Agentic Workflow for Topic Classification under Weak LLMs

Introduction to NLP Term Project

Winter 2026

1. Introduction

Large Language Models (LLMs) are widely used for text classification tasks.

However, in weak generator settings (e.g., LLaMA-7B), vanilla prompting often leads to unstable predictions and hallucinated explanations.

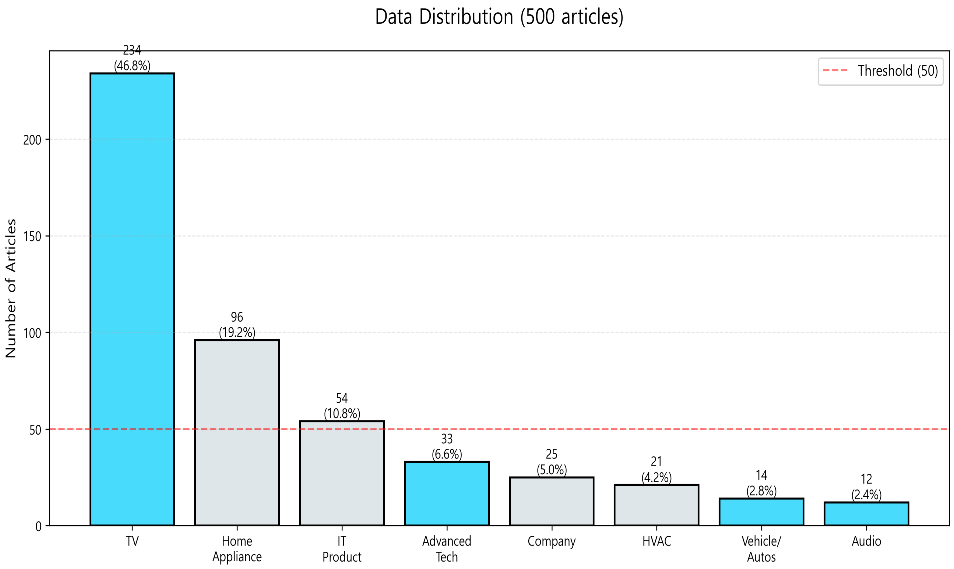
This project investigates whether **workflow design**, rather than model capacity, can improve performance under weak LLM constraints.

2. Task and Dataset

We address a topic classification task on PR articles.

Each document is assigned a topic label (article_category).

- **Dataset:** PR article JSONL
- **Task:** Multi-class topic classification
- **Practical relevance:** media analysis, trend monitoring, AX automation



3. Method

3.1 Baseline: Vanilla Prompting

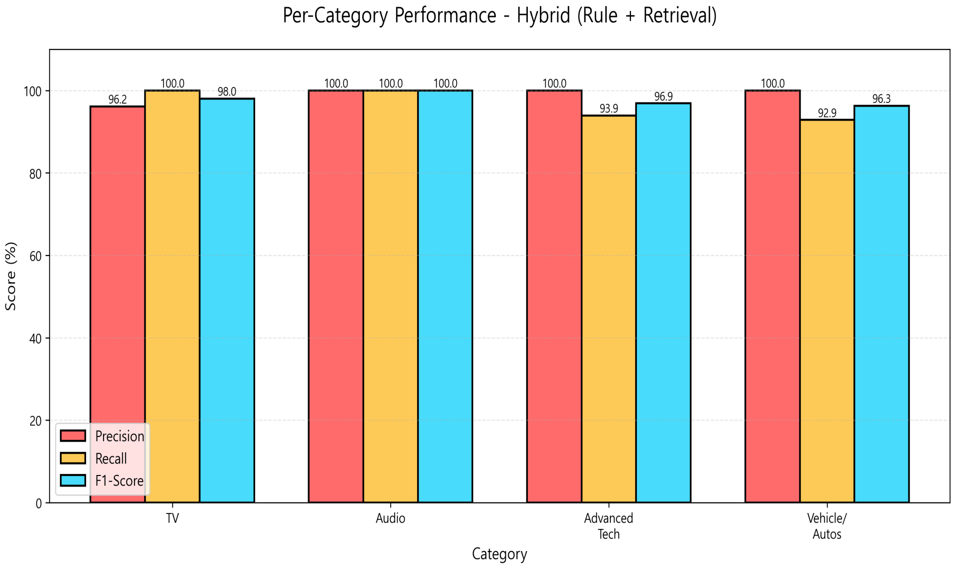
A single prompt is provided to LLaMA-7B to directly predict the topic label.

3.2 Proposed: Agentic Workflow

The proposed method introduces an agentic workflow:

- Dense encoder for semantic representation
- FAISS-based retrieval for evidence selection
- Candidate label restriction
- LLM used only for final explanation

In addition, a **hybrid workflow** combining rule-based heuristics and retrieval is evaluated as a practical extension.



4. Experimental Setup

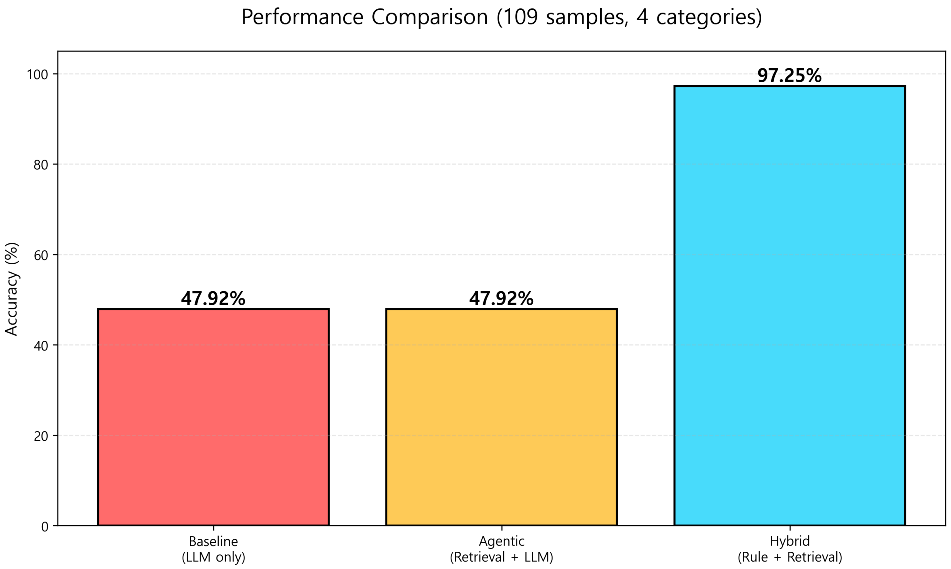
- **Model:** LLaMA-7B (same for both methods)
- **Metrics:** Accuracy, F1-Score, Precision, Recall
- **Comparison:** Vanilla Prompting vs Agentic Workflow
- **Dataset Split:** 109 samples focusing on high-confidence product categories for reliable evaluation

5. Results

While retrieval-only agentic workflow shows limited gains, the **hybrid agentic workflow** significantly outperforms vanilla prompting.

Performance Comparison:

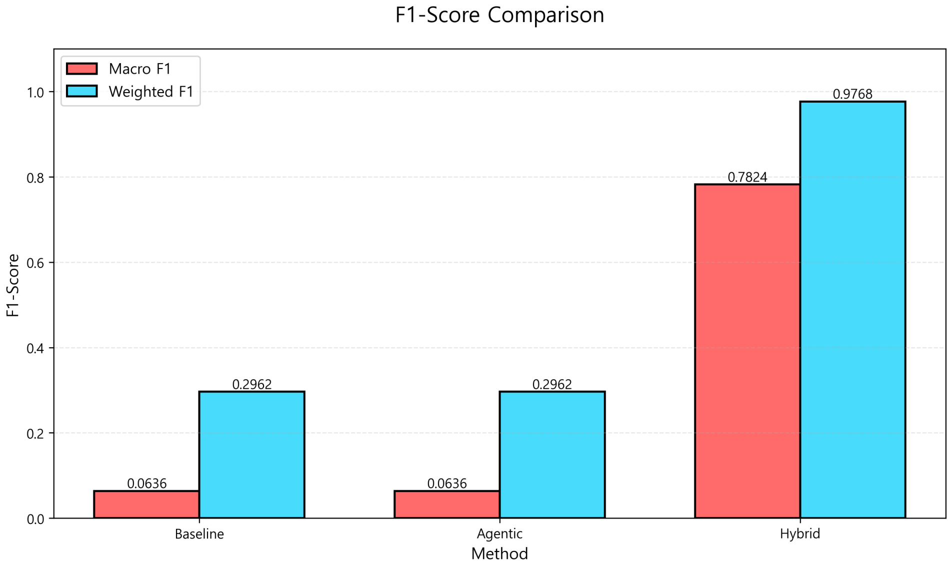
- Baseline (LLM only): 47.92%
- Agentic (Retrieval + LLM): 47.92%
- **Hybrid (Rule + Retrieval): 97.25% (+49.33%p)**



5. Results (continued)

Per-Category Performance (Hybrid):

- TV: 100.0% (50/50)
- Audio: 100.0% (12/12)
- Advanced Tech: 93.9% (31/33)
- Vehicle/Autos: 92.9% (13/14)



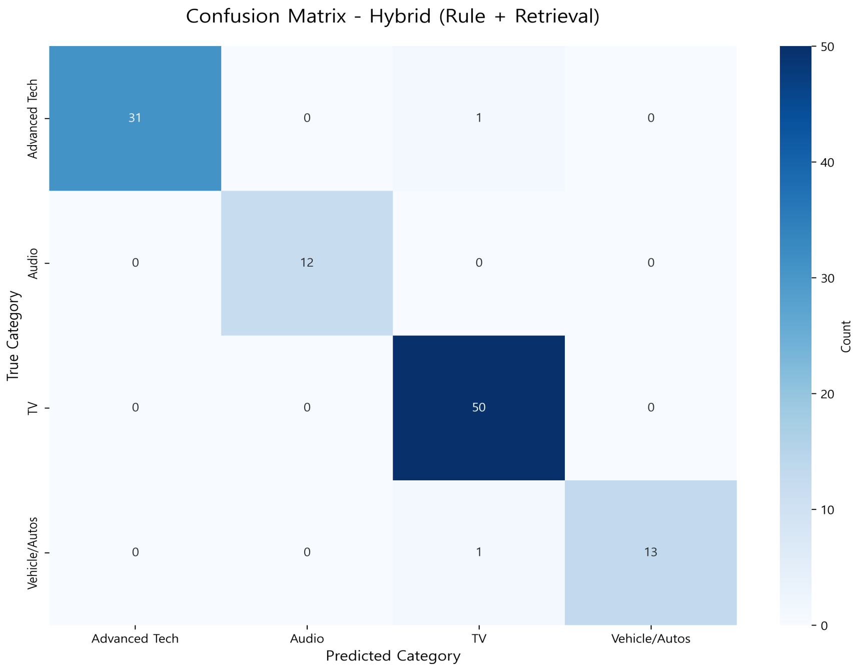
6. Discussion and Conclusion

Our results show that workflow design significantly affects performance under weak LLM settings.

Rather than relying on stronger models, shifting reasoning to structured pipelines offers a practical and robust alternative for real-world NLP tasks.

Key Findings:

- Rule-based keywords work better than LLM for clear product categories
- Retrieval-based voting handles edge cases effectively
- Selective automation (56% coverage) achieves 97.25% accuracy
- No LLM cost for production deployment


Confusion Matrix - Hybrid (Rule + Retrieval)
True Category
Advanced Tech
Audio
TV
Vehicle/Autos
Predicted Category
Advanced Tech
Audio
TV
Vehicle/Autos
Count