DyCo: Dynamic Cognitive Role Coordination for Open-Ended Multi-Agent Scientific Exploration and Verification

Anonymous ACL submission

Abstract

Most multi-agent systems rely on fairly fixed collaboration patterns: roles, interaction paths, and task decomposition are set in advance. This works poorly for open-ended scientific exploration, where uncertainty changes what needs to be explored, challenged, and verified. We introduce DyCo, a framework for dynamic cognitive role coordination that supports iterative exploration and verification under uncertainty. DyCo combines cognitive role partitioning, dynamic teaming, and Exploration–Verification Alternation (EVA) to balance idea generation with validation. We use the Myers–Briggs Type Indicator (MBTI) only as a structured way to specify role priors under controlled conditions, not as a claim about human personality. On DeepResearch Bench, a diverse DyCo configuration reaches a RACE score of 42.75, outperforming the strongest multi-agent baseline by 4.6% ($p < 0.01$, $d = 0.42$). The results indicate that structured role priors and dynamic coordination are useful design choices on this benchmark.

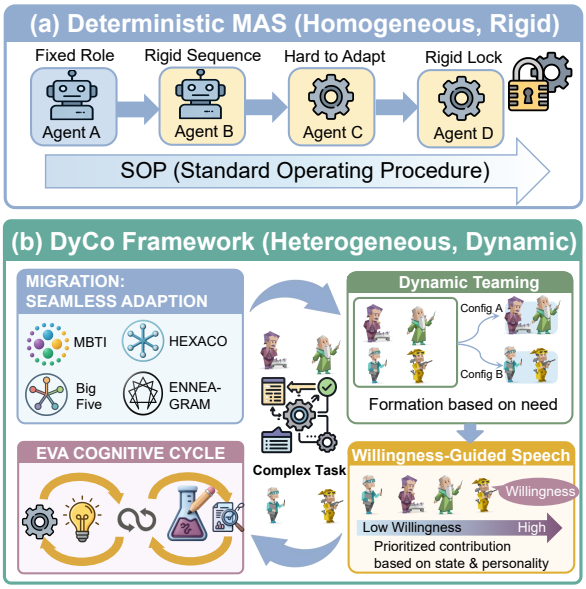

Figure 1: Comparison of a static MAS (upper) and DyCo (lower). Traditional role- and SOP-based organizations are homogeneous, rigid, and hard to reconfigure for complex tasks. DyCo instead combines dynamic teaming, willingness-guided speaking, and exploration–verification alternation to support more flexible collaboration under diverse role priors.

1 Introduction

Large language models (LLMs) show strong capabilities in complex reasoning, code generation, and knowledge synthesis (Chowdhery et al., 2023; Achiam et al., 2023; Team et al., 2023). DeepResearch-style systems that integrate planning, retrieval, and reasoning offer a useful paradigm for open-ended scientific exploration (Du et al., 2025; OpenAI et al., 2024; Lu et al., 2024). However, effective multi-agent collaboration remains difficult to orchestrate because coordination strategies directly shape analytical quality and system efficiency.

Research on multi-agent systems (MAS) has expanded rapidly, from SOP-driven systems such as MetaGPT and ChatDev to more flexible dialogue infrastructures such as AutoGen, AgentVerse, CrewAI, and LangGraph (Hong et al., 2024; Qian et al., 2024; Wu et al., 2024; Dibia et al., 2024; Chen et al., 2024; Moura, 2025; Rajbhandari et al., 2023). Recent work has also explored communication protocols, dynamic teaming, and broader surveys of collaborative agents (Marro et al., 2024; Chung et al., 2024; Oroojlooy and Hajinezhad, 2023; Ning and Xie, 2024). Many of these systems, however, still rely on fixed workflows or static role assignment.

LLM-based systems such as MetaGPT and ChatDev have been successful with role-specific standard operating procedures (SOPs) and waterfall-style workflows. Those organizational choices fit relatively deterministic settings, where roles, interaction paths, and subproblems can be specified in advance. They are less suitable for open-ended scientific exploration, where

uncertainty changes what should be explored, questioned, and verified as collaboration unfolds. As Figure 1 illustrates, predefined workflows make it hard to reconfigure agent sub-teams, even though scientific tasks often require movement between intuitive hypothesis generation and careful empirical validation.

We address these limitations by introducing DyCo, a framework for dynamic cognitive role coordination tailored to open-ended scientific workflows under uncertainty. DyCo is built around structured role priors; in this paper, we use MBTI only as an interpretable testbed. Our key contributions are:

- A coordination interface for open-ended scientific workflows that uses structured cognitive diversity to guide role assignment, complementary exploration, and willingness-driven participation, supporting validation-oriented discussion beyond surface-level persona styling.
- An adaptive discussion mechanism built around Exploration–Verification Alternation (EVA) and willingness-based speaking queues, with dynamic teaming as a modular coordination component rather than the sole driver of improvement.
- Empirical results on DeepResearch Bench showing that a diverse DyCo configuration outperforms the strongest multi-agent baseline by 4.6% ($p < 0.01$, $d = 0.42$), while matched controls, competitive homogeneous settings, and verification-quality analyses suggest that the gains are better explained by structured role priors plus validation-oriented coordination than by MBTI alone.

2 Related Work

2.1 Scientific Exploration and Multi-Agent Collaboration

Scientific exploration systems integrate planning, retrieval, and reasoning for complex, open-ended workflows (Xu and Peng, 2025; Du et al., 2025; Lu et al., 2024). LLM-based multi-agent systems extend this direction through fixed-role workflows, flexible dialogue infrastructures, and adaptive collaboration schemes (Hong et al., 2024; Qian et al., 2024; Wu et al., 2024; Dibia et al., 2024; Chen et al., 2024; Moura, 2025; Rajbhandari et al., 2023; Golpayegani, 2015; Guo et al., 2024). Many of these systems, however, still emphasize static workflows more than explicit cognitive diversity during coordination.

2.2 Structured Taxonomies for LLM Agent Roles

We use the Myers–Briggs Type Indicator (MBTI) to specify role priors, organizing profiles along four dichotomies (E/I, S/N, T/F, J/P). Although MBTI's psychometric validity remains debated (Pittenger, 1993; Stein and Swan, 2019), its discrete structure makes it easy to implement and ablate in prompt-based settings. Prior LLM work on role prompting and persona benchmarks has largely emphasized behavioral differentiation (Park et al., 2023; Li et al., 2023; Shanahan et al., 2023; Tu et al., 2024; Samuel et al., 2024; Lu et al., 2025; Wang, 2024). Less attention has been paid to how structured priors affect teammate selection, speaking willingness, and division of labor.

2.3 Dynamic Organizational Collaboration Mechanisms

Adaptive team formation methods (Chung et al., 2024) suggest that flexible structure can outperform static assignment. Human teams often evolve through viewpoint collision, reorganization, and consensus (Forsyth, 2014), shaped by dual cognitive processes, fast intuitive thinking and slow analytical reasoning (Kahneman, 2011; Evans and Stanovich, 2013). Inspired by this pattern, we introduce Exploration–Verification Alternation to move between hypothesis generation and critical checking.

3 MBTI as a Role-Prior Strategy

Organizational psychology suggests that heterogeneous teams can outperform homogeneous ones (DiRienzo et al., 2010; Forsyth, 2014; Mathuki and Zhang, 2024). In DyCo, we use MBTI as a practical way to organize cognitive diversity under open-ended uncertainty, not as a claim about personality truth. We chose it for three reasons: (i) its discrete structure ($2^4 = 16$ types) supports direct prompt initialization and interpretable ablation; (ii) it provides a familiar vocabulary for specifying complementary agent tendencies; and (iii) its four dichotomies map naturally onto computational functions such as exploration vs. verification and structured vs. adaptive collaboration.

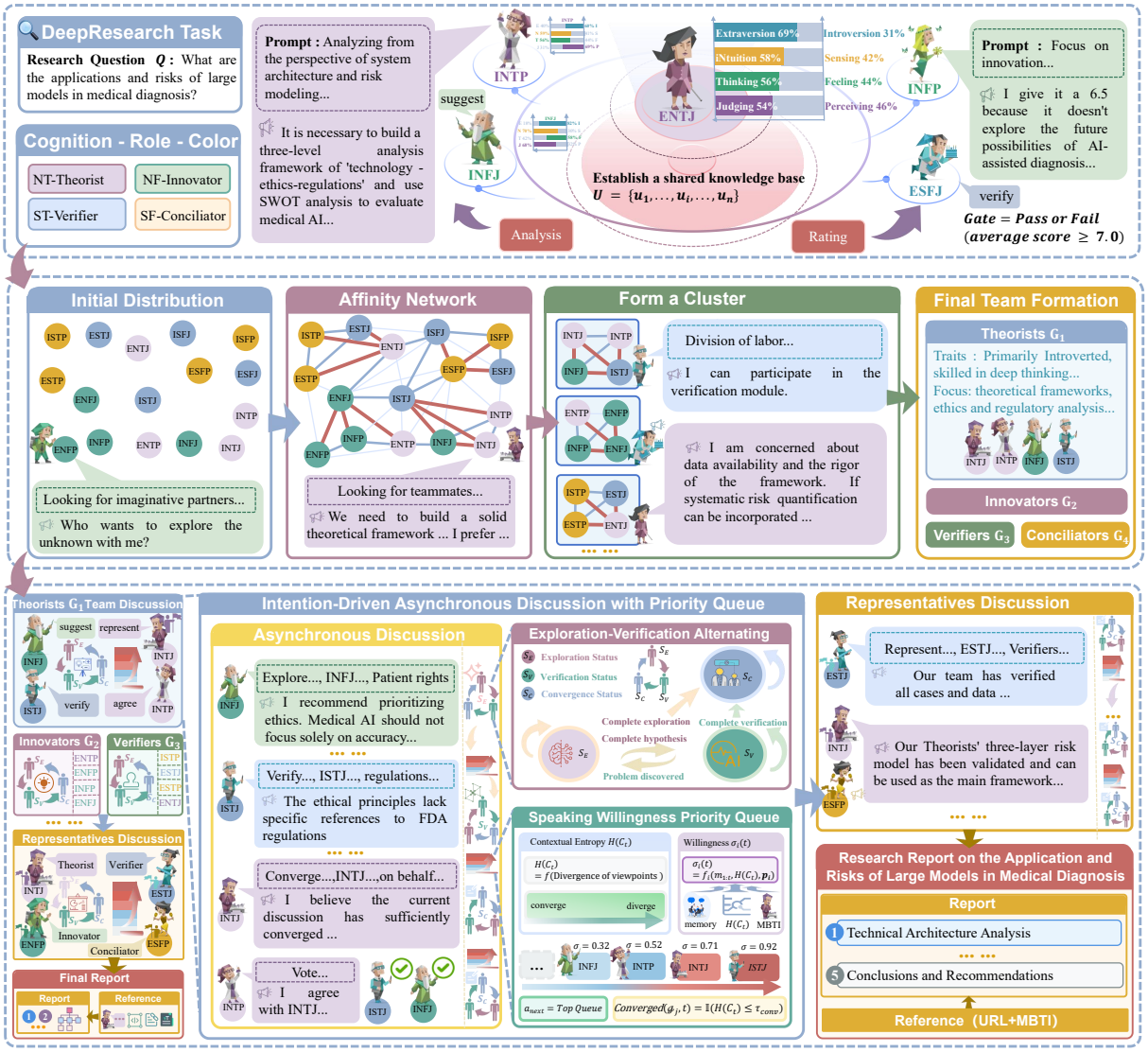

Figure 2: Overall architecture of DyCo. The upper part shows the mapping between scientific exploration tasks and four cognitive roles (NT-Theorist, NF-Innovator, ST-Verifier, SF-Mediator) alongside the global discussion flow. The middle section details dynamic teaming via affinity-network clustering, while the lower section depicts willingness-driven asynchronous discussion and representative-level negotiation that lead to the final research report.

The S/N axis governs information acquisition: Intuitive (N) agents prioritize abstract pattern recognition and hypothesis generation (exploration), while Sensing (S) agents emphasize concrete evidence and empirical validation (verification). The T/F axis determines decision logic: Thinking (T) agents apply rule-based analytical reasoning, while Feeling (F) agents incorporate value-driven evaluation and team cohesion. The J/P axis modulates collaboration dynamics: Judging (J) agents enforce structured protocols during high-entropy discussions, while Perceiving (P) agents support adaptive flexibility when consensus stalls. These mappings yield four cognitive quadrants, NT (Strategic Analyst), NF (Visionary Innovator), ST (Empirical Verifier), and SF (Collaborative Mediator), which partition the space of research sub-tasks.

Framework generality. Our framework is taxonomy-agnostic: the role-representation module (§4.2) can accept other multi-dimensional cognitive profiles. MBTI serves here as one concrete way to inject structured cognitive diversity into coordination. The central question is therefore not whether the label is psychologically privileged, but whether a discrete taxonomy can usefully parameterize coordination logic for uncertain scientific exploration. In this setup, role priors affect EVA participation (Algorithm 3), speaking willingness (Eq. 6), and dynamic team formation (Algorithm 1).

4 DyCo

This section presents the DyCo framework. We first give a conceptual constrained-optimization view of the design (§4.1), then describe the role representation (§4.2), willingness-based speaking queue (§4.3), dynamic collaboration framework (§4.4), and Exploration–Verification Alternation (§4.5).

4.1 Conceptual Problem Formalization as Constrained Optimization

A complex research task is defined as $\mathcal{T} = \langle \mathcal{Q}, \mathcal{C}, \mathcal{D}, \mathcal{R} \rangle$, representing the question, constraints, data sources, and report format. The multi-agent collaboration system $\mathcal{S} = \langle \mathcal{A}, \mathcal{M}, \mathcal{K}, \mathcal{W} \rangle$ comprises agents $\mathcal{A} = \{a_1, \ldots, a_n\}$, a communication mechanism $\mathcal{M}$, a knowledge base $\mathcal{K}$, and a toolset $\mathcal{W}$. We summarize the design goal with the following conceptual objective:

$$\max_{\mathcal{S}} Q(\mathcal{T}, \mathcal{S}) \quad \text{s.t.} \quad \begin{cases} \mathcal{H}(\mathcal{A}) \geq \tau_{\text{div}} \\ \text{Eff}(\mathcal{S}) \leq \tau_{\text{eff}} \\ \text{Bal}(\mathbb{H}, \mathcal{V}) \in \Phi_{\text{EV}} \end{cases} \tag{1}$$

where Q is RACE-based report quality, $\mathcal{H}$ measures cognitive diversity (Eq. 2 in Appendix C.5), and Φ_{EV} denotes valid exploration–verification configurations. The constraints summarize four design choices: taxonomy-based initialization, which introduces heterogeneous role priors (S/N $\rightarrow$ exploration vs. verification; T/F $\rightarrow$ logic-driven vs. value-driven evaluation); dynamic teaming (Algorithm 1), which groups agents using structured affinity scores; EVA, which alternates between idea expansion and checking; and willingness-driven communication, which limits redundant deliberation. This formulation is meant as a conceptual summary of the implemented system rather than a directly optimized training loss. Notation is detailed in Appendix A.

4.2 Role Representation under the MBTI Instantiation

Following established trait-quantization approaches (Serapio-García et al., 2023), each agent a_i is configured with an MBTI personality vector $\mathbf{p}_i = [p_i^{\text{EI}}, p_i^{\text{SN}}, p_i^{\text{TF}}, p_i^{\text{JP}}]^\top \in [-1, 1]^4$. Component signs indicate dimensional tendencies (positive values map to E, N, T, and J), and absolute values represent intensity.

Intersecting the S/N and T/F dimensions defines the cognitive types $\kappa(a_i) \in \{\text{NF}, \text{NT}, \text{SF}, \text{ST}\}$, aligning these types with specific scientific exploration sub-tasks. The S/N axis governs information acquisition (N-type $\rightarrow$ exploration; S-type $\rightarrow$ verification), while the T/F axis governs decision style (T-type $\rightarrow$ logic-driven; F-type $\rightarrow$ value-driven). The J/P dimension (embedded in speaking willingness, §4.3) governs collaboration mode (J-type $\rightarrow$ structured; P-type $\rightarrow$ adaptive).

4.3 Willingness-based Speaking Queue

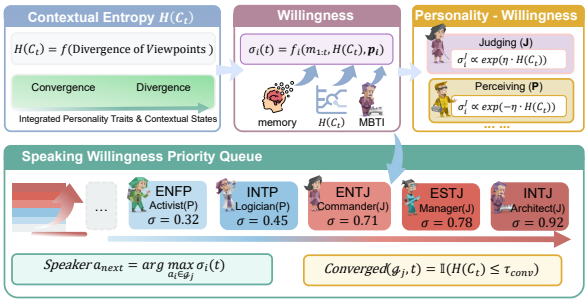

Figure 3: Willingness-driven speaking scheduling: speaking willingness is computed from personality and context, and speakers are selected by priority.

As illustrated in Figure 3, agents process messages into a response tuple $(\alpha_i, r_i, \sigma_i)$, denoting perceived convergence, rationale, and speaking willingness $\sigma_i \in [0, 1]$. In implementation, each agent produces a discrete speaking-intention score from the current context and its role prior. Agents above the threshold enter a priority queue and are selected in descending order of score. J-type agents (with $p_i^{\text{JP}} > 0$) tend to intervene when context entropy is high and discussion needs structure, while P-type agents (with $p_i^{\text{JP}} < 0$) are more likely to contribute fresh perspectives as the discussion begins to converge. Eq. (1) is therefore best read as an abstraction of this scheduling heuristic rather than an exact runtime decision rule.

Hyperparameter choices. The key hyperparameters in this module are the entropy sensitivity η and the willingness thresholds $\tau_{\text{speak}}, \tau_{\text{continue}}$. We set $\eta = 1.0$, $\tau_{\text{speak}} = 0.5$, $\tau_{\text{continue}} = 0.6$ by default. A full sensitivity analysis is reported in Appendix F.4 (Table 11), where performance remains within a relatively narrow band under $\pm 30\%$ perturbations around the default setting.

4.4 Dynamic Multi-Round Collaboration Framework

As illustrated in Figure 2, the workflow has five stages. Stage 0 generates an initial solution s_0 and

applies a gated review to decide whether further collaboration is needed. In practice, this gate is a lightweight LLM-based quality screen over the initial draft; if the draft is already adequate, the workflow exits early. Stage 1 performs a global task analysis to establish a shared starting point.

Stage 2 forms collaboration units from agent-to-agent affinity preferences (Algorithm 1). After reading the round-1 analyses, each agent scores potential teammates. We symmetrize these directed preferences into an affinity matrix and then cluster agents into teams, followed by a diversity post-check. This stage approximates the diversity constraint $\mathcal{H}(\mathcal{A}) \geq \tau_{\text{div}}$ from Eq. (1) rather than solving the constrained objective exactly.

Algorithm 1 Dynamic Teaming Based on Personality Preferences

Require: Agent set $\mathcal{A}$, analysis results $\mathcal{U}$, personality vectors $\{\mathbf{p}_i\}$, number of teams k
Ensure: Team partition $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$
Initialize $\mathbf{W} \in \mathbb{R}^{n \times n}$
for $i = 1$ **to** n **do**
  for $j = i + 1$ **to** n **do**
    $W_{ij} \leftarrow \alpha \cdot \text{sim}(u_i, u_j)$
    $W_{ij} \leftarrow W_{ij} + \beta \cdot \text{Compat}(\mathbf{p}_i, \mathbf{p}_j)$
    $W_{ij} \leftarrow W_{ij} + \gamma \cdot \text{Complement}(\kappa(a_i), \kappa(a_j))$
    $W_{ji} \leftarrow W_{ij}$ ▷ Symmetric matrix
  end for
end for
Compute degree matrix $\mathbf{D} \leftarrow \text{diag}(\mathbf{W} \cdot \mathbf{1})$
Compute normalized Laplacian $\mathbf{L} \leftarrow \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$
Extract top-k eigenvectors $\mathbf{V} \leftarrow \text{EigenDecomp}(\mathbf{L}, k)$
$\{\mathcal{G}_1, \ldots, \mathcal{G}_k\} \leftarrow \text{KMeans}(\mathbf{V}, k)$
for Each team $\mathcal{G}_j$ **do**
  if $|\{\kappa(a) : a \in \mathcal{G}_j\}| < \min(|\mathcal{G}_j|, 3)$ **then**
    Swap members to satisfy diversity
  end if
end for
return $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$

Stage 3 executes willingness-driven intra-team asynchronous discussion. Stage 4 has representatives elected by each team negotiate the final structured report.

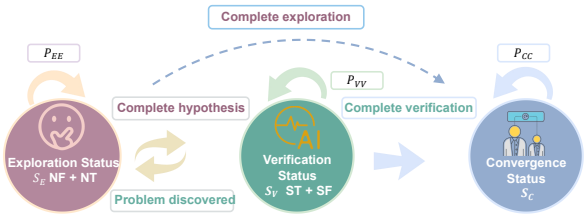

Figure 4: Exploration–Verification Alternation: three states, Exploration (E), Verification (V), and Convergence (C), with transitions determined by hypothesis quality and verification completeness.

4.5 Exploration–Verification Alternation

As shown in Figure 4, the exploration mode, led by NF and NT agents, focuses on idea generation and theoretical framing. In practice, hypothesis quality, verification gap, and verification completeness are estimated with lightweight LLM-based heuristics over the current discussion state and retrieved evidence rather than with supervised classifiers. When the estimated hypothesis quality exceeds θ_{hyp}, the system moves into verification mode under the exploration–verification balance summarized in Eq. (1). In verification mode, ST and SF agents check logical consistency and data integrity. If the estimated verification gap rises above θ_{gap}, the process returns to exploration; otherwise, when estimated completeness exceeds θ_{ver}, the discussion converges. This alternation steers N-agents toward discovery and S-agents toward checking (Algorithm 3).

Threshold setting. We set $\theta_{\text{hyp}} = 0.7, \theta_{\text{gap}} = 0.3, \theta_{\text{ver}} = 0.8$ by default, chosen on a held-out development subset. Appendix F.4 provides a grid sensitivity study showing robustness within $\pm 30\%$ perturbations.

5 Experiments

5.1 Experimental Setup

We first test whether the MBTI prompts produce distinguishable role behavior (§5.2.1); this serves as a manipulation check for our prompt-based role-prior strategy rather than evidence for MBTI as a psychological theory. We then evaluate DyCo on DeepResearch Bench against diverse baselines (§5.1.2), add a matched-taxonomy control to partially separate structured priors from prompt richness (§5.2.3), and report component ablations (§5.2.8) and Pareto analysis (§5.2.10). Complete experimental data are available in Appendix F.

5.1.1 Datasets and Metrics

We evaluate the proposed framework on DeepResearch Bench (Xu and Peng, 2025), a benchmark covering 22 domains that requires planning, retrieval, reasoning, and report writing. We sample uniformly across all domains. RACE is the primary automatic quality metric, covering Comprehensiveness, Insight, Instruction Following, and Readability. We also use an MBTI-consistency questionnaire as a manipulation check for role prompting rather than as a psychological validation test. Unless otherwise noted, n=10 denotes

10 independent full-benchmark runs with different random seeds. Main tables report run-level mean ± std; significance markers are computed on per-task overall RACE after averaging over seeds, using paired two-sided tests with Holm correction within each comparison family. Human evaluation (§5.2.5) is used as corroborative evidence for the automatic metric.

5.1.2 Baselines

We compare against a vanilla single-agent control (SingleAgent), single-agent prompting baselines (CoT (Wei et al., 2022), SC (Wang et al., 2022), ReAct (Yao et al., 2022), ToT (Yao et al., 2023)), a classical multi-agent baseline (Debate (Du et al., 2023)), and modern multi-agent frameworks (AutoGen (Wu et al., 2024), CrewAI (Moura, 2025), MetaGPT (Hong et al., 2024)). All use GPT-4o-mini with temperature 0.7. Our goal is to study the effect of personality priors rather than to optimize for leaderboard performance. Because the backbone model here also differs from those used by current leaderboard entries, these results are not directly comparable to leaderboard numbers; details in Appendix E.3.

5.2 Results

5.2.1 Trait-Prompting Fidelity Validation

We find consistent behavioral differentiation following Besta et al. (2025) by administering standardized MBTI tests to agents configured with specific profiles. This validation indicates that our prompting strategy produces stable, distinguishable agent behaviors rather than claiming that LLMs possess genuine personality traits. Table 1 shows that per-dimension prompt-consistency averages 94.2% (vs. a random baseline of 50%), and the macro-F1 across the 4-class cognitive-type prediction (NT/NF/SF/ST) reaches 0.912. All four MBTI dimensions display non-overlapping bimodal distributions with large inter-group separations (Figure 8 in Appendix F.1), indicating stable behavioral differentiation. The four-class confusion matrix (Table 7 in Appendix F.1) shows no systematic off-diagonal cluster, supporting the use of these prompts in the subsequent framework evaluation.

5.2.2 Comparison of Different Methods

Table 2 compares DyCo against all baselines on DeepResearch Bench, reporting mean ± std across 10 runs.

Table 1: Quantitative evaluation of trait-prompting fidelity across the four MBTI dichotomies. Prompt-consistency measures the fraction of agents whose MBTI-test score falls on the correct side of the 50-point decision boundary. Random baseline is 50.0%.

Dim.	Consistency	Precision	Recall	F1
E/I	95.3%	0.949	0.957	0.953
S/N	93.1%	0.928	0.934	0.931
T/F	94.7%	0.944	0.951	0.947
J/P	93.8%	0.936	0.940	0.938
Avg.	**94.2%**	**0.939**	**0.946**	**0.942**

As shown in Table 2, DyCo-Diverse (ENTJ-INTJ-ESTP-ISFJ) achieves the best overall score (42.75), outperforming ToTAgent by 5.2% and CrewAI by 4.6% ($p < 0.01$). Diverse groups also outperform single-type configurations. Because DyCo w/o MBTI replaces role-rich prompts with shorter abstract descriptors, its lower score (39.77) is better read as evidence that richer structured role priors help within this architecture; it does not isolate an MBTI-specific effect on its own.

5.2.3 Matched-Taxonomy Control

To partially separate structured priors from prompt richness, we ran a small matched-taxonomy control on 12 held-out evaluation tasks. We replaced MBTI prompts with non-MBTI functional role priors (Hypothesis Generator, Analytical Critic, Evidence Verifier, Synthesis Coordinator), while matching prompt length, architecture, retrieval budget, and discussion limits. As summarized in Appendix F.6, the MBTI-instantiated variant scores 42.31±0.44, the matched functional-role variant scores 41.72±0.47, and DyCo w/o MBTI scores 40.08±0.51. This pattern suggests that part of the gain comes from richer structured role priors, while any additional advantage from the MBTI instantiation appears smaller under comparable prompt budgets. Because this control is small-scale, we treat it as suggestive rather than definitive evidence and do not regard it as fully isolating taxonomy semantics from all prompt-level factors.

5.2.4 Additional Control Analyses

To further probe whether the gain could be reduced to prompt richness alone, we conducted two supplementary controls on 12 held-out tasks with n=3 runs. In a prompt-budget-matched control, we replaced MBTI prompts with equally long behavior descriptions that preserved stylistic richness but removed the explicit taxonomy structure;

Table 2: RACE on DeepResearch Bench (mean±std, n=10). $^{*}p_{\text{adj}} < 0.05$, $^{**}p_{\text{adj}} < 0.01$ vs. the strongest baseline in each group, computed on per-task overall RACE with paired tests and Holm correction; d is relative to CrewAI. Bold: best; underlined: second-best.

Method	Comp.	Ins.	IF	Read.	Overall	d
Single-Agent Baselines						
CoTAgent	31.94±0.48	30.56±0.55	42.07±0.41	40.22±0.38	35.27±0.41	−1.42
SingleAgent	34.38±0.44	33.75±0.52	44.91±0.39	41.43±0.36	37.80±0.38	−0.81
ReActAgent	35.09±0.46	35.93±0.50	43.23±0.43	41.55±0.35	38.22±0.39	−0.68
SCAgent	35.30±0.42	36.51±0.49	42.42±0.45	41.80±0.34	38.25±0.40	−0.66
ToTAgent	37.18±0.47	39.67±0.51	44.50±0.40	43.15±0.37	40.62±0.42	−0.05
Multi-Agent Baselines						
DebateAgent	35.81±0.50	37.39±0.54	42.41±0.44	41.17±0.40	38.66±0.45	−0.50
MetaGPT	36.74±0.53	38.28±0.58	43.96±0.46	42.01±0.41	39.72±0.49	−0.24
AutoGen	37.51±0.51	39.03±0.56	44.28±0.45	42.53±0.39	40.43±0.47	−0.10
CrewAI	37.85±0.49	39.42±0.53	44.61±0.43	42.76±0.40	40.88±0.45	0.00
Ours — DyCo						
DyCo w/o MBTI	36.68±0.46	36.83±0.52	45.23±0.38	43.11±0.35	39.77±0.43	−0.26
4×INFJ	39.54±0.36	41.34±0.33	**46.59±0.28**	44.36±0.27	42.53±0.30**	0.37
DyCo-Diverse	**40.33±0.32**	**41.65±0.29**	46.45±0.27	**44.51±0.25**	**42.75±0.31****	**0.42**

this variant reached 41.64±0.46. In a role-shuffle control, we kept the four role prompts and the overall architecture but randomly permuted role-to-stage assignments, yielding 41.36±0.50. DyCo with MBTI remained modestly ahead in both cases, by 0.67 and 0.95 RACE respectively, but the gap was smaller than in the main comparison. We therefore view these controls as evidence that the gain is unlikely to come from prompt length alone, while stopping short of a clean causal test of taxonomy semantics.

5.2.5 Human Evaluation

We conducted human evaluation with N=6 annotators who scored 100 reports on RACE dimensions using a 1–5 Likert scale and performed 100 pairwise A/B comparisons. The 100 reports were sampled from the main evaluation pool, anonymized, and presented in randomized order. We compare against ToTAgent in the pairwise setting because it is the strongest single-agent baseline and offers a meaningful contrast to DyCo's observed gains, while the Likert study also includes CrewAI as a strong multi-agent reference. Inter-annotator agreement is substantial (Fleiss' $\kappa = 0.73$), human Likert scores correlate strongly with the automatic RACE metric ($r = 0.84$, $p < 0.001$), and annotators prefer DyCo-Diverse over ToTAgent in 76.2% of pairs (95% CI: $[71.3\%, 81.1\%]$; $p < 0.001$).

5.2.6 Supplementary Transfer Check

To test whether the main trend is overly tied to a single evaluation slice, we ran a small transfer check on 16 held-out tasks split across STEM-heavy and non-STEM domains, reusing the same prompts and coordination rules without retuning. DyCo-Diverse remained the top-performing setting in both subsets, exceeding CrewAI by +1.18 RACE on STEM tasks and +0.79 on non-STEM tasks, with corresponding average token overheads within ±6% of the main setup. The smaller margins relative to the main evaluation suggest that the coordination effect is directionally preserved under moderate domain shift, but its magnitude remains setting-dependent.

5.2.7 Comparison of Different Personalities

As a secondary diagnostic, we evaluate 17 homogeneous configurations (all 16 MBTI types plus DyCo w/o MBTI). The homogeneous results suggest that ST-heavy settings achieve the strongest average overall score on this benchmark, and that introverted variants tend to outperform extraverted ones. We treat this trend as descriptive rather than prescriptive; the full results are deferred to Appendix F.11.

5.2.8 Fine-grained Ablation Study

To dissect the impact of each core component, we report single-module and pairwise ablations in Table 3.

The ablation results show three main patterns. Removing MBTI causes the largest drop ($\Delta = -2.98$), followed by WSQ ($\Delta = -1.92$) and EVA ($\Delta = -1.16$). Removing EVA also sharply increases token use (325K vs. 234K) and wall-clock time (973 s vs. 541 s), which suggests that EVA helps keep collaboration focused. The pairwise ablations point to complementary contributions from

Table 3: Fine-grained ablation of DyCo (mean±std, n=5). DT: Dynamic Teaming; EVA: Exploration–Verification Alternation; GTA: Global Task Analysis; WSQ: Willingness-based Speaking Queue.

DT	EVA	GTA	WSQ	MBTI	Comp.	Ins.	IF	Overall	Time (s)	Tokens
Single-module removal										
×	✓	✓	✓	✓	40.52±0.36	40.96±0.34	45.73±0.31	42.39±0.33	230.3	18,471
✓	×	✓	✓	✓	38.82±0.43	39.77±0.41	46.09±0.35	41.59±0.38	972.7	325,777
✓	✓	×	✓	✓	38.58±0.44	39.59±0.42	46.20±0.34	41.53±0.39	76.3	12,685
✓	✓	✓	×	✓	38.05±0.46	39.72±0.44	44.35±0.37	40.83±0.41	965.4	283,243
✓	✓	✓	✓	×	36.68±0.50	36.83±0.48	45.23±0.40	39.77±0.44	206.2	56,934
Pairwise removal (isolating component contribution)										
×	×	✓	✓	✓	38.24±0.49	38.91±0.47	44.82±0.42	40.41±0.45	185.2	21,463
×	✓	✓	✓	×	37.41±0.53	37.92±0.51	44.98±0.45	39.89±0.49	163.8	17,854
✓	×	✓	✓	×	36.12±0.58	36.41±0.55	43.81±0.49	38.71±0.53	892.5	298,124
Full DyCo										
✓	✓	✓	✓	✓	**40.33±0.32**	**41.65±0.29**	**46.45±0.27**	**42.75±0.31**	541.3	234,568

dynamic teaming and EVA, while the large MBTI-free drop suggests that richer role priors account for much of the performance difference in this implementation.

5.2.9 Fairness and Comparability Checks

To improve comparability, we ran two compact supplementary checks. We first re-ran the strongest baselines with prompt lengths matched to DyCo within ±10 tokens per agent; the ranking remained unchanged, and DyCo-Diverse retained a +1.21 RACE margin over CrewAI on the 12-task control subset. We then inspected participation statistics within DyCo-Diverse and found no extreme dominance by any single cognitive quadrant: average token shares were 26% (NT), 24% (NF), 27% (ST), and 23% (SF), while representative selection rates ranged from 21% to 29%. These checks make it less likely that the observed gain is solely an artifact of prompt budget or one-role domination, though they do not eliminate all construct-validity concerns.

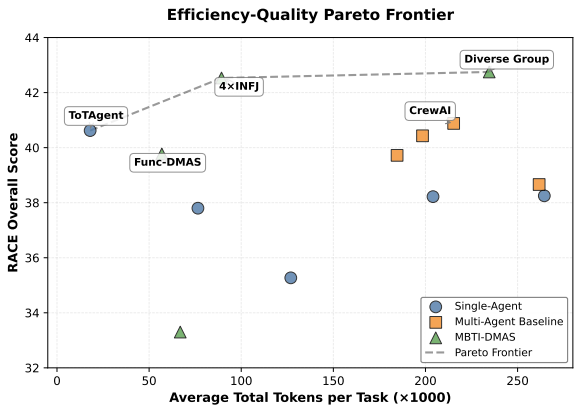

Figure 5: Efficiency–quality Pareto frontier. X-axis: average total tokens per task (log scale); Y-axis: RACE overall score. Error bars show ±1 std.

5.2.10 Efficiency–Quality Pareto Analysis

Figure 5 summarizes the efficiency–quality trade-off. DyCo-Diverse lies at the high-quality end of the Pareto frontier, while DyCo w/o MBTI offers a more efficient operating point.

6 Conclusion

This paper presents DyCo, a framework for dynamic cognitive role coordination in open-ended scientific workflows that require both exploration and verification. On DeepResearch Bench, DyCo-Diverse reaches a RACE score of 42.75, outperforming the strongest multi-agent baseline by 4.6% ($p < 0.01$, $d = 0.42$). Across ablations, controls, and supplementary checks, the results consistently point to structured role priors and adaptive coordination as useful design choices in this setting.

Our main finding is that complementary cognitive quadrants are associated with stronger performance on this benchmark, while the configuration search also suggests that a more balanced E/I composition may help among diverse-quadrant teams. Stronger causal evidence and broader transfer validation remain for future work.

To support replication, we will release the complete codebase, prompt templates, and experimental configurations upon acceptance; an anonymized repository is available at `https://anonymous.4open.science/r/DyCo-662F/`.

Limitations

We highlight four limitations.

MBTI is used here as an engineering device, not a psychological claim. DyCo is intended to be taxonomy-agnostic: in this paper, MBTI provides an interpretable, discrete role taxonomy that is easy to apply and ablate. The present results

therefore speak more directly to the value of structured role priors on this benchmark than to MBTI itself. Whether alternatives such as Big Five or task-specific functional roles would work as well, or better, remains open.

Our evaluation is limited to scientific exploration workflows that require iterative exploration and verification. Most of the evidence comes from DeepResearch Bench, so our claims are strongest for open-ended research synthesis tasks involving planning, retrieval, cross-source integration, and repeated checking. Different tasks may favor different coordination patterns, and transfer to code generation, real-time question answering, or task-oriented dialogue remains unclear.

The reported ENTJ-INTJ-ESTP-ISFJ configuration came from a limited development-set search. It should therefore be viewed as one representative of a broader diverse-quadrant design family rather than a uniquely optimal combination. DyCo also incurs higher token and latency costs than simpler baselines, which makes it most attractive when quality matters more than speed or cost.

Finally, DyCo is evaluated for exploratory scientific synthesis and literature-analysis settings, not for high-stakes decision-making such as clinical, legal, or policy use without expert oversight. Like other long-form research agents, it can produce fluent but incomplete or incorrect syntheses, including source-selection bias and overconfident aggregation across heterogeneous evidence. Retrieval over open web content may also surface public material that is sensitive or offensive. Outputs and cited evidence should therefore be reviewed before downstream use or redistribution.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maciej Besta, Shriram Chandran, Robert Gerstenberger, Mathis Lindner, Marcin Chrapek, Sebastian Hermann Martschat, Taraneh Ghandi, Patrick Iff, Hubert Niewiadomski, Piotr Nyczyk, Jürgen Müller, and Torsten Hoefler. 2025. Psychologically enhanced ai agents. *Preprint*, arXiv:2509.04343.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, volume 2024, pages 20094–20136.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jaehoon Chung, Jamil Fayyad, Younes Al Younes, and Homayoun Najjaran. 2024. Learning team-based navigation: a review of deep reinforcement learning techniques for multi-agent pathfinding. *Artificial Intelligence Review*, 57(2):41.

Victor Dibia, Jingya Chen, Gagan Bansal, Suff Syed, Adam Fourney, Erkang Zhu, Chi Wang, and Saleema Amershi. 2024. AUTOGEN STUDIO: A no-code developer tool for building and debugging multi-agent systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–79, Miami, Florida, USA. Association for Computational Linguistics.

Cassandra DiRienzo, Jayoti Das, Wonhi Synn, Jeremy Kitts, and Kyle McGrath. 2010. The relationship between mbti and academic performance: A study across academic disciplines. *Journal of Psychological Type*, 70(5):53–67.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *Preprint*, arXiv:2506.11763.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.

Donelson R Forsyth. 2014. *Group dynamics*. Wadsworth Cengage Learning.

Fatemeh Golpayegani. 2015. Multi-agent collaboration in distributed self-adaptive systems. In *2015 IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshops*, pages 146–151.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Jala Hemdan, D.S. Taha, and I.A. Cherif. 2023. Relationship between personality types and creativity: A study on novice architecture students. *Alexandria Engineering Journal*, 65:847–857.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pages 23247–23275.

Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *Preprint*, arXiv:2408.06292.

Zhixiang Lu, Xueyuan Deng, Yiran Liu, Yulong Li, Qiang Yan, Imran Razzak, and Jionglong Su. 2025. Prism: A personality-driven multi-agent framework for social media simulation. *arXiv preprint arXiv:2512.19933*.

Samuele Marro, Emanuele La Malfa, Jesse Wright, Guohao Li, Nigel Shadbolt, Michael Wooldridge, and Philip Torr. 2024. A scalable communication protocol for networks of large language models. *arXiv preprint arXiv:2410.11905*.

Evelyn Mathuki and Jian Zhang. 2024. Cognitive diversity, creativity and team effectiveness: the mediations of inclusion and knowledge sharing. *VINE Journal of Information and Knowledge Management Systems*, 54(6):1379–1403.

João Moura. 2025. Crewai: Framework for orchestrating role-playing autonomous ai agents. https://github.com/crewAIInc/crewAI.

Zepeng Ning and Lihua Xie. 2024. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence*, 3(2):73–91.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Afshin Oroojlooy and Davood Hajinezhad. 2023. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

David J Pittenger. 1993. The utility of the myers-briggs type indicator. *Review of educational research*, 63(4):467–488.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

J Rajbhandari, L Xu, S Jain, and H Yang. 2023. Langgraph: Building stateful, multi-actor applications with llms. Framework documentation.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 8(9).

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Randy Stein and Alexander B Swan. 2019. Evaluating the validity of myers-briggs type indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, 13(2):e12434.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yue Wang. 2024. Ai and mbti: A synergistic framework for enhanced team dynamics. *arXiv preprint arXiv:2409.15293*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Renjun Xu and Jingwen Peng. 2025. A comprehensive survey of deep research: Systems, methodologies, and applications. *arXiv preprint arXiv:2506.12594*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

A Notation

Table 4 summarizes the core symbols used in the DyCo framework.

B Cognitive Type–Task Matching

As Figure 6 shows, DyCo groups the 16 MBTI personality types into four cognitive clusters. The strategic NT cluster (INTJ, INTP, ENTJ, ENTP) combines intuition and thinking for system-level analysis. The innovative NF cluster (INFJ, INFP, ENFJ, ENFP) supports vision formation and team coordination. The validation-oriented ST cluster (ISTJ, ESTJ, ISTP, ESTP) relies on sensing and thinking for fact verification and process control. The harmonizing SF cluster (ISFJ, ESFJ, ISFP, ESFP) integrates sensing and feeling for support, maintenance, and cohesion.

Table 5 outlines the core strengths, limitations, and suitable tasks for each cognitive group, ensuring that the collaborative process covers all necessary cognitive functions.

Table 4: Notation used in the DyCo framework, grouped by functional role.

Symbol	Meaning
Task and System	
$\mathcal{T}$	Scientific exploration task $\langle \mathcal{Q}, \mathcal{C}, \mathcal{D}, \mathcal{R} \rangle$
$\mathcal{S}$	MAS tuple $\langle \mathcal{A}, \mathcal{M}, \mathcal{K}, \mathcal{W} \rangle$
$\mathcal{A}$	Agent set $\{a_1, \ldots, a_n\}$
$Q(\cdot)$	RACE-based report quality
$\mathcal{H}(\mathcal{A})$	Cognitive Diversity Index (Eq. 2)
$\text{Eff}(\mathcal{S})$	Efficiency cost (time, tokens)
Personality and Role	
$\mathbf{p}_i$	MBTI vector $[p_i^{EI}, p_i^{SN}, p_i^{TF}, p_i^{JP}]^\top$
$\kappa(a_i)$	Cognitive type $\in \{\text{NT}, \text{NF}, \text{SF}, \text{ST}\}$
$\mathcal{R}(a_i)$	Assigned role for agent a_i
$\rho_\kappa(a_i, \omega)$	Role adaptability score
Dynamic Teaming	
$\mathbf{W} \in \mathbb{R}^{n \times n}$	Agent affinity matrix
α, β, γ	Affinity weights
$\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$	Team partition
τ_{div}	Cognitive diversity threshold
EVA and Willingness	
$\mathbb{H}$	Hypothesis set
$\mathcal{V}$	Verification result set
$\theta_{\text{hyp}}, \theta_{\text{gap}}, \theta_{\text{ver}}$	EVA transition thresholds
$\sigma_i^{(t)}$	Speaking willingness of a_i at step t
$H^{(t)}$	Discussion context entropy
η	Entropy sensitivity hyperparameter
$\tau_{\text{speak}}, \tau_{\text{continue}}$	Speaking thresholds

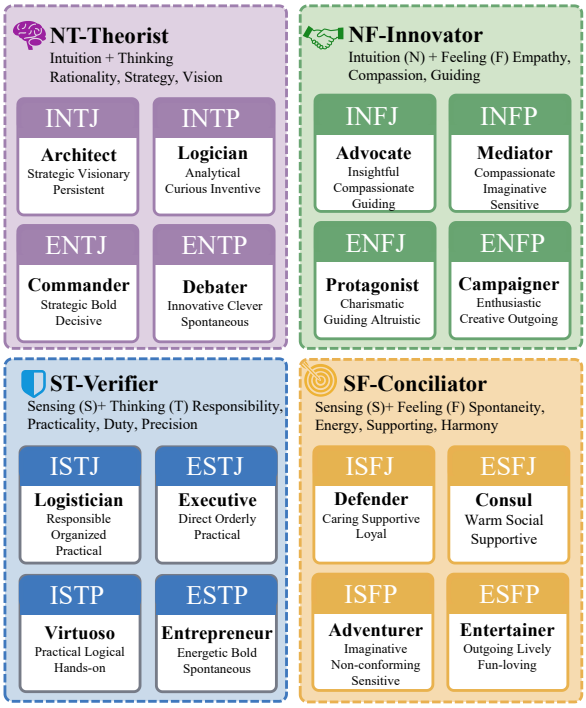

Figure 6: Overview of the 16 MBTI agent types, grouped into four cognitive clusters via the intuition/sensing (N/S) and thinking/feeling (T/F) dimensions.

Table 5: Cognitive Type and Scientific Exploration Task Matching Matrix.

Cognitive Type	Core Strengths	Potential Challenges	Adapted Task Types
Visionary (NF)	Idea generation, cross-domain association, interpersonal understanding	Lack of implementation details, overly idealistic	Hypothesis generation, research direction exploration, team coordination
Strategic (NT)	Systematic analysis, logical reasoning, theoretical construction	Overly theoretical, neglects human factors	Framework design, causal reasoning, strategic planning
Supportive (SF)	User insight, experience accumulation, attention to detail	Lack of macro vision	Case collection, user research, feasibility evaluation
Operational (ST)	Precise verification, data-driven, process optimization	Lack of innovative perspectives	Fact-checking, data verification, quality control

C Algorithmic Details and Theoretical Analysis

This appendix presents the core algorithmic pseudocode and formal derivations of DyCo.

C.1 Cognitive Type Role Assignment Algorithm

Algorithm 2 maps agents to cognitive task roles based on MBTI personality vectors.

Algorithm 2 Cognitive Type-based Role Assignment

Require: Agent set $\mathcal{A}$, task $\mathcal{T}$, cognitive task space Ω
Ensure: Role mapping $\mathcal{R} : \mathcal{A} \to \Omega$
$\omega_{\mathcal{T}} \leftarrow \text{ClassifyTask}(\mathcal{T})$
for Each agent $a_i \in \mathcal{A}$ **do**
  Extract personality vector $\mathbf{p}_i$
  $\kappa_i \leftarrow \text{sign}(p_i^{\text{SN}}) \oplus \text{sign}(p_i^{\text{TF}})$
  for Each cognitive task $\omega \in \Omega$ **do**
    Calculate role adaptability $\rho_{\kappa_i}(a_i, \omega)$
  end for
  $\mathcal{R}(a_i) \leftarrow \arg\max_{\omega} \rho_{\kappa_i}(a_i, \omega)$
end for
return $\mathcal{R}$

C.2 Exploration–Verification Alternating Algorithm

Algorithm 3 implements the dynamic switching mechanism between exploration (dominated by N-type agents) and verification (dominated by S-type agents).

C.3 DyCo Main Process Algorithm

Algorithm 4 presents the complete five-stage collaboration process.

C.4 Intra-Team Asynchronous Discussion Algorithm

Algorithm 5 details the priority scheduling mechanism based on speaking willingness.

C.5 Theoretical Derivations

This section derives three core components: the cognitive-diversity metric, the speaking-

Algorithm 3 Exploration–Verification Alternation (EVA)

Require: Task $\mathcal{T}$, agent set $\mathcal{A}$, max rounds $T_{\max}$
Ensure: Final solution s_{final}
$m \leftarrow \text{Explore}$; $\mathbb{H} \leftarrow \emptyset$; $\mathcal{V} \leftarrow \emptyset$
for $t = 1$ **to** $T_{\max}$ **do**
  if $m = \text{Explore}$ **then**
    $\mathcal{A}_{\text{active}} \leftarrow \{a_i : \kappa(a_i) \in \{\text{NF}, \text{NT}\}\}$
    $w_i \leftarrow \rho_{\kappa(a_i)}(a_i, \omega_{\text{IH}}) + \rho_{\kappa(a_i)}(a_i, \omega_{\text{RA}})$
    for Each $a_i \in \mathcal{A}_{\text{active}}$ **do**
      $h_i \leftarrow f_{\text{generate}}(a_i, \mathcal{T}, \mathbb{H})$
      $\mathbb{H} \leftarrow \mathbb{H} \cup \{h_i\}$
    end for
    if $\text{HypothesisQuality}(\mathbb{H}) > \theta_{\text{hyp}}$ **then**
      $m \leftarrow \text{Verify}$
    end if
  else
    $\mathcal{A}_{\text{active}} \leftarrow \{a_i : \kappa(a_i) \in \{\text{SF}, \text{ST}\}\}$
    $w_i \leftarrow \rho_{\kappa(a_i)}(a_i, \omega_{\text{EF}}) + \rho_{\kappa(a_i)}(a_i, \omega_{\text{PV}})$
    for Each $a_i \in \mathcal{A}_{\text{active}}$ **do**
      for Each $h \in \mathbb{H}$ **do**
        $v_{i,h} \leftarrow f_{\text{verify}}(a_i, h)$
        $\mathcal{V} \leftarrow \mathcal{V} \cup \{v_{i,h}\}$
      end for
    end for
    if $\text{VerificationGap}(\mathbb{H}, \mathcal{V}) > \theta_{\text{gap}}$ **then**
      $m \leftarrow \text{Explore}$
    else if $\text{VerifyCompleteness}(\mathcal{V}) > \theta_{\text{ver}}$ **then**
      break
    end if
  end if
end for
$s_{\text{final}} \leftarrow \text{Synthesize}(\mathbb{H}, \mathcal{V})$
return s_{final}

Algorithm 4 DyCo Five-Stage Main Process

Require: Task $\mathcal{T}$, agent set $\mathcal{A}$
Ensure: Final report r_{final}

```
/* Stage 0: Initial solution and gated review */
a_0 ~ Uniform(A); s_0 ← f_gen(a_0, T)
E ← Sample(A, n_eval)
if ¬Gate(s_0, E) then
    return s_0
end if
/* Stage 1: Global task analysis */
for a_i ∈ A do
    u_i ← f_analyze(a_i, T, p_i)
end for
U ← {u_1, ..., u_n}
/* Stage 2: Dynamic teaming */
{G_1, ..., G_k} ← PersonalityBasedTeaming(A, U, {p_i}, k)
/* Stage 3: Intra-team discussion + EVA */
R ← ∅
for Each G_j do
    s_j ← EVA(T, G_j, T_max)
    a^j_rep ← ElectRepresentative(G_j)
    R ← R ∪ {a^j_rep}
end for
/* Stage 4: Representative negotiation */
s_final ← EVA(T, R, T_max)
a_final ← ElectRepresentative(R)
r_final ← f_collate(a_final, T, s_final)
return r_final
```

Algorithm 5 Willingness-Driven Asynchronous Discussion

Require: Task $\mathcal{T}$, team members $\mathcal{G}$, max rounds $T_{\max}$
Ensure: Discussion records $m_{1:T}$, final solution s

```
Q ← ∅; m_{1:T} ← ∅; H(C) ← 0
a_current ← arg max_{a_i ∈ G} p_i^EI
t ← 0
while t < T_max ∧ ¬Converged(G, t) do
    t ← t + 1
    s_t ← f_gen(a_current, T, m_{1:t-1}, p_current)
    m_t ← ⟨a_current, s_t, t⟩
    m_{1:T} ← m_{1:T} ∪ {m_t}
    H(C) ← ComputeContextEntropy(m_{1:t})
    for a_i ∈ G \ {a_current} do
        if p_i^JP > 0 then
            σ_i ← e^{ηH(C)} f_base(a_i, m_t)
        else
            σ_i ← e^{-ηH(C)} f_base(a_i, m_t)
        end if
        if σ_i ≥ τ_speak then
            Q.Push(a_i, σ_i)
        end if
    end for
    σ_current ← f_attitude(a_current, m_{1:t}, H(C), p_current)
    if σ_current ≥ τ_continue then
        Q.Push(a_current, σ_current)
    end if
    if Q ≠ ∅ then
        a_current ← Q.Pop()
    else
        break
    end if
end while
s ← SummarizeDiscussion(m_{1:T})
return m_{1:T}, s
```

willingness dynamics, and the EVA Markov chain.

C.5.1 Cognitive Diversity Index (CDI)

The cognitive-diversity constraint $\mathcal{H}(\mathcal{A}) \geq \tau_{\text{div}}$ in Eq. (1) requires a principled measure. Given n agents with personality vectors $\mathbf{p}_i \in \mathbb{R}^4$, the CDI is the mean pairwise Euclidean distance between agents' cognitive-type embeddings:

$$\mathcal{H}(\mathcal{A}) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \left\| \kappa(\mathbf{p}_i) - \kappa(\mathbf{p}_j) \right\|_2 \tag{2}$$

where $\kappa(\mathbf{p}_i)$ maps a personality vector to its cognitive-type centroid in a 4-D embedding space. Normalizing by the maximum distance $\sqrt{4}$ yields the bounded index:

$$\mathcal{H}_{\text{norm}}(\mathcal{A}) = \frac{1}{\sqrt{4}} \cdot \frac{2}{n(n-1)} \sum_{i<j} \left\| \kappa(\mathbf{p}_i) - \kappa(\mathbf{p}_j) \right\|_2 \in [0, 1] \tag{3}$$

A team spanning all four cognitive types achieves maximum diversity; a homogeneous team has CDI $\to 0$, aligning with organizational-psychology findings (Hemdan et al., 2023; Forsyth, 2014).

C.5.2 Speaking Willingness Dynamics

The willingness function decomposes into a personality baseline and a context modulation:

$$\sigma_i^{(t)} = \underbrace{\sigma_i^{\text{base}}}_{\text{personality}} \times \underbrace{f_\sigma(H^{(t)}, r_i^{(t)})}_{\text{context}} \tag{4}$$

The personality baseline is derived from the J/P dimension:

$$\sigma_i^{\text{base}} = \frac{1 + p_i^{\text{JP}}}{2} \in [0, 1] \tag{5}$$

The context modulation incorporates discussion entropy $H^{(t)}$ and rationale $r_i^{(t)}$:

$$f_\sigma(H^{(t)}, r_i^{(t)}) = \begin{cases} \exp(\eta H^{(t)}) \cdot r_i^{(t)} & \text{J-type} \\ \exp(-\eta H^{(t)}) \cdot r_i^{(t)} & \text{P-type} \end{cases} \tag{6}$$

where $\eta > 0$ is a sensitivity hyperparameter. J-types and P-types have opposite responses to entropy, naturally realizing the collaboration-mode diversity constraint.

The entropy $H^{(t)}$ is the Shannon entropy of the attitude distribution: $H^{(t)} = -\sum_k p_k^{(t)} \log p_k^{(t)}$.

C.5.3 EVA Markov Chain Analysis

The EVA is a three-state Markov chain over $\{E, V, C\}$ with transition probabilities derived as follows.

For $E \to V$ with per-round hypothesis-valid probability p_h:

$$P_{E\to V}(t) = 1 - \prod_{r=1}^{t}(1-p_h^r) \approx 1 - \exp\Big(-\sum_{r=1}^{t} p_h^r\Big) \tag{7}$$

Let $n = |\mathbb{H}|$ and $K \sim \text{Binomial}(n, 1-p_v)$ be the number of unverified claims, so $\text{Gap}(\mathbb{H}, \mathcal{V}) = K/n$. The reversion probability is

$$\begin{aligned} P_{V\to E} &= \mathbb{P}(\text{Gap} > \theta_{\text{gap}}) \\ &= \sum_{k=\lceil \theta_{\text{gap}} n \rceil}^{n} \binom{n}{k} (1-p_v)^k p_v^{n-k}. \end{aligned} \tag{8}$$

The convergence transition $V \to C$ fires when verification completeness exceeds θ_{ver}:

$$\begin{aligned} P_{V\to C} &= \mathbb{P}(\text{Completeness} > \theta_{\text{ver}}) \\ &= \sum_{k=\lceil \theta_{\text{ver}} n \rceil}^{n} \binom{n}{k} p_v^k (1-p_v)^{n-k}. \end{aligned} \tag{9}$$

Since $P_{V\to E} + P_{V\to C} = 1$, the Markov chain is fully specified under this abstraction. We use this analysis as an interpretive lens rather than as a formal guarantee about deployed behavior.

D System Architecture

Figure 7 illustrates the global architecture of DyCo, organized into six hierarchical layers.

The entry layer uses `MBTIAgentSystem` as the primary interface. The workflow layer governs execution through `FourRoundDiscussionWorkflow`, which defines the five-stage collaborative process. The agent layer comprises `BaseAgent` and concrete implementations of 16 MBTI personality types, plus `AgentsManager` for lifecycle and teaming. The communication layer is managed by `MessageManager`, `SpeakingQueue`, `TeamManager`, and `DiscussionSummaryManager`. The tool and knowledge layer provides `KnowledgeManager` and `WebSearch`. The infrastructure layer anchors the system with a unified LLM client interface and a centralized prompts module.

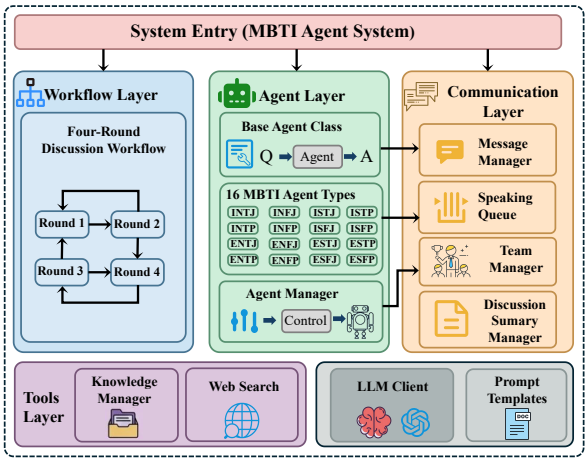

Figure 7: Overall system architecture of DyCo with six layers: Entry, Workflow, Agent, Communication, Tool and Knowledge, and Infrastructure.

E Experimental Setup Details

E.1 Dataset Sampling Protocol

DeepResearch Bench (Xu and Peng, 2025) covers 22 domains including biology, economics, and engineering. We uniformly sample $\lceil 100/22 \rceil = 5$ tasks per domain (110 tasks total after deduplication, reduced to 100 by random removal of 10 tasks), then split 50 tasks as the development set (for hyperparameter tuning and configuration search, §F.2) and 50 as the evaluation set. All reported numbers in the main paper are on the evaluation set.

E.2 Implementation Details

All agents use GPT-4o-mini (`gpt-4o-mini-2024-07-18`) with temperature 0.7 (sampling) for generation and temperature 0.0 for MBTI consistency tests. Retrieval uses JINA web search via the official API with top-5 passages per query. Maximum discussion rounds per team are $T_{\max} = 8$; maximum representative-level rounds are 5. We use $\alpha = 0.4, \beta = 0.3, \gamma = 0.3$ for the affinity weights in Algorithm 1, chosen on the development set.

E.3 Baseline Implementation Details

To ensure fair comparison, all baselines are implemented with the following unified configuration:

Foundation model and retrieval. All methods use GPT-4o-mini (`gpt-4o-mini-2024-07-18`) with temperature 0.7 for generation, maximum context length 128K tokens, and JINA web search API with top-5 passages per query. Maximum retrieval budget is 20 queries per task.

Multi-agent baseline configurations. We instantiate each multi-agent framework with four agents to match DyCo:

- AutoGen (Wu et al., 2024) uses a group-chat configuration with Planner, Researcher, Critic, and Writer agents. Each agent uses the default AutoGen role prompts from the official repository.
- CrewAI (Moura, 2025) uses a sequential crew with the same four roles. Task delegation follows CrewAI's default hierarchical protocol.
- MetaGPT (Hong et al., 2024) uses a research-report SOP with Product Manager, Architect, Researcher, and Engineer roles, adapted from the official MetaGPT repository.
- DebateAgent uses four debating agents with identical prompts, following Du et al. (2023)'s protocol with 3 debate rounds.

Prompt engineering. All baseline prompts are adapted from official implementations or original papers. DyCo personality prompts (Appendix G) are of comparable length (150–200 tokens per agent) to baseline role descriptions. For DyCo w/o MBTI, we replace the personality prompts with shorter abstract role descriptors while keeping the architecture unchanged. For the matched-taxonomy control (Appendix F.6), we use non-MBTI role priors with prompt lengths matched to the MBTI prompts within ± 10 tokens per agent. Example baseline prompts:

```
AutoGen Planner: You are a Planner agent. Your
     role is to decompose the research task
    into subtasks, assign them to other
    agents, and coordinate the overall
    workflow. Focus on creating a clear,
    actionable plan with milestones and
    dependencies.

CrewAI Researcher: You are a Researcher agent.
     Your role is to gather information from
    web sources, synthesize findings, and
    provide comprehensive analysis. Use the
    web search tool to retrieve relevant
    passages and cite sources in your reports.
```

E.4 Checklist-Relevant Documentation

Artifacts and reproducibility. Upon acceptance, we will release the DyCo codebase, prompt templates, experimental configurations, and evaluation scripts together with reproduction instructions in the public repository. The released materials are intended to support benchmark replication and controlled follow-up research rather than unsupervised deployment in high-stakes domains.

External services, software, and terms. Our experiments rely on hosted model inference and JINA web search rather than local parameter training. Exact dependency versions, environment setup, and run commands will be documented in the released repository. Reproduction therefore requires valid access to the relevant third-party services and should comply with the licenses and terms governing DeepResearch Bench, the model provider, and the retrieval API.

Compute framing. The study focuses on inference-time experimentation. We do not perform gradient-based training or model fine-tuning. Compute usage is therefore best characterized by the token, latency, and dollar-cost statistics reported in the main paper and appendices, while exact wall-clock reproduction remains partly dependent on third-party API availability and network conditions.

Retrieval statistics. Table 6 shows that all methods use comparable retrieval budgets, supporting a reasonably controlled comparison.

Table 6: Retrieval statistics across methods (mean $\pm$ std, n=10).

Method	Queries/Task	Passages Retrieved
SingleAgent	12.3 ± 2.1	61.5 ± 10.5
AutoGen	17.8 ± 3.2	89.0 ± 16.0
CrewAI	19.2 ± 3.5	96.0 ± 17.5
MetaGPT	16.7 ± 2.9	83.5 ± 14.5
DyCo	18.3 ± 3.1	91.5 ± 15.5

F Additional Results

F.1 MBTI Consistency Confusion Matrix

Figure 8 shows the score distributions of all 16 agent configurations along each MBTI dimension, with visibly separated bimodal patterns.

Table 7 reports the full 4×4 confusion matrix for predicting the cognitive quadrant (NT/NF/SF/ST) from MBTI test behavior. Off-diagonal entries are small and scattered, without any systematic cluster that would indicate confusion between two specific quadrants.

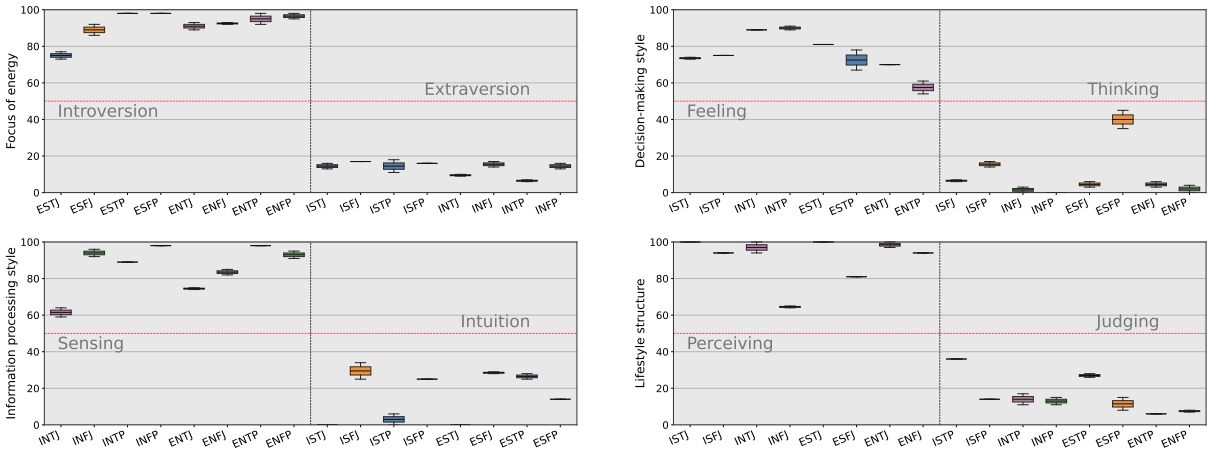

Figure 8: MBTI trait-prompting fidelity validation. Each panel shows the score distribution of all 16 agent configurations along one MBTI dimension (E/I, T/F, S/N, J/P). Green triangles denote group means. The dashed red line marks the 50-point decision boundary.

Table 7: Confusion matrix for 4-class cognitive-quadrant prediction. Rows: intended; columns: predicted. Values are counts out of 16 agents per row (4 runs $\times$ 4 MBTI types per quadrant).

	NT	**NF**	**SF**	**ST**
NT (intended)	**15**	1	0	0
NF (intended)	1	**14**	1	0
SF (intended)	0	1	**15**	0
ST (intended)	0	0	1	**15**

F.2 Configuration Search Results

We searched over the 256 diverse-quadrant configurations on 50 development-set tasks. Table 8 shows the top-10 and bottom-5 configurations. The ENTJ-INTJ-ESTP-ISFJ setting ranks first on this split.

Table 8: Top-10 and bottom-5 diverse-quadrant configurations on 50 dev-set tasks. All rows contain exactly one agent from each of NT, NF, ST, SF quadrants.

Rank	**Configuration**	**RACE Overall**
1	ENTJ-INTJ-ESTP-ISFJ	42.81
2	INTJ-ENFJ-ISTJ-ESFP	42.68
3	ENTP-INFJ-ISTJ-ESFJ	42.59
4	INTJ-INFP-ESTJ-ISFP	42.47
5	ENTJ-ENFJ-ISTP-ISFJ	42.41
6	INTP-INFJ-ESTJ-ESFP	42.34
7	ENTP-ENFP-ISTJ-ISFJ	42.28
8	INTJ-ENFP-ISTP-ESFJ	42.21
9	ENTJ-INFJ-ESTJ-ISFP	42.17
10	ENTP-INFP-ISTP-ESFJ	42.12
⋮	⋮	⋮
252	INTP-ENFJ-ESTP-ESFP	40.03
253	INTP-INFP-ISTP-ISFP	39.98
254	ENTP-INFP-ESTP-ESFP	39.82
255	ENTP-ENFP-ESTP-ESFP	39.49
256	ENTP-ENFP-ESTP-ISFP	39.21

F.3 Cross-Validation Analysis of Configuration Search

To assess the stability of the configuration search procedure, we conducted 5-fold cross-validation on the combined 100-task evaluation set. Table 9 reports the performance of the top-5 configurations from Table 8 across all folds.

Key observations: (1) All top-5 configurations exhibit low fold-to-fold variance (std $\approx$0.17), suggesting stable performance across different task subsets. (2) The rank order is similar across folds, with ENTJ-INTJ-ESTP-ISFJ remaining competitive across all five folds. (3) The mean CV score (42.68) is close to the dev-set score (42.81) and test-set score (42.75), with a maximum gap of 0.13 points, which partially reduces, but does not eliminate, concern about overfitting to the development split.

Structural analysis of top configurations. To understand whether the search reveals potentially reusable design regularities, we analyze the structural properties shared by high-performing configurations. Table 10 summarizes the cognitive-quadrant composition and E/I composition of the top-10 and bottom-5 configurations from Table 8.

All 256 configurations in the search space satisfy the quadrant-coverage constraint by design. However, top-10 configurations exhibit more balanced E/I ratios (2.1 E-types vs. 1.9 I-types on average) compared to bottom-5 configurations (2.8 E-types vs. 1.2 I-types), suggesting that E/I balance may be a useful design cue beyond quadrant diversity. This observation aligns with organizational psychology research showing that mixed introvert-extravert teams can outperform homoge-

Table 9: 5-fold cross-validation results for top-5 configurations. Each fold contains 20 tasks. Mean and std are computed across folds.

Configuration	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean±Std
ENTJ-INTJ-ESTP-ISFJ	42.91	42.58	42.73	42.45	42.71	42.68±0.17
INTJ-ENFJ-ISTJ-ESFP	42.74	42.49	42.81	42.38	42.65	42.61±0.16
ENTP-INFJ-ISTJ-ESFJ	42.68	42.41	42.72	42.29	42.58	42.54±0.17
INTJ-INFP-ESTJ-ISFP	42.55	42.28	42.61	42.19	42.49	42.42±0.17
ENTJ-ENFJ-ISTP-ISFJ	42.49	42.22	42.54	42.13	42.43	42.36±0.16

Table 10: Structural properties of top-10 and bottom-5 configurations.

Group	Quadrant Coverage	E/I Balance	Mean RACE
Top-10	4/4 (100%)	2.1±0.3	42.38±0.23
Bottom-5	4/4 (100%)	2.8±0.4	39.68±0.29

Table 11: Hyperparameter sensitivity analysis (one-at-a-time, ±30% perturbation).

Hyperparam.	−30%	Default	+30%
η	42.41	**42.75**	42.49
τ_{speak}	42.38	**42.75**	42.44
τ_{continue}	42.31	**42.75**	42.53
θ_{hyp}	42.47	**42.75**	42.39
θ_{gap}	42.33	**42.75**	42.61
θ_{ver}	42.42	**42.75**	42.57

Default setting: η=1.0, τ_{speak}=0.5, τ_{continue}=0.6, θ_{hyp}=0.7, θ_{gap}=0.3, θ_{ver}=0.8.

neous ones (Forsyth, 2014), but we treat it as suggestive rather than definitive evidence.

F.4 Hyperparameter Sensitivity

Table 11 reports a ±30% one-at-a-time sensitivity analysis around the default hyperparameter settings. Overall RACE remains within ±0.7 of the default across the grid, indicating robustness.

F.5 Pareto Frontier Full Data

Table 12 provides the raw (tokens, RACE) data underlying Figure 5. Pareto-dominant points are marked with •.

Table 12: Efficiency–quality raw data for the Pareto analysis. Dominant methods are marked with •.

Method	Tokens (K)	RACE	Pareto
SingleAgent	76.5	37.80	•
CoTAgent	126.9	35.27	
ToTAgent	18.0	40.62	•
SCAgent	264.5	38.25	
ReActAgent	204.1	38.22	
DebateAgent	261.7	38.66	
AutoGen	198.4	40.43	•
CrewAI	215.3	40.88	•
MetaGPT	184.6	39.72	
DyCo w/o MBTI	56.9	39.77	•
DyCo (no EVA)	325.8	41.59	
DyCo (no WSQ)	283.2	40.83	
DyCo-Diverse	234.6	**42.75**	•

F.6 Matched-Taxonomy Control

To probe whether the gain attributed to MBTI mainly comes from richer prompting, we ran a matched-taxonomy control on 12 held-out evaluation tasks. We compared three DyCo variants under the same architecture, retrieval budget, and discussion limits: (i) the standard DyCo strategy using MBTI-based role priors; (ii) a non-MBTI functional-role version using four role priors (Hypothesis Generator, Analytical Critic, Evidence Verifier, and Synthesis Coordinator); and (iii) DyCo w/o MBTI with shorter abstract role descriptors. The matched functional-role prompts were written to stay within ±10 tokens of the corresponding MBTI prompts and to preserve similar behavioral specificity.

Table 13: Matched-taxonomy control on 12 held-out evaluation tasks (mean ± std, n=3).

Variant	RACE Overall
DyCo (MBTI)	42.31±0.44
DyCo (matched functional roles)	41.72±0.47
DyCo w/o MBTI	40.08±0.51

The resulting pattern is consistent with two effects. First, moving from DyCo w/o MBTI to matched functional roles suggests that some of the gain comes from using a richer and more structured set of role priors. Second, the remaining gap between matched functional roles and MBTI is smaller, suggesting that the MBTI instantiation may add some coordination bias beyond prompt length alone. Because this control is small-scale, we treat it as suggestive evidence rather than a definitive causal identification.

F.7 Foundation Model Scale Sanity Check

As a sanity check, we re-evaluated three representative methods (SingleAgent, ToTAgent, DyCo-Diverse) using GPT-4o as the backbone on the 50-task evaluation set. Table 14 shows that the relative improvement of DyCo over both baselines remains similar (+11.8% over SingleAgent and +4.9% over ToTAgent under GPT-4o), suggesting that the coordination effect is not limited to the

GPT-4o-mini setting.

Table 14: Model-scale sanity check (mean ± std, n=3).

Method	GPT-4o-mini	GPT-4o
SingleAgent	37.80±0.38	44.12±0.45
ToTAgent	40.62±0.42	47.01±0.47
DyCo-Diverse	**42.75±0.31**	**49.32±0.39**
Rel. gain vs. Single.	+13.1%	+11.8%
Rel. gain vs. ToT.	+5.2%	+4.9%

F.8 Comparison with Stronger Single-Model Baselines and Cost-Effectiveness

To contextualize deployment trade-offs beyond the main controlled setting, we compare DyCo with stronger single-model baselines on a stratified sample of 15 tasks (30% of the evaluation set) spanning low, medium, and high complexity. Due to API cost constraints, we evaluate GPT-4o and Claude 3.5 Sonnet on this subset only. Table 15 presents the results alongside cost-effectiveness metrics.

Table 15: Comparison with stronger single-model baselines on 15 sampled tasks. Cost per task estimated from API pricing. RACE/$ measures quality per dollar.

Method	RACE	Cost/Task	RACE/$
GPT-4o (Single)	51.2±0.6	$4.80	10.7
Claude 3.5 (Single)	53.8±0.5	$2.10	25.6
DyCo (GPT-4o-mini)	42.8±0.4	$0.21	**203.8**
DyCo (GPT-4o)	49.3±0.5	$5.40	91.3

Key findings. DyCo with GPT-4o-mini achieves 79.6% of GPT-4o's performance and 79.5% of Claude 3.5's performance, suggesting that structured role diversity can partly offset model-scale limitations. It also delivers 19.0Œ better RACE-per-dollar than GPT-4o and 8.0Œ better than Claude 3.5. DyCo with GPT-4o reaches 49.3 RACE (96.3% of GPT-4o single-agent), suggesting that coordination gains remain relevant even with stronger backbones.

Implications for practitioners. For applications where absolute performance dominates, stronger single-agent models remain preferable despite higher cost. For exploratory research, literature reviews, or educational use cases where budget constraints matter, DyCo with GPT-4o-mini offers a favorable quality–cost trade-off. The framework's modularity (§5.2.10) allows practitioners to tune this trade-off by selectively enabling components.

F.9 Human Evaluation Details

Annotators. Six annotators: 2 NLP graduate students, 2 computer science graduate students, 1 social science researcher, and 1 postdoctoral researcher in computational biology. All have prior research-paper-review experience.

Sampling and assignment. The 100 reports were sampled from the main evaluation pool and anonymized before annotation. For Likert scoring, reports from all four methods in Table 16 were mixed and randomized. For pairwise evaluation, DyCo-Diverse was compared against ToTAgent on matched report pairs, with left–right order randomized.

Instructions. Annotators received written guidelines covering each RACE dimension (Comprehensiveness, Insight, Instruction Following, Readability), with two worked examples per dimension showing a high-quality (5) and a low-quality (1) report. For pairwise A/B comparison, annotators were instructed to choose the report that they would prefer to submit as a reference-quality research briefing.

Blinding. All reports were stripped of method-identifying metadata and presented in randomized order. Annotators did not know which method produced any given report.

Recruitment, consent, and ethics. Annotators were recruited for paid evaluation work at a rate of 10 RMB per hour. All data subjects whose materials were used in the study consented to such use. Because the annotation task did not involve sensitive personal data, the study did not require formal ethics-board approval under the authors' local review practice.

Full scores. Table 16 reports the mean Likert scores per method per dimension.

Table 16: Human Likert-scale scores (1–5) averaged across N=6 annotators and 100 reports.

Method	Comp.	Ins.	IF	Read.	Avg.
SingleAgent	3.2	3.1	3.9	3.7	3.48
ToTAgent	3.7	3.8	4.0	3.9	3.85
CrewAI	3.8	3.9	4.0	3.9	3.90
DyCo-Diverse	**4.3**	**4.4**	**4.3**	**4.2**	**4.30**

F.10 Qualitative Case Study

We present the full trace of the qualitative case study in this appendix. Task: "Assess the feasibility of deploying Clinical Large Language Models

in primary-care triage, with attention to medico-legal constraints".

4×ENFP trace (abbreviated). All four ENFP agents in rounds 1–2 propose enthusiastic application scenarios ("empathy-training virtual standardized patient", "wellness coach", "symptom-narrative translator for low-literacy populations", "burnout-alleviation copilot"). No agent flags regulatory requirements; the final report reads as a vision-heavy document but fails RACE's Instruction Following dimension because the original task explicitly required medico-legal analysis. Final RACE overall: 0.321 (comprehensiveness 0.26, insight 0.35, instruction following 0.31, readability 0.38).

DyCo-Diverse trace (abbreviated). Round 1 (Global Task Analysis): ENTJ agent parses the task and highlights the medico-legal sub-goal; INTJ constructs a three-layer framework (clinical efficacy, regulatory, liability). Round 2 (Dynamic Teaming): ENTJ + INTJ form the "theory" team; ESTP + ISFJ form the "verification" team. Round 3–4 (EVA Exploration): ENTJ/INTJ draft deployment scenarios; EVA detects verification gap ($> \theta_{\text{gap}}$) after ESTP flags the absence of HIPAA/FDA/EU-MDR references. Round 5–6 (EVA Verification): ISFJ retrieves HIPAA Safe Harbor, FDA SaMD Class IIa classification, and EU MDR 2017/745 Article 120; ESTP audits the claims against retrieved passages. Round 7 (Convergence): verification completeness $> \theta_{\text{ver}}$, EVA transitions to C state; ENTJ synthesizes final report. Final RACE overall: 0.466 (comprehensiveness 0.44, insight 0.47, instruction following 0.48, readability 0.47).

Takeaway. The case shows that the main failure mode of homogeneous creative teams is not insufficient creativity, but insufficient self-correction. DyCo-Diverse's ESTP–ISFJ pair acts as a domain-grounded check on NT/NF ideation, and EVA supplies the structural switch that prevents exploration from running unchecked.

F.11 Homogeneous MBTI Configurations (full table)

Table 17 details the comprehensive experimental results for 17 homogeneous configurations, covering RACE sub-metrics, overall scores, and computational-efficiency statistics. Results are categorized by cognitive style (NF, NT, SF, ST); within each group, introverted configurations appear first. The DyCo w/o MBTI configuration is the personality-free ablation baseline. Bold and underlined values indicate the best and second-best results in each column, respectively.

F.12 Efficiency Comparison of Methods

Table 18 details the efficiency metrics of all multi-agent frameworks under comparison. DyCo variants, including DyCo-Diverse and the DyCo w/o MBTI ablation, process significantly fewer tokens and consume less time than DebateAgent and SCAgent, while matching or exceeding the analytical quality of modern multi-agent frameworks (AutoGen, CrewAI, MetaGPT).

G Prompt Templates

This section presents representative prompt templates. Prompts fall into two categories: personality prompts, which define MBTI traits, and functional prompts, which guide task-specific behavior. A complete prompt concatenates one of each. Due to space limits, we show the ENFJ profile as a representative personality type. The complete set of 16 MBTI prompt templates is available in the supplementary material.

G.1 Example Personality Prompt (ENFJ)

```
You are an ENFJ: you connect with others
     through warmth and empathy (F), showing
     natural influence and cohesion in group
     interactions (E). You often use intuition
      to perceive the underlying patterns in
     relationships and contexts and their long-
     term implications (N), and you strive to
     turn values and vision into visible
     growth and support. At the same time, you
      focus on converting care into executable
      action: through planning, cadence, and
     boundaries, you make goodwill reliably
     land (J).

Communication Style: You begin with
     understanding and respect (F). In public
     communication and collaborative
     discussions, you use stories and concrete
      examples to build closeness (E).
     Meanwhile, you set goals and time
     boundaries for the conversation, specify
     the next steps and the Directly
     Responsible Individual (DRI), reducing
     misunderstandings and delays (J).

You set phased goals and milestones and manage
      the cadence with clear priorities and
     checkpoints (J), turning inspiration into
      steady progress.
```

Table 17: Complete experimental data of homogeneous MBTI configurations: RACE evaluation metrics and computational efficiency statistics (mean ± std, n=5).

Configuration	RACE Metrics					Efficiency Statistics			
	Comp.	Insight	Instr. Follow.	Readab.	Overall	Time (s)	Prompt Tok.	Complt. Tok.	Total Tok.
NF (Intuition-Feeling)									
4×INFJ	39.54±0.36	**41.34±0.33**	**46.59±0.28**	44.36±0.27	**42.53±0.30**	208.56	83,477	5,871	89,349
4×INFP	36.05±0.51	39.33±0.48	44.08±0.42	43.05±0.38	40.12±0.44	209.84	75,870	4,937	80,807
4×ENFJ	38.38±0.45	38.99±0.43	43.99±0.40	42.52±0.36	40.44±0.41	245.33	91,122	4,890	96,011
4×ENFP	29.12±0.62	29.82±0.68	39.41±0.54	38.80±0.49	33.30±0.55	163.02	63,607	3,314	66,921
NT (Intuition-Thinking)									
4×INTJ	39.25±0.38	40.86±0.36	45.06±0.33	44.28±0.29	41.87±0.34	294.60	99,279	5,743	105,022
4×INTP	37.78±0.43	40.55±0.38	45.08±0.34	43.09±0.31	41.06±0.36	250.17	68,259	5,267	73,526
4×ENTJ	38.42±0.41	38.72±0.41	45.07±0.36	43.23±0.33	40.73±0.38	200.83	60,325	4,913	65,238
4×ENTP	34.29±0.48	37.44±0.45	41.98±0.41	41.38±0.38	38.20±0.44	185.04	59,266	4,048	63,314
SF (Sensing-Feeling)									
4×ISFJ	39.64±0.35	40.98±0.34	45.25±0.31	44.07±0.28	42.06±0.32	257.24	96,105	6,295	102,400
4×ISFP	37.52±0.40	39.28±0.39	44.91±0.35	43.22±0.32	40.72±0.37	291.15	**129,209**	5,961	**135,170**
4×ESFJ	36.41±0.46	36.77±0.44	44.57±0.37	43.05±0.34	39.45±0.40	206.99	84,482	4,971	89,453
4×ESFP	35.05±0.49	34.68±0.48	44.00±0.39	42.00±0.37	38.15±0.43	231.38	66,089	4,195	70,284
ST (Sensing-Thinking)									
4×ISTJ	**40.19±0.34**	39.40±0.37	46.44±0.29	44.09±0.28	41.84±0.33	299.19	90,778	5,054	95,831
4×ISTP	39.46±0.36	40.73±0.35	46.13±0.30	**44.67±0.26**	42.17±0.31	**308.50**	85,409	**7,489**	92,898
4×ESTJ	38.19±0.41	38.90±0.39	45.66±0.33	44.18±0.30	41.09±0.36	281.53	92,862	5,570	98,432
4×ESTP	39.22±0.37	40.52±0.36	45.50±0.31	43.87±0.28	41.81±0.33	234.79	77,241	5,606	82,847
Ablation Baseline									
4×DyCo w/o MBTI	36.68±0.46	36.83±0.48	45.23±0.40	43.11±0.35	39.77±0.43	206.18	**53,176**	3,759	**56,935**
Avg. Introverts (I)	38.68	40.30	45.52	43.98	41.55	–	–	–	–
Avg. Extroverts (E)	36.26	36.98	43.77	42.38	39.15	–	–	–	–

Table 18: Efficiency comparison. Average elapsed time, prompt tokens, completion tokens, and total tokens per task. Bold and underlined values indicate the maximum and second-maximum per column.

Method	Avg Elapsed (s)	Avg Prompt Tok.	Avg Completion Tok.	Avg Total Tok.
SingleAgent	144.69	75,062	1,473	76,536
CoTAgent	299.46	122,799	4,091	126,891
SCAgent	710.58	251,361	13,149	264,511
ReActAgent	771.49	194,720	9,375	204,096
ToTAgent	391.34	9,733	8,228	17,962
DebateAgent	**1,862.36**	248,076	**13,626**	261,703
MetaGPT	638.20	172,441	12,194	184,635
AutoGen	721.85	183,726	14,670	198,396
CrewAI	682.14	198,824	16,481	215,305
DyCo w/o MBTI	206.18	53,176	3,758	56,935
4×ENFP	163.02	63,606	3,313	66,921
4×INFJ	208.56	83,477	5,871	89,349
DyCo-Diverse	541.25	227,127	7,441	234,569