computers

MDPI

Article

Cognitive Grounding for Perspective Integration in Multi-LLM Systems

Lev Sukherman [1,*], Yetunde Longe-Folajimi [2] and Marina Konkol [3]

[1] Computer Science Department, Worcester Polytechnic Institute, Worcester, MA 01609, USA
[2] School of Computing and Data Science, Wentworth Institute of Technology, Boston, MA 02115, USA; folajimiy@wit.edu
[3] Department of English Language No. 3, Moscow State Institute of International Relations (MGIMO University), Moscow 119454, Russia; m.konkol@my.mgimo.ru
* Correspondence: lsukherman@wpi.edu

Abstract

This paper investigates whether structured collaboration between multiple large language models (LLMs), each assigned a distinct cognitive role grounded in psychological theory, produces benefits beyond simple answer aggregation. We propose the Parallel Synthesis architecture, in which three cognitively specialized roles Analyzer (hierarchical decomposition), Creative (divergent thinking), and Critic (critical evaluation) process each task independently and in parallel, and a Synthesizer integrates their outputs into a final response. To evaluate collaborative reasoning, we introduce the Emergent Reasoning Score (ERS), a composite metric that separates perspective integration (Synthesis Effectiveness) from novel concept generation (Emergent Value). Experiments on Experiments on the AI2 Reasoning Challenge (ARC-Challenge) (1172 questions) and and the Massive Multitask Language Understanding benchmark (MMLU) (1531 questions) show two consistent findings. First, the architecture achieves high Synthesis Effectiveness ($SE = 0.711$–0.744), indicating reliable integration of all three cognitive perspectives. Second, Emergent Value remains low ($EV = 0.096$–0.112), indicating that synthesis primarily recombines existing concepts rather than generating substantial novel content. A Majority Voting baseline achieves comparable or slightly higher answer accuracy than the Synthesizer on both benchmarks, showing that the architecture's main contribution lies not in answer selection but in producing integrated reasoning traces that draw on multiple perspectives. These findings suggest that the practical value of cognitively grounded multi-agent architectures lies in reliable perspective integration, while ERS provides a reusable framework for distinguishing integration from genuinely novel reasoning in multi-agent LLM systems. The empirical results reported here constitute a pilot validation of the proposed framework on closed-form benchmarks, intended to establish a proof of concept and motivate larger-scale evaluation.

Keywords: large language models; natural language processing; multi-agent LLM systems; reasoning evaluation; prompt engineering; cognitive role specialization; question answering

Check for updates

Academic Editor: Paolo Bellavista

Received: 18 March 2026
Revised: 18 April 2026
Accepted: 23 April 2026
Published: 27 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As large language models continue to demonstrate increasingly sophisticated reasoning capabilities, attention has begun to shift from improving individual models to exploring how multiple models can collaborate effectively. A key open question is whether such collaboration, when grounded in principled role differentiation, can produce qualitatively

richer reasoning than any single model achieves alone. This paper approaches this question by assigning each participating model a distinct cognitive role informed by established psychological theory and examining the nature of the resulting collaborative output.

1.1. Research Context and Motivation

Large language models have achieved remarkable progress across a wide range of tasks [1,2], with recent advances focused on enhancing individual model capabilities through retrieval augmentation [3], parameter-efficient fine-tuning [4], and advanced prompting strategies including chain-of-thought reasoning [5,6]. Despite their diversity, these approaches share a common limitation: they operate within the reasoning process of a single model [5]. This raises an important question that remains insufficiently resolved in the current literature: can structured collaboration between multiple LLMs generate emergent reasoning patterns or insights that are absent from the outputs produced by any individual model?

This question draws inspiration from well-established findings in cognitive science and organizational psychology [7,8]. Research on collective intelligence has demonstrated that human teams with appropriate role specialization consistently produce solutions that surpass the capabilities of any individual team member [7,8]. Empirical evidence further supports the existence of a general collective intelligence factor in human groups [9–11]. Crucially, this factor depends not on the highest individual cognitive ability of group members but on the structural properties of collaboration, including cognitive diversity among participants [9]. While other predictors, such as social sensitivity and conversational dynamics, also contribute [10,11], we argue that cognitive diversity is particularly relevant to computational settings where role differentiation can be explicitly designed. These findings suggest that how roles are divided and contributions are integrated plays a more decisive role in group performance than individual cognitive abilities.

A growing body of research on multi-agent LLM systems has begun to explore collaborative approaches to reasoning [12], spanning debate [13], task decomposition [14,15], role-playing [16], and consensus paradigms [17] (reviewed in detail in Section 2). Despite their architectural variety, many of these systems share a common limitation: they treat participating models as general-purpose reasoners and rely on model diversity or iterative negotiation rather than theoretically motivated cognitive specialization to produce collaborative gains.

This limitation reflects three broader gaps in the current literature.

First, existing approaches typically employ homogeneous models or assign superficial persona-based roles (e.g., “doctor”, “engineer”) [18] that mimic social identities rather than cognitive functions. Although established cognitive science frameworks describe the cognitive mechanisms that underlie effective collaboration in human teams (Section 3), these theories have rarely been systematically applied to role design in multi-agent LLM systems, as recent reviews confirm [18].

Second, the parallel versus sequential nature of agent interaction has not been treated as an explicit architectural choice in reasoning-oriented multi-agent systems. While parallel execution is common in task decomposition frameworks for efficiency [14,15], systems designed for collaborative reasoning rely on sequential exchanges where later responses are conditioned on earlier ones [13,17]. This sequential design risks introducing anchoring bias and suppressing the diversity of reasoning paths. It also makes it difficult to cleanly isolate the contribution of each individual agent.

Third, there is currently no standardized metric for distinguishing genuine emergent reasoning—that is, novel insights that arise from the interaction of multiple perspectives—from simple performance aggregation or majority voting. Recent surveys of both prompting

techniques [5] and multi-agent architectures [12] confirm the absence of such evaluation frameworks, highlighting this as an open methodological gap.

1.2. Research Questions and Contributions

This paper addresses these gaps through two research questions:

RQ1: Does structured multi-LLM collaboration in which each agent is assigned a distinct cognitive role grounded in psychological theory (Analyzer: hierarchical decomposition; Creative: divergent thinking; Critic: critical evaluation) produce concepts in the synthesized output that are absent from every individual role output, as measured by the Emergent Value component of ERS, on closed-form reasoning benchmarks?

RQ2: Can a composite metric that separates perspective integration (Synthesis Effectiveness) from novel concept generation (Emergent Value) distinguish architectures that aggregate existing perspectives from those that generate genuinely new reasoning, and do the two components correlate differently with task accuracy across benchmarks?

To answer these questions, we make the following contributions:

1. We propose a Parallel Synthesis architecture that assigns four open-source LLMs distinct cognitive roles, each grounded in established psychological frameworks (Section 3, Table 1).
   The three primary roles operate in parallel without access to each other's outputs, ensuring independence of perspectives and eliminating anchoring bias inherent in sequential interaction designs.
2. We introduce the Emergent Reasoning Score (ERS), a composite metric for multi-agent LLM systems that combines two components with empirically derived weights. The first component, Synthesis Effectiveness (SE), measures the degree to which the Synthesizer integrates concepts from all roles. The second component, Emergent Value (EV), quantifies the proportion of genuinely novel concepts that are absent from all individual role outputs. The weights assigned to SE and EV are determined empirically by their correlation with task accuracy, rather than being set arbitrarily, which makes ERS a data-driven evaluation framework for distinguishing perspective integration from novel concept generation in multi-agent LLM systems.
3. We conduct an empirical evaluation on two complementary benchmarks—ARC-Challenge [19] (1172 questions) and MMLU [20] (1531 questions)—with paired t-tests, Bonferroni correction, and Cohen's d effect size analysis [21] across three independent runs with distinct random seeds. Results show that the Synthesizer performs comparably to the strongest individual roles and significantly outperforms the weakest role on both benchmarks, while not exceeding a simple Majority Voting baseline in answer accuracy. This design is intended as a proof-of-concept validation of the ERS framework and the Parallel Synthesis architecture rather than a definitive empirical assessment.
4. Through ERS analysis, we show that the architecture achieves high perspective integration ($SE = 0.711$–0.744) but limited novel concept generation ($EV = 0.096$–0.112), indicating that approximately 90% of synthesized concepts originate from individual role outputs rather than emerging through synthesis. This reframes the architecture's primary function from creativity amplification to reliable perspective integration.
5. We show that the benefits of multi-LLM collaboration are task-dependent and that characterizing the limits of emergent reasoning is itself informative for system design. This finding helps clarify when cognitively grounded multi-agent architectures are most useful and when simpler aggregation methods may suffice.

Table 1. Mapping of cognitive theories to LLM roles in the Parallel Synthesis architecture.

Role	Theoretical Framework	Cognitive Function	Prompting Strategy
Analyzer	Simon [22]	Systematic decomposition	Break the task into subproblems, identify constraints, and dependencies
Creative	Torrance [23]	Divergent thinking	Generate multiple unconventional perspectives; prioritize fluency and originality
Critic	Kahneman [24]	Critical evaluation (System 2)	Identify potential weaknesses, ambiguities, and hidden complexities in the given task
Synthesizer	Hutchins [8]	Integrative synthesis	Integrate all role outputs; resolve contradictions; construct unified response

The remainder of this paper is organized as follows. Section 2 reviews related work on emergent abilities in large language models, multi-agent LLM systems, and cognitive role specialization in AI. Section 3 presents the theoretical foundations underlying our cognitive role design. Section 4 describes the Parallel Synthesis architecture, cognitive role design, and the Emergent Reasoning Score framework. Section 5 details the dataset selection, experimental setup, and statistical testing protocol. Section 6 presents role-level performance, statistical significance analysis, and ERS analysis. Section 7 discusses interpretation of results, implications for multi-agent system design, limitations, and future work. Section 8 concludes the paper.

2. Related Work

This section reviews three areas of research that inform our approach: the debate on emergent abilities in large language models, the design of multi-agent LLM systems, and the application of cognitive role specialization to AI agents.

2.1. Emergent Abilities in Large Language Models

The concept of emergent abilities in large language models, defined as capabilities that appear only when models exceed a certain scale threshold, was first formalized through the observation that smaller models show near-zero accuracy on certain tasks while larger models solve them with high accuracy, with no gradual transition in between [25]. Among the most prominent examples of such abilities are multi-step arithmetic, logical reasoning, and the capacity to follow chain-of-thought prompting [6], where models produce intermediate reasoning steps before arriving at a final answer. These capabilities were largely absent in earlier, smaller models and manifested without explicit training for the specific tasks, which suggested that scaling alone could unlock qualitatively new reasoning capabilities [26].

However, this interpretation has been challenged by evidence that many apparent emergent abilities can be attributed to the choice of evaluation metrics rather than to fundamental changes in model behavior [27]. Specifically, when nonlinear metrics (e.g., exact-match accuracy, where any partially correct answer scores zero) are replaced with continuous ones (e.g., measuring the proportion of correct tokens in the model's output), the sharp transitions largely disappear, revealing smooth, predictable improvement curves. This finding reframes the question of emergence: the issue is not only whether new capabilities arise but whether our evaluation tools are sensitive enough to capture the true trajectory of model improvement.

The broader landscape of emergence research reflects this tension. The BIG-bench benchmark evaluated language model capabilities across 204 tasks, revealing that performance scaling patterns vary substantially depending on task type and evaluation methodology [28]. Some tasks showed smooth improvement with scale, others exhibited apparent

phase transitions, and a third group showed no meaningful improvement regardless of model size. These results suggest that emergence is not a uniform phenomenon but depends on the interaction between task structure and model capacity.

Early analyses of GPT-4 suggested capabilities in mathematical reasoning, spatial understanding, and theory of mind that appeared qualitatively different from prior models [29]. However, these claims have been widely contested on methodological grounds, and whether such observations reflect genuine cognitive leaps or incremental gains made visible by particular evaluation choices remains an open question.

This debate has direct implications for multi-agent research: if apparent emergence in single models can be partly attributed to evaluation methodology [27], then claims of emergent reasoning in multi-agent systems require equally rigorous metric design. This observation motivates the Emergent Reasoning Score framework developed in Section 4.3, which separates perspective integration from novel concept generation to avoid conflating aggregation effects with genuine emergence.

2.2. Multi-Agent LLM Systems

Research on multi-agent LLM systems has explored several interaction paradigms, each offering a different mechanism for combining the outputs of multiple models. These approaches can be broadly categorized into debate-based, consensus-based, and task-decomposition systems.

Debate-based systems. The multi-agent debate framework [13] demonstrated that multiple instances of the same model, exchanging and critiquing each other's responses over several rounds, can improve mathematical reasoning and factual accuracy by exposing models to counterarguments. This paradigm was further extended by encouraging divergent thinking through structured debate [30], showing that explicitly prompting for disagreement can increase the diversity of generated solutions. However, both approaches rely on homogeneous models and sequential interaction, where each round is conditioned on the previous one.

Consensus-based systems. While debate-based approaches focus on refining individual responses through argumentation, a parallel line of work explores collective agreement among diverse models. ReConcile [17] introduced a round-table discussion protocol among heterogeneous LLMs (ChatGPT 3.5, Bard, Claude), where models iteratively revise their answers based on the group's weighted confidence scores, demonstrating that model diversity can improve reasoning accuracy. However, subsequent work showed that simply scaling the number of agents and applying majority voting can yield similar performance gains [31], raising the question of whether observed improvements reflect genuine collaborative reasoning or statistical aggregation effects—a distinction that motivates the Emergent Value component of our evaluation framework (Section 4.3.3). Mixture-of-Agents [32] addressed this by constructing a layered architecture in which each agent refines its response using outputs from the previous layer, reportedly surpassing GPT-4 on selected open-ended generation benchmarks using only open-source models [32]. While this confirms the value of multi-model integration, the approach treats all agents as general-purpose reasoners without cognitively differentiated roles.

Task-decomposition systems. A third line of work shifts from improving reasoning quality to managing task complexity by dividing complex tasks among specialized agents. AutoGen [14] provides a general-purpose framework for multi-agent conversations with customizable interaction patterns, while MetaGPT [15] structures software development workflows by assigning agents roles such as product manager, architect, and engineer. CAMEL [16] explores autonomous cooperation between agents through role-playing with inception prompting. In the medical domain, specialist personas have been used to improve

zero-shot clinical reasoning [33], and adaptive agent recruitment based on task complexity has shown further gains [34].

Several comprehensive surveys reflect the rapid growth of this field, systematically categorizing multi-agent architectures by their communication topology, role assignment strategy, and evaluation methodology [12,35,36]. Across all three paradigms, two patterns reinforce the gaps identified in Section 1.1. First, role assignments remain either absent, based on professional personas, or defined by software engineering functions; none are grounded in cognitive science theories of human reasoning. Second, evaluation relies exclusively on downstream task accuracy and offers no way to assess whether collaboration produces qualitatively new reasoning. Notably, these limitations persist regardless of interaction topology (sequential or parallel), the number of participating agents, or the degree of model heterogeneity, suggesting that the gaps are structural rather than incidental to any particular framework. We note that the present study does not include direct empirical comparisons against these architectures. This is consistent with the pilot scope of the evaluation: a controlled comparison with debate-based systems [13], consensus protocols [17], or layered mixture architectures [32] would require identical datasets, matched model scales, and equivalent inference budgets across all systems. These requirements exceed the scope of the current study. We treat such comparisons as a primary direction for future work once the Parallel Synthesis architecture and ERS framework have been validated at a larger scale.

Recent work has begun to address evaluation gaps in multi-agent LLM research through standardized benchmarking that measures not only task completion but also the quality of collaboration and competition [37], as well as information-theoretic frameworks for measuring emergent coordination [38] and comprehensive reviews of framework design and open challenges [39]. These contributions reflect a shared recognition that task accuracy alone is insufficient for evaluating multi-agent collaboration and directly motivate the development of reusable metrics such as ERS that separate integration quality from novelty generation.

2.3. Cognitive Role Specialization in AI

The idea of assigning distinct cognitive or social roles to AI agents has gained attention through several lines of work. Generative agents that simulate believable human behavior in an interactive virtual environment demonstrated that LLMs can maintain consistent persona-based behavior, including planning, reflection, and social interaction [40]. While this work established that LLMs can sustain differentiated roles over extended interactions, the roles modeled social identities rather than cognitive functions.

A step closer to cognitively motivated design is the Mindstorms framework [41], which draws on Minsky's Society of Mind theory to organize LLM agents into specialized modules that collaborate on complex tasks. However, the mapping between cognitive theory and agent specialization in this framework remains loosely defined: while it invokes Minsky's Society of Mind as a conceptual metaphor, individual agent roles are not derived from specific, testable cognitive theories, and their correspondence to distinct reasoning processes is not empirically validated. More broadly, a systematic study of persona assignment in LLMs found that superficial role labels often fail to change model reasoning behavior in the expected way, and that their effectiveness varies widely across tasks [18]. This suggests that meaningful role differentiation requires grounding in cognitive theory rather than superficial labels.

Our work aims to bridge multi-agent LLM systems and the cognitive science of collaboration by operationalizing established psychological frameworks as LLM role prompts. Unlike prior approaches that assign professional or social personas, we design each role

to target a specific cognitive function grounded in theories of problem solving [22], creative thinking [23], critical evaluation [24], and distributed cognition [8]. This grounding enables a controlled investigation of whether cognitively differentiated collaboration in LLM systems produces reasoning properties that parallel the collective intelligence patterns observed in human teams [9]. The following section describes each of these theoretical frameworks in detail and explains how they inform the design of our architecture.

Relative to prior multi-agent LLM systems, the present study contributes two elements in combination: cognitively grounded role design derived from established psychological theory and an explicit evaluation framework that distinguishes perspective integration from genuinely novel conceptual contribution.

3. Theoretical Foundation

The Parallel Synthesis architecture draws on well-established cognitive science theories that explain how complementary reasoning processes support effective problem solving in human teams. This section outlines the key theoretical models that shape the design of each cognitive role and the overall architectural choices.

3.1. Cognitive Problem-Solving Models

Human problem solving typically draws on several complementary cognitive processes [22,24]. Three of these frameworks are especially relevant to our design of cognitively specialized roles in LLMs.

The first is the theory of hierarchical problem decomposition, which models complex problem solving as a process of recursively breaking a problem into smaller, more manageable subproblems [22]. On this account, effective reasoning begins not with generating solutions but with structuring the problem space, identifying constraints, and establishing relationships between components. In practice, this means that a solver who first maps the structure of a problem tends to avoid errors that come from jumping directly to an answer. This framework motivates the *Analyzer* role in our architecture, which is prompted to perform systematic decomposition before attempting to solve a given task.

The second framework concerns the nature of creative thought. The distinction between convergent thinking, which narrows toward a single correct answer, and divergent thinking, which generates multiple alternative solutions, was originally proposed as a fundamental dimension of human intellect [42]. Torrance [23] further developed this distinction into a multidimensional model of creativity characterized by fluency (the number of ideas produced), flexibility (the variety of conceptual categories explored), and originality (the statistical rarity of responses). These dimensions motivate the *Creative* role, which is operationalized along all three dimensions: fluency is instantiated by requiring at least two to three distinct interpretive angle s per task; flexibility by explicitly encouraging analogies and framings drawn from domains outside the immediate subject matter; and originality by instructing the model to deviate from the most salient or conventional reading of the question before converging on an answer.

The third framework is dual-process theory, which distinguishes between two modes of cognitive processing [24]. System 1 operates automatically and quickly, relying on heuristics and pattern recognition. System 2 engages in deliberate analytical processing, including complex reasoning, logical evaluation, and the monitoring of System 1 outputs for potential errors and unwarranted assumptions. Our *Critic* role draws specifically on this evaluative dimension of System 2: it is prompted to identify potential weaknesses, ambiguities, and hidden complexities in the given task before a solution is proposed.

Taken together, these three frameworks provide complementary reasoning functions: problem structuring (Analyzer), perspective exploration (Ceative), and critical evaluation

(Critic). To realize the full potential of these diverse perspectives, they must be integrated into a coherent response through a principled synthesis process [43]. Distributed cognition theory provides the theoretical basis for this integration, as described in the following section [8].

3.2. Distributed Cognition Theory

Distributed cognition theory proposes that cognitive processes are not confined to individual minds but are distributed across individuals, artifacts, and the environment in which they operate [8]. In this framework, the unit of analysis is not the individual thinker but the entire cognitive system, including the interactions between its components. A central insight of this theory is that the coordination of diverse specialized contributions can produce system-level cognitive properties that no single component possesses on its own [8].

This perspective motivates two key design decisions in our architecture. First, it motivates the *Synthesizer* role, which receives the independent outputs of the *Analyzer*, *Creative*, and *Critic* and is prompted to integrate them into a unified response. The Synthesizer does not simply select the best individual answer but follows a three-step integration process. First, it identifies points of convergence across role outputs: when two or more roles reach the same conclusion through different reasoning paths, that agreement is treated as a reliability signal and weighted accordingly. Second, it evaluates divergences by examining which perspective is better supported by the task's structural constraints (factual, logical, or domain-specific) rather than defaulting to a majority. Third, it constructs a unified response that explicitly draws on complementary contributions from each role, preserving the reasoning chain that leads to the final answer rather than collapsing it into a bare selection.

This design directly operationalizes the distributed cognition principle that integration across specialized components can produce capabilities beyond those of any single component.

Second, the theory supports parallel execution of the three primary roles. In human teams, cognitive diversity contributes most effectively when perspectives are formed independently, before any group interaction takes place [9,44]. Sequential interaction risks anchoring effects that diminish effective diversity [45]. By running the Analyzer, Creative, and Critic in parallel, without access to each other's outputs, the architecture preserves this independence, which distributed cognition identifies as essential for collective performance [8].

3.3. Key Definitions

Two terms used throughout this paper require clarification before the architecture is described in detail. First, we use *cognitive role* to refer to a prompt designed to elicit a specific cognitive function (e.g., decomposition, divergent thinking), as distinct from a *persona*, which assigns a social or professional identity (e.g., 'doctor', 'engineer') without targeting a particular reasoning process. Second, we define *emergent reasoning* operationally as the appearance of concepts in the synthesized output that are absent from every individual role output, measured through the Emergent Value component of the ERS framework (Section 4.3). This definition intentionally sets a high bar: mere performance improvement over individual roles does not constitute emergence unless it is accompanied by genuinely novel conceptual content. These definitions guide both the architectural decisions described in Section 4 and the evaluation criteria developed in Section 4.3.

4. Methodology

This section describes the Parallel Synthesis architecture, the design of cognitively specialized roles, and the Emergent Reasoning Score framework used to evaluate multi-agent collaboration.

4.1. Parallel Synthesis Architecture

The Parallel Synthesis architecture processes each task through a two-stage pipeline. In the first stage, three cognitively specialized roles, the Analyzer, Creative, and Critic, receive the same task input and generate their responses independently and in complete isolation from one another. In the second stage, the Synthesizer receives all three role outputs along with the original task and produces a final integrated response. In effect, the first stage preserves the independence of reasoning paths by preventing any mutual influence between roles, while the second stage selects and combines the strongest elements from each.

This two-stage design reflects the theoretical principles established in Section 3. Parallel execution preserves the independence of perspectives essential for collective intelligence (Section 3.2), an independence that sequential designs could compromise through anchoring effects, thereby reducing the effective cognitive diversity of the system. The separation between generation (Stage 1) and integration (Stage 2) mirrors the distinction between divergent and convergent cognitive processes [42], with the three primary roles exploring the problem from different angles, while the Synthesizer converges toward a unified answer.

Each role is implemented as a separate LLM call with a role-specific system prompt. All roles use identical generation parameters ($temperature = 0.7$, $top_p = 0.9$) across all experimental conditions; the rationale for these settings is described in Section 5.3. Operationally, the pipeline proceeds as follows. For each task, the question and answer options are formatted into a standardized prompt and passed simultaneously to the Analyzer, Creative, and Critic via three independent API calls. No output from one role is visible to any other during this stage. Once all three responses are received, a fourth API call passes the original task together with all three role outputs to the Synthesizer. The Synthesizer produces a single integrated response that includes both a reasoning trace and a final answer extracted from a structured output field. A fifth API call then passes each of the four outputs independently to the concept extraction model. The total inference cost per task is therefore up to eight LLM calls: three concurrent role calls, one synthesis call, and four extraction calls (one per output). The role calls and extraction calls can each be parallelized, reducing the critical path to two sequential stages: parallel role execution followed by synthesis. The architecture is model-agnostic and can be applied to any combination of instruction-following LLMs. Model selection criteria, including the rationale for controlling model scale, are described in Section 5.1. The overall pipeline is illustrated in Figure 1.

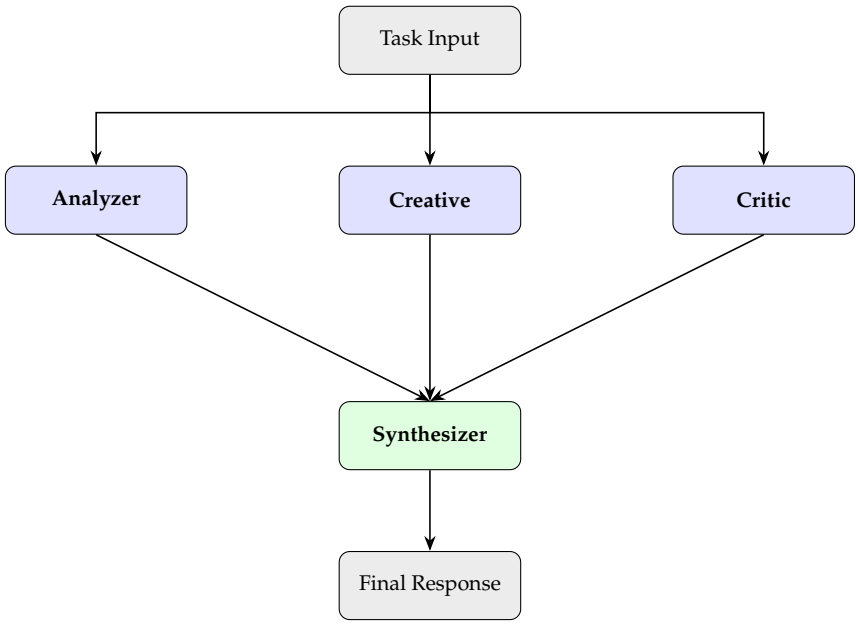


Figure 1. The Parallel Synthesis architecture. Arrows indicate the direction of information flow. Three cognitively specialized roles receive the task input and process it independently and in parallel. Their outputs are then passed to the Synthesizer, which has access to all three perspectives but the primary roles have no access to each other's outputs.

4.2. Cognitive Role Design

Each cognitive role is operationalized in two steps: first, a model is assigned a specific cognitive function based on the theoretical frameworks described in Section 3.1; second, a dedicated system prompt translates that function into concrete behavioral instructions.

The *Analyzer* prompt directs the model to decompose the task into subproblems, identify constraints, and reason through each component systematically before arriving at an answer. The *Creative* prompt instructs the model to generate multiple alternative perspectives, emphasizing fluency, flexibility, and originality rather than converging on a single interpretation. The *Critic* prompt directs the model to examine the task for potential ambiguities, hidden assumptions, and complexities before proposing a solution. The *Synthesizer* prompt instructs the model to identify points of agreement across the three role outputs, resolve contradictions, and construct a unified response that integrates complementary insights rather than simply selecting one role's answer.

Crucially, each prompt targets a cognitive process rather than a professional identity (Section 3.3): the Analyzer is not prompted to act as a scientist or engineer but to apply hierarchical decomposition regardless of domain. This ensures that role differentiation stems from distinct reasoning strategies rather than from domain-specific knowledge associations.

The choice of three primary roles reflects a principled balance between cognitive diversity and synthesis complexity. Fewer roles would limit the range of complementary reasoning strategies available to the Synthesizer, while additional roles would increase the risk of conceptual redundancy and place a greater integration burden on the synthesis stage. The three selected roles correspond to three well-established and non-overlapping cognitive functions, namely problem structuring, perspective exploration, and critical evaluation (Section 3), providing broad coverage of complementary reasoning processes.

All prompts were iteratively refined through manual inspection of role outputs on a held-out set of approximately 20 tasks per benchmark, sampled to cover a range of difficulty levels and subject domains. Over three rounds of revision, we verified that each role produced outputs consistent with its intended cognitive function: specifically, that the Analyzer decomposed problems into subcomponents, the Creative generated multiple

alternative perspectives, and the Critic identified potential weaknesses before proposing an answer. Table 1 summarizes the mapping from theory to role design. The final versions of all cognitive role assignments are provided in Appendix A, and the corresponding system prompts in Appendix B.

4.3. Emergent Reasoning Score

The Emergent Reasoning Score (ERS) is a composite metric designed to quantify emergent properties in multi-agent LLM systems. ERS captures two distinct dimensions of multi-agent collaboration: how effectively the Synthesizer integrates concepts from all roles and whether the synthesis process generates genuinely novel concepts absent from any individual role output. For a dataset of N tasks, the final score is computed as:

$$\text{ERS} = \overline{SE} \cdot w_{SE} + \overline{EV} \cdot w_{EV} \tag{1}$$

where SE and EV denote dataset-level Synthesis Effectiveness and Emergent Value, respectively, and w_{SE}, w_{EV} are empirically derived weights. The computation of ERS proceeds through four steps: concept extraction (Section 4.3.1), calculation of Synthesis Effectiveness (Section 4.3.2), calculation of Emergent Value (Section 4.3.3), and empirical weight derivation (Section 4.3.4).

Illustrative example.

Consider the following concrete instantiation of the ERS pipeline on a single task. The Analyzer yields the concepts {*constraint checking*, *elimination strategy*, *unit conversion*, *answer consistency*, *dependency mapping*, *logical decomposition*, *step sequencing*, *constraint satisfaction*, *option elimination*, *structural analysis*}; the Creative yields {*analogy*, *alternative framing*, *elimination strategy*, *estimation*, *lateral thinking*, *domain transfer*, *reframing*, *unconventional approach*, *cross-domain reasoning*, *creative synthesis*}; and the Critic yields {*answer consistency*, *ambiguity check*, *distractor analysis*, *hidden assumption*, *edge case*, *logical fallacy*, *bias detection*, *assumption challenge*, *error identification*, *robustness check*}.

Concept extraction. The extraction model processes each output independently and returns exactly ten concepts per role, normalized to lowercase and limited to three words each.

SE calculation. The Synthesizer output contains {*constraint checking*, *elimination strategy*, *ambiguity check*, *answer consistency*, *estimation*, *unit conversion*, *hidden assumption*, *alternative framing*, *distractor analysis*, *causal inference*}. Computing overlap per role under the relaxed matching criterion: $\text{overlap}_A = 4/10 = 0.40$ (*constraint checking*, *elimination strategy*, *unit conversion*, and *answer consistency* matched); $\text{overlap}_C = 3/10 = 0.30$ (*alternative framing*, *elimination strategy*, and *estimation* matched); $\text{overlap}_E = 4/10 = 0.40$ (*answer consistency*, *ambiguity check*, *hidden assumption*, and *distractor analysis* matched). Thus $SE = (0.40 + 0.30 + 0.40)/3 = 0.367$.

EV calculation. The union C_{all} contains all unique concepts across the three roles. The only synthesized concept absent from C_{all} is *causal inference*, a cross-role inference that plausibly emerges from combining the Analyzer's structural decomposition with the Critic's assumption evaluation, yet was not explicitly present in either output. Thus $C_{\text{novel}} = \{\textit{causal inference}\}$ and $EV = 1/10 = 0.10$.

Interpretation. Moderate SE (0.367) indicates partial integration across perspectives. Low but nonzero EV (0.10) indicates that synthesis introduced one genuinely novel concept not present in any individual role output. This constructed example illustrates the distinction between perspective integration (captured by SE) and novel concept generation (captured by EV); empirical results across both benchmarks are reported in Section 6.

4.3.1. Concept Extraction

To enable quantitative comparison between role outputs and the synthesized response, we employ an automated concept extraction pipeline. A separate LLM processes each of the four outputs (Analyzer, Creative, Critic, and Synthesizer) independently using a standardized prompt that instructs the model to extract exactly ten key concepts focusing on solution approaches, key principles, and specific reasoning techniques. Using the same extraction model and prompt across all outputs ensures consistency in concept identification methodology.

To support reliable set-theoretic comparison, all extracted concepts are normalized to lowercase. Since concept boundaries in free-text extraction are inherently imprecise, we apply two constraints to ensure reliable comparison. First, the extraction prompt limits each concept to a maximum of three words, reducing variability in how concepts are expressed. Second, we apply a relaxed matching criterion: two concepts are considered equivalent if one is a substring of the other or if they differ by at most one word. Operationally, "differ by at most one word" is computed as a word-level Levenshtein distance of at most one edit—a single-token insertion, deletion, or substitution. Under this rule, "constraint satisfaction" and "constraint checking" are treated as equivalent (one substitution), whereas "heat transfer" and "thermal conduction" are treated as distinct (two edits). Together, these constraints account for minor wording variation (e.g., 'problem decomposition' matching 'systematic problem decomposition') while preserving most meaningful semantic distinctions. This criterion may occasionally produce false positives by matching concepts that share surface-level similarity but differ semantically (e.g., 'energy transfer' and 'heat transfer'); however, we prioritize recall over precision in matching to avoid artificially inflating Emergent Value through superficial wording differences across independently generated outputs. The impact of this design choice on ERS reliability is discussed in Section 7.

It is important to acknowledge that this matching criterion is lexical rather than semantic: two concepts that are expressed in different words but denote the same reasoning operation may be counted as distinct, which would inflate Emergent Value. Lexical matching was chosen for this pilot for three practical reasons: it is fully deterministic, it requires no additional model calls or embedding infrastructure, and it yields results that can be reproduced across experimental conditions without hyperparameter tuning. The principal alternative, embedding-based semantic similarity with a cosine threshold, introduces its own design choices—model selection, threshold calibration, and the need for a separate validation study of those choices.

To bound the influence of this choice on our conclusions, we conducted two robustness analyses on the existing runs. The first varies the matching rule across the full spectrum from strict exact match to the paper's relaxed rule; the second replaces lexical matching with semantic similarity. Both are reported in Section 6.4, alongside a construct-validity analysis that tests whether EV responds to the structural property it is designed to capture.

Formally, for each task t in a dataset D of N tasks, the extraction pipeline produces four concept sets:

$$C_A(t),\ C_C(t),\ C_E(t),\ C_S(t) \tag{2}$$

where C_A, C_C, C_E, and C_S denote the concept sets extracted from the Analyzer, Creative (C for creative reasoning), Critic (E for evaluative reasoning), and Synthesizer outputs, respectively, with $|C_i(t)| = 10$ for each role i. This number was determined through preliminary testing on a subset of 30 tasks, comparing extraction counts of 5, 8, 10, 12, and 15. Below ten, extracted concept sets frequently omitted reasoning steps that were identifiable through manual review of the original outputs. Above twelve, concepts became increasingly generic (e.g., 'logical reasoning', 'systematic analysis') or overlapped with other extracted concepts

from the same output. Ten provided the best balance between coverage of key reasoning elements and specificity of individual concepts. This intermediate representation transforms free-text responses into structured concept sets that support the set-theoretic operations underlying the *SE* and *EV* calculations described below.

4.3.2. Synthesis Effectiveness

Synthesis Effectiveness (*SE*) measures the degree to which the Synthesizer integrates concepts from all three primary roles. For each role $R \in \{A, C, E\}$, we compute the concept overlap between the Synthesizer output and the role output as the proportion of role concepts that appear in the synthesis:

$$\text{overlap}_R(t) = \frac{|C_S(t) \cap C_R(t)|}{|C_R(t)|} \tag{3}$$

SE for a single task is then defined as the average overlap across all three roles:

$$SE(t) = \frac{1}{3} \sum_{R \in \{A,C,E\}} \text{overlap}_R(t) \tag{4}$$

$SE(t)$ ranges from 0 to 1, where $SE = 1$ indicates that every concept from every role is represented in the synthesized output, and $SE = 0$ indicates no concept overlap. The dataset-level Synthesis Effectiveness is computed as the sum across all tasks:

$$\overline{SE} = \frac{1}{N} \sum_{t=1}^{N} SE(t) \tag{5}$$

SE measures concept presence rather than concept importance: a high *SE* score indicates that the Synthesizer draws broadly on all role perspectives but does not guarantee that the most task-relevant concepts are prioritized. This limitation is addressed indirectly through the empirical weight derivation (Section 4.3.4), which ties each ERS component to task accuracy, thereby ensuring that the overall score reflects integration quality that is predictive of correct answers rather than integration breadth alone.

4.3.3. Emergent Value

Emergent Value (*EV*) quantifies the proportion of concepts in the Synthesizer output that are genuinely novel, meaning they do not appear in any individual role output. We first compute the union of all individual role concepts:

$$C_{\text{all}}(t) = C_A(t) \cup C_C(t) \cup C_E(t) \tag{6}$$

The set of novel concepts is defined as the set difference between the Synthesizer concepts and the union of all individual role concepts:

$$C_{\text{novel}}(t) = C_S(t) \setminus C_{\text{all}}(t) \tag{7}$$

EV for a single task is then the proportion of novel concepts in the synthesized output:

$$EV(t) = \frac{|C_{\text{novel}}(t)|}{|C_S(t)|} \tag{8}$$

$EV(t)$ ranges from 0 to 1, where $EV = 0$ indicates that every synthesized concept was already present in at least one role output, and $EV = 1$ indicates that the synthesis produced entirely new concepts. The dataset-level Emergent Value is

$$\overline{EV} = \frac{1}{N} \sum_{t=1}^{N} EV(t) \tag{9}$$

The relationship between SE and EV is illustrated in Figure 2.

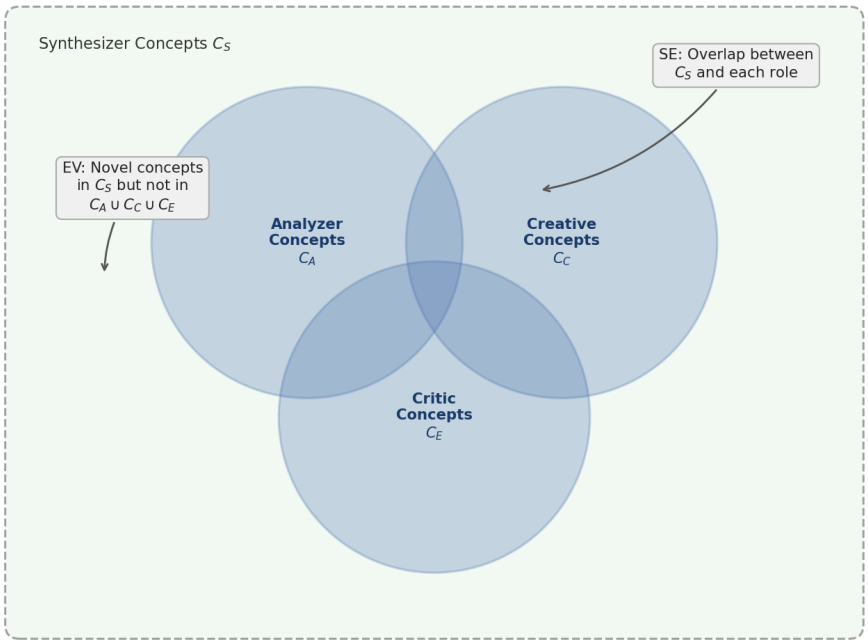


Figure 2. Conceptual illustration of the Emergent Reasoning Score framework. The three lower circles represent the concept sets extracted from each role output (Analyzer C_A, Creative C_C, Critic C_E). The upper region represents the Synthesizer's concept set C_S. Overlapping areas between C_S and the role circles represent Synthesis Effectiveness (SE): the proportion of role concepts integrated into the synthesis. The non-overlapping portion of C_S represents Emergent Value (EV): novel concepts present in the Synthesizer output but absent from all individual role outputs ($C_A \cup C_C \cup C_E$).

4.3.4. Weight Derivation

Rather than assigning fixed weights to SE and EV, we derive them empirically based on each component's predictive relationship with task accuracy. We compute Pearson correlation coefficients between each component and task-specific accuracy across all N tasks in the dataset:

$$\rho_{SE} = \text{corr}(SE, \text{Accuracy}), \quad \rho_{EV} = \text{corr}(EV, \text{Accuracy}) \tag{10}$$

Since task-specific accuracy is binary (1 for correct, 0 for incorrect), these correlations are equivalent to point-biserial correlations, which are a special case of Pearson's r when one variable is dichotomous [21]. This equivalence ensures that the weight derivation remains statistically valid despite the binary nature of the outcome variable.

Weights are then determined proportionally to the absolute correlation strength:

$$w_{SE} = \frac{|\rho_{SE}|}{|\rho_{SE}| + |\rho_{EV}|}, \quad w_{EV} = \frac{|\rho_{EV}|}{|\rho_{SE}| + |\rho_{EV}|} \tag{11}$$

This normalization ensures that $w_{SE} + w_{EV} = 1$. The dataset-level Emergent Reasoning Score is then computed as:

$$\text{ERS} = \overline{SE} \cdot w_{SE} + \overline{EV} \cdot w_{EV} \tag{12}$$

This data-driven approach estimates the relative contribution of integration quality and novelty generation using their observed association with task accuracy, rather than assigning fixed theoretical weights in advance. A potential concern with this approach is that weights derived from the same dataset used for evaluation may overfit to the specific benchmark distribution, inflating the apparent predictive validity of ERS. We note three mitigating factors. First, the weights are derived from task-level correlations across thousands of instances (1172 for ARC-Challenge, 1531 for MMLU), which limits the instability that arises from small-sample weight estimation. Second, the weights are used only for reporting the composite ERS score; the *SE* and *EV* components are reported independently (Section 6.3) and constitute and constitute the primary interpretive basis for our conclusions. Third, the empirical weight derivation is presented as a proof-of-concept mechanism rather than a validated calibration procedure. Cross-dataset weight stability, in which weights derived on one benchmark are applied to another, is an important robustness check that we identify as a concrete direction for future work.

The resulting weights reflect which dimension of emergent reasoning, whether perspective integration or novel concept generation, more strongly contributes to solution quality on a given benchmark. The following section describes the experimental setup in which ERS, together with task-specific accuracy, is applied to evaluate the Parallel Synthesis architecture across two benchmarks.

5. Experiments

This section describes the experimental conditions under which the Parallel Synthesis architecture and the ERS framework are evaluated, including model selection, dataset selection, experimental setup, and the statistical testing protocol.

5.1. Model Selection

All models used in this study are open-source and publicly available on Hugging Face, ensuring complete experimental reproducibility. Model selection was guided by three criteria: architectural diversity, open licensing, and strong instruction-following capabilities. To control for the effect of model scale, all dense models used in this study operate at approximately 70B parameters. The one exception is DeepSeek-V3, a Mixture-of-Experts model with 671B total parameters but only 37B parameters activated per token [46]. Since MoE models activate only a fraction of their parameters during inference, the effective computational scale of DeepSeek-V3 is comparable to that of the dense 70B models, while its distinct architecture provides additional diversity in reasoning profiles.

The three primary cognitive roles are assigned as follows. The *Analyzer* role is assigned to Llama 3.3 70B Instruct [47], selected for its strong performance on structured reasoning tasks. The *Creative* role is assigned to DeepSeek-V3 [46], a Mixture-of-Experts architecture that provides a distinct reasoning profile from standard dense transformers. The *Critic* role is assigned to Llama 3.1 70B Instruct [47]. Although both the Analyzer and Critic use models from the same organization (Meta AI), they belong to different model generations (3.3 vs. 3.1), and their distinct cognitive prompts are designed to elicit fundamentally different reasoning behaviors: systematic decomposition for the Analyzer versus identification of weaknesses and hidden complexities for the Critic.

The *Synthesizer* role is assigned to Qwen 2.5 72B Instruct [48], which receives the outputs of all three primary roles and produces the final integrated response. For concept

extraction (Section 4.3.1), we use Mistral Small 24B Instruct to process all four outputs (Analyzer, Creative, Critic, and Synthesizer). A smaller model produces more deterministic extractions at the same temperature setting, which improves consistency in concept identification across experimental conditions. Using a separate model family for extraction further reduces the risk of systematic bias that could result from the same model both generating and evaluating its own outputs.

This combination spans four distinct organizations (Meta AI, DeepSeek, Alibaba Cloud, Mistral AI) and two architectural paradigms (dense transformers and Mixture-of-Experts), providing sufficient diversity to ensure that observed results are not artifacts of a single model family or training methodology.

5.2. Dataset Selection

We evaluate the Parallel Synthesis architecture on two complementary benchmarks that test different reasoning capabilities.

ARC-Challenge [19] consists of 1172 multiple-choice science questions drawn from standardized exams. These questions require knowledge integration across scientific domains and multi-step inference, making them suitable for evaluating whether cognitively specialized roles contribute complementary perspectives to scientific reasoning tasks.

MMLU [20] consists of 1531 multiple-choice questions from the validation set, spanning 57 subjects across humanities, social sciences, and STEM. The breadth of this benchmark allows us to assess whether the benefits of multi-agent collaboration generalize across diverse knowledge domains or are confined to specific task types.

Since the two benchmarks use different answer formats, ARC-Challenge provides labeled options (A, B, C, D) while MMLU provides only answer texts, we standardize all questions to a uniform multiple-choice format with explicitly labeled options (A, B, C, D) during preprocessing. This ensures consistent prompting across both benchmarks and enables fair comparison of model performance.

Together, these benchmarks provide 2703 evaluation instances covering both specialized scientific reasoning (ARC-Challenge) and broad general knowledge (MMLU), enabling analysis of whether emergent reasoning patterns vary across task characteristics.

5.3. Experimental Setup

For each benchmark, we execute the complete Parallel Synthesis pipeline three times with $temperature = 0.7$ and $top_p = 0.9$. The stochasticity inherent in LLM sampling at these settings introduces run-to-run variability, providing the variance necessary for statistical comparison across conditions. Each run uses a distinct, arbitrarily chosen random seed (42, 123, 456) to ensure that this variability is reproducible.

This three-run design constitutes a pilot experimental protocol. The number of runs was constrained by the cumulative API inference costs across four models and 2703 tasks per run, which made a larger number of replications infeasible within the scope of this study. Accordingly, the statistical results reported in Section 6 should be interpreted as preliminary evidence supporting a proof of concept rather than as definitive quantitative conclusions. The primary goal of this evaluation is to demonstrate the operational feasibility of the Parallel Synthesis architecture and the ERS framework and to identify patterns that motivate larger-scale empirical follow-up.

All models are accessed through the OpenRouter API, which provides inference access to the selected open-source models while maintaining the availability of original weights on Hugging Face for local deployment and verification. To ensure independent model outputs across runs, even for identical input prompts, we appended a unique cache-busting identifier to each request, encoding the random seed, question index, and role assignment.

This prevented API-level caching from returning identical responses across experimental conditions. Generation parameters are held constant across all roles and runs to ensure that differences in output are attributable to the cognitive role prompts rather than to variation in sampling configuration.

All four experimental conditions are derived from a single execution pipeline. For each task and run, we execute the complete Parallel Synthesis process: three parallel role executions followed by synthesis integration. Individual role baselines (Analyzer-only, Creative-only, Critic-only) use the respective role outputs generated during this process, while the Synthesizer condition uses the final integrated output. This design guarantees that baseline comparisons evaluate identical role outputs, isolating the contribution of synthesis integration from any variance in individual model responses.

To contextualize the Synthesizer's performance relative to simpler aggregation methods, we also compute a Majority Voting baseline [31] from the same three role outputs. For each task, Majority Voting selects the answer endorsed by at least two of the three primary roles. When all three roles select different answers (no majority exists), the Analyzer's answer is used as the tiebreaker, reflecting its consistently strong individual performance (Section 6.1). This baseline requires no additional model inference and represents the lowest-cost aggregation strategy available from the same set of role outputs, providing a reference point for assessing whether the Synthesizer's integration process adds value beyond simple answer aggregation.

Task-specific accuracy is determined by exact match against the reference answer for all multiple-choice questions. Each role output and the synthesized response include a standardized "FINAL ANSWER" field from which the selected answer is extracted for evaluation.

5.4. Statistical Testing Protocol

To establish statistical rigor, we perform significance testing across the three independent runs for each experimental condition. For all metrics, we report mean $\pm$ standard deviation across runs and calculate 95% confidence intervals to quantify uncertainty in our estimates.

We employ paired t-tests to compare the Synthesizer performance against each of the three role baselines (Analyzer-only, Creative-only, Critic-only) across the three runs. Given three pairwise comparisons per benchmark, we apply Bonferroni correction to control the family-wise error rate, setting the significance threshold at $\alpha = 0.05/3 \approx 0.017$. The Majority Voting comparison is treated as a separate exploratory analysis rather than included in the Bonferroni family, since it represents an aggregation baseline rather than an individual role condition. This comparison is evaluated at the unadjusted threshold $\alpha = 0.05$. We note that including the Majority Voting comparison in the Bonferroni family (four comparisons, $\alpha = 0.0125$) would not change the pattern of results: the Synthesizer versus Critic difference ($p = 0.013$) would fall marginally above this stricter threshold ($p > 0.0125$), consistent with the limited statistical power of three-run designs discussed in Section 7.3. Beyond p-values, we calculate Cohen's d effect sizes to quantify the practical magnitude of performance differences, following standard interpretation thresholds: small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) [21].

This combination of Bonferroni-corrected significance testing and effect size analysis ensures that reported differences are both statistically reliable and practically meaningful, addressing the risk of overstating results from a limited number of runs.

To supplement the parametric analysis, we compute bootstrap confidence intervals for SE and EV using the three observed run values as the sampling distribution. Bootstrap resampling (10,000 iterations with replacement) yields 95% bootstrap CIs of $[0.738, 0.750]$

for ARC-Challenge *SE*, $[0.088, 0.108]$ for ARC-Challenge *EV*, $[0.709, 0.715]$ for MMLU *SE*, and $[0.111, 0.114]$ for MMLU *EV*. These intervals are consistent with the parametric estimates reported in Section 6.3 and confirm that the SE–EV contrast is stable across resampling. We acknowledge that bootstrap CIs derived from three observations are themselves imprecise; they are reported as a supplementary robustness check rather than a replacement for larger-scale replication.

6. Results

This section presents the experimental results across three dimensions: role-level accuracy, statistical significance of performance differences, and Emergent Reasoning Score analysis.

6.1. Role-Level Performance

Table 2 summarizes the mean accuracy and standard deviation across three independent runs for each role and the Synthesizer on both benchmarks.

Table 2. Mean accuracy (%) ± standard deviation across three runs for each role, the Majority Voting baseline, and the Synthesizer on ARC-Challenge and MMLU. Bold values indicate the highest mean accuracy on each benchmark.

Condition	ARC-Challenge	MMLU
Analyzer	95.67 ± 0.30	86.67 ± 0.53
Creative	94.53 ± 0.23	86.26 ± 0.92
Critic	94.16 ± 0.05	83.49 ± 1.15
Majority Voting	$\mathbf{95.90 \pm 0.23}$	$\mathbf{87.30 \pm 0.34}$
Synthesizer	95.70 ± 0.35	86.26 ± 0.85

On ARC-Challenge, the Majority Voting baseline achieves the highest mean accuracy (95.90%), followed by the Synthesizer (95.70%), Analyzer (95.67%), Creative (94.53%), and Critic (94.16%). The difference between Majority Voting and the Synthesizer is small (0.20 percentage points), while both substantially outperform the Critic by over 1.5 percentage points.

On MMLU, Majority Voting again achieves the highest mean accuracy (87.30%), exceeding the Analyzer (86.67%) by 0.63 percentage points and the Synthesizer (86.26%) by 1.05 percentage points. The Critic lags behind at 83.49%, resulting in a 3.81 percentage point gap between Majority Voting and the weakest role. Notably, the Synthesizer performs below both Majority Voting and the Analyzer on MMLU, suggesting that the additional synthesis step does not improve answer selection on this benchmark.

The Majority Voting baseline achieves the highest mean accuracy on both benchmarks among all conditions. Across both benchmarks, the Majority Voting answer coincides with the Synthesizer's answer in approximately 94–98% of tasks, indicating that the Synthesizer's answer selection closely mirrors a majority consensus mechanism. All three roles agree on the same answer in 92% of ARC-Challenge tasks and 81% of MMLU tasks, with complete disagreement (no majority) occurring in fewer than 2% of tasks on either benchmark.

Across both benchmarks, three consistent patterns emerge. First, the Critic is the weakest-performing role on both benchmarks. Second, the Synthesizer performs at or near the level of the strongest individual role, indicating that the synthesis process can preserve competitive answer quality while producing an integrated reasoning trace. Third, the Majority Voting baseline matches or slightly exceeds the Synthesizer, suggesting that much of the Synthesizer's answer-selection behavior can be approximated by simple plurality voting. The interpretation of these patterns is discussed in Section 7.1.

6.2. Statistical Significance Analysis

Table 3 reports the results of paired t-tests comparing the Synthesizer against each role baseline with Bonferroni-corrected significance threshold ($\alpha = 0.017$). The Majority Voting comparison is treated as a separate exploratory analysis evaluated at $\alpha = 0.05$.

On both benchmarks, the only statistically significant difference is between the Synthesizer and the Critic (ARC-Challenge: $p = 0.013$, Cohen's $d = 4.99$; MMLU: $p = 0.013$, Cohen's $d = 4.98$). The large Cohen's d values reported for this comparison should not be interpreted as evidence of a practically large effect in the conventional sense. With only three observations per condition, the denominator of Cohen's d—the standard deviation of paired differences—is estimated from an extremely small sample, making the resulting effect size highly sensitive to low within-condition variance (standard deviations of 0.05–1.15 percentage points, as shown in Table 2) [21]. In this case, the inflated d values reflect the high consistency of the performance gap across runs rather than its absolute magnitude: the underlying differences of 1.54 and 2.77 percentage points are modest. We therefore treat Cohen's d here as an indicator of result stability rather than a reliable estimate of effect magnitude and caution against cross-study comparisons based on these values. The statistical significance ($p = 0.013$, Bonferroni-corrected) provides a more interpretable basis for the conclusion that the Synthesizer reliably outperforms the Critic on both benchmarks.

Table 3. Statistical comparison of Synthesizer versus each role baseline and the Majority Voting baseline. Role comparisons use a Bonferroni-corrected threshold ($p < 0.017$); the Majority Voting comparison is evaluated as a separate exploratory analysis ($p < 0.05$). Significant results are marked with *. Italicized rows are benchmark section headers.

Comparison	Δ Mean	t	p	Cohen's d
ARC-Challenge				
vs. Analyzer	+0.03	0.18	0.874	0.10
vs. Creative	+1.17	3.48	0.074	2.01
vs. Critic *	+1.54	8.65	0.013	4.99
vs. Majority Voting	−0.20	−1.94	0.192	−1.12
MMLU				
vs. Analyzer	−0.41	−1.69	0.234	−0.97
vs. Creative	−0.00	−0.00	1.000	−0.00
vs. Critic *	+2.77	8.62	0.013	4.98
vs. Majority Voting	−1.05	−3.55	0.071	−2.05

The Synthesizer versus Analyzer comparison is not significant on either benchmark (ARC-Challenge: $p = 0.874$; MMLU: $p = 0.234$). On ARC-Challenge, the effect size is negligible ($d = 0.10$), confirming that the two conditions perform nearly identically. On MMLU, the negative effect size ($d = -0.97$) suggests a practical advantage for the Analyzer, although this difference does not reach significance. As with the Synthesizer versus Critic comparison, these d values should be interpreted cautiously given the three-run design.

The comparison between the Synthesizer and the Majority Voting baseline reveals a consistent pattern: the Synthesizer performs slightly below Majority Voting on both benchmarks (ARC-Challenge: $\Delta = -0.20$, $p = 0.192$; MMLU: $\Delta = -1.05$, $p = 0.071$), though neither difference reaches statistical significance at the exploratory threshold. The negative Cohen's d values ($d = -1.12$ on ARC-Challenge, $d = -2.05$ on MMLU) suggest a practical advantage for Majority Voting, particularly on MMLU; however, as noted above, these estimates should not be compared directly against conventional effect size benchmarks given the limited number of runs. These results indicate that the Synthesizer's answer-selection behavior largely mirrors majority consensus among the three roles and that the

additional LLM call for synthesis does not improve answer accuracy beyond what simple voting achieves.

The Synthesizer versus Creative comparison presents a notable case on ARC-Challenge: despite a large effect size ($d = 2.01$) and a mean difference of 1.17 percentage points, the p-value (0.074) does not meet the Bonferroni-corrected threshold. This discrepancy between effect size and p-value reflects the limited statistical power inherent in three-run designs, where even consistent practical differences may not achieve significance. On MMLU, the Synthesizer and Creative produce virtually identical results.

In summary, the Synthesizer significantly and consistently outperforms the Critic across both benchmarks while performing comparably to the Analyzer and Creative roles. The Majority Voting baseline matches or slightly exceeds the Synthesizer on both benchmarks, indicating that synthesis does not improve answer accuracy beyond what simple answer aggregation achieves under these conditions. Across all comparisons, Cohen's d values are reported for completeness but should be interpreted as indicators of cross-run consistency rather than estimates of absolute effect magnitude.

6.3. Emergent Reasoning Score Analysis

Table 4 presents the ERS components across both benchmarks.

Table 4. ERS components: mean ± standard deviation across three runs.

Metric	ARC-Challenge	MMLU
Synthesis Effectiveness ($\overline{SE}$)	0.744 ± 0.006	0.711 ± 0.003
Emergent Value ($\overline{EV}$)	0.096 ± 0.010	0.112 ± 0.001
Emergent Reasoning Score ($\overline{\text{ERS}}$)	0.484 ± 0.157	0.441 ± 0.076

Synthesis Effectiveness is high on both benchmarks ($SE = 0.744$ on ARC-Challenge, 0.711 on MMLU), indicating that the Synthesizer consistently incorporates approximately 71–74% of the concepts generated by the three primary roles. The low standard deviations (0.003–0.006) confirm that integration quality is stable across runs. The slightly higher SE on ARC-Challenge suggests that scientific reasoning tasks, which draw on a more constrained knowledge domain, may produce more compatible concept sets that are easier to integrate.

Emergent Value is low on both benchmarks ($EV = 0.096$ on ARC-Challenge, 0.112 on MMLU), meaning that approximately 90% of the Synthesizer's concepts were already present in at least one individual role output. Only about one in ten synthesized concepts is genuinely novel. This finding indicates that the Synthesizer functions primarily as an integrator of existing perspectives rather than a generator of new reasoning.

The composite ERS scores (0.484 on ARC-Challenge, 0.441 on MMLU) reflect the dominant contribution of SE over EV. The higher variance of ERS on ARC-Challenge (std = 0.157 versus 0.076 on MMLU) is at least partly attributable to the greater variance in EV on that benchmark (std = 0.010 versus 0.001), as well as potential instability in the empirically derived weights across runs. This pattern suggests that novel concept generation is less stable than perspective integration and that the weight derivation process may amplify small fluctuations in component scores. Figure 3 visualizes the contrast between SE and EV across both benchmarks.

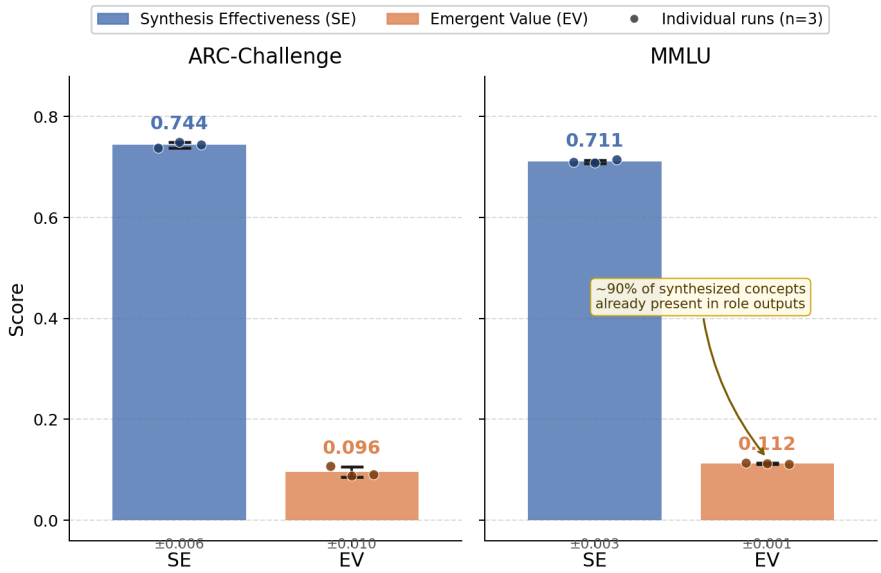


Figure 3. ERS components across benchmarks. Synthesis Effectiveness (*SE*) is consistently high (0.711–0.744), indicating successful integration of role perspectives. Emergent Value (*EV*) remains low (0.096–0.112), confirming that the Synthesizer primarily recombines existing concepts rather than generating novel ones. Error bars indicate standard deviation across three independent runs; dots show individual run values ($n = 3$).

Taken together, the ERS analysis reveals a clear pattern: the Parallel Synthesis architecture is effective at integrating diverse cognitive perspectives (high *SE*) but does not reliably produce genuinely novel concepts (low *EV*). These results are examined further, together with the robustness analyses in Section 6.4, in Section 7.

6.4. Robustness of ERS Components

The results in Section 6.3 rely on three operational choices: a relaxed lexical matching rule (Section 4.3.1), a specific concept-extraction pipeline (Section 4.3.1), and empirically derived weights (Section 4.3.4). This section reports three targeted analyses conducted on the existing runs to bound the sensitivity of our findings to the first two of these choices and to probe the construct validity of the Emergent Value component.

6.4.1. Sensitivity to the Matching Rule

We recomputed *SE* and *EV* under five alternative matching rules spanning the spectrum from strict exact-string equality to the relaxed rule used throughout the main paper. The intermediate rules are substring containment alone (one concept is a character substring of the other), token-set symmetric difference of at most one (a strict reading of "differ by at most one word" in which a substitution counts as two edits), and word-level Levenshtein distance of at most one (the permissive reading in which a substitution counts as one edit). Figure 4 summarizes the results. The qualitative finding $SE > EV$ holds under every relaxed rule on both benchmarks, with the ratio SE/EV ranging from approximately 1.6 (symmetric-difference rule on ARC-Challenge) to 7.2 (combined rule on ARC-Challenge). Under strict exact match, *SE* and *EV* fall to comparable values in the 0.35–0.37 range, showing that the gap observed under relaxed rules comes from concepts that share meaning but differ in surface wording—precisely the wording variation the relaxed rule was designed to absorb. Absolute magnitudes, therefore, depend on the rule, but the direction of the effect is consistent across all four relaxed specifications.

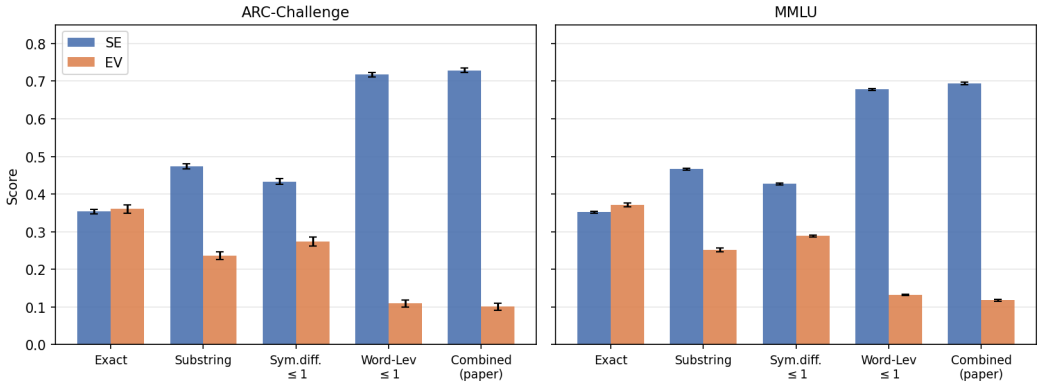


Figure 4. Sensitivity of *SE* and *EV* to the concept matching rule. Bars show the mean across three runs; error bars indicate one standard deviation. The qualitative pattern $SE > EV$ holds under every relaxed rule on both benchmarks. Under strict exact match (leftmost pair), *SE* and *EV* are comparable, illustrating that the relaxed rules absorb wording variation rather than reasoning variation.

6.4.2. Semantic Robustness

To test whether the $SE > EV$ pattern depends specifically on lexical matching, we replaced it with semantic cosine similarity between concept embeddings produced by `sentence-transformers/all-MiniLM-L6-v2` (384-dimensional, L2-normalized) on a random sample of 100 tasks per benchmark drawn from the seed-42 run. Two concepts are treated as equivalent when their cosine similarity meets or exceeds a threshold τ. Table 5 reports *SE*, *EV*, and task-level Pearson correlations with the lexical combined rule at $\tau \in \{0.70, 0.80, 0.85\}$. The $SE > EV$ ordering holds at every threshold on both benchmarks, replicating the pattern observed under lexical matching. Task-level Pearson correlations between the two methods are moderate, in the 0.40–0.65 range, indicating that they measure related but not identical constructs: semantic matching credits paraphrases that share meaning without shared tokens, while lexical matching credits one-word-edit equivalences between morphologically distant forms. Neither method is uniformly more permissive than the other. Correlations are consistently higher on MMLU (0.56–0.65) than on ARC-Challenge (0.40–0.48), which we attribute to the more domain-general vocabulary of MMLU compared with the specialized scientific terminology of ARC-Challenge; under more specialized vocabulary, lexical morphological matches and embedding-based neighborhoods diverge more frequently. The qualitative conclusions of Section 6.3 are preserved under semantic matching.

Table 5. Semantic robustness check using sentence-transformer embeddings (`all-MiniLM-L6-v2`) on a 100-task sample per benchmark from the seed-42 run. Two concepts are treated as equivalent when their cosine similarity meets or exceeds τ. Task-level Pearson *r* columns compare the semantic metric with the lexical combined rule from the main paper.

Benchmark	Method	*SE*	*EV*	*r* (*SE*)	*r* (*EV*)
ARC-Challenge	Lexical (paper)	0.739	0.097	–	–
	Semantic $\tau = 0.70$	0.516	0.190	+0.484	+0.398
	Semantic $\tau = 0.80$	0.435	0.271	+0.450	+0.397
	Semantic $\tau = 0.85$	0.409	0.296	+0.450	+0.415
MMLU	Lexical (paper)	0.672	0.119	–	–
	Semantic $\tau = 0.70$	0.496	0.213	+0.628	+0.621
	Semantic $\tau = 0.80$	0.428	0.279	+0.647	+0.622
	Semantic $\tau = 0.85$	0.400	0.309	+0.622	+0.561

6.4.3. Construct Validity of Emergent Value

A residual concern for EV is whether it responds to the structural conditions that synthesis is meant to address, or whether it is generated uniformly across task types and therefore functions as noise. To test this, we computed, for each task, the mean pairwise Jaccard distance between the three role concept sets under the paper's matching rule, $D(t) = 1 - \bar{J}(C_A, C_C, C_E)$. $D(t)$ captures the conceptual divergence among the independently produced role outputs—the structural condition under which synthesis is expected to do the most reconciliation work. If EV tracks this work, we expect D and EV to correlate positively. Across all task $\times$ seed observations, the Pearson correlation is $+0.065$ on ARC-Challenge ($p < 10^{-3}$, $N = 3510$) and $+0.097$ on MMLU ($p < 10^{-10}$, $N = 4584$). Spearman ρ values are slightly higher (0.085 and 0.112 respectively). The correlation is small in magnitude—D explains under one percent of task-level variance in EV—but is in the predicted direction and highly statistically significant on both benchmarks. We interpret this as modest but unambiguous construct validity: EV is responsive to inter-role diversity, consistent with its role-theoretic motivation, while the small magnitude is itself consistent with the paper's central finding that the architecture's novel-concept generation is limited in scale. As a finer check, we also compared EV between tasks on which all three primary roles produce the same final letter answer and tasks on which at least one dissents; EV did not differ between the two conditions on either benchmark (ARC: $\Delta EV = -0.009$, $p = 0.85$; MMLU: $\Delta EV = -0.005$, $p = 0.94$). The null at the answer level, combined with the positive correlation at the concept level, indicates that synthesis responds to reasoning diversity in the underlying concept sets rather than to surface disagreement over the final letter.

7. Discussion

The results of this study offer a nuanced perspective on the potential and current limitations of cognitively grounded multi-agent LLM collaboration. This section discusses the implications of our findings for multi-agent system design, acknowledges the limitations of the present study, and identifies directions for future research.

7.1. Interpretation of Results

The results address both research questions posed in Section 1.2, though the answers are more nuanced than initially anticipated.

Regarding RQ1, the evidence supports reliable perspective integration but only limited emergent reasoning under our operational definition. While the Synthesizer successfully draws on all three cognitive perspectives (Section 6.3), genuinely novel concepts remain rare. The large Cohen's d values reported in Section 6.1 reflect the high consistency of this pattern across runs rather than extraordinarily large performance gaps, reinforcing the interpretation of synthesis as a reliability mechanism rather than a source of dramatic performance improvement. These interpretations are supported by three robustness analyses reported in Section 6.4: a sensitivity check across five matching rules, a semantic robustness check using sentence-transformer embeddings, and a construct-validity check correlating EV with inter-role concept diversity. Together, they show that the ordering $SE > EV$ is robust to the choice of matching criterion and that EV responds in the predicted direction to the structural pressure of divergent role perspectives, while confirming that the absolute magnitudes of both components are sensitive to the matching rule and should not be interpreted as rule-invariant. Two factors may explain this pattern.

First, the choice of closed-form benchmarks was deliberate for a pilot study: multiple-choice questions with unambiguous ground truth provide a clean evaluation signal and enable exact-match accuracy as a reliable validity anchor for the ERS weight derivation. However, this design choice imposes a structural ceiling on Emergent Value. When the

correct answer is one of four labeled options, the solution space available to the Synthesizer is inherently constrained, and the conceptual distance between any synthesized response and the pre-existing role outputs is bounded by the vocabulary of the answer options themselves. Low *EV* under these conditions, therefore, should not be interpreted as evidence that the architecture is incapable of novel reasoning in principle: it reflects the limited opportunity for conceptual novelty afforded by the task format. Open-ended generation tasks (where the solution space is unbounded and multiple valid responses are possible) would provide a more discriminating test of whether cognitively grounded synthesis produces genuinely novel conceptual contributions [19,20].

Second, the parallel execution design itself limits emergent generation. Because the three primary roles operate independently, the Synthesizer must reconcile perspectives formed in isolation. While this preserves cognitive diversity [44], it prevents iterative building on each other's ideas. Sequential interaction designs, despite their susceptibility to anchoring bias [13], may offer greater potential for emergent concept generation through iterative refinement. Of these two factors, we consider benchmark constraints to be the more likely explanation. Even in sequential multi-agent systems that allow iterative refinement [13,17], genuine emergence on closed-form tasks has not been convincingly demonstrated; the solution space of multiple-choice questions inherently limits the opportunity for novel concept generation regardless of interaction topology.

The comparison between parallel and sequential interaction designs warrants a more precise characterization of the trade-off. Sequential designs such as multi-agent debate [13] allow each agent to condition its response on prior outputs, which enables iterative refinement but risks anchoring: once an early agent commits to a direction, subsequent agents are more likely to converge toward it regardless of its quality [45]. Parallel designs preserve independence but sacrifice the ability to build incrementally on prior reasoning. The present results suggest that for closed-form tasks, the independence benefit of parallel execution is sufficient to preserve accuracy while achieving high *SE*. Whether sequential designs would yield higher *EV* on open-ended tasks by allowing agents to generate genuinely new concepts through iterative exchange is an open empirical question that the current study cannot answer. A direct parallel-versus-sequential comparison within the same cognitive role framework, holding models and tasks constant, is identified as a priority for future work.

Setting aside the parallel–sequential trade-off, which future work will need to resolve empirically, the ERS framework already demonstrates practical utility by revealing a pattern that accuracy alone would obscure. Without ERS, the Synthesizer's comparable accuracy to the Analyzer and Creative could be interpreted as evidence that synthesis adds no value. The ERS analysis reveals instead that the Synthesizer draws on all three perspectives while maintaining accuracy, a qualitatively different mode of arriving at the correct answer than relying on a single reasoning strategy.

The Majority Voting comparison provides further insight into the Synthesizer's contribution. The near-identical answer selections between the two conditions (94–98% agreement) suggest that the Synthesizer implicitly performs a form of majority voting when selecting its final answer. On the roughly 2–6% of tasks where the Synthesizer departs from the majority, it does not consistently improve accuracy, as evidenced by the slight advantage of Majority Voting on both benchmarks. However, this finding does not diminish the value of the Synthesizer as measured by ERS. While Majority Voting selects an answer through simple plurality, the Synthesizer produces an integrated reasoning trace that draws on all three cognitive perspectives, as reflected in the high *SE* scores (0.711–0.744). This distinction is consequential for applications where the reasoning process matters alongside answer correctness, such as open-ended generation, explanation tasks, or domains where

transparent justification is required. In sum, the Synthesizer's primary contribution lies not in answer selection, where it approximates voting, but in perspective integration, which simple aggregation cannot provide. This distinction is the central empirical finding of the paper and should be read as a clarification of scope rather than a negative result: the architecture was not designed to exceed majority voting in accuracy but to produce reasoning traces that draw on multiple cognitive perspectives. ERS is the instrument that makes this contribution visible. Practitioners who require only answer correctness should use majority voting; practitioners who require interpretable, multi-perspective reasoning traces should use the Synthesizer and evaluate it with *SE* rather than accuracy.

The Critic's consistent underperformance, observed on both benchmarks (Section 6.1), warrants specific attention. A cognitive prompt focused on identifying weaknesses and complexities appears less suited to tasks that require direct answer selection. The Critic's value may lie not in standalone accuracy but in its contribution to the Synthesizer's integration process, as reflected in the high *SE* scores.

Finally, the discrepancy between benchmarks suggests that the benefit of synthesis depends on the uniformity of individual role performance [9]. On ARC-Challenge, where all roles perform above 94%, synthesis achieves the highest accuracy. On MMLU, where the Critic drops to 83.49%, the Synthesizer falls slightly below the Analyzer. One possible explanation is that a substantially weaker role pulls the synthesis toward a compromise that dilutes the strongest perspective, an effect analogous to the degradation of collective accuracy observed when individual judgments are influenced by lower-quality contributions [45]. Testing this hypothesis would require systematic variation of individual role quality, which we leave to future work.

More broadly, these results suggest that characterizing the boundaries of emergence is itself informative for the field. Understanding where multi-agent collaboration does not produce novel reasoning is as valuable for system design as identifying where it does, as it prevents over-investment in architectures that primarily replicate capabilities already available from individual models.

7.2. Implications for Multi-Agent System Design

Our findings carry several practical implications for the design and deployment of multi-agent LLM systems.

First, the high Synthesis Effectiveness confirms that cognitively grounded role design, combined with parallel execution, provides a viable mechanism for preserving diverse reasoning perspectives through the synthesis process [13,17]. For practitioners, this supports the value of enforcing independence between agents during generation and delegating integration to a dedicated synthesis component.

Second, the low Emergent Value challenges the implicit expectation in multi-agent research that structured collaboration inherently produces emergent reasoning [12,31]. The primary value of the architecture appears to lie in reliability: the Synthesizer consistently outperforms the weakest individual role while performing comparably to the strongest. Multi-agent architectures may therefore be most beneficial not as a mechanism for exceeding peak individual performance, but as a safeguard against poor performance by ensuring that the final output draws on multiple reasoning strategies [9,44].

The comparison with Majority Voting reinforces this point: since the Synthesizer's accuracy does not exceed that of simple voting, practitioners seeking only answer correctness may prefer Majority Voting as a more cost-effective alternative that eliminates the additional inference call. However, for applications where reasoning coherence and perspective integration are valued—such as generating explanations, addressing open-

ended problems, or supporting human decision-making—the Synthesizer's integration capabilities, as captured by *SE*, may justify the additional computational cost.

Third, practitioners should evaluate whether their target task is likely to benefit from cognitive diversity before deploying multi-agent architectures. Our results suggest that the overhead of multiple model calls is best justified when individual roles perform at comparable levels, as in ARC-Challenge, rather than when one role substantially underperforms, as with the Critic on MMLU. For tasks where a single strong model already achieves near-ceiling accuracy, simpler approaches may be more cost-effective [28].

Finally, the ERS framework offers a reusable evaluation tool for the multi-agent research community. By separating integration quality (*SE*) from novelty generation (*EV*), ERS enables researchers to diagnose how a multi-agent system produces its results, distinguishing between effective aggregation and genuine emergence. The empirical weight derivation ensures that ERS adapts to different benchmarks without manual calibration.

7.3. Limitations

This study has several limitations that should be considered when interpreting the results.

Given the fixed computational budget (three full pipeline runs across four 70B-scale models on 2703 tasks), we prioritized depth of analysis on the existing runs over additional replication. The sensitivity analysis, the semantic robustness check, and the construct-validity analysis reported in Section 6.4 are the principal products of this prioritization; they do not substitute for larger-scale replication but materially constrain the space of interpretations that the available data can support.

First, and most significantly, the experimental design does not include an ablation comparing cognitively grounded roles against generic (non-specialized) prompts within the same Parallel Synthesis architecture. This is the principal limitation of the current study. Without this comparison, we cannot determine whether the observed integration patterns are attributable to cognitive role design specifically or whether they would emerge from any set of structurally distinct prompts regardless of their grounding in cognitive theory [18]. Two considerations informed the decision to defer this ablation. First, establishing a working baseline with confirmed operational feasibility was a prerequisite: an ablation conducted against a non-functional or poorly calibrated architecture would not yield interpretable results. Second, a meaningful ablation requires carefully matched generic prompts that preserve structural parallelism and output format while removing cognitive specificity. Designing such prompts that are structurally matched but cognitively neutral is itself a distinct methodological challenge. We treat this design challenge as a primary objective of the follow-up study and prioritize it as the most important direction for future work (Section 7.4). Accordingly, the present study should not be interpreted as demonstrating that cognitive grounding is necessary for the observed integration behavior.

Second, while we compare the Synthesizer against a Majority Voting baseline (Section 6.1), this represents only the simplest aggregation method. More sophisticated ensemble strategies, such as weighted voting based on role-specific confidence scores or learned aggregation functions, remain unexplored and could provide stronger baselines for evaluating the added value of LLM-based synthesis.

Third, although all dense models operate at approximately 70B parameters, the Creative role uses DeepSeek-V3, a Mixture-of-Experts architecture with 671B total parameters and 37B activated per token. While the effective computational scale is comparable to the dense models, the architectural difference means that model scale is not perfectly controlled across all roles. More broadly, the behavior of the architecture may differ with substantially larger or smaller models or with proprietary models that exhibit different reasoning characteristics.

Fourth, concept extraction relies on a single LLM (Mistral Small 24B Instruct) to identify key concepts from each output. This introduces potential systematic biases in how concepts are identified and matched. Although each concept is limited to three words and the relaxed matching criterion (substring or ± 1 word) mitigates wording variation, the pipeline may still miss semantically equivalent concepts expressed in different terms or, conversely, match concepts that are superficially similar but semantically distinct.

Fifth, the ERS framework has not been directly validated against human judgment, and a properly powered human-annotation study remains a gold-standard complement to the automated pipeline. We identify such a study as a priority for follow-up work. We chose not to conduct a small human validation within the scope of this revision because the resulting inter-rater variance on 30–50 tasks with two or three annotators would risk obscuring rather than clarifying the validity of ERS; a correctly scoped human study with multiple expert annotators, a power-analyzedd sample size, and pre-registered coding guidelines deserves its own methodological treatment. In the interim, we rely on three non-human validity signals. The first is convergent validity against an independent semantic similarity method (Section 6.4.2): the $SE > EV$ ordering is preserved when lexical matching is replaced with sentence-transformer cosine similarity. The second is criterion validity via the empirical weight derivation (Section 4.3.4), which ties each ERS component to task accuracy on thousands of instances. The third is construct validity via the correlation of EV with inter-role concept diversity (Section 6.4.3), which places EV on the predicted side of a theoretically motivated structural contrast. None of these signals replaces human validation, but together they constrain the space of interpretations in which ERS could be spurious, and their convergence supports the interpretive claims we make.

Sixth, the statistical analysis is based on three independent runs per condition. While Bonferroni correction, effect size analysis, and bootstrap resampling (Section 5.4) provide rigor within this design, three runs are insufficient for reliable estimation of sampling distributions or for conventional power analysis. The cumulative API inference costs across four models and 2703 tasks per run constrained the feasible number of replications within the scope of this pilot study. All quantitative conclusions should therefore be interpreted as directional evidence motivating larger-scale replication rather than as definitive empirical claims. We recommend a minimum of ten runs for any follow-up study that aims to make stronger statistical claims about ERS component stability.

Seventh, all models used in this study are instruction-tuned and likely share substantial overlap in their pretraining data, which may constrain the effective cognitive diversity achievable through prompting alone. While the use of models from four distinct organizations (Section 5.1) mitigates this concern to some extent, the degree to which prompt-based role differentiation can overcome similarities in underlying model representations remains an open question.

Eighth, the architecture requires eight LLM calls per task (three parallel role calls, one synthesis call, and four extraction calls—one per output), which substantially increases inference cost and latency relative to single-model baselines. At the scale of this study (2703 tasks across three runs), the cumulative cost of four models accessed via a commercial API was the primary constraint on the number of experimental replications. For deployment at a larger scale, the cost of the synthesis and extraction stages would need to be weighed against the integration benefits demonstrated by the SE scores. One cost-reduction strategy is to replace the synthesis LLM call with a deterministic aggregation function for answer selection while retaining the LLM-based synthesis only for tasks requiring an integrated reasoning trace. Latency can be partially mitigated by parallelizing the three role calls and the four extraction calls, reducing the critical path to two sequential LLM

calls (roles in parallel, then synthesis). A full cost-benefit analysis across task types and deployment contexts is left for future work.

Ninth, the Emergent Value component measures lexical novelty rather than semantic novelty: a concept is counted as emergent only if its surface form does not match any concept in the role outputs under the substring-or-$\pm$1-word criterion. This creates two failure modes. First, *EV* may overestimate novelty when the extraction model generates lexically distinct but semantically equivalent concepts across outputs (e.g., *causal inference* and *causal reasoning* counted as distinct). Second, *EV* may underestimate novelty when genuinely new concepts happen to share surface substrings with role concepts. The relaxed matching criterion is conservative in its effect: it reduces *EV* by treating more near-matches as non-novel, meaning the reported *EV* values (0.096–0.112) represent a lower bound on lexical novelty. This does not resolve the gap between lexical and semantic novelty, which remains the primary limitation of the current *EV* operationalization and motivates the embedding-based similarity approach identified in Section 7.4.

Tenth, the assignment of specific models to specific cognitive roles introduces potential confounds between model capability and role function. The Analyzer (Llama 3.3 70B) and Creative (DeepSeek-V3) achieve higher individual accuracy than the Critic (Llama 3.1 70B), and this performance gap may partly reflect differences in model generation rather than differences in cognitive role prompting. Specifically, the Critic's lower accuracy could be attributable to the older model generation (3.1 vs. 3.3) rather than to the inherent difficulty of the critical evaluation role. Disentangling these confounds would require a factorial design in which each cognitive prompt is applied to each model, holding architecture and scale constant across roles. Such a design was not feasible within the computational scope of this study but is identified as a necessary control condition for any follow-up work that makes strong claims about the contribution of cognitive role design specifically.

7.4. Future Work

The limitations identified above point to several directions for future research.

The highest-priority next step is a controlled ablation study that replaces cognitively grounded prompts with generic role descriptions (e.g., "provide your perspective on this task") while keeping the Parallel Synthesis architecture, models, and evaluation pipeline identical. This experiment is necessary to isolate the contribution of cognitive specialization from the structural benefits of parallel generation and synthesis, directly testing whether theoretical grounding in cognitive science produces measurably different collaboration dynamics compared to ungrounded role assignment [18].

Second, evaluating the architecture with models of varying scale, from smaller models (7B–13B) to larger proprietary models, would reveal whether cognitive role effectiveness depends on a minimum level of model capability or whether stronger models benefit differently from role specialization [25,26]. This line of investigation could help identify the conditions under which cognitively grounded collaboration is most effective.

Third, the concept extraction pipeline could be strengthened through human validation. An inter-rater reliability study, in which human annotators independently extract concepts and their judgments are compared against the LLM-based extraction, would quantify the accuracy of the automated pipeline and identify systematic biases. Additionally, semantic similarity methods could replace or supplement the current substring matching approach, potentially capturing deeper conceptual equivalences.

Fourth, extending the evaluation to additional task types, particularly open-ended generation tasks, mathematical reasoning, and tasks requiring creativity, would test whether the low Emergent Value observed in this study is inherent to the architecture or specific to

multiple-choice benchmarks where the solution space is constrained [19,20]. Open-ended tasks may provide a greater opportunity for genuine novel concept generation.

Finally, exploring alternative cognitive role configurations, such as adding roles based on other psychological frameworks or varying the number of parallel roles, would help determine the optimal balance between cognitive diversity and synthesis complexity. The current four-role design represents one theoretically motivated configuration, but the space of possible role combinations remains largely unexplored. Additionally, evaluating the architecture with models trained on substantially different data distributions, such as domain-specific or multilingual models, would help determine the extent to which prompt-based role differentiation depends on diversity in underlying model representations.

8. Conclusions

This paper investigated whether structured collaboration between multiple large language models, each assigned a cognitively specialized role grounded in established psychological theory, can produce emergent reasoning properties absent from individual model outputs.

We proposed the Parallel Synthesis architecture, in which three primary roles, Analyzer (Simon [22]), Creative (Torrance [23]), and Critic (Kahneman [24]), process each task independently and in parallel, and a Synthesizer (Hutchins [8]) integrates their outputs into a final response. To evaluate emergent properties, we introduced the Emergent Reasoning Score (ERS), a composite metric that separates perspective integration (Synthesis Effectiveness) from novel concept generation (Emergent Value).

Experiments on ARC-Challenge and MMLU revealed that the architecture achieves high Synthesis Effectiveness ($SE = 0.711$–0.744) but low Emergent Value ($EV = 0.096$–0.112). The Synthesizer significantly outperforms the weakest role (Critic, $p = 0.013$) while performing comparably to the Analyzer and Creative roles. A Majority Voting baseline achieves comparable accuracy to the Synthesizer on both benchmarks, confirming that the architecture's value lies not in answer selection but in perspective integration. These results suggest that the primary value of cognitively grounded multi-agent architectures lies in reliable perspective integration rather than emergent creativity.

The ERS framework contributes a reusable evaluation tool that enables researchers to distinguish between integration and novelty in multi-agent systems. We hope that this metric, together with the findings reported here, will inform the design and evaluation of future multi-agent LLM systems. More broadly, our results demonstrate that characterizing the boundaries of emergence to understand where it does and does not occur is as informative for the field as demonstrating its presence. In this sense, the study contributes to ongoing work in NLP and large language modeling by clarifying how multi-agent prompting architectures grounded in cognitive science should be evaluated when answer accuracy and reasoning quality do not align.

The empirical results reported here represent a pilot validation on closed-form benchmarks, sufficient to establish a proof of concept and identify consistent directional patterns. We invite the research community to build on the ERS framework and the Parallel Synthesis architecture through larger-scale replications, open-ended task evaluations, and ablation studies identified in Section 7.4.

Author Contributions: Conceptualization, L.S. and Y.L.-F.; methodology, L.S.; software, L.S.; validation, L.S., Y.L.-F. and M.K.; formal analysis, L.S.; investigation, L.S.; resources, L.S.; data curation, L.S.; writing—original draft preparation, L.S.; writing—review and editing, L.S., Y.L.-F. and M.K.; visualization, L.S.; project administration, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Data Availability Statement: The data presented in this study were generated using large language models and analyzed during the research process. Raw model responses, the extracted concept sets, analytical results (including the robustness and construct-validity analyses reported in Section 6.4), and the experimental pipeline code are available upon reasonable request from the corresponding author. No new empirical datasets were created beyond model-generated outputs.

Acknowledgments: The authors thank the open-source community for providing the large language models and frameworks that made this research possible. Generative AI systems were used as the subject of study and as experimental components in the reported methodology; they were not used to draft manuscript content. All manuscript content was carefully reviewed, edited, and approved by the authors, who take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
ARC-Challenge	AI2 Reasoning Challenge
CoT	Chain-of-Thought
ERS	Emergent Reasoning Score
EV	Emergent Value
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MMLU	Massive Multitask Language Understanding
MoA	Mixture-of-Agents
MoE	Mixture-of-Experts
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
SE	Synthesis Effectiveness

Appendix A. Cognitive Role Assignments

This appendix details the cognitive role assignments used in the Parallel Synthesis architecture. Each entry specifies the assigned model, the cognitive function the role operationalizes, and the behavioral description provided in the system configuration.

Appendix A.1. Analyzer

Model: Llama 3.3 70B Instruct (`meta-llama/llama-3.3-70b-instruct`)
Cognitive Function: Systematic decomposition

```
You are an analytical reasoning expert.  Your role is to:
- Break down problems into logical components
- Identify explicit and implicit constraints
- Map dependencies and relationships
- Provide systematic, structured analysis
Focus on precision, logic, and methodical reasoning.
```

Appendix A.2. Creative

Model: DeepSeek-V3 (`deepseek/deepseek-chat`)
Cognitive Function: Divergent thinking

```
You are a creative problem-solver.  Your role is to:
- Explore unconventional approaches and alternatives
```

```
- Think beyond obvious solutions
- Consider multiple perspectives and possibilities
- Generate innovative ideas and connections
Focus on originality, flexibility, and lateral thinking.
```

Appendix A.3. Critic

Model: Llama 3.1 70B Instruct (`meta-llama/llama-3.1-70b-instruct`)
Cognitive Function: Critical evaluation

```
You are a critical evaluator.  Your role is to:
- Identify potential flaws and edge cases
- Challenge assumptions and reasoning gaps
- Assess risks and limitations
- Ensure robustness and completeness
Focus on rigor, skepticism, and quality assurance.
```

Appendix A.4. Synthesizer

Model: Qwen 2.5 72B Instruct (`qwen/qwen-2.5-72b-instruct`)
Cognitive Function: Integrative synthesis

```
You are a synthesis expert.  Your role is to:
- Integrate insights from multiple specialized perspectives
- Identify connections and complementary viewpoints
- Resolve conflicts between different analyses
- Produce a coherent, comprehensive solution
Focus on integration, coherence, and holistic reasoning.
```

Appendix A.5. Concept Extractor

Model: Mistral Small 24B Instruct (`mistralai/mistral-small-24b-instruct-2501`)
Cognitive Function: Concept identification

```
You are a concept extraction specialist.  Your role is to:
- Identify key concepts and ideas in text
- Extract domain-specific terminology
- Recognize important themes and patterns
- Generate structured concept lists
Focus on precision, completeness, and semantic understanding.
```

Appendix B. System Prompts

This appendix provides the full system prompts used for each cognitive role. All prompts follow a standardized structure: the task input is presented first, followed by role-specific instructions organized as a set of guiding questions, and concluding with a required output format that enforces structured extraction of both reasoning and the final answer. The formatting instructions (tag syntax, placement rules) are identical across all roles and are omitted for brevity.

Appendix B.1. Analyzer Prompt

The Analyzer role assignment is defined in Appendix A.1.

```
Question:  {question}

Options:
{choices}

ROLE: [Analyzer role description, see Appendix A.1]
```

```
Analyze this problem systematically. Break it down by addressing the following:
- What are the key components and subproblems?
- What explicit and implicit constraints apply?
- What dependencies or relationships exist between components?
- What reasoning steps lead from the problem structure to the answer?
Work through each component in order before arriving at a conclusion.

REQUIRED OUTPUT FORMAT (follow exactly):

<ANALYSIS>
[Your analytical reasoning here]
</ANALYSIS>

<ANSWER>
your answer letter (A, B, C, or D)
</ANSWER>
```

Appendix B.2. Creative Prompt

The Creative role assignment is defined in Appendix A.2.

```
Question: {question}

Options:
{choices}

ROLE: [Creative role description, see Appendix A.2]

Think about this problem from multiple creative angles. Consider at least 2-3
different perspectives:
- What if the most obvious interpretation is wrong?
- Could there be unexpected indirect or secondary effects?
- Are there analogies from entirely different domains that reframe the problem?
- What would someone from a completely different field focus on that a special-
ist might overlook?
Choose the most insightful creative angle and explore it fully.

REQUIRED OUTPUT FORMAT (follow exactly):

<ANALYSIS>
[Your creative exploration - be bold and original]
</ANALYSIS>

<ANSWER>
your answer letter (A, B, C, or D)
</ANSWER>
```

Appendix B.3. Critic Prompt

The Critic role assignment is defined in Appendix A.3.

```
Question:
{question}

Options:
{choices}

ROLE: [Critic role description, see Appendix A.3]

Critically evaluate this problem before proposing an answer.
Examine the following:
- What assumptions are embedded in the problem or the obvious answer?
- Are there potential flaws, edge cases, or ambiguities?
- Which answer options are most likely to mislead, and why?
- What hidden complexities might a hasty solver overlook?
Identify the most significant weakness in the problem and reason from it to
your answer.
```

```
REQUIRED OUTPUT FORMAT (follow exactly):
<ANALYSIS>
[Your critical evaluation]
</ANALYSIS>
<ANSWER>
your answer letter (A, B, C, or D)
</ANSWER>
```

Appendix B.4. Synthesizer Prompt

The Synthesizer role assignment is defined in Appendix A.4.

```
Question: {question}
Options:
{choices}
ROLE: [Synthesizer role description, see Appendix A.4]
You have received three specialized analyses. Integrate them by addressing
the following:
- Where do the three perspectives agree, and what does that convergence imply?
- Where do they diverge, and how should those contradictions be resolved?
- What complementary insights emerge from combining all three perspectives?
- Which reasoning elements from each role are most relevant to the final answer?
Construct a unified response that draws on all three perspectives rather than
selecting one.
=== ANALYZER ===
{analyzer_response}
=== CREATIVE ===
{creative_response}
=== CRITIC ===
{critic_response}
REQUIRED OUTPUT FORMAT (follow exactly):
<SYNTHESIS>
[Your integrated analysis combining all three perspectives]
</SYNTHESIS>
<ANSWER>
your answer letter (A, B, C, or D)
</ANSWER>
```

Appendix B.5. Concept Extractor Prompt

The Concept Extractor role assignment is defined in Appendix A.5.

```
ROLE: [Concept Extractor role description, see Appendix A.5]
Extract exactly 10 key concepts from the text below. Apply the following rules:
- Prioritize domain-specific terms and important reasoning ideas
- Avoid generic words (problem, solution, approach, question, answer, analysis)
- Use lowercase; limit each concept to 1-3 words
- Ensure concepts are distinct and non-overlapping
Separate all concepts with commas and spaces.
TEXT:
{text}
REQUIRED OUTPUT FORMAT (follow exactly):
<CONCEPTS>
concept1, concept2, concept3, concept4, concept5, concept6, concept7, concept8,
```

```
concept9, concept10
</CONCEPTS>
```

References

1. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223. https://doi.org/10.48550/arXiv.2303.18223.
2. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.A.; Socher, R.; Amatriain, X.; Gao, J. Large Language Models: A Survey. *arXiv* **2024**, arXiv:2402.06196. https://doi.org/10.48550/arXiv.2402.06196.
3. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; Curran Associates: Red Hook, NY, USA, 2020; pp. 9459–9474. https://doi.org/10.48550/arXiv.2005.11401.
4. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022), Online, 25–29 April 2022.
5. Schulhoff, S.; Ilie, M.; Balepur, N.; Kahadze, K.; Liu, A.; Si, C.; Li, Y.; Gupta, A.; Han, H.; Schulhoff, S.; et al. The Prompt Report: A Systematic Survey of Prompting Techniques. *arXiv* **2024**, arXiv:2406.06608. https://doi.org/10.48550/arXiv.2406.06608.
6. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*; Curran Associates: Red Hook, NY, USA, 2022.
7. Malone, T.W.; Laubacher, R.; Dellarocas, C. The Collective Intelligence Genome. *IEEE Eng. Manag. Rev.* **2010**, *38*, 38–52. https://doi.org/10.1109/EMR.2010.5559142.
8. Hutchins, E. *Cognition in the Wild*; MIT Press: Cambridge, MA, USA, 1995.
9. Woolley, A.W.; Chabris, C.F.; Pentland, A.; Hashmi, N.; Malone, T.W. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* **2010**, *330*, 686–688. https://doi.org/10.1126/science.1193147.
10. Cui, H.; Yasseri, T. AI-enhanced Collective Intelligence. *Patterns* **2024**, *5*, 101074. https://doi.org/10.1016/j.patter.2024.101074.
11. Malone, T.W.; Bernstein, M.S. (Eds.) *Handbook of Collective Intelligence*; The MIT Press: Cambridge, MA, USA, 2022; p. 232.
12. Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N.V.; Wiest, O.; Zhang, X. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI 2024), Jeju Island, Republic of Korea, 3–9 August 2024*; IJCAI Organization: Marina Del Rey, CA, USA, 2024; pp. 8048–8057. https://doi.org/10.24963/ijcai.2024/890.
13. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, 21–27 July 2024*; PMLR: Cambridge, MA, USA, 2024; Volume 235.
14. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In Proceedings of the 1st Conference on Language Modeling (COLM 2024), Philadelphia, PA, USA, 7–9 October 2024.
15. Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S.K.S.; Lin, Z.; et al. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
16. Li, G.; Hammoud, H.A.A.K.; Itani, H.; Khizbullin, D.; Ghanem, B. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023*; Curran Associates: Red Hook, NY, USA, 2023.
17. Chen, J.; Saha, S.; Bansal, M. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 7066–7085. https://doi.org/10.18653/v1/2024.acl-long.381.
18. Chen, J.; Wang, X.; Xu, R.; Yuan, S.; Zhang, Y.; Shi, W.; Xie, J.; Li, S.; Yang, R.; Zhu, T.; et al. From Persona to Personalization: A Survey on Role-Playing Language Agents. *arXiv* **2024**, arXiv:2404.18231.
19. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv* **2018**, arXiv:1803.05457. https://doi.org/10.48550/arXiv.1803.05457.
20. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring Massive Multitask Language Understanding. In Proceedings of the International Conference on Learning Representations (ICLR 2021), Virtual Event, 3–7 May 2021.
21. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988; p. 567.
22. Simon, H.A. *Models of Thought*; Yale University Press: New Haven, CT, USA, 1979.
23. Torrance, E.P. *The Torrance Tests of Creative Thinking: Norms-Technical Manual*; Personnel Press: Princeton, NJ, USA, 1974.

24. Kahneman, D. Maps of Bounded Rationality: Psychology for Behavioral Economics. *Am. Econ. Rev.* **2003**, *93*, 1449–1475. https://doi.org/10.1257/000282803322655392.
25. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent Abilities of Large Language Models. *arXiv* **2022**, arXiv:2206.07682. https://doi.org/10.48550/arXiv.2206.07682.
26. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling Laws for Neural Language Models. *arXiv* **2020**, arXiv:2001.08361. https://doi.org/10.48550/arXiv.2001.08361.
27. Schaeffer, R.; Miranda, B.; Koyejo, S. Are Emergent Abilities of Large Language Models a Mirage? In *Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023*; Curran Associates: Red Hook, NY, USA, 2023.
28. Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, M.A.; Abid, A.; Fisch, A.; Brown, A.R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv* **2023**, arXiv.2206.04615. https://doi.org/10.48550/arXiv.2206.04615.
29. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv* **2023**, arXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712.
30. Tian, L.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; Tu, Z. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 17889–17904. https://doi.org/10.18653/v1/2024.emnlp-main.992.
31. Li, J.; Zhang, Q.; Yu, Y.; Fu, Q.; Ye, D. More Agents Is All You Need. *arXiv* **2024**, arXiv.2402.05120. https://doi.org/10.48550/arXiv.2402.05120.
32. Wang, J.; Wang, J.; Athiwaratkun, B.; Zhang, C.; Zou, J. Mixture-of-Agents Enhances Large Language Model Capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025.
33. Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; Gerstein, M. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 599–621. https://doi.org/10.18653/v1/2024.findings-acl.33.
34. Kim, Y.H.; Park, C.; Jeong, H.; Chan, Y.S.; Xu, X.; McDuff, D.; Breazeal, C.; Park, H.W. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In *Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024*; Curran Associates: Red Hook, NY, USA, 2024.
35. Li, X.; Wang, S.; Zeng, S.; Wu, Y.; Yang, Y. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth* **2024**, *1*, 9. https://doi.org/10.1007/s44336-024-00009-2.
36. Plaat, A.; van Duijn, M.; van Stein, N.; Preuss, M.; van der Putten, P.; Batenburg, K.J. Agentic Large Language Models, a Survey. *J. Artif. Intell. Res.* **2025**, *84*, 29. https://doi.org/10.1613/jair.1.18675.
37. Zhu, K.; Du, H.; Hong, Z.; Yang, X.; Guo, S.; Wang, Z.; Wang, Z.; Qian, C.; Tang, X.; Ji, H.; et al. MultiAgentBench: Evaluating the Collaboration and Competition of LLM Agents. *arXiv* **2025**, arXiv:2503.01935.
38. Riedl, C. Emergent Coordination in Multi-Agent Language Models. *arXiv* **2025**, arXiv:2510.05174.
39. Shaikh, S.H. LLM-Based Multi-agent Systems: Frameworks, Evaluation, Open Challenges, and Research Frontiers. In *Communications in Computer and Information Science, Proceedings of the Computational Intelligence (IJCCI 2025), Marbella, Spain, 22–24 October 2025*; Springer: Cham, Switzerland, 2026; Volume 2827, pp. 149–170. https://doi.org/10.1007/978-3-032-15632-7_9.
40. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST 2023), San Francisco, CA, USA, 29 October–1 November 2023*; ACM: New York, NY, USA, 2023. https://doi.org/10.1145/3586183.3606763.
41. Zhuge, M.; Liu, H.; Faccio, F.; Ashley, D.R.; Csordás, R.; Gopalakrishnan, A.; Hamdi, A.; Hammoud, H.A.A.K.; Herrmann, V.; Irie, K.; et al. Mindstorms in Natural Language-Based Societies of Mind. *Comput. Vis. Media* **2025**, *11*, 29–81. https://doi.org/10.26599/CVM.2025.9450460.
42. Guilford, J.P. *The Nature of Human Intelligence*; McGraw-Hill Book Company: New York, NY, USA, 1967.
43. Brown, T. Design Thinking. *Harv. Bus. Rev.* **2008**, *86*, 84–92.
44. Surowiecki, J. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*; Anchor Books: New York, NY, USA, 2005; p. 336.
45. Lorenz, J.; Rauhut, H.; Schweitzer, F.; Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9020–9025. https://doi.org/10.1073/pnas.1008636108.
46. DeepSeek-AI. DeepSeek-V3 Technical Report. *arXiv* **2024**, arXiv:2412.19437. https://doi.org/10.48550/arXiv.2412.19437.

47. Llama Team. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783.
48. Qwen Team. Qwen2.5 Technical Report. *arXiv* **2024**, arXiv:2412.15115. https://doi.org/10.48550/arXiv.2412.15115.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.