token prediction conditioned on in-batch attention, without requiring query-document pairs or hypothetical document generation. The technical approaches and objectives differ fundamentally.

Contribution 2: In-batch attention mechanism weighted by retriever similarity scores

Description: The framework introduces an in-batch attention mechanism where next-token prediction is conditioned on both the input sequence and other sequences within the same batch. The attention weights are determined by retriever-computed similarity scores, enabling joint optimization of the retriever and language model.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Current Limitations of Language Models: What You Need is Retrieval

URL: View paper

Brief Assessment

Retrieval Limitations[67] discusses retrieval-augmented approaches broadly but does not describe an in-batch attention mechanism weighted by retriever similarity scores for joint training. The candidate focuses on general retrieval strategies rather than the specific architectural innovation of in-batch attention.

2. Retrieve anything to augment large language models

URL: View paper

Brief Assessment

Retrieve Anything[64] focuses on training embedding models for diverse retrieval tasks (knowledge, memory, tools, examples) but does not employ an in-batch attention mechanism for joint retriever-LM training. The candidate uses contrastive learning and knowledge distillation for retriever training, not attention-based conditioning during next-token prediction.

3. Supporting Retriever's Training by Joint Likelihood-Based Soft-Label Generation

URL: View paper

Brief Assessment

Soft-Label Generation[68] focuses on contrastive learning with in-batch negatives for retriever training, not on an in-batch attention mechanism for joint language modeling. The candidate does not describe conditioning next-token prediction on other sequences via attention weighted by retriever scores.

4. A Survey on Efficient Protein Language Models

URL: View paper

Brief Assessment

Protein Language Models[66] is a survey paper on efficient protein language models. It does not describe in-batch attention mechanisms for joint retriever-language model training or any retrieval framework.

5. Retrieval augmented language model pre-training

URL: View paper

Brief Assessment

REALM[1] uses a different architecture where retrieval is performed before prediction, with documents retrieved via MIPS and then fed to a knowledge-augmented encoder. The candidate does not employ an in-batch attention mechanism where tokens attend to other sequences within the same batch during training, as described in the original paper.

6. Retrieval-Native Language Models: Integrating Parametric and Vector Memory with Bayesian Attention

URL: View paper

Brief Assessment

Retrieval-Native[65] focuses on integrating parametric and vector memory with Bayesian attention for retrieval-augmented generation. The provided context is too limited to assess whether it implements in-batch attention weighted by retriever similarity scores for joint optimization.

Contribution 3: Superior performance without query-document pairs across multiple benchmarks

Description: Revela achieves state-of-the-art results on domain-specific (CoIR), reasoning-intensive (BRIGHT), and general-domain (BEIR) benchmarks without using annotated or synthetic query-document pairs. It outperforms larger supervised models and proprietary APIs while using significantly less training data and compute resources.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. From matching to generation: A survey on generative information retrieval

URL: View paper

Brief Assessment

Generative IR Survey[56] is a survey paper that reviews existing generative information retrieval methods. It does not present a novel retrieval training method that could refute Revela's claims about unsupervised training performance.

2. Documentâ□□toâ□□Document Retrieval Using Selfâ□□Retrieval Learning and Automatic Keyword Extraction

URL: View paper

Brief Assessment

Document-to-Document[62] focuses on legal document retrieval using keyword extraction for self-supervised learning, while the original paper addresses general dense retrieval across diverse domains (code, reasoning-intensive, general) using language modeling with in-batch attention. The technical approaches and evaluation contexts differ substantially.

---

3. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval

URL: View paper

Brief Assessment

GPL[57] focuses on unsupervised domain adaptation using synthetic query generation and pseudo-labeling, not on general self-supervised retriever learning via language modeling as in the original paper. GPL requires a pre-trained query generator and cross-encoder from MS MARCO, whereas the original paper trains retrievers directly from raw text without any query-document pairs or pre-existing models trained on such pairs.

---

4. Laprador: Unsupervised pretrained dense retriever for zero-shot text retrieval

URL: View paper

Brief Assessment

Laprador[10] focuses on unsupervised pretraining for zero-shot retrieval using contrastive learning with ICT and DAPI objectives, not on language modeling-based retriever training. The technical approach differs fundamentally from Revela's in-batch attention mechanism conditioned on language modeling objectives.

---

5. Unsupervised learning of semantic audio representations

URL: View paper

Brief Assessment

Semantic Audio[61] focuses on unsupervised learning of audio representations using triplet loss, not text retrieval. The domains (audio vs. text/code retrieval) and methods (triplet loss for audio vs. language modeling for documents) are fundamentally different.

---

6. Inpars: Unsupervised dataset generation for information retrieval

URL: View paper

Brief Assessment

Inpars[58] uses synthetic query-document pairs generated by language models for training, which is fundamentally different from Revela's approach of training without any query-document pairs (annotated or synthetic). The candidate explicitly generates synthetic datasets for fine-tuning, while the original paper trains directly on raw text via language modeling.

---

7. Transformer-based Clipped Contrastive Quantization Learning for Unsupervised Image Retrieval

URL: View paper

Brief Assessment

Clipped Contrastive[63] addresses unsupervised image retrieval using visual features and contrastive learning, while the original paper focuses on text/code retrieval using language modeling without query-document pairs. These are fundamentally different domains and methodologies.

---

8. Unsupervised Dense Retrieval Training with Web Anchors

URL: View paper

Brief Assessment

Web Anchors[60] focuses on using web anchor-document pairs for unsupervised training, which are still structured pairs (albeit naturally occurring). The original paper claims novelty in training without ANY query-document pairs using language modeling objectives, which is a fundamentally different approach from contrastive learning on anchor-document pairs.

---

9. Unsupervised dense information retrieval with contrastive learning

URL: View paper

Prior Art Analysis

Contrastive Learning[52] demonstrates that unsupervised dense retrievers trained with contrastive learning can achieve competitive performance with BM25 and outperform it on multiple benchmarks without using query-document pairs. The paper reports that their unsupervised model outperforms BM25 on 11 out of 15 datasets on BEIR for recall@100, and achieves strong results on domain-specific and general benchmarks. This establishes prior work showing unsupervised retrievers without query-document pairs can match or exceed supervised baselines.

Evidence

Evidence 1 - **Rationale**: Both demonstrate unsupervised models without query-document pairs achieving competitive or superior performance to baselines on multiple benchmarks. - **Original**: without annotated or synthetic query-document pairs, revela surpasses larger supervised models and proprietary apis on coir and matches them on bright. it achieves beir's unsupervised sota with ˜ 1000x less training data and 10x less compute. - **Candidate**: we observe that in this setting, contriever is competitive compared to bm25 on all datasets, but trec-covid and tóuche-2020. in particular, it obtains better performance than bm25 on 11 out of 15 datasets from the benchmark for the recall@100. contriever also outperforms previously proposed unsuperv...

Evidence 2 - **Rationale**: Both papers claim their unsupervised approach achieves state-of-the-art or competitive results without query-document pairs, showing prior demonstration of this capability. - **Original**: without query-document pairs, revela surpasses e5-mistral-7b-instruct on coir by 2.8% and outperforms unsupervised baselines by 9.7%, while also matching the performance of proprietary apis on bright - **Candidate**: when used as pre-training before fine-tuning on ms marco, our technique leads to strong results, in particular for recall@100. based on that observation, we use a cross-encoder to re-rank documents retrieved with our model, leading to new state-of-the-art on the competitive beir benchmark.

---

10. SCENIR: Visual Semantic Clarity through Unsupervised Scene Graph Retrieval

URL: View paper

Brief Assessment

SCENIR[59] addresses image-to-image retrieval using scene graphs in an unsupervised manner, while the original paper focuses on text-based dense retrieval for information retrieval tasks. These are fundamentally different domains (visual vs. textual retrieval) with distinct technical approaches.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Revela: Dense Retriever Learning via Language Modeling View paper
- [1] Retrieval augmented language model pre-training View paper
- [2] Precise zero-shot dense retrieval without relevance labels View paper
- [3] Designing accurate retrieval systems using language models View paper
- [4] Unsupervised dense retrieval with conterfactual contrastive learning View paper
- [5] SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features View paper
- [6] Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval View paper
- [7] RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder View paper
- [8] Self-retrieval: End-to-end information retrieval with one large language model View paper
- [9] Iterative self-supervised learning for legal similar case retrieval View paper
- [10] Laprador: Unsupervised pretrained dense retriever for zero-shot text retrieval View paper
- [11] Self-supervised fine-tuning for efficient passage re-ranking View paper
- [12] Self-supervised Contrastive BERT Fine-tuning for Fusion-Based Reviewed-Item Retrieval View paper
- [13] ReGen: Zero-Shot Text Classification via Training Data Generation with Progressive Dense Retrieval View paper
- [14] Unsupervised Text Representation Learning via Instruction-Tuning for Zero-Shot Dense Retrieval View paper
- [15] Easyrec: Simple yet effective language models for recommendation View paper
- [16] Self-supervised information retrieval trained from self-generated sets of queries and relevant documents View paper
- [17] Llama2Vec: Unsupervised Adaptation of Large Language Models for Dense Retrieval View paper
- [18] Questions Are All You Need to Train a Dense Passage Retriever View paper
- [19] Knowledge Enhanced Pre-training for Cross-lingual Dense Retrieval View paper
- [20] Noisy Self-Training with Synthetic Queries for Dense Retrieval View paper
- [21] A study of situational reasoning for traffic understanding View paper
- [22] Structure-aware language model pretraining improves dense retrieval on structured data View paper
- [23] Lexicon-enhanced self-supervised training for multilingual dense retrieval View paper
- [24] A Self-supervised Learning Algorithm for Unsupervised Information Retrieval in Big Data Corpora View paper
- [25] Challenging Decoder helps in Masked Auto-Encoder Pre-training for Dense Passage Retrieval View paper
- [26] Conditioned Masked Language and Image Modeling for Image-Text Dense Retrieval View paper
- [27] Improving first-stage retrieval of point-of-interest search by pre-training models View paper
- [28] RaDeR: Reasoning-aware Dense Retrieval Models View paper
- [29] Domain Adaptation for Dense Retrieval through Self-Supervision by Pseudo-Relevance Labeling View paper
- [30] Advancements in Self-Supervised Learning: A Review of Unsupervised Information Retrieval Techniques for Big Data View paper
- [31] Boosting search engines with interactive agents View paper
- [32] Self-supervised information retrieval: a novel approach based on Deep Metric Learning and Neural Language Models
- [33] Strategies for Efficient Text Clustering and Retrieval based on Language Models View paper
- [34] Novel language models and methods for semantic representation learning, self-supervised retrieval, and summarization View paper
- [35] Long: Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval View paper
- [36] Domain Adaptation for Dense Retrieval and Conversational Dense Retrieval through Self-Supervision by Meticulous Pseudo-Relevance Labeling View paper
- [37] Drop your Decoder: Pre-training with Bag-of-Word Prediction for Dense Passage Retrieval. View paper
- [38] Enhancing Fusion Techniques and Language Model Fine-tuning for Natural Language-based Reviewed-Item Retrieval View paper
- [39] Unsupervised Context Aware Sentence Representation Pretraining for Multi-lingual Dense Retrieval View paper
- [40] Efficient Long-range Language Modeling with Self-supervised Causal Retrieval View paper
- [41] Adapter des modÃ¨les de recherche d'information basÃ©s sur les rÃ©seaux neuronauxprofonds pour les documents longs et les nouveaux domaines View paper
- [42] Pre-training with Large Language Model-based Document Expansion for Dense Passage Retrieval View paper
- [43] ConTextual Masked Auto-Encoder for Dense Passage Retrieval View paper
- [44] ConTextual Mask AutoEncoder for Dense Passage Retrieval View paper
- [45] Embedding Earth: Self-supervised contrastive pre-training for dense land cover classification View paper
- [46] Making Large Language Models A Better Foundation For Dense Retrieval View paper
- [47] High-Order Correlation-Guided Slide-Level Histology Retrieval With Self-Supervised Hashing. View paper
- [48] Aggretriever: A Simple Approach to Aggregate Textual Representations for Robust Dense Passage Retrieval View paper
- [49] Discriminative Language Model via Self-Teaching for Dense Retrieval View paper
- [50] Retrieval Oriented Masking Pre-training Language Model for Dense Passage Retrieval View paper
- [51] Condenser: a pre-training architecture for dense retrieval View paper
- [52] Unsupervised dense information retrieval with contrastive learning View paper
- [53] Unsupervised dense retrieval with relevance-aware contrastive pre-training View paper
- [54] Unleashing the Power of LLMs in Dense Retrieval with Query Likelihood Modeling View paper
- [55] Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction View paper
- [56] From matching to generation: A survey on generative information retrieval View paper
- [57] GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval View paper
- [58] Inpars: Unsupervised dataset generation for information retrieval View paper
- [59] SCENIR: Visual Semantic Clarity through Unsupervised Scene Graph Retrieval View paper
- [60] Unsupervised Dense Retrieval Training with Web Anchors View paper
- [61] Unsupervised learning of semantic audio representations View paper
- [62] Documentâ□□toâ□□Document Retrieval Using Selfâ□□Retrieval Learning and Automatic Keyword Extraction View paper
- [63] Transformer-based Clipped Contrastive Quantization Learning for Unsupervised Image Retrieval View paper

- [64] Retrieve anything to augment large language models View paper
- [65] Retrieval-Native Language Models: Integrating Parametric and Vector Memory with Bayesian Attention View paper
- [66] A Survey on Efficient Protein Language Models View paper
- [67] Current Limitations of Language Models: What You Need is Retrieval View paper
- [68] Supporting Retriever's Training by Joint Likelihood-Based Soft-Label Generation View paper