The LLM Team Composition Paradox: Why the Right Team Isn't Always Diverse

Developing a Contingency Model on the ARC-AGI Benchmark with an OODA-Belbin Framework Inspired by Multicultural Human Teams

Youssef Hariri
Rennes School of Business and Upgrad

Solving ill-structured, "wicked problems" remains a grand challenge for artificial intelligence, demanding novel agentic architectures. The Abstraction and Reasoning Corpus (ARC) serves as a canonical benchmark for this challenge. Crucially, our analysis moves beyond the standard ARC metric of a 'perfect pixel match' to include a rigorous audit of the underlying code, revealing that many apparent successes are 'false positives' with non-generalizable logic. To investigate the optimal approach, this research first conducts a systematic comparison of six distinct LLM-based agentic conditions, ranging from a monolithic single agent to a real-time adaptive multi-agent system inspired by the cognitive synergy of human multicultural teams. Our findings challenge the prevailing assumption that architectural complexity guarantees superior performance, revealing instead a data-driven contingency model. The results demonstrate that: (1) a significant portion of ARC tasks are most effectively solved by a single agent specializing in holistic, abstract reasoning; (2) a recurring "Expert Anomaly" shows a non-diverse, homogeneous team of 'Implementer' agents uniquely solving high-precision procedural tasks; and (3) the most algorithmically complex "Cognitive Labyrinth" problems are mastered not by the most complex system, as initially hypothesized, but by a cognitively diverse, heterogeneous static team. This surprising outcome is starkly contrasted by the poorest-performing condition, our most complex adaptive system, whose results provide strong evidence for a "cost of complexity" in agentic reasoning.

An initial contingency model, derived from preliminary pilot studies, was then operationalized in a novel Hierarchical OODA-Belbin Framework. Its strategic layer, a Strategic Selector Agent, applies this model to deploy the optimal architecture. However, testing this system against the most difficult challenges revealed a profound limitation: "Generative Exhaustion." To combat this, the framework's tactical layer, a Self-Correction Loop, was developed, achieving the first successful solve on these previously unsolvable tasks. The complete system transforms our descriptive model into a functional, prescriptive one, signaling a necessary shift toward a portfolio-based approach and highlighting the critical need to augment standard AI benchmarks with logical code validation.

Keywords: Multi-Agent Systems (MAS), LLM Agents, Agentic AI, Team Composition, Cognitive Diversity, Contingency Model, Abstraction and Reasoning Corpus (ARC), Wicked Problems, Design Science Research (DSR), Qualitative Code Audit, Cost of Complexity, OODA Loop, Belbin Team Roles

1. Introduction

The pursuit of Artificial General Intelligence (AGI) is increasingly focused on creating systems capable of solving ill-structured, "wicked problems"—challenges characterized by ambiguity, emergent properties, and the absence of predefined solution paths. The Abstraction and Reasoning Corpus (ARC) stands as a canonical benchmark in this domain, designed to measure an AI's capacity for genuine fluid intelligence and on-the-fly skill acquisition. A prevailing assumption in the field is that tackling such complexity necessitates a corresponding increase in the complexity of the AI solution. This has fueled a drive toward developing sophisticated multi-agent systems (MAS) that are diverse, collaborative, and adaptive, often drawing inspiration from the cognitive synergy of high-performing human multicultural teams. The implicit hypothesis is that a more complex, human-like architecture will yield superior results.

However, is this assumption universally valid? While the benefits of diversity and adaptation are well-theorized, the inherent costs of architectural complexity—including computational overhead, communication friction, and coordination challenges—are less frequently interrogated. This research challenges the monolithic pursuit of complexity by asking a more fundamental question: What is the relationship between an agentic framework's design and its performance on problems of varying structure? Does the heterogeneity advantage always hold, or are there classes of wicked problems for which simpler, more focused architectures are superior?

To answer these questions, this paper presents a systematic, empirical comparison of six distinct LLM-based agentic conditions on the ARC benchmark. Crucially, our analysis moves beyond the standard 'perfect pixel match' metric to include a rigorous audit of the underlying code—a step that proves essential for uncovering the true nature of agentic performance. This deeper analysis reveals a surprising, data-driven contingency model that directly challenges several prevailing assumptions. Rather than a simple hierarchy of performance, our results reveal a series of paradoxes: a monolithic single agent proves uniquely capable on tasks requiring holistic abstract reasoning; a non-diverse, homogeneous team validates a recurring "Expert Anomaly" on Local Geometric & Procedural Analysis tasks; and, most significantly, the most algorithmically complex "Cognitive Labyrinths" are mastered not by the most complex adaptive system, as might be expected, but by a cognitively diverse static team. This study's primary contribution is therefore the articulation of this evidence-based contingency model, which is operationalized in our final artifact, a Hierarchical OODA-Belbin Framework. Ultimately, our findings force a necessary re-evaluation of the costs of complexity and the methodological rigor required for the evaluation of agentic AI[1].

2. Literature Review

2.1 The Nature of the ARC Challenge

2.1.1 The Frontier of AI Reasoning

The Abstraction and Reasoning Corpus is meticulously designed to assess an AI system's capacity for genuine, on-the-fly skill acquisition (Chollet, 2019). Introduced in 2019 by the French Engineer and AI Researcher, François Chollet, ARC distinguishes itself by presenting tasks that necessitate the derivation of underlying rules and patterns from minimal examples, as an emphasis on abstraction and reasoning to facilitate fair comparisons between AI systems and human-like

[1] All data, code, and supplementary materials associated with this research are permanently archived on Zenodo (DOI: https://doi.org/10.5281/zenodo.17490945), and are also available on GitHub at: https://github.com/youssef-hariri/OODA-Belbin-Framework

general intelligence (Chollet, 2019; Chollet et al., 2025). The ongoing challenge posed by ARC, even after several years, underscores its significance in driving progress toward Artificial General Intelligence by focusing on generalization to novel tasks, which is considered the essence of intelligence (Chollet et al., 2024). Recent advancements have led to the development of ARC-AGI-2, an upgraded version that seeks even finer-grained evaluation at higher levels of cognitive complexity, while maintaining the core task format of its predecessor (Chollet et al., 2025).

This continuous evolution of the ARC benchmark highlights its critical role in pushing the boundaries of what AI can achieve in abstract reasoning and dynamic problem-solving (Cole & Osman, 2025; Yang et al., 2025).

The ARC corpus is made of two parts: the first part contains a set of demonstration pairs (from one to five). Each pair consists of an "input" grid and a corresponding "output" grid. The LLM must analyze these examples to infer the underlying abstract rule or transformation that maps each input to its respective output.

The second part contains a "test" input grid that is previously unseen by the LLM. The LLM must then apply the inferred transformation to this grid to generate a final solution. The evaluation criterion is strict and binary: a task is considered successfully solved only if the generated output grid is a 100% pixel-perfect match with the ground-truth solution; any deviation results in failure. Two examples of ARC-AGi 2 problems are presented in Figure 1 and Figure 2

2.1.2 From Well-Defined to Ill-Structured Problems

Situating ARC within the broader cognitive science typology of problems, it becomes apparent that the benchmark represents a hybrid challenge that transcends the traditional dichotomy between well-defined and ill-structured problems (Laxman, 2011; Schraw et al., 1995). As articulated by (Jonassen, 2009), well-defined problems are characterized by clear initial states, defined goals, and identifiable solution paths, often with a single, guaranteed solution (Laxman, 2011; Schraw et al., 1995). Conversely, ill-structured problems, as discussed by (Schraw

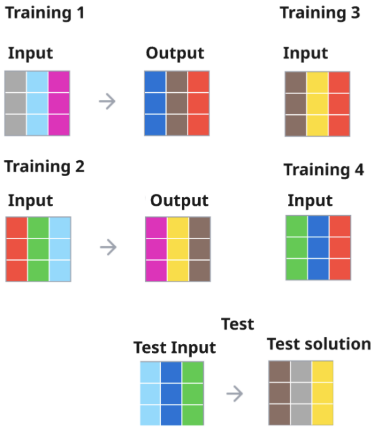

Figure 1

An example of a task from the Abstraction and Reasoning Corpus (ARC), Problem 0d3d703e (colors transformation pattern) training and test sets

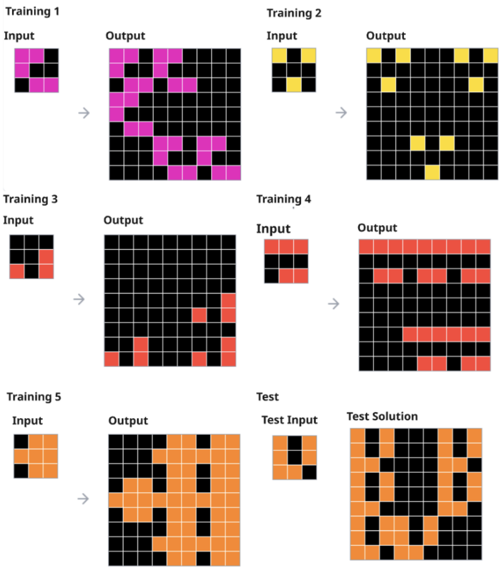

Figure 2

An example of a task from the Abstraction and Reasoning Corpus (ARC), Problem 007bbfb7 (creation of a bigger tile using the initial one) training and test sets et al., 1995), are often vague, open-ended, and possess multiple, non-guaranteed solution pathways, frequently due to inadequate information regarding their components or transformations (Laxman, 2011; Taylor, 1974). ARC, therefore, presents a unique two-phase structure: an initial ill-structured phase demanding novel rule discovery and pattern inference, where AI systems must effectively "induce programmatic skills" or "infer rules from data" under potentially noisy observations (Li et al., 2025; Z. Z. Wang et al., 2025), followed by a more well-defined phase of implementing these discovered rules to generate solutions. This intricate blend requires AI systems to not only learn structured declarative rule sets (Fürnkranz et al., 2020) but also to engage in sophisticated knowledge discovery (Sarker et al., 2024), thus positioning ARC as a complex problem that necessitates adaptive problem-solving strategies rather than static, pre-programmed approaches.

2.1.3 Characterizing ARC as a "Wicked Problem"

Formally defining the Abstraction and Reasoning Corpus as a "wicked problem" provides a robust framework for understanding its inherent challenges, aligning with the seminal work of Rittel and Webber (Rittel & Webber, 1973). These problems, often found in social policy, are distinguished from "tame" problems by several key characteristics: there is no definitive formulation, no stopping rule to determine when a solution is reached, and solutions are not true-or-false but rather good-or-bad (Ojasalo & Koski, 2019). Furthermore, wicked problems involve multiple stakeholders with potentially conflicting perspectives and evolving requirements, making their scope and scale indeterminate (Head & Alford, 2013). Applying this framework to ARC, the challenge lies in the absence of a pre-defined algorithm or clear set of rules for each new task; instead, the system must grapple with ill-defined components and transformations to discover underlying patterns, making "a priori planning" an inherently flawed strategy (Ojasalo & Koski, 2019). Just as with complex social issues like climate change or public health (Newman & Head, 2017; Stavros et al., 2021), ARC requires a continuous learning and adaptive approach, where solutions are provisional and subject to refinement, reflecting the "wicked tendencies" of such problems (Newman & Head, 2017). This framing positions ARC as a problem that necessitates novel approaches in artificial intelligence, moving beyond conventional problem-solving methods to embrace the complexities of genuine fluid intelligence (H.-Y. Liu & Maas, 2021; Locklear, 2025).

This framing of ARC as a wicked problem necessitates novel approaches, but it also raises a critical question: is a single type of advanced architecture, such as a diverse, adaptive team, the universal solution, or does the nature of wickedness itself demand a more contingent approach to problem-solving?

2.1.4 Implications for Problem-Solving

The reframing of benchmarks like the Abstraction and Reasoning Corpus as "wicked problems" carries profound implications for the design and implementation of problem-solving strategies in artificial intelligence. This perspective explains why initial static planners, which rely on *a priori* defined rules and fixed operations, have documented limitations when confronted with such challenges. Traditional planning methods, often hand-coded or based on comprehensive, predefined models, prove insufficient and lack robustness or generality when applied to dynamic or unknown environments (Geffner, 2018; Matveev et al., 2015). These systems typically assume stable environments and require prior knowledge, leading to failures when assumptions are violated by unanticipated external pressures or novel circumstances (Matveev et al., 2015; Usenko et al., 2017).

Instead, the inherent "wickedness" of ARC tasks necessitates a fundamental shift towards adaptive frameworks. This is not merely an engineering improvement but a necessary response to the nature of problems where traditional methods fail and solutions are not easily defined (Gruetzemacher, 2018). AI systems designed for such complex, real-world problems must be able to adapt to unknown situations and continually learn from novelties, highlighting the critical need for adaptability in dynamic environments (Dellermann et al., 2021; Dusparić & Cardozo, 2021; B. Liu et al., 2023). This implies a move towards continuous learning and iterative adjustment, where solutions are provisional and evolve through engagement with the problem space (Locklear, 2025).

Furthermore, the limitations of current large language models in complex planning tasks further underscore this point. While LLMs show impressive generative capabilities, they often struggle with the precise resource management, consistent state tracking, and strict constraint compliance required for more complex planning scenarios (Aghzal et al., 2025; Goebel & Zips, 2025). Studies indicate that LLMs perform poorly in complex and long-horizon reasoning tasks, highlighting their current inability to achieve human-level planning abilities in many real-world benchmarks (Aghzal et al., 2025; Lin et al., 2024; Xie et al., 2024). These shortcomings reinforce the argument that effective solutions for wicked problems like ARC will require adaptive AI architectures capable of continuous learning and dynamic strategy refinement, extending beyond the current capabilities of static or less adaptive LLM approaches.

2.2 Competing Theories of AI Team Composition

2.2.1 The Rationale for Multi-Agent Systems and the Heterogeneity Advantage

The foundational logic of Multi-Agent Systems for tackling complex problems, particularly those akin to the Abstraction and Reasoning Corpus, is rooted in the "heterogeneity advantage," where diverse teams consistently outperform homogeneous ones on complex tasks (Hong & Page, 2004). This concept, famously articulated by (Hong & Page, 2004), posits that when selecting a problem-solving team from a diverse population of intelligent agents, a team of randomly selected agents can even outperform a team comprised of individually high-ability, but less diverse, agents. The strength of this approach lies in the functional diversity of agents, encompassing differences in how they represent problems and the algorithms or heuristics they employ to generate solutions (Hong & Page, 2001).

In the context of AI, heterogeneous multi-agent systems offer significant practical advantages compared to homogeneous systems, particularly in addressing complex scenarios that require varied abilities and functionalities (Fu et al., 2023; X. Liu et al., 2024). By leveraging the distinct capabilities and roles of individual agents, MAS can collectively address intricate tasks through collaboration, proving to be a robust substitute for traditional software systems in certain application domains (Han et al., 2024; Šalamon, 2011). This collaborative paradigm is especially pertinent for problems that demand dynamic adaptation and nuanced problem-solving, as seen in the challenges posed by ARC. Studies across various fields, from embodied multi-agent collaboration to collective problem solving, reinforce that diversity among agents, whether in strategies or perspectives, can lead to superior performance and more complex model production for challenging problems (Aminpour, Gray, et al., 2021; Boroomand & Smaldino, 2021; X. Liu et al., 2024). This makes MAS a compelling framework for designing AI systems capable of navigating the complexities of abstract reasoning.

However, the cognitive and computational overhead associated with multi-agent coordination is non-trivial. This raises the critical question of whether the heterogeneity advantage is absolute or if, for certain problem types, the costs of collaboration outweigh the benefits, making a monolithic or even a non-diverse approach superior—a central question this research aims to answer.

2.2.2 The Contingency Thesis: Challenging the Universality of the Heterogeneity Advantage

While the "heterogeneity advantage" emphasizes the benefits of diverse teams for complex problem-solving, it is crucial to advance a more nuanced, contingency-based view of cognitive diversity. This perspective acknowledges that, under specific conditions, a homogeneous team might even outperform a diverse one, highlighting the "double-edged sword" nature of diversity (Martins et al., 2012).

The "Double-Edged Sword" of Diversity. Cognitive diversity, characterized by variations in how individuals think, process information, and solve problems, is often lauded for bringing a broader range of perspectives and knowledge resources to a team (Aminpour, Schwermer, & Gray, 2021; Heidari et al., 2023). This can lead to enhanced decision-making, innovation, and economic growth by promoting deeper information processing and complex thinking (Aggarwal & Woolley, 2018; Galinsky et al., 2015). However, this advantage comes with inherent costs, primarily in the form of increased coordination overhead (Schimmelpfennig et al., 2021).

The literature treats cognitive diversity as a trade-off: while it offers the benefits of richer cognitive resources, it can also incite detrimental forms of conflict and resentment, leading to challenges in communication and coordination (Chen, Liu, et al., 2019; Galinsky et al., 2015; Schimmelpfennig et al., 2021). Managing diverse teams effectively requires significant effort to overcome potential negative effects such as interpersonal conflict and decreased cohesion, especially if conflict management is low (J. Liu et al., 2023; Mello & Delise, 2015; Nowak, 2020). Some research even suggests that executive diversity can inhibit comprehensive examinations and extensive long-range planning (Miller et al., 1998). For instance, a study by (Horwitz & Horwitz, 2007b) highlights the complex and sometimes inconsistent findings regarding the impact of team diversity on outcomes, underscoring that the benefits are not always realized without effective management strategies (Mehta & Saxena, 2025). The initial composition of a team can even determine whether its dynamics will move towards or away from optimal performance, emphasizing the tension between the benefits of informational diversity and potential affinity biases (Heidari et al., 2023).

Task Complexity as a Key Moderator. The impact of cognitive diversity on team performance is significantly moderated by task complexity (Higgs et al., 2005; Zabalandikoetxea et al., 2021). A contingency model suggests that cognitive diversity is positively related to performance for complex tasks and negatively related for straightforward tasks (Dinwoodie, 2005; Higgs et al., 2005).

For complex, ill-defined, or creative problems, diverse cognitive styles allow teams to explore a wider range of solutions and perspectives, leading to more innovative and robust outcomes (Aggarwal & Woolley, 2018; Aminpour, Schwermer, & Gray, 2021). Cognitive diversity brings unique combinations of cognitive resources, such as variations in how people think and solve problems, which can be critical for improving group performance in complex problem-solving (Aminpour, Schwermer, & Gray, 2021; Dong et al., 2021). Research indicates that for tasks requiring broad exploration and novel solutions, cognitive diversity amplifies the ability to identify critical cognitive resources within the team and enhances team creativity (Aggarwal & Woolley, 2018).

Conversely, for straightforward or execution-focused tasks, high cognitive diversity can introduce inefficiencies (Higgs et al., 2005). In such scenarios, the benefits of varied perspectives are outweighed by the increased coordination costs and potential for miscommunication (Dinwoodie, 2005). Studies have shown that while heterogeneous task cognition is beneficial in the strategizing phase of a task, homogeneous task cognition can be more useful during the implementation and adjustment phases (S. Wang et al., 2019). This highlights that for tasks primarily focused on execution with well-defined parameters, a more homogeneous team might achieve higher efficiency due to reduced communication and coordination overhead (Aggarwal & Woolley, 2013; Nowak, 2020). Therefore, understanding the nature of the task is crucial for determining the optimal level and type of cognitive diversity for a given team (Higgs et al., 2005).

2.3 Methodological Foundations for an Adaptive Framework

2.3.1 Realizing Heterogeneity via Cultural Personas

To effectively realize the "heterogeneity advantage" within AI multi-agent systems, particularly in tackling complex problems like ARC, this research proposes leveraging persona-conditioning through cultural frameworks. This approach allows for the creation of functionally specialized agents that effectively mimic multicultural human teams (Bhalerao et al., 2025). The concept of imbuing AI agents with human-like personalities or personas is not new; conversational agents, for example, have long been designed with distinct traits like gender or a backstory to establish rapport and enhance user interaction (Bickmore & Gruber, 2010; Pradhan & Lazar, 2021). Furthermore, recent research has explored how projecting human personality traits onto agents can guide their behavior and how generative AI personas can increase collective diversity in human ideation (Lim et al., 2025; Wan & Kalman, 2025).

Building on this, conditioning agents with personas derived from diverse cultural perspectives, such as those inspired by cultural geography or through multi-agent debate frameworks for cultural alignment, allows AI systems to achieve a more nuanced understanding and approach to problem-solving (Ki et al., 2025; Kovač et al., 2023; J. Yuan et al., 2024). This method ensures that AI agents, akin to human team members from varied backgrounds, bring distinct cognitive styles and problem representations to the collective effort, moving beyond predominantly Western-centric AI models (Anderson et al., 2021; Chhikara et al., 2025; Prabhakaran et al., 2022). This cultural attunement is crucial for problems where a lack of diverse understanding could lead to less effective or even biased solutions (Anik et al., 2025; Feng et al., 2025; Villanueva et al., 2025).

The effectiveness of such an approach is further underscored by the inherent trade-off in designing diverse teams, often termed the "persona-criterion trade-off," which suggests that a team-based methodology is essential to achieve a balanced solution that accounts for various facets of a problem (Hariri, 2025). Research into multi-agent systems highlights the importance of heterogeneous agents, demonstrating that performance gains are often limited when frameworks rely on a single LLM to drive all agents, thus constraining the system's overall intelligence (Cemri et al., 2025; Ye et al., 2025). Therefore, by integrating diverse LLMs into heterogeneous multi-agent systems, the collective intelligence is elevated, as different LLMs can contribute varied capabilities and roles (Ye et al., 2025). This strategy directly addresses the need for adaptive problem-solving by ensuring that the AI team encompasses a broad spectrum of "thinking styles" and approaches, thereby enabling a more comprehensive and robust engagement with complex, ill-defined tasks (Wan & Kalman, 2025).

2.3.2 The State-of-the-Art Search Algorithm: Adaptive Branching MCTS

A significant challenge in scaling large language models for complex problem-solving, especially for "wicked problems" like ARC, lies in the "unbounded branching" of their solution spaces. Unlike traditional algorithms with defined steps, LLMs can generate a near-infinite number of responses from a single prompt, making systematic exploration and efficient navigation of possibilities difficult (Inoue et al., 2025). While LLMs have demonstrated impressive reasoning abilities through test-time computation techniques, current models often lack the ability to systematically explore these vast solution spaces, leading to issues like invalid reasoning steps or redundant explorations (J. Lu et al., 2025). This "wandering" rather than systematic exploration can hinder their effectiveness in complex, multi-step problems (J. Lu et al., 2025).

To address this, Adaptive Branching Monte Carlo Tree Search, proposed by (Inoue et al., 2025) and developed by Sakana AI, emerges as a state-of-the-art technical framework for orchestrating multi-agent search. AB-MCTS is a novel inference-time framework that generalizes repeated sampling with principled multi-turn exploration and exploitation. Its core mechanism lies in its dynamic decision-making process at each node of the search tree: it intelligently decides whether to "go wider" by expanding new candidate responses (exploration) or "go deeper" by revisiting and refining existing ones (exploitation), based on external feedback signals to allow the system to navigate the LLM solution spaces more effectively (Inoue et al., 2025).

The empirical validation of the **Adaptive Branching Monte Carlo Tree Search (AB-MCTS)** is particularly compelling, especially in its application to the ARC-AGI benchmark. The original study demonstrated that while AB-MCTS can achieve Pass@2 scores as high as **18.3%**, performance increases significantly to a **30% success rate with 250 iterations** (Pass@250) (Inoue et al., 2025). This work, which surpasses previous methods like repeated sampling and standard MCTS, establishes a powerful technical precedent. However, the performance gap indicates that substantial room for improvement exists. Our research builds on this foundation by exploring how different **LLM team structures** can enhance problem-solving effectiveness within the **same computational budget**

2.3.3 Prevailing Limitation: A Priori Dynamic Team Composition

Current advanced Multi-Agent System architectures frequently adhere to a paradigm where team composition and agent roles are determined *a priori*, or before any execution commences. This often involves a human designer or a "manager" agent analyzing the task and then assembling a bespoke team with predefined goals and behaviors (Lhaksmana et al., 2018). This approach is highly effective and optimized for "tame" problems—those with clear objectives, defined initial states, and predictable environments. For such scenarios, traditional automated systems, reliant on fixed rules, can achieve robust performance (Ren et al., 2025). Early AI agents, for instance, operated with predefined decision trees, a strategy sufficient for structured and unchanging environments, albeit lacking the self-learning and adaptability seen in more recent agentic AI (Sapkota et al., 2025).

However, this *a priori* paradigm presents a significant limitation when confronted with "wicked problems" like the Abstraction and Reasoning Corpus. The inherent ill-definition, evolving requirements, and emergent properties of wicked problems mean that the optimal team structure, agent capabilities, or even the full scope of necessary actions often cannot be fully determined at the outset. Existing MAS engineering methodologies that require predefined goals and agent behaviors are not suited for self-organizing systems that need to adapt to dynamic changes (Lhaksmana et al., 2018). The scale and uncertainty of possible actions and coordination strategies in complex, dynamic environments can be too vast for agents to reason about directly *a priori*, necessitating adaptable and decentralized approaches (Hoang et al., 2017). Moreover, challenges remain in optimizing task allocation and managing complex, layered context information in multi-agent systems, particularly when confronted with unforeseen circumstances (Han et al., 2024).

While agentic AI systems are designed for goal-directed behavior, dynamic adaptation, and self-improvement (Hughes et al., 2025), their optimization often still requires labor-intensive, manual adjustments to refine roles, tasks, and interactions (Yüksel & Sawaf, 2024). This implies that despite the adaptive capabilities of the agents themselves, the overarching orchestration or initial setup of the multi-agent system may retain an element of static, *a priori* design. This limitation can hinder a system's ability to genuinely adapt and discover solutions in real-time when faced with the true novelty and ill-defined nature of problems like ARC.

2.4 The Theoretical Shift Towards Real-Time Adaptation

2.4.1 The Next Frontier: Real-Time Adaptation in Complex Systems

The limitations of *a priori* team composition in Multi-Agent Systems when faced with "wicked problems" necessitate a paradigm shift towards dynamic, real-time adaptation. This section introduces theoretical foundations from established cross-domain theories to support this shift.

Adaptive Management as a Strategic Response

Adaptive Management is a structured, iterative process for robust decision-making under conditions of uncertainty. Instead of viewing actions as fixed solutions, it treats them as experiments from which to learn and refine future strategies (Williams & Johnson, 2017). This framework is designed to simultaneously manage and learn about complex systems by promoting continuous learning-based decision-making, particularly when there is uncertainty about how a system will respond to interventions. It emphasizes iterative sequences of decision-making, monitoring, and assessment of system responses, integrating learned insights into subsequent actions (Williams, 2010). This contrasts with traditional management approaches that may not closely monitor implemented strategies or incorporate lessons learned effectively (Linkov et al., 2006). The core of adaptive management lies in its ability to adapt to evolving risks and changing environments, making it a powerful approach for situations with deep uncertainties where traditional "predict and act" decision-making is ineffective (Lee et al., 2018; Stanton & Roelich, 2021).

2.4.2 Theoretical Precedents for Adaptation

Organizational Learning and Team Adaptation. Connecting adaptive management to the principles of organizational learning, effective feedback mechanisms enable teams to continuously adapt their processes in response to environmental cues. Organizational learning involves the dynamic evolution and adaptation of teams as members interact over time and as situational demands unfold (Kozlowski & Ilgen, 2006). This process moves beyond classical theories of omniscient rationality, acknowledging that learning occurs even when faced with ambiguity about what happened, why it happened, and whether it was successful (March & Olsen, 1975).

Modern organizational structures, particularly those centered around teams, are increasingly designed for rapid, flexible, and adaptive responses to unexpected challenges (Kozlowski & Ilgen, 2006). The critical element is the ability to maintain technical competence and adapt to changes, often influenced by external pressures (Ployhart & Bliese, 2005). This is particularly relevant in AI systems, where self-initiated continual learning and adaptation to novel circumstances are crucial for agents to become more knowledgeable and self-sustainable over time (B. Liu et al., 2023). The idea of organizational learning also involves integrating various learning styles within functional specialties and aligning them with environmental demands (Kolb, 1976).

In the context of multi-agent AI systems, organizational learning facilitates collaboration and promotes group resilience, allowing agents to adapt to environmental perturbations (Abu et al., 2021). AI agents are evolving from passive tools to active collaborators in human-AI teams, requiring new interaction protocols and delegation strategies to optimize their adaptive capabilities (Lou et al., 2025). While LLM agents can simulate human behaviors and spontaneous collaborations, they often struggle with higher-order cognition for adaptive collaboration and lack mechanisms for inferring others' mental states, which are fundamental to effective knowledge sharing and coordination (Kostka & Chudziak, 2025). Therefore, advanced frameworks for AI multi-agent systems incorporate mechanisms such as dynamic task routing and bidirectional feedback to enable adaptiveness and iterative improvement (Xia et al., 2025).

Computational Precedents in MAS. The concept of real-time adaptation in MAS has strong computational precedents across various domains, particularly in robotics and logistics. In robotics, multi-agent systems are designed to be inherently adaptive, utilizing feedback loops to monitor performance and dynamically reallocate tasks in response to unforeseen events. For instance, distributed task allocation algorithms enable multi-robot teams to adjust their roles and responsibilities dynamically in response to changing conditions, balancing workload and performance (Ji et al., 2006; Jiang & Wang, 2019; Mina et al., 2020). This includes solutions for dynamic multi-robot task allocation problems with critical time constraints, where task information and robot states can change unpredictably (Chen, Zhang, et al., 2019).

Similarly, in logistics and supply chain management, MAS are employed for real-time adaptation to disruptions. Agent-based approaches allow for decentralized, reactive, and collaborative decision-making, enabling supply chain systems to react dynamically to changing environments (Gunasekaran & Ngai, 2004; Lau et al., 2007). This includes scenarios where immediate replanning is necessary due to disruptions, with multi-agent scheduling mechanisms adapting existing plans based on local changes and negotiations among agents (Tan et al., 2020). Such systems aim to achieve supply chain synchronization through adaptive control, ensuring minimal delays and interruptions by dynamically coordinating production schedules, inventory control, and delivery plans (Kegenbekov & Jackson, 2021). The overarching goal is to enable autonomous learning and decision-making without constant human intervention, leading to self-adaptive and autonomous supply chains (Xu et al., 2024). These examples underscore the practical feasibility and significant benefits of embedding real-time adaptation mechanisms within multi-agent systems.

2.5 A Socio-Cognitive Blueprint for a Culturally-Attuned Adaptive Team Manager

2.5.1 The Cognitive Engine of Adaptation: The OODA Loop

To address the complexities of "wicked problems" and enable real-time adaptation in Multi-Agent Systems, this section outlines a socio-cognitive blueprint for an adaptive team manager. This blueprint details how human-centric theories can provide the cognitive architecture for such a manager, explaining its ability to translate performance feedback into intelligent, culturally-informed team recomposition. The Observe-Orient-Decide-Act loop, developed by military strategist John Boyd, serves as a formal model for rapid decision-making in dynamic and uncertain environments (Lubitz & Wickramasinghe, 2006). This iterative four-stage process emphasizes continuous learning and adaptation:

- **Observe:** Gathering information from the environment.

- **Orient:** Analyzing and synthesizing observed information, shaping one's perspective, and forming hypotheses (X. Lu et al., 2022).
- **Decide:** Selecting a course of action based on the orientation.
- **Act:** Implementing the chosen action and observing its effects, which feeds back into the next "Observe" phase (X. Lu et al., 2022).

In the context of an adaptive AI team manager, this loop is operationalized as an adaptive cycle. Crucially, the "Orient" phase functions as a "Cultural Center of Gravity," signifying that an agent's interpretation of observations and subsequent decision-making is profoundly shaped by its cultural traditions and prior experiences (Colas et al., 2022; Hariri, 2025; Jin et al., 2023) . This means that for a multi-agent system composed of distinct cultural personas, their orientation to a given problem will be inherently different (Hariri, 2025).

Just as human cognitive processes differ across cultures, leading to varied interpretations and judgments (Jin et al., 2023), an AI agent imbued with cultural personas would "orient" itself differently based on its culturally-attuned worldview, influencing its understanding of problems and potential solutions (J. Yuan et al., 2024). This cultural influence on the "Orient" phase is critical, as military and strategic analyses of the OODA loop in multinational contexts identify this phase as the most susceptible to cultural impact (Boyd, 1986; Wendt, 2024). For instance, a culturally-attuned "OODA Reconnaissance" would prompt a large language model to not only observe objective features of a problem but also to orient to it by considering how different cognitive styles might approach it, eliciting a rich, structured understanding of the task's demands, including cultural considerations (Hariri, 2025). This highlights the need for AI systems to move beyond monocultural biases to effectively engage with nuanced cultural semantics and diverse worldviews (Saha et al., 2025). While traditionally conceptualized as a single decision cycle, this research will later demonstrate the framework's application at multiple levels of abstraction, forming the basis of a **Hierarchical OODA-Belbin Framework** that governs both high-level strategy and low-level tactical refinement.

2.5.2 A Diagnostic Framework for Team Function: Belbin's Team Roles

Belbin's Team Role theory offers a robust model for diagnosing functional deficits within a team. Belbin identified nine distinct team roles, each contributing a specific behavioral tendency to the team's overall effectiveness (Townend, 2007). An effective team ideally comprises a balance of these roles, ensuring that various functions, from idea generation to task completion, are covered (Omar et al., 2016). The theory posits that understanding individual team members' preferred roles can lead to improved communication, better working relationships, and enhanced team performance (Townend, 2007).

Incorporating cross-cultural research, it becomes clear that Belbin's roles are "culturally modulated" (A. M. R. Rodriguez et al., 2005). This means that while the underlying functions of the roles may be universal, their expression, preference, and even perceived importance can vary significantly across different cultural contexts (i Klett & Arnulf, 2020). Therefore, the selection of an AI agent for a specific role must be culturally informed to ensure genuine heterogeneity and effective team functioning. For instance, an agent embodying a "Shaper" role might express its drive differently depending on its cultural persona, impacting how it interacts with other agents and contributes to the team's dynamic. This cultural modulation of roles ensures that the AI team manager can dynamically recompose teams not just based on functional requirements, but also on the cultural compatibility and interpretive biases introduced by culturally-attuned personas.

2.5.3 The Cognitive Infrastructure for Coordinated Action

For a team, particularly one undergoing dynamic recomposition, to function smoothly and effectively, robust cognitive infrastructure is essential. Two key concepts underpin this: Shared Mental Models and Transactive Memory Systems.

Shared Mental Models: SMMs are "organized mental representations of the key elements within a team's relevant environment that are shared across team members" (Mohammed et al., 2010). These shared understandings enable teams to anticipate each other's needs and coordinate actions efficiently, even implicitly, especially under increased workload (Hoeft et al., 2005; Mohammed et al., 2010). The convergence of shared teamwork mental models has been shown to improve team performance (Singh et al., 2016; Ying & Wang, 2010). In multi-agent systems, SMMs can lead to mutual awareness, allowing team members to reason about their own situation, as well as the status and activities of teammates and the team's progress toward its goal, which can enhance teamwork behaviors (Kamali et al., 2006). Research also indicates that the degree of consensus in transactive memory systems, a component of shared cognition, is positively related to group performance (Robertson, 2008). For AI agents, applying the principles of SMMs would involve creating common representations of the problem space, task dependencies, and agent capabilities, allowing for more coherent collective action (Shao et al., 2017). When individual AI agents' mental models align, they are more effective (Andrews et al., 2022).

Transactive Memory Systems: TMS describe how knowledge is distributed among team members and how the team coordinates access to that knowledge (Wegner, 1995). A TMS operates through processes like "directory updating" (learning who knows what), "information allocation" (assigning memory items to members), and "retrieval coordination" (planning how to find information efficiently) (Wegner, 1995). This system allows group members to effectively query the knowledge contained in another's mind (M. Fisher et al., 2015), which is crucial for efficient knowledge exchange and to avoid redundant effort (Olabisi & Lewis, 2018; Yan et al., 2020). While traditionally studied in human teams, the principles of TMS are highly relevant for AI multi-agent systems, where agents need to be aware of each other's knowledge bases and processing capabilities to optimize collaboration (Hu et al., 2023a, 2023b). AI agents could benefit from episodic memory, which would allow them to store and retrieve records of what they do, similar to human cognition, leading to improved capabilities for agents interacting with the world (DeChant, 2025; Murphy et al., 2020). However, current AI systems often struggle with higher-order cognition required for adaptive collaboration and lack mechanisms like the ability to infer others' mental states, which are fundamental to effective knowledge sharing and coordination in human teams (Çelikok et al., 2019; Kostka & Chudziak, 2025; Oguntola et al., 2023). This includes the challenge of agents effectively modeling the mental states of others to achieve performance advantages in cooperation (L. Yuan et al., 2021). Therefore, incorporating SMMs and TMS principles provides the cognitive scaffolding necessary for AI agent teams to not only adapt to changing task requirements but also to maintain seamless coordination during real-time recomposition.

2.6 Synthesis and Defined Research Gap

The preceding review establishes a clear logical progression. The Abstraction and Reasoning Corpus (ARC) represents a class of "wicked problems" that defy traditional, static planning methods. A compelling theoretical pathway for tackling such problems is the development of multi-agent systems that mimic the cognitive synergy of **human multicultural teams**, operationalized through the use of culturally-attuned LLM personas. While this approach is supported by the well-documented "heterogeneity advantage," the contingency thesis introduces a critical counterpoint, suggesting that the benefits of diversity are moderated by task complexity and are not universally guaranteed (Higgs et al., 2005; Martins et al., 2012). State-of-the-art technical frameworks like Adaptive Branching Monte Carlo Tree Search provide powerful search capabilities (Inoue et al., 2025), while a socio-cognitively grounded adaptive manager, inspired by frameworks like OODA and Belbin, represents a significant step toward providing the necessary strategic intelligence (Omar et al., 2016; Townend, 2007).

However, despite these parallel advancements, a significant and more fundamental research gap remains. The literature has largely focused on designing progressively more sophisticated adaptive and heterogeneous systems, implicitly assuming that architectural complexity is the primary path to success. What is critically absent is a **systematic, empirical understanding of the performance trade-offs between fundamentally different agentic architectures when applied to wicked problems.** The question is not simply how to build a better adaptive team, but whether an adaptive team is always the right tool for the job. The relative costs and benefits of simplicity, homogeneity, diversity, and adaptation have not been rigorously compared in a controlled experimental setting.

Contribution. This paper aims to fill this critical gap. Our primary contribution is not the creation of a single, superior problem-solving framework, but the development of an **evidence-based contingency model derived from the direct comparison of six distinct agentic architectures.** By systematically testing these conditions against the ARC benchmark, we undertake a rigorous scientific process. We begin with an initial hypothesis that recurring problem archetypes—which we term "Direct Procedure," "Meticulous Execution," and "Cognitive Labyrinth"—map to specific architectural strategies.

However, the core contribution of this research lies in the discoveries that emerged from the deep, qualitative audit of the experimental results, which moved beyond the benchmark's standard 'perfect pixel match' metric. This analysis reveals a more nuanced reality, uncovering profound paradoxes such as a "cost of complexity" where the most complex adaptive systems underperform, and statistically insignificant differences in raw success rates that hide critical variations in solution quality. The paper's final, most significant contribution is the refinement of our initial problem taxonomy into a more precise, four-category model. This new model successfully explains the observed paradoxes and provides the definitive, evidence-based justification for a hierarchical, portfolio-based approach to advanced AI problem-solving. This research, therefore, transforms an initial set of hypotheses into a validated and nuanced theory of agentic performance.

3. Research Hypotheses

Based on the surprising empirical findings and abductive reasoning from our iterative pilot studies, we formulated a new contingency-based theory of AI problem-solving. The following hypotheses were derived from this new theory and were tested in the final, full-scale confirmatory experiment.

1. **The Simplicity Hypothesis:** For a significant subset of "Direct Procedure" problems, where the core logic is singular and the primary challenge is execution, a single agent (`Baseline_Single_Agent`) will be the most effective and efficient framework, as the cognitive and computational overhead of teamwork provides a net-negative value.

2. **The "Expert Anomaly" Hypothesis:** A non-diverse, homogeneous team of 'Implementer' agents

(Baseline_Static_Homogeneous) will outperform all diverse and adaptive frameworks on a specific class of "Meticulous Execution" problems that require high-fidelity procedural replication and where creative deviation is detrimental.

3. **The Adaptation Hypothesis:** The real-time adaptive framework (Adaptive_CP-ATS_Budget_Pool) will possess an exclusive capability to solve a unique class of "Cognitive Labyrinth" problems—tasks with ambiguous, multi-stage, or deeply relational logic that are intractable to any static framework.

4. Research Methodology

4.1. Research Philosophy: Design Science Research

This study is positioned within the **Design Science Research (DSR)** paradigm. DSR is a well-established research philosophy in information systems and related engineering fields that is fundamentally a problem-solving paradigm (Hevner et al., 2004; Peffers et al., 2007). It seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts, such as constructs, models, methods, and instantiations (Simon, 2019). Unlike behavioral science, which aims to develop and verify theories that *explain* or *predict* phenomena, the DSR paradigm asserts that knowledge and understanding of a problem domain and its solution are achieved through the iterative process of building and applying a designed artifact (Hevner et al., 2004; Simon, 2019) .

This philosophy is particularly apposite for the present research for several key reasons:

- **Problem-Centric Orientation:** The primary objective of this research is not merely to describe the limitations of current AI systems but to design, build, and evaluate a novel computational artifact—the **Hierarchical OODA-Belbin Framework** (which utilizes Cultural Persona-based Adaptive Team Search, or CP-ATS, as its core mechanism)—to address the identified “wicked problem” of abstract reasoning as presented by the ARC-AGI-2 benchmark. DSR is explicitly oriented toward solving such practical and relevant problems through the creation of innovative solutions (Peffers et al., 2007; Vaishnavi, 2007).

- **Artifact as the Core Contribution:** The central contribution of this work is the novel **Hierarchical OODA-Belbin Framework**, which operationalizes our contingency model. This artifact encompasses multiple agentic architectures, culminating in its strategic layer, the Strategic Selector Agent. In DSR, the artifact is the primary research output, and its utility and efficacy serve as the core validation of the research (Gregor & Hevner, 2013).

- **Iterative Design and Evaluation:** The DSR process is inherently iterative, involving a cycle of problem identification, solution design, demonstration, and evaluation (Peffers et al., 2007; Vaishnavi, 2007). This is precisely mirrored in the documented evolution of this study. The initial heuristic-based framework was designed and demonstrated in a pilot study; its subsequent failure was evaluated, which in turn motivated a return to the design phase to develop a more theoretically robust, adaptive artifact. This progression from a static planner to a real-time adaptive framework embodies the cyclical nature of DSR.

- **Integration of Relevance and Rigor:** A successful DSR project must satisfy two key principles: relevance and rigor (Hevner et al., 2004). The *relevance* of this research is established by addressing the significant and unsolved challenge of the ARC-AGI-2 benchmark. The *rigor* is ensured by grounding the artifact's design in established "kernel theories" from adjacent disciplines. The re-architecture of the `TeamManagerAgent`, for instance, was explicitly informed by socio-cognitive theories such as the OODA loop and Belbin's Team Roles, ensuring the created artifact is not an ad-hoc solution but a theory-informed innovation.

By adopting the DSR philosophy, this research frames the creation of a novel socio-technical system as a valid and rigorous contribution to scientific knowledge. The focus is placed squarely on the utility and effectiveness of the designed solution in its intended problem domain, thereby bridging the gap between theoretical understanding and practical application.

4.2 Research Approach: An Iterative Cycle of Discovery

Situated within the Design Science Research (DSR) paradigm established in the previous section, the logical pathway of this research inquiry did not follow a simple, linear trajectory . Instead, it embraced a dynamic and iterative cycle of reasoning, a process essential for tackling the novel and ill-structured nature of the "wicked problem" posed by the Abstraction and Reasoning Corpus (ARC). The research is characterized by a primary reliance on **abductive reasoning**, which then matures into a phase of rigorous **deductive validation**. Abduction, or inference to the best explanation, was employed in response to surprising empirical findings, allowing the research to generate novel theoretical models and artifacts. Deduction was then used to derive specific, falsifiable hypotheses from these new models and test them through controlled experimentation. This cyclical process of theory generation and validation is the intellectual engine that drove the development of the study's core contributions.

The evolution of the research approach itself serves as a powerful case study for the DSR paradigm in action. The journey from a simple heuristic planner to a real-time adaptive framework, and finally to a proposed strategic meta-agent, is not a story of correcting errors but of progressively deepening the understanding of the problem domain. Each "failure" was not a setback but a critical data point that invalidated a simpler theory and demanded the creation of a more sophisticated one. This progression demonstrates that in DSR, knowledge is built through the iterative creation and evaluation of artifacts in response to real-world problem-solving attempts, elevating the research from a simple report of results to a methodological contribution in its own right .

4.2.1. The Abductive Leap: From Empirical Anomaly to Theoretical Reframing

The research commenced with a deductive plan but was fundamentally reshaped by an abductive leap in response to a critical, unexpected experimental outcome. This pivot is central to the study's intellectual contribution, as it forced a reframing of the problem and the subsequent development of a more robust theoretical and computational artifact.

4.2.2 The Initial Deductive Stance and its Empirical Test

The research began with a clear, theory-driven hypothesis grounded in the literature on cognitive diversity and problem-solving styles. The initial artifact, a `TeamManagerAgent` for the Cultural Persona-based Adaptive Team Search (CP-ATS) framework, was designed to operate on a set of heuristics derived from Kirton's Adaption-Innovation (A-I) Theory. This initial design represented a deductive approach: a general theory (Kirton's A-I) was used to derive specific, testable rules for composing teams to solve ARC problems. These initial heuristics, detailed in Section 4.4.3, translated qualitative findings from the literature into a quantitative system for calculating an `innovation_score` based on machine-extractable features of each ARC problem. The first pilot study was designed as a straightforward deductive validation of this artifact's efficacy.

4.2.3. The Surprising Anomaly: A Catalyst for Abductive Inference

The initial 5-problem pilot study produced a "surprising empirical finding" that served as the catalyst for abduction. The proposed CP-ATS framework, designed to be more intelligent through its dynamic, heuristic-based team selection, was outperformed by simpler static teams on problem `00dbd492` (See Figure 3).

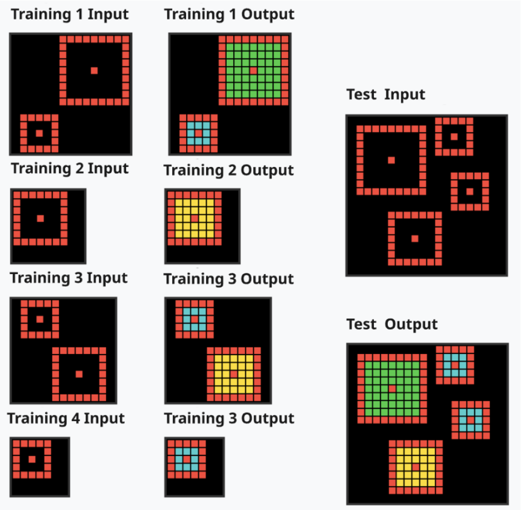

Figure 3

Problem 00dbd492 training sets, test and solution. The rule is to color the squares depending on their sizes (big ones in green, medium ones in yellow and small ones in blue).

This was a profound anomaly because it directly contradicted the primary hypothesis that a more tailored team would yield superior results. The failure was traced to the "a priori team selection problem," a risk that had been identified in the research plan. Analysis of the experimental logs revealed that the heuristic-based manager, in attempting to match the problem's perceived need for innovation, selected a team of novelty-generating personas but critically omitted the 'Implementer' persona (`Culture_Expert`) that was present in the successful static teams.

4.2.4. Inference to the Best Explanation: From Cognitive Style to Team Function

Faced with this anomaly, the research pivoted from a deductive to an abductive mode of reasoning. The guiding question became: "What is the most plausible explanation for the failure of the 'smarter' system and the success of the 'dumber' ones?" The abductive inference was that the initial theoretical lens—a single dimension of cognitive style (Adaption-Innovation)—was insufficient to capture the complexities of team function. The "best explanation" was that effective team composition is not just about matching a cognitive style to a problem, but about ensuring a balance of functional roles.

This led to a search for new "kernel theories" in adjacent disciplines, specifically organizational science and military strategy, to inform a more robust artifact design. This search yielded two critical frameworks: Belbin's Team Role theory, a model for diagnosing functional deficits within a team, and the OODA (Observe-Orient-Decide-Act) loop, a model for rapid decision-making in uncertain environments. The synthesis of these theories into a new "Culturally-Attuned OODA-Belbin Planner" was the direct outcome of this abductive leap. This new artifact represented a far more sophisticated theory of team management, one that accounted for both team balance and the cultural dynamics that modulate the expression of team roles.

4.2.5 The Second Abductive Cycle: The Insufficiency of A Priori Planning

An expanded pilot study testing this new, more sophisticated planner led to a second, even more critical failure. Despite using a superior planning model grounded in richer theory, all attempts across all conditions failed to achieve the 100% pixel-perfect score required by the official ARC metric. This universal failure triggered a second abductive cycle. The question now was: "What is the best explanation for why even a sophisticated, theory-grounded *a priori* planner is insufficient?"

The abductive inference was that the foundational vulnerability was the reliance on *any* purely *a priori* strategy. The generative nature of the LLM agents introduces a level of uncertainty and variability that no amount of upfront analysis can fully mitigate. The "best explanation" for the universal failure was the lack of a real-time feedback loop. This led to the creation of the study's central technical artifact: a real-time adaptive framework capable of in-mission learning and recomposition. This artifact was built upon a multi-objective fitness function that generates a diagnostic "failure signature" ($V_{fitness}$). This signature is then mapped to Belbin role deficits to guide dynamic team changes, transforming the Team Manager from a static planner into an adaptive problem-solver that learns and iterates within a single task instance. This artifact itself represents the product of the second abductive leap.

This research narrative demonstrates that abductive reasoning is not just a tool for discovery but a crucial risk mitigation strategy in DSR, especially when dealing with high-uncertainty technologies like LLMs. The initial research plan identified the "a priori team selection problem" as a potential risk. The first pilot study was designed to test for this risk early and on a small scale. When the risk materialized, the project did not fail. Instead, the abductive pivot allowed the research to incorporate the failure into a new, more robust theory and artifact. This proactive embrace of potential failure is a hallmark of mature scientific inquiry, where small, iterative pilot studies are deliberately used to surface "surprising facts" and trigger necessary abductive pivots before committing to computationally expensive, large-scale experiments .

4.2.6. The Deductive Turn: From a New Theory to Falsifiable Hypotheses

Following the abductive generation of the real-time adaptive framework, the research approach pivoted back to a deductive mode. The new, complex artifact was no longer just a solution to a problem; it embodied a comprehensive and testable theory about how to solve “wicked problems” with multi-agent AI systems. The subsequent large-scale experiments were therefore designed specifically to validate this new theory. As detailed in Section 5, this validation was structured to rigorously and thematically test the three deductively derived hypotheses that formed the core of the emergent contingency model.

4.2.7. A New Theory of Adaptive Problem-Solving

The final adaptive framework, with its diagnostic fitness function, Belbin-mapped error analysis, and dynamic recomposition logic, represents a new, prescriptive theory. This theory posits that for complex, ill-defined problems, an effective AI system must possess the ability to: (1) perform a deep, diagnostic analysis of its own performance failures, moving beyond binary success/failure; (2) map these diagnostic signatures to specific functional deficits in its problem-solving team; and (3) dynamically adapt its team composition in real-time to rectify these deficits.

4.2.8. The Experimental Design as Deductive Validation

The final, full-scale experiments on the complete 120-problem dataset were the instruments of this deductive validation phase. The analysis of these experiments focused explicitly on testing the above hypotheses. A focus on “Exclusive Solves” was a direct consequence of this deductive approach. The goal was not just to compare aggregate solve rates but to find definitive evidence for the specialized capabilities predicted by the new theory. The **scrutiny** of these exclusive solves for each predicted archetype—a process that involved both confirmation and, crucially, refutation—provided the definitive evidence needed to **construct and validate** the final, contingency-based view of problem-solving that emerged from the research.

4.2.9. Synthesis: A Cyclical Path to Knowledge Contribution

The research approach of this study is best understood as a virtuous cycle, moving between abductive inference and deductive validation. This iterative path was not a deviation from a plan but the core methodological process itself, enabling the generation of robust and novel contributions.

The study began with a simple deductive test of an existing theory. An empirical anomaly triggered an **abductive leap**, leading to the creation of a new, more powerful theoretical model and computational artifact. This new artifact was then subjected to rigorous **deductive validation**, which in turn produced more nuanced empirical findings, such as the "Expert Anomaly". These findings were then incorporated back into the artifact through refined heuristics and logic, such as the preemptive heuristic for the "Expert Anomaly" and the refined diagnostic rules for real-time adaptation.

The initial two cycles culminated in the **design and implementation of the Strategic Selector Agent**, which operationalized the contingency model derived from the 120-problem experiment. The development of this artifact, and the subsequent challenges it faced, led to the **conceptualization of a full Hierarchical OODA-Belbin Framework** to describe the system's multi-level decision-making process. This initial series of cycles provided a robust model for the majority of ARC problems. However, as the subsequent sections will detail, testing this model against the most difficult challenges in the official evaluation set revealed a final, more subtle 'local optima trap.' This triggered a **third and final abductive-deductive cycle**, leading to the most advanced methodological refinements and the ultimate test of the framework's limits. This complete, three-cycle journey demonstrates the power of the DSR paradigm, showing how the final, sophisticated conclusions were not assumed *a priori* but were earned through a structured and reflective process of building, testing, failing, and reframing.

4.3. Experimental Design

To execute the experimental research strategy, a formal design was established to ensure a clear, unambiguous, and replicable comparison between the different problem-solving frameworks. This design explicitly defines the independent, dependent, and control variables, providing the rigorous structure necessary to isolate the causal effects of each agent architecture on performance and efficiency.

4.3.1 Independent Variable

The core manipulation in this study is the **problem-solving framework**. This is a categorical variable with six distinct levels, each representing a unique theoretical approach to agent-based problem-solving:

1. `Baseline_Single_Agent`: A monolithic agent representing a non-collaborative approach, serving as a benchmark for the value of teamwork itself.

2. `Baseline_Static_Homogeneous`: A non-diverse team composed of identical 'Implementer' agents, designed to test the "Expert Anomaly" hypothesis.

3. `Baseline_Static_Heterogeneous`: A diverse team of agents assembled without any guiding composition logic, representing the baseline effect of unmanaged diversity.

4. `CP-ATS_Static`: A diverse team composed via the *a priori* heuristic-based planner, testing the efficacy of an initial, static strategic analysis.

5. `Adaptive_CP-ATS_Fixed_Cycle`: A real-time adaptive framework with a fixed number of adaptation cycles, representing a structured approach to in-mission learning.

6. `Adaptive_CP-ATS_Budget_Pool`: The most advanced adaptive framework, operating with a fluid total computational budget, designed to maximize the exploration of the team composition space.

4.3.2 Dependent Variables

The performance of each framework was operationalized and measured using two primary dependent variables:

1. **Task Success Rate**: The primary outcome measure, defined by the official Abstraction and Reasoning Corpus (ARC) benchmark as a strict binary score: a value of 1 is awarded for a 100% pixel-perfect match between the proposed solution and the ground-truth grid, and a value of 0 is awarded for any deviation. The analysis of this variable focuses on both aggregate solve rates across the full dataset and, more critically, the identification of "Exclusive Solves"—problems solved by only one framework, which serve to highlight specialized capabilities.

2. **Qualitative Solution Efficiency**: A secondary but crucial measure designed to assess the reasoning efficiency and resource cost associated with each framework's problem-solving strategy. Due to experimental limitations preventing the measurement of iterations to first success, efficiency is assessed qualitatively through a comparative analysis of the code generated for 'real successes'. Solutions are evaluated based on criteria such as **parsimony** (directness and conciseness), **algorithmic purity**, and **robustness**, as detailed in Section 4.6.2. This qualitative variable allows for a comparison of the reasoning overhead introduced by more complex strategies versus specialist approaches.

4.3.3 Control Variables

To ensure the internal validity of the experiment, several key factors were held constant across all six experimental conditions. This control allows for the confident attribution of any observed differences in the dependent variables to the manipulation of the independent variable. The control variables were:

- **The Problem Set**: All frameworks were tested on the identical set of problems from the standardized ARC-AGI-2 dataset to ensure a fair and consistent challenge.

- **The Computational Budget**: A total computational budget (e.g., 250 MCTS iterations for the main experiment) was enforced for each problem instance across all conditions, ensuring that no single framework had an unfair advantage in terms of processing resources.

- **The Underlying Large Language Model (LLM)**: The same foundational LLM, Deepseek-coder-v2-instruct, was used for all agents in all conditions to ensure that performance differences were attributable to the strategic framework and team composition, not to variations in the core generative capabilities of the agents.

4.4. Heuristic-Driven Team Composition

To facilitate the dynamic team composition capabilities of the Cultural Persona-based Adaptive Team Search (CP-ATS) framework, a `TeamManagerAgent` governed by a set of heuristics was designed. The development of these heuristics was not arbitrary; rather, it was the result of a formal, multi-phase research protocol intended to ensure academic rigor, transparency, and replicability. This section provides a detailed account of this protocol.

4.4.1 Phase 1: Systematic Literature Search

The initial phase consisted of a systematic literature search to identify established academic frameworks that link problem typologies to the cognitive styles required for their effective resolution.

- **Search Protocol:** The search was conducted across primary academic databases, including Google Scholar, ACM Digital Library, IEEE Xplore, PsycINFO, and arXiv. A structured set of keywords was utilized, organized into three conceptual groups: (A) Problem Typology (e.g., ‘"ill-structured problem"‘), (B) Cognitive and Team Style (e.g., ‘"cognitive diversity"‘, ‘"Kirton Adaption-Innovation"‘), and (C) Task Domain (e.g., ‘"abstract reasoning"‘).

- **Outcome:** The search identified a robust body of literature. The subsequent analysis was anchored by a core set of six foundational and meta-analytic works, selected for their direct contribution to the methodology's theoretical underpinnings. These key sources include (Kirton, 1976), which provides the seminal theory of cognitive style (Adaption-Innovation); (Jonassen, 2000), which offers a comprehensive problem typology suitable for classifying ARC tasks; (Horwitz & Horwitz, 2007a), whose meta-analytic review confirms the value of managed cognitive diversity; (Rittel & Webber, 1973), who define the "wicked problems" that represent the upper bound of ARC task difficulty; (Somech & Drach-Zahavy, 2013), who link team composition to the implementation of creative solutions; and (Armstrong, 2000), which validates the application of cognitive styles in complex, technical domains.

4.4.2 Phase 2: Framework Selection

Following a synthesis of the literature, **Kirton's Adaption-Innovation (A-I) Theory** (Kirton, 1976) was formally selected as the primary theoretical framework for the design of the heuristics.

- **Justification:** This framework was determined to be the most suitable for three principal reasons:

  1. **Direct Mapping:** The theory provides a direct and powerful mapping to the five creativity dimensions used in the cultural personas (Novelty, Usefulness/Feasibility, Flexibility, Elaboration and Cultural Appropriateness/Sensitivity) (Hariri, 2025). The **Innovator** style corresponds to a preference for **Novelty** and **Flexibility**, while the **Adaptor** style corresponds to a preference for **Usefulness/Feasibility** and **Elaboration** (Kirton, 1976).
  2. **Empirical Validity:** It is a well-established and empirically validated framework for cognitive style, supported by several decades of research.
  3. **Alignment with Research Goals:** The theory's focus on problem-solving *style*—as opposed to level of competence—is perfectly aligned with the CP-ATS objective of composing a team of diverse specialists.

4.4.3 Phase 3: Formalization of Heuristics

The final phase translated the selected academic framework into a quantifiable and executable algorithm for the `TeamManagerAgent`. To ensure the generated targets were mathematically sound and achievable by the available personas, a **Profile Blending** methodology was developed.

Archetype Definition. Two "archetype" profiles were defined based on the extreme styles of Kirton's A-I Theory, utilizing the empirical scores from the available cultural personas (Hariri, 2025).

- **The "Pure Innovator" Archetype ($\mathbf{P_{in}}$):** The profile vector of the persona exhibiting the highest scores in Novelty and Flexibility (`Culture_5`), with scores of {Novelty: 3.53, Usefulness: 4.26, Flexibility: 4.09, Elaboration: 4.23, Sensitivity: 4.55}.
- **The "Pure Adaptor" Archetype ($\mathbf{P_{ad}}$):** The profile vector of the persona exhibiting the highest scores in Usefulness and Elaboration (`Culture_Expert`), with scores of {Novelty: 3.37, Usefulness: 4.38, Flexibility: 4.00, Elaboration: 4.53, Sensitivity: 4.48}.

Problem Classification and Profile Generation. A central component of our methodology is the classification of ARC problems into distinct archetypes to facilitate a contingency-based analysis. Based on an initial thematic analysis of the ARC training set, we hypothesized that the problems could be broadly categorized into three core archetypes, which formed the basis for the Strategic Selector Agent's initial decision-making model:

- **Direct Procedure:** Tasks that can be solved by applying a clear, deterministic, step-by-step rule. The logic is explicit in the examples and requires no deep abstraction or interpretation of ambiguous patterns.
- **Meticulous Execution:** Tasks where the core challenge lies in the precise, often complex, implementation of a known goal. This includes tasks with intricate geometric patterns, ambiguous edge cases, or rules that depend on a careful analysis of local pixel environments.
- **Cognitive Labyrinth:** The most difficult class of problems, requiring the synthesis of a non-trivial, multi-step, abstract algorithm. Success often depends on a key "eureka" insight and the correct sequencing of several distinct logical steps (e.g., filter -> sort -> reconstruct).

Algorithmic Implementation of Problem Classification. To operationalize this theoretical classification, the Team Manager executes the following algorithm for each ARC-AGI problem:

1. **Feature Extraction:** A set of machine-extractable features is calculated from the problem specification.
2. **Innovation Score Calculation:** An `innovation_score` is initialized to 0.0. The heuristic rules, derived from A-I theory, modify this score. The final score is clamped to a range of $[-1.0$ (purely Adaptive) to $+$ 1.0 (purely Innovative)$]$.
3. **Profile Blending:** The `innovation_score` is utilized to calculate a weighted average of the two archetype profiles ($\mathbf{P_{in}}$ and $\mathbf{P_{ad}}$), producing a final, continuous $\mathbf{P_{target}}$ vector that is mathematically guaranteed to be within the achievable range of the personas.

Calibrated Heuristic Weighting. The quantitative rules used to modify the `innovation_score` were not derived from a direct formula but were established through a process of **Calibrated Heuristic Weighting**. This process translates the qualitative findings from the literature into a rank-ordered, quantitative system, wherein the weights represent the relative theoretical impact of each feature on shifting a problem from an "Adaptive" to an "Innovative" type.

- **Strongest Signals:** The largest weights were assigned to the most unambiguous indicators of problem type. A **Symmetry-Based** problem (`Rule C3: -0.6`) is considered the clearest signal of a well-structured task requiring an Adaptive approach. Conversely, a **Single Training Example** (`Rule C1: +0.5`) creates the most ambiguity, demanding an Innovative approach, as described by (Jonassen, 2000) for ill-structured problems.
- **Weaker Signals:** Features that are less definitive receive smaller weights. For example, **Significant Grid Growth** (`Rule C4: +0.2`) could result from either innovative or adaptive processes; thus, its impact on the score is moderated.

This calibrated system ensures that the most theoretically significant problem features exert the greatest mathematical influence on the final team composition.

Heuristic rules details. The `innovation_score` is calculated based on the following set of Composition rules (henceforth labeled with a 'C' prefix):

- **C1: Single Training Example**
  - *Impact:* Strongly Innovative
  - *Justification:* High ambiguity requires a paradigm-shifting solution.
  - *Rule:* `innovation_score += 0.5`
- **C2: New Colors Introduced**
  - *Impact:* Strongly Innovative
  - *Justification:* Requires generating entirely new elements.
  - *Rule:* `innovation_score += 0.4`
- **C3: Symmetry-Based**
  - *Impact:* Strongly Adaptive
  - *Justification:* A well-defined, procedural task within an existing paradigm.
  - *Rule:* `innovation_score -= 0.6`
- **C4: Significant Grid Growth**
  - *Impact:* Leans Innovative
  - *Justification:* Requires creating new structures.
  - *Rule:* `innovation_score += 0.2`
- **C5: Object Disassembly**
  - *Impact:* Leans Innovative
  - *Justification:* The solution breaks down existing structures.
  - *Rule:* `innovation_score += 0.3`
- **C6: Stable Object/Color Count**
  - *Impact:* Leans Adaptive
  - *Justification:* Focuses on rearranging existing elements.
  - *Rule:* `innovation_score -= 0.3`

Optimal Team Selection. The final step consists of selecting a real-world team of three agents that best matches the generated $\mathbf{P_{target}}$. This task is framed as an optimization problem, a formulation consistent with research suggesting that the benefits of cognitive diversity must be actively managed to be realized (Horwitz & Horwitz, 2007a).

1. **Define Aggregated Team Profile:** The profile of any given team ($\mathbf{P_{team}}$) is defined as the vector average of the profiles of its three member personas.

2. **Define Distance Metric:** The proximity of a team's profile to the target profile is quantified using **Euclidean Distance**.

3. **Combinatorial Search:** The agent performs a brute-force search of all unique three-person teams possible from the pool of available personas. For each combination, the Euclidean distance between its $\mathbf{P}_{\mathbf{team}}$ and the $\mathbf{P}_{\mathbf{target}}$ is calculated.

4. **Final Selection:** The team yielding the **minimum Euclidean distance** is selected as the optimal team for the given problem. This method is deterministic and guaranteed to find the best possible match.

4.4.4. Preliminary Experimentation and Rationale for the Full Study

To validate our experimental protocol and gain initial insights into our core hypotheses, we conducted a staged pilot study before committing to the full 120-problem ARC-AGI-2 benchmark. This preliminary phase was designed to de-risk the main experiment by testing the efficacy of the AB-MCTS search framework and gathering an early signal on the performance of the proposed dynamic team composition model.

4.4.5. Initial 5-Problem Pilot Study

The first research cycle, summarized in Figure 4, began with an initial pilot study involving four experimental conditions (Single Agent, Static Homogeneous, Static Heterogeneous, and the proposed CP-ATS) on a small, randomly selected subset of five problems from the training dataset.

The objectives of this initial run were twofold:

1. To confirm that a structured search framework (AB-MCTS) provides a tangible benefit over unguided exploration (Repeated Sampling).

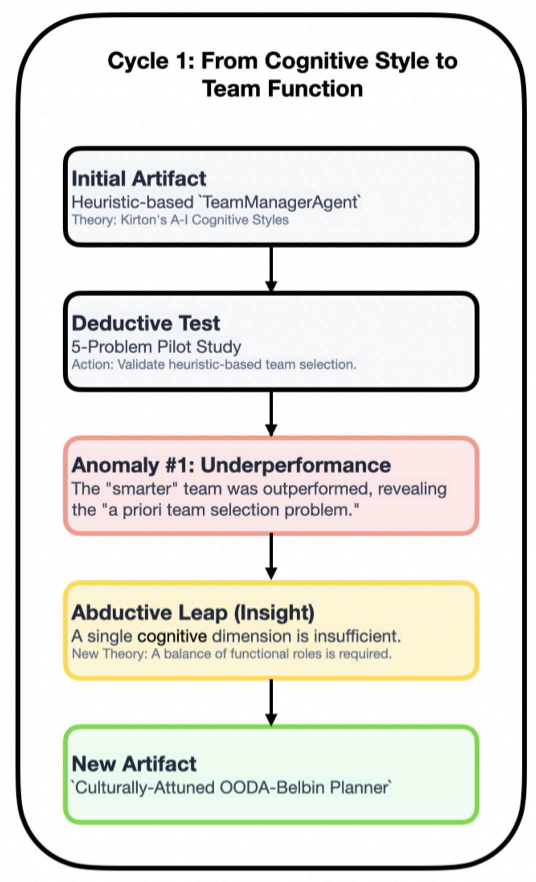
Cycle 1: From Cognitive Style to Team Function
Initial Artifact
Heuristic-based `TeamManagerAgent`
Theory: Kirton's A-I Cognitive Styles
Deductive Test
5-Problem Pilot Study
Action: Validate heuristic-based team selection.
Anomaly #1: Underperformance
The "smarter" team was outperformed, revealing the "a priori team selection problem."
Abductive Leap (Insight)
A single cognitive dimension is insufficient.
New Theory: A balance of functional roles is required.
New Artifact
`Culturally-Attuned OODA-Belbin Planner`

Figure 4

Research Cycle 1 – From Cognitive Style to Team Function. *This flowchart illustrates the first DSR cycle, beginning with an initial artifact based on cognitive style theory. The key "anomaly" was the framework's underperformance in a pilot study, which triggered an "abductive leap": the insight that balancing functional team roles was more critical than matching a one-dimensional cognitive profile. This directly led to the development of the more sophisticated* `OODA-Belbin Planner`*.*

2. To obtain a preliminary indication of the performance of the dynamic CP-ATS framework against its static baselines.

The results of this 5-problem experiment were highly informative. The search-based frameworks (all three AB-MCTS conditions) solved 60-80% of the problems, whereas the Single Agent baseline solved only 40%. This provided initial validation for the use of the

AB-MCTS architecture.

More critically, however, the results provided a direct counter-example to our primary hypothesis. The proposed CP-ATS framework, with its dynamic team selection, solved only 3 of the 5 problems (60%). In contrast, both the Static Homogeneous and Static Heterogeneous teams successfully solved 4 of the 5 problems (80%). The point of divergence occurred on problem `00dbd492` (See Figure 3), which was solved by both static teams but failed by the CP-ATS system.

This failure appears to be a direct manifestation of the “a priori team selection problem,” identified as the single greatest potential point of failure in our research plan. Analysis of the experimental logs revealed that the heuristic-based Team Manager for CP-ATS selected a team heavily skewed towards novelty-generating personas, omitting the crucial “implementer” persona (`Culture_Expert`) that was present in the successful static teams.

4.4.6. Rationale for a New Planner and Confirmatory Test

Given the counter-intuitive nature of these initial findings—specifically, the underperformance of our proposed method due to a foreseen but unconfirmed risk—we determined that a direct progression to the full 120-problem experiment would be premature. A key question arose: was the observed failure of the CP-ATS framework an anomaly specific to the small 5-problem sample, or was it indicative of a more fundamental flaw in the heuristic-based Team Manager?

Simply expanding the pilot study to a larger set of problems would be an inefficient use of computational resources if the Team Manager’s logic was indeed flawed. The more rigorous scientific path was to first address the inferred root cause of the failure directly. Therefore, we concluded that the initial, simple heuristic planner was insufficient.

To address this, we pivoted to redesigning the core of the framework: the Team Manager agent itself. This led to the development of the "Culturally-Attuned OODA-Belbin Planner," a more sophisticated artifact grounded in richer organizational and strategic theories. The objective of the next experimental step was thus refined: to conduct a direct, confirmatory test of this new planner against the original set of 5 problems. This approach provides the clearest and most direct evidence as to whether our re-architecture successfully mitigated the identified failure mode, thereby providing a more solid empirical justification before undertaking the final, computationally expensive experiment.

4.4.7. Resolving the Team Manager’s Critical Failure: A Culturally-Attuned OODA-Belbin Planner

The failure of the initial CP-ATS framework in the **5-problem pilot study** to outperform both, the Baseline_Static_Heterogeneous and the Baseline_Static_Homogeneous conditions, provided an unequivocal directive: the simple, heuristic-based Team Manager is a critical point of failure. The manager’s inability to form a balanced team for problem ‘00dbd492‘—selecting a team of innovators (‘Plants‘) where an ‘Implementer‘ was required—highlighted the naivety of a one-dimensional approach to team composition.

To resolve this, we conducted a comprehensive review of academic and professional literature on team formation, extending our search beyond multi-agent systems to organizational science and military strategy. A key constraint of our search was to find frameworks that could be adapted to the explicitly *multicultural* nature of our CP-ATS agents. While no single, pre-existing framework for an "OODA-Belbin Planner for multicultural teams" was found, our research revealed that principles from Belbin’s Team Role theory and the OODA loop have been separately studied in cross-cultural contexts, providing a strong theoretical foundation for a novel, integrated approach.

Theoretical Underpinnings from Cross-Cultural Research. Our literature review yielded two critical insights that directly inform our revised methodology:

1. **Belbin's Roles are Culturally Modulated, Not Invalidated:** Research by (Aritzeta et al., 2007; D. V. Rodriguez et al., 2024) confirms that the nine Belbin Team Roles are recognizable archetypes of behavior that manifest across different national and organizational cultures. However, the *expression* and *valuation* of these roles are culturally dependent. For instance, a culture high in individualism may value the assertive 'Shaper' role, while a collectivist culture may prioritize the 'Teamworker' role. This implies that a multicultural team's effectiveness hinges not on seeking cultural homogeneity, but on consciously balancing these archetypal roles while being aware of their varied cultural expressions.

2. **The OODA Loop's "Orient" Phase is a Cultural Center of Gravity:** Military and strategic analyses of the OODA loop in multinational or coalition contexts identify the 'Orient' phase as the most susceptible to cultural influence (Boyd, 1986; Wendt, 2024). How an actor or agent orients themselves—interpreting observations based on their genetic heritage, cultural traditions, and prior experiences—is the fulcrum of decision-making. For a multi-agent system composed of distinct 'Cultural Personas', this means that their orientation to a given ARC problem will be inherently different. A successful Team Manager cannot ignore this; it must leverage it.

The Proposed Solution: A Two-Stage, Culturally-Informed Planner. Based on these findings, we re-architected the Team Manager with a two-stage planning process that explicitly accounts for both team balance and cultural dynamics.

Stage 1: Culturally-Attuned "OODA Reconnaissance". The Team Manager's first step is to perform a deep analysis of the ARC problem. It now prompts the base LLM to not only observe the problem's objective features but to *orient* to it by considering how different cognitive styles might approach it. This prompt is designed to elicit a rich, structured understanding of the task's demands.

```
% Conceptual OODA Reconnaissance
Prompt
[SYSTEM] You are a master analyst
for the ARC-AGI benchmark...
...
"orientation": {
    "problem_type":
    "{Classify as 'Constructive',
    'Destructive', etc.}",
    "core_challenge":
    "{Describe the central
    difficulty...}",
    "required_cognitive_style":
    { ... },
    "cultural_considerations":
    {
      "collectivist_vs_
      individualist_approach":

      "{Describe potential
      approaches}",
      "harmony_vs_
      confrontation_approach":

      "{Describe
      potential approaches}"
    }
}
```

It is important to clarify the relationship between the 'OODA Reconnaissance' prompt presented here and its implementation in the system's code. The prompt should be understood as a *conceptual model* that outlines the strategic analysis the Team Manager is designed to perform—classifying the problem's nature to determine the required cognitive style.

For reasons of computational efficiency and reliability, this analysis was *operationalized* in the `team_manager.py` module through a set of deterministic, machine-extractable heuristics. For instance, abstract concepts from the prompt, such as classifying a problem as 'Constructive' versus 'Destructive', are proxied in the code by the `_extract_features` function, which checks for concrete properties like symmetry (`symmetry_based`) or the introduction of new colors (`new_colors_introduced`).

The output of these heuristics is an `innovation_score`, which serves the same functional purpose as the conceptual prompt's analysis: it guides the initial selection of an 'Innovator' or 'Adaptor' oriented team, as detailed in the `select_initial_team` method. This heuristic approach provides a reliable and reproducible proxy for the deeper qualitative analysis envisioned in the OODA-Belbin framework.

Stage 2: Culturally-Informed Belbin Team Selection. The manager then selects a team using the Belbin framework as its foundation, but with rules informed by the cultural context from Stage 1. The core logic is to ensure team balance (e.g., pairing a 'Plant' with an 'Implementer') while using cultural information to select the most appropriate persona for that role.

The selection process is governed by a refined set of rules:

- **Rule 1 (The Anti-Failure Clause):** To prevent the failure mode observed in the pilot, if the 'OODA Reconnaissance' identifies the problem as requiring high novelty ('Plant' behavior), the team **must** also include an 'Implementer' ('Culture_Expert'). This ensures that creative ideas can be practically executed.

- **Rule 2 (Culturally-Informed Role Selection):** The manager selects the specific agent for a required Belbin role based on the problem's cultural orientation. For example:
  - If a 'Teamworker' role is needed for a task requiring robust group consensus, the manager will select the 'Culture_4' (collectivist) persona.
  - If a 'Teamworker' is needed for a task requiring subtle social navigation to avoid conflict, the manager will select the 'Culture_11' (harmony-focused) persona.

- **Rule 3 (Default Balanced Team):** In ambiguous cases, the manager defaults to a balanced team of a 'Plant' ('Culture_5'), an 'Implementer' ('Culture_Expert'), and a 'Teamworker' ('Culture_4'), ensuring a baseline of creativity, practicality, and cooperation.

This new methodology for the Team Manager is a direct response to the empirical results of our pilot study. It is academically rigorous, grounding our agent design in established, cross-domain theories and adapting them to the unique, multicultural nature of our CP-ATS framework. We hypothesize that this theory-grounded manager will be significantly more robust and effective than the initial heuristic, and this will be tested in the pilot study.

Adaptive CP-ATS Experimental Conditions. This critical failure motivated the development of two new, fully adaptive experimental frameworks designed to overcome the limitations of a purely *a priori* strategy. These frameworks, implemented in `main.py`, transform the Team Manager from a static planner into a real-time problem-solver:

- **Adaptive_CP-ATS_Fixed_Cycle:** An adaptive framework where the system can recompose the team a fixed number of times (`MAX_ADAPTATION_CYCLES` that was set to 5 cycles), with each cycle receiving a discrete computational budget (`BUDGET_PER_CYCLE` set to 50 attempts).

- **Adaptive_CP-ATS_Budget_Pool:** The most advanced framework, which operates with a total fluid computational budget. This allows the system to dynamically allocate resources, enabling more adaptation cycles if it detects promising pathways, thereby maximizing the exploration of the team composition space.

These new adaptive conditions, along with the static baselines, form the basis for the subsequent experiments designed to test the efficacy of real-time, diagnostic-driven team recomposition.

4.4.8. Dynamic Performance Evaluation and Real-Time Adaptation

The initial team selection methodology, while grounded in established theory, operates on an *a priori* analysis of the problem's static features. This approach is vulnerable to misclassification, as demonstrated in the pilot study, where an incorrect initial assessment led to a suboptimal team composition and subsequent task failure. To mitigate this critical vulnerability and enhance the adaptability of the CP-ATS framework, we introduce a real-time performance feedback loop. This loop is enabled by a multi-objective computational fitness function designed to provide a nuanced, diagnostic evaluation of a proposed solution's quality.

This fitness function moves beyond the insufficient binary metric of correct/incorrect, which provides no actionable information for adaptation on failed attempts. Instead, it deconstructs solution quality into a hierarchical set of metrics that generate a high-dimensional "failure signature." This signature is fed back into the Team Manager's OODA loop, augmenting the **Observe** phase with the direct results of the team's actions. This allows the **Orient** phase to perform a diagnostic analysis of the failure, leading to a more informed **Decide** phase that can dynamically recompose the team to address specific, identified weaknesses before the next **Act** cycle. (Belbin & Brown, 2012; S. Fisher et al., 2001) This transforms the Team Manager from a static planner into an adaptive problem-solver capable of learning and iterating within a single task instance.

A Hierarchical Framework for Solution Quality. The fitness function is structured as a four-level hierarchy, with each level assessing the solution against core knowledge priors inherent to ARC tasks, such as objectness, goal-directedness, and basic geometry.

Level 1: Pixel and Colorimetric Fidelity (The Canvas). This level assesses the raw, rendered output grid (G_{prop}) against the ground-truth grid (G_{truth}).

- **Pixel-wise Accuracy ($f_{\textbf{pixel\_accuracy}}$):** A direct, cell-by-cell comparison, as used in the official Kaggle ARC Prize evaluation (De Dreu & Weingart, 2003). Let G_{prop} and G_{truth} be two grids of dimensions $H \times W$. The pixel-wise accuracy is:

$$f_{\text{pixel\_accuracy}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \delta(G_{prop}(i,j), G_{truth}(i,j)) \tag{1}$$

where $\delta(a,b) = 1$ if $a = b$ and 0 otherwise.

- **Structural Similarity Index Measure for Discrete Colormaps ($f_{\textbf{ssim\_discrete}}$):** Adapted from the standard SSIM (Z. Wang et al., 2004), this metric assesses local structural similarity, treating ARC's discrete color values (0-9) as luminance levels. The dynamic range L is 9. For a local window x from G_{prop} and y from G_{truth}, the formula is:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2}$$

where μ is the mean, σ^2 is the variance, σ_{xy} is the covariance, and C_1, C_2 are stabilizing constants derived from L

(Armstrong, 2000; Partington & Harris, 1999).

- **Color Histogram Similarity ($f_{\textbf{hist\_bc}}$):** This metric compares the color distributions of the two grids, ignoring spatial arrangement, using the Bhattacharyya coefficient on the normalized color histograms ($H_{prop\_norm}, H_{truth\_norm}$).

$$f_{\text{hist\_bc}} = \sum_{k=0}^{9} \sqrt{H_{prop\_norm}(k) \cdot H_{truth\_norm}(k)} \tag{3}$$

Level 2: Object-Centric Structural Integrity (The Components). This level moves from pixels to discrete objects, assessing whether the solution correctly represents the constituent components of the ground truth. Objects are first identified using a Connected-Component Labeling (CCL) algorithm (Rosenfeld & Pfaltz, 1966).

- **Object Count Fitness ($f_{\textbf{obj\_count}}$):** Measures the accuracy of the number of objects produced. Let O_{prop} and O_{truth} be the sets of objects in the proposed and ground-truth grids, respectively.

$$f_{\text{obj\_count}} = \max\left(0, 1 - \frac{||O_{prop}| - |O_{truth}||}{|O_{truth}|}\right) \tag{4}$$

- **Jaccard Index for Object Sets ($f_{\textbf{jaccard\_objects}}$):** Also known as Intersection over Union (IoU), this metric measures the similarity between the two sets of objects after establishing a matching between them (JACCARD, 1901).

$$f_{\text{jaccard\_objects}} = \frac{|O_{prop} \cap O_{truth}|}{|O_{prop} \cup O_{truth}|} \tag{5}$$

Level 3: Topological and Symmetrical Congruence (The Blueprint). This level assesses global geometric and topological properties, which are central to many ARC tasks.

- **Symmetry Congruence ($f_{\textbf{symmetry}}$):** Algorithms for detecting reflectional and rotational symmetry (Marola, 1989) are applied to both grids. The fitness score is a function of both the matching of the primary symmetry type (e.g., 90-degree rotational) and the degree of similarity in the symmetry itself.

- **Euler Characteristic ($f_{\textbf{euler}}$):** A topological invariant (Abbena et al., 2017), the Euler characteristic (χ) for a 2D binary image is calculated as:

$$\chi = (\text{number of connected components}) - (\text{number of holes}) \tag{6}$$

  This is computed for each color channel and compared between grids to detect topological errors, such as incorrectly filled areas. The calculation can be performed efficiently using libraries such as `scikit-image` (Walt et al., 2014).

Level 4: Graph-Based Relational Similarity (The Abstract Logic). The highest level of abstraction, this assesses the relational logic of the solution by converting grids into attributed graphs and comparing their structure.

- **Grid-to-Graph Conversion:** Each object from Level 2 becomes a node, with attributes for color, size, etc. Edges represent relationships like adjacency, relative position, or containment (JACCARD, 1901).

- **Graph Similarity ($f_{\textbf{graph\_spectral}}$):** While Graph Edit Distance (GED) is a comprehensive but computationally expensive metric (Sanfeliu & Fu, 1983), we utilize a more efficient spectral method. This involves comparing the eigenvalues (spectrum) of the graph Laplacian matrices of G_{prop} and G_{truth} (Von Cybernetics, 2007). The spectral distance d_{spectral} is the Euclidean distance between the sorted eigenvalue vectors, which is then normalized into a fitness score:

$$d_{\text{spectral}} = \sqrt{\sum_{i=1}^{|V|}(\lambda_i(L_{prop}) - \lambda_i(L_{truth}))^2} \implies f_{\text{graph\_spectral}} = e^{-\beta \cdot d_{\text{spectral}}} \tag{7}$$

A Unified Fitness Score via Dynamically Weighted Summation. The component metrics (f_i), each normalized to a range of $[0,1]$, are combined into a single fitness score, F, via a weighted sum.(Sanfeliu & Fu, 1983; Walt et al., 2014)

$$F(G_{\text{prop}}, G_{\text{truth}}) = \frac{\sum_i w_i \cdot f_i(G_{\text{prop}}, G_{\text{truth}})}{\sum_i w_i} \tag{8}$$

Crucially, the weights (w_i) are not static. They are determined dynamically for each task based on the output of the "Culturally-Attuned OODA Reconnaissance" phase. For instance, if the analysis identifies a task as symmetry-based (Rule C3), the weight w_{symmetry} is significantly increased, focusing the evaluation on the most salient feature of the problem. This transforms the fitness function into an active component of the problem-solving strategy, aligning the evaluation with the hypothesized problem constraints.

Adaptive Strategy via Diagnostic Error-to-Role Mapping. The unweighted vector of component scores,

$$V_{fitness} = [f_{\text{pixel\_accuracy}}, \ldots, f_{\text{graph\_spectral}}] \tag{9}$$

, serves as a diagnostic "failure signature" that is fed back into the Team Manager's OODA loop. In the **Orient** phase, this signature is mapped to a specific failure mode, which is then attributed to a functional deficit as described by Belbin's Team Role theory (Belbin & Brown, 2012). This mapping provides a prescriptive basis for team recomposition in the **Decide** phase.

Diagnostic Error-to-Belbin Role Mapping for Adaptive Team Recomposition.

- **Ideation Failure**
  - *Fitness Vector Signature:* Low scores across all levels.
  - *Inferred Deficit:* **'Plant'** (Creativity).
  - *Prescriptive Action:* Strongly increase weight for the 'Plant' role in the target team profile. Prioritize new persona with high novelty.
- **Compositional Error**
  - *Fitness Vector Signature:* Low Level 2 (Object) scores, potentially high Level 1 (Color) scores.
  - *Inferred Deficit:* **'Implementer'** (Practicality).
  - *Prescriptive Action:* Strongly increase weight for the 'Implementer' role. Enforce inclusion of 'Culture_Expert' (The Anti-Failure Clause).
- **Arrangement Error**
  - *Fitness Vector Signature:* High Level 2 scores, low Level 3 (Topology/Symmetry) & 4 (Graph) scores.
  - *Inferred Deficit:* **'Monitor Evaluator'** (Strategy).
  - *Prescriptive Action:* Increase weight for the 'Monitor Evaluator' role. Select persona excelling at logical analysis.
- **Precision Error**

- *Fitness Vector Signature:* High scores on Levels 2, 3, 4; low scores on Level 1 (Pixel Accuracy).
- *Inferred Deficit:* **'Completer Finisher'** (Perfectionism).
- *Prescriptive Action:* Strongly increase weight for the 'Completer Finisher' role. Select a meticulous, error-checking persona.

This mechanism closes the feedback loop, allowing the system to learn not only about the problem but about the optimal team composition required to solve it, thereby directly addressing the critical failure mode identified in the pilot study.

4.4.9. Analysis of the Expanded Pilot Study

Table 1

Summary of Top-Performing Conditions in the Second Pilot Study

Problem ID	**Top-Performing Condition(s) (Highest Score)**
576224	CP-ATS (0.9783)
007bbfb7	Single Agent (0.8256), Static Heterogeneous (0.8256)
009d5c81	CP-ATS (0.9762)
00d62c1b	CP-ATS (0.9762)
00dbd492	Single Agent (0.9667), Static Homogeneous (0.9667)

Following the initial 5-problem pilot, which revealed a critical failure in the heuristic-based Team Manager, an expanded pilot study was conducted. This second study tested the newly developed "Culturally-Attuned OODA-Belbin Planner" on the same set of the initial 5 problems. This second research cycle, summarized in Figure 5, aimed to determine if this more theoretically robust manager could overcome the *a priori* team selection problem identified in the first run. The performance of the four experimental conditions—CP-ATS, Single Agent, Static Homogeneous, and Static Heterogeneous—was evaluated using the multi-objective fitness function (F), which provides a nuanced score indicating a solution's proximity to the ground truth.

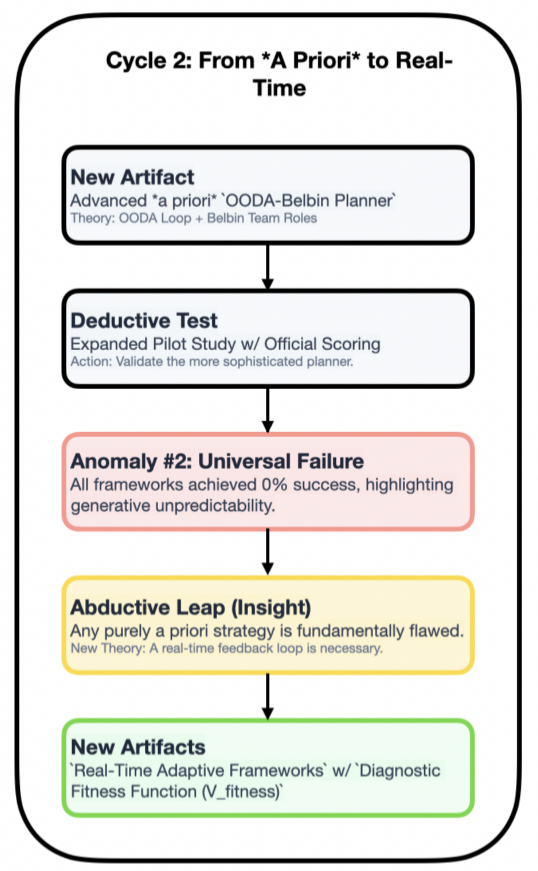
Cycle 2: From *A Priori* to Real-Time
New Artifact
Advanced *a priori* `OODA-Belbin Planner`
Theory: OODA Loop + Belbin Team Roles
Deductive Test
Expanded Pilot Study w/ Official Scoring
Action: Validate the more sophisticated planner.
Anomaly #2: Universal Failure
All frameworks achieved 0% success, highlighting generative unpredictability.
Abductive Leap (Insight)
Any purely a priori strategy is fundamentally flawed.
New Theory: A real-time feedback loop is necessary.
New Artifacts
`Real-Time Adaptive Frameworks` w/ `Diagnostic Fitness Function (V_fitness)`

Figure 5

Research Cycle 2 – From A Priori Planning to Real-Time Adaptation. *This diagram shows the second research cycle, which tested the advanced a priori planner from Cycle 1. The "anomaly" was a universal failure against the official ARC metric, highlighting the inherent unpredictability of the generative agents. This finding motivated the "abductive leap" that any purely upfront planning is fundamentally flawed, necessitating the creation of the* `Real-Time Adaptive Frameworks` *capable of in-mission learning.*

The results, summarized in Table 1, show a competitive performance from the revised

CP-ATS framework based on this diagnostic metric. The CP-ATS condition achieved the highest fitness score on three of the five problems (576224, 009d5c81, and 00d62c1b). However, on problem ‘00dbd492‘—the same problem that caused the initial failure—the CP-ATS framework once again achieved the lowest fitness score (0.9408) of all four conditions. This outcome is particularly significant, as it suggests that even with a more sophisticated, theory-grounded planning stage, the framework remained vulnerable to suboptimal initial team selection. While the diagnostic scores indicate that the generated solutions were often structurally close to the correct answer, this metric does not confirm success under the official ARC benchmark's strict criteria.

4.4.10. Verification Against the Official ARC Scoring Metric

The Abstract Reasoning Corpus (ARC) defines success with a strict, binary metric: a proposed solution is only considered correct if it achieves 100% pixel-wise accuracy against the ground-truth grid. Any deviation, however minor, results in a score of 0 (fail). The diagnostic fitness function (F), while essential for guiding the search process, is insufficient for final validation.

To ascertain the official performance of the agents in the second pilot study, we developed a validation script to systematically execute the Python code generated during each experimental run. For every attempt logged by the agents, the script executed its `solve()` function using the official test `input` grid and compared the resulting `output` grid against the official solution. This process provides an unambiguous, binary score (0 or 1) for every attempt, mirroring the conditions of the official ARC test.

The results of this verification were definitive and revealing: **all attempts, across all five problems and all four experimental conditions, failed to achieve a perfect score.** Every generated solution received an official ARC score of 0.

4.4.11. Analysis and Rationale for a New Research Direction

The universal failure to pass the official ARC test in the second pilot study, especially when contrasted with the partial success observed in the initial pilot, presents a critical finding. This discrepancy does not stem from a change in the definition of success, which remained a constant pixel-perfect match. Rather, it highlights the inherent variability and unpredictability of the agents' generative code capabilities between experimental runs. The agents in the second study, despite being guided by a more advanced Team Manager, simply did not produce code that was 100% correct for any given problem.

This outcome provides the strongest evidence yet that the foundational vulnerability of the CP-ATS framework is its reliance on a purely *a priori* team selection strategy. The results demonstrate that even a sophisticated, theory-grounded initial analysis of a problem's features is insufficient to guarantee the formation of an optimal team. The generative nature of the agents introduces a level of uncertainty that static, upfront planning cannot fully mitigate. The critical failure on problem ‘00dbd492‘ was not an anomaly but a clear signal of this fundamental limitation.

Therefore, the next logical evolution of this research is to pivot from static planning to **real-time adaptation**. The system must be able to not only select a team but also to evaluate its performance *during* a task and dynamically recompose it in response to failure. This requires transforming the Team Manager from a static planner into an adaptive problem-solver that learns and iterates within a single problem instance.

To achieve this, we will leverage the existing multi-objective fitness function not as a final score, but as a **diagnostic tool**. As detailed in the methodology, the vector of component scores (V_{fitness}) serves as a “failure signature.” By feeding this signature back into the Team Manager's OODA loop, the system can map a specific type of failure to a team role deficit, as defined by the Diagnostic Error-to-Belbin Role Mapping (Defined in Section 4.4.8). For example, a solution with high colorimetric fidelity (Level 1) but low object integrity (Level 2) signifies a “Compositional Error,” inferring a deficit of the ’Implementer’ role. Armed with this diagnosis, the Team Manager can dynamically swap team members—for instance, substituting in the ‘Culture_Expert‘ persona—to address the identified weakness before initiating the next attempt. This closed-loop, diagnostic-driven approach directly targets the primary failure mode observed in our pilot studies and represents the necessary next step in developing a truly adaptive multi-agent framework. This theoretical shift, born from the pilot study results, directly informed the core hypotheses that were subsequently tested in our main experiment.

4.5. Real-Time Adaptation via Diagnostic Failure Analysis

To overcome the limitations of a purely *a priori* team selection strategy, we developed a real-time adaptive framework. This framework leverages a multi-objective fitness function not as a simple score, but as a diagnostic "failure signature" ($V_{fitness}$). This signature is fed back into the `TeamManagerAgent`’s OODA loop, which uses the following set of academically-grounded rules to map a specific failure mode to a team role deficit as described by Belbin’s Team Role theory (Belbin & Brown, 2012). This enables the agent to dynamically recompose the team to address identified weaknesses.

4.5.1. Diagnostic Rules for Team Recomposition

Adaptive Role Intensification. A critical refinement to the recomposition logic was developed to address situations where a team fails despite already possessing the persona corresponding to the diagnosed deficit. The initial implementation only checked for the *presence* of a required Belbin role, which could lead to a diagnostic stalemate if, for example, a team with one ’Plant’ still suffered from an ’Ideation Failure’.

To overcome this, we enhanced the `TeamManagerAgent` with the capability for **adaptive role intensification**. The agent no longer simply ensures that a required role is present; it can now decide to increase the *influence* of that role by adding a second, identical persona to the team. Within the AB-MCTS framework, this is a significant strategic decision, as it effectively doubles the opportunities for that specific cognitive style to contribute during the search process. This allows the system to not merely swap agents, but to dynamically adjust the team’s cognitive balance, enabling a more aggressive focus on a failing strategy (e.g., creating a team with two ’Plant’ agents to overcome a severe creativity deficit). This transforms the recomposition from a simple personnel change into a more nuanced adjustment of the team’s overall problem-solving paradigm.

The diagnostic mapping from failure signature to prescriptive action is governed by the following set of Diagnostic rules (labeled with a ’D’ prefix):

D1 Creativity Deficit (’Plant’): Triggered by two primary failure modes:

- *Ideation Failure:* Indicated by persistently low scores across all fitness levels.
- *Stalled Progress:* Indicated by two or more consecutive adaptation cycles with no score improvement.
- **Action:** Increase the weight for the ’Plant’ role to introduce novel, paradigm-shifting ideas and break cognitive fixation (Kirton, 1976).

D2 Execution Deficit (’Implementer’): Triggered by a ’Compositional Error’.

- *Trigger:* Low Level 2 (Object) scores, even if color fidelity is high.

  - **Action:** Enforce the inclusion of the `Culture_Expert` persona, the archetypal implementer.

- **D3 Strategic Deficit ('Monitor Evaluator'):** Triggered by an 'Arrangement/Shape Error'.
  - *Trigger:* High Level 2 scores (correct objects) but low scores on Levels 1, 3, or 4 (incorrect placement, topology, or relational logic).
  - **Action:** Increase the weight for the 'Monitor Evaluator' role to improve strategic oversight.

- **D4 Team Dynamics & Focus Deficit ('Shaper'/'Teamworker'):** Triggered by two distinct modes:
  - *Over-Engineering:* The `obj_count` fitness is low due to producing too many objects. This infers a 'Shaper' deficit.
  - *Team Disintegration:* A significant score drop occurs immediately after a team recomposition. This infers a 'Teamworker' deficit.
  - **Action:** Increase the weight for the corresponding role ('Shaper' to restore focus (Partington & Harris, 1999) or 'Teamworker' to mitigate team friction (S. Fisher et al., 2001)).

- **D5 Precision Deficit ('Completer Finisher'):** Triggered by a 'Precision Error'.
  - *Trigger:* A solution achieves high overall similarity but fails the strict binary check (`pixel_accuracy` > 0.9 but < 1.0).
  - **Action (Escalating):** First, add a 'Completer Finisher'. If the error persists, escalate by also adding an 'Implementer' to address a deeper procedural flaw.

4.5.2. Heuristic for the "Expert Anomaly"

Our pilot studies revealed an "Expert Anomaly" where a homogeneous team of three 'Implementer' agents (`Culture_Expert`) consistently solved a highly complex, procedural problem (‘00dbd492‘) that all diverse teams failed. To address this, we integrated a preemptive heuristic into the initial team selection logic.

Heuristic Trigger: The *a priori* "OODA Reconnaissance" phase classifies a problem as strongly "Adaptive" (e.g., an ‘innovation_score‘ <= -0.5), often signaled by strong symmetrical properties (Rule C3). **Prescriptive Action:** The system overrides the standard diversity-focused team builder and preemptively deploys the homogeneous 'Implementer' team.

4.6. Data Analysis Approach

The data generated from the full-scale experiment was analyzed using a multi-pronged approach designed to move from quantitative observation to qualitative explanation in order to rigorously test the three core hypotheses.

4.6.1. Quantitative and Qualitative Performance Analysis

The primary analysis proceeds in two stages. It is crucial to distinguish our analytical definition of success from the standard ARC-AGI benchmark metric. While the ARC benchmark defines success simply as a perfect pixel match, our methodology incorporates a manual validation step: a detailed analysis of the underlying code for every solution that achieved this benchmark. This validation allowed us to classify these outputs into two groups for our analysis: *real successes*, defined as solutions with logically sound and generalizable code, and *failures*, defined as solutions whose code was flawed and non-generalizable, having passed the test case only by coincidence. The quantitative and qualitative analyses presented herein are based on this refined classification.

First, a quantitative analysis was conducted to establish a baseline of performance. It is important to note that the total number of solutions achieving a perfect pixel match under the ARC-AGI2 standard was not uniform across all six experimental conditions, with the count of such outputs ranging from 25 to 33. To enable a standardized and fair comparison across these uneven sample sizes, all primary performance metrics in this analysis, such as success and contamination rates, are reported as percentages.

A Chi-Squared test of independence was then performed on the raw counts of *real* successes versus failures to determine if the observed variations in these normalized rates were statistically significant. The test revealed no significant difference in the ‘real success’ rates between conditions ($\chi^2(5,N = 177) = 3.86, p = 0.57$). This finding is critical: it demonstrates that a simple, quantitative measure like ‘success rate’ is insufficient to truly differentiate the performance of these complex agentic systems.

This result reinforces the necessity of the subsequent qualitative analysis. This second stage of analysis involves a deep examination of the logic and structure of the generated code to build a “Taxonomy of Failure” and an “Anatomy of Success.” These components together allow for the construction of a detailed cognitive profile for each experimental condition. Furthermore, this process revealed a more nuanced structure within the problem space itself, leading to the refinement of our initial three-category problem taxonomy into a more descriptive four-category model. This refined taxonomy, presented in Section 5, is instrumental in explaining the observed performance paradoxes and validating our contingency model.

4.6.2. Qualitative Efficiency and Solution Quality Analysis

To evaluate the resource cost and reasoning efficiency associated with each framework, a qualitative analysis was conducted based on the comprehensive audit of all ’real success’ solutions. Instead of a quantitative comparison of iteration counts, which was not feasible due to the experimental setup, this analysis uses comparative case studies to assess efficiency through the lens of solution quality.

For problems solved by multiple frameworks, the generated code was compared based on criteria such as **parsimony** (directness and conciseness), **algorithmic purity** (use of an optimal or standard algorithm), and **robustness** (avoidance of over-specified or brittle logic). This qualitative assessment serves as a powerful proxy for efficiency, as a framework that produces a more elegant and direct solution is demonstrating a more efficient reasoning path. These case studies are used to provide concrete, mechanistic evidence for the paper’s core hypotheses.

4.7. Validity and Reliability

To ensure the trustworthiness of the study’s findings, the research design incorporates specific measures to address internal validity, external validity, and reliability. This section explicitly outlines how these principles are upheld.

4.7.1 Internal Validity

Internal validity, the extent to which causal claims can be confidently made, is ensured through the strict, controlled experimental design. By holding several key factors constant across all experimental conditions, the study isolates the independent variable as the sole source of systematic variation. The control variables include:

- **The Problem Set:** All frameworks were tested on the identical set of problems from the standardized ARC-AGI-2 dataset.

- **The Computational Budget:** A consistent total computational budget of 250 iterations was enforced for each problem instance across all conditions.

- **The Underlying Large Language Model (LLM):** The same foundational LLM, Deepseek-coder-v2-instruct, was used for all agents in all frameworks.

Because these variables are controlled, any observed differences in the dependent variables (Task Success Rate and Computational Efficiency) can be attributed with high confidence to the manipulation of the independent variable—the problem-solving framework.

4.7.2 External Validity (Generalizability)

External validity refers to the extent to which the study's findings can be generalized to other contexts. It is acknowledged that the results of this research are directly generalizable only to the Abstraction and Reasoning Corpus (ARC) benchmark. However, a strong argument is made for the broader relevance of the findings. ARC is not an arbitrary task set; it is meticulously designed as a proxy for general fluid intelligence and embodies the core characteristics of "wicked problems"—ill-structured challenges with no definitive formulation or stopping rule.

Therefore, the study's central conclusions regarding the contingency-based nature of problem-solving—and the specific trade-offs between simplicity, homogeneity, and adaptation—have significant theoretical and practical implications for designing AI systems intended to solve other complex, ill-structured problems in diverse real-world domains.

4.7.3 Reliability

Reliability, or the consistency and replicability of the research, is ensured through two primary mechanisms:

1. **Use of a Standardized Benchmark:** The reliance on the public, standardized ARC benchmark ensures that the tasks are consistent, well-defined, and available for other researchers to use for replication and comparative studies.

2. **Transparent and Replicable Protocol:** The detailed documentation of the methodology within this paper—including the agent architectures, the specific heuristics, the diagnostic rules for adaptation, and the formal experimental design—provides a clear and replicable protocol. This transparency allows other researchers to verify the findings and build upon the work with confidence.

4.8. Role of AI Assistance in the Research Process

To ensure full academic transparency, this section details the integral role of generative AI as an assistive tool throughout the research lifecycle. The primary tool utilized was Google's Gemini 2.5 Pro, supplemented by Jenni AI for specific literature search tasks. The human author directed all research activities, critically validated all AI-generated outputs, and retains full intellectual responsibility for the final work.

4.8.1. Scope and Application of AI Assistance

The application of AI assistance can be categorized by research phase:

- **Literature Review and Conceptualization:** Jenni AI and Gemini 2.5 Pro were employed to conduct broad searches for relevant academic literature and to synthesize foundational concepts. Gemini 2.5 Pro was further utilized to advise on the overall structure and narrative arc of the paper.

- **Experimental Design and Implementation:** Gemini 2.5 Pro was instrumental in designing and writing the Python code for the experimental frameworks. It provided crucial advice on technical decisions, including hardware selection, the choice of code-running environments, and a cost-benefit analysis of using API-based versus local models. It also served as a continuous debugging partner, helping to identify and resolve errors as they appeared.

- **Data Analysis and Manuscript Preparation:** AI assistance was used in the analysis of the experimental data, particularly in identifying patterns in the results. Subsequently, Gemini 2.5 Pro was used extensively to draft, edit, and refine the different sections of this manuscript to meet a high standard of academic writing.

4.8.2. Reflections on Best Practices for Human-AI Collaboration

The experience of using a state-of-the-art LLM as a research assistant revealed several characteristic failure modes and led to the development of a set of best practices for ensuring rigor and quality. These reflections are offered as a contribution to the emerging field of human-AI research collaboration:

- **Managing Generative Consistency:** The model occasionally exhibited a tendency to revert to previous, outdated responses. A best practice was to re-prompt with the same query or, if unsuccessful, to start a new session to reset the context. Periodically reattaching key data files was also crucial to prevent the model from referencing cached, older versions of the data.
- **Ensuring Factual and Numerical Accuracy:** The model demonstrated limitations in precise numerical counting and could occasionally introduce unsupported assumptions or "fake data." All quantitative outputs required manual verification or validation with a separate script. Prompts were explicitly engineered to instruct the model to rely *only* on the provided data files.
- **Controlling for Omissions and Complexity:** To prevent the model from omitting key details in complex tasks, a two-step prompting process was adopted: first, requesting a high-level design plan, and then prompting for the execution of each step sequentially. To manage complexity, prompts were designed to include explicit constraints (e.g., time, budget, technical limitations) and to request the simplest viable solution.
- **Systematic Code Validation:** When generating or debugging code, the model sometimes produced code with placeholders or introduced errors with cascading dependencies. The most effective mitigation strategy was to provide the model with all relevant program files, prompt it to perform a holistic error check, and then re-verify all dependencies after any modification.

Ultimately, this research confirms that while generative AI is a powerful accelerator for complex research, it functions as an assistant, not arbiter of truth. Continuous and meticulous human oversight, critical validation, and a structured, iterative prompting strategy are essential to ensure the validity, accuracy, and overall quality of the final research output.

4.9. Methodological Refinements for the Frontier Challenge

To address the “local optima trap” and the challenge of “Generative Exhaustion” identified in the final research cycle, two critical refinements were implemented as artifacts. These methodological enhancements, designed to address the **’Local Optima Trap’ (Anomaly #3, Section 6.2)** by combating its inferred root cause, **’Generative Exhaustion’ (Section 6.3)**, were applied only in the final multi-run study detailed in **Section 6.4**.

4.9.1. Refinement 1: A Gated Scoring Framework

To enforce a stricter evaluation hierarchy, a Gated Scoring Framework was implemented. This mechanism acts as a preemptive check before the standard weighted-average score is calculated, prioritizing fundamental logical correctness over superficial similarity.

Gate 1: Object Count Integrity. This gate assesses whether the solution has produced a plausible number of objects. It is triggered if the object count fitness, f_{obj_count}, is less than 0.5, which corresponds to a solution producing either less than half or more than double the number of objects in the ground-truth grid.

Gate 2: Object Structure Integrity. This gate assesses the structural correctness of the generated objects using the Jaccard Index. It is triggered if the Jaccard similarity for the object sets, $f_{jaccard\_objects}$, is less than 0.6, indicating a major error in shape or position.

The Penalty Function. If either gate is triggered, the standard evaluation is bypassed and a penalized score is calculated to signal to the search algorithm that the solution path is a dead end. To provide a minimal gradient, the penalty function is defined as:

$$F_{penalized} = 0.1 \times f_{pixel\_accuracy} \tag{10}$$

.

4.9.2. Refinement 2: The Self-Correction Loop

To combat generative exhaustion directly, the `Self-Correction Loop` was implemented as the tactical layer of the Hierarchical OODA-Belbin Framework. This artifact operationalizes a tactical "micro-loop," limited to a maximum of three correction attempts within a single MCTS expansion step. This allows an agent to learn from its own immediate failures through a rapid, diagnostic-driven OODA cycle. The process leverages specific prompting strategies, including a **"Chain-of-Thought"** prompt to force an analysis of the error, and a forceful **"Meta-cognitive"** prompt on repeated failures to break cognitive fixation.

- **Chain-of-Though prompt** This prompt is made of 4 parts; an introductory part that calls the LLm to analyze the feedback, the previous failed code, the failure analysis with its explanation and the new task including the chain of thoughts instructions. This example is from run 4, problem 8f3a5a89, API Call 11/250:

"You are an expert AI programmer debugging a solution for an Abstraction and Reasoning Corpus (ARC) problem. Your previous attempt failed. You must analyze the provided error feedback and your previous code to generate a corrected solution.

— YOUR PREVIOUS FAILED CODE —

python code...

— FAILURE ANALYSIS —
Your code was executed and failed the evaluation. Here is the diagnostic feedback:
- Failure Status: GATE_1_FAILED
- Detailed Fitness Scores: {...}

A 'GATE_1_FAILED' status means your code produced the wrong number of objects.

A 'GATE_2_FAILED' status means the objects had the wrong shape or structure.

— YOUR TASK —
1. **Analyze the Failure**: In a few sentences, explain why your previous code failed based on the feedback.

2. **Formulate a Plan**: Outline your step-by-step plan to fix the code.

3. **Write the Corrected Code**: Provide the new, corrected Python function 'solve(grid)'.

Your entire response must follow this structure and place the final Python code inside a markdown block."

- **Meta-cognitive prompt**

This prompt is made of 4 parts; an initial prompt containing a "Critical Feedback", the previous failed strategy, new instructions to force the LLM to abandon its current strategy and change its approach and the original prompt that presents the problem to solve:

The example is taken from This example is from run 4, problem 8f3a5a89, API Call 12/250:

"You are an expert AI programmer debugging a solution for an Abstraction and Reasoning Corpus (ARC) problem.

CRITICAL FEEDBACK: Your previous attempts have failed with the same logical error. Your current approach is fundamentally flawed and must be abandoned.

— FLAWED STRATEGY —
Your previous code, shown below, failed with the error: GATE_1_FAILED.
“‘python . . . code. . . “‘

— NEW INSTRUCTIONS —
DO NOT try to fix or modify your previous code. It is incorrect.

You **MUST** abandon your current strategy and devise a completely new and different method for solving the problem.

Re-examine the original problem description below and come up with a fresh, alternative approach.

— ORIGINAL PROBLEM DESCRIPTION —
. . . .. original prompt. . . ."

5. Results: The Abductive Discovery of the Contingency Model

Following the development of the real-time adaptive frameworks, a comprehensive 120-problem experiment was conducted to test the full suite of six agentic architectures. The initial hypothesis, grounded in prevailing assumptions, was that the most complex framework, `Adaptive_CP-ATS_Budget_Pool`, would prove universally superior.

The early phases of this research quickly refuted this simple assumption, revealing instead a nuanced contingency model where the optimal architecture is dependent on the specific nature of the problem. This discovery led to a reframing of the primary research objective toward validating this emergent model.

The remainder of this section is therefore structured to present a quantitative overview of the successful solutions, then the definitive deductive validation for each of the three core hypotheses that form this contingency model:

1. **The Simplicity Hypothesis:** A significant portion of ARC problems are best suited to simpler architectures, where the overhead of teamwork is a detriment.
2. **The “Expert Anomaly” Hypothesis:** A non-diverse, homogeneous team of specialists provides a unique and essential capability for a specific class of high-precision tasks.
3. **The Adaptation Hypothesis:** The true value of a highly adaptive framework lies in its exclusive ability to solve the most complex, multi-stage problems that are intractable to all other approaches.

5.1. Quantitative Overview: A Story of Paradoxes

A comprehensive quantitative analysis of the 177 audited solutions reveals a series of paradoxes. To provide a clear basis for this analysis, we first define the different tiers of solution quality observed in the study.

Defining Solution Quality. The solutions that achieved a ’perfect pixel match’ were categorized based on a rigorous audit of their underlying code to assess their logical soundness and generalizability. **High-Quality Successes** are solutions that are not just correct, but are well-reasoned and robust. This category includes solutions deemed *Optimal &*

Robust (representing the canonical, most efficient algorithm) and *Optimal & Parsimonious* (correctly identifying the simplest possible rule). In contrast, **Lower-Quality Successes** are solutions that passed the specific test case but contained logical flaws. This includes solutions that were *Correct but Inefficient*, *Correct but Inefficient & Convoluted*, or *Robust but Over-Specified*. These are counted as 'real successes' for the overall success rate but are excluded from the 'high-quality' metric.

Table 2

Real Success and Solution Quality Rates per Condition

Condition	Real Success %	High-Quality %
Adaptive_CP-ATS_Budget_Pool	48.3%	85.7%
Adaptive_CP-ATS_Fixed_Cycle	64.0%	93.8%
Baseline_Single_Agent	60.6%	80.0%
Baseline_Static_Heterogeneous	**71.9%**	91.3%
Baseline_Static_Homogeneous	60.7%	83.3%
CP-ATS_Static	56.7%	**94.1%**
Overall	**60.5%**	**88.0%**

Note: Percentages are rounded to one decimal place.

Table 3

Data contamination per condition

Condition	Percentage
Adaptive_CP-ATS_Budget_Pool	**71.4%**
Adaptive_CP-ATS_Fixed_Cycle	60.0%
Baseline_Single_Agent	66.7%
Baseline_Static_Heterogeneous	66.7%
Baseline_Static_Homogeneous	50.0%
CP-ATS_Static	50.0%
Overall	**61%**

Note: Percentages are rounded to one decimal place.

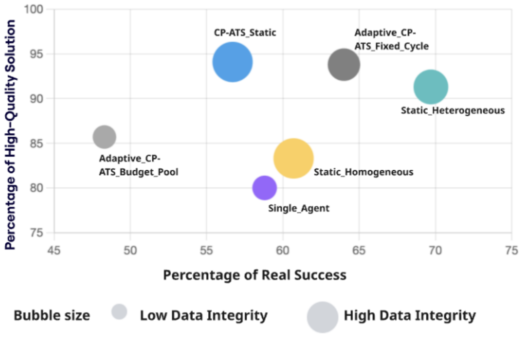

Figure 6

Real Success, Reliability, and Integrity Matrix

Quantitative Findings. The primary finding is that a simple, quantitative measure like 'success rate' is insufficient to truly differentiate the performance of these complex agentic systems. While the 'real success' rates varied from 48.3% to 71.9%, a Chi-Squared test of independence found that these differences were not statistically significant ($\chi^2(5, N = 177) = 3.86, p = 0.57$).

The truly meaningful distinctions emerge from the analysis of solution quality and vulnerability to data contamination (see Tables 2 and 3). These results reveal a central tension between the *quantity* of success and the *quality* of the resulting solutions (See Figure 6). For instance, the condition with the highest success rate, `Baseline_Static_Heterogeneous` (71.9%), was not the one that produced the most reliable solutions. That distinction belongs to `CP-ATS_Static`, which delivered high-quality, robust code in 94.1% of its successful attempts.

This quantitative puzzle necessitates the deep qualitative analysis that follows to validate our hypotheses.

A critical methodological finding was the "Ghost Problem" anomaly. For problem 0520fde7, three conditions were recorded as successful despite the problem being absent from the V2 dataset. In these cases, the system defaulted to its memory of the V1 version. This anomaly, a form of *Default-based Data Contamination*, underscores the critical need to augment standard AI benchmarks with the logical code validation performed in this study, as a 'perfect pixel match' can completely mask underlying data pipeline failures.

5.2. Overview of Exclusive Solves and Emergent Archetypes

The most critical insights from the 120-problem experiment emerge from an analysis of the problems solved exclusively by a single framework. The definitive data, presented in Table 4, provides clear evidence for the specialized value of each architectural class and confirms the trade-offs between simplicity and complexity.

Table 4

Unique Problems Solved Exclusively by a Single Condition

Condition	Unique Solves	Total
Baseline_Single_Agent	1c786137	1
Baseline_Static_Homogeneous	05269061	1
Adaptive_CP-ATS_Fixed_Cycle	1d398264	1
Adaptive_CP-ATS_Budget_Pool	22168020, 0becf7df, 00d62c1b	3

At first glance, these exclusively solved problems seem to offer strong support for our three core hypotheses, suggesting a clear mapping between problem archetype and architectural strategy. The following sections will now undertake a deep, qualitative analysis of these and other solutions to rigorously test this initial observation.

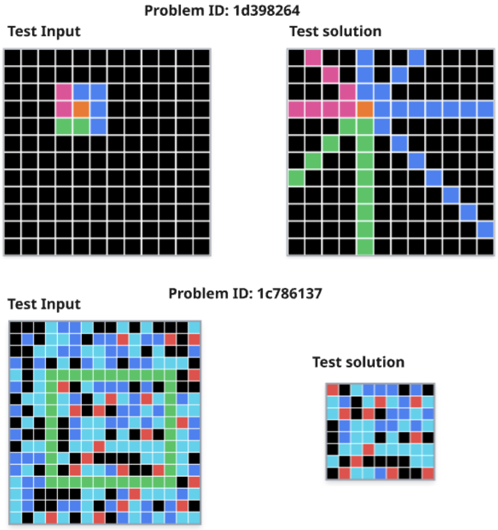

Figure 7

Direct Procedure Tasks: Problem 1d398264 (to solve it, the agent must extend the colors from the square diagonally) and problem 1c786137 (the agent must extract the pixels inside the square)

Direct Procedure Tasks. Two of the exclusive solves fall under this archetype, where the core logic is a singular, global rule (See Figure 7). Problem 1d398264[2], solved only by the Adaptive_CP-ATS_Fixed_Cycle, involves extending the shape in all directions. Similarly, problem 1c786137, solved only by the Baseline_Single_Agent, requires a global find-and-replace operation . The apparent success of these simpler architectures presents a compelling apparent case for the Simplicity Hypothesis, which the following analysis will explore in detail.

Meticulous Execution Tasks. The analysis provides strong initial evidence for the "Expert Anomaly." Problem 05269061, solved exclusively by the

[2] The test of Problem 1d398264 has two sets. Only one set is presented in its figure

`Baseline_Static_Homogeneous` team, is a canonical “Meticulous Execution” task, requiring the precise, hierarchical replication of a small pattern into a larger structure (See Figure 8). This finding suggests that for high-fidelity procedural tasks, cognitive homogeneity may be a strategic advantage—a claim we will rigorously test.

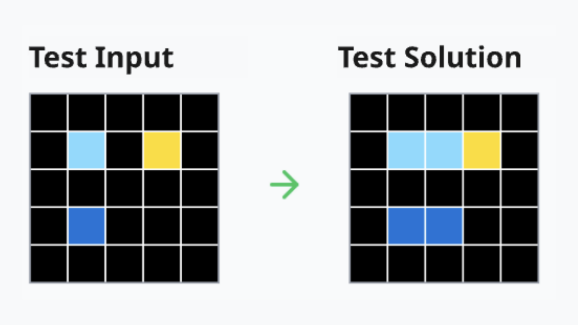

Figure 8

Meticulous Execution task: Problem 05269061 (To solve this problem, the agent must duplicate the colored pixels that are on the left side only)

Cognitive Labyrinth Tasks. Initially, the results for the most complex problems seemed to offer a clear victory for architectural complexity. The `Adaptive_CP-ATS_Budget_Pool` was the sole recorded solver of three "Cognitive Labyrinth" problems (See Figure 9). Each requires multi-stage, relational reasoning: `00d62c1b` involves noise reduction and shape completion; `0becf7df` requires complex spatial-relational coloring; and `22168020` involves shapes identification and completion. On the surface, this would appear to confirm the Adaptation Hypothesis. However, as the detailed audit in Section 5.5 will reveal, these apparent successes are the most powerful illustration of the danger of ’false positives,’ serving as the crucial evidence that ultimately falsifies this hypothesis.

5.3. Deductive Validation I: The Value of Simplicity for Direct Procedure

Our first hypothesis posited that for a significant subset of **"Direct Procedure"** problems, a single agent would be the most effective and efficient framework. The comprehensive analysis of all 177 audited solutions provides strong validation for this hypothesis.

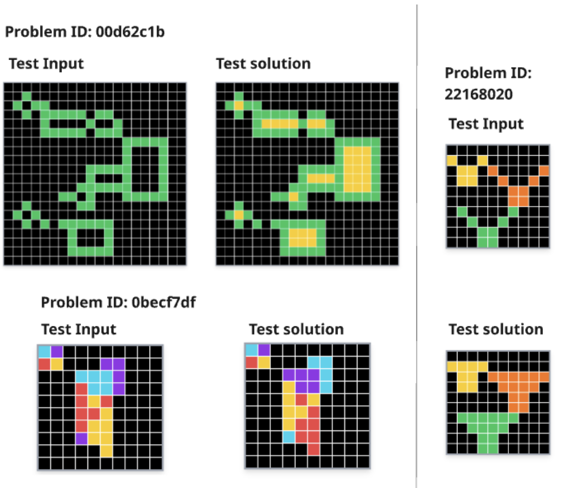

Figure 9

Problems 00d62c1b, 0becf7df and 22168020: Cognitive Labyrinth inputs and solutions

Our analysis confirms that for tasks defined by a clear, deterministic, and often global rule (e.g., ‘1e0a9b12‘, ‘0c786b71‘), all six experimental conditions demonstrated competence. This finding supports the core of the Simplicity Hypothesis: for such straightforward tasks, the cognitive and computational overhead of multi-agent teamwork is unnecessary, making a low-overhead single agent the most efficient choice by definition.

Furthermore, the `Baseline_Single_Agent`ś unique cognitive style was powerfully demonstrated through its exclusive, optimal solves of problems requiring **Holistic Abstract Reasoning** (‘12997ef3‘[3], ‘0962bcdd‘). This proven specialization in finding a single, "big picture" insight reinforces its value, confirming that for problems where the core logic is monolithic, the `Baseline_Single_Agent` is not just a sufficient choice, but a uniquely capable one, making it the optimal architecture for this entire class of problems. However,

[3] The test of Problem 12997ef3 has two sets. Only one set is presented in the figure

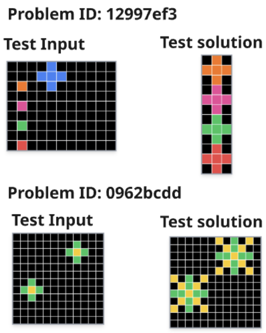

Figure 10

Problems 12997ef3 (To solve this problem, the agent must duplicate the shape and color it according to the pixels presented with it) and 0962bcdd (To solve this problem, the agent must extend the shape)

this specialization comes at a significant cost. The "Taxonomy of Failure" reveals that the `Baseline_Single_Agent` suffers a catastrophic "cognitive cliff" when faced with *Cognitive Labyrinths*, where its inability to manage multiple interlocking steps leads to fundamentally broken logic. Moreover, its holistic pattern-matching strength makes it the most susceptible condition to *Data Contamination*. This duality defines its profile as a high-risk, high-reward "Monolithic Reasoner", making it the optimal choice for one class of problems (See Figure 10) but dangerously unreliable for others.

5.4. Deductive Validation II: Definitive Evidence for the “Expert Anomaly” on Meticulous Execution Tasks

Our second hypothesis predicted an "Expert Anomaly," where a non-diverse, homogeneous team of ’Implementer’ agents would outperform all other conditions on a specific class of "Meticulous Execution" problems. The initial evidence for this counterintuitive phenomenon emerged from the analysis of exclusive solves, where the `Baseline_Static_Homogeneous` condition was the unique solver for problem `05269061`, a canonical “Meticulous Execution” task requiring the precise replication of a hierarchical pattern.

The comprehensive analysis of all 177 audited successful solutions provides even more definitive validation, revealing the condition’s cognitive profile as a highly specialized "Expert Proceduralist." The most powerful evidence for this specialization comes from its performance on high-precision procedural tasks. In the case of problem `08ed6ac7`, which required the identification of connected components, the `Baseline_Static_Homogeneous` condition was one of only two to produce an optimal solution that more creative and complex conditions failed to achieve. This confirms that for well-defined procedural problems where creative deviation is detrimental, the focused cognitive style of a homogeneous team is a strategic advantage.

However, our qualitative analysis also reveals that this expertise is exceptionally narrow. The "Taxonomy of Failure" shows that this condition is highly prone to cognitive biases when faced with problems outside its procedural domain. It frequently produced bizarrely inefficient and convoluted logic for simple tasks (e.g., `1e0a9b12`), and invented flawed numerical rules when confronted with abstract problems that did not fit its rigid, step-by-step worldview. This duality is the essence of the "Expert Anomaly": the condition’s cognitive homogeneity is the source of both its unique strength on a narrow set of problems and its profound brittleness on all others.

5.5. Falsification and Refinement: Cognitive Diversity, Not Complexity, Unlocks “Cognitive Labyrinths”

Our third and final hypothesis, the **Adaptation Hypothesis**, predicted that our most complex, real-time adaptive framework (`Adaptive_CP-ATS_Budget_Pool`) would possess an exclusive capability to solve the most difficult "Cognitive Labyrinth" problems. The initial evidence appeared to support this, with the framework achieving three exclusive solves on problems `22168020`, 0becf7df, and 00d62c1b.

However, the comprehensive audit of all the successful solutions definitively **falsifies this hypothesis**. The rigorous code analysis revealed that these apparent exclusive solves were, in fact, "false positives." They were brittle, non-generalizable solutions, or were instances of data contamination where the system had solved the wrong problem version entirely. Furthermore, the complete quantitative analysis reveals that the Adaptive_CP-ATS_Budget_Pool was the poorest-performing condition overall, with the lowest 'real success' rate (48.3%) and a high susceptibility to data contamination (71.4%). This provides powerful evidence for a **"cost of complexity,"** where the cognitive overhead of managing a highly dynamic architecture can hinder, rather than help, effective problem-solving.

The crucial discovery from our analysis is that the true key to unlocking "Cognitive Labyrinths" is not adaptive complexity, but **cognitive diversity**. The Baseline_Static_Heterogeneous condition emerged as the clear specialist for this problem archetype. Its cognitive profile as a "Master Algorithm Designer" enabled it to be the only condition to produce optimal, robust solutions for some of the most algorithmically demanding problems, such as devising the correct multi-step sorting algorithm for 19bb5feb.

This finding forces a refinement of our initial theory: while adaptation has its place, the synthesis of multiple, distinct cognitive styles within a stable, heterogeneous team is the essential ingredient for the decomposition and algorithmic synthesis required to solve the most wicked problems in the ARC dataset. This insight is the final key to our nuanced, portfolio-based model of agentic performance.

5.6 Synthesis: A Refined Contingency Model and the Justification for a Hierarchical Approach

The deductive validation of our three initial hypotheses provides the definitive explanation for the quantitative paradoxes established in Section 5.1. The initial Chi-Squared test of independence, which found no statistically significant difference in the 'real success' rates between conditions ($\chi^2(5, N = 177) = 3.86, p = 0.57$), confirmed that a simple success metric is insufficient to differentiate performance. The qualitative validation now reveals why: the evidence confirms the "Simplicity Hypothesis" and the "Expert Anomaly" while decisively falsifying the "Adaptation Hypothesis." This complexity is not a contradiction, but rather evidence of a deeper underlying structure in the problem space itself. The key to resolving these paradoxes and building a truly effective contingency model lies in a refined taxonomy of ARC problems, derived from our comprehensive qualitative analysis.

A Refined Four-Category Taxonomy of ARC Problems. Our initial three-category model, while a useful starting point, proved incapable of explaining the full spectrum of results. The deep analysis of the audited solutions forces the refinement of this model into a more precise, four-category taxonomy:

1. **Direct Procedure Tasks:** Problems solvable by a single, clear, deterministic rule. Our analysis confirms all conditions are competent here.

2. **Local Geometric & Procedural Analysis:** A refinement of "Meticulous Execution," these tasks require a detailed, procedural analysis of a pixel's immediate neighbors, often involving ambiguity. Our evidence shows Baseline_Static_Homogeneous excels here.

3. **Holistic Abstract Reasoning:** A new category, formerly hidden within

"Meticulous Execution," these tasks require a single, non-obvious "big picture" insight. Our evidence proves the `Baseline_Single_Agent` is the unique specialist for this archetype.

4. **Cognitive Labyrinths:** The most algorithmically demanding problems, requiring the synthesis of a non-trivial, multi-step algorithm. Our evidence proves that the cognitively diverse `Baseline_Static_Heterogeneous` team is the master of this domain.

The Duality of "Meticulous Execution": The Final Insight. The explanatory power of this refined taxonomy is most profound when we re-examine the initial "Meticulous Execution" category. Its splitting into two distinct types—*Local Geometric Analysis* and *Holistic Abstract Reasoning*—is the central discovery of our research. It perfectly explains the paradoxical performance of the specialized conditions. It reveals why the `Baseline_Single_Agent` ("Monolithic Reasoner") sometimes dramatically outperformed complex teams, and why the focused, non-diverse `Baseline_Static_Homogeneous` team ("Meticulous Executor") could succeed where more creative architectures failed.

This discovery provides the definitive, evidence-based justification for the paper's central thesis: **a high-level Strategic Selector is essential.** No single baseline condition can be optimal for both sub-types of the original "Meticulous Execution" category, as they require opposite architectural solutions. Only a hierarchical system with a strategic layer capable of diagnosing a problem at this deeper level can hope to make the correct, contingent choice.

Quantitative Validation of the Portfolio Approach. This theoretical necessity for a portfolio approach is empirically validated by an aggregate analysis of the unique 'real' problems solved. The collective portfolio of all six conditions, in the 120-problems experiment, solved a total of 30 unique problems (See Table 5). The best-performing individual condition in the same experiment, `Baseline_Static_Heterogeneous`, solved 21. This represents a substantial **42.9% increase** in problem-solving breadth over the best single architecture. This finding empirically invalidates a monolithic 'best framework' approach and provides the strongest possible quantitative validation for a portfolio-based strategy.

This validated need for a portfolio-based strategy provides the direct rationale for the third and final DSR cycle, detailed in Section 6: to operationalize our *initial three-category contingency model* into a functional artifact and test its predictive power against the most difficult frontier problems.

Table 5

Real success data for the 120 problems experiment

Condition	Real success
Adaptive_CP-ATS_Budget_Pool	13
Adaptive_CP-ATS_Fixed_Cycle	16
Baseline_Single_Agent	17
Baseline_Static_Heterogeneous	21
Baseline_Static_Homogeneous	15
CP-ATS_Static	15
Total unique successes	**30**

6. The Frontier Challenge: Evolving the Framework to Solve "Trap" Problems

The development of the Hierarchical Framework for the "Frontier Challenge" represents the third and final cycle of our Design Science Research methodology. This system, particularly its Strategic Selector Agent, was designed to operationalize the initial three-category contingency model (*Direct Procedure, Meticulous Execution, Cognitive Labyrinth*) derived from our preliminary pilot studies. It is crucial to note that this experimental phase was conducted based on this initial theoretical understanding, *before* the discovery of the more refined four-category taxonomy detailed in the preceding analysis. The results of this Frontier Challenge, therefore, serve a dual purpose: they test the efficacy of a portfolio-based approach while also revealing the limitations of the initial problem classification, thereby providing the final piece of evidence for the necessity of the refined model.

6.1. Deductive Test: Deploying the Strategic Selector Agent

The consolidated analysis of the 120-problems experiment, which validated our nuanced portfolio model, provides the direct rationale for the third and final Design Science Research (DSR) cycle: to operationalize this contingency model into a functional artifact. This final phase of the research, summarized in Figure 11, involved the design and implementation of a `Strategic Selector Agent`. This meta-agent's role is not to select individual personas, but to select the entire problem-solving condition based on an initial analysis of the task, transforming our descriptive model into a prescriptive, functional system.

The agent leverages the same "OODA Reconnaissance" phase as the adaptive models to classify a problem. Crucially, its deployment logic for this test was governed by the **initial three-category contingency model** and the corresponding hypotheses formulated from our preliminary studies (Section 3), not the refined, final conclusions presented in Section 5. It makes its high-level strategic decision based on the following set of empirically grounded Strategy rules (labeled with an 'S' prefix):

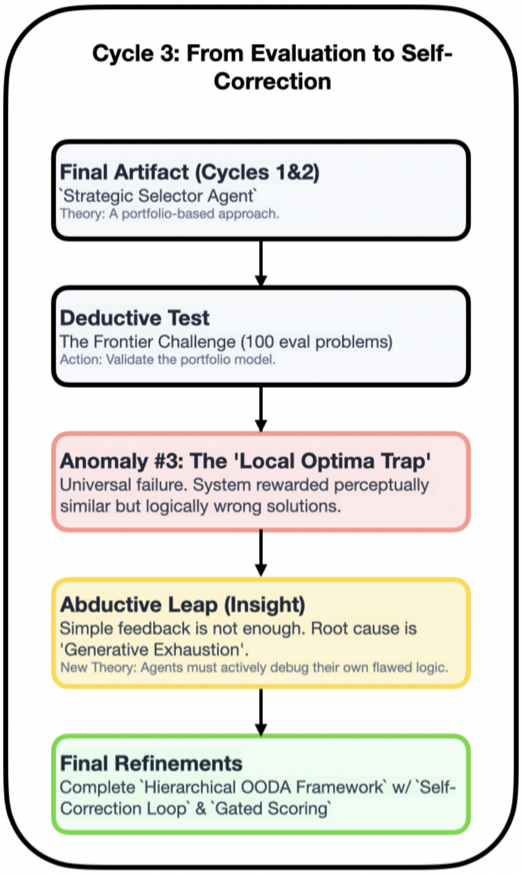
Cycle 3: From Evaluation to Self-Correction
Final Artifact (Cycles 1&2)
`Strategic Selector Agent`
Theory: A portfolio-based approach.
Deductive Test
The Frontier Challenge (100 eval problems)
Action: Validate the portfolio model.
Anomaly #3: The 'Local Optima Trap'
Universal failure. System rewarded perceptually similar but logically wrong solutions.
Abductive Leap (Insight)
Simple feedback is not enough. Root cause is 'Generative Exhaustion'.
New Theory: Agents must actively debug their own flawed logic.
Final Refinements
Complete `Hierarchical OODA Framework` w/ `Self-Correction Loop` & `Gated Scoring`

Figure 11

Research Cycle 3 – From Evaluation to Self-Correction. *This final cycle operationalizes the portfolio model into a* `Strategic Selector Agent`*. Its failure on the frontier challenge (the 'Local Optima Trap') led to the deepest "abductive leap": diagnosing 'Generative Exhaustion' as the root cause of the failure. This insight drove the development of the paper's most advanced artifacts—the tactical* `Self-Correction Loop` *and* `Gated Scoring`*—completing the final* `Hierarchical OODA-Belbin Framework`*.*

- **S1 (Monolithic Task Strategy):** If a problem was classified as a *Direct Procedure* task, the agent deployed the `Baseline_Single_Agent`, in line with the *Simplicity Hypothesis*.
- **S2 (Expert Anomaly Strategy):** If a problem was classified as a *Meticulous Execution* task (a

strongly "Adaptive" task approximated by an `innovation_score <= -0.5`), the agent deployed the `Baseline_Static_Homogeneous` team, in line with the *"Expert Anomaly" Hypothesis*.

- **S3 (Cognitive Labyrinth Strategy):** Based on our initial (and subsequently falsified) *Adaptation Hypothesis*, if a problem was classified as a highly innovative *Cognitive Labyrinth* (approximated by an `innovation_score >= +0.5`), the agent deployed the `Adaptive_CP-ATS_Budget_Pool`.

- **S4 (Default Diversity Strategy):** For all other problems that did not strongly fit the specialized archetypes, the agent defaulted to deploying the cognitively diverse `Baseline_Static_Heterogeneous` team.

Following its implementation, the `Strategic Selector Agent` was deployed for a rigorous deductive test against a distinct set of 100 challenges from the official ARC-AGI-2 evaluation dataset. This test, therefore, serves not only to validate the efficacy of a portfolio-based approach but also to stress-test the predictive power of our initial, and as we will see, incomplete, problem taxonomy against the most challenging tasks in the ARC corpus.

6.2. Anomaly #3: The 'Local Optima Trap'

The outcome of this deductive test was both definitive and unexpected, producing the third major empirical anomaly of the research. The system failed to successfully solve *any* of the 100 problems in the evaluation set.[4]

A detailed analysis of the experimental logs revealed the root cause was a profound failure in the Strategic Selector Agent's diagnostic capabilities, operating with its initial three-category model (Direct Procedure, Meticulous Execution, Cognitive Labyrinth). The agent deployed the `Adaptive_CP-ATS_Budget_Pool` framework (intended for Cognitive Labyrinths) for only 2 problems. For the remaining 98 problems, its internal heuristics failed to confidently classify them as matching any specific archetype strongly enough to trigger a specialist deployment (Rules S1, S2, or S3). Consequently, it invoked its default strategy (Rule S4), deploying the generalist `Baseline_Static_Heterogeneous` framework.

This reveals a drastic underestimation of the evaluation set's complexity and significant flaws in the agent's heuristics, even within its own 3-category model. Our refined analysis shows the evaluation set actually comprised **40 Simple Procedural Tasks, 18 Local Geometric & Procedural Analysis tasks, 18 Holistic Abstract Reasoning tasks, and 24 Cognitive Labyrinths**. The agent's heuristics failed profoundly in two ways:

- It failed to identify 22 out of the 24 problems that should have qualified as 'Cognitive Labyrinths' under its operating model, incorrectly assessing them as less complex.

- It failed to identify the 40 'Simple Procedural Tasks' for which Rule S1 (deploying the `Single_Agent`) should have been triggered, instead defaulting to the generalist team.

Furthermore, the 36 problems requiring specialist solutions under the refined taxonomy ('Holistic Abstract Reasoning' and 'Local Geometric Analysis') were inevitably misclassified or defaulted upon, as the agent's 3-category model lacked the necessary representational capacity. This widespread failure

[4] Of the 100 challenges in the evaluation dataset, 99 resulted in failure. The run for the final problem was not completed due to a system crash, and its result is therefore considered inconclusive.

to correctly assess problem type led to the inappropriate application of the default strategy (Rule S4) in 98% of cases.

This universal failure stems from a **dual-layer misdiagnosis**:

1. **Heuristic Inaccuracy:** The agent's primary failure was its inability to recognize complexity or specialization using its feature-extraction and scoring mechanism. It failed to confidently classify 98% of the problems, including the vast majority requiring specialist or advanced architectures, thus defaulting to a suboptimal generalist approach.

2. **Model Insufficiency:** Compounding this heuristic failure, the underlying 3-type model itself lacks the necessary granularity identified by our refined 4-type taxonomy. Even with perfect heuristics, the model has no categories corresponding to 'Holistic Abstract Reasoning' or 'Local Geometric Analysis,' forcing the agent to either misclassify these problems or trigger the default deployment of the generalist `Static_Heterogeneous` framework instead of the optimal specialist architectures (`Single_Agent` or `Static_Homogeneous`, respectively).

This profound misdiagnosis led to the deployment of architectures cognitively ill-equipped for the tasks. These suboptimal frameworks then became susceptible to the "local optima trap" created by the multi-objective fitness function (originally detailed in Section 4.4.8). While effective on the training dataset, the fitness function's reliance on a simple weighted average systematically rewarded solutions that were perceptually similar but logically incorrect when applied by an inappropriate architecture. This critical failure, therefore, was not an endpoint but a crucial diagnostic finding that directly motivated the subsequent abductive leap and methodological refinements. It proves that the failure was not just in the execution (fitness function) but originated in the diagnostic coarseness of both the agent's heuristics and its underlying strategic model, providing the final piece of evidence that a deeper, more refined understanding of the problem space is essential for successful strategic deployment.

6.3. Abductive Leap #3: A Hierarchical Framework to Combat Generative Exhaustion

Faced with the universal failure on the evaluation set, the research entered its final abductive cycle. The "inference to the best explanation" for the 'local optima trap' was that the dual-layer failure identified in the preceding analysis—a profound **strategic failure** rooted in both inaccurate heuristics and an insufficient problem taxonomy, compounded by a **tactical failure** of a permissive fitness function—led to a more fundamental limitation in the agentic process itself: **Generative Exhaustion**. This concept describes the state where an agent, having been incorrectly deployed for a task it is cognitively ill-equipped to handle, is unable to generate a candidate solution that is fundamentally sound enough to pass basic logical checks, or it generates plausible solutions but lacks the final inventive step to achieve pixel-perfect accuracy.

This abductive leap led to the development of two critical methodological refinements **detailed in Section 4.9**: the Gated Scoring Framework and the Self-Correction Loop. They were designed to combat this exhaustion by advancing the system's cognitive architecture. The first refinement, the **Gated Scoring Framework**, was designed to fix the immediate evaluation flaw. This mechanism enforces a stricter evaluation hierarchy, prioritizing fundamental logical correctness (e.g., object count and structure integrity) over superficial perceptual similarity. It acts as a pre-emptive check, applying a significant penalty to any solution that fails these critical logical gates, thereby providing a more discerning feedback signal to the search algorithm.

The second and more advanced refinement completed the conceptualization of the **Hi-**

erarchical OODA-Belbin Framework by implementing its tactical layer: the **Self-Correction Loop**. This artifact transforms the agents from simple generators into more sophisticated reasoners capable of iterative refinement. It operationalizes a tactical "micro-loop," **limited to a maximum of three correction attempts within a single MCTS expansion step**, allowing an agent to learn from its own immediate failures. The process is structured as a rapid OODA cycle:

- **Observe:** The agent gathers the full diagnostic feedback from a failed attempt, including the high-dimensional fitness vector and the specific gate status.
- **Orient:** The agent is prompted to analyze its previously generated code in light of this specific failure signature, forcing a "Chain-of-Thought" analysis of its error.
- **Decide:** Based on this orientation, the agent formulates a correction strategy. A forceful "meta-cognitive" prompt is triggered on repeated failures to break cognitive fixation.
- **Act:** The agent generates a new, corrected version of its solution, which is then immediately re-evaluated.

This methodological enhancement, born from the final abductive leap, shifts the research focus from the passive evaluation of generated artifacts to the active and intelligent guidance of the generative process itself.

6.4. Final Deductive Validation: Achieving a Hard-Won Success on Intractable "Trap" Problems

The final stage of the research was a rigorous deductive test of the most advanced version of the system—the complete Hierarchical OODA-Belbin Framework with its Gated Scoring and Self-Correction Loop **(see Section 4.9 for implementation details)**. The objective was to assess whether this new artifact could overcome the "local optima trap" and solve the problems identified as intractable in Anomaly #3. To account for the potential impact of LLM stochasticity, the experiment was executed five times under identical, deterministic conditions against a curated set of 20 "trap" problems.

Table 6

Summary of Solved Problems Across Five Experimental Runs

Run	**Problems Solved**	**Success Rate**	**Solved Problem IDs**
1	1	5%	1818057f
2	0	0%	—
3	0	0%	—
4	1	5%	9bbf930d
5	1	5%	7b5033c1
Average	**0.6**	**3%**	

Aggregate Performance and the Pervasiveness of "False Positives". Across the five independent runs, the system demonstrated a tangible, albeit inconsistent, capability to solve these highly complex challenges. As detailed in Table 6, the framework achieved a 'perfect pixel match' on a total of three unique problems: `1818057f`, `9bbf930d`, and `7b5033c1`. The resulting average success rate across all runs was 3%.

However, a rigorous code audit, consistent with the methodology of this paper, revealed a critical distinction. Of the three apparent successes, two were "false positives" where the correct pixel output was achieved through flawed, non-generalizable logic. Only one solution—the solve for problem `7b5033c1` —was a **'real success'** with robust, optimal, and generalizable code.

This outcome, while modest, is highly significant: it represents the **first-ever, robust successful solve** on this class of previously unsolvable problems, providing a hard-won

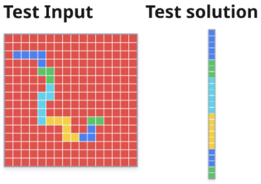

Figure 12

Problem 7b5033c1 (To solve it, the agent must recognize the line of pixels inside the square and output it as a straight line).

validation of the final artifact's effectiveness (See Figure 12). Simultaneously, it serves as a final, powerful warning of the pervasive danger of 'false positives' in AI evaluation.

Qualitative Analysis of Agentic Behavior. A granular analysis of the experimental logs reveals the mechanistic reasons for both the successes and failures, providing a deep insight into the new framework's capabilities and limitations.

Success via Structured Refinement. The most compelling evidence for the new framework's effectiveness comes from the single, robustly solved problem, `7b5033c1`. The logs for this successful run show a clear pattern of structured, iterative refinement. The agent's initial attempt failed with a `GATE_1_FAILED` error. However, after being presented with the **"Chain-of-Thought" correction prompt (see example in Section 4.9.2)**, the agent was able to correctly analyze its failure, formulate a new plan, and generate a pixel-perfect and logically sound solution on its second attempt. This pattern provides direct validation that the Self-Correction Loop can enable agents to overcome initial logical flaws and achieve a true, generalizable solution.

Failure Mode 1: Persistent Cognitive Inertia. The most common failure mode observed was Persistent Cognitive Inertia. This occurs when an agent, despite being given the forceful "meta-cognitive" prompt to abandon its flawed strategy, generates a new solution that is only a minor variation of its failed attempt. This demonstrates that while the prompt to "think differently" is correctly delivered, the agent's generative process can remain anchored to its initial, flawed premise.

Failure Mode 2: The "Whack-a-Mole" Problem. A more nuanced failure pattern occurs when an agent successfully corrects one error but introduces a new, often more subtle, one in the process. For example, an agent would successfully pass the logical gates, but in its subsequent attempt to achieve pixel-perfection, its correction would break the object's core structure, causing the new solution to fail a gate. This mirrors a complex, human-like debugging challenge where local optimizations have unintended negative consequences on the global solution.

The Impact of LLM Stochasticity and Creative Variance. The high degree of variance in the results, observed under deterministic settings, is a direct manifestation of LLM stochasticity. This finding does not detract from the study's success but rather enriches it, providing a powerful parallel to the paper's central theme. The observed variance mirrors the creative variance and unpredictability inherent in human collaboration. Just as two identical human teams are not guaranteed to produce the same solution to a "wicked problem," the agentic framework exhibits a similar emergent property. This suggests that in building this complex system, we have created a computational artifact that models not only the strengths of human team synergy but also its characteristic unpredictability.

Final Synthesis: The Tactical Imperative as the Ultimate Validation. The hard-won, robust success of the `Self-Correction Loop` on one previously intractable "trap" problem, set against the backdrop of the system still producing "false positives," is the final and most powerful validation of our paper's central thesis. The initial experiment in this cycle proved that a high-level strategic choice, when based on an incomplete model of the problem space, is insufficient. The only path to a *true, verifiable* success was through a lower-level, tactical, and diagnostic-driven process that could iteratively refine its own logic.

This provides the ultimate, practical proof for the necessity of both the refined, four-category taxonomy and a rigorous, code-auditing methodology. The failure of the initial strategic deployment and the subsequent, singular success of the tactical loop demonstrate that a deep, nuanced understanding of the problem space is an operational imperative for achieving robust AI reasoning. It confirms that the future of agentic problem-solving lies in hierarchical frameworks that possess both the strategic wisdom to select an approach and the tactical intelligence to correct and validate the solutions that emerge.

7. Discussion

The results of our comprehensive analysis, which moved beyond simple success metrics to a deep audit of solution quality, provide a powerful validation for a contingency-based model of agentic problem-solving. The quantitative finding that raw success rates were not statistically significant served as a crucial signpost, indicating that the most meaningful differences between the experimental conditions lie not in their aggregate performance, but in their distinct cognitive styles. This discussion synthesizes our qualitative findings to construct a detailed explanation for these differences, connects them to established theoretical frameworks, and culminates in a refined theoretical model of the ARC problem space that justifies a hierarchical approach to agentic design.

7.1. Interpreting the Contingency Model: From Cognitive Profiles to Validated Hypotheses

The qualitative analysis allows us to resolve the quantitative paradoxes by constructing a detailed "cognitive profile" for each experimental condition (See Figure 13). These profiles, which synthesize the findings from our "Taxonomy of Failure" and "Anatomy of Success," provide the direct, evidence-based validation (and in one case, falsification) of our three core hypotheses.

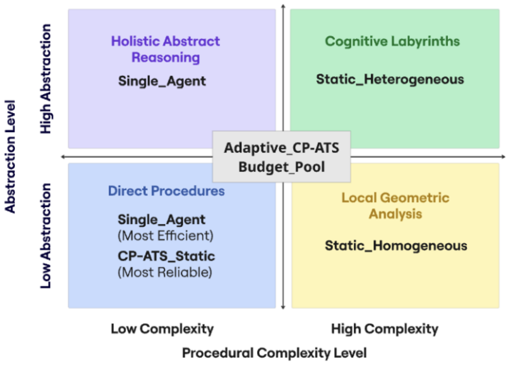

Figure 13

Agentic Specialization Matrix

The Primacy of Simplicity. The Simplicity Hypothesis was unequivocally supported. Our finding that the `Baseline_Single_Agent` is the uniquely capable specialist for problems requiring Holistic Abstract Reasoning provides a crucial empirical grounding for the concept of cognitive and computational overhead discussed in the literature (Schimmelpfennig et al., 2021). For tasks where the logical path is singular, the communication and coordination required for teamwork introduce friction and potential for misinterpretation (Chen, Liu, et al., 2019; Galinsky et al., 2015), making a focused, monolithic approach more effective. This result aligns perfectly with the contingency thesis, which posits that for straightforward or less complex tasks, the performance benefits of cognitive diversity can be negative (Dinwoodie, 2005; Higgs et al., 2005).

Case Study: The Cost of Complexity. The superior efficiency of simpler architectures on straightforward tasks is qualitatively evident in the code generated for 'Direct Procedure Task' `1e0a9b12`. The data reveals that while all six conditions achieved a 'real success,' their solution quality differed dramatically. The `Baseline_Single_Agent` and `CP-ATS_Static` frameworks produced code that our audit classified as 'Optimal & Parsimonious.' In stark contrast, the four other frameworks, including both complex adaptive systems, produced solutions that were 'Correct but Inefficient & Convoluted.' This direct comparison demonstrates that the overhead of multi-agent coordination and adaptation led to less efficient reasoning paths, providing powerful, qualitative evidence for the 'cost of complexity' on tasks that do not require it.

The "Expert Anomaly". The *"Expert Anomaly" Hypothesis* was also definitively validated. The unique success of the `Baseline_Static_Homogeneous` condition on high-precision "Meticulous Execution" tasks validates the **"double-edged sword" theory of diversity** (Martins et al., 2012). For problems that demand high-fidelity procedural replication, the cognitive diversity that is beneficial for ideation becomes a liability. The absence of competing perspectives in the homogeneous team ensures a focused, unwavering execution, a clear instance where the moderating role of task complexity makes homogeneity the superior approach (Dinwoodie, 2005; Higgs et al., 2005; S. Wang et al., 2019), particularly for tasks primarily focused on implementation rather than strategy (S. Wang et al., 2019).

Case Study: The Specialist's Algorithmic Purity. The unique expertise of the homogeneous team is powerfully demonstrated in the solutions for 'Local Geometric Analysis' task `08ed6ac7`. Our audit revealed that the specialist for this archetype, the `Baseline_Static_Homogeneous` team, was one of only two conditions to produce an 'Optimal & Robust' solution. The `Baseline_Static_Heterogeneous` team, while also successful, produced a solution that was merely 'Correct but Inefficient.' This case provides definitive evidence for the "Expert Anomaly," showing that for 'Local Geometric Analysis' tasks, the focused cognitive style of the homogeneous team leads to a more algorithmically pure and efficient solution than a more diverse team.

Falsification of the Adaptation Hypothesis. The most profound finding of this study is the decisive falsification of the *Adaptation Hypothesis.* The initial prediction that the most complex adaptive framework (`Adaptive_CP-ATS_Budget_Pool`) would master "Cognitive Labyrinths" was proven false. Instead, this condition was the poorest performer, providing strong evidence for a **"cost of complexity."** The true master of these problems was the cognitively diverse `Baseline_Static_Heterogeneous` team. This finding suggests that for the most wicked problems, which cannot be solved by a single specialist, the key is not architectural complexity but the synthesis of multiple, distinct cognitive styles within a stable team (Aminpour, Gray, et al., 2021; Hong & Page, 2001, 2004).

Case Study: The Failure of Complexity. The qualitative data provides the definitive explanation for the falsification of the Adaptation Hypothesis. On the difficult 'Cognitive Labyrinth' problem `19bb5feb`, the true specialist, the `Baseline_Static_Heterogeneous` team, was the only framework to produce a 'real success,' which was audited as 'Optimal & Robust.' In stark contrast, both the `Adaptive_CP-ATS_Fixed_Cycle` and `Adaptive_CP-ATS_Budget_Pool` frameworks failed on this problem, generating solutions that were audited as 'Brittle & Overly Complex.' This is not merely a difference in efficiency between two successes; it is a demonstration of a fundamental difference in reasoning capability. It provides the strongest possible evidence that for the most wicked problems, the stable synthesis of diverse cognitive styles is a more effective path to a robust solution than a complex adaptive architecture burdened by its own overhead.

7.2. A Refined Taxonomy of ARC Problems: The Key Theoretical Insight

The validation and falsification of our hypotheses force a refinement of our initial three-category problem taxonomy into a more

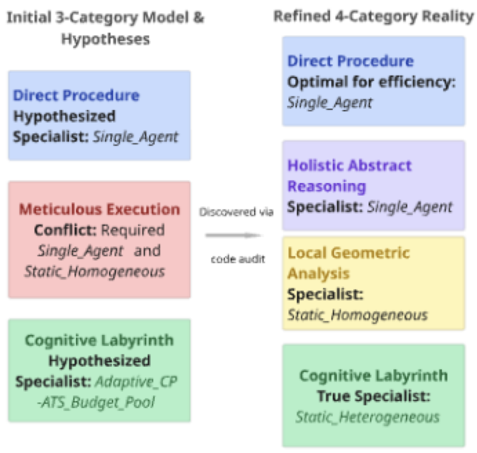
Initial 3-Category Model & Hypotheses
Refined 4-Category Reality
Direct Procedure
Hypothesized Specialist: Single_Agent
Meticulous Execution
Conflict: Required Single_Agent and Static_Homogeneous
Cognitive Labyrinth
Hypothesized Specialist: Adaptive_CP-ATS_Budget_Pool
Discovered via
code audit
Direct Procedure
Optimal for efficiency: Single_Agent
Holistic Abstract Reasoning
Specialist: Single_Agent
Local Geometric Analysis
Specialist: Static_Homogeneous
Cognitive Labyrinth
True Specialist: Static_Heterogeneous

Figure 14

The Refinement of the ARC Problem Taxonomy

precise, four-category model that is the central theoretical contribution of this research (See Figure 14):

1. **Direct Procedure Tasks:** Problems solvable by a single, clear, deterministic rule where all conditions are competent.

2. **Local Geometric & Procedural Analysis:** A refinement of "Meticulous Execution," these tasks require a detailed, procedural analysis of a pixel's immediate neighbors. Our evidence shows `Baseline_Static_Homogeneous` teams excel here.

3. **Holistic Abstract Reasoning:** A new category, formerly hidden within "Meticulous Execution," these tasks require a single, non-obvious "big picture" insight. Our evidence proves the `Baseline_Single_Agent` is the unique specialist for this archetype.

4. **Cognitive Labyrinths:** The most algorithmically demanding problems, requiring the synthesis of a non-trivial, multi-step algorithm. Our evidence proves that the cognitively diverse `Baseline_Static_Heterogeneous` team is the master of this domain.

7.3. The Duality of "Meticulous Execution": The Ultimate Justification for Hierarchy

The explanatory power of this refined taxonomy is most profound when we re-examine the initial "Meticulous Execution" category. Its splitting into two distinct types—*Local Geometric Analysis* and *Holistic Abstract Reasoning*—is the central discovery of our research. It perfectly explains the paradoxical performance of the specialized conditions. It reveals why the "Monolithic Reasoner" (`Baseline_Single_Agent`) sometimes dramatically outperformed complex teams, and why the "Meticulous Executor" (`Baseline_Static_Homogeneous`) could succeed where more creative architectures failed. This discovery provides the definitive, evidence-based justification for the paper's central thesis: **a high-level Strategic Selector is essential,** as no single baseline condition can be optimal for both sub-types of the original "Meticulous Execution" category.

7.4. Broader Implications and Emergent Properties

Our findings carry several broad implications for the field. First, the consistent underperformance of our most complex adaptive system provides strong empirical evidence for the "cost of complexity," challenging the prevailing assumption that greater architectural complexity is the optimal path to solving wicked problems.

Second, the prevalence of "false positives," "ghost problems," and other anomalies revealed only through our rigorous code-auditing process serves as a powerful methodological warning. It highlights the critical need for the field to move beyond a reliance on simple, output-based benchmarks like a 'perfect pixel match' and toward more sophisticated evaluation regimes that include the validation of a solution's logical and algorithmic integrity.

Third, the final "Frontier Challenge" experiments revealed a critical limitation we term **"Generative Exhaustion."** The observed failure modes of **"Cognitive Inertia"** and the **"Whack-a-Mole" problem** represent the current frontier of this challenge (J. Lu et al., 2025), suggesting that for the most difficult problems, the solution space is either too vast or the required logical leap is too great for current refinement processes to navigate effectively (Aghzal et al., 2025; Lin et al., 2024; Xie et al., 2024).

Finally, the high variance in performance across multiple runs highlights the profound impact of **LLM stochasticity**. In setting out to mimic **human multicultural teams** (Bhalerao et al., 2025; Hong & Page, 2004), we created a system that also mimics their characteristic **creative variance and unpredictability**. This suggests we have created not a deterministic machine, but a computational artifact that models not only the strengths of human synergy but also its inherent unpredictability. This aligns perfectly with contingency theories, which posit that the optimal team structure is moderated by the nature of the task itself (Higgs et al., 2005; Zabalandikoetxea et al., 2021), and provides a clear direction for future work: enhancing the robustness and reasoning of the agents *within* a portfolio-based system (B. Liu et al., 2023).

8. Conclusion

This research set out to challenge the prevailing assumption that architectural complexity is the universal solution to the complexity of "wicked problems." Through a systematic, empirical comparison of six distinct agentic conditions, supplemented by a rigorous audit of solution quality that moved beyond the standard 'perfect pixel match' metric, we have provided a definitive, evidence-based answer. The findings unequivocally refute a monolithic approach to agentic design and validate a more nuanced, contingency-based model of performance.

Our central contribution is the articulation of this contingency model, a discovery made possible by the development of the refined, four-category problem taxonomy that now forms its evidence-based foundation. Our research has demonstrated that for each archetype, a different agentic strategy is optimal: (1) for *Direct Procedure Tasks*, where all conditions are competent, a low-overhead single agent is the most efficient choice; (2) for problems of *Holistic Abstract Reasoning*, that same monolithic agent is uniquely capable; (3) for tasks requiring *Local Geometric & Procedural Analysis*, a non-diverse, homogeneous team proves the "Expert Anomaly"; and (4) for the most algorithmically demanding *Cognitive Labyrinths*, the key to success is the *cognitive diversity* of a heterogeneous team. This is starkly contrasted by the consistent underperformance of our most complex adaptive system, which provides strong empirical evidence for a "cost of complexity."

This study also serves as a candid exploration of the current frontiers and limitations of agentic AI. The achievement of only a single, robust, generalizable success on the most difficult "trap" problems, even with our most advanced Self-Correction Loop, underscores the profound challenge of "Generative Exhaustion." Furthermore, the high variance in results due to LLM stochasticity revealed an inherent and fascinating property of these systems. Far from being a flaw, this variance mirrors the creative unpredictability of human collaboration itself, suggesting that in attempting to mimic human synergy, we have created an artifact that also models its characteristic unpredictability.

Ultimately, the discovery that our initial "Meticulous Execution" category was a duality of two distinct problem types requiring opposite solutions provides the most powerful justification for a hierarchical, portfolio-based approach to AI problem-solving. It proves that the highest form of intelligence in this domain lies not in any single architecture, but in the strategic capacity to correctly diagnose a problem and deploy the most appropriate specialist. The future of AGI, therefore, lies in the design of such hierarchical systems. As demonstrated in our final experiment, the artifact that operationalized our initial three-category model proved insufficient precisely because of its limited taxonomy. A critical and direct direction for **future work**, therefore, is to implement a new Strategic Selector based on the refined, four-category taxonomy discovered herein A parallel and equally vital research avenue is to automate the solution validity check by integrating the principles of our code audit directly into the framework's tactical layer. This, combined with a fundamental re-evaluation of how we measure success—shifting away from superficial, output-based metrics and toward a more rigorous, logic-based validation of an agent's reasoning process—charts a clear path forward for the field.

References

References

Abbena, E., Salamon, S., & Gray, A. (2017, September). *Modern Differential Geometry of Curves and Surfaces with Mathematica* (3rd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315276038

Abu, O., Gerstgrasser, M., Rosenschein, J. S., & Keren, S. (2021). Collaboration promotes group resilience in multi-agent AI [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2111.06614

Aggarwal, I., & Woolley, A. W. (2013). Do you see what I see? The effect of members' cognitive styles on team processes and errors in task execution [Publisher: Elsevier BV]. *Organizational Behavior and Human Decision Processes*, *122*(1), 92–99. https://doi.org/10.1016/j.obhdp.2013.04.003

Aggarwal, I., & Woolley, A. W. (2018). Team creativity, cognition, and cognitive style diversity [Publisher: Institute for Operations Research and the Management Sciences]. *Management Science*, *65*(4), 1586–1599. https://doi.org/10.1287/mnsc.2017.3001

Aghzal, M., Plaku, E., Stein, G., & Yao, Z. (2025). A survey on large language models for automated planning [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.12435

Aminpour, P., Gray, S. A., Singer, A., Scyphers, S. B., Jetter, A., Jordan, R., Murphy, R., & Grabowski, J. H. (2021). The diversity bonus in pooling local knowledge about complex problems [Publisher: National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *118*(5). https://doi.org/10.1073/pnas.2016887118

Aminpour, P., Schwermer, H., & Gray, S. (2021). Do social identity and cognitive diversity correlate in environmental stakeholders? A novel approach to measuring cognitive distance within and between groups [Publisher: Public Library of Science]. *PLoS ONE*, *16*(11). https://doi.org/10.1371/journal.pone.0244907

Anderson, A. W., Li, T., Vorvoreanu, M., & Burnett, M. (2021). Human-AI interaction for diverse humans: What cognitive style disaggregation reveals [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2108.00588

Andrews, R. W., Lilly, J. M., Srivastava, D., & Feigh, K. M. (2022, April). The role of shared mental models in human-AI teams: A theoretical review [ISSN: 1463-922X, 1464-536X Issue: 2 Pages: 129-175 Volume: 24]. https://doi.org/10.1080/1463922x.2022.2061080

Anik, M. A., Rahman, A., Wasi, A. T., & Ahsan, M. M. (2025). Preserving cultural identity with context-aware translation through multi-agent AI systems, 51–60. https://doi.org/10.18653/v1/2025.lm4uc-1.7

Aritzeta, A., Swailes, S., & Senior, B. (2007). Belbin's team role model: Development, validity and applications for team building* [Publisher: Wiley]. *Journal of Management Studies*, *44*(1), 96–118. https://doi.org/10.1111/j.1467-6486.2007.00666.x

Armstrong, S. J. (2000). The Influence of Individual Cognitive Style on Performance in Management Education [Publisher: Routledge _eprint: https://doi.org/10.1080/01443410075018020]. *Educational Psychology*, *20*(3), 323–339. https://doi.org/10.1080/014434100750018020

Belbin, R., & Brown, V. (2012). Team roles at work. Routledge.

Bhalerao, P., Yalamarty, M., Trinh, B., & Ignat, O. (2025). Multi-agent multimodal models for multicultural text to image generation [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2502.15972

Bickmore, T., & Gruber, A. J. (2010). Relational agents in clinical psychiatry [Publisher: Lippincott Williams & Wilkins]. *Harvard Review of Psychiatry*, *18*(2), 119–130. https://doi.org/10.3109/10673221003707538

Boroomand, A., & Smaldino, P. E. (2021). Hard work, risk-taking, and diversity in a model of collective problem solving [Publisher: University of Surrey]. *Journal of Artificial Societies and Social Simulation*, *24*(4). https://doi.org/10.18564/jasss.4704

Boyd, J. R. (1986). Patterns of conflict. *Unpublished Paper, December*.

Çelikok, M. M., Peltola, T., Daee, P., & Kaski, S. (2019). Interactive AI with a theory of mind [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.1912.05284

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., & Stoica, I. (2025). Why do multi-agent LLM systems fail? [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2503.13657

Chen, X., Liu, J., Zhang, H., & Kwan, H. K. (2019). Cognitive diversity and innovative work behaviour: The mediating roles of task reflexivity and relationship conflict and the moderating role of perceived support [Publisher: Wiley]. *Journal of Occupational and Organizational Psychology*, *92*(3), 671–694. https://doi.org/10.1111/joop.12259

Chen, X., Zhang, P., Du, G., & Li, F. (2019). A distributed method for dynamic multi-robot task allocation problems with critical time constraints [Publisher: Elsevier BV]. *Robotics and Autonomous Systems*, *118*, 31–46. https://doi.org/10.1016/j.robot.2019.04.012

Chhikara, G., Kumar, A., & Chakraborty, A. (2025). Through the prism of culture: Evaluating llms' understanding of indian subcultures and traditions [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.16748

Chollet, F. (2019). On the measure of intelligence [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.1911.01547

Chollet, F., Knoop, M., Kamradt, G., & Landers, B. (2024). ARC prize 2024: Technical report [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.04604

Chollet, F., Knoop, M., Kamradt, G., Landers, B., & Pinkard, H. (2025). ARC-AGI-2: A new challenge for frontier AI reasoning systems [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2505.11831

Colas, C., Karch, T., Moulin-Frier, C., & Oudeyer, P.-Y. (2022). Language and culture internalization for human-like autotelic AI [Publisher: Nature Portfolio]. *Nature Machine Intelligence*, *4*(12), 1068–1076. https://doi.org/10.1038/s42256-022-00591-4

Cole, J., & Osman, M. (2025). Don't throw the baby out with the bathwater: How and why deep learning for ARC [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2506.14276

De Dreu, C. K. W., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis [Publisher: American Psychological Association]. *Journal of Applied Psychology*, *88*(4), 741–749. https://doi.org/10.1037/0021-9010.88.4.741

DeChant, C. (2025). Episodic memory in AI agents poses risks that should be studied and mitigated, 321–332. https://doi.org/10.1109/satml64287.2025.00024

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2021). The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2105.03354

Dinwoodie, D. (2005). Solving the dilemma: A leader's guide to managing diversity [Publisher: Wiley]. *Leadership in Action*, *25*(2), 3–6. https://doi.org/10.1002/lia.1107

Dong, F., Peng, J., Wang, X., & Tang, M. (2021). The development and validation of a cognitive diversity scale for chinese academic research teams [Publisher: Frontiers Media]. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.687179

Dusparić, I., & Cardozo, N. (2021). Adaptation to unknown situations as the holy grail of learning-based self-adaptive systems: Research directions [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2103.06908

Feng, S., Chan, W.-C., Chouhan, S., Ayala, J. F. G., Medicherla, S., Clark, K., & Shi, M. (2025). Whispers of many shores: Cultural alignment through collaborative cultural expertise [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2506.00242

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. [Publisher: American Psychological Association]. *Journal of Experimental Psychology General*, *144*(3), 674–687. https://doi.org/10.1037/xge0000070

Fisher, S., Hunter, T., & MacRosson, W. (2001). A validation study of Belbin's team roles [Publisher: Routledge _eprint: https://doi.org/10.1080/13594320143000591]. *European Journal of Work and Organizational Psychology*, *10*(2), 121–144. https://doi.org/10.1080/13594320143000591

Fu, Q., Ai, X., Yi, J., Qiu, T., Yuan, W., & Pu, Z. (2023). Learning heterogeneous agent cooperation via multiagent league training [Publisher: Elsevier BV]. *IFAC-PapersOnLine*, *56*(2), 3033–3040. https://doi.org/10.1016/j.ifacol.2023.10.1431

Fürnkranz, J., Hüllermeier, E., Mencía, E. L., & Rapp, M. (2020). Learning structured declarative rule sets – a challenge for deep discrete learning [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2012.04377

Galinsky, A. D., Todd, A. R., Homan, A. C., Phillips, K. W., Apfelbaum, E. P., Sasaki, S. J., Richeson, J. A., Olayon,

J. B., & Maddux, W. W. (2015). Maximizing the gains and minimizing the pains of diversity [Publisher: SAGE Publishing]. *Perspectives on Psychological Science*, *10*(6), 742–748. https://doi.org/10.1177/1745691615598513

Geffner, H. (2018). Model-free, model-based, and general intelligence, 10–17. https://doi.org/10.24963/ijcai.2018/2

Goebel, K., & Zips, P. (2025). Can LLM-reasoning models replace classical planning? A benchmark study [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2507.23589

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact [Publisher: MIS Quarterly]. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.25300/misq/2013/37.2.01

Gruetzemacher, R. (2018). Rethinking AI strategy and policy as entangled super wicked problems, 122–122. https://doi.org/10.1145/3278721.3278746

Gunasekaran, A., & Ngai, E. W. (2004, December). Build-to-order supply chain management: A literature review and framework for development [ISSN: 0272-6963, 1873-1317 Issue: 5 Pages: 423-451 Volume: 23]. https://doi.org/10.1016/j.jom.2004.10.005

Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., & He, C. (2024). LLM multi-agent systems: Challenges and open problems [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2402.03578

Hariri, Y. (2025). *LLM-Creativity: Measuring and Explaining Creativity in Large Language Models: A Mixed-Methods Investigation into the Effect of Cultural Persona Prompts on Business Solution Generation* [Doctoral dissertation]. https://doi.org/10.5281/zenodo.17407805

Head, B., & Alford, J. (2013). Wicked problems [Publisher: SAGE Publishing]. *Administration & Society*, *47*(6), 711–739. https://doi.org/10.1177/0095399713481601

Heidari, H., Barocas, S., Kleinberg, J., & Levy, K. (2023). Informational diversity and affinity bias in team growth dynamics, 1–10. https://doi.org/10.1145/3617694.3623238

Hevner, March, M., Park, J., & Ram, R. (2004). Design science in information systems research [Publisher: MIS Quarterly]. *MIS Quarterly*, *28*(1), 75–75. https://doi.org/10.2307/25148625

Higgs, M., Plewnia, U., & Ploch, J. (2005). Influence of team composition and task complexity on team performance [Publisher: Emerald Publishing Limited]. *Team Performance Management*, *11*, 227–250. https://doi.org/10.1108/13527590510635134

Hoang, T. N., Xiao, Y., Sivakumar, K., Amato, C., & How, J. P. (2017). Near-optimal adversarial policy switching for decentralized asynchronous multi-agent systems [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.1710.06525

Hoeft, R. M., Jentsch, F., Smith-Jentsch, K. A., & Bowers, C. (2005). Exploring the role of shared mental models for implicit coordination in teams [Publisher: SAGE Publishing]. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(21), 1863–1867. https://doi.org/10.1177/154193120504902110

Hong, L., & Page, S. E. (2001). Problem solving by heterogeneous agents [Publisher: Elsevier BV]. *Journal of Economic Theory*, *97*(1), 123–163. https://doi.org/10.1006/jeth.2000.2709

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers [Publisher: National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Horwitz, S. K., & Horwitz, I. B. (2007a). The Effects of Team Diversity on Team Outcomes: A Meta-Analytic Review of Team Demography [Publisher: SAGE Publications Inc]. *Journal of Management*, *33*(6), 987–1015. https://doi.org/10.1177/0149206307308587

Horwitz, S. K., & Horwitz, I. B. (2007b, November). The effects of team diversity on team outcomes: A meta-analytic review of team demography [ISSN: 0149-2063, 1557-1211 Issue: 6 Pages: 987-1015 Volume: 33]. https://doi.org/10.1177/0149206307308587

Hu, Y., Stegner, L., & Mutlu, B. (2023a). Knowing who knows what: Designing socially assistive robots with transactive memory system [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2305.05115

Hu, Y., Stegner, L., & Mutlu, B. (2023b). Knowing who knows what: Designing socially assistive robots with transactive memory system [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.05115

Hughes, L., Dwivedi, Y. K., Malik, T., Shawosh, M., Albashrawi, M., Jeon, I., Dutot, V., Appanderanda, M., Crick, T., Dé, R., Fenwick, M., Gunaratnege, S. M., Jurčys, P., Kar, A. K., Kshetri, N., Li, K., Mutasa, L. S., Samothrakis, S., Wade, M., & Walton, P. (2025). AI agents and agentic systems: A multi-expert analysis [Publisher: Taylor & Francis]. *Journal of Computer Information Systems*, 1–29. https://doi.org/10.1080/08874417.2025.2483832

i Klett, T. C., & Arnulf, J. K. (2020, August). Are chinese teams like western teams? Indigenous management theory to leapfrog essentialist team myths [ISSN: 1664-1078 Volume: 11]. https://doi.org/10.3389/fpsyg.2020.01758

Inoue, Y., Misaki, K., Imajuku, Y., Kuroki, S., Nakamura, T., & Akiba, T. (2025, June). Wider or deeper? Scaling LLM inference-time compute with adaptive branching tree search.

JACCARD, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, *37*, 547–579. Retrieved September 10, 2025, from https://cir.nii.ac.jp/crid/1573387450552842240

Ji, M., Azuma, S.-i., & Egerstedt, M. (2006). Role-assignment in multi-agent coordination. https://smartech.gatech.edu/handle/1853/38457

Jiang, L., & Wang, Y. (2019). Respect your emotion: Human-multi-robot teaming based on regret decision model. *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, 936–941. https://doi.org/10.1109/coase.2019.8843206

Jin, C., Zhang, S., Shu, T., & Cui, Z. (2023). The cultural psychology of large language models: Is ChatGPT a holistic or analytic thinker? [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.14242

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, *48*(4), 63–85. https://doi.org/10.1007/BF02300500

Jonassen, D. H. (2009). Learning to solve problems: An instructional design guide [Publisher: Taylor & Francis]. *Gifted and Talented International*, *24*(2), 153–154. https://doi.org/10.1080/15332276.2009.11673538

Kamali, K., Fan, X., & Yen, J. (2006). Multiparty proactive communication: A perspective for evolving shared mental models, 685–690. https://www.aa ai.org/Papers/AAAI/2006/AAAI06-109.pdf

Kegenbekov, Z., & Jackson, I. (2021). Adaptive supply chain: Demand–supply synchronization using deep reinforcement learning [Publisher: Multidisciplinary Digital Publishing Institute]. *Algorithms*, *14*(8), 240–240. https://doi.org/10.3390/a14080240

Ki, D., Rudinger, R., Zhou, T., & Carpuat, M. (2025). Multiple LLM agents debate for equitable cultural alignment [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2505.24671

Kirton, M. (1976). Adaptors and innovators: A description and measure [Place: US Publisher: American Psychological Association]. *Journal of Applied Psychology*, *61*(5), 622–629. https://doi.org/10.1037/0021-9010.61.5.622

Kolb, D. (1976). Management and the learning process [Publisher: SAGE Publishing]. *California Management Review*, *18*(3), 21–31. https://doi.org/10.2307/41164649

Kostka, A., & Chudziak, J. A. (2025). Towards cognitive synergy in LLM-based multi-agent systems: Integrating theory of mind and critical evaluation [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2507.21969

Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F., & Oudeyer, P.-Y. (2023). Large language models as superpositions of cultural perspectives [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.07870

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams [Publisher: SAGE Publishing]. *Psychological Science in the Public Interest*, *7*(3), 77–124. https://doi.org/10.1111/j.1529-1006.2006.00030.x

Lau, H. C., Agussurja, L., & Thangarajoo, R. (2007). Real-time supply chain control via multi-agent adjustable autonomy [Publisher: Elsevier BV]. *Computers & Operations Research*, *35*(11), 3452–3464. https://doi.org/10.1016/j.cor.2007.01.027

Laxman, K. (2011). Cognitive learning processes undergirding design-based ill-structured problem solving [Publisher: Inderscience Publishers]. *International Journal of Innovation and Learning*, *11*(1), 60–60. https://doi.org/10.1504/ijil.2012.044329

Lee, J. Y., Burton, H., & Lallemant, D. (2018). Adaptive decision framework for civil infrastructure exposed to evolving risks [Publisher: Elsevier BV]. *Procedia Engineering*, *212*, 435–442. https://doi.org/10.1016/j.proeng.2018.01.056

Lhaksmana, K. M., Murakami, Y., & Ishida, T. (2018). Role-based modeling for designing agent behavior in self-organizing multi-agent systems [Publisher: World Scientific]. *International Journal of Software Engineering and Knowledge Engineering*, *28*(1), 79–96. https://doi.org/10.1142/s0218194018500043

Li, C., Wang, W., Zheng, T., & Song, Y. (2025). Patterns over principles: The fragility of inductive reasoning in llms under noisy observations [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2502.16169

Lim, S., Lee, S., Min, D., & Yu, Y. (2025). Persona dynamics: Unveiling the impact of personality traits on agents in text-based games [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2504.06868

Lin, F., Malfa, E. L., Hofmann, V., Yang, E. M., Cohn, A. G., & Pierrehumbert, J. B. (2024). Graph-enhanced large language models in asynchronous plan reasoning [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2402.02805

Linkov, I., Satterstrom, F. K., Kiker, G. A., Batchelor, C. J., Bridges, T. S., & Ferguson, E. (2006, August). From comparative risk assessment to multi-criteria decision analysis and adaptive management: Recent developments and applications [ISSN: 0160-4120, 1873-6750 Issue: 8 Pages: 1072-1093 Volume: 32]. https://doi.org/10.1016/j.envint.2006.06.013

Liu, B., Mazumder, S., Robertson, E., & Grigsby, S. S. (2023). AI autonomy: Self-initiated open-world continual learning and adaptation [Publisher: Association for the Advancement of Artificial Intelligence]. *AI Magazine*, *44*(2), 185–199. https://doi.org/10.1002/aaai.12087

Liu, H.-Y., & Maas, M. M. (2021). 'Solving for X?' Towards a problem-finding framework to ground long-term governance strategies for artificial intelligence [Publisher: Elsevier BV]. *Futures*, *126*, 102672–102672. https://doi.org/10.1016/j.futures.2020.102672

Liu, J., Zhu, Y., & Wang, H. (2023). Managing the negative impact of workforce diversity: The important roles of inclusive HRM and employee learning-oriented behaviors [Publisher: Frontiers Media]. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1117690

Liu, X., Guo, D., Zhang, X., & Liu, H. (2024). Heterogeneous embodied multi-agent collaboration [Publisher: Institute of Electrical and Electronics Engineers]. *IEEE Robotics and Automation Letters*, *9*(6), 5377–5384. https://doi.org/10.1109/lra.2024.3390588

Locklear, K. (2025). Wicked problems: A novel approach using artificial intelligence and scenarios [Publisher: SAGE Publishing]. *Journal of Leadership & Organizational Studies*. https://doi.org/10.1177/15480518251330728

Lou, B., Lu, T., Raghu, T. S., & Zhang, Y. (2025, January). Unraveling human-AI teaming: A review and outlook. https://doi.org/10.2139/ssrn.5211067

Lu, J., Xu, Z., & Kankanhalli, M. (2025). Reasoning llms are wandering solution explorers [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2505.20296

Lu, X., Ma, H., & Wang, Z. (2022). Analysis of OODA loop based on adversarial for complex game environments [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.15502

Lubitz, D. V., & Wickramasinghe, N. (2006). Dynamic leadership in unstable and unpredictable environments [Publisher: Inderscience Publishers]. *International Journal of Management and Enterprise Development*, *3*(4), 339–339. https://doi.org/10.1504/ijmed.2006.009085

March, J. G., & Olsen, J. P. (1975). The uncertainty of the past: Organizational learning under ambiguity* [Publisher: Wiley]. *European Journal of Political Research*, *3*(2), 147–171. https://doi.org/10.1111/j.1475-6765.1975.tb00521.x

Marola, G. (1989). On the detection of the axes of symmetry of symmetric and almost symmetric planar images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(1), 104–108. https://doi.org/10.1109/34.23119

Martins, L. L., Schilpzand, M. C., Kirkman, B. L., Ivanaj, S., & Ivanaj, V. (2012). A contingency view of the effects of cognitive diversity on team performance [Publisher: SAGE Publishing]. *Small Group Research*, *44*(2), 96–126. https://doi.org/10.1177/1046496412466921

Matveev, A. S., Hoy, M., & Savkin, A. V. (2015). A globally converging algorithm for reactive robot navigation among moving and deforming obstacles [Publisher: Elsevier BV]. *Automatica*, *54*, 292–304. https://doi.org/10.1016/j.automatica.2015.02.012

Mehta, M., & Saxena, A. (2025). Entanglement of cultural diversity and future of work: Thematic analysis [Publisher: Taylor & Francis]. *Cogent Arts and Humanities*, *12*(1). https://doi.org/10.1080/23311983.2025.2451500

Mello, A. L., & Delise, L. A. (2015). Cognitive diversity to team outcomes [Publisher: SAGE Publishing]. *Small Group Research*, *46*(2), 204–226. https://doi.org/10.1177/1046496415570916

Miller, C. C., Burke, L., & Glick, W. H. (1998). Cognitive diversity among upper-echelon executives: Implications for strategic decision processes [Publisher: Wiley]. *Strategic Management Journal*, *19*(1), 39–58. https://doi.org/10.1002/(sici)1097-0266(199801)19:1<39::aid-smj932>3.0.co;2-a

Mina, T., Kannan, S. S., Jo, W., & Min, B.-C. (2020). Adaptive workload allocation for multi-human multi-robot teams for independent and homogeneous tasks [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2007.13897

Mohammed, S., Ferzandi, L., & Hamilton, K. (2010, February). Metaphor no more: A 15-year review of the team mental model construct [ISSN: 0149-2063, 1557-1211 Issue: 4 Pages: 876-910 Volume: 36]. https://doi.org/10.1177/0149206309356804

Murphy, D., Paula, T. S., Staehler, W., Vacaro, J., Paz, G. A., Marques, G. F., & dos Santos Oliveira, B. A. (2020). A proposal for intelligent agents with episodic memory [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2005.03182

Newman, J., & Head, B. (2017). Wicked tendencies in policy problems: Rethinking the distinction between social and technical problems [Publisher: Elsevier BV]. *Policy and Society*, *36*(3), 414–429. https://doi.org/10.1080/14494035.2017.1361635

Nowak, R. (2020). "Process of strategic planning and cognitive diversity as determinants of cohesiveness and performance" [Publisher: Emerald Publishing Limited]. *Business Process Management Journal*, *27*(1), 55–74. https://doi.org/10.1108/bpmj-09-2019-0401

Oguntola, I., Campbell, J., Stepputtis, S., & Sycara, K. (2023). Theory of mind as intrinsic motivation for multi-agent reinforcement learning [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2307.01158

Ojasalo, J., & Koski, J. (2019). Designing a concept of a capability development program for building scalable catering business for refugees [Publisher: International Academy of Technology, Education and Development]. *INTED proceedings*, *1*, 7912–7920. https://doi.org/10.21125/inted.2019.1958

Olabisi, J., & Lewis, K. (2018). Within- and between-team coordination via transactive memory systems and boundary spanning [Publisher: SAGE Publishing]. *Group & Organization Management*, *43*(5), 691–717. https://doi.org/10.1177/1059601118793750

Omar, M., Hasan, B., Ahmad, M., Yasin, A., Baharom, F., Mohd, H., & Darus, N. M. (2016). Towards a balanced software team formation based on Belbin team role using fuzzy technique [Publisher: American Institute of Physics]. *AIP conference proceedings*, *1761*, 20082–20082. https://doi.org/10.1063/1.4960922

Partington, D., & Harris, H. (1999). Team role balance and team performance:

An empirical study. *Journal of Management Development*, *18*(8), 694–705. https://doi.org/10.1108/02621719910293783

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research [Publisher: Taylor & Francis]. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/mis0742-1222240302

Ployhart, R. E., & Bliese, P. D. (2005, December). Individual adaptability (i-ADAPT) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability [ISSN: 1479-3601]. In *Advances in human performance and cognitive engineering research* (pp. 3–39). https://doi.org/10.1016/s1479-3601(05)06001-7

Prabhakaran, V., Qadri, R., & Hutchinson, B. (2022). Cultural incongruencies in artificial intelligence [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2211.13069

Pradhan, A., & Lazar, A. (2021). Hey google, do you have a personality? Designing personality and personas for conversational agents. https://doi.org/10.1145/3469595.3469607

Ren, Y., Liu, Y., Ji, T., & Xu, X. (2025). AI agents and agentic AI-navigating a plethora of concepts for future manufacturing [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2507.01376

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning [Publisher: Springer Science+Business Media]. *Policy Sciences*, *4*(2), 155–169. https://doi.org/10.1007/bf01405730

Robertson, P. J. (2008). Transactive memory systems and group performance: A simulation study of the role of shared cognition. [Publisher: Academy of Management]. *Academy of Management Proceedings*, *2008*(1), 1–6. https://doi.org/10.5465/ambpp.2008.33718594

Rodriguez, A. M. R., Cañas, R., & Martinez, V. L. (2005). Meridith belbin team roles and modes of conflict behaviour: A study in work teams from the basque country organizations [Publisher: RELX Group (Netherlands)]. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.736185

Rodriguez, D. V., Andreadis, K., Chen, J., Gonzalez, J., & Mann, D. (2024). Development of a GenAI-Powered Hypertension Management Assistant: Early Development Phases and Architectural Design [ISSN: 2575-2634]. *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 350–359. https://doi.org/10.1109/ICHI61247.2024.00052

Rosenfeld, A., & Pfaltz, J. L. (1966). Sequential Operations in Digital Picture Processing. *Journal of the ACM*, *13*(4), 471–494. https://doi.org/10.1145/321356.321357

Saha, S., Pandey, S. K., & Choudhury, M. (2025). Meta-cultural competence: Climbing the right hill of cultural awareness [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.09637

Šalamon, T. (2011, September). *Design of agent-based models*. https://openlibrary.org/books/OL30696678M/Design_of_Agent-Based_Models

Sanfeliu, A., & Fu, K.-S. (1983). A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(3), 353–362. https://doi.org/10.1109/TSMC.1983.6313167

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *2*(3). https://doi.org/10.70777/si.v2i3.15161

Sarker, I. H., Janicke, H., Ferrag, M. A., & Abuadbba, A. (2024). Multi-aspect rule-based AI: Methods, taxonomy, challenges and directions towards automation, intelligence and transparent cybersecurity modeling for critical infrastructures [Publisher: Elsevier BV]. *Internet of Things*, *25*, 101110–101110. https://doi.org/10.1016/j.iot.2024.101110

Schimmelpfennig, R., Razek, L., Schnell, E., & Muthukrishna, M. (2021). Paradox of diversity in the collective brain [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B Biological Sciences*, *377*(1843). https://doi.org/10.1098/rstb.2020.0316

Schraw, G., Dunkle, M. E., & Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem solving [Publisher: Wiley]. *Applied Cognitive Psychology*, *9*(6), 523–538. https://doi.org/10.1002/acp.2350090605

Shao, Y., Yang, J., & Chen, G. (2017, November). Research on networked collaborative operations based on multi-agent system [ISSN: 2194-5365, 2194-5357]. In *Advances in intelligent systems and computing* (pp. 517–522). Springer Nature. https://doi.org/10.1007/978-3-319-65978-7_78

Simon, H. A. (2019, August). *The Sciences of the Artificial, reissue of the third edition with a new introduction by John Laird* [Google-Books-ID: yYSkDwAAQBAJ]. MIT Press.

Singh, R., Sonenberg, L., & Miller, T. (2016, September). Communication and shared mental models for teams performing interdependent tasks [ISSN: 0302-9743, 1611-3349]. In *Lecture notes in computer science* (pp. 81–97). Springer Science+Business Media. https://doi.org/10.1007/978-3-319-46882-2_10

Somech, A., & Drach-Zahavy, A. (2013). Translating Team Creativity to Innovation Implementation: The Role of Team Composition and Climate for Innovation [Publisher: SAGE Publications Inc]. *Journal of Management*, *39*(3), 684–708. https://doi.org/10.1177/0149206310394187

Stanton, M. C. B., & Roelich, K. (2021, June). Decision making under deep uncertainties: A review of the applicability of methods in practice [ISSN: 0040-1625, 1873-5509 Pages: 120939-120939 Volume: 171]. https://doi.org/10.1016/j.techfore.2021.120939

Stavros, E. N., Iglesias, V., & DeCastro, A. (2021). The wicked wildfire problem and solution space for detecting and tracking the fires that matter. https://doi.org/10.1002/essoar.10506888.1

Tan, J., Braubach, L., Jander, K., Xu, R., & Chen, K. (2020). A novel multi-agent scheduling mechanism for adaptation of production plans in case of supply chain disruptions [Publisher: IOS Press]. *AI Communications*, *33*(1), 1–12. https://doi.org/10.3233/aic-200646

Taylor, R. N. (1974). Nature of problem ill-structuredness: Implications for problem formulation and solution [Publisher: Wiley]. *Decision Sciences*, *5*(4), 632–643. https://doi.org/10.1111/j.1540-5915.1974.tb00642.x

Townend, A. (2007, January). Belbin team roles. In *Palgrave Macmillan UK eBooks* (pp. 111–117). Palgrave Macmillan. https://doi.org/10.1057/9780230582019_11

Usenko, V., von Stumberg, L., Pangercic, A., & Cremers, D. (2017). Real-time trajectory replanning for MAVs using uniform B-splines and a 3D circular buffer. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. https://doi.org/10.1109/iros.2017.8202160

Vaishnavi, V. K. (2007, October). *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. Auerbach Publications. https://doi.org/10.1201/9781420059335

Villanueva, I., Bobinac, T., Yao, B., Hu, J., & Chen, K. (2025). AI as a deliberative partner fosters intercultural empathy for Americans but fails for Latin American participants [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2504.13887

Von Cybernetics, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *7*, 4.

Walt, S. v. d., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image: Image processing in Python [Publisher: PeerJ Inc.]. *PeerJ*, *2*, e453. https://doi.org/10.7717/peerj.453

Wan, Y., & Kalman, Y. M. (2025). Using generative AI personas increases collective diversity in human ideation [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2504.13868

Wang, S., Sauer, S. J., & Schryver, T. D. (2019). The benefits of early diverse and late shared task cognition [Publisher: SAGE Publishing]. *Small Group Research*, *50*(3), 408–439. https://doi.org/10.1177/1046496419835917

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Wang, Z. Z., Gandhi, A., Neubig, G., & Fried, D. (2025). Inducing programmatic skills for agentic tasks [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2504.06821

Wegner, D. M. (1995). A computer network model of human transactive memory [Publisher: Guilford Press]. *Social Cognition*, *13*(3), 319–339. https://doi.org/10.1521/soco.1995.13.3.319

Wendt, D. W. (2024). The OODA Loop Revisited. In D. W. Wendt (Ed.), *The Cybersecurity Trinity: Artificial Intelligence, Automation, and Active Cyber Defense* (pp. 299–307). Apress. https://doi.org/10.1007/979-8-8688-0947-7_11

Williams, B. K. (2010). Adaptive management of natural resources—framework and issues [Publisher: Elsevier BV]. *Journal of Environmental Management*, *92*(5), 1346–1353. https://doi.org/10.1016/j.jenvman.2010.10.041

Williams, B. K., & Johnson, F. A. (2017). Frequencies of decision making and monitoring in adaptive resource management [Publisher: Public Library of Science]. *PLoS ONE*, *12*(8). https://doi.org/10.1371/journal.pone.0182934

Xia, C., Wu, Q., Tian, S., & Hao, Y. (2025). Parallelism meets adaptiveness: Scalable documents understanding in multi-agent LLM systems [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2507.17061

Xie, J., Zhang, K., Chen, J., Yuan, S., Zhang, K., Zhang, Y., Li, L., & Xiao, Y. (2024). Revealing the barriers of language agents in planning [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.12409

Xu, L., Mak, S., Minaricova, M., & Brintrup, A. (2024). On implementing autonomous supply chains: A multi-agent system approach [Publisher: Elsevier BV]. *Computers in Industry*, *161*, 104120–104120. https://doi.org/10.1016/j.compind.2024.104120

Yan, B., Hollingshead, A. B., Alexander, K. S., Cruz, I., & Shaikh, S. J. (2020, December). Communication in transactive memory systems: A review and multidimensional network perspective [ISSN: 1046-4964, 1552-8278 Issue: 1 Pages: 3-32 Volume: 52]. https://doi.org/10.1177/1046496420967764

Yang, Y., Chen, M., Liu, Q., Hu, M., Chen, Q., Zhang, G., Hu, S., Zhai, G., Qiao, Y., Wang, Y., Shao, W., & Luo, P. (2025). Truly assessing fluid intelligence of large language models through dynamic reasoning evaluation [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2506.02648

Ye, R., Liu, X., Wu, Q., Pang, X., Yin, Z., Bai, L., & Chen, S. (2025). X-MAS: Towards building multi-agent systems with heterogeneous llms [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2505.16997

Ying, Z., & Wang, E. (2010). Shared mental models as moderators of team process-performance relationships [Publisher: Scientific Journal Publishers Limited]. *Social Behavior and Personality An International Journal*, *38*(4), 433–444. https://doi.org/10.2224/sbp.2010.38.4.433

Yuan, J., Di, Z., Zhao, S., & Naseem, U. (2024). Cultural palette: Pluralising culture alignment via multi-agent palette [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.11167

Yuan, L., Fu, Z., Zhou, L., Yang, K., & Zhu, S.-C. (2021). Emergence of theory of mind collaboration in multiagent systems [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2110.00121

Yüksel, K. A., & Sawaf, H. (2024). A multi-AI agent system for autonomous optimization of agentic AI solutions via iterative refinement and LLM-driven feedback loops [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.17149

Zabalandikoetxea, S. U., Fernández–Sainz, A., & Merino, J. D. G. (2021). Team diversity and performance in management students: Towards an integrated model [Publisher: Elsevier BV]. *The International Journal of Management Education*, *19*(2), 100478–100478. https://doi.org/10.1016/j.ijme.2021.100478