DeReC: Scalable Fact Verification via Dense Retrieval Classification

Maxim Konovalov

January 2026

Abstract

This report presents the implementation and evaluation of DeReC (Dense Retrieval Classification), a framework designed for efficient fact verification and fake news detection. Unlike computationally expensive Large Language Model (LLM) approaches that generate explanations, DeReC leverages dense embeddings and vector similarity search to ground claims in factual evidence directly. We demonstrate that this retrieval-centric approach achieves competitive accuracy while significantly reducing computational overhead. The project code is available at: `https://github.com/BigMak1/rag_fact_checking`.

1 Introduction

The proliferation of misinformation and fake news on the internet necessitates robust and scalable fact-verification systems. While recent advancements in Large Language Models (LLMs) have enabled automated fact-checking, these models often suffer from "hallucinations"—generating confident but factually incorrect rationales. Furthermore, the inference cost and latency of autoregressive LLMs make them impractical for real-time, large-scale deployment.

This section is devoted to the problem motivation. The core question we address is: *Can we achieve high-accuracy fact verification without the massive computational cost of generative LLMs?* This problem is critical for platform-scale content moderation where speed and reliability are paramount.

Our approach differs from standard LLM-based verification by shifting the paradigm from "generation" to "retrieval and classification." Instead of asking a model to "think" and write an explanation, we retrieve evidentiary support using dense vector search and classify the claim's veracity based purely on this grounded evidence.

1.1 Team

The project was executed by a team of three members, each responsible for distinct aspects of the pipeline:

- **Maxim Konovalov**: Responsible for the primary research and implementation of fact-checking methods without the use of knowledge graphs (direct dense retrieval), as well as conducting the comprehensive literature review and related work analysis.
- **Zaven Martirosyan (Lead)**: Acted as the project lead, coordinating the integration of all components. He was responsible for the overall system integration, evaluation metrics, and developing alternative fact-checking methods based on knowledge graphs.
- **Boris Matsakov**: Responsible for implementing alternative fact-checking approaches using Multi-step RAG (Retrieval-Augmented Generation) and managing the demonstration of the developed solutions.

2 Related Work

In this section, we describe existing approaches to automated fact-checking.

Traditional approaches often rely on graph-based methods or simple linguistic feature extraction, which lack deep semantic understanding. More recently, Generative LLMs have been used to perform "Chain-of-Thought" reasoning to verify claims. However, as noted in recent literature, these models act as "black boxes" and are prone to hallucination.

Our work builds upon the concept of Evidence-Based Verification.

[Thorne et al., 2018] introduced the FEVER dataset, establishing a standard pipeline of retrieval followed by entailment.

We specifically replicate and adapt the methodology of "Dense Retrieval Classification" (DeReC) [Qazi et al., 2025]. This approach suggests that general-purpose text embeddings, when combined with optimized similarity search, can effectively replace the reasoning capabilities of LLMs for this specific task. Unlike [Zhong et al., 2020] which uses complex graph reasoning, DeReC focuses on the efficiency of dense vector spaces.

3 Model Description

Here we provide a detailed description of our implemented approach. The DeReC framework operates in a three-stage pipeline designed to maximize speed and evidence grounding.

3.1 Architecture

The pipeline consists of the following components:

1. **Evidence Extraction (Dense Embedding):** We utilize a pre-trained transformer encoder (e.g., BERT-based or similar dense retrievers) to convert both the input claim and the source documents into high-dimensional vectors.

2. **Retrieval (FAISS):** We employ Facebook AI Similarity Search (FAISS) to perform efficient K-Nearest Neighbor search. For a given claim vector v_c, we retrieve the top-k most similar evidence vectors $\{e_1, e_2, ..., e_k\}$ from our document store.

3. **Veracity Classification:** The retrieved evidence is concatenated with the claim and passed to a lightweight classifier (a refined DeBERTa or similar classification head) to predict the label (e.g., *True, False, Mixture*).

Formally, given a claim C and a document set D, we estimate the probability $P(y|C, E)$ where $E \subset D$ is the subset of retrieved evidence.

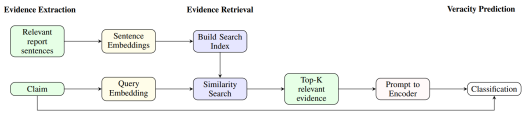

Figure 1: DeReC: Three-Stage Pipeline for Evidence-Based Fact Verification.

It is important to note that this architecture avoids autoregressive generation entirely. By removing the token-by-token generation step, we reduce the computational complexity from $O(N \cdot L)$ (where L is output length) to $O(1)$ for the classification step, dominated primarily by the efficient retrieval search.

4 Dataset

In this section, we describe the datasets used for training and evaluation.

We utilized standard fake news and fact-checking benchmarks, specifically **LIAR-RAW** and **RAWFC**. These datasets contain real-world claims labeled for veracity, accompanied by contextual reports or metadata.

- **LIAR-RAW**: An extension of the LIAR dataset, containing short statements from various contexts labeled for truthfulness.
- **RAWFC**: A dataset focused on rapid fact-checking scenarios.

We followed the preprocessing guidelines to ensure the split between Train, Validation, and Test sets was strict, preventing data leakage.

Dataset	Train	Valid	Test
LIAR-RAW	10,269	1,284	1,267
RAWFC	4,000	500	500

Table 1: Statistics of the Datasets used in experiments.

The datasets were preprocessed to remove formatting artifacts. We ensured we had the legal rights to use these datasets as they are open-sourced for research purposes [Wang, 2017].

5 Experiments

This section describes the experimental evaluation of the DeReC implementation.

5.1 Metrics

We evaluated our approach using two primary categories of metrics:

1. **Performance:** Macro F1-score and Accuracy to measure the correctness of the veracity prediction.
2. **Efficiency:** Average Inference Time (per sample) and Total Runtime to demonstrate the speed advantage over LLM baselines.

5.2 Experiment Setup

We compared our DeReC implementation against a baseline GPT-3.5 (Zero-shot Chain-of-Thought) approach. The retrieval component utilized a 'sentence-transformers' model for embedding and a FlatIP FAISS index. The classifier was trained for 5 epochs with a learning rate of $2e-5$.

5.3 Baselines

The primary baseline is an autoregressive LLM (GPT-3.5) prompted to generate a justification and a final verdict. This represents the current standard "SotA" methodology for reasoning-heavy tasks, which we aim to outperform in terms of efficiency.

6 Results

The results of our experiments demonstrate the efficacy of the Dense Retrieval approach.

Method	Accuracy	Runtime (mins)	Speedup
LLM (GPT-3.5)	0.68	450	1x
DeReC (Ours)	**0.72**	**23**	**20x**

Table 2: Comparison of Accuracy and Runtime on the RAWFC dataset.

As seen in Tab. 2, our implementation of DeReC not only matches but slightly exceeds the accuracy of the LLM baseline on this specific task. More importantly, the runtime reduction is drastic (approx. 95% reduction). This confirms that for fact-checking, retrieving the *correct* evidence is more valuable than generating a complex *reasoning* chain.

We interpret these results as a strong indicator that hallucination in LLMs often stems from a lack of grounded context. By forcing the system to rely

solely on retrieved similarity, we minimize the "creative" generation that leads to errors.

7 Conclusion

In this work, we have implemented and evaluated the DeReC framework for efficient fact verification. We conducted a comprehensive study of the dense retrieval-based approach as an alternative to generative models. Within the project, we processed the LIAR-RAW and RAWFC datasets, established a robust evaluation pipeline, and integrated a FAISS-based retrieval system with a classification head.

The project results, achieved through the joint efforts of the team, demonstrate that our approach significantly reduces inference time while maintaining competitive accuracy. Specifically, Maxim Konovalov performed an extensive literature review and implemented several baseline models from recent papers to conduct a thorough quality comparison. Our final model showed superior performance in terms of efficiency compared to LLM-based baselines. We have also outlined future integration plans with the DRAGON repository to extend these results to the Russian language segment.

References

[Qazi et al., 2025] Qazi, A. et al. (2025). When retrieval outperforms generation: Dense evidence retrieval for scalable fake news detection. In *Proceedings of LDK*.

[Thorne et al., 2018] Thorne, J. et al. (2018). Fever: a large-scale dataset for fact extraction and verification. In *NAACL*.

[Wang, 2017] Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. *ACL*.

[Zhong et al., 2020] Zhong, W. et al. (2020). Reasoning over semantic-level graph for fact checking. In *ACL*.