Research Proposal

Title: Fake News Detection using LLaMA-based Large Language Models

I. Background and Motivation

Fake news continues to be a serious problem for the societies. It has the potential to influence how people think, stir social instability, and even undermine the stability of democratic systems. Although transformer-based models such as BERT, RoBERTa have demonstrated strong baseline performance, the latest advancement of Large Language Models (LLMs) goes further by offering richer contextual reasoning, better use of commonsense, and stronger adaptability across different languages. My previous work surveyed 50 LLM-based approaches to fake news detection, identifying promising strategies such as entity extraction, retrieval-augmentation, adversarial training, and explainable reasoning. However, most studies rely on closed-source models like GPT-3.5/4, which limit reproducibility and impose high computational costs.

The LLaMA family of models is open-source, efficient, and scalable, offering a great opportunity to build transparent, cost-effective, and adaptable fake news detection systems. With parameter-efficient fine-tuning and the addition of retrieval and reasoning modules, LLaMA models are particularly well-suited for low-resource environments.

II. Research Objectives

This project aims to explore the use of LLaMA models for fake news detection, with the following objectives:

- Build a LLaMA-based cost-efficient framework that can accurately classify news as real or fake.
- Explore parameter-efficient fine-tuning methods (e.g., LoRA, QLoRA) to adapt LLaMA for domain-specific misinformation.
- Integrate retrieval-augmented generation (RAG) to verify predictions using external knowledge.
- Enhance interpretability, providing human-understandable rationales for each prediction.

III. Methodology

- Datasets: ISOT, LIAR, PolitiFact, GossipCop, etc.
- Model: Fine-tune LLaMA (7B/13B) with LoRA/QLoRA for parameter efficient fine tuning; integrate retrieval-based verification from sources like Wikipedia and fact-checking databases.
- Evaluation matrices: Accuracy, F1-score, explainability (rationale quality), and efficiency metrics.
- Comparisons: Benchmark against GPT-based baselines and smaller models (BERT, RoBERTa).

IV. Expected Contributions

- A cost-efficient and reproducible framework for misinformation research.
- Enhancing prediction reliability by mitigating LLaMA's hallucinations through retrieval-augmented verification.
- Insights on retrieval-augmented fake news detection.
- Open-sourced fine-tuned models and training scripts for community use.