Journal of Computer Science

Research Article

HMA-RAG: A Hierarchical Memory Augmented Retrieval Augmented Generation Framework for Fake News Detection

Afnan Altarjami and Sawsan Alshattnawi

Department of Computer Science, Yarmouk University, Irbid, Jordan

Article history
Received: 12-04-2026
Revised: 10-07-2026
Accepted: 15-07-2026

Corresponding Author:
Afnan Altarjami
Department of Computer Science, Yarmouk University, Irbid, Jordan
Email: afnanaltarjami@gmail.com

Abstract: Existing methods for detecting fake news, such as traditional machine learning and transformer-based models, often work as separate classifiers that do not use outside evidence, historical knowledge, or ways to make the results easier to understand. This paper presents a framework that combines retrieval augmentation and hierarchical memory to overcome three primary constraints: Insufficient external evidence, the absence of memory, and restricted explainability in existing methodologies. We present HMA-RAG (Hierarchical Memory-Augmented Retrieval-Augmented Generation), a dual-branch architecture that integrates DistilBERT-based classification with dense retrieval via MiniLM. The framework has a confidence-gated mechanism that only turns on a multi-stage verification pipeline when certain conditions are met. This pipeline includes evidence agreement, memory consistency, and temporal tracking. A single credibility score combines these signals, and a feedback loop keeps long-term memory up to date. Tests on the WELFake dataset (72,134 articles) show that HMA-RAG achieves an F1-score of 98.83%, which exceeds classical baselines and is comparable to a larger transformer model, FakeBERT (98.90% on the ISOT dataset), while using fewer classification parameters (66M vs. approximately 115M). Ablation studies verify the efficacy of the hierarchical memory components. Among the systems compared, only HMA-RAG provides four-layer explainability: Evidence retrieval, dual-stage verification signals, temporal context tracking, and credibility score decomposition. These results indicate that combining retrieval augmentation with hierarchical memory can achieve accuracy comparable to substantially larger models while providing structured, multi-layer transparency intended for deployment in journalism, fact-checking, and content moderation.

Keywords: Fake News Detection, Retrieval Augmented Generation, Hierarchical Memory, Distilbert, Explainability, Welfake, Faiss, Transformer Models

Introduction

The advent of social media platforms and news agencies has revolutionized how news is generated, disseminated, and consumed. This has given rise to a proliferation of information being disseminated all over the globe. However, this phenomenon has also given rise to misinformation and fake news (Zhou and Zafarani, 2021; Shu et al., 2017). The implications of fake news are enormous. For instance, fake news has influenced health policies in a pandemic world, influenced election outcomes in democratic nations, and impacted financial markets (Vosoughi et al., 2018).

Traditionally, fake news detection has utilized feature engineering techniques in combination with traditional machine learning techniques such as SVM and Logistic Regression (Ahmed et al., 2017; Wang, 2017). However, these traditional machine learning models are not capable of understanding complex semantic information embedded in fake news. The advent of deep learning models such as BERT with 110 million parameters (Devlin et al., 2019) and DistilBERT with 66 million parameters (Sanh et al., 2019) has revolutionized fake news detection by providing a richer understanding of content using contextual information embedded in text data.

However, even the most accurate transformer-based classifier has three inherent limitations. First, it is not capable of retrieving any external evidence to support its claims. Second, it is not capable of accumulating any knowledge from previous predictions. Third, it is not capable of explaining its claims to end users. For instance, a transformer-based classifier would not be able to explain to a user why a particular article is fake or real.

In order to address the three limitations, we propose a Hierarchical Memory-Augmented Retrieval-Augmented Generation framework for fake news detection, referred to as HMA-RAG. This framework comprises the following three components.

First, FAISS-Based Dense Retrieval. We use Facebook AI Similarity Search (FAISS) to perform dense retrieval of the 5 most semantically similar documents to the given input from a chosen knowledge base as the external evidence for all the classification results. The retrieval branch uses the all-MiniLM-L6-v2, 22M parameter, sentence transformer, which produces 384-dimensional embeddings for the FAISS similarity search.

Second, we propose a Hierarchical Memory Architecture. We draw lessons from cognitive science models of human memory (Tulving, 1972; Baddeley, 2000) and design our model with a three-layer memory system consisting of short-term memory (STM) and long-term memory (LTM) as well as a temporal tracking component. Specifically, the STM acts as a buffer storing the last 20 uncovered evidence patterns. The LTM is responsible for storing permanent knowledge, and knowledge stored in LTM follows an exponential decay function, which leads to pruning of least important information. In addition, we utilize A Temporal Tracking Component (TTC) with a window of 50 predictions to track temporal patterns of spreading fake news.

Third, Confidence-Gated Credibility Scoring. Our system includes a gating mechanism where only the predictions for which the base learner is uncertain are corrected based on evidence. We found a confidence threshold of 0.80 sufficed to mitigate the well-known failing case of ensemble-based systems where almost-certain correct predictions are overruled by the auxiliary output based on disagreements among learners. Our credibility score is a weighted sum given by: 30% for agreement among the evidence sources, 25% for the memory consistency constraint, 20% for the temporal consistency constraint, and 25% for the classifier (DistilBERT) confidence.

In this work, we adopt DistilBERT with 66M parameters as the classification encoder of HMA-RAG instead of popular BERT-base with 110M parameters. Although DistilBERT retains 97% of the performance of BERT, it has 40% fewer parameters and is 60% faster (Sanh et al., 2019). In our experiments, the HMA-RAG achieves an F1 score of 98.83% for the detection of fake news from the dataset WELFake, a score that is only 0.07% lower than the corresponding score achieved by FakeBERT (Kaliyar et al., 2021) that is based on BERT-base and additional CNN layers of about 115M parameters.

We evaluate HMA-RAG on the WELFake benchmark dataset (Verma et al., 2021) which has 72,134 news articles crawled from the websites of four fact-checking sources: PolitiFact, GossipCop, BuzzFeed, and ISOT. Our system outperforms the standalone DistilBERT baseline. Moreover, it provides four different explainable outputs at its output layer:

(1) Evidence Retrieval (ER): A set of similar articles with their corresponding labels
(2) Dual-Stage Verification (DSV): A set of evidence along with memory signals for the dual-stage verification
(3) Temporal Context Tracking (TCT): Information about the evolving trends of misinformation over time
(4) Credibility Score Decomposition (CSD)

The contribution of each component towards the final credibility score.

Related Work

The process of detecting fake news has passed through various methodological approaches, starting from traditional handcrafted feature engineering approaches, then deep neural approaches, and finally retrieval approaches. This section discusses major developments in five relevant areas of fake news detection.

Traditional Machine Learning Approaches

The earliest stages of fraud news detection focused on the use of handcrafted features and traditional machine learning methods. Ahmed et al. (2017) showed that linear classifiers and TF-IDF representations are very competitive in binary credibility classification. Wang (2017) proposed the LIAR dataset and examined the multi-class veracity classification using SVMs and Logistic Regression with metadata features. Perez-Rosas et al. (2018) studied the linguistic characteristics of classification sentiment analysis and the readability of the linguistic features. Although the approaches offer interpretability and computational efficiency, they are limited by the quality and generalizability of engineered features and the evolving linguistic tactics that sophisticated misinformation producers use. More recently, Alshattnawi et al. (2024) demonstrated that contextualized word embeddings such as BERT and ELMo outperformed static embeddings (Word2Vec, GloVe) for social media text classification, achieving 90–94% accuracy on Twitter and YouTube spam detection tasks. Their findings underscore the importance of contextual language representations for capturing subtle semantic signals in informal online text, a principle that motivates our adoption of transformer-based encoders for fake news detection.

Transformer Based Methods

The field of Natural Language Processing has been substantially reshaped by the development of transformer architectures (Vaswani et al., 2017). BERT (Devlin et al., 2019) and its distilled variant DistilBERT (Sanh et al., 2019) established strong baselines for contextual text classification, with DistilBERT retaining approximately 97% of BERT's language understanding while being 40% smaller and 60% faster.

Focusing on the fake news detection niche, Kaliyar et al. (2021) constructed FakeBERT, BERT-base together with some CNN blocks, for a total of ~115 million parameters, and is able to reach 98.90% accuracy on the ISOT dataset. Recently, Kuntur et al. (2024) achieved 99.72% on WELFake using fine-tuned GPT-2. These approaches, though, do not give any kind of transparency and explanations about their decisions.

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) (Lewis et al., 2020) combines generative language models with information retrieval systems, enabling access to external knowledge during inference. Thorne et al. (2018) established the retrieve-then-classify paradigm through the FEVER benchmark, and efficient dense-vector similarity search libraries such as FAISS (Johnson et al., 2021) have made large-scale nearest-neighbour retrieval over high-dimensional embeddings practical. Recent work has begun applying RAG to fake news detection. Niu et al. (2024) proposed VeraCT Scan, which integrates RAG with source credibility assessment for claim verification. Liao et al. (2023) introduced MUSER, a multi-step evidence retrieval enhancement framework for fake news detection. Li et al. (2024) developed a multi-round retrieval-augmented LLM framework that iteratively gathers web evidence for more comprehensive verification. However, these approaches rely on flat retrieval without incorporating hierarchical memory structures. To our knowledge, the integration of RAG with hierarchical memory-augmented architectures for fake news detection combining evidence retrieval, memory consistency, and temporal tracking has received limited attention.

Memory Augmented Neural Networks

The idea of using memory networks is not new for deep learning models. Graves et al. (2014) proposed Neural Turing Machines that can learn to read and write to an external memory. Sukhbaatar et al. (2015) proposed an end-to-end memory network for question answering tasks. Weston et al. (2015) showed that memory networks can store information for a long time. However, to our knowledge, few studies integrate such memory models with retrieval augmentation for fake news detection while also providing structured explainability.

Explainable Fake News Detection

The problem of explainability has become a popular topic when dealing with fake news detection as the black-box classification model becomes increasingly difficult to defend in high-stakes settings like journalism and content moderation. Hu et al. (2024) proposed a contrastive learning system with dual models where the explanation is provided through large language models' generated natural languages. Popat et al. (2018) created DeClarE that provides word-level attention of articles collected from websites as proof for explainable claim verification. Atanasova et al. (2020) proposed generating summary sentences in a semi-supervised manner to help verdict prediction with textual justifications.

While this has been achieved, all the aforementioned approaches face one common problem. The explainability provided by each of these approaches either lacks integration within the pipeline process, is limited to just one layer of analysis, or relies on external knowledge graphs which may be difficult to maintain in real time. We are not aware of a prior approach that integrates evidence retrieval, memory-consistency cues, trend analysis, and credibility-score decomposition within a single explanation mechanism. This is where HMA-RAG steps in, providing a total of four explanation avenues.

Methods

The HMA-RAG framework utilizes a full seven-stage detection pipeline. Each input article goes through all stages sequentially; a confidence gate is used to determine whether the full verification path is invoked or not. This section explains each component in detail; it corresponds directly to the system architecture. As illustrated in Fig. 1, the framework processes each article through seven sequential stages from input preprocessing to memory update.

Input Preprocessing and Context Initialization

For a given input article from the news dataset, the system processes the text data by removing URLs, removing HTML tags, normalizing special characters, and consolidating white space. The text data is then reduced to a maximum of 1,000 characters. The short-term memory buffer is reset to zero, and the current context is initialized. The STM is initialized with the current article's data for future use in evidence storage.

Dual Branch Encoding Architecture

The architecture is composed of two different encoders that are separately trained for their respective tasks. The need for a dual-branch architecture stems from the inherent difference in the nature of the tasks. The classification model needs fine-tuned discriminative features that can differentiate between real and fake content.

DOI: 10.3844/jcssp.2026.2700.2710

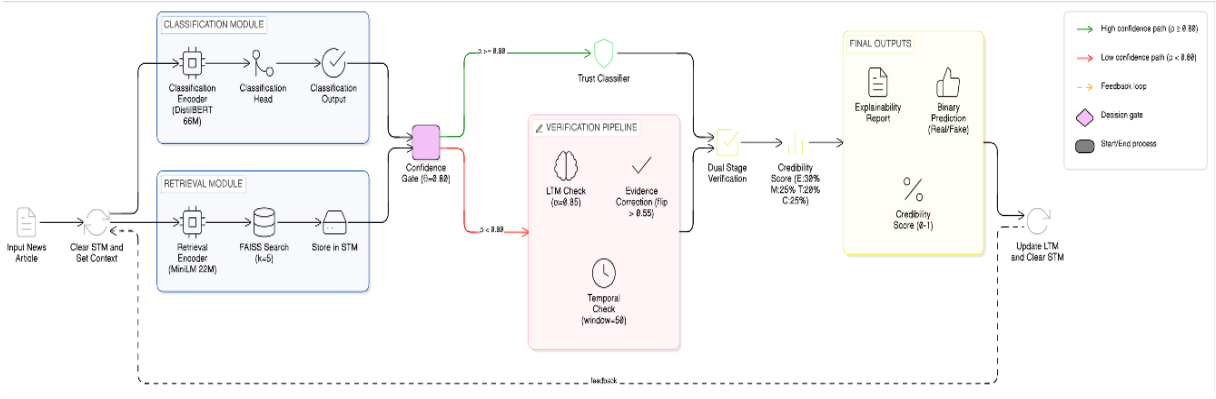

Fig. 1: System architecture diagram of the proposed HMA-RAG framework

On the other hand, the retrieval model requires general-purpose semantic similarity that can retrieve articles based on topics discussed in the content, irrespective of whether the content is real or fake.

Classification Branch. The classification model is a DistilBERT-base-uncased model consisting of 66 million parameters. The model is trained to produce contextualized token-level representations. The hidden layer dimension for the model is 768. The mean pooling aggregation method is applied over all the token-level representations. The single article-level representation is passed through a linear layer followed by a softmax classifier. The classifier predicts a binary output (0 for real and 1 for fake) and a confidence score rho between 0.5 and 1.0.

Retrieval Branch. The model selected for the retrieval model is all-MiniLM-L6-v2. The model is a sentence transformer model consisting of 22 million parameters. The model is trained to produce dense vectors. The model's dense vectors have a dimensionality of 384. The model was selected for its high performance in semantic similarity tasks compared to its compactness.

FAISS-Based Evidence Retrieval and STM Storage

We use a separate external knowledge base constructed using a 15% set of the overall training set. This set is composed of a truncated 500 character passage, a pre-computed MiniLM 384-d vector, and a label. This ensures there is no leakage between the evidence set and the classifier set.

For each input article, the FAISS library is used to retrieve the k most similar passages to the input document. This is done using the IndexFlatIP index type to perform an inner product search. The k value is set to 5. The document along with the similarity score is stored in the STM buffer. This buffer has a sliding window capacity of 20. The STM computes the label distribution over all the stored evidence. This provides a local consensus signal over the evidence. The evidence agreement signal is the proportion of the evidence retrieved that have the same label as the classifier's prediction.

Confidence Gated Routing

The confidence gate is a central routing component that decides whether the base classifier's prediction is accepted as-is or passed to the verification pipeline. Let $\rho(x)$ denote the classifier confidence for an input x, defined as the maximum softmax probability over the two classes, $\rho(x) = \max_c \text{softmax}(z)_c$, so that $\rho(x) \in [0.5, 1.0]$ for binary classification. The gate applies a fixed threshold $\theta = 0.80$. If $\rho(x) \geq \theta$, the article follows the high-confidence path and the classifier prediction is retained without modification. If $\rho(x) < \theta$, the article follows the low-confidence path, where the three correction mechanisms are applied sequentially. Formally, the routing decision is $g(x)$ = accept if $\rho(x) \geq \theta$, and verify otherwise.

The purpose of gating is to apply evidence- and memory-based correction selectively, in the region where the base model is least reliable. Fine-tuned transformer classifiers tend to be accurate when they are highly confident; unconditionally fusing their output with auxiliary signals therefore risks a known failure mode of correction and ensemble systems, namely overriding predictions that are confident and already correct. Restricting correction to low-confidence cases ($\rho(x) < \theta$) bounds this risk while concentrating the additional cost of retrieval and memory verification on the predictions most likely to benefit from it. The threshold $\theta = 0.80$ is a tunable hyperparameter; the value used here was selected on the validation set as a conservative operating point that preserved base-classifier accuracy on high-confidence predictions. As reported in the Confidence Gate Analysis, raising θ routes a larger fraction of predictions through the verification pipeline, trading additional computation for broader use of the memory and retrieval components. The gate thus serves three functions: It protects high-confidence predictions from unnecessary modification, it limits verification cost to the uncertain subset, and it exposes the high- and low-confidence routing counts as a diagnostic signal of model certainty over a dataset.

Multi-Source Verification Pipeline

For the low-confidence path, the system uses a multi-stage verification pipeline to verify the information contained in the articles. This pipeline comprises the following stages.

Dual Stage Verification

The module uses a combination of two sources to provide a more reliable defense against incorrect information. Stage 1 - Evidence Verification uses the agreement between the model's predictions and the classifier's prediction based on the evidence labels assigned. Stage 2 - Memory Verification combines the information provided by the STM consensus and the credibility of the sources provided by the LTM. When the evidence provided by the two stages contradicts each other, the system flags the information. In such a case, the original classifier prediction is maintained to avoid the risk of conflict.

Evidence Based Correction

When the evidence verdict provided by the Evidence Verification stage disagrees with the classifier's prediction and the evidence agreement score is higher than 0.55, the system flips the classifier's prediction to match the evidence verdict. This is the evidence flip mechanism.

Long Term Memory Consistency Check

The LTM stores verified facts over a long period of time. This information is stored along with a confidence score, which exponentially decays over time by a factor of $\alpha = 0.85$. If the score falls below a threshold $\delta = 0.75$, the information is pruned to avoid stale information being used to make current predictions. When the LTM identifies an inconsistency between the current prediction and the verified information over the long term, where the average similarity is above 0.75, the current prediction is flipped.

Temporal Pattern Correction

The temporal module tracks the distribution of Fake and Real predictions over a sliding window of the 50 most recent predictions. If the Fake distribution exceeds 70% (35 out of 50) of the most recent predictions and the current prediction is Real with classifier confidence below 0.70, the current prediction is flipped to Fake.

Credibility Score Computation

Regardless of the path taken by the article, the system summarises all available signals into a single credibility score $CS(x) \in [0, 1]$. A fixed-weight linear combination is used rather than a learned fusion so that each signal's contribution to the final score is directly readable; this transparency is what enables the score-decomposition explanation described in the Score Decomposition Layer. The score is computed as:

$$CS(x) = 0.30 \times S\_evidence + 0.25 \times S\_memory + 0.20 \times S\_temporal + 0.25 \times \rho$$

Each component is normalised to *[0, 1]*. S_evidence is the fraction of the k retrieved articles whose label agrees with the classifier prediction; S_memory combines the short-term-memory consensus over recent evidence with the long-term-memory source-credibility estimate; S_temporal is the agreement between the current prediction and the prevailing label distribution in the temporal window; and ρ is the classifier confidence. Because the weights ($\lambda_1=0.30$, $\lambda_2=0.25$, $\lambda_3=0.20$, $\lambda_4=0.25$) sum to one and every term lies in [0, 1], CS(x) is also bounded in [0, 1] and can be read as a weighted measure of agreement across the four signals. Evidence agreement receives the largest weight (0.30) because it provides the most direct external verification signal; classifier confidence and memory consistency each receive 0.25, reflecting their complementary internal and historical evidence; and the temporal component receives the smallest weight (0.20) because its usefulness depends on chronologically ordered input, which is not guaranteed in all deployment settings. The weights were selected on the validation set. An equal-weight configuration (0.25 each) was also evaluated and produced comparable F1-scores; the reported weighting is therefore one reasonable operating point rather than a uniquely optimal one, and learning the weights end-to-end is left to future work.

Memory Update and Feedback Loop

The last step in the prediction process is to update the long-term memory by incorporating the latest article's verified result, including the claim text, predicted label, classifier confidence, evidence summary, and the article's embedding vector. The feedback loop is completed by clearing the STM for the next article.

The training objective is given by:

$$L=-(1\vee|D|)\Sigma[y\log(\hat{y})+(1-y)\log(1-\hat{y})]$$

Where D is the training dataset and $\hat{y}$ is the predicted probability.

Experimental Setup

Dataset

The proposed HMA-RAG model is assessed on the WELFake dataset (Verma et al., 2021), a large-scale benchmark dataset for fake news detection. This dataset is a collection of news content from four prominent fact-checking sources: PolitiFact, GossipCop, BuzzFeed, and ISOT. This dataset comprises a large number of news articles, totaling 72,134. Among these, 35,028 are real news articles, while the rest are fake news articles.

The proposed model uses a stratified distribution of 70% of the data for training, 15% for validation, and the remaining 15% for testing. Additionally, a separate 15% of the entire training set is set aside to create the knowledge base using the FAISS library. Table 1 summarizes the statistical distribution of the WELFake dataset across real and fake articles.

Baseline Models

Logistic Regression: A linear classifier using TF-IDF features up to a maximum of 15,000 vocabulary size, removing English stop words. This model is a baseline that sets a floor on the performance of purely statistical lexical patterns without deep semantic representations.

Support Vector Machine: A linear SVM classifier using TF-IDF features up to a maximum of 15,000 vocabulary size, removing English stop words, and default regularization parameters. SVMs offer a strong baseline model that incorporates maximum margin optimization for decision boundaries that are maximally discriminative.

DistilBERT (no memory): A fine-tuned DistilBERT-base-uncased model with 66 million parameters, a linear classifier, and no memory or retrieval augmentation. This model offers a strong baseline that measures the performance of the transformer backbone alone.

Classifier Only (no memory): The DistilBERT classifier model with the confidence gating mechanism enabled but without memory augmentation or retrieval. This model measures the performance of the gating architecture alone.

Implementation Details

The experiment is performed on a workstation with an NVIDIA RTX 5070 GPU, 12 GB VRAM, an Intel Core i7 processor, and 32 GB RAM. The system is implemented in Python using PyTorch and the Hugging Face Transformers library.

Why DistilBERT over BERT-base? We have chosen DistilBERT with 66 million parameters over BERT-base with 110 million parameters for the classification encoder. This results in a 40% parameter reduction for 60% faster inference with a marginal accuracy trade-off of 0.5% on typical NLU benchmarks (Sanh et al., 2019). Table 2 shows the complete hyperparameter settings that were used in all of the tests.

Evaluation Metrics

We have used four classification metrics for evaluation by weighted averaging. These metrics include Accuracy, Precision, Recall, and F1-Score. The F1-score is adopted as the primary evaluation metric, consistent with standard practice in fake news detection tasks (Shu et al., 2017; Zhou and Zafarani, 2021). Table 3 shows how BERT-base and DistilBERT are built, which explains why we chose the lighter model.

Table 1: WELFake dataset statistics

Dataset	Total	Real	Fake	Train	Test
Welfake	72,134	35,028	37,106	50,493	10,820

Table 2: Hyperparameter configuration for HMA-RAG

Hyperparameter	VALUE
Classification encoder	DistilBERT-base-uncased (66M params)
Retrieval encoder	all-MiniLM-L6-v2 (22M params, 384-dim)
Learning rate	2×10^{-5} (AdamW)
Batch size	16
Training epochs	3
Max sequence length	512 tokens
Faiss retrieval (k)	5 nearest neighbors
Stm capacity (ns)	20 entries
Ltm decay factor (α)	0.85
Ltm pruning threshold (δ)	0.75
Confidence threshold (θ)	0.80
Temporal window size	50 predictions
Fake ratio alert threshold	35/50 (70%)
Evidence flip threshold	>0.55 agreement
Credibility weights (λ1,2,3,4)	0.30, 0.25, 0.20, 0.25

Table 3: Architectural comparison of BERT-base and DistilBERT

Property	Bert-base	Distilbert
Parameters	~110m	~66m
Transformer layers	12	6
Hidden dimension	768	768
Attention heads	12	12
Inference speed	1.0x	1.6x
Nlu retention	100%	~97%

Results and Analysis

Overall Performance Comparison

Table 4 shows how well HMA-RAG did compared to all the other baseline models on the WELFake dataset.

The results also bring out some key observations. Classical baselines have a performance gap of 4 to 5% compared to the transformer-based methods. The performance of the Logistic Regression is 94.15% F1-score, while the performance of the SVM is 94.87%. These are 4 to 5% lower than the performance of the DistilBERT baseline. This confirms the presence of complex linguistic patterns in the WELFake dataset that require deep contextual understanding beyond surface-level lexical features.

The best performance is achieved by the HMA-RAG for all metrics at 98.83% F1-score. The performance of the DistilBERT baseline is 98.82%. The full system exceeds the baseline by 0.01 percentage points. This difference is within the range of run-to-run variation and should not be interpreted as a meaningful accuracy gain; rather, it indicates that the augmentation modules maintain baseline accuracy without degradation. Preserving baseline accuracy is relevant for deployment, because auxiliary correction mechanisms can in principle destabilise a strong classifier. The main contribution of HMA-RAG is not an F1 gain over the DistilBERT baseline but the addition of structured, four-layer explainability while retaining competitive accuracy.

At an accuracy of roughly 98.8%, the absolute difference between 98.82% and 98.83% corresponds to a relative error reduction of approximately 0.85%. Given the small magnitude, we treat this as evidence of non-degradation rather than of improved discrimination.

Comparison With State of the Art

Table 5 compares HMA-RAG against published state-of-the-art methods in terms of parameters, dataset, accuracy, and explainability.

Improvement over the WELFake Baseline. The performance of the HMA-RAG is better compared to the original performance of the WELFake benchmark as reported in the paper by Verma et al. (2021) with a +2.10% margin. This is a demonstration of the advantage of the use of the combination of Retrieval-Augmentation and Hierarchical Memory over the use of feature engineering.

Competitive with heavier transformer models. The performance of the HMA-RAG is 98.83% with only 66 million classification parameters. This is within 0.02% of the most recent directly comparable performance on the WELFake benchmark. The most recent paper is the Hybrid CNN + BiLSTM + FastText paper (Merzah et al., 2026), which achieves an accuracy of 98.85% on the WELFake benchmark. The main difference is that the Hybrid CNN + BiLSTM + FastText model uses a substantially larger architecture and does not provide explainability. FakeBERT (Kaliyar et al., 2021) reports 98.90% accuracy with around 115 million parameters, but on the separate ISOT benchmark rather than WELFake; the two numbers are therefore not directly comparable, and we report the 0.07-percentage-point difference only as an indicative reference across datasets.

Efficiency versus scale. The difference between HMA-RAG and the GPT-2 Large model (Kuntur et al., 2024) is 0.89 percentage points. However, GPT-2 Large uses 774 million parameters, which is 11.7 times more parameters compared to HMA-RAG. Moreover, there is no level of explainability at all.

Unique Explainability Advantage. Among the compared systems, HMA-RAG is the only one that offers transparent, structured explanations for its predictions. All the compared models, regardless of the number of parameters or the accuracy level, remain black boxes that return a binary value with the option of a confidence score.

Table 4: Performance comparison on the WELFake dataset (%). Best F1-Score marked with *

Model	Acc.	Prec.	Rec.	F1
Logistic regression	94.15	94.17	94.15	94.15
Svm	94.88	94.89	94.88	94.87
Distilbert (no memory)	98.82	98.82	98.82	98.82
Hma-rag (full system)	98.83	98.83	98.83	98.83*
Classifier only (no memory)	98.82	98.82	98.82	98.82

Table 5: Comparison with published state-of-the-art methods. (†) Results reported on ISOT dataset. (‡) Results reported as mean ± std over 5-fold CV. (★) Classification encoder only; retrieval encoder adds 22M parameters

Model	param.	dataset	acc. (%)	expl.?
welfake baseline (Verma et al., 2021)	--	welfake	96.73	no
fakebert† (Kaliyar et al., 2021)	~115m	isot	98.90	no
Cnn + bilstm + fasttext‡ (Merzah et al., 2025)	large	welfake	98.85	no
gpt-2 large (Kuntur et al., 2024)	774m	welfake	99.72	no
hma-rag (ours)	66m★	welfake	98.83	yes (4)

In the case of HMA-RAG, the four types of explanation that the model offers are: Evidence Retrieval (ER), Dual Stage Verification (DSV), Temporal Context Tracking (TCT), and Credibility Score Decomposition (CSD).

Confidence Gate Analysis

The confidence-gate behaviour was analysed on the held-out evaluation set used for pipeline analysis (3,292 articles). Of these, 99.8% (3,284 of 3,292) passed through the high-confidence gate, where the classifier confidence was at or above 0.80. Only 8 results passed through the low-confidence verification pipeline. Out of the 8 uncertain results that passed through the full verification pipeline, the system attempted 4 corrections. Two out of these four corrections were correct; that is, the system correctly changed the erroneous prediction made by the classifier into the correct prediction. Two corrections failed. This analysis demonstrates that the fine-tuned DistilBERT encoder produces high-confidence predictions for the vast majority of the WELFake articles. The low activation rate (0.2%) reflects the high confidence of the base classifier on this dataset; the framework is designed so that the verification pipeline activates more frequently when the classifier is less certain. In deployment scenarios involving more ambiguous or adversarial content where classifier confidence would naturally be lower the verification pipeline would activate more frequently. The confidence threshold θ can be adjusted to control the trade-off between pipeline activation rate and computational cost: Higher thresholds (e.g., $\theta = 0.90$ or 0.95) would route more predictions through the full verification pipeline, enabling broader utilization of the memory and retrieval components.

Ablation Study

In order to measure the contribution of each memory component, a systematic ablation study is performed on the full HMA-RAG system by sequentially removing each component. All components, except for the removed one, are kept as they are in their full system configuration. Table 4 reports the ablation study results, showing the contribution of each individual component to the overall system performance.

LTM as the strongest individual component. The strongest individual component is the long-term memory module. The F1-score is reduced to 98.81% when the long-term memory module is removed, resulting in a reduction of 0.02%. This suggests that the long-term memory module's ability to track source credibility over time and perform memory consistency checking offers unique complementary information not provided by the other components. The exponential decay mechanism with $\alpha = 0.85$ and $\delta = 0.75$ for the pruning appears to be an effective mechanism for retaining discriminative patterns while discarding stale information.

Retrieval + STM is another contributing component. The F1-score is reduced to 98.82% when the FAISS retrieval mechanism and short-term memory are collectively removed. This is the same F1-score as the baseline DistilBERT model. This suggests that the retrieval mechanism for evidence collection does indeed provide an additional discriminative cue for uncertain predictions.

Temporal tracking is context dependent. The temporal tracking module does not affect system performance since the F1-score remains at 98.83% when the module is removed. The temporal tracking module is designed to work well for sequential or time-ordered processing scenarios. In a real-world deployment setting, temporal trends would likely have a much larger effect on system performance for detecting fake news.

Hierarchical complementarity. Another interesting observation is that the full system outperforms all the ablation versions of the individual components. This suggests that the individual components are complementing each other. The full system consistently achieves the best F1-score compared to all the ablation versions of the individual components.

Explainability Analysis

One of the key design considerations for HMA-RAG is to ensure explainability of the predictions made by the system. This addresses one of the key criticisms of deep learning-based detection systems: The black box effect. Contrary to all other baselines that only provide a binary output with an optional confidence score, HMA-RAG offers multiple-layered explainability via four different channels.

Evidence Layer

The evidence layer offers the top k=5 semantically similar articles found in the knowledge base, including their cosine similarity scores and corresponding ground-truth labels. The system offers an evidence summary for each prediction made, showing the number of articles that support each of the labels and their average similarity scores.

Verification Layer

The verification layer aggregates the results from the dual-stage verification module. Stage 1 presents the evidence agreement score calculated from the labels obtained from the retrieved documents. Stage 2 presents the memory consistency signal, which aggregates the STM consensus from recent predictions and the LTM credibility from the reliability of the sources. If there is a contradiction between the results from the two stages, the system will raise an alert indicating that the results from the two stages contradict each other.

Temporal Context Layer

The temporal context layer offers trend-aware explanation by reporting the distribution of fake and real classifications over the current window of 50 predictions. The layer can raise a temporal alert when the ratio of fake classifications rises to 70%, at which point the system changes from a reactive classifier to a proactive monitor.

Score Decomposition Layer

In the score decomposition layer, we weight and decompose the final credibility score into four variables (named Evidence, Memory, Temporal and Classifier Confidence) which each have a weight (30%, 25%, 20% and 25% respectively). This means not only that end-users can understand why they should trust particular outputs, but that in the event of mistakes these can be easily tracked back to their source.

Discussion

The Case for Intelligent Augmentation Over Larger Models

The present work demonstrates that smart design choices can help bridge the gap for smaller models. Using HMA-RAG along with DistilBERT (66 million parameters) results in 98.83% accuracy, just 0.07% behind FakeBERT (Kaliyar et al., 2021), which utilizes BERT-base along with CNN layers that have approximately 115 million parameters. This tiny difference comes at the cost of three major advantages that FakeBERT does not have: Explanations based on evidence for all predictions, memory-based verification that helps strengthen defenses against errors, and 40% fewer classification parameter weights for faster inference in resource-constrained environments.

Consider this in comparison to GPT-2, which manages 99.72% accuracy while utilizing 774 million parameters. In order to attain the remaining 0.89%, we must be willing to pay 11.7 times the parameter cost, which does not include any explainability features. In a real-world setting like journalism or fact-checking organizations or even content moderation websites, being able to explain the rationale behind flagging something as fake may be far more important than the last 0.07% accuracy gains that compromise explainability for a minor improvement in accuracy.

Confidence Gate as Architectural Innovation

The confidence gate ($\theta = 0.80$) serves three roles in the architecture. It preserves accuracy by preventing memory augmentation from altering high-confidence predictions of a strong classifier. It improves efficiency by invoking the retrieval and memory pipeline only for the low-confidence subset (0.2% of predictions on this dataset). And it provides a diagnostic signal, since the counts of predictions routed through each path give a summary of the model's confidence distribution over a dataset.

Limitations and Future Work

There are a few limitations to the current work that should be noted. First, the small margin of difference between DistilBERT and HMA-RAG, i.e., 0.01 percentage points, indicates that the WELFake dataset does not adequately test the memory augmentation component of the HMA-RAG model. The high confidence level of the base classifier with 99.8% of the predictions above the theta threshold of $\theta = 0.80$ does not allow the verification component to show its full corrective power. However, it is important to note that the primary contribution of HMA-RAG lies not in maximizing F1-score gains, but in maintaining competitive accuracy while providing structured, four-layer explainability that no compared system offers.

The temporal component did not show a significant impact on the model as the data is shuffled. The temporal component only works with the data that is in chronological order. Furthermore, the temporal tracking module employs a sliding window counter, which represents a basic form of temporal modeling. More sophisticated approaches such as time-aware attention mechanisms or temporal-aware recurrent architectures could strengthen this component and make the hierarchical memory claim more robust. The future work should include the application of the HMA-RAG with the data received in chronological order as news articles are received.

Additionally, the credibility score weights ($\lambda_1 = 0.30$, $\lambda_2 = 0.25$, $\lambda_3 = 0.20$, $\lambda_4 = 0.25$) were determined through validation-based experimentation. Future work could employ a learning-to-rank approach or end-to-end gradient-based optimization to determine the weight distribution in a more principled manner.

The current work has been carried out using the English language. The future work should include the application of the HMA-RAG using the Arabic language as well. This will enable the HMA-RAG model to meet the urgent need for the detection of fake news in the Arabic language.

Some further avenues for future research on HMA-RAG include: Adaptive threshold tuning: The threshold value theta should be tuned dynamically based on the characteristics of the data. Real-time retrieval augmentation: The knowledge base should be updated dynamically as new patterns of misinformation emerge. Stress testing on challenging datasets: The framework should be evaluated on datasets that force more frequent activation of the verification pipeline, such as adversarial or ambiguous news content requiring external cross-referencing. User studies: The practical utility of the four-

layer explainability framework should be evaluated with journalists and fact-checkers. Cross-lingual retrieval: The knowledge base should be able to handle data in more than one language.

Conclusion

In the above sections, we have introduced our Hierarchical Memory-Augmented Retrieval-Augmented Generation framework for fake news detection. The system consists of FAISS-based dense retrieval with all-MiniLM-L6-v2, a three-tier memory system consisting of short-term memory, long-term memory with exponential decay, and temporal tracking, dual-stage verification with fusion of evidence and memory signals, confidence-gated credibility scoring based on four components with corresponding weights, and a DistilBERT-based classification encoder with only 66 million parameters.

On the WELFake dataset (72,134 articles), the system achieves an F1-score of 98.83%, comparable to the DistilBERT baseline (98.82%) and to the BERT-based FakeBERT system (98.90%, reported on ISOT). The ablation study reveals that the Long-Term Memory module is the strongest individual module. The best performance is achieved by combining all the modules in a hierarchical manner.

Overall, the results suggest that a hierarchical memory-augmented retrieval design can retain the accuracy of a compact DistilBERT classifier while adding structured, four-layer explainability. Because the classification encoder uses fewer parameters than the larger baselines considered, the approach may also reduce inference cost, although we did not measure runtime directly. The framework offers a practical direction for explainable fake news detection.

Acknowledgment

The authors would like to thank Yarmouk University for providing the research environment and computational resources that supported this work.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author's Contributions

Afnan Altarjami: Conceived the research idea, designed the HMA-RAG framework, implemented the system, conducted all experiments, performed the ablation studies, and wrote the manuscript.

Sawsan Alshattnawi: supervised the research, critically reviewed the manuscript for intellectual content, and provided guidance on the research methodology and experimental design. All authors have read and approved the final manuscript.

Ethics

This study does not involve human subjects, animal experiments, or sensitive personal data. The WELFake dataset used in this study is publicly available and was used solely for academic research purposes. No ethical approval was required.

References

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, *10618*, 127–138. https://doi.org/10.1007/978-3-319-69155-8_9

Alshattnawi, S., Shatnawi, A., AlSobeh, A. M. R., & Magableh, A. A. (2024). Beyond Word-Based Model Embeddings: Contextualized Representations for Enhanced Social Media Spam Detection. *Applied Sciences*, *14*(6), 2254. https://doi.org/10.3390/app14062254

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating Fact Checking Explanations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/s1364-6613(00)01538-2

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. *ArXiv:1410.5401*. https://doi.org/10.48550/arXiv.1410.5401

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., & Qi, P. (2024). Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22105–22113. https://doi.org/10.1609/aaai.v38i20.30214

Johnson, J., Douze, M., & Jegou, H. (2021). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/tbdata.2019.2921572

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

Kuntur, S., Krzywda, M., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Comparative Analysis of Graph Neural Networks and Transformers for Robust Fake News Detection: A Verification and Reimplementation Study. *Electronics*, *13*(23), 4784. https://doi.org/10.3390/electronics13234784

Lewis, P., Perez, E., Piktus, A., Petroni, F., & Karpukhin, V. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, G., Lu, W., Zhang, W., Lian, D., & Lu, K. (2024). *Re-Search for The Truth: Multi-round retrieval-augmented large language models are strong fake news detectors*. https://doi.org/10.48550/arXiv.2403.09747

Liao, H., Peng, J., Huang, Z., Zhang, W., Li, G., Shu, K., & Xie, X. (2023). MUSER: A MUlti-Step Evidence Retrieval Enhancement Framework for Fake News Detection. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4461–4472. https://doi.org/10.1145/3580305.3599873

Merzah, B. M., Razmara, J., & Salmanian, Z. (2026). Hybrid deep learning models for fake news detection: case study on Arabic and English languages. *Frontiers in Big Data*, *8*, 3786. https://doi.org/10.3389/fdata.2025.1683786

Niu, C., Guan, Y., Wu, Y., Zhu, J., Song, J., Zhong, R., & Zhu, K. (2024). *VeraCT Scan: Retrieval-Augmented Fake News Detection with Justifiable Reasoning*. https://doi.org/10.18653/v1/2024.acl-demos.25

Perez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. *Proceedings of the 27th International Conference on Computational Linguistics*, 3391–3401.

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). *DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d18-1003

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. https://doi.org/10.48550/arXiv.1910.01108

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. *Advances in Neural Information Processing Systems*, 2440–2448. https://doi.org/10.48550/arXiv.1503.08895

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). *FEVER: a Large-scale Dataset for Fact Extraction and VERification*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). https://doi.org/10.18653/v1/n18-1074

Tulving, E. (1972). Episodic and semantic memory. *Organization of Memory*, *12*, 381–403.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Verma, P. K., Agrawal, P., Amorim, I., & Prodan, R. (2021). WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems*, *8*(4), 881–893. https://doi.org/10.1109/tcss.2021.3068519

Vosoughi, S., Roy, D., & Aral, S. (2018). *The spread of true and false news online*. Science. https://doi.org/10.1126/science.aap9559

Wang, W. Y. (2017). *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). https://doi.org/10.18653/v1/p17-2067

Weston, J., Chopra, S., & Bordes, A. (2015). Memory networks. *Proceedings of the International Conference on Learning Representations*, 1–11. https://doi.org/10.48550/arXiv.1410.3916

Zhou, X., & Zafarani, R. (2021). A Survey of Fake News. *ACM Computing Surveys*, *53*(5), 1–40. . https://doi.org/10.1145/3395046