Interpretable RAG for Fact-Checking: Comparing LLMs and Fine-Tuning Methods

Saurav, Lahari, Savithri, Taniksha

Graduate Students, Information Science
University of Arizona

Abstract—We introduce an interpretable, retrieval-augmented fact-checking system that achieves strong performance and transparency on the FEVER dataset. Our pipeline integrates **multiple retriever types** (dense, multi-query, hybrid) and evaluates both proprietary (GPT-4.1, GPT-4o, 4.1 mini, 3.5 Turbo) and open-source **LLMs** (Mistral-7B, Qwen-7B). Utilizing Parameter-Efficient Fine-Tuning (**PEFT**/LoRA), we elevate **Mistral-7B accuracy from ~56% to 74%** with limited training. Notably, we explore Direct Preference Optimization (**DPO**) within the RAG framework for fact-checking, a relatively novel application, demonstrating its potential to enhance alignment with human judgments. Our experiments show GPT-4o achieving the highest fact-checking accuracy (~80%), while fine-tuned 7B models approach this performance with significantly lower resource demands. **Extensive ablation studies** compare retriever configurations and performance, fine-tuning methods, prompt sensitivities, and RAG versus non-RAG setups and more. To enhance interpretability, we incorporate **Explainable AI** (XAI) tools (**Captum**, **LIME**, **SHAP**) that elucidate token-level contributions to model decisions, though integration challenges persist for certain models with some of these tools. We evaluate using standard metrics and integrate the state-or-art **RAGAS** framework (faithfulness, relevancy, and others) to affirm the effectiveness of our approach. Overall, our study demonstrates that a retrieval-enhanced, explainable fact-checking system can combine accuracy and transparency with model choice, fine-tuning, and retrieval design - each playing a vital role.

The full implementation and supporting notebooks are publicly available at:
https://github.com/sauravverma78/InterpretableRAGFactCheck

Index Terms - RAG, PEFT, DPO, XAI, LLM, LORA, RLHF, SHAP, LIME, ICL, RAGAS

I. INTRODUCTION

The rapid spread of misinformation poses a significant challenge in the digital age, necessitating robust and automated fact-checking mechanisms. While Large Language Models (LLMs) show promise in natural language understanding, they often lack grounding in factual evidence and can hallucinate. This motivates the use of Retrieval-Augmented Generation (RAG), which couples an LLM with a retrieval mechanism that supplies relevant evidence from an external knowledge source to both improve factual accuracy and reduce hallucinations.

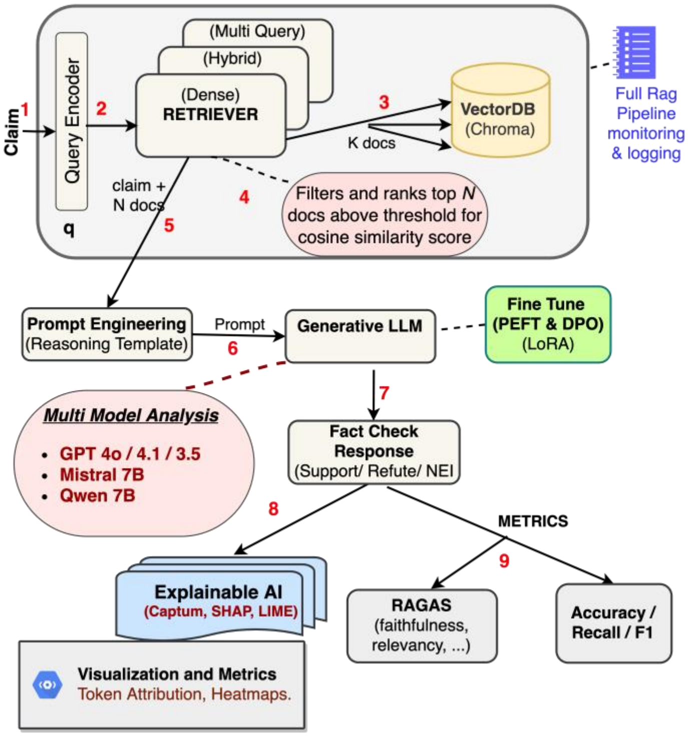


Fig 1 Modular Interpretable RAG Pipeline Architecture

However, the "black-box" nature of LLMs hinders trust and adoption. In fact-checking, the system's decision process must be transparent and should indicate *why* a claim was judged true or false, ideally by pointing to evidence. Explainable AI (XAI) techniques offer a pathway to understanding and interpreting model predictions. Prior works have explored providing natural language justifications or highlighting relevant evidence sentences to explain model predictions [2]. In **our work**, we leverage token-level attribution methods to shed light on which parts of the input (claim and evidence) most influenced the model's verdict. We also **build** full **logging aspect of RAG pipeline** which logs retrieved documents, filtered documents on threshold, LLM prompt, LLM response and its reasoning.

Beyond building the pipeline, we explore the performance of various LLMs and fine-tuning strategies. While proprietary models like GPT-4 excel, they are costly and opaque. We investigate if smaller, open-source models like Mistral-7B, when fine-tuned on fact verification tasks using Parameter-Efficient Fine-Tuning (PEFT) techniques such as LoRA and QLoRA, can achieve comparable results. Additionally, we experiment with Direct Preference Optimization (DPO), an alignment method that trains models using pairwise preference data rather than traditional labels. A high-level overview of the proposed fact-checking pipeline is shown in Fig [1], illustrating the retrieval, reasoning, and explanation components.

Our key contributions are as follows:

- We develop an end-to-end RAG-based fact-checking system evaluated on the FEVER (1.0) dataset, with **integrated explainability** using Captum, SHAP, LIME.
- We conduct a **comparative evaluation of multiple LLMs** (GPT-4.1, 4.1 mini, 4o, 3.5, Mistral-7B, Qwen-7B) under identical RAG settings.
- We **fine-tune Mistral-7B using PEFT** (LoRA/QLoRA) and demonstrate significant claim verification accuracy improvements from **56% to 73.5%.**
- We explore **DPO fine-tuning** for fact-checking, providing one of the earliest experimental insights into its effectiveness in this domain.
- We integrate RAG pipeline with state-of-art **RAGAS** for additional metrics like Faithfulness, answer relevancy.
- We perform **ablation studies** on retriever types, retriever threshold cutoffs, asynchronous retrieval, and prompt engineering strategies, analyzing their impact on verification performance and explainability.

II. RELATED WORK

Automated Fact-Checking: The task of claim verification has been extensively studied in NLP. The FEVER dataset [11] introduced a benchmark for fact-checking, consisting of claims paired with evidence from Wikipedia and labelled as Supported, Refuted, or Not Enough Information. Early approaches [2] used IR techniques (TF-IDF or BM25) for retrieval and then a neural classifier for verification. Subsequent research improved retrieval via neural embeddings and explored generating human-readable explanations [18], [27]. Atanasova et al. [19] proposed ranking evidence and producing textual justifications for the prediction. FEVER-KILT [26], [29] benchmarks the state of the art, with Re2G [28] currently topping the leader board at 89% accuracy. In our study, we achieved 80% accuracy using (un-tuned) GPT-4o within a RAG-based system, showing competitive performance with emphasis on transparency and modularity. Tan et al. [20] present an explainable fact-checking system (CorXFact) that explicitly models claim-evidence correlations, while Krishna et al. [26] design a seq2seq model to generate natural logic-based inferences as proofs. **Our work** aligns with this trend by focusing on explanation, but instead of generating rationale text, **we highlight** which parts of the given evidence influenced the model providing a distinct and interpretable layer of transparency.

Retrieval-Augmented Generation (RAG): Lewis *et al.* [1] introduced RAG for open-domain question answering, showing that retrieving text from Wikipedia and feeding it into a model significantly improved answer accuracy. In fact-checking, RAG is a natural fit. Recent works have evaluated RAG pipelines in the fact-checking context. Khaliq *et al.* [3] propose RAGAR, a RAG-augmented reasoning approach for political fact-checking that integrates multimodal evidence. They found that augmenting an LLM with retrieval and additional reasoning steps improved accuracy on political claims. He et al. [22] introduce CoV-RAG, a chain-of-verification approach that enhances RAG by correcting retrieval and generation errors through query revision, improving performance. In **our work**, we use RAG as the backbone and specifically analyze different retriever configurations (dense, multi-query, hybrid) to understand their impact on evidence quality. We also integrate with the **RAGAS** framework by Es *et al.* [21] for evaluation, which provides metrics like faithfulness, and answer relevancy for RAG. This allows us to assess not just accuracy, but also how faithfully the LLM uses the retrieved evidence.

Explainability for LLMs and RAG: To explain the outputs of neural networks is an active research area and common approaches include **LIME** [8] and **SHAP** [7]. **Captum's** LLM attribution [17] module allows computing token-level importance for transformer models by treating them as sequence classifiers. Prior fact-checking systems, particularly those leveraging Retrieval-Augmented Generation (RAG) on datasets like FEVER, often provide explanations by outputting a single justification sentence or a subset of retrieved evidence to support predictions such as "SUPPORTED" or "REFUTED" [23], [24]. In contrast, **our approach** employs Captum, LIME and SHAP, to **probe the internal decision-making** of large language models (LLMs), identifying specific tokens and evidence segments that drive classification outcomes. This fine-grained **token-level attribution** complements higher-level explanation generation. However, our analysis reveals limitations: LLM attributions can occasionally overemphasize irrelevant tokens, such as formatting markers or high-frequency words, due to biases in learned representations.

Parameter-Efficient Fine-Tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) [5], enable efficient adaptation of large language models (LLMs) by injecting trainable low-rank matrices into frozen model layers, reducing computational costs for domain-specific tasks. **We apply LoRA** with 4-bit quantization (QLoRA) to fine-tune Mistral-7B on 1,000 FEVER claim-evidence pairs, achieving a **17% accuracy improvement**, consistent with PEFT's success in resource-constrained settings. Similarly, Direct Preference Optimization (DPO) [6], a streamlined alternative to reinforcement learning with human feedback (RLHF), aligns LLMs by optimizing preferences for correct outputs. **We explore DPO** atop SFT, following recent applications of DPO in knowledge-intensive tasks [14]. These methods highlight the potential of PEFT and DPO to customize LLMs for fact-checking.

III. Problem Definition and Methodology

A. Problem Definition

We address the task of evidence-based claim verification. Given a textual claim *C* and access to a corpus *D* of evidence sentences, the goal is to assign a label

$$y \in \{\text{Supports}, \text{Refutes}, \text{Not Enough Info}\} \qquad (1)$$

indicating the veracity of *C* based on available evidence. If the label is Supports or Refutes, some sentences in *D* must substantiate the decision; if Not Enough Info (NEI), no conclusive evidence exists. This corresponds to the FEVER dataset.

B. Methodology

Our methodology revolves around a RAG pipeline combined with XAI and fine-tuning, implemented primarily using LangChain, Hugging Face Transformers, PyTorch, and associated libraries.

The workflow comprises:

1. **Dataset Preparation:** Utilize the FEVER (1.0) dataset, specifically the gold evidence sentences. Pre-processing involves cleaning text data to remove noise and ensure consistency before loading into Vector DB.
2. **Knowledge Base Indexing:** As shown in Fig [2], embedding the cleaned evidence sentences using a sentence transformer model (BAAI/bge-m3) and indexing them into FAISS backed Chroma vector store for efficient retrieval.
3. **Retrieval:** Implementing and comparing different retrieval strategies to fetch relevant evidence for a given claim:
   - **Dense Retriever (Single Query):** Encodes the claim and retrieves top-k nearest neighbours based on cosine similarity.
   - **Multi-Query Dense Retriever:** Generates multiple paraphrased queries using an LLM, retrieves candidates for each, and merges results to enhance recall.
   - **Hybrid Retriever (Dense + Sparse):** Combines BM25 lexical search with dense retrieval to improve coverage, particularly for named entities and keywords

   A threshold (typically τ=0.6) is applied post-retrieval to filter irrelevant documents.
4. **Prompt Styles**: We experiment with two prompt styles: a simple direct-label prompt and an elaborate prompt requesting explanations followed by a label. The prompt specific to each LLM is carefully crafted.
5. **Generation/Classification:** Feeding the retrieved evidence and the claim into various LLMs (Mistral-7B, Qwen2-7B, GPT series, fine-tuned Mistral) prompted to perform the three-way classification. Strict prompting was used, especially for GPT series, to encourage caution and minimize guessing.
6. **Fine-Tuning:** We apply PEFT (LoRA/QLoRA) and DPO techniques to adapt the Mistral-7B model specifically for the FEVER fact-checking task, using a subset of the training data. We use ***few-shot In context learning (ICL)*** for LLM training prompt
7. **Evaluation:**
   - **Accuracy:** Evaluate using classification accuracy and confusion matrices.
   - **RAGAS:** Additionally, RAGAS metrics [Context Recall, Faithfulness, Answer Correctness, Response Relevancy, Factual Correctness] are computed.
8. **Attribution & Explainability:** Integrating XAI tools (Captum, SHAP, LIME) to generate token-level attributions for open source models.
9. **Ablation Studies:** Systematically evaluating the impact of different components: retriever types, retrieval thresholds, fine-tuning methods, and other pipeline components.

The entire pipeline incorporates verbose logging for transparency and debugging. Refer [Appendix E] for example

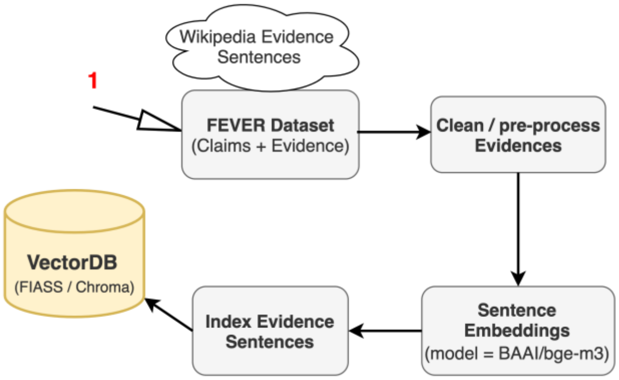

Wikipedia Evidence Sentences
1
FEVER Dataset (Claims + Evidence)
Clean / pre-process Evidences
VectorDB (FIASS / Chroma)
Index Evidence Sentences
Sentence Embeddings (model = BAAI/bge-m3)

Fig 2 Data preparation and store

C. Label Distribution.

Initial analysis shows a relatively balanced distribution among the labels in the validation subsets used.

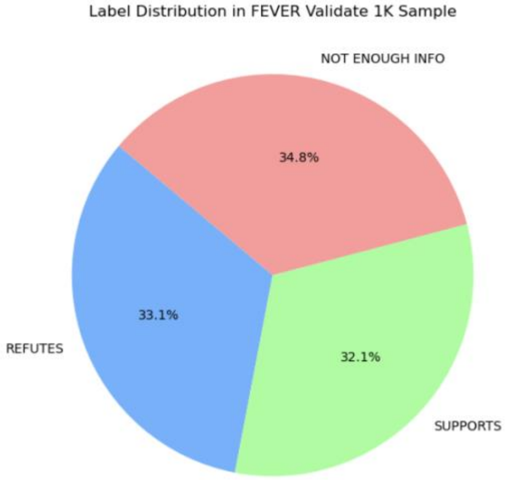


Fig 3: Validation dataset label distribution

D. Mathematical Pipeline

Given a claim embedding $q \in R^d$, we retrieve the top - k documents from the vector database.

$$D_{retrieved} = \{d_1, d_2, \ldots, d_k\} \qquad (2)$$

We then filter the retrieved documents based on cosine similarity with threshold r.

$$D_{filtered} = \{d_i \in \ D_{retrieved} \mid cosine\ (q, d_i) \geq r \} \qquad (3)$$

A prompt P is constructed by combining the original claim text with the text of the filtered documents:

$$P \ = PromptTemplate\{\ q_{text}\ , (d_{i,text})\ d_i \in D_{filtered}\} \quad (4)$$

Finally, the prompt is passed to a generative model M to predict the fact verification label

$$y = M\ (P) \qquad (5)$$

IV. MODEL ARCHITECTURE AND EXPERIMENTAL SETUP

A. High level architecture

The high-level architecture of our RAG-based system is illustrated in Fig [1]. For sentence embeddings, we employ pre-trained transformers that output 768-dimensional vectors, BAAI/bge-m3-large which is optimized for search tasks.

The evidence corpus is indexed using FAISS with HNSW (Hierarchical Navigable Small World) graph indexing for efficient cosine similarity search. All vector indexing is performed using the Chroma vector database.

Following **LLMs were evaluated within the RAG** pipeline:

- Mistral-7B-Instruct-v0.2 : An open-source 7B parameter model. It demonstrates good understanding but needed structured prompting for factual accuracy [https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2]
- Mistral-7B-Instruct-v0.2 (PEFT-tuned on FEVER)
- Mistral-7B-Instruct-v0.2 (DPO-tuned on FEVER)
- Qwen2-7B-Instruct: A multilingual instruction-tuned model by Alibaba. Performs quite well on fact-checking without fine-tuning. [https://huggingface.co/Qwen/Qwen2-7B-Instruct]
- OpenAI GPT-3.5-Turbo-0125
- OpenAI GPT-4o
- OpenAI GPT-4.1
- OpenAI GPT-4.1-mini

B. Fine-Tuning with LoRA (Supervised)

We fine-tune Mistral-7B using QLoRA via **PEFT** and transformers libraries. LoRA settings are: rank r=32 & alpha=32, dropout = 0.1. The base model weights remain frozen except for layer norms, following QLoRA standards.
We **use few-shot ICL examples in the prompt** [Appendix H] while training and that results in improvement of resultant model as against, without using it. Training is performed on 1,000 claim-evidence-label examples over 5 epochs, using:

- Batch size: 4
- Learning rate: 5×10^{-4}
- Optimizer: AdamW with weight decay = 0.01
- Mixed precision (bf16)
- No gradient accumulation and no evaluation due to GPU resource constraints

The model converged within 2 epochs, peaking at ~73.5% in the 5th epoch. The fully merged tuned model is loaded to HuggingFace for reproducibility. [https://huggingface.co/sauravverma78/mistral7bInstruct02_fever_trained]

C. Fine-Tuning with DPO (Preference-based)

We also fine-tune Mistral using **Direct Preference Optimization (DPO)** from the TRL library. For this, we generate ~1,000 response pairs: each with a "chosen" (correct) and "rejected" (plausible but incorrect) output for the same prompt. The incorrect responses are LLM-generated to mimic common errors.

We apply DPO loss using the same hyperparameters as LoRA tuning (5×10^{-4} LR, 2 epochs), training on the same hardware setup. This allows us to compare supervised fine-tuning with preference-based alignment under controlled conditions. Tuned model is available in HuggingFace [https://huggingface.co/sauravverma78/mistral7bInstruct02_fever_dpo].

D. Experimental setup and Libraries used

We conduct **Experiments** using two main environments:

- **Local:** Macbook Pro M3 with 128GB RAM and 40-core Neural Engine. Used for running models (Mistral, Qwen2) and initial prototyping.
- **Cloud:** Google Colab PRO (paid) with A100 GPU. Used for computationally intensive tasks like model fine-tuning (PEFT/DPO) and running RAG pipelines with OpenAI models and RAGAS computations

Note : Due to constraints in accessing advanced GPU environments and associated compute costs, certain experiments—such as full-scale fine-tuning and large-batch metric evaluations—are conducted on a reduced dataset. This limitation may have an impact on the achievable upper bound of model performance and ablation depth. Moreover GPT models are not fine-tuned for the cost issues.

Frameworks and Libraries

The project leveraged a range of Python libraries:

- Core-ML/NLP: torch, transformers, accelerate, bitsandbytes, sentence-transformers.
- Fine-Tuning: peft, trl
- RAG & Vector Store: langchain, chromadb, faiss-gpu
- Evaluation: scikit-learn, ragas
- XAI: captum, lime, shap, ,lit
- Data Handling: datasets, numpy
- Utilities: tqdm, psutil, asyncio, nest_asyncio, openai, tiktoken, matplotlib, seaborn

V. RESULTS AND COMPARATIVE EVALUATION

A. Overall Results

Model performance is evaluated primarily based on classification accuracy on a held-out validation set (400 - 500 samples). Note: Open-AI model sizes are not published.

Model (size)	Accuracy	F1	Recall		
			Support	Refute	NEI
Mistral (7B)	0.56	0.48	0.97	0.07	0.64
QWEN (7B)	0.70	0.66	0.92	0.85	0.31
GPT 4.1 (?)	0.79	0.78	0.88	0.86	0.61
GPT 4o (?)	**0.80**	**0.79**	0.91	0.92	0.59
GPT 4.1 mini (?)	0.77	0.76	0.89	**0.93**	0.50
GPT 3.5 Turbo (175B)	0.59	0.56	0.96	0.53	0.29
Mistral Peft tuned (5M)	**0.74**	0.74	0.81	0.66	**0.74**

Table 1 Comparative evaluation of LLMs under the RAG framework

Table [1] shows that GPT-4o and GPT-4.1 led in overall performance (Accuracy ≈ 80), showing high recall across all classes. The base Mistral-7B model performs poorly, especially on the Refutes class (recall = 0.07), indicating over-reliance on the Supports label. **PEFT-tuning** significantly improves Mistral's performance-raising accuracy and F1 to 0.74, and notably achieves the **highest NEI recall (0.74)** among all models. This highlights the effectiveness of task-specific fine-tuning in reducing hallucinations and enhancing the model's ability to abstain when insufficient evidence. The untuned Qwen-7B model also performs strongly with 70% accuracy Qwen-7B improves REFUTES recall but struggles with NEI classification.

Overall, while GPT models dominate in raw performance, the **PEFT-tuned Mistral** stands out for its NEI classification ability and balanced performance, key traits for building explainable and trustworthy fact verification systems.

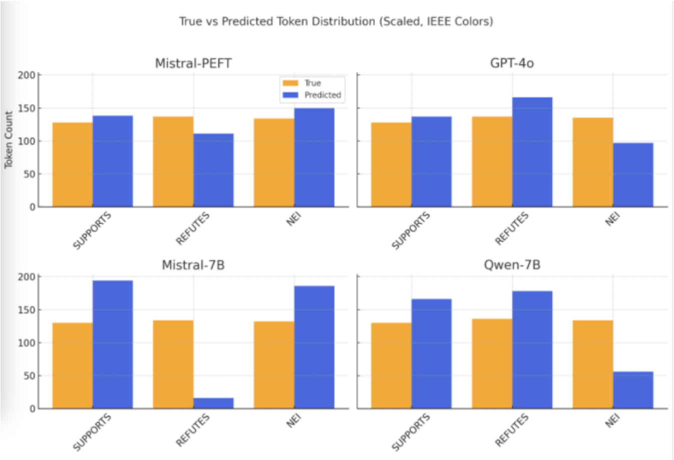


Fig 4: Predicted Vs Actual label distribution across models

Fig. [4] presents a side-by-side comparison of true vs predicted class distributions for four LLMs. Mistral-PEFT tuned model not only heavily improves on base Mistral model but it even surpasses GPT-4o to demonstrate well balanced predictions across all three classes. Qwen-7B shows improved REFUTES alignment but misclassifies many NEI instances. These results highlight the effectiveness of PEFT tuning and the strength of larger LLMs in maintaining balanced label predictions

B. RAGAS evaluation metrics

RAGAS metrics are computed for base Mistral and Qwen2 models on a 25-sample subset to provide qualitative insights beyond simple accuracy. Due to resource constraints, only faithfulness and answer relevancy are calculated.

Model	Faithfulness	Answer Relevancy
Mistral RAG	0.80	0.56
QWEN RAG	0.83	0.82

Table 2 : RAGAS metrics for open source models

Qwen RAG achieves the **highest scores on both faithfulness (0.83)** and **answer relevancy (0.82)**, indicating that it not only retrieves contextually appropriate evidence but also generates responses that align well with the claim. In comparison, **Mistral** RAG shows slightly lower faithfulness (0.80) and a notably **weaker answer relevancy (0.56)**. This suggests that while Mistral is somewhat grounded in evidence, its responses are less focused or coherent in relation to the claim. Overall, Qwen demonstrates a more consistent and reliable generation behavior under the RAG framework.

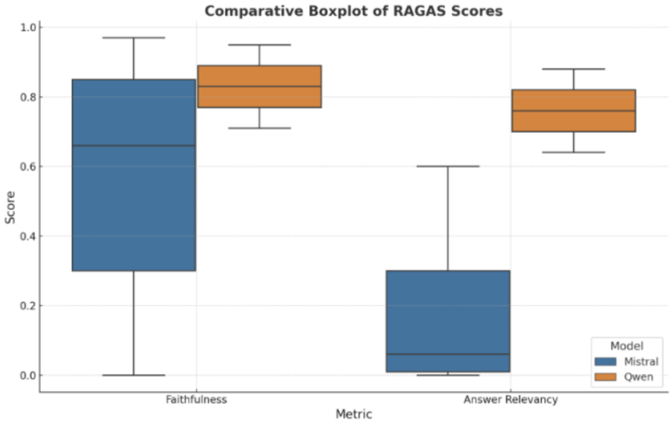


Fig 5: RAGAs metrics compare for Mistral Vs Qwen RAG pipelines

C. PEFT and DPO Fine-Tuning

To evaluate the effect of tuning strategies, we compare Mistral in four configurations: base, PEFT (LoRA), DPO on base, and DPO on PEFT. Trained on ~1,000 samples. PEFT improves accuracy to 74% (from 56%) and NEI recall to 0.74. DPO on top of PEFT retained accuracy (70%) and shows smoother convergence. DPO on base gives marginal gains (60%)

These experiments validate the power of PEFT for low-resource fine-tuning. Achieving a **17%** gain with *just 1,000 samples* demonstrates strong task-specific adaptation. Notably, **this study is among the first** to explore **DPO within a RAG**

pipeline for classification, showing it retains benefits when layered on SFT but offers little when used alone. Loss curves as in Fig [6], further support this: DPO exhibits faster, more stable convergence over PEFT, highlighting its alignment efficiency when paired with prior supervision.

Model	Accuracy	F1	Recall		
			Support	Refute	NEI
Base model	0.56	0.48	0.97	0.07	0.64
PEFT tuned	0.74	0.74	0.81	0.66	0.74
DPO on Peft Tuned	0.70	0.71	0.79	0.62	0.69
DPO on base model	0.60	0.60	0.65	0.33	0.70

Table 3 : Model performance comparison after Fine-Tuning

Both tuning approaches are conducted on modest compute (Colab A100 40GB GPU) without evaluation during training due to GPU memory limits. Our study validates the feasibility of parameter-efficient fine-tuning even for mid-sized open LLMs

.

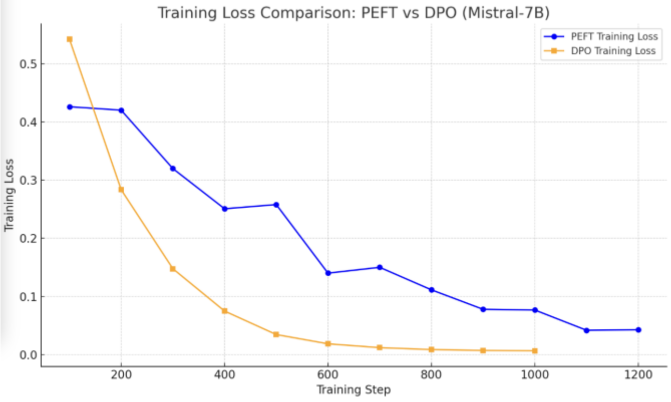
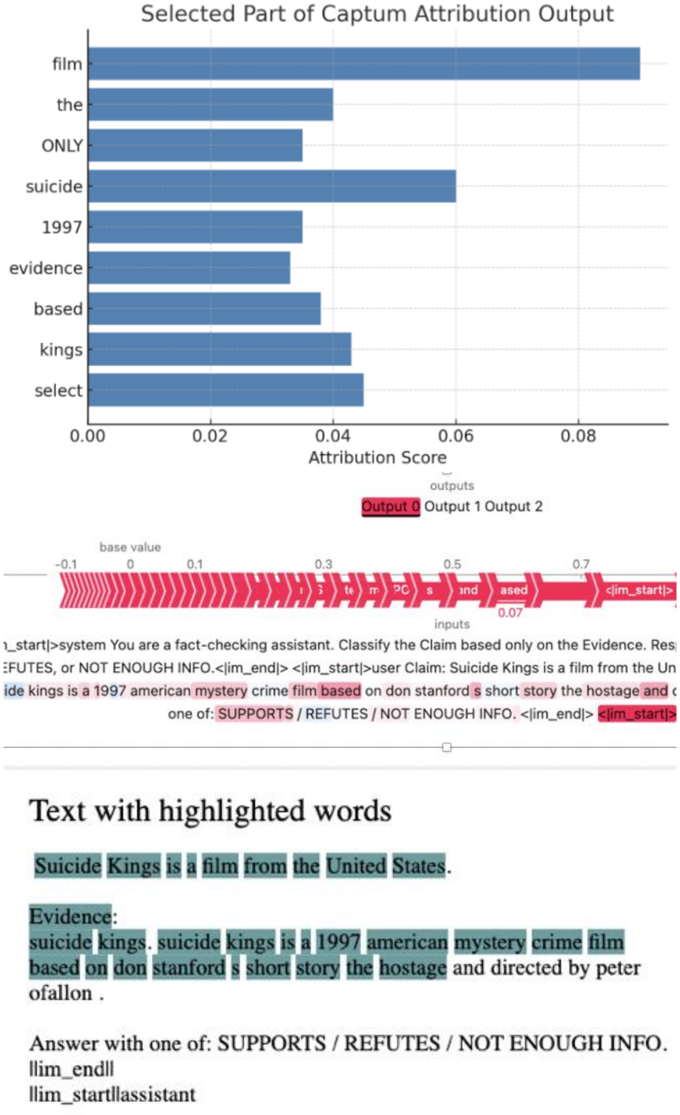


Fig 6: Training Loss Comparison PEFT vs DPO tuning on Mistral

D. ***Explainability*** *Analysis with Captum, SHAP and LIME*

To improve transparency and interpretability in our RAG-based fact verification system, we integrate several XAI frameworks like Captum (Integrated Gradients), SHAP, and LIME for token-level attribution**.** We also experiment with Soft XAI. We apply these methods to both Mistral and Qwen models on the same sample:

Claim: Suicide Kings is a film from the United States.
Evidence: Suicide Kings is a 1997 American mystery crime film based on Don Stanford's short story...
Ground Truth and Model Answer: SUPPORTS



ı_start|>system You are a fact-checking assistant. Classify the Claim based only on the Evidence. Res
:FUTES, or NOT ENOUGH INFO.<|im_end|> <|im_start|>user Claim: Suicide Kings is a film from the Un
ide kings is a 1997 american mystery crime film based on don stanford s short story the hostage and c
one of: SUPPORTS / REFUTES / NOT ENOUGH INFO. <|im_end|> <|im_start|>

Text with highlighted words

Suicide Kings is a film from the United States.

Evidence:
suicide kings. suicide kings is a 1997 american mystery crime film based on don stanford s short story the hostage and directed by peter ofallon .

Answer with one of: SUPPORTS / REFUTES / NOT ENOUGH INFO.
|im_end|
|im_start|assistant

Fig 7: (Top) Mistral-Captum Token Level Attributions; (Middle) Qwen SHAP visualization and scores for top contributors; (Bottom) Qwen : LIME visualization for critical word spans

Explainability Insights

- **Captum** (Mistral): Captum attribution scores reveal "film," "suicide," "kings," and "1997" as dominant tokens influencing Mistral's prediction. It works well with decoder-only models like Mistral, though interpretability is limited by context size and integration steps.
- **SHAP** (Qwen): SHAP analysis highlights "film," "American," and "mystery" as strong contributors to the SUPPORTS label. SHAP's token-level contributions aligns with the model's output confidence. It could not be used with Mistral possibly due to lack of explicit class logits in generative outputs.
- **LIME** (Qwen/Mistral): LIME emphasizes critical spans in both claim and evidence ("1997 American mystery crime film"). It ignores instruction boilerplate and focusses on semantically relevant text, demonstrating faithful local attribution.

- **Soft XAI / Prompt-Based Rationales**: Adding a reasoning prompt (“Explain your reasoning before answering”) results in Qwen and Mistral models’ improved interpretability. Outputs included quoted evidence, aligning with attribution scores. Refer [Appendix D] for examples.

Conclusion: Captum is best suited for models like Mistral. SHAP and LIME provide stronger visual explanations for Qwen. Prompt-based soft XAI provides complementary transparency, often surfaces rationale snippets aligned with attributions – good as a lightweight attribution layer.

XAI analysis summary:

Model	XAI Method	Compatibliity	Key Insights
Mistral	Captum	✓	Highlights claim and evidence tokens (e.g., 'film', '1997')
Mistral	SHAP	X	Decoder-only model limits SHAP application
Mistral	LIME	✓	Highlights task-relevant spans (claim, year, genre terms)
Qwen	Captum	✓	Highlights semantically meaningful words
Qwen	SHAP	✓	Provides token-wise relevance using color bands
Qwen	LIME	✓	Shows precise word influence; ignores filler text

Table 4 : XAI Compatibility Summary Table, comparing Captum, SHAP, and LIME across Mistral-7B and Qwen-7B

Note: These XAI visualizations are based on a representative claim-evidence pair. While only one example is presented, similar attribution trends are observed across multiple samples. Additionally, XAI tools such as Captum, SHAP, and LIME require substantial computation time and memory, especially on large language models, and may limit their practicality for batch-scale analysis without proper tuning.

VI. ABLATION STUDIES

A. RAG performance by Single Vs Multi Query Retriever

We compare single vs multi-query retrievers on a very small sample, Fig. [8], on two LLMs using the FEVER dataset. **Mistral-7B** shows negligible gains (F1: 0.36 → 0.35), suggesting limited benefit for smaller models. In contrast, **GPT-3.5 Turbo** improves notably (Accuracy: 0.70 → 0.80, F1: 0.67 → 0.79), indicating that multi-query retrieval enhances performance by enriching evidence diversity. [Appendix D]

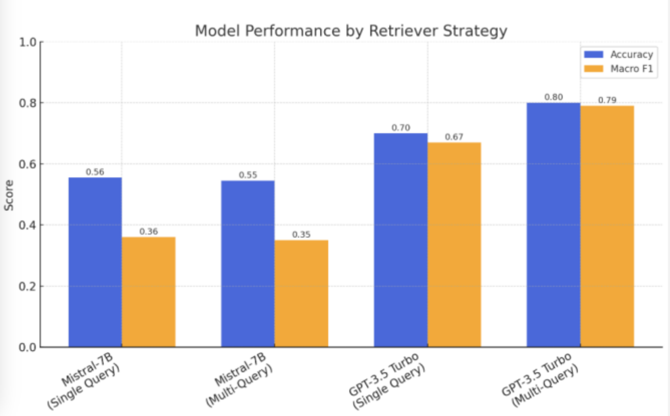


Fig 8: RAG Performance by retriever strategy

B. Retrieval Threshold and Document Quality

Retrieval Cutoff	Faithfulness	Answer Relevancy
0.6	0.80	0.56
0.3	0.88	0.31

Table 5 : RAGAS metrics for retrieval threshold cutoff on Mistral 7B

We evaluate the impact of retrieval cutoff thresholds on Mistral RAG pipeline performance using RAGAS metrics as in [Table 5]. Lowering the threshold from 0.6 to 0.3 slightly increases faithfulness but decreases average answer relevancy. This reveals a precision-recall trade-off in retrieval, i.e. more documents ≠ better answers. Threshold tuning is critical for stable answer quality.

C. Prompt [Engineering] Sensitivity

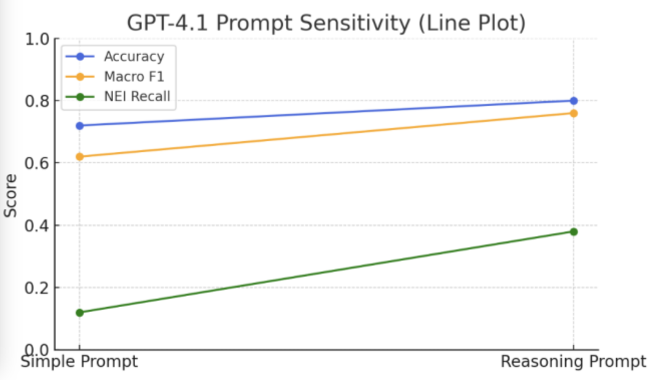


Fig 9: Model Performance by retriever strategy

We examine the impact of prompt structure on GPT-4.1’s classification performance. Using a cautious reasoning-style prompt improves both **accuracy** (from 0.72 to 0.80), F1 and especially **NEI recall** (0.12 → 0.38) compared to a minimal prompt, demonstrating that even robust models benefit from structured guidance. This highlights the role of instruction tuning and prompt optimization in high-risk domains like fact-checking. Refer [Appendix C] for more details.

D. Claim - Evidence Similarity across Retriever types

We compare three retriever strategies using cosine similarity scores across the top 5 ranked evidences per claim on Mistral RAG pipeline. As shown in Fig. [10], the **dense retriever** provides the most relevant top-ranked evidence (avg. similarity ~ 0.71), but declines after rank 2, leading to poor recall. The **multi-query retriever**, by aggregating results from multiple reformulations, maintains higher average similarity at mid-ranks (ranks 2–4), improving evidence coverage and recall. The **hybrid retriever** shows moderate performance at rank 0–1 but suffers at lower ranks. These findings support multi-query retrieval for balanced precision and recall.

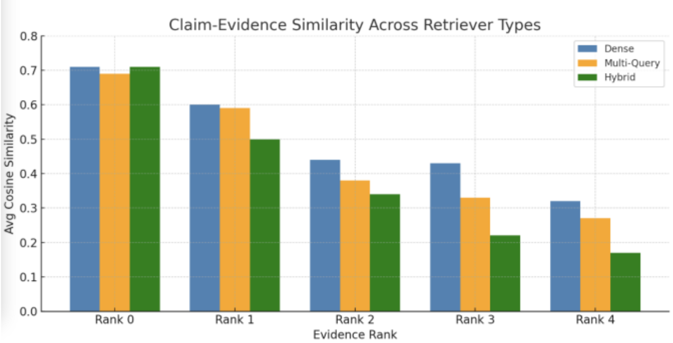


Fig 10: Claim-Fetch Evidence Similarity across Retriever types

E. RAG Pipeline vs Direct LLM Comparison

We compares the performance of the Mistral-7B model in two configurations: (1) integration within a RAG pipeline with retrieved evidence, and (2) as standalone without any external documents. As shown in [Table 6], the RAG-enabled setup achieves significantly higher accuracy (0.55 vs. 0.45) and F1 score. This confirms RAG's core value of grounding outputs in external evidence is essential for factual tasks. Without it, hallucinations increase significantly

Model	Accuracy	F1 score
Mistral with RAG	0.55	0.48
Mistral standalone	0.45	0.39

Table 6 : Performance comparison of the base Mistral model with and without RAG evidence retrieval on fact verification

F. Impact of Asynchronous Retrieval on RAG Latency

While conducted on a small test batch, an asynchronous implementation of the retrieval reduces average execution time from **3.7s to 2.4s**, as shown in Fig. [11]. Asynchronous retrieval can enhance system throughput with minimal engineering effort—a practical takeaway for real-time RAG deployments.

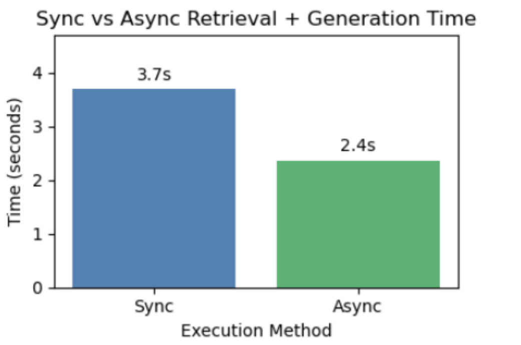


Fig 11: Response time Sync Vs Async Retrieval

VII. CONCLUSION

A. This project presents the successful design and evaluation of a retrieval-augmented generation (RAG) pipeline for automated and interpretable fact-checking using the FEVER dataset. We investigate a range of large language models (LLMs), explore multiple retrieval configurations, apply fine-tuning strategies including PEFT and DPO, and integrate explainability techniques to enhance transparency, supported by comprehensive instrumentation of the RAG pipeline. We also acknowledge our limitations due to constrained GPU resources and the modest scale of the training dataset, which may cap model performance and limit the depth of tuning and ablation. Key conclusions:

- **RAG improves factual grounding and trust:** Even untuned 7B models like Mistral and Qwen **perform** significantly better when provided with retrieved evidence. A direct comparison shows that removing retrieval reduces Mistral's accuracy by 11%. RAG reduces hallucinations and enables explainable, traceable predictions.
- **Parameter-Efficient Fine-Tuning (PEFT) yields major performance gains:** With just 1,000 examples and limited compute, PEFT tuning improves Mistral-7B's accuracy by ~17% (56% → 74%). NEI recall improves the most, demonstrating the model's enhanced ability to abstain when evidence is insufficient—an essential trait for trustworthy fact-checking.
- **Novel application of Direct Preference Optimization (DPO):** Our study is among the first to explore DPO within a RAG framework for fact-checking. While DPO alone offers limited benefits, applying it on top of PEFT-tuned models improved training stability and preserved alignment gains—showing synergy between supervised tuning and preference learning.
- **Retriever design affects evidence quality and outcomes:** Multi-query retrieval enhances accuracy and evidence diversity, particularly for stronger models like GPT-3.5. Threshold tuning reveals that faithfulness and relevancy often trade off, underscoring the need for careful evidence filtering
- **Explainability tools enhance transparency**: We integrate Captum, SHAP, LIME, and prompt-based soft-XAI to interpret model decisions at the token level. While Captum

works best with decoder-only models like Mistral, LIME and SHAP offer clearer attributions for Qwen. Prompt-based rationales align well with attribution results and added interpretability at minimal cost.

- **Performance vs. Transparency Trade-offs**: While GPT-4o achieves the highest accuracy (~80%), the fine-tuned Mistral model offers a competitive and explainable alternative using open tools. Our experiments demonstrate that smaller models can be made viable with proper tuning and careful pipeline engineering. Large closed-source models like GPT4o offer limited attribution analysis or explainability.

VIII. FUTURE WORK

This project opens several promising directions for extending the RAG + XAI fact-checking pipeline:

- **Benchmark our RAG on KILT**: Benchmark this RAG for fact verification on KILT (Knowledge Intensive Language Tasks) to compare our performance against state-of-art.
- **Scaling Fine-Tuning**: Our PEFT tuning on Mistral-7B yields strong gains. With more resources, full fine-tuning of 7B models or larger models (e.g., LLaMA-2/3 13B or 70B) on the complete FEVER dataset (>100K claims) can push accuracy beyond GPT-4.1 levels. Instruction tuning with rationale supervision may further improve explanation quality.
- **Advanced Retrieval Techniques**: Current retrievers we use off-the-shelf dense embeddings. Fine-tuning retrievers using claim-evidence pairs, distillation from oracle retrievers, or implementing cross-encoder re-rankers can improve precision/recall. Refer [25]
- **Generalization to Complex Tasks**: Extend the pipeline to multi-hop fact-checking (e.g., HoVer dataset) requires handling cross-document reasoning. Can explore multimodal RAG setups (e.g., RAGAR) to verify claims involving images or structured data is another direction.
- **XAI Enhancements**: Captum and LIME work well, but SHAP faces limitations with Mistral. Future work can test counterfactual explanations, or develop better attribution methods for generative LLMs. Explore other frameworks like LIT and transformer-explain etc.
- **Human Evaluation & Interface Design**: A UI demo allowing users to input claims and visualize evidence + explanations will help validate trust and usability.
- **Optimization for Deployment**: To scale this system, real-time performance needs improvement. We can explore Strategies like embedding optimization, asynchronous pipeline components, and GPU batching. Latency/throughput benchmarking is essential for production readiness.
- **Multi dataset Validation**: Beyond FEVER, the pipeline can be validated on other fact-checking datasets (e.g. LIAR, COVID-19 misinformation sets) to assess generalization and robustness across domains.

REFERENCES

[1] Patrick Lewis et el, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 793), 2020. *https://arxiv.org/abs/2005.11401v4*

[2] Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. Improving Explainable Fact-Checking with Claim-Evidence Correlations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1600–1612, Abu Dhabi, UAE. Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.108/

[3] M. Abdul Khaliq et el, "RAGAR, Your Falsehood Radar: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models". Seventh Fact Extraction and VERification Workshop (FEVER), pages 280–296, USA. Association for Computational Linguistics. *https://arxiv.org/abs/2404.12065*

[4] Russo, Daniel & Menini, Stefano & Staiano, Jacopo & Guerini, Marco. (2024). Face the Facts! Evaluating RAG-based Fact-checking Pipelines in Realistic Settings. *https://arxiv.org/abs/2412.15189*

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685. https://doi.org/10.48550/arXiv.2106.09685

[6] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv preprint:arXiv:2305.18290. https://doi.org/10.48550/arXiv.2305.18290

[7] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 4765–4774). https://arxiv.org/abs/1705.07874

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

[9] FEVER: *https://huggingface.co/datasets/copenlu/fever_gold_evidence*

[10] RAGAS for additional RAG performance metrics: https://github.com/explodinggradients/ragas

[11] Fever datasets : *https://fever.ai/*

[12] FAISS *: https://github.com/facebookresearch/faiss*

[13] Chroma*: https://github.com/chroma-core/chroma*

[14] X. Wu, Y. Zhang, Z. Wang, and J. Li, "Enhancing large language models with direct preference optimization: A comprehensive survey," arXiv preprint arXiv:2410.01808, Oct. 2024, doi: 10.48550/arXiv.2410.01808

[15] RAG Driven Generative AI by Denis Rothman, packt, 2024. *https://www.amazon.com/RAG-Driven-Generative-retrieval-generation-LlamaIndex/dp/1836200919*

[16] HOVER dataset : *https://hover-nlp.github.io/*

[17] Captum : *https://captum.ai/*

[18] Nie, Yixin & Chen, Haonan & Bansal, Mohit. (2018). Combining Fact Extraction and Verification with Neural Semantic Matching Networks. 10.48550/arXiv.1811.07039. https://arxiv.org/abs/1811.07039

[19] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "Generating fact-checking explanations," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), Online, Jul. 2020, pp. 7352–7364, doi: 10.18653/v1/2020.acl-main.614. https://aclanthology.org/2020.acl-main.656/

[20] Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. Improving Explainable Fact Checking with Claim Evidence Correlations. In Proceedings of the 31st International Conference on Computational Linguistics, pages 1600–1612, Abu Dhabi, UAE. Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.108/

[21] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in Proc.

18th Conf. Eur. Chapter Assoc. Comput. Linguistics: Syst. Demonstrations, St. Julians, Malta, Mar. 2024, pp. 150–158, doi: 10.18653/v1/2024.eacl-demo.16. https://arxiv.org/abs/2309.15217

[22] B. He et al., "Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation," arXiv preprint arXiv:2410.05801, Oct. 2024, doi: 10.48550/arXiv.2410.05801. https://arxiv.org/abs/2410.05801

[23] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "Generating fact-checking explanations," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), Online, Jul. 2020, pp. 7352–7364, doi: 10.18653/v1/2020.acl-main.614. https://aclanthology.org/2020.acl-main.656/

[24] N. Kotonya and F. Toni, "Explainable automated fact-checking: A survey," in Proc. 28th Int. Conf. Comput. Linguistics (COLING), Online, Dec. 2020, pp. 5430–5443, doi: 10.18653/v1/2020.coling-main.474. https://aclanthology.org/2020.coling-main.474/

[25] Pezzuti, F., MacAvaney, S., & Tonellotto, N. (2025). Exploring the Effectiveness of Multi-stage Fine-tuning for Cross-encoder Re-rankers. arXiv preprint arXiv:2503.22672. https://arxiv.org/abs/2503.22672

[26] FEVER Fact Verification benchmarks on KILT : https://paperswithcode.com/sota/fact-verification-on-kilt-fever

[27] Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. Transactions of the Association for Computational Linguistics, 10:1013–1030. https://aclanthology.org/2022.tacl-1.59/

[28] M. Glass, G. Rossiello, M. F. M. Chowdhury, A. R. Naik, P. Cai, and A. Gliozzo, "Re^{2}G: Retrieve, Rerank, Generate," arXiv preprint arXiv:2207.06300, 2022. [Online]. Available: https://arxiv.org/pdf/2207.06300

[29] F. Petroni et al., "KILT: a Benchmark for Knowledge-Intensive Language Tasks," in *Proc. North Am. Assoc. Comput. Linguist. - Hum. Lang. Technol.: Long Papers*, Online, Jun. 2021, Available: https://aclanthology.org/2021.naacl-main.106

APPENDIX

A. *Model Performance Comparison by Accuracy*

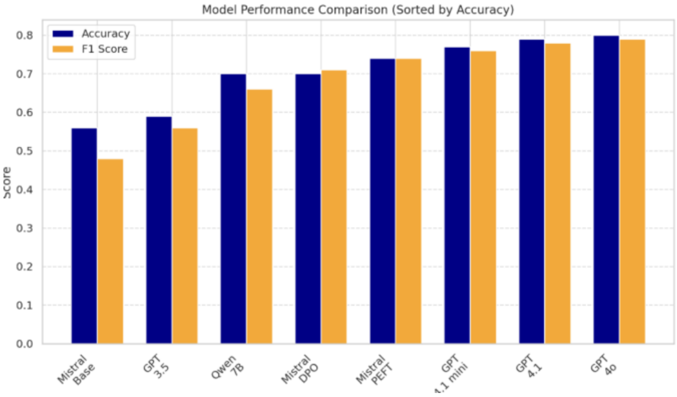


Fig 12: Model performance by Query

B. *PEFT and DPO Tuned RAG/Model Evaluation Metrics*

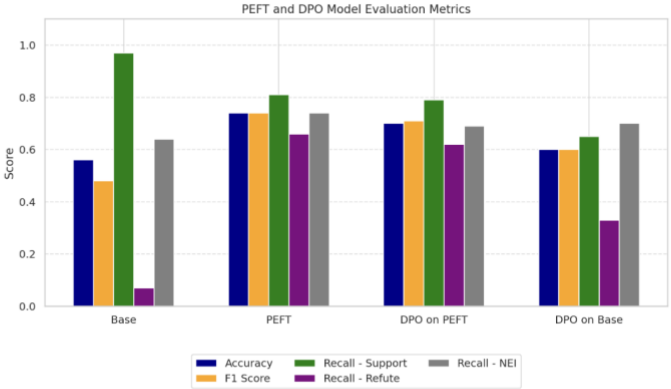
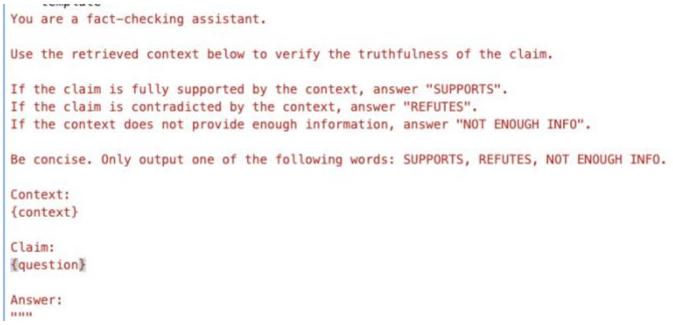


Fig 13: PEFT and DPO tuned compared to base Mistral model

C. *GPT 4o Prompt Sensitivity*

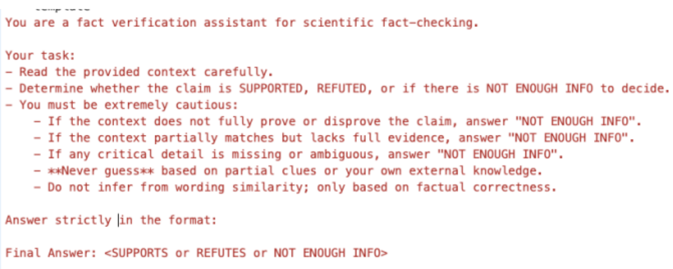
We examine the impact of prompt structure on GPT-4.1's classification performance. Below is a simple prompt (lower accuracy, 0.72)



```
You are a fact-checking assistant.

Use the retrieved context below to verify the truthfulness of the claim.

If the claim is fully supported by the context, answer "SUPPORTS".
If the claim is contradicted by the context, answer "REFUTES".
If the context does not provide enough information, answer "NOT ENOUGH INFO".

Be concise. Only output one of the following words: SUPPORTS, REFUTES, NOT ENOUGH INFO.

Context:
{context}

Claim:
{question}

Answer:
"""
```

Fig 14: GPT simple prompt

We change to more reasoning style prompt, accuracy improves from 0.72 to ~ 0.80.



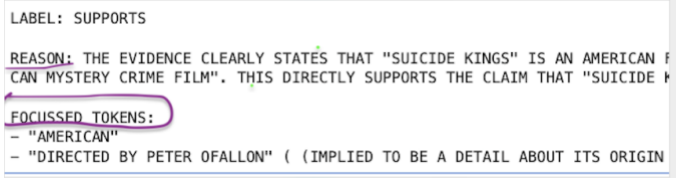
```
You are a fact verification assistant for scientific fact-checking.

Your task:
- Read the provided context carefully.
- Determine whether the claim is SUPPORTED, REFUTED, or if there is NOT ENOUGH INFO to decide.
- You must be extremely cautious:
    - If the context does not fully prove or disprove the claim, answer "NOT ENOUGH INFO".
    - If the context partially matches but lacks full evidence, answer "NOT ENOUGH INFO".
    - If any critical detail is missing or ambiguous, answer "NOT ENOUGH INFO".
    - **Never guess** based on partial clues or your own external knowledge.
    - Do not infer from wording similarity; only based on factual correctness.

Answer strictly in the format:

Final Answer: <SUPPORTS or REFUTES or NOT ENOUGH INFO>
```

Fig 15 GPT reasoning prompt

D. *Soft XAI prompt and response*

An additional prompt encourages the model to give context to why it chose a label. For Qwen, the underlined is that prompt. "*…classify the claim as SUPPORTS, … Then briefly explain why. Also tell which tokens from claim and evidence it focussed most on or were of highest attention*"

For same Claim "Suicide kings is a film from the United…", Model responded with Reason and Token information.

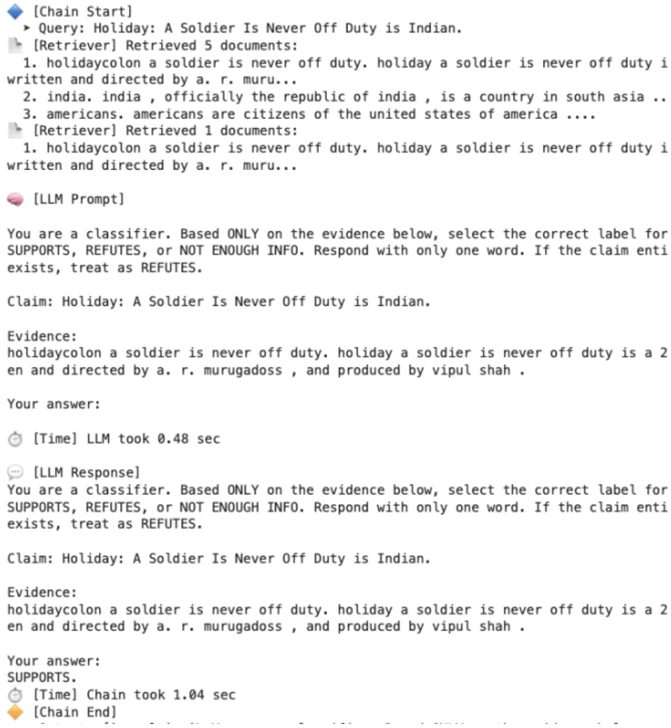


```
LABEL: SUPPORTS

REASON: THE EVIDENCE CLEARLY STATES THAT "SUICIDE KINGS" IS AN AMERICAN F
CAN MYSTERY CRIME FILM". THIS DIRECTLY SUPPORTS THE CLAIM THAT "SUICIDE K

FOCUSSED TOKENS:
- "AMERICAN"
- "DIRECTED BY PETER OFALLON" ( (IMPLIED TO BE A DETAIL ABOUT ITS ORIGIN
```

Fig 16: Additional Prompt to Model for added XAI interpretability

E. *Full RAG pipeline Logging*

Below is a trimmed example of full RAG pipeline logging. It includes the documents Retriever fetches, the filtered documents based on threshold cutoff, the LLM prompt, LLM's full response and final classifier label. Times taken at each stage of pipeline is logged.



```
🔷 [Chain Start]
  ➤ Query: Holiday: A Soldier Is Never Off Duty is Indian.
📄 [Retriever] Retrieved 5 documents:
  1. holidaycolon a soldier is never off duty. holiday a soldier is never off duty i
written and directed by a. r. muru...
  2. india. india , officially the republic of india , is a country in south asia ..
  3. americans. americans are citizens of the united states of america ....
📄 [Retriever] Retrieved 1 documents:
  1. holidaycolon a soldier is never off duty. holiday a soldier is never off duty i
written and directed by a. r. muru...

🧠 [LLM Prompt]

You are a classifier. Based ONLY on the evidence below, select the correct label for
SUPPORTS, REFUTES, or NOT ENOUGH INFO. Respond with only one word. If the claim enti
exists, treat as REFUTES.

Claim: Holiday: A Soldier Is Never Off Duty is Indian.

Evidence:
holidaycolon a soldier is never off duty. holiday a soldier is never off duty is a 2
en and directed by a. r. murugadoss , and produced by vipul shah .

Your answer:

⏱ [Time] LLM took 0.48 sec

💬 [LLM Response]
You are a classifier. Based ONLY on the evidence below, select the correct label for
SUPPORTS, REFUTES, or NOT ENOUGH INFO. Respond with only one word. If the claim enti
exists, treat as REFUTES.

Claim: Holiday: A Soldier Is Never Off Duty is Indian.

Evidence:
holidaycolon a soldier is never off duty. holiday a soldier is never off duty is a 2
en and directed by a. r. murugadoss , and produced by vipul shah .

Your answer:
SUPPORTS.
⏱ [Time] Chain took 1.04 sec
🔶 [Chain End]
```

Fig 17: Example of Full RAG pipeline logging for transparency and introspection.

F. Notes on Challenges in XAI for LLMs in our study

We encountered a few challenges: (1) Context Length – Dealing with long chunks of text for explanations is tricky to visualize, so we focused on the most relevant bits.(2) Computation Cost – running *Captum IntegratedGradients* on a 7B model with many tokens is very slow (even with 30 steps). We only did this for analysis on sample cases, as it's not practical to do for every query in real-time without speedup. (3) Multi-token Phrases – Sometimes the meaning comes from a combo of tokens, like "United States" (split into "United" and "States"). On their own, they don't seem key, but together they matter, a way to group token importance would help (4) Model bias - As we noticed, the model sometimes overvalued random stuff like punctuation, perhaps due to its learned biases.

G. Sample Prompts for LLMs

For this sample Claim-Evidence pair, we show the prompts for different models

Claim: Suicide Kings is a film from the United States.
Evidence: Suicide Kings is a 1997 American mystery crime film based on Don Stanford's short story...

Mistral-7B Prompt

You are a classifier. Based ONLY on the evidence below, select the correct label for the claim: **SUPPORTS**, **REFUTES**, or **NOT ENOUGH INFO**. Respond with only one word. If the claim entity is mentioned but no positive link exists, treat as REFUTES.

Claim: Suicide Kings is a film from the United States.

Evidence: suicide kings. suicide kings is a 1997 american mystery crime film based on don stanford's short story….

Your answer:

Qwen-7B Prompt

<|im_start|>**system** You are a fact-checking assistant.
Classify the Claim based only on the Evidence.
Respond with exactly one of: **SUPPORTS, REFUTES, or NOT ENOUGH INFO**.
<|im_end|>

<|im_start|>**user**
Claim: Suicide Kings is a film from the United States.

Evidence: suicide kings. suicide kings is a 1997 american mystery crime film based on don stanford's short story….

Answer with one of: SUPPORTS / REFUTES / NOT ENOUGH INFO
<|im_end|> <|im_start|>**assistant**

GPT Prompt

You are a fact verification assistant for fact-checking claims using retrieved evidence. **Instructions:** Read the claim and the provided evidence (context). Decide if the claim is **SUPPORTED**, **REFUTED**, or **NOT ENOUGH INFO.**

Be cautious: a) If evidence does not fully support or contradict the claim, choose "NOT ENOUGH INFO". **b)** Do not infer from partial clues or use your own knowledge.

Output Format:
Final Answer: <SUPPORTS / REFUTES / NOT ENOUGH INFO> **Most Influential Passage**: "<key phrase from context>" **Context**: suicide kings. suicide kings is a 1997 american mystery crime film based on don stanford's short story….. **Claim**: Suicide Kings is a film from the United States

H. Few-shot ICL Prompt for PEFT Tuning

Using the same Claim example, for PEFT tuning, we use Few shot examples as an ICL learning within the prompt and this showed some improvement in the training accuracy.

PEFT Tuning Prompt

You are a classifier. Based ONLY on the evidence below, select the correct label for the claim: **SUPPORTS, REFUTES, or NOT ENOUGH INFO.**

Example 1: Claim: The moon is made of cheese. **Evidence**: Scientific studies show the moon is made of rock and dust. **Your answer**: REFUTES

Example 2: Claim: The sun is hot. **Evidence**: The sun's surface temperature is about 5,500°C. **Your answer**: SUPPORTS

Example 3: Claim: Gabrielle Union won an Oscar for her role in Bring It On. **Evidence**: Gabrielle Union starred in Bring It On, a 2000 American teen cheerleading comedy film. **Your answer**: NOT ENOUGH INFO

Claim: Suicide Kings is a film from the United States

Evidence: Suicide Kings is a 1997 American mystery crime film based on Don Stanford's short story... **Your answer**: SUPPORTS: