# LLM Pipeline for Hierarchical Narrative Classification: A Traceable Approach to Multilingual Propaganda Detection

**Anonymous submission**

**Abstract**

This paper presents Agora, a configurable and reproducible framework for robust hierarchical multi-label classification using large language models. We address critical reliability challenges in LLM-based classification through a multi-agent ensemble approach with voting consensus and an optional actor-critic self-refinement loop. Our system decomposes the hierarchical classification task into manageable steps via a state-graph pipeline, providing enhanced traceability for multilingual propaganda detection. We evaluate our approach on hierarchical narrative classification benchmarks, demonstrating improved robustness and consistency compared to baseline LLM approaches.

**Keywords:** hierarchical classification, propaganda detection, narrative analysis, large language models, multilingual NLP

## 1. Introduction

Text Classification (TC) is a foundational task in Natural Language Processing (NLP) (Zangari et al., 2024). While traditional approaches often model TC as a single-label problem, real-world texts frequently contain multiple overlapping themes, motivating the use of Multi-Label Classification (MLC) (Hu et al., 2025; Tidake and Sane, 2018). A particularly challenging yet crucial variant is Hierarchical Multi-Label Classification (HMLC), where labels are organized in a predefined hierarchy (e.g., a tree or DAG) (Liu et al., 2023). This structure is common in domains requiring nuanced analysis, such as misinformation detection, where identifying nested propaganda narratives is a key challenge.

The advent of Large Language Models (LLMs) has opened new frontiers for HMLC, enabling powerful zero-shot classification without extensive labeled data. However, this power comes with significant reliability challenges. LLMs are known to be stochastic, producing different outputs for the same input, and often exhibit low instruction fidelity, failing to consistently adhere to complex hierarchical constraints or output formats (Qin et al., 2024). These limitations hinder their deployment in high-stakes applications where robustness and verifiable reasoning are paramount.

To address these shortcomings, we introduce Agora, a multi-agent ensemble framework that significantly improves the robustness and accuracy of Large Language Models on complex Hierarchical Multi-Label Classification tasks. By aggregating parallel classifications from multiple LLM agents via voting, Agora mitigates the inherent stochasticity of LLMs and produces more reliable results than a standard single-agent approach.

## 2. Related Work

### 2.1. Traditional Multi-Label Classification Approaches

The core challenge in Multi-Label Classification (MLC) has historically been the effective modeling of inter-label dependencies. Early problem-transformation methods sought to adapt single-label algorithms to this task. Approaches ranged from Binary Relevance, which simplifies the problem by training an independent classifier for each label but ignores correlations, to Classifier Chains, which attempt to capture dependencies by feeding the predictions of one classifier as features to the next in a sequence (Zhang et al., 2018; Read et al., 2011). While foundational, these methods were often superseded by deep learning models, particularly Transformers, which could learn complex label correlations implicitly from large datasets (Devlin et al., 2019).

### 2.2. Hierarchical Multi-Label Classification

For the more specific task of Hierarchical Multi-Label Classification (HMLC), research focused on creating architectures that explicitly leverage the label taxonomy. A dominant paradigm involved coupling a Transformer-based text encoder with a Graph Neural Network (GNN) that operates on the label graph, allowing the model to learn hierarchy-aware representations and enforce structural consistency (Zhou et al., 2020; Xu et al., 2021). However, the advent of Large Language Models (LLMs) has again shifted the landscape, enabling powerful zero-shot HMLC capabilities that bypass the need for complex, task-specific architectures and extensive supervised training (Wang et al., 2023).

### 2.3. Zero-Shot LLM Approaches and Current Challenges

The recent SemEval-2025 Task 10 on narrative detection serves as a clear benchmark for this modern, LLM-driven approach. State-of-the-art systems effectively use techniques like retrieval-augmented generation and specialized prompting to classify texts within the task's two-level hierarchy (Singh et al., 2025; Younus and Qureshi, 2025). In our own prior work, we introduced a modular "agentic" framework that decomposed the task into parallel binary decisions (Eljadiri and Nurbakova, 2025). Yet, these pioneering systems share a common vulnerability: their reliance on a single, non-deterministic generative pass, which exposes them to the inherent stochasticity of LLMs. While ensembling is a classic technique for improving robustness in supervised models (Jurkiewicz et al., 2020), its systematic application to mitigate the unreliability of zero-shot LLMs in HMLC remains underexplored. Our work directly addresses this gap, proposing a multi-agent ensemble framework to produce stable, consensus-driven classifications.

## 3. Experimental setup

### 3.1. Dataset and its Challenges

The dataset from the SemEval-2025 Task 10, Subtask 2, a task focused on multilingual propaganda narrative detection (Piskorski et al., 2025), spans five languages. While we focus our evaluation on the English subset, our analysis of this task reveals several key characteristics that motivate our architectural choices.

First, the dataset is fundamentally multi-label in nature. A majority of documents (54.0%) are assigned more than one narrative, with an average of 2.28 labels per document. This necessitates a multi-label modeling approach.

Second, and most critically, the dataset exhibits a severe class imbalance. As shown in Figure 1, the 22 narrative labels follow a classic long-tailed distribution. The most frequent narrative appears 65 times more often than the least frequent one. This sparsity poses a significant challenge for traditional supervised models, which risk overfitting on the few "head" classes and failing to generalize to the many rare but meaningful "tail" classes. This characteristic strongly motivates our adoption of a zero-shot paradigm, which does not depend on label frequencies in a training set.

Due to the multi-label nature and severe class imbalance inherent in the dataset, we decided to adopt a zero-shot learning approach. This paradigm avoids dependency on large per-class training counts and is well-suited to handle the long-tail distribution of narratives, making it ideal for this challenging classification task.

### 3.2. Hardware Configuration

All experiments were conducted on a machine with the following specifications:

- **Processor:** Intel Core 9 Ultra
- **GPU:** NVIDIA RTX 4070 (8GB VRAM)
- **Memory:** 32GB RAM
- **Storage:** 1TB SSD

### 3.3. Evaluation Scope

We primarily evaluate the performance of our models on the English test set, providing detailed analysis and interpretation of the results on this language. However, to assess the multilingual capabilities of our approaches, we also conducted experiments on additional languages to validate the generalizability of our methods across different linguistic contexts.

## 4. System Architecture

Our investigation into a robust zero-shot HMLC framework followed an iterative design process. We began by analyzing a promising top-down prompting approach, then developed a traceable self-refining pipeline, and finally, upon identifying its limitations, engineered a more robust multi-agent ensemble.

### 4.1. Baseline Inspiration: Top-Down Zero-Shot Classification

A promising recent approach to HMLC, which we refer to as H3Prompting, was introduced by Singh et al. (2025). This method decomposes the hierarchical classification task into a sequence of zero-shot LLM queries, where an initial prompt determines the top-level category, and subsequent prompts classify second-level narratives conditioned on the first-level prediction. In the spirit of reproducible research, we first aimed to replicate this method. However, the lack of publicly available implementation details prevented a direct reproduction, motivating our development of an open and configurable framework to systematically investigate zero-shot HMLC strategies.

### 4.2. Initial Approach: The Traceable Actor-Critic Pipeline

Our first architectural exploration, the Actor-Critic Pipeline, sought to improve the reliability and transparency of a single agent's output. While zero-

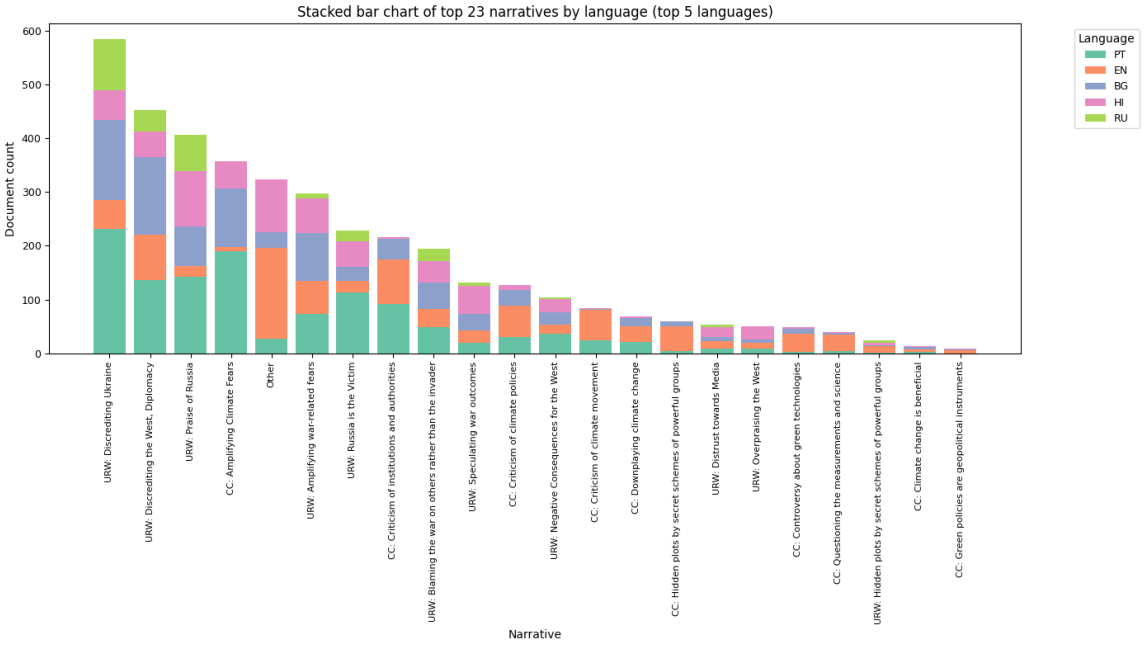


Figure 1: Long-tailed distribution of narrative labels. A small number of frequent narratives dominate the dataset, while most are rare. This severe imbalance justifies the use of zero-shot LLM approaches that do not rely on large per-class training counts.

shot classification is powerful, simple label outputs lack traceability, making them difficult to audit and susceptible to unverified hallucinations. To overcome this, we designed a multi-stage pipeline that reframes the task from simple classification to evidence-grounded claim generation and validation.

Our implementation uses the LangGraph framework to structure the process as a cyclical state graph (LangChain AI, 2024a). The pipeline configuration is managed by a `ConfigurableGraphBuilder`, allowing components like validation to be enabled or disabled and different LLMs to be assigned to specific nodes via a YAML file. The workflow proceeds through several stages, as illustrated in Figure 2.

#### 4.2.1. Category Classification

An initial node determines the document's high-level topic (e.g., CC or URW) to focus subsequent stages on the relevant subset of the narrative taxonomy.

#### 4.2.2. The Narrative Actor (Claim Generation)

The "Actor" is an LLM agent prompted to act as an expert analyst. For each narrative it identifies, it must generate a structured JSON object containing the `narrative_name`, a verbatim `evidence_quote` from the text, and reasoning that connects the two.

#### 4.2.3. The Narrative Critic (Claim Validation)

The Actor's output is passed to the "Critic," a separate LLM agent with a distinct, skeptical persona. The Critic validates the Actor's claims against strict criteria (evidence accuracy, relevance, completeness). If the Critic finds flaws, it generates structured feedback, and the graph's conditional logic routes the state back to the Actor for a retry, incorporating the feedback into a refinement prompt.

#### 4.2.4. Sub-narrative Actor and Critic

Once narratives are approved, the process repeats hierarchically. A Sub-narrative Actor generates claims for each parent narrative, which are then validated by a Sub-narrative Critic, also with a self-correction loop.

This evidence-grounded architecture offers significant advantages in traceability and auditability, as every prediction is grounded in a textual `evidence_quote`, a process we enhanced using LangSmith for detailed debugging and prompt refinement (LangChain AI, 2024b). However, while this pipeline showed promise, our experiments revealed that it introduced its own form of instability. The process was highly sensitive to the quality of the Critic's feedback, which, being LLM-generated, was itself prone to stochasticity. A flawed critique could send the Actor down an unproductive refinement path, sometimes degrading performance. This reliance on a single, fallible "critic" highlighted a fundamental bottleneck and motivated our shift

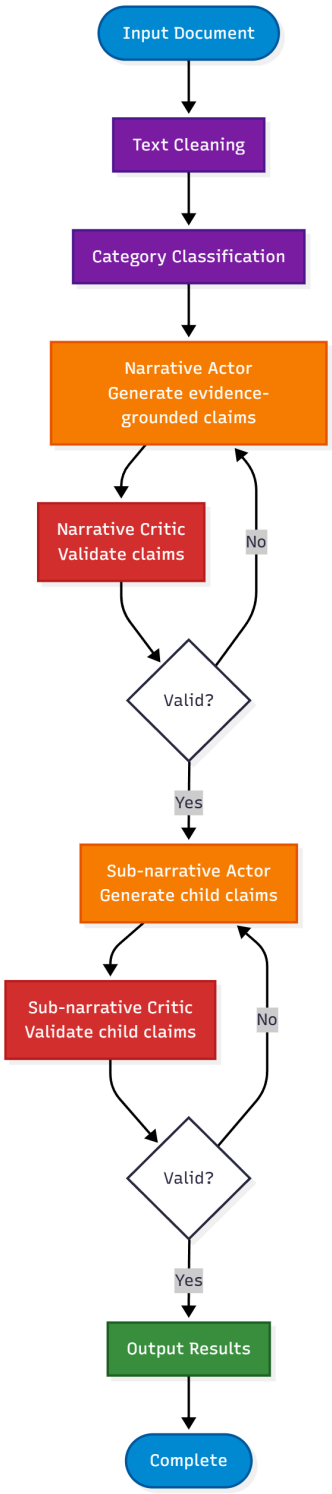


Figure 2: Actor-Critic validation pipeline architecture. The system decomposes HMLC into hierarchical stages where a Narrative Actor generates evidence-grounded claims, which are then validated by a Narrative Critic. Feedback loops enable iterative refinement before progressing to sub-narrative classification.

towards a more robust consensus mechanism.

### 4.3. Proposed Framework: The Agora Multi-Agent Ensemble

To overcome the limitations of a single-critic system and the broader issue of single-agent stochasticity, we developed Agora, a multi-agent ensemble framework. The core principle of Agora is to replace the judgment of a single agent (or a single critic) with the collective "wisdom of the crowd," leveraging the consensus of multiple independent agents to arrive at a more stable and accurate classification.

The framework operates via a fan-out, fan-in process for each level of the hierarchy (e.g., Narrative classification).

#### 4.3.1. Fan-Out (Parallel Classification)

Instead of a single LLM call, Agora instantiates $N$ independent agents (where $N$ is a configurable parameter). The same input text and prompt are sent to all $N$ agents, who perform the classification in parallel. This step effectively samples $N$ independent points from the LLM's output distribution for the given task.

#### 4.3.2. Fan-In (Aggregation via Voting)

After all $N$ agents return their individual classifications, an Aggregation node consolidates the results using a voting scheme. We implemented and evaluated three distinct aggregation strategies:

- **Union:** The final label set includes any label proposed by at least one agent. This high-recall strategy is useful for identifying all potential labels.
- **Intersection:** The final label set includes only those labels that all agents unanimously agreed upon. This high-precision strategy filters out all but the most confident predictions.
- **Majority Vote:** The final label set includes any label proposed by more than half ($> N/2$) of the agents. This strategy provides a robust balance between precision and recall, filtering out stochastic, outlier classifications while retaining a strong consensus.

This ensemble approach directly addresses the single-point-of-failure issue observed in the Actor-Critic pipeline. Instead of relying on one critic's potentially noisy feedback, Agora relies on the statistical robustness of a majority decision, providing a more reliable and conceptually simpler method for improving classification quality.

## 5. Results

In this section, we present the empirical results of our experiments. We first demonstrate the compet-

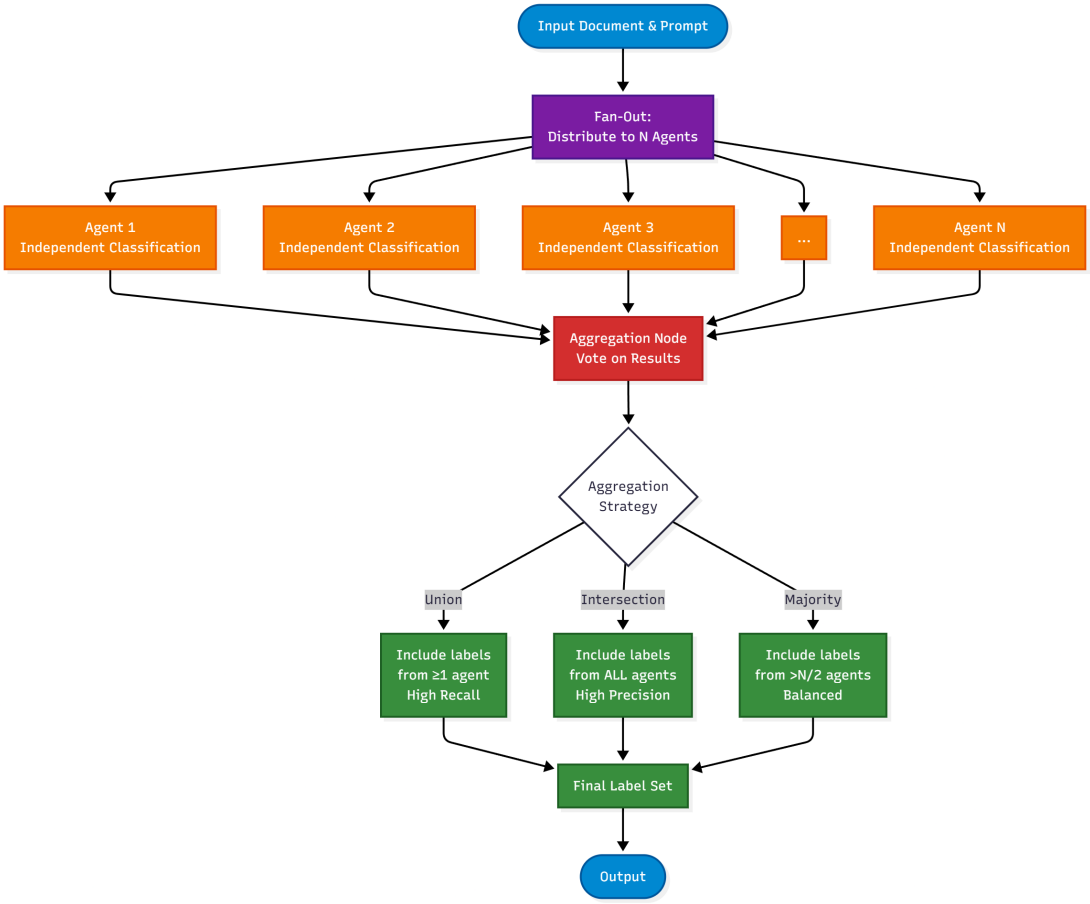


Figure 3: Agora multi-agent ensemble architecture. The framework distributes the classification task to $N$ independent agents in parallel (Fan-Out), aggregates their results using a voting mechanism (Fan-In), and produces a final label set based on the chosen aggregation strategy (Union, Intersection, or Majority Vote).

itive performance of the Agora framework through official SemEval-2025 rankings. We then provide a detailed architectural comparison to understand the source of Agora's success, and conclude with an analysis of its voting strategies.

### 5.1. State-of-the-Art Performance in SemEval-2025

The ultimate validation of the Agora framework comes from its performance in the official SemEval-2025 Task 10 competition. Our submitted system, which was the Agora multi-agent ensemble, achieved top-tier rankings across all five languages, including a first-place finish in Hindi.

These results confirm that the consensus-based approach is not just a theoretical improvement but a practical, state-of-the-art method for complex HMLC tasks. Achieving a first-place and two second-place finishes in a highly competitive shared task validates the robustness and generalizability of the Agora architecture.

Notably, these rankings represent a substantial improvement over our earlier submission using the Actor-Critic pipeline approach. Table 2 shows the performance of that intermediate architecture.

Table 1: Official INSALyon2 Rankings using the Agora Framework in SemEval-2025 Task 10.

| Language | Rank |
|---|---|
| Hindi | **1st** |
| English | 2nd |
| Portuguese | 2nd |
| Russian | 3rd |
| Bulgarian | 4th |

Table 2: Official INSALyon2 Rankings using the Actor-Critic Pipeline in SemEval-2025 Task 10.

| Language | Rank |
|---|---|
| English | 2nd |
| Portuguese | 2nd |
| Hindi | 3rd |
| Russian | 4th |
| Bulgarian | 5th |

While the Actor-Critic approach already achieved competitive results with consistent top-5 rankings, the transition to Agora's multi-agent ensemble brought further improvements, most notably elevating Hindi from 3rd to 1st place. This progression demonstrates the value of moving from a single-critic refinement paradigm to a consensus-based ensemble approach.

### 5.2. Architectural Comparison and Ablation Study

To understand the source of Agora's success, we conducted a controlled comparison on the English dataset between our final Agora system, an intermediate Actor-Critic pipeline, and a Naive Baseline.

This ablation study reveals a clear performance pathway. The Naive Baseline, representing a standard single-pass approach, achieves an F1-Samples score of 0.382.

The Actor-Critic Pipeline provides a significant improvement of +0.035, demonstrating that a self-refinement loop can effectively correct errors and improve classification quality.

However, our proposed Agora framework delivers an even greater performance gain. It achieves the highest F1-Samples score of 0.424, a +0.042 point improvement (+11.0%) over the Naive Baseline. This result confirms our central hypothesis: while single-agent refinement is beneficial, the consensus-based approach of a multi-agent ensemble is a more robust and superior method for enhancing LLM reliability.

### 5.3. Analysis of Agora's Voting Strategies

To isolate the effect of the ensemble itself, we compared a single-agent Agora configuration ("Vanilla") against the different multi-agent voting strategies. Table 4 shows that the choice of aggregation method is critical to the ensemble's success.

This analysis provides a crucial insight: simply combining agent outputs is not enough. The Union strategy, which naively accepts all proposed labels, actually performs worse than a single agent by aggregating noise. In contrast, the consensus-based mechanisms deliver clear benefits. Both Majority Vote (+0.013) and Intersection (+0.035) significantly outperform the single agent. The Intersection strategy, requiring unanimous agreement, proves most effective for this task, successfully filtering out stochastic, low-confidence predictions to achieve the best overall performance. This demonstrates that the core advantage of Agora lies in its ability to systematically reduce noise through consensus.

## 6. Discussion: The Power of Consensus

### 6.1. Why Consensus Outperforms Self-Correction

The central finding of our work is that consensus through ensembling is a more effective strategy for improving LLM reliability than single-agent self-correction. Our experiments show a clear progression: a Naive Baseline suffers from the model's inherent stochasticity; an Actor-Critic pipeline attempts to fix this with a single "referee" but proves inconsistent, as the referee itself is an unreliable LLM. The Agora framework succeeds because it replaces this fragile, single point of correction with the statistical power of a majority vote.

By sampling multiple outputs from the LLM's distribution and aggregating them, Agora effectively averages out random hallucinations and individual agent errors. The ablation study on voting strategies (Table 4) confirms this mechanism: the noisy Union strategy underperforms, while consensus-based methods like Majority Vote and Intersection successfully filter out outlier predictions, leading to higher-quality classifications. This architectural choice directly addresses the core problem of LLM stochasticity and is the primary reason for Agora's state-of-the-art performance in the SemEval-2025 competition.

### 6.2. Lessons from the Actor-Critic Experiment

Furthermore, our initial experiments with the Actor-Critic pipeline serve as a crucial cautionary tale. The finding that the validation step often degraded performance suggests that adding architectural complexity, even when well-motivated, does not guarantee improvement. In multi-step reasoning chains, errors can accumulate, and a flawed "critic" can be more harmful than no critic at all. This reinforces the elegance of the Agora approach: it improves robustness through a conceptually simple and statistically grounded method rather than a complex and potentially brittle reasoning loop.

## 7. Limitations and Future Work

### 7.1. Limitations

Our study, while demonstrating the effectiveness of the Agora framework, is subject to several limitations that offer avenues for future work.

**Model and API Dependency:** Our zero-shot framework relies on proprietary LLMs (e.g., gpt-5-nano). This limits full reproducibility, as performance is tied to specific, closed-source model versions. Future work should explore the effectiveness

Table 3: Main Performance Comparison on the English Dataset (F1-Samples).

| System Configuration | F1-Samples Score | Improvement over Baseline |
|---|---|---|
| 1. Naive Baseline (Single-Pass) | 0.382 | — |
| 2. Actor-Critic Pipeline | 0.417 | +0.035 |
| 3. Agora (3-Agent Intersection) | **0.424** | **+0.042** |

Table 4: Ablation of Agora Configurations on the English Dataset (F1-Samples).

| Agora Configuration | F1-Samples Score | Improvement over Single Agent |
|---|---|---|
| 1. Single Agent (Vanilla) | 0.389 | — |
| 2. 3-Agent Union | 0.375 | -0.014 |
| 3. 3-Agent Majority Vote | 0.402 | +0.013 |
| 4. 3-Agent Intersection | **0.424** | **+0.035** |

of this ensemble approach with open-source models to ensure broader accessibility and control.

**Prompt Sensitivity:** The performance of any zero-shot system is highly sensitive to prompt engineering. While we standardized prompts across configurations, it is possible that the specific phrasing influenced the degree of stochasticity observed and the effectiveness of the voting mechanism.

**Computational Cost:** Deploying a multi-agent ensemble incurs a direct multiplication of inference cost and latency compared to a single-agent system. In our $N = 3$ configuration, the cost is roughly triple that of the baseline. While the performance gains justified this trade-off in a competitive setting, a critical area for future research is exploring cost-reduction techniques, such as using smaller, distilled models for some agents or implementing more sophisticated routing where an ensemble is only triggered for high-uncertainty cases.

### 7.2. Future Work

Based on our findings, we propose several directions for future research. First, exploring a weighted majority vote, where agents' votes are weighted by a confidence score, could further refine the consensus mechanism. Second, a hybrid approach that uses the fast Naive Baseline for simple cases and dynamically invokes the more robust Agora ensemble for difficult or ambiguous texts could optimize the cost-performance trade-off. Finally, applying the Agora framework to other complex, hierarchical domains beyond propaganda detection, such as legal text or biomedical literature analysis, would further validate its generalizability.

Beyond the zero-shot paradigm, a promising direction involves fine-tuning LLMs on task-specific data to achieve higher performance. However, this approach faces a significant challenge in the multilingual propaganda detection context: the need for high-quality labeled data across multiple languages while preserving cultural nuances. As noted in our prior work (Eljadiri and Nurbakova, 2025), machine translation of training data often loses cultural context and domain-specific subtleties, making direct translation-based data augmentation problematic. Future work must explore methods for generating culturally-aware training data, such as collaborative annotation with native speakers or culturally-grounded synthetic data generation, to enable effective fine-tuning while respecting the nuanced differences between languages.

## 8. Conclusion

In this work, we addressed the critical challenge of stochasticity and unreliability in Large Language Models when applied to complex Hierarchical Multi-Label Classification tasks. While the zero-shot capabilities of LLMs are powerful, their inconsistent outputs present a significant barrier to their deployment in real-world, high-stakes applications like propaganda detection.

We introduced Agora, a multi-agent ensemble framework designed to directly mitigate this issue. By leveraging the consensus of multiple independent LLM agents through a robust voting mechanism, Agora transforms a brittle, single-pass classification into a stable, statistically-grounded decision process.

Our experiments, conducted on the challenging SemEval-2025 narrative detection task, provided three key findings. First, we demonstrated that a naive single-pass LLM baseline is outperformed by more sophisticated architectures. Second, we showed that a seemingly intuitive Actor-Critic self-refinement pipeline can be counterproductive, as the critic's own unreliability can introduce noise and degrade performance. Finally, we proved that our Agora framework, using a 3-agent majority vote, delivers substantial and consistent performance gains over both other approaches.

The state-of-the-art performance of Agora in the official SemEval competition, including a first-place finish, validates our central claim: ensembling is a powerful and practical method for building more

robust and accurate zero-shot classification systems. As the field increasingly relies on powerful but imperfect LLMs, frameworks like Agora that prioritize reliability through consensus will be essential for creating trustworthy and deployable NLP solutions.

## Acknowledgements

We would like to thank the SemEval-2025 organizers for providing the challenging Task 10 on multilingual propaganda narrative detection, which served as an excellent testbed for our research. We are particularly grateful to Diana Nurbakova for her invaluable guidance and mentorship throughout our learning journey, which was instrumental in achieving these results.

## References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohamed Nour Eljadiri and Diana Nurbakova. 2025. Team INSALyon2 at SemEval-2025 task 10: A zero-shot agentic approach to text classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 965–980, Vienna, Austria. Association for Computational Linguistics.

W. Hu, Q. Fan, H. Yan, X. Xu, S. Huang, and K. Zhang. 2025. A survey of multi-label text classification under few-shot scenarios. *Applied Sciences*, 15:8872.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

LangChain AI. 2024a. Langgraph: Build resilient language agents as graphs. Accessed: 2025-09-19.

LangChain AI. 2024b. Langsmith: Developer platform for building reliable llm applications. Accessed: 2025-09-19.

Rundong Liu, Wenhan Liang, Weijun Luo, Yuxiang Song, He Zhang, Ruohua Xu, Yunfeng Li, and Ming Liu. 2023. Recent advances in hierarchical multi-label text classification: A survey.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval 2025)*, Vienna, Austria.

Tianyi Qin, Yaliang Wu, Yaliang Zhang, Weiliang Liu, Jundong Li, and Philip S. Yu. 2024. Infobench: A benchmark for information disorder detection in social media. *IEEE Transactions on Knowledge and Data Engineering*, 36(5):2132–2145.

J. Read, B. Pfahringer, G. Holmes, and E. Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Iknoor Singh, Carolina Scarton, and Kalina Bontcheva. 2025. GateNLP at SemEval-2025 task 10: Hierarchical three-step prompting for multilingual narrative classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 148–154, Vienna, Austria. Association for Computational Linguistics.

Vaishali S. Tidake and Shirish S. Sane. 2018. Multi-label classification: a survey. *International Journal of Engineering & Technology*, 7(4.19):1045.

Fengjun Wang, Moran Beladev, Ofri Kleinfeld, Elina Frayerman, Tal Shachar, Eran Fainman, Karen Lastmann Assaraf, Sarai Mizrachi, and Benjamin Wang. 2023. Text2Topic: Multi-label text classification system for efficient topic detection in user generated content with zero-shot capabilities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 93–103, Singapore. Association for Computational Linguistics.

Linli Xu, Sijie Teng, Ruoyu Zhao, Junliang Guo, Chi Xiao, Deqiang Jiang, and Bo Ren. 2021. Hierarchical multi-label text classification with horizontal and vertical category correlations. In *Proceedings of the 2021 Conference on Empirical*

*Methods in Natural Language Processing*, pages 2459–2468, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arjumand Younus and Muhammad Atif Qureshi. 2025. nlptuducd at SemEval-2025 task 10: Narrative classification as a retrieval task through story embeddings. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1742–1746, Vienna, Austria. Association for Computational Linguistics.

A. Zangari, M. Marcuzzo, M. Rizzo, L. Giudice, A. Albarelli, and A. Gasparetto. 2024. Hierarchical text classification and its foundations: A review of current research. *Electronics*, 13(7):1199.

Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.