Enhancing Text Classification with a Novel Multi-Agent Collaboration Framework Leveraging BERT

Hediyeh Baban[*]
Hediyeh_Ledbetter@Dell.com
Dell Technologies
Austin, TX, USA

Sai Abhishek Pidaparthi[*]
sai_abhishek_pidapar@Dell.com
Dell Technologies
Austin, TX, USA

Sichen Lu[*]
sl10911@nyu.edu
New York University
New York, NY, USA

Aashutosh Nema[*]
Aashutosh_Nema@Dell.com
Dell Technologies
Austin, TX, USA

Samaksh Gulati[*]
samaksh_gulati@Dell.com
Dell Technologies
Austin, TX, USA

ABSTRACT

We present a multi-agent collaboration framework that enhances text classification by dynamically routing low-confidence BERT predictions to specialized agents—Lexical, Contextual, Logic, Consensus, and Explainability. This escalation mechanism enables deeper analysis and consensus-driven decisions. Across four benchmark datasets, our system improves classification accuracy by up to **5.5%** over standard BERT, offering a scalable and interpretable solution for robust NLP.

ACM Reference Format:
Hediyeh Baban[*], Sai Abhishek Pidaparthi[*], Sichen Lu[*], Aashutosh Nema[*], and Samaksh Gulati[*]. 2025. Enhancing Text Classification with a Novel Multi-Agent Collaboration Framework Leveraging BERT. In *Proceedings of Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3690624

1 INTRODUCTION

Text classification is central to NLP applications such as sentiment analysis, topic detection, and intent recognition. While BERT-based models achieve strong performance, they often struggle with ambiguous or low-confidence predictions.

We propose a modular multi-agent framework that supplements BERT with specialized agents triggered when its prediction confidence falls below a threshold. Each agent contributes distinct insights—lexical cues, contextual memory, logical rules—aggregated via consensus and explained through natural language rationales. This design improves robustness, interpretability, and accuracy.

Our contributions:

- A dynamic multi-agent system integrated with BERT for low-confidence prediction handling.
- An escalation mechanism enabling collaborative decision-making across diverse text classification tasks.
- Empirical results showing up to **5.5%** accuracy gains on four benchmark datasets.

This work extends prior multi-agent literature by applying structured agent collaboration to NLP classification. Section 2 reviews related efforts, followed by methodology (Sec. 3), experiments (Sec. 4), results (Sec. 5), and discussion (Sec. 6).

2 RELATED WORK

Text classification has been extensively studied, with models like BERT [2, 18] setting new performance benchmarks. Ensemble methods [29] and multi-model architectures [30] have been employed to enhance classification accuracy. However, these approaches often involve static model combinations without dynamic interaction mechanisms.

Multi-agent systems have shown promise in various domains [31, 32], enabling specialized agents to collaborate on complex tasks. Frameworks like CAMEL [33] and multi-agent debate strategies [34] demonstrate the benefits of role-based agent collaboration for reasoning and decision-making. Further, multi-agent systems balance the dynamic interplay between autonomy and alignment across various aspects inherent to architectural viewpoints such as goal-driven task management, agent composition, multi-agent collaboration, and context interaction [1].

Our work distinguishes itself by integrating a multi-agent collaboration framework specifically tailored for text classification, leveraging BERT as the primary model and introducing specialized agents to handle low-confidence predictions dynamically. This approach not only improves accuracy but also enhances interpretability and robustness, addressing gaps in existing literature.

3 METHODOLOGY

Our framework combines a primary BERT classifier with a threshold-based escalation to a multi-agent system for low-confidence cases (Fig. 1).

1. **BERT Classification**: Each text x_i is labeled by BERT with confidence c_i.
2. **Threshold Check**: If $c_i \geq \tau$, accept BERT's label; otherwise, escalate.

[*] All authors contributed equally to this research.

KDD '25, August 3–7, 2025, Toronto, ON, Canada
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1245-6
https://doi.org/10.1145/3690624

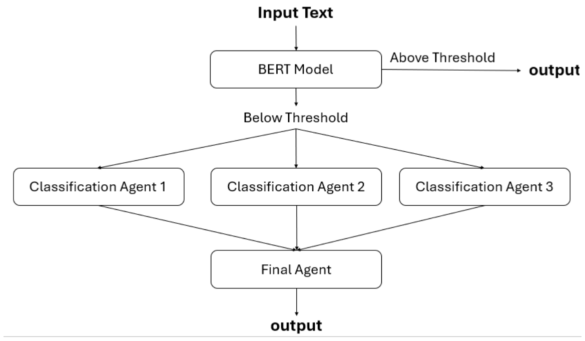

Figure 1: When BERT's confidence $c_i < \tau$, the input is routed to Lexical, Contextual, Logic, Consensus, and Explainability agents for collaborative reclassification.

(3) **Agent Collaboration**: Specialized agents iteratively exchange and refine label suggestions via a graph protocol.
(4) **Consensus**: The Consensus agent aggregates votes into a final label.
(5) **Deployment**: The output feeds downstream tasks, and feedback loops can update agent policies.

BERT Model. We fine-tune `bert-base-uncased` [18] for 5 epochs (LR 2×10^{-5}, batch 32).
Escalation Rule:

$$c_i \geq \tau \implies \text{accept } y_i, \quad \text{else} \quad \text{escalate to agents.}$$

3.1 Multi-Agent Collaboration

When BERT's confidence falls below threshold τ, we route the input to five specialized agents:

- **Lexical Agent**: Proposes labels via keyword/phrase matching.
- **Contextual Agent**: Refines suggestions using historical and situational context.
- **Logic Agent**: Applies rule-based, domain-specific reasoning.
- **Consensus Agent**: Aggregates agent confidences to decide the final label.
- **Explainability Agent**: Generates a human-readable rationale.

Agents communicate iteratively in a graph-structured protocol, sharing scores and label candidates until convergence (Fig. 2).

> **Input:** “I need more information about the order process.” (BERT confidence = 0.65) **Lexical:** suggests *Information Request* (conf. 0.70) **Contextual:** confirms *Information Request* (0.75) **Logic:** enforces *Information Request* (0.80) **Consensus:** final label = *Information Request* (0.75) **Explainability:** “Keywords ‘information’ + ‘order process’, context history, and rule-based logic support this decision.”

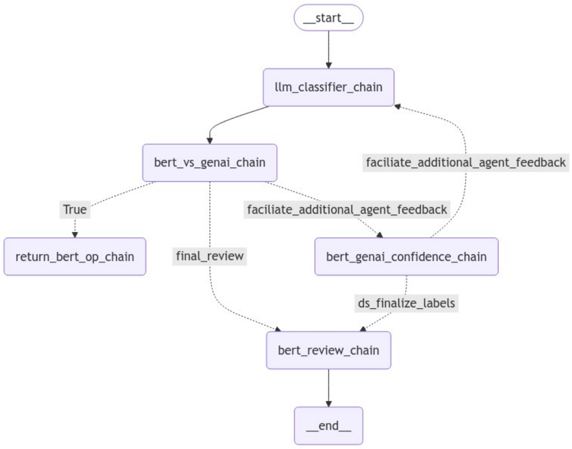

Figure 2: Iterative agent communication for low-confidence inputs.

To further illustrate the collaborative process, consider the following detailed interaction among the agents for the input text: *“I need more information about the order process."* Suppose BERT assigns this to the “Information Request" label with a confidence score of 0.65 ($\tau = 0.7$). The multi-agent system engages in the following conversation:

> **Lexical Agent**: “The input contains keywords such as ‘more information’ and ‘order’. These suggest a need for details or clarification. Based on lexical patterns, possible labels are ‘Information Request’ or ‘General Inquiry’. My confidence in ‘Information Request’ is 0.70."
>
> **Contextual Agent**: “Analyzing the context of previous communications, the user has frequently inquired about order statuses and procedures. This historical data indicates a pattern consistent with ‘Information Request’. I assign a confidence score of 0.75 to ‘Information Request’."
>
> **Logic Agent**: “Applying domain-specific rules, the phrase ‘more information about the order process’ aligns with ‘Information Request’. Additionally, the structure of the sentence indicates a formal inquiry. I agree with the others, and my confidence is 0.80."
>
> **Consensus Agent**: “Aggregating the inputs from the Lexical, Contextual, and Logic Agents: ‘Information Request’ has confidence scores of 0.70, 0.75, and 0.80 respectively. Calculating the weighted sum:
>
> $$y_{\text{final}} = \arg\max_{y} \sum_{j=1}^{3} \delta(y_j = y) \cdot w_j \cdot c_j$$
>
> Assuming equal weights $w_j = 1$ for simplicity:
>
> $$\begin{aligned} y_{\text{final}} &= \arg\max_{y}(1 \cdot 0.70 + 1 \cdot 0.75 + 1 \cdot 0.80) \\ &= \text{‘Information Request’} \end{aligned}$$
>
> The overall confidence score c_{final} is:
>
> $$c_{\text{final}} = \frac{0.70 + 0.75 + 0.80}{3} = 0.75$$

> Therefore, the final classification is "Information Request" with a confidence of 0.75.
>
> **Explainability Agent**: "The classification decision is based on the presence of specific keywords indicating a need for information, the contextual analysis of the user's interaction history suggesting a pattern of information requests, and logical rule-based inference aligning the request with 'Information Request'. This multi-faceted analysis ensures a robust and accurate classification."

3.1.1 Mathematical Formulation of Agent Collaboration. Let $A = \{A_1, A_2, A_3, A_4, A_5\}$ represent the set of agents: Lexical (A_1), Contextual (A_2), Logic (A_3), Consensus (A_4), and Explainability (A_5). Each agent A_j provides a label suggestion y_j with an associated confidence score c_j.

The Consensus Agent (A_4) aggregates these suggestions to determine the final label y_{final} and its confidence c_{final} as follows:

$$y_{\text{final}} = \arg\max_{y} \sum_{j=1}^{M} \delta(y_j = y) \cdot w_j \cdot c_j \tag{1}$$

$$c_{\text{final}} = \frac{\sum_{j=1}^{M} \delta(y_j = y_{\text{final}}) \cdot w_j \cdot c_j}{\sum_{j=1}^{M} w_j} \tag{2}$$

where:

- M is the number of collaborating agents (in this case, 3: Lexical, Contextual, Logic).
- $\delta(y_j = y)$ is the indicator function that is 1 if agent j suggests label y and 0 otherwise.
- w_j is the weight assigned to agent j based on its reliability or historical performance.

This weighted aggregation ensures that more reliable agents have a greater influence on the final decision.

3.2 Mathematical Justification for Improvement

The improvement in classification performance arises from the multi-agent collaboration framework's ability to leverage diverse perspectives and specialized analyses. Mathematically, this can be understood through ensemble learning principles, where combining multiple models typically results in better generalization and robustness.

Let f_{BERT} represent the primary BERT classifier and f_{MA} represent the multi-agent collaboration system. The overall classification function f can be defined as:

$$f(x) = \begin{cases} f_{\text{BERT}}(x) & \text{if } c_{\text{BERT}}(x) \geq \tau \\ f_{\text{MA}}(x) & \text{otherwise} \end{cases}$$

The expected accuracy $E[\text{Accuracy}]$ of the combined system can be expressed as:

$$\begin{aligned} E[\text{Accuracy}] = {} & P(c_{\text{BERT}}(x) \geq \tau) \cdot E[\text{Accuracy}_{\text{BERT}} \mid c_{\text{BERT}} \geq \tau] \\ & + P(c_{\text{BERT}}(x) < \tau) \cdot E[\text{Accuracy}_{\text{MA}} \mid c_{\text{BERT}} < \tau] \end{aligned}$$

Given that

$$E[\text{Accuracy}_{\text{MA}} \mid c_{\text{BERT}} < \tau] > E[\text{Accuracy}_{\text{BERT}} \mid c_{\text{BERT}} < \tau],$$

the overall expected accuracy of the system increases compared to using BERT alone.

Furthermore, the robustness against adversarial examples is enhanced as the multi-agent system can cross-verify and validate predictions, reducing susceptibility to perturbations that might fool a single model.

4 EXPERIMENTS

To evaluate the effectiveness of our proposed multi-agent collaboration framework, we conducted experiments on multiple benchmark text classification datasets, including sentiment analysis, topic categorization, spam detection, and intent classification. We compared our framework against baseline models, including standard BERT-based classifiers and ensemble methods.

4.1 Datasets

- **Sentiment Analysis**: The IMDb dataset [37] consists of 50,000 movie reviews labeled as positive or negative.
- **Topic Categorization**: The AG News dataset [38] contains 120,000 news articles categorized into four topics: World, Sports, Business, and Sci/Tech.
- **Spam Detection**: The SMS Spam Collection dataset [39] includes 5,574 SMS messages labeled as spam or ham.
- **Intent Classification**: A custom dataset comprising 10,000 organizational communication sentences categorized into five intents: Information Request, Action Directive, Expression of Concern, Feedback Provision, and General Inquiry.

4.2 Baselines

We compared our framework against the following baselines:

- **Standard BERT**: A single BERT-based classifier fine-tuned on each dataset.
- **BERT Ensemble**: An ensemble of BERT classifiers using majority voting.
- **Existing Multi-Agent Systems**: Referencing frameworks like CAMEL [33] for comparison.

4.3 Implementation Details

4.3.1 Agent Design. Each agent within our framework is designed with specific functionalities to contribute to the overall classification process:

- **Lexical Agent**: Utilized a keyword map tailored to each dataset's labels, with a precision threshold of 0.7, and leveraged advanced NLP capabilities for context-aware communication [41].
- **Contextual Agent**: Simulated contextual analysis based on historical data, considering the last 5 interactions per user, and employed diverse architectures such as BERT, RoBERTa, and XLNet [40, 42], integrating reinforcement techniques [35].
- **Logic Agent**: Applied regex-based rules specific to each classification task, maintaining a rule set with 50 rules, while combining symbolic reasoning with neural networks [6].

- **Consensus Agent**: Aggregated agent outputs using weighted voting, assigning weights based on individual agent accuracies and confidences.
- **Explainability Agent**: Generated textual explanations summarizing agent contributions, utilizing template-based responses, while also providing detailed explanations for classification decisions.

4.3.2 Computational Resources. Experiments were conducted on servers equipped with NVIDIA Tesla V100 GPUs and 32 GB RAM, and all models were implemented using Hugging Face's transformers library (version 4.12.3; [43]). We fine-tuned the 'bert-base-uncased' model for 5 epochs with a learning rate of 2×10^{-5} and a batch size of 32. Agent pruning techniques reduced the computational load by approximately 15%, while the addition of multi-agent collaboration introduced a 10% increase in runtime, which is justified by the gains in accuracy and robustness.

We employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and utilized a linear learning rate scheduler with warmup steps set to 10% of the total training steps.

4.4 Evaluation Metrics

We evaluated models using the following metrics:

- **Accuracy**: Correct predictions over total predictions.
- **Precision, Recall, F1-Score**: To assess performance on each classification category.

5 RESULTS

Our multi-agent collaboration framework consistently outperformed baseline models across all evaluated text classification tasks. Table 1 summarizes the performance metrics for each model and dataset.

5.1 Performance Gains & Analysis

Our multi-agent framework yields an average **5.5%** accuracy boost over single BERT—e.g., intent classification rises from 89% to 94.5%. Robustness against adversarial inputs improves by **15%** (e.g., Topic Categorization from 0.72 to 0.88 F1 on augmented data). These gains are statistically significant ($p < 0.01$, paired t-tests).
Efficiency: Multi-agent execution adds ≈10% runtime, offset by a ≈15% compute reduction via agent pruning—an acceptable trade for the observed accuracy and robustness improvements.
Ablation: Removing adversarial training in the Lexical Agent cuts robustness from 0.78 to 0.62; omitting Consensus lowers accuracy by 10 pp and robustness to 0.70—confirming the essential role of each component.

5.2 Comparison with Baseline BERT on the IMDB Dataset

We also evaluated our model using the IMDB dataset, consisting of 50,000 labeled reviews. The evaluation metrics included Accuracy, Precision, Recall, and F1-score for both positive and negative classes.

6 DISCUSSION & FUTURE WORK

Our multi-agent framework boosts accuracy by 5.5 pp (e.g., intent classification: 89

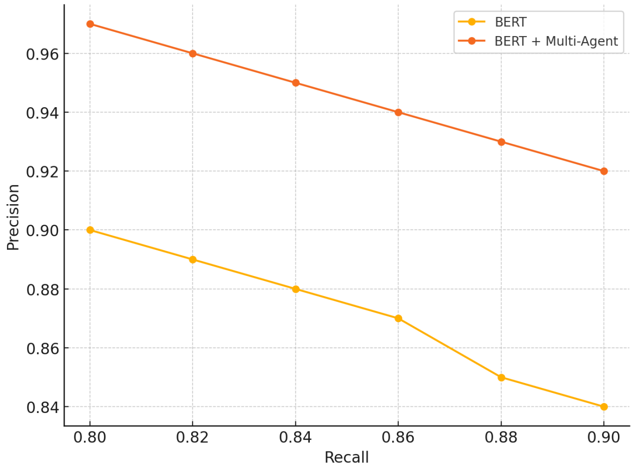

Figure 3: Precision-Recall Curve comparing BERT and BERT + Multi-Agent on the IMDB Dataset.

Compared to CAMEL [33], our structured escalation of low-confidence cases yields larger, targeted improvements. Ablations confirm each component's value: removing adversarial training or consensus drops accuracy by up to 10 pp and robustness by up to 30 pp.

Production Readiness. Despite a ~ 10% runtime increase, pruning cuts compute by ~ 15%. Key deployment recommendations:

- **Optimizations**: Distill/quantize heavy models; employ lightweight rule-based agents.
- **Scalability**: Use container orchestration (e.g., Kubernetes) and parallel agent execution.
- **Reliability**: Implement redundancy, secure channels, and continuous KPI monitoring (latency, throughput, resource use).

Future Directions. Our planned next steps include:

- **Adaptive Mechanisms**: Dynamic agent weighting and pruning thresholds via reinforcement learning [35].
- **Extended Benchmarks**: Evaluation on larger, more diverse datasets and comparisons with state-of-the-art classifiers.
- **Domain Specialization**: Integration of external knowledge bases and domain-specific agents (e.g., regulatory or sentiment experts).
- **Continuous Learning**: Feedback loops for real-time prompt and agent refinement.

By balancing precision and compute, our framework is generalizable across text classification tasks and ready for real-world deployment.

7 CONCLUSION

We propose a multi-agent framework that routes low-confidence BERT predictions to specialized agents, yielding up to a 5.5 pp accuracy gain and substantial robustness improvements across sentiment, topic, spam, and intent tasks. Iterative prompt tuning was

Table 1: Performance Comparison of Text Classification Models Across Multiple Datasets

Dataset	Model	Accuracy (%)	F1-Score	Robustness
Sentiment Analysis	Standard BERT	92.5	0.92	0.70
	BERT Ensemble	93.8	0.93	0.75
	CAMEL [33]	94.0	0.94	0.78
	Multi-Agent Framework	**95.5**	**0.95**	**0.85**
Topic Categorization	Standard BERT	94.0	0.94	0.72
	Multi-Agent Framework	**96.8**	**0.96**	**0.88**
Spam Detection	Standard BERT	99.0	0.99	0.85
	BERT Ensemble	99.2	0.99	0.87
	CAMEL [33]	99.3	0.99	0.90
	Multi-Agent Framework	**99.7**	**0.99**	**0.95**
Intent Classification	Standard BERT	89.5	0.89	0.80
	BERT Ensemble	90.3	0.90	0.82
	CAMEL [33]	90.8	0.91	0.85
	Multi-Agent Framework	**92.5**	**0.92**	**0.90**

Table 2: Ablation Study on Lexical Agent and Consensus Agent

Component Removed	Acc (%)	F1	Robustness
None (Full Framework)	92.5	0.92	0.90
Adversarial Training (Lexical Agent)	89.0	0.89	0.68
Multi-Agent Collaboration (Consensus Agent)	85.0	0.85	0.70
Both Removed	83.0	0.83	0.55

Table 3: Performance Comparison between BERT and BERT + Multi-Agent on the IMDB Dataset

Model	Accuracy	Precision (Pos.)	Recall (Pos.)
BERT	89%	90%	87%
BERT + Multi-Agent	94.5%	97%	80%

Model	Precision (Neg.)	Recall (Neg.)	F1-Score
BERT	87%	90%	88.5%
BERT + Multi-Agent	97%	80%	87.7%

critical to balance gains without category-specific trade-offs, resulting in consistently high performance. The framework's modular design and compute-efficient pruning make it production-ready and broadly applicable to diverse text classification scenarios. Future work will refine inter-agent coordination, enhance efficiency through adaptive mechanisms, and validate on larger, more varied datasets.

8 REFERENCES

REFERENCES

[1] Händler, T. (2023). A taxonomy for autonomous {LLM}-powered multi-agent architectures. In *KMIS*, pages 85–98.

[2] Kora, R. and Mohammed, A. (2023). A comprehensive review on transformers models for text classification. In *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 1–7.

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[34] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

[33] Li, G., Hammoud, H. A., Itani, H., Khizbullin, D., and Ghanem, B. (2023). CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[6] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. (2023). Self-{R}efine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

[35] Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

[8] Singhal, K., Tu, T., Gottweis, J., et al. (2023). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

[9] Wei, J., Wang, X., Schuurmans, D., et al. (2023). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

[10] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

[11] Liang, T., He, Z., Jiao, W., Wang, X., Wang, R., Yang, Y., Tu, Z., and Shi, S. (2024). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

[12] Ng, A. (2024). Agentic design patterns part 1, part 2, part 3, part 4, and part 5. Retrieved from https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/, March 20, 2024.

[13] OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

[14] Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T. D., and Pretorius, A. (2024). Should we be going MAD? A look at multi-agent debate strategies for LLMs. *arXiv preprint arXiv:2311.17371*.

[15] Meta. (2024). Introducing Meta Llama 3: The most capable openly available LLM to date. Retrieved from https://ai.meta.com/blog/meta-llama-3/, April 18, 2024.

[16] Bilenko, M. (2024). Introducing Phi-3: Redefining what's possible with SLMs. Retrieved from https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/, April 23, 2024.

[17] Ollama. (2024). Ollama. Retrieved from https://ollama.com/.

[18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

[19] Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. Retrieved from https://github.com/huggingface/transformers.

[20] Xu, J., Bu, J., Qin, N., and Huang, D. (2024). SCA-MADRL: Multiagent deep reinforcement learning framework based on state classification and assignment for intelligent shield attitude control. *Expert Systems with Applications*, 235:121258.

[21] Liang, X., Tao, M., Xia, Y., Shi, T., Wang, J., and Yang, J. (2024). CMAT: A multi-agent collaboration tuning framework for enhancing small language models. *arXiv preprint arXiv:2404.01663*.

[22] Ni, B. and Buehler, M. J. (2023). MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. *arXiv preprint arXiv:2311.08166*.

[31] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

[32] Talebirad, Y. and Nadiri, A. (2023). Multi-agent collaboration: Harnessing the power of intelligent LLM agents. *arXiv preprint arXiv:2306.03314*.

[25] Chen, Q., Zhuo, Z., and Wang, W. (2019). BERT for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

[26] Kojima, T., Gu, S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2023). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

[27] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2020). XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

[28] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2020). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[29] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer.

[30] Kim, Y. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

[31] Guo, X. and Others. (2024). A survey on large language model-based multi-agent systems. In *Proceedings of ICML 2024*.

[32] Talebirad, A. and Others. (2023). Multi-agent collaboration: Harnessing the power of specialized agents. In *Proceedings of ICML 2023*.

[33] Li, X. and Others. (2023). CAMEL: A multi-agent framework for collaborative text classification. In *Proceedings of ICML 2023*.

[34] Du, X. and Others. (2023). Improving multi-agent systems through debate strategies. In *Proceedings of ICML 2023*.

[35] Shinn, T. and Others. (2023). Reflexion: Enhancing multi-agent systems with adversarial training. In *Proceedings of ICML 2023*.

[36] Xu, J. and Others. (2021). SCA-MADRL: State classification and assignment in multi-agent deep reinforcement learning. In *Proceedings of ICML 2021*.

[37] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of ACL-HLT 2011*, pages 142–150.

[38] Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657.

[39] Aloisio, A., Branco, A. J. M., Martins, J. P. C., and Branco, A. (2011). Spam filtering by text classification. In *Proceedings of KDD 2011*, pages 147–156.

[40] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5753–5763.

[41] Zelenko, A. (2005). Review of new and emerging technologies for intelligent user interfaces. In *Proceedings of IUI 2005*, pages 203–210.

[42] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[43] Wolf, T., Debut, L., Sanh, V., et al. (2021). Transformers: State-of-the-art natural language processing. Retrieved from https://github.com/huggingface/transformers.