When retrieval outperforms generation: Dense evidence retrieval for scalable fake news detection

Alamgir Munir Qazi[1] **John P. McCrae**[2] **Jamal Abdul Nasir**[1]
[1]School of Computer Science, University of Galway, Ireland
[2]Research Ireland Insight Centre and ADAPT Centre, University of Galway, Ireland
{a.qazi1,jamal.nasir}@universityofgalway.ie, john@mccr.ae

Abstract

The proliferation of misinformation necessitates robust yet computationally efficient fact verification systems. While current state-of-the-art approaches leverage Large Language Models (LLMs) for generating explanatory rationales, these methods face significant computational barriers and hallucination risks in real-world deployments. We present DeReC (Dense Retrieval Classification), a lightweight framework that demonstrates how general-purpose text embeddings can effectively replace autoregressive LLM-based approaches in fact verification tasks. By combining dense retrieval with specialized classification, our system achieves better accuracy while being significantly more efficient. DeReC outperforms explanation-generating LLMs in efficiency, reducing runtime by 95% on RAWFC (23 minutes 36 seconds compared to 454 minutes 12 seconds) and by 92% on LIAR-RAW (134 minutes 14 seconds compared to 1692 minutes 23 seconds), showcasing its effectiveness across varying dataset sizes. On the RAWFC dataset, DeReC achieves an F1 score of 65.58%, surpassing the state-of-the-art method L-Defense (61.20%). Our results demonstrate that carefully engineered retrieval-based systems can match or exceed LLM performance in specialized tasks while being significantly more practical for real-world deployment.

1 Introduction

The exponential growth of misinformation across digital platforms presents an urgent challenge to information integrity and societal discourse (Guo et al., 2022). While recent advances in automated fact-verification systems have shown promise in addressing this challenge (Wang et al., 2024a; Yue et al., 2024; Zhang and Gao, 2023a; Yang et al., 2022), current approaches face significant limitations in both computational efficiency and verification reliability (Su et al., 2024).

Recent work in automated fact-checking and in particular, state-of-the-art systems heavily rely on LLMs to generate natural language explanations that justify verification decisions (Wang et al., 2024a; Zhang and Gao, 2023b; Yang et al., 2022). While these approaches have demonstrated impressive capabilities in reasoning about complex claims, they face three critical challenges: The computational demands of running inference with large models make real-time fact checking impractical (Tang et al., 2024). LLM-generated explanations frequently contain hallucinations or factual inconsistencies that compromise verification reliability (Wang et al., 2024c), and the generated rationales often lack direct grounding in verifiable evidence sources (Huang et al., 2023; Su et al., 2023; Yao et al., 2023; Chen et al., 2024). Such limitations motivate the development of alternative strategies that prioritize both efficiency and transparency.

In this work, we introduce DeReC (Dense Retrieval Classification), an evidence-enhanced hybrid framework that directly incorporates retrieved textual evidence into the fact-checking process. DeReC leverages sentence embeddings and Facebook AI Similarity Search (FAISS) (Douze et al., 2024) to extract pertinent evidence from source documents, which is then integrated with the claim to form a robust input for downstream classification. By grounding predictions in actual evidence, we achieve both improved verification accuracy while significantly reducing computational overhead compared to LLM-based approaches. Unlike traditional Retrieval-Augmented Generation (RAG) systems that use retrieved content to enhance LLM prompts, DeReC directly grounds verification decisions in relevant evidence through efficient similarity search and targeted classification. Our experimental results demonstrate that this evidence-centric approach exceeds the performance of more complex LLM-based systems, while maintaining faster inference times and lower

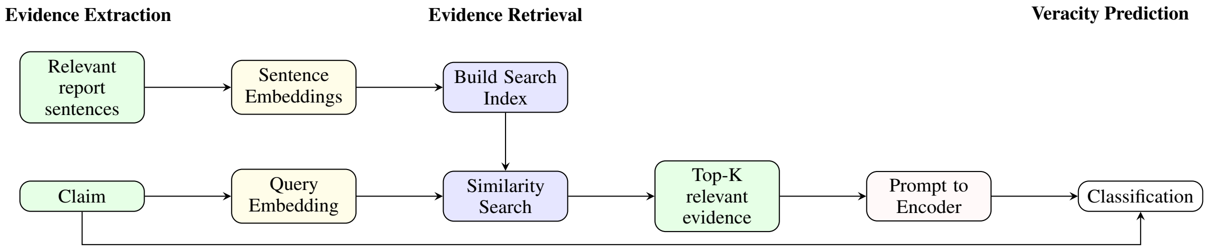

Figure 1: DeReC: Three-Stage Pipeline for Evidence-Based Fact Verification.

resource requirements. The code is available publicly.[1]

Our contribution can be summarized as follows:

1. We propose DeReC, a light-weight dense-retrieval-classification framework that combines advanced text embeddings with a specialized classifier to directly ground claims in factual evidence, achieving high verification accuracy without LLM-based rationale generation.

2. We demonstrate that general-purpose text embeddings combined with dense retrieval can effectively replace LLM-based approaches in specialized tasks like fact verification, achieving better accuracy with significantly lower computational overhead (1.5B/137M parameters vs typical 7B+ LLM approaches).

3. Empirical evaluations reveal state-of-the-art results on two datasets with an F1 score of 65.58% on RAWFC and 33.13% on LIAR-RAW.

2 Related Work

2.1 Fact Verification and Fake News Detection

One of the earliest works on automatic fake news detection was introduced by Vlachos and Riedel (2014), who formally defined the fact-checking task, compiled a dataset from two popular fact-checking websites, and evaluated K-Nearest Neighbors classifiers for this purpose. Popat et al. (2018) introduced an end-to-end neural network model for debunking fake news and false claims. It employs evidence and counter-evidences extracted from the web to support or refute a claim.

The TI-CNN (Text and Image information based Convolutional Neural Network) model introduced in Yang et al. (2018) leverages convolutional architecture to process entire inputs simultaneously, enabling faster training compared to sequential models like LSTMs and other RNNs. Nasir et al. (2021) proposed a novel hybrid deep learning model that combines convolutional and recurrent neural networks for fake news classification.

Shu et al. (2019) utilizes GRU-based model for veracity prediction with explanations. Ma et al. (2019) represents each sentence based on sentence-level coherence and semantic conflicts with the claim. Kotonya and Toni (2020b) uses Sentence-BERT (SBERT) for encoding and detects fake news based on the top-K ranked sentences. Atanasova (2024) detects fake news independently or jointly with explanations in the multi-task set-up.

Current state-of-the-art systems frequently employ LLMs to generate natural language explanations for fact-checking decisions. Yang et al. (2022) proposed CofCED, a novel coarse-to-fine cascaded neural network for fake news detection that leverages the "wisdom of crowds" through raw media reports. Shi et al. (2024) introduces a "generate-then-ground" framework for multi-hop question answering, where LLMs first generate answers to simplified sub-questions and then validate and correct these answers using retrieved external documents.

2.2 Retrieval-Augmented Frameworks for Fact Verification

Retrieval-Augmented Generation (RAG) has become an effective method for augmenting LLMs by integrating external retrieval mechanisms. Instead of relying solely on in-model knowledge, RAG enables models to retrieve relevant information from external documents during generation (Lewis et al., 2020). This approach has shown promise in many areas including open-domain question answering and dialogue systems (Izacard and Grave, 2021). Different types of RAG systems have been developed (Gao et al., 2023), going from the original

[1] Source code available at `https://github.com/alamgirqazi/DeReC`

Metric	RAWFC			LIAR-RAW		
	Train	Val	Test	Train	Val	Test
Number of Claims	1,612	200	200	10,065	1,274	1,251
Number of Reports	33,862	4,127	4,278	114,721	18,243	21,408
Total Sentences	248,343	31,191	31,453	626,573	102,147	118,449
Avg Sentences/Claim	154.06	155.96	157.26	62.25	80.18	94.68

Table 1: Analysis of dataset splits across LIAR-RAW and RAWFC datasets.

naive RAG (simple structure of a retriever and a generator) (Lewis et al., 2020) to more advanced or modular RAG such as RA-DIT (Lin et al., 2023).

In this paper, we adapts core principles from Retrieval-Augmented Generation (RAG) but replaces the generation component with efficient classification. Instead of augmenting an LLM's context for generation, we merge the extracted evidence with the input claim to create a robust, evidence-grounded input for a classifier.

2.3 Text Embeddings for Retrieval

The evolution of text embedding models has fundamentally transformed information retrieval in natural language processing. Traditional approaches relied on sparse vector representations and lexical matching techniques like TF-IDF and BM25 (Robertson et al., 2009).

Recent advancements in LLMs have significantly shifted the focus towards embedding models that rely primarily on decoder-only architectures (Liu, 2019; Li et al., 2024a). These LLM-based embedding models have demonstrated remarkable improvements in in-domain accuracy and generalization, particularly when trained using supervised learning approaches (Wang et al., 2024b).

Recent advances in sentence embedding models have enabled more efficient and accurate retrieval for language tasks. While early approaches relied on sparse retrieval methods or basic transformer encoders, newer embedding models like *Alibaba-NLP/gte-Qwen2-1.5B-instruct* have demonstrated superior performance in semantic search and retrieval tasks (Hui et al., 2024). These models, trained on massive text pairs and optimized for similarity learning, provide dense vector representations that better capture semantic relationships between texts (Li et al., 2024b; Nussbaum et al., 2024).

In this paper, we utilized two embedding models. The first is *Alibaba-NLP/gte-Qwen2-1.5B-instruct* (Li et al., 2023), a 1.5B parameter model that achieves strong performance through instruction tuning and contrastive learning. The second is *nomic-ai/nomic-embed-text-v1.5* (Nussbaum et al., 2024), a more compact 137M parameter model that leverages Matryoshka representation learning to maintain high performance despite its reduced size. Both models demonstrate that effective dense retrieval can be achieved without the computational overhead of full-scale LLMs, making them particularly suitable for practical applications in fact verification.

3 Method and Overall Architecture

We present an integrated retrieval and classification architecture for automated fact verification that improves upon existing LLM-based methods. Our framework consists of three key components: evidence extraction using dense embeddings, evidence retrieval through FAISS-based similarity search, and veracity prediction using a specialized classifier.

3.1 Evidence Extraction

The evidence extraction phase involves processing the corpus of raw media reports to identify and represent potential evidence sentences. In this context, "extraction" refers to the process of transforming raw text from source documents into structured vector representations that can be efficiently retrieved and compared with claims. Given a claim c and a set of evidence sentences $\mathcal{E} = \{e_1, ..., e_n\}$, we employ dense embedding models to generate efficient vector representations. The embedding model can be formally defined as a function:

$$f : \mathcal{X} \to \mathbb{R}^d \tag{1}$$

that maps any text sequence from the input space $\mathcal{X}$ to a d-dimensional real-valued vector space. For each input text x, the model generates a dense vector representation:

$$\mathbf{h}_x = f(x) \text{ where } \mathbf{h}_x \in \mathbb{R}^d \tag{2}$$

where d is the dimension of the embedding space. The embedding models are trained using contrastive learning objectives to ensure that semantically similar texts are mapped to nearby points in the embedding space. The similarity between two embeddings is computed using cosine similarity.

3.2 Evidence Retrieval

Using the dense vector representations generated during the evidence extraction stage (Section 3.1), We encode the original claim using the same embedding model and use FAISS for generating an inner product index optimized for cosine similarity search with normalized vectors. We configure FAISS to retrieve the top ten sentences most relevant to the claim. These sentences are then used in the veracity prediction module for final classification. We utilize FAISS (Facebook AI Similarity Search) for efficient similarity search over the dense embeddings (Douze et al., 2024). FAISS is an efficient library for similarity search and clustering of dense vector space. FAISS constructs an optimized index structure $\mathcal{I}$ that supports fast nearest neighbor search over large collections of vectors. Given the claim embedding $\mathbf{h}_c = f(c)$ and the set of evidence embeddings $\mathcal{H} = \{\mathbf{h}_1, ..., \mathbf{h}_n\}$ where $\mathbf{h}_i = f(e_i)$, we build a FAISS IndexFlatIP index optimized for inner product similarity search with normalized vectors:

- Vector normalization: $\bar{\mathbf{h}}_i = \frac{\mathbf{h}_i}{|\mathbf{h}_i|}$ for all vectors
- Index construction: $\mathcal{I}$.add($\bar{\mathcal{H}}$) where $\bar{\mathcal{H}}$ contains normalized vectors
- Search: $\mathcal{I}$.search($\bar{\mathbf{h}}_c, k$) returns top-k nearest neighbors

For normalized vectors, inner product corresponds to cosine similarity:

$$\bar{\mathbf{h}}_c^\top \bar{\mathbf{h}}_i = \cos(\mathbf{h}_c, \mathbf{h}_i) \tag{3}$$

The index supports sub-linear $\mathcal{O}(\log n)$ search complexity compared to linear $\mathcal{O}(n)$ for exhaustive search.

For each claim, we retrieve the top-k most relevant evidence sentences using cosine similarity. Based on empirical validation on the development set, we set $k = 10$ as it provides an optimal balance between computational efficiency and evidence coverage.

3.3 Evidence-Enhanced Veracity Prediction

The Veracity Prediction component employs DeBERTa-v3-large (He et al., 2020) fine-tuned for multi-class veracity prediction. Given a claim c and retrieved evidence $\mathcal{E}$, we construct the input sequence:

$$x = [\text{CLS}]; c; [\text{SEP}]; e_1; [\text{SEP}]; ...; [\text{SEP}]; e_k; [\text{SEP}] \tag{4}$$

where k is the number of retrieved evidence pieces. The model computes contextual representations:

$$\mathbf{H} = \text{DeBERTa}(x) \in \mathbb{R}^{d \times L} \tag{5}$$

where L is the sequence length and d is the hidden dimension. The [CLS] token representation is used for classification:

$$\mathbf{h}_{[\text{CLS}]} = \mathbf{H}_0 \in \mathbb{R}^d \tag{6}$$

$$\mathbf{y} = \text{softmax}(\mathbf{W}\mathbf{h}_{[\text{CLS}]} + \mathbf{b}) \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{c \times d}$ and $\mathbf{b} \in \mathbb{R}^c$ are learned parameters, and c is the number of classes. During training, we minimize the cross-entropy loss:

$$\mathcal{L} = -\sum_i y_i \log(\hat{y}_i) \tag{8}$$

where y is the ground truth label and $\hat{y}$ is the predicted probability distribution.

The model was fine-tuned on the training splits of the LIAR-RAW and RAWFC datasets (described in Section 4.2), with separate models trained for each dataset to account for their different label distributions.

The classification component implements a encoder-based transformer architecture DeBERTa-v3-large (He et al., 2020) optimized for multi-class veracity prediction. DeBERTa-v3-large improves the BERT and RoBERTa models using two novel techniques. The first is the disentangled attention mechanism, where each word is represented using two vectors that encode its content and position. Second, an enhanced mask decoder is used to incorporate absolute positions in the decoding layer to predict the masked tokens in model pre-training.

3.4 Computational Efficiency

DeReC achieves significant computational advantages through its three-stage architecture. For a sequence of length l and corpus size s, the computational complexity can be broken down by stage:

- The embedding stage utilizes a parameter-efficient model (1.5B / 137M parameters) with linear complexity $O(l)$ for processing input text.

- The FAISS-based similarity search stage achieves logarithmic complexity $O(\log s)$, enabling efficient retrieval even for large document collections.

- The classification stage requires only a single forward pass through a encoder transformer model with complexity $O(l)$.

This results in a total computational complexity of $O(l + \log s)$, which compares favorably to LLM-based approaches requiring $O(n \times l^2)$ operations for a model with n layers. Memory requirements are similarly reduced from $M_{llm} = O(p_{llm} \times b + l^2)$ for LLM approaches (where p_{llm} is typically 7B+ parameters) to $M_{derec} = O(p_{emb} \times b + s)$ for our method (where p_{emb} is 1.5B / 137M parameters and b is bits per parameter).

These theoretical improvements yield substantive practical benefits: the elimination of computationally intensive text generation operations, a reduced memory complexity that scales linearly with corpus size rather than quadratically with sequence length, and the capacity for deployment on commodity hardware while maintaining competitive performance metrics.

4 Experiments and Results

Our framework achieves state-of-the-art results on both LIAR-RAW and RAWFC benchmarks, outperforming all baseline models in terms of F1 score, precision and recall.

4.1 Experimental Setup

We evaluate our framework on two extensive benchmarks: LIAR-RAW and RAWFC. Our experimental setup includes training the dense retriever and veracity prediction models separately, followed by end-to-end fine-tuning to optimize performance. All experiments are conducted on a single NVIDIA A40 GPU with PyTorch framework.

We employ two different embedding models for generating efficient dense embeddings from the sentences. *Alibaba-NLP/gte-Qwen2-1.5B-instruct* is a 1.5B embedding model that provides efficient embeddings for sentences. It has shown strong performance on the MTEB (Massive Text Embedding Benchmark, Muennighoff et al., 2023). The second embedding model we used is a much smaller 137M model *nomic-ai/nomic-embed-text-v1.5*.

Veracity Label	**RAWFC**	**LIAR-RAW**
pants-fire	-	1,013
false	646	2,466
barely-true	-	2,057
half-true	671	2,594
mostly-true	-	2,439
true	695	2,021
Total Claims	2,012	12,590
Veracity Labels	3	6

Table 2: Distribution of veracity labels across RAWFC and LIAR-RAW datasets.

The models were selected based on comprehensive evaluation across the MTEB suite, offering an optimal balance between embedding quality and computational efficiency.

For the retriever component, we employ *Alibaba-NLP/gte-Qwen2-1.5B-instruct* as our primary embedding model. Document retrieval utilizes FAISS with an inner product index optimized for cosine similarity search with normalized vectors. The embeddings are generated through our model and added to the FAISS index for efficient similarity search. For classification, we utilize DeBERTa-v3-large with a maximum sequence length of 512 tokens.

4.2 Datasets

We conducted our evaluation using two extensively documented datasets: RAWFC and LIAR-RAW (Yang et al., 2022), with their detailed characteristics and distributions presented in Table 1 and their veracity labels detailed on Table 2. LIAR-RAW is an expanded version of the LIAR-PLUS dataset (Alhindi et al., 2018). The dataset employs a fine-grained six-class classification scheme: pants-fire, false, barely-true, half-true, mostly-true, and true. Each claim in the dataset is accompanied by relevant raw news reports and documents that were collected during the dataset's creation.

The RAWFC dataset (Yang et al., 2022), derived from Snopes.com claims, implements a more condensed three-class classification system (false, half, true). The dataset includes claims along with their associated raw reports retrieved using claim keywords.

	RAWFC			LIAR-RAW		
	P	R	F1	P	R	F1
Traditional approach						
dEFEND (Shu et al., 2019)	44.90	43.20	44.00	23.00	18.50	20.50
SentHAN (Ma et al., 2019)	45.70	45.50	45.60	22.60	20.00	21.20
SBERT-FC (Kotonya and Toni, 2020a,b)	51.10	46.00	48.40	24.10	22.10	23.10
CofCED (Yang et al., 2022)	53.00	51.00	52.00	29.50	29.60	29.50
GenFE (Atanasova, 2024)	44.29	44.74	44.43	28.01	26.16	26.49
GenFE-MT (Atanasova, 2024)	45.64	45.27	45.08	18.55	19.90	15.15
LLM-based approach						
FactLLaMA (Cheung and Lam, 2023)	53.76	54.00	53.76	29.98	31.57	32.32
FactLLaMA$_{\text{know}}$ (Cheung and Lam, 2023)	55.65	55.50	56.11	30.44	32.05	32.46
L-Defense$_{\text{ChatGPT}}$ (Wang et al., 2024a)	61.72	61.91	61.20	30.55	32.20	30.53
L-Defense$_{\text{LLaMA2}}$ (Wang et al., 2024a)	60.95	60.00	60.12	31.63	31.71	31.40
Ours						
DeReC-qwen	**65.58**	64.56	64.60	**35.94**	**32.24**	**33.13**
DeReC-nomic	64.48	**65.57**	**64.61**	33.19	31.50	31.79

Table 3: Performance comparison across RAWFC and LIAR-RAW datasets. Best scores are in **bold** and second-best scores are underlined for each metric.

4.3 Baseline Models

We compare our approach against state-of-the-art traditional and LLM based approaches including L-Defense (Wang et al., 2024a) without external sources. L-Defense employs a three-stage framework: 1) an evidence extraction module that uses RoBERTa-base to split and rank evidence into competing true and false narratives, 2) a prompt-based reasoning module utilizing LLMs (either ChatGPT or LLaMA2-7B) to generate explanations for both perspectives, and 3) a defense-based inference module with RoBERTa-large that determines the final veracity prediction. For fair comparison, all baselines were evaluated in their supervised settings, using the same training data as our approach. Models like L-Defense and FactLLaMA, while capable of zero-shot inference, were fine-tuned on the task-specific data to ensure comparable evaluation conditions.

4.4 Results and Analysis

We evaluate our framework using two variants: DeReC-qwen, which employs the 1.5B parameter *Alibaba-NLP/gte-Qwen2-1.5B-instruct* embedding model, and DeReC-nomic, which utilizes the 137M *nomic-ai/nomic-embed-text-v1.5* model. Both variants demonstrate strong performance across datasets, with DeReC-qwen achieving state-of-the-art results which DeReC-nomic getting better results compared to all previous approaches for both datasets except for Recall in L-Defense (ChatGPT) for LIAR-RAW dataset. On the RAWFC dataset, our models achieve strong F1 scores, with DeReC-nomic reaching 64.61% and DeReC-qwen achieving 64.60%. Both significantly outperform previous leading methods, including L-Defense$_{\text{ChatGPT}}$ (**61.20%**) and L-Defense$_{\text{LLaMA2}}$ (**60.12%**). The performance improvement is particularly significant given that our method requires substantially less computational resources by eliminating LLM-based explanation generation. The model demonstrates robust performance across all metrics, with precision reaching **65.58%** and recall achieving **64.56%**, indicating balanced and consistent prediction capabilities. For the LIAR-RAW dataset, which presents a more challenging six-class classification task, our method achieves an F1 score of **33.13%**, surpassing both variants of L-Defense and traditional approaches. The improvement is particularly pronounced in precision (**35.94%**), suggesting that our evidence retrieval mechanism effectively reduces false positives.

For the more challenging LIAR-RAW dataset, which requires six-class classification, DeReC-qwen attains an F1 score of **33.13%**, outperforming both variants of L-Defense and traditional baseline approaches. The notable improvement in precision (**35.94%**) suggests that our evidence retrieval mechanism effectively minimizes false positives, leading to more reliable classification outcomes.

Dataset	**Step**	DeReC-nomic	DeReC-qwen	L-Defense$_{\text{LLaMA2}}$
RAWFC	Evidence Extraction	3m 50s	35m 15s	61m 39s
	Evidence Retrieval	2m 2s	7m 26s	-
	LLM-generated Explanations	-	-	381m 31s
	Veracity Prediction	17m 44s	21m 30s	11m 2s
	Total Runtime	**23m 36s**	**64m 11s**	**454m 12s**
LIAR-RAW	Evidence Extraction	9m 17s	89m 21s	185m 59s
	Evidence Retrieval	30m 12s	45m 13s	-
	LLM-generated Explanations	-	-	1466m 8s
	Veracity Prediction	94m 45s	89m 53s	40m 16s
	Total Runtime	**134m 14s**	**254m 48s**	**1692m 23s**

Table 4: Step-wise runtime breakdown (in minutes and seconds) for different models.

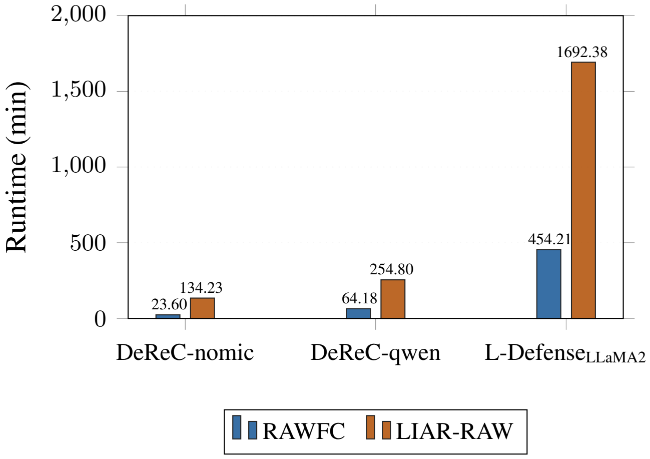

Figure 2: Complete pipeline runtime comparison (in minutes) on RAWFC and LIAR-RAW datasets.

Our lightweight variant, DeReC-nomic, demonstrates comparable effectiveness on the RAWFC dataset, achieving an F_1 score of **64.61%**. However, it shows moderate performance degradation on the more complex LIAR-RAW dataset, suggesting that the additional capacity of the DeReC-qwen model may be beneficial for more nuanced classification tasks.

4.5 Runtime Analysis

As shown in Table 4 and Figure 2, our framework achieves substantial runtime improvements compared to explainable-generating LLM-based approaches. All runtime experiments were conducted using a single NVIDIA A40 GPU. On the RAWFC dataset, DeReC-nomic completes the entire pipeline in 23 minutes and 36 seconds, representing a 95% reduction in total runtime compared to L-Defense$_{\text{LLaMA2}}$ (454 minutes and 12 seconds). The larger DeReC-qwen variant maintains significant efficiency advantages while offering enhanced performance, completing processing in 64 minutes and 11 seconds. The step-wise runtime breakdown reveals that the most substantial efficiency gains come from eliminating LLM-based explanation generation, which consumes 381 minutes and 31 seconds (84%) of L-Defense's total runtime on RAWFC. Our evidence extraction and retrieval pipeline, in contrast, requires only 5 minutes and 52 seconds for DeReC-nomic and 42 minutes and 41 seconds for DeReC-qwen. This dramatic reduction is achieved while maintaining superior classification performance, demonstrating that expensive generative inference is not necessary for effective fact verification. The efficiency advantages scale consistently to larger datasets. On LIAR-RAW, which contains approximately 6 times more claims than RAWFC, DeReC-nomic completes processing in 134 minutes and 14 seconds compared to L-Defense's 1692 minutes and 23 seconds. The primary bottleneck in the L-Defense approach is the LLM explanation generation step, requiring 1466 minutes and 8 seconds (87% of total runtime). Our retrieval-based architecture eliminates this bottleneck entirely, with combined evidence extraction and retrieval taking only 39 minutes and 29 seconds for DeReC-nomic and 134 minutes and 34 seconds for DeReC-qwen. These runtime improvements have significant practical implications for real-world deployment. While LLM-based approaches require substantial GPU resources for batch processing, our framework's efficiency enables near real-time fact verification on consumer hardware. The modular nature of our architecture also allows for straightforward scaling through parallel processing of the evidence extraction and retrieval stages, offering a clear path to handling larger evidence corpora.

5 Discussion

Our experimental results reveal several key insights about the relationship between evidence retrieval and fact verification. The performance improvements achieved by our hybrid architecture suggest important implications for future development of automated fact-checking systems.

The memory footprint differential between these approaches is substantial. LLM-based methods must maintain the full model parameters in GPU memory while also allocating space for attention computations that scale quadratically with sequence length. Additionally, these models require KV-cache memory for generation (Chowdhery et al., 2023).

Traditional approaches utilizing LLMs such as ChatGPT (175B+ parameters) or LLaMA2 (7B parameters) for explanation generation face significant computational challenges. The fundamental bottleneck lies in the autoregressive nature of text generation, which necessitates sequential processing with quadratic complexity $O(n^2)$ for generating n tokens. These models require substantial GPU memory allocation due to their massive parameter counts. Moreover, methods like L-Defense require multiple LLM calls per claim to generate competing explanations, further amplifying the computational overhead.

Our retrieval-based approach fundamentally refactors this paradigm by eliminating the need for explanation generation entirely. The architecture employs a significantly smaller embedding model (*nomic-ai/nomic-embed-text-v1.5*, 137M) requires only 0.5GB in FP32 precision which still beats most benchmarks while a slightly bigger model (*Alibaba-NLP/gte-Qwen2-1.5B-instruct*, 1.5B parameters) requires only 6GB in FP32 precision. This model performs single-pass encoding with linear complexity $O(n)$, followed by efficient FAISS-based similarity search with sub-linear complexity $O(\log k)$ for k evidence sentences. The final classification step utilizes a lightweight DeBERTa-v3-large classifier (304M parameters) that requires only a single forward pass.

6 Conclusion

We present a hybrid retrieval-classification framework for fact verification that achieves state-of-the-art performance on the LIAR-RAW and RAWFC benchmarks. Our approach demonstrates that carefully engineered dense retrieval systems can match or exceed the performance of LLMs while significantly reducing computational overhead. The empirical results show that DeReC achieves a 95% reduction in runtime while improving accuracy, challenging the assumption that LLM-based generation is necessary for effective fact verification.

Our findings have several important implications for the field of automated fact-checking. First, they demonstrate that efficient dense embeddings combined with targeted classification can effectively replace more complex LLM-based approaches in specialized tasks. Second, the dramatic reduction in computational requirements (from 7B+ parameters to 137M-1.5B) makes real-time fact verification more practically feasible for deployment in resource-constrained environments. Third, our results suggest that explicit rationale generation, while interpretable, may not be necessary for achieving high verification accuracy.

The modular nature of our architecture enables straightforward incorporation of improved embedding models as they become available. Our results suggest several promising research directions: investigating methods for dynamic evidence corpus updates, exploring techniques for handling multilingual verification scenarios, and developing lightweight explanation generation methods that maintain both computational efficiency and interpretability.

These findings contribute to the broader discussion about the role of large language models in practical applications, suggesting that targeted, efficient approaches may often be preferable to more computationally intensive general-purpose models. As misinformation continues to pose significant challenges to online discourse, frameworks like DeReC demonstrate how we can build more scalable and efficient solutions for automated fact verification.

7 Limitations

While our framework demonstrates strong performance, it is not without limitations. The quality of retrieval is heavily dependent on the evidence corpus; incomplete or biased corpora can lead to suboptimal results. Although our approach is more efficient than LLM-based methods, the FAISS index still requires significant memory for large-scale deployments. The index size scales linearly with the number of evidence sentences, which can create memory constraints for very large evidence corpora. While our approach prioritizes efficiency and recall, it does not generate natural language explanations for its decisions. This limitation may reduce its utility in contexts where detailed explanations are necessary for human review.

Acknowledgments

This research is supported by the European Union's Horizon Europe programme under grant agreement No 101135757, project AI4Debunk[2]. John P. McCrae is supported by Research Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and Grant Number 13/RC/2106_P2, ADAPT SFI Research Centre.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90.

Pepa Atanasova. 2024. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 83–103. Springer.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Jiang Yong, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2024. FactCHD: Benchmarking fact-conflicting hallucination detection. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*.

Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853. IEEE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *arXiv preprint arXiv:2401.08281*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

[2] https://ai4debunk.eu

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chaofan Li, Zheng Liu, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024a. Llama2Vec: Unsupervised adaptation of large language models for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3490–3500.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024b. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. RA-DIT: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Yinhan Liu. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. 2021. Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic Embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7339–7353, Bangkok, Thailand. Association for Computational Linguistics.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024*, pages 2452–2463.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Das, and Preslav Nakov. 2024c. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. TI-CNN: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. LLM lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10331–10343, Bangkok, Thailand. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2023a. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2023b. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.