Reliable Decision-Making for Multi-Agent LLM Systems

Xian Yeow Lee, Shunichi Akatsuka, Lasitha Vidyaratne, Aman Kumar, Ahmed Farahat, Chetan Gupta

Industrial A.I. Laboratory, Hitachi America, Ltd.

HITACHI
Inspire the Next

Introduction

- **What are Multi-Agent LLM Systems?**
  - Collaborative AI systems leveraging multiple Large Language Models (LLMs).
  - Applications in logistics, robotics, and industrial decision-making.
- **Why Reliability Matters**
  - High-stakes environments (supply chains, emergency response) require **consistent** performance.
  - Complex architectures risk error propagation and reduces robustness.
- **Research Focus**
  - Investigate how different **aggregation strategies** impact **reliability.**

Multi-Agent Architectures

- **Single Agent (Baseline)** – One LLM agent making independent decisions.
- **Majority Voting** – Aggregates multiple LLM agents' outputs based on majority consensus.
- **Averaging** – Computes the mean of LLM agent's numerical outputs.
- **Decentralized** – LLM Agents iteratively refine responses until consensus.
- **Decentralized (Feedback)** –LLM agents incorporate prior responses into iterations.
- **Spoke & Wheel** – Central "hub" LLM agent integrates independent LLM agents' decisions.
- **Spoke & Wheel (Feedback)** – Central LLM agent's feedback guides future responses of multiple independent LLM agents

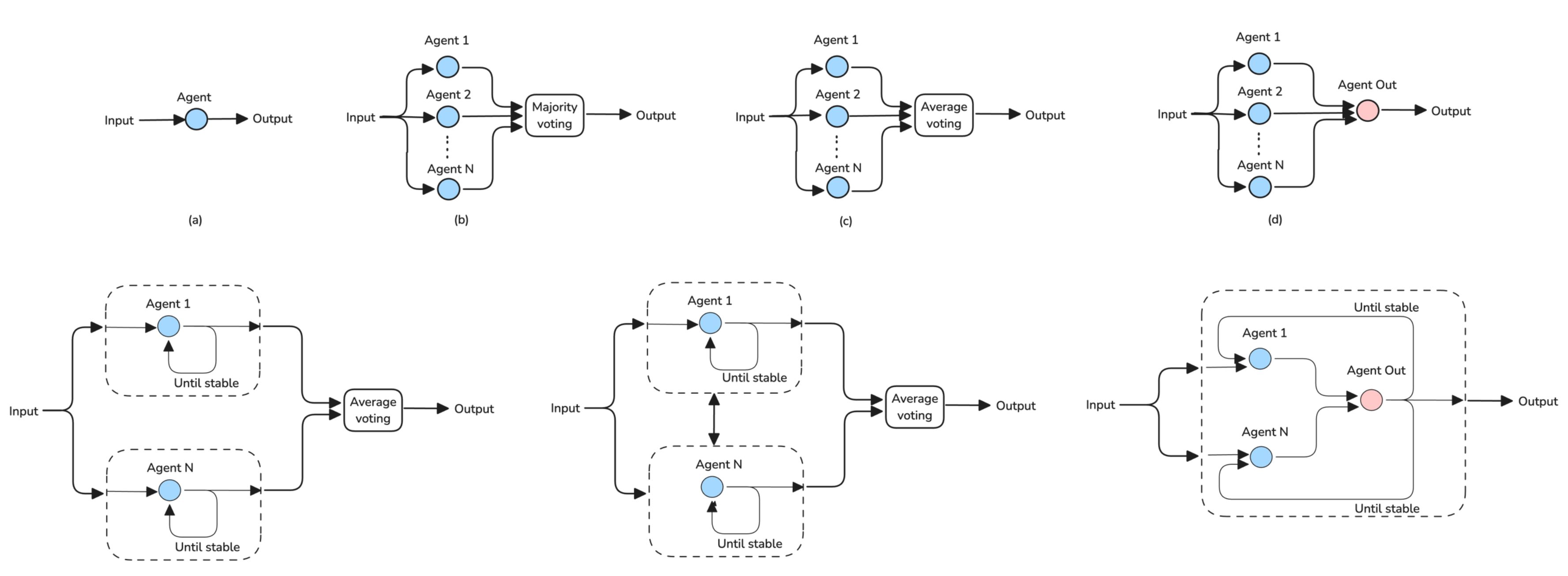

Figure 1: Illustration of different output aggregation strategies

Experimental Setup

- **Tasks Evaluated:**
  - **Resource Allocation** – Distributing limited resources across regions.
  - **Question Answering** – Answering SQuAD 2.0 questions with specific formatting.
  - **Topic Classification** – Categorizing news articles into predefined topics.
  - **Text Summarization** – Generating concise summaries from news articles.
- **Evaluation Metrics:**
  - **Task-specific Performance Metrics (S):** Allocation satisfaction, accuracy, correctness, ROUGE scores
  - **Reliability Metric κ(τ)** – Measures consistency across multiple trials at threshold τ.
  - **Area under Reliability Curve (AURC)** – Measures area under the reliability curve for reliability metric across all thresholds

$$\kappa(\tau) = \frac{\sum_{t=1}^{T} \mathbb{I}\left(S^{(t)} \geq \tau\right)}{T}$$

$S^{(t)}$ = Task – specific performance metric at trial t

T = Number of total trials

τ = Performance threshold

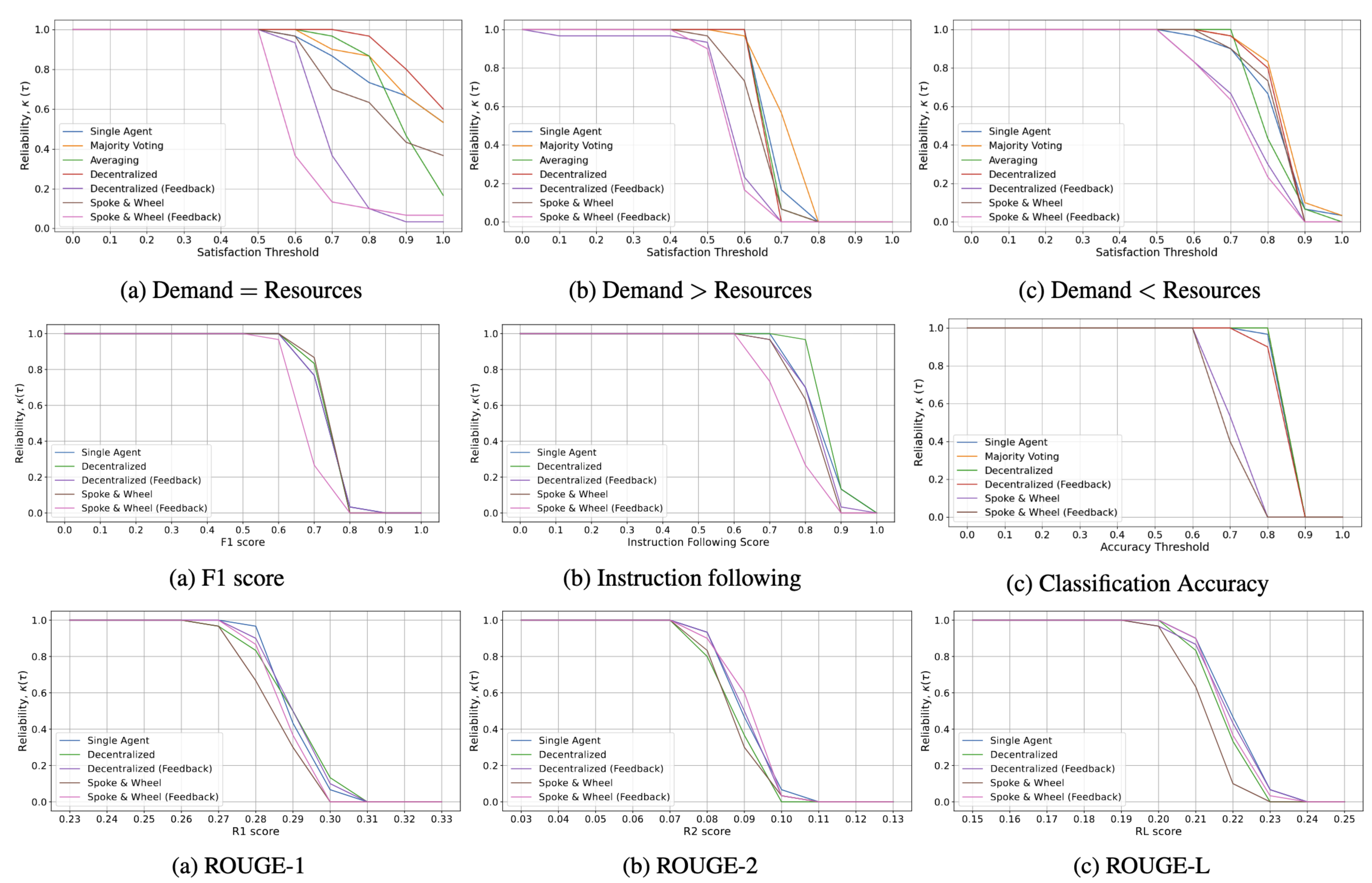

Figure 2: Reliability curves for evaluated task. Row 1: Resource allocation. Row 2: Question Answering & Topic Classification, Row 3: Text Summarization

			Single Agent	Majority Voting	Avera-ging	Decen-tralized	Feedback Decen-tralized	Spoke & Wheel	Feedback Spoke & Wheel
Resource Allocation	Small	Equal	0.999	**1.000**	0.999	**1.000**	0.999	**1.000**	0.988
		Lack	0.680	0.688	0.654	0.699	0.688	0.702	**0.706**
		Excess	**0.709**	0.669	0.705	0.665	0.702	0.669	0.666
	Medium	Equal	0.897	0.915	0.890	**0.961**	0.695	0.835	0.627
		Lack	0.689	**0.691**	0.668	0.673	0.534	0.623	0.532
		Excess	0.802	**0.826**	0.801	0.816	0.729	0.805	0.722
	Large	Equal	0.640	**0.711**	0.626	0.623	0.531	0.567	0.447
		Lack	0.564	**0.671**	0.593	0.574	0.532	0.559	0.424
		Excess	0.672	**0.816**	0.682	0.648	0.582	0.579	0.450
Question Answering		Correctness	0.722	–	–	**0.732**	0.725	0.729	0.670
		Instruction Following	0.824	–	–	**0.838**	0.811	0.801	0.744
Classification		Accuracy	0.831	**0.833**	–	**0.833**	0.826	0.704	0.686
Text Summarization		ROUGE-1	**0.290**	–	–	0.289	0.289	0.284	0.287
		ROUGE-2	0.089	–	–	0.087	0.089	0.087	**0.090**
		ROUGE-L	**0.219**	–	–	0.217	0.218	0.213	0.218

Table 1: Summary of AURC for all experiments

Key Findings

Resource Allocation:

- Majority Voting and Decentralized methods consistently achieved higher reliability.
- Feedback-based approaches amplified errors, reducing robustness.

Question Answering & Topic Classification:

- Decentralized & Majority Voting approaches improved performance and consistency.
- Spoke & Wheel methods performed the worst due to over-dependence on a central agent.

Text Summarization:

- No significant difference between aggregation strategies due to limitations in ROUGE evaluation.

Key Takeaways

- **Simplicity Outperforms Complexity** – Majority Voting and Decentralized methods provide higher reliability.
- **Feedback Loops Can Hurt Reliability** – Risk of error propagation in iterative feedback mechanisms.
- **Redundancy is Key** – Independent decision-making prevents system-wide failures.
- **Evaluation Metrics Matter** – Traditional NLP metrics may not capture reliability effectively.

Conclusion & Future Work

- Majority Voting & Decentralized strategies offer the best balance of accuracy and reliability.
- Future research:
  - Better aggregation strategies for tasks where simple voting isn't feasible.
  - Advanced evaluation metrics to better assess reliability