From Coordination Failure to Scalable AI SWARMS:

The MSCFT Protocol for Structured Forecasting and Multi-Agent Alignment Among Large Language Model Systems

Brian Helip

Santa Monica College

Phi Theta Kappa International Honor Society

This paper argues that while large language model (LLM) agents possess the potential to coordinate at scales far beyond human capability, they lack the structural discipline necessary to do so reliably. Drawing on insights from recent research into agent societies and swarm-based behavior, this work presents the Master SWARM Consensus Forecasting Template (MSCFT) as a functional protocol that enables transparent, auditable, and aligned reasoning across multiple AI agents. By providing a shared cognitive and formatting framework, MSCFT operationalizes what swarm theory lacks: a stable infrastructure for scalable, multi-agent intelligence.

I. Introduction

Large language models (LLMs) have rapidly evolved from static text generators to dynamic reasoning systems capable of answering questions, summarizing complex content, and even simulating dialogue across multiple domains. Since the release of transformer-based architectures like GPT-2 and GPT-3, these models have been increasingly embedded in decision-making workflows in business, government, and scientific research. LLMs can now generalize tasks and interact with users, making them cognitive systems capable of supporting or replacing aspects of human reasoning.

As these models grow more capable, so too does the ambition to network them into larger coordinated systems. Multi-agent frameworks, sometimes referred to as “AI swarms,” envision groups of LLMs reasoning, debating, and collaborating in parallel. These systems promise a form of scalable collective intelligence: one in which AI agents, rather than humans, perform coordinated forecasting, analysis, and judgment under structured protocols. But with this promise comes a critical question: can LLM agents reliably align their reasoning without breaking down into inconsistency, error propagation, or decision drift?

In April 2025, researchers published a pivotal study (arXiv:2409.02822v4) investigating the limitations of large-scale LLM coordination using swarm-like simulations. The findings were striking beyond a certain scale; AI agent societies became unstable without external structure. Opinion groups fractured, consensus decayed, and even highly capable models failed to maintain alignment across iterations. The paper served as a warning: that emergent reasoning is not enough—and that any serious use of LLM swarms requires a formal structure to stabilize reasoning and enforce epistemic clarity across agents.

As LLMs are deployed in groups, whether for reasoning tasks, collaborative planning, or distributed decision-making, they face escalating challenges of coordination. Without a shared input structure, agents may interpret the same question differently. Without a standard output format, they may produce responses that cannot be compared, ranked, or synthesized. And without transparency into how each agent reaches its conclusion, even agreement can mask error. These limitations are not merely inconvenient; they pose fundamental threats to trust, accountability, and the use of AI in high stakes forecasting environments such as medicine, defense, and public policy.

In April 2025, a team of researchers released a study on arXiv (arXiv:2409.02822v4) that modeled the behavior of LLM-based agent societies using swarm coordination methods. The paper found that beyond a critical group size, AI agents began to fragment into smaller factions, unable to maintain stable consensus. Coordination decayed rapidly without explicit structural constraints. Despite the growing sophistication of models like GPT-4, Claude, and LLaMA, the researchers showed that scale alone does not yield reliable group behavior—only structure does.

This insight became the catalyst for the development of the Master SWARM Consensus Forecasting Template (MSCFT). Rather than rely on emergent coordination, MSCFT enforces a disciplined structure that standardizes input, reasoning, and output across LLM agents. It provides the scaffolding necessary for multi-agent systems to coordinate transparently, auditably, and at scale. By implementing a common cognitive framework, MSCFT bridges the gap between theoretical agent societies and real-world forecasting reliability.

II. Background and Motivation

The idea for the Master SWARM Consensus Forecasting Template (MSCFT) began not with theory, but with experience. While working through structured forecasting exercises and observing how large language models (LLMs) performed across different environments, it became clear that even highly capable systems like GPT-4, Claude, and Gemini lacked consistent alignment when asked to perform identical reasoning tasks in parallel. These inconsistencies were not due to errors in capability, but there was a lack of structure; there was no fixed format for input framing, evidence logging, or probability assignment. These gaps opened the door to noise, drift, and failure to converge, even on well-defined questions.

A turning point came after reviewing the 2025 Swarm coordination paper by De Marzo, Castellano, and Garcia (2025), which showed through simulation that agent societies, even when composed of powerful LLMs, become unstable past a certain group size unless strict coordination rules are applied. The failure modes identified in that paper—fragmentation, emergent bias, and rational collapse—mirrored the same failure patterns encountered in practical forecasting. The message was simple: LLMs can coordinate, but only under constraint. Swarms don’t scale unless structure is enforced.

This message was reinforced during Zoom-based seminars hosted by the Centre for Cognition, Computation and Modelling (CCCM) at Birkbeck College. In these sessions, researchers explored the cognitive limitations of generative AI models, especially their difficulty with internal consistency, source tracking, and self-reflective reasoning. What stood out was the repeated observation that LLMs are not inherently epistemic; they do not understand *why* they’re right or wrong. When placed in multi-agent settings, these cognitive gaps compound. It became clear that no matter how advanced the model, coordination would require more than emergent interaction. It would require a protocol.

These threads; Swarm limitations, CCCM cognitive research, and firsthand experience with reasoning failures—came together to inspire MSCFT. The goal was to produce a repeatable, format-enforcing structure that could allow different LLM agents (or humans) to operate in parallel while remaining transparent, auditable, and aligned. It was not just a formatting tool. It was designed to be a scaffolding for thought.

III. Core Challenges in Multi-Agent LLM Systems

One of the most critical challenges facing multi-agent deployments of large language models (LLMs) is their tendency toward opinion fragmentation, epistemic drift, and rational inconsistency. While individual LLMs can generate plausible outputs, their coordination across parallel agents often leads to divergence. In forecasting tasks, for example, slight differences in prompt interpretation, stochastic token selection, or framing bias can cause agents to arrive at mutually incompatible conclusions; even when given the same question. As more agents are introduced, this divergence compounds, resulting in clusters of models that reinforce one another’s errors, overlook key assumptions, or omit crucial counterarguments. These breakdowns mimic cognitive failure modes observed in human groups but occur more rapidly and without awareness. This breakdown underscores that emergent coordination is fundamentally different and less reliable than enforced structural alignment.

While some researchers have explored ensemble methods that assume convergence will emerge naturally through repeated interaction or averaging, practical results consistently show the opposite; in the absence of a shared structure, LLM swarms tend to destabilize. De Marzo, Castellano, and Garcia (2025) demonstrated this formally, showing that beyond a certain group size, model collectives cannot maintain consensus without structural enforcement. Unlike human teams that can fall back on shared memory, norms, or interpersonal feedback, LLMs require explicit formatting to anchor their reasoning and make their internal logic accessible.

These structural gaps have become especially problematic in high-stakes domains such as medicine, public policy, and critical forecasting. In such contexts, divergence isn't just inconvenient, it is dangerous. When multiple agents produce conflicting risk assessments or incompatible conclusions about the same scenario, trust is eroded, accountability is obscured, and downstream decisions may be compromised. Transparency, traceability, and auditability are not optional in these domains; they are foundational. Without a protocol to surface assumptions, reveal uncertainty, and enable side-by-side comparison, coordination among LLMs becomes unscalable and opaque. These realities underscore the urgent need for forecasting templates that not only improve performance but encode structure at the heart of multi-agent reasoning.

IV. The MSCFT Framework

The Master SWARM Consensus Forecasting Template (MSCFT) was designed to solve a specific problem: how to enable large language model (LLM) agents to reason independently while still producing outputs that are logically consistent, epistemically transparent, and structurally comparable. The goal was not simply to standardize formatting, but to create a protocol for cognitive alignment—one that could scale across multiple agents and high-stakes domains. The MSCFT achieves this by enforcing mirrored reasoning steps, embedding diagnostic prompts, and producing a clean output format optimized for both human and machine evaluation.

Each component of the template serves a specific function in guiding, capturing, and surfacing structured thought. The sections are arranged sequentially to mirror how a forecaster—or an LLM agent—would move from question interpretation to probability allocation and self-reflection.

Initial Question Framing

This section ensures that every agent interprets the forecast question in the same way. It requires the forecaster to restate the prompt in plain terms, define key terms or time frames, and flag any ambiguities. This prevents subtle divergences in interpretation from undermining downstream reasoning and enables consistent grounding across agents.

Refinement & Analysis

Here, the forecaster expands on the context, breaking down what is known, what is uncertain, and what variables are likely to influence outcomes. This section supports structured deliberation, incorporates relevant domain knowledge, and separates speculative reasoning from factual input. It acts as the "thinking out loud" stage.

Data Anomaly & Source Integrity Log

Forecasts built on flawed or misunderstood data can yield confident but incorrect results. This section flags any anomalies in source data, unusual statistical signals, or conflicts between primary sources. It also prompts the agent to assess the credibility and bias of the inputs used to shape its forecast.

Probability Allocation

This section forces the forecaster to commit to a clear quantitative view of the likely outcomes. For categorical questions, probabilities must sum to 100% or for binary questions (Yes/No) the output still must be equal to 100%, this stage forces clarity. It also creates a reference point that can later be compared across agents to detect convergence or persistent disagreement.

Final Forecast Summary

The forecaster condenses its reasoning into a single narrative paragraph that explains what it believes will happen and why. This section should avoid jargon and aim for clarity. It is designed for evaluators—both human and AI—to quickly assess the logical coherence of the forecast and its justification.

Why Might You Be Wrong?

This diagnostic section compels the forecaster to step outside its own frame and enumerate failure modes, blind spots, or misinterpreted signals. It forces the model to engage in structured doubt, which is often absent in generative outputs but essential for robust forecasting.

Inside/Outside View Integration

This section draws from dual system forecasting literature and behavioral science. The inside view focuses on the specific case details, while the outside view generalizes across base rates, historical trends, or reference class forecasting. Both are necessary for anchoring judgment and avoiding overfitting to recent information.

Integrating the BIN Model: Diagnosing Reasoning Quality

A distinctive feature of the MSCFT is its embedded alignment with the BIN model—Bias, Information, and Noise—which serves as a diagnostic framework for evaluating reasoning quality in uncertain or data-limited contexts. Originally developed in cognitive science and applied forecasting domains, the BIN model emphasizes three principal error sources: systematic distortion (bias), incomplete or incorrect data (information gaps), and random variation in judgment (noise). Rather than treating these dimensions as abstract theory, the MSCFT operationalizes them through its structural components.

Bias is directly addressed through the “Why Might You Be Wrong?” section, which prompts agents to identify cognitive blind spots, motivated reasoning, or overconfidence. This encourages self-reflection and counters the common tendency of LLMs to reinforce the assumptions embedded in their prompts. Information quality is surfaced in the “Refinement and Analysis” and “Data Anomaly & Source Integrity Log” sections, which force the forecaster to distinguish between high-quality sources and unverifiable claims. These sections also support external validation by creating a transparent citation trail.

Noise—arguably the hardest problem in forecasting—is mitigated not by eliminating subjectivity, but by enforcing repeatable structure across instances. When each agent follows the same decision pathway, variations in output can be diagnosed rather than dismissed. The “Probability Allocation” section further isolates stochastic judgment, enabling ensemble comparison and aggregation without collapsing diverse viewpoints into meaningless averages.

By building the BIN framework into its architecture, MSCFT transforms forecasting from a purely generative task into an epistemic discipline. It guides both human and artificial agents to clarify what they know, how well they know it, and where their confidence may be unjustified. This reinforces the template's role as not just a protocol for alignment, but a framework for metacognitive audit.

GJO Paste Block and Clean Output Mode

To ensure usability and rapid integration into collaborative forecasting environments like Good Judgment Open (GJO), MSCFT includes a final, clean output block. This section strips away internal diagnostics and presents a single, paste-ready paragraph with bucketed probabilities and a concise rationale. It allows rapid posting while preserving the rigor of the full reasoning trail.

V. Use Case: Roche AI Challenge Forecasting

The first comprehensive test of the MSCFT protocol took place during the 2025 Roche AI Challenge, a structured forecasting competition hosted on Good Judgment Open (GJO). The challenge focused on six complex, high-stakes questions involving artificial intelligence in medicine—specifically FDA guidance, clinical trials, diagnostic tools, acquisitions, and LLM benchmarking in healthcare. Good Judgment Open is powered by Cultivate Labs, the software platform responsible for hosting structured forecasting environments and enabling real-time probability tracking and collaboration among forecasters. Each question required precise interpretation, legally and medically grounded reasoning, and defensible probability assignments. These conditions made the challenge an ideal proving ground for the MSCFT framework. Note, Roche AI Challenge is still an open forecasting project and closes Sep 30, 2025, 12:01AM .

Over the course of multiple days, the template was applied to each of the six questions in isolation. Every forecast began with an explicit restatement of the question, followed by structured analysis of the underlying domain. In cases involving drug candidates, clinical stage definitions, or regulatory classification, the MSCFT enforced a disciplined separation between what was publicly verifiable and what was inferential or speculative. In parallel, the "Why Might You Be Wrong?" and "Inside/Outside View" sections forced the LLM forecaster to examine alternative scenarios and reflect on sources of uncertainty—outputs that were often missing from crowd submissions.

The challenge also provided a high-pressure environment where time constraints and scoring incentives risked undermining structure. Despite these constraints, the MSCFT ensured consistency: each forecast included full rationale, numerical bucket allocation, and diagnostic self-assessment. This compliance held even under scoring pressure, helping to preserve auditability and reduce the risk of retroactive bias. Unlike many crowd forecasts, which evolve over time in response to groupthink or feedback loops, MSCFT outputs were fixed at the time of submission and independently traceable.

A final concern addressed by MSCFT was the issue of crowd contamination. Because GJO is a shared platform, forecasters often see each other's estimates and explanations. This can create echo effects and reduce diversity in the prediction set. The MSCFT counters this by enforcing independent rationale before exposure to crowd consensus. In effect, it creates a pre-commitment device: reasoning is generated before social influence can distort it. This preserves the epistemic independence of each forecast and supports more reliable group aggregation.

In sum, the Roche AI Challenge did more than test forecasting skill—it demonstrated that structured alignment tools like MSCFT are essential when LLMs are used in public, medical, or competitive forecasting settings. Without such structure, transparency, traceability, and performance suffer. With it, even complex tasks can be broken down, evaluated, and improved, one section at a time.**

VI. MSCFT as a Protocol for Agent Coordination

One of the most overlooked problems in multi-agent AI systems is the absence of a shared protocol for structured input-output coordination. Without it, even the most capable language models—such as GPT, Claude, and Gemini—cannot reliably collaborate or be meaningfully compared. The Master SWARM Consensus Forecasting Template (MSCFT) was explicitly designed to fill this gap by serving as a coordination protocol: it enforces structural regularity, isolates interpretational ambiguity, and standardizes the way information flows into and out of each agent.

The most fundamental contribution of the MSCFT is its ability to standardize input interpretation and output formatting. In typical deployments, even small variations in prompt phrasing or contextual framing can cause models to diverge. MSCFT neutralizes this risk by beginning each forecast with an "Initial Question Framing" section, which requires agents to restate the problem in their own terms. Downstream sections such as "Probability Allocation," "Inside/Outside View," and "Final Forecast Summary" ensure that outputs are presented in a consistent, readable structure that enables side-by-side comparison between agents—regardless of model architecture or developer.

This structure also enables auditability across different AI systems. When multiple agents respond to the same forecasting task using MSCFT, their reasoning becomes legible, testable, and traceable. Whether a model assigns 70% confidence to an event or 30%, the underlying rationale is surfaced for inspection. In ensemble settings, this makes it possible to isolate why certain agents are converging and why others are diverging, something that would otherwise remain hidden in the black-box generation.

Finally, MSCFT promotes shared epistemic grounding. When agents are required to surface base rates, identify cognitive failure modes, and reflect on potential misinterpretations, their reasoning becomes epistemically transparent. This supports not just better outputs but more defensible and replicable ones. As LLMs become integrated into workflows with medical, legal, or regulatory stakes, the demand for transparent and explainable reasoning will only grow. MSCFT satisfies this requirement by embedding diagnostic prompts and epistemic checkpoints directly into the forecasting process.

In this way, MSCFT is more than a formatting aid. It acts as a protocol for structured, shared cognition across intelligent systems. MSCFT offers a structured way for LLMs to reason in parallel without fragmenting or distorting their decision logic.

VII. Discussion

The deployment of the MSCFT in live multi-agent forecasting environments has revealed both its strengths and limitations. In practical use—such as during the Roche AI Challenge—the template demonstrated the ability to stabilize reasoning, enforce epistemic discipline, and maintain transparent forecast structures across different sessions, domains, and agents. Its modular design allowed forecasters, whether human or LLM-based, to work independently while still producing logically comparable outputs. These characteristics are essential when running structured competitions or high-consequence decision-making simulations involving multiple autonomous systems.

Among its greatest strengths is MSCFT's ability to maintain internal coherence across diverse reasoning agents. In settings where GPT, Claude, and Gemini were used to respond independently to identical questions, the template ensured that the logical trail, probability assignments, and counterfactual awareness could be directly compared. No additional parsing or formatting was required, and the inclusion of explicit diagnostic sections ("Why Might You Be Wrong?" and "Inside/Outside View") made model transparency a built-in feature rather than a retrofit.

However, the current version of the MSCFT requires manual enforcement. This creates friction in deployment and limits scalability. Human users must copy, paste, and adjust the structure by hand. LLMs, even when trained on the format, occasionally omit sections or interpret the structure loosely unless directly prompted. In forecasting environments that demand real-time responsiveness or involve dozens of agents, manual compliance can introduce inconsistencies. These challenges highlight the need for automation—ideally through API-level wrappers, enforced input/output schemas, or integration with orchestration platforms that can guarantee structural fidelity.

Looking forward, the MSCFT has the potential to evolve from a template into forecasting infrastructure. Rather than thinking of it as a document format, it can be understood as a modular scaffold for coordinating distributed reasoning. With proper implementation, the MSCFT could serve as a middleware layer for agent-based forecasting systems, enabling plug-and-play compatibility across AI models, human participants, and institutional governance frameworks. Its logic could also be embedded into user interfaces, prompting users or agents dynamically based on incomplete sections or missing reasoning.

Future applications of the MSCFT protocol are already being considered for alternative forecasting platforms such as Metaculus, which uses a slightly different structure for question formatting and resolution criteria. While Metaculus emphasizes time-series trends and crowd-weighted scoring mechanisms, the core forecasting challenges remain the same: interpretation variance, reasoning opacity, and alignment drift. By adapting the MSCFT template to Metaculus standards, forecasters could retain the same structural discipline while integrating with the site's resolution mechanics, allowing for both comparability and transparency across platforms.

In short, the MSCFT is more than a tool for generating forecasts—it is a blueprint for making structured cognition scalable and comparable. Its design reflects the assumption that the future of reasoning, especially in high-stakes domains, will not be determined by raw model capacity, but by the frameworks that allow those models to think together without falling apart.

This growing interest in structured AI coordination is echoed in institutional efforts such as the RAND Forecasting Initiative, (RFI) seeking scalable methods to improve forecasting accuracy and decision-making frameworks involving both artificial and human agents.

VIII. Conclusion

The Master SWARM Consensus Forecasting Template (MSCFT) was developed to meet a growing need in AI coordination: how to enable multiple agents—both human and artificial—to forecast in parallel without drifting into incoherence, redundancy, or epistemic opacity. Through its deployment in structured environments like the Roche AI Challenge, MSCFT demonstrated its practical value as a scaffolding tool for reliable, auditable, and scalable forecasting under competitive and time-sensitive conditions.

More than a formatting aid, the MSCFT represents an enabling protocol—one that transforms unstructured generative output into coordinated, testable reasoning chains. By requiring consistent inputs, enforcing mirrored output structure, and embedding diagnostic checkpoints throughout the reasoning process, it provides the minimal infrastructure necessary for LLM-based forecasting systems to behave predictably and transparently. It also promotes intellectual humility and self-reflection, qualities essential for safety in high-stakes domains such as medicine, policy, and risk management.

Looking ahead, the MSCFT can serve as a foundational element in broader research into multi-agent AI orchestration. Future work may focus on automating compliance using model-level fine-tuning, plug-in modules, or front-end scaffolds that ensure fidelity to the structure without user micromanagement. Benchmarking projects could also explore comparative performance between structured and unstructured forecasting across domains and agent architectures. Most importantly, the MSCFT opens the possibility of auditable agent reasoning—allowing stakeholders to review not just what a model predicted, but how and why it reached that view.

In this way, the MSCFT is not merely a stopgap—it is a pathway toward safer, more reliable AI reasoning at scale. It provides the protocol-level thinking necessary to align large language model agents without suppressing their diversity or creativity. As the AI field moves toward decentralized, multi-agent architectures, the need for such tools will only grow.

IX. References

Cultivate Labs. (2025). About Us. https://www.cultivatelabs.com/

De Marzo, A., Castellano, G., & García, E. (2025). *Agent societies and epistemic instability in LLM swarms* (arXiv:2409.02822v4). arXiv. https://arxiv.org/abs/2409.02822

Centre for Cognition, Computation and Modelling. (2024). *CCCM Seminar Series on the Cognitive Science of Generative AI* [Zoom meeting materials]. Birkbeck College, University of London.

Good Judgment Open. (2025). *Roche AI Challenge Questions and Forecasts.* https://www.gjopen.com/challenges/roche-ai-challenge

Helip, B. (2025). *Master SWARM Consensus Forecasting Template (MSCFT)* [GitHub repository]. https://github.com/captbullett65/MSCFT

National Cancer Institute. (2025). *Oncology overview.* https://www.cancer.gov/about-cancer/understanding/what-is-cancer

U.S. Food and Drug Administration. (2025). *Federal Register guidance documents.* https://www.federalregister.gov/agencies/food-and-drug-administration

SkyLab, UC Berkeley. (2025). *LM Arena: Text Arena Leaderboard and Blog.* https://lmarena.org

RAND Corporation. (2024). *RAND Forecasting Initiative | RAND - RAND Corporation*. https://www.rand.org/rfi-forecasting-2024

Appendix

MASTER CONSENSUS FORECASTING TEMPLATE (MSCFT v3.1B)

Forecast Title: [Paste or paraphrase the forecast question title]

Initial Question Framing (Copilot/GPT Input):

[Rephrase the question in natural language. Break down key components. Clarify terms. Add definitions if needed.]

Refinement and Analysis: [Summarize key issues. Use outside view and inside view comparison. Highlight precedent data, trends, known drivers, and edge cases.]

Data Anomaly and Source Integrity Log: [Log source credibility, inconsistencies, ambiguity, or unusual patterns. Include time-stamped audit details if needed.]

Probability Allocation:

[List forecast buckets and assign whole number percentages totaling 100. Use precise phrasing.]

Final Forecast Summary (150-character rationale): [Write a concise, accurate rationale no longer than 150 characters. Use sentence case. Do not restate the question.]

Inside View (Case-Specific Factors): [Detail case-specific facts, pipeline history, known behavior, company actions, or context-specific reasoning.]

Outside View (Base Rate / Historical Pattern):

[Provide comparative historical data, base rate trends, or structural reference class evidence.]

Why might you be wrong?

[Write a few plausible reasons this specific forecast might be wrong.]

Why might this forecast fail structurally (Error Mode Notes):

[Document how process breakdowns, ambiguous definitions, or hidden assumptions could cause forecast error.]

Appendix C: Attribution Record – Influence of MSCFT on Raif

In June 2025, Cultivate Labs released an open-source Ruby AI framework called Raif (https://github.com/cultivatelabs/raif), designed to support LLM-powered task routing, tool use, and structured conversation handling in a Rails environment.

This framework shares substantial architectural overlap with the Master SWARM Consensus Forecasting Template (MSCFT), which was independently developed and documented by Brian Helip (Captbullett). The MSCFT was shared directly with Cultivate Labs representative Ben Roesch on May 29, 2025, through email and a GitHub link (https://github.com/captbullett65/MSCFT).

Key MSCFT structures reflected in Raif include:

- Structured ReAct-style reasoning loops (`run! do |conversation history|`)