

DyCo: Dynamic Cognitive Role Coordination for Open-Ended Multi-Agent Research Synthesis and Verification

Zhiyu Zhang¹ Leheng Wu¹ Yijun Mo^{1*} Mingchu Zhong²
Yulin Zhang¹ CHEN XI³

¹Huazhong University of Science and Technology

²Wuhan University of Technology

³Nanyang Technological University

{zhangzhiyu, leheng, moyj, u202315542}@hust.edu.cn

zhongmingchu@whut.edu.cn CHEN1944@e.ntu.edu.sg

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, literature-grounded research synthesis, where uncertainty changes what needs to be explored, challenged, and verified. We introduce DyCo, a framework for dynamic cognitive role coordination that 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% ($p < 0.01$, $d = 0.42$), and a budget-matched non-MBTI control shows that this gain is not explained by additional computation. Extended matched-role controls on all 50 evaluation tasks indicate that most of the improvement comes from rich, structured role priors combined with verification-oriented coordination rather than the MBTI taxonomy itself, and the diverse configuration is not significantly better than the strongest homogeneous one. We further present DyCo-Efficient, which removes dynamic teaming and preserves most of the observed quality at roughly one thirteenth of the token cost, positioning Full DyCo as an optional high-budget configuration.

1 Introduction

Large language models (LLMs) show strong capabilities in complex reasoning, code generation, and knowledge synthesis (Chowdhery et al., 2023; OpenAI, 2023; Gemini Team, 2023). DeepResearch-style systems that integrate planning, retrieval, and reasoning offer a useful paradigm for open-ended, literature-grounded re-

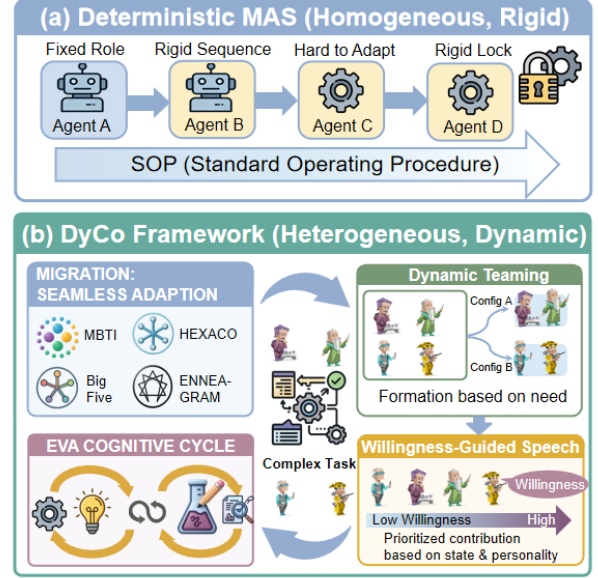


Figure 1: Comparison of a static MAS (upper) and DyCo (lower). Traditional role- and SOP-based organizations are homogeneous, rigid, and hard to re-configure for complex tasks. DyCo instead combines dynamic teaming, willingness-guided speaking, and exploration-verification alternation to support more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

search synthesis (Du et al., 2026; Zheng et al., 2025; Ye et al., 2026). 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; CrewAI Inc, 2024; LangChain Inc, 2024). Recent work has also explored communication protocols, dynamic teaming, and broader surveys of collaborative agents (Wang et al., 2025; Marro et al., 2024; Chung et al., 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 research synthesis, 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 research 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 research synthesis and verification under uncertainty. Throughout, our empirical scope is long-form, literature-grounded research synthesis—multi-stage planning, open-ended retrieval, cross-source integration, candidate explanation generation, critical examination, conflict resolution, and iterative verification—rather than scientific exploration in general. 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 research synthesis workflows that uses structured role priors 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, optional component. This yields two operating points: the low-cost DyCo-Efficient and the optional high-budget Full DyCo (§5.2.6).
- Empirical results on DeepResearch Bench: a diverse DyCo configuration outperforms the strongest multi-agent baseline by 4.6% ($p < 0.01$, $d = 0.42$). Extended matched-role controls on all 50 evaluation tasks indicate that

the gains are better explained by rich, structured role priors plus verification-oriented coordination than by the MBTI taxonomy itself, and that the advantage over the strongest homogeneous configuration is small and not statistically significant. A blinded pilot audit of factuality and source faithfulness complements the automatic metric.

2 Related Work

2.1 Research Synthesis and Multi-Agent Collaboration

Research synthesis systems integrate planning, retrieval, and reasoning for complex, open-ended literature-grounded workflows (Xu and Peng, 2025; Xi et al., 2026; Du et al., 2026; Wu et al., 2025; Ye et al., 2026). 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; CrewAI Inc, 2024; LangChain Inc, 2024; Wang et al., 2025; 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 (Wang et al., 2024; Hu and Collier, 2024; Park et al., 2023; Li et al., 2023; Shanahan et al., 2023; Tu et al., 2024; Samuel 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 coordination has been studied in multi-agent settings such as team-based navigation and multi-agent pathfinding (Chung et al., 2024). Human teams often evolve through viewpoint collision, reorganization, and consensus (Forsyth,

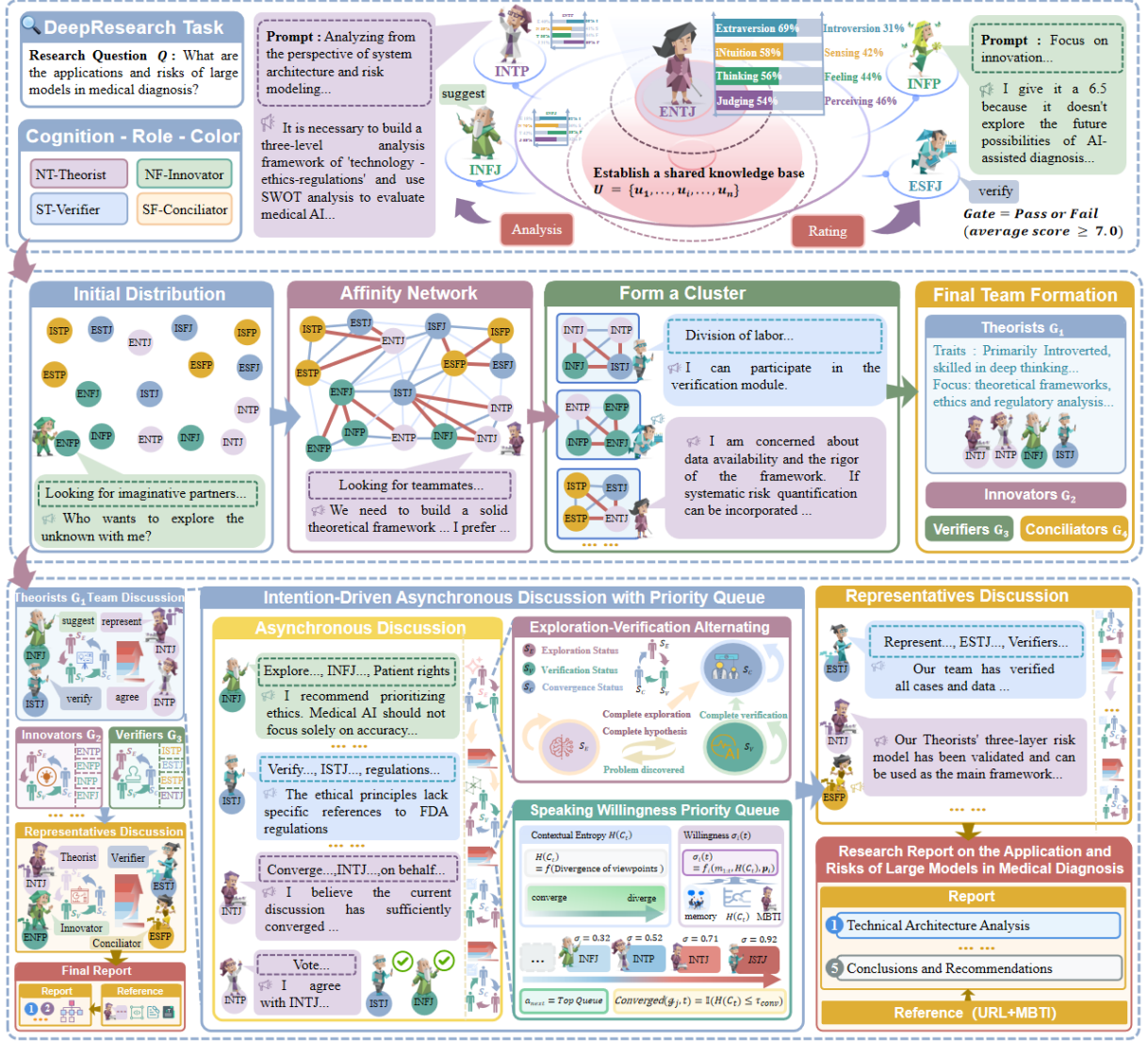


Figure 2: Overall architecture of DyCo. The upper part shows the mapping between research synthesis tasks and four cognitive roles (NT-Theorist, NF-Innovator, ST-Verifier, SF-Mediator) alongside the global discussion flow. The middle section details dynamic teaming via affinity-network clustering, while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

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

Prior work on problem-solving diversity and work-group composition suggests that heterogeneous groups can sometimes outperform more homogeneous groups, although the effect depends on task demands and team context (Hong and Page, 2004; van Knippenberg et al., 2004; 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.

The S/N axis governs information acquisition: Intuitive (N) agents prioritize abstract pat-

tern 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, and MBTI serves here as one concrete way to inject structured cognitive diversity into coordination. The central question is whether a discrete taxonomy can usefully parameterize coordination logic for uncertain research synthesis, not whether the label is psychologically privileged. In this setup, role priors affect EVA participation (Algorithm 3), speaking willingness (Eq. 6), and dynamic team formation (Algorithm 1).

4 Dynamic Cognitive Coordination Framework

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 $\mathcal{A} = \{a_1, \dots, a_n\}$, 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:

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

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

Inspired by prior work on quantitatively evaluating and shaping personality traits in LLMs (Serapio-García et al., 2025), we instantiate each agent a_i with a continuous MBTI role-prior 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 research 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

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

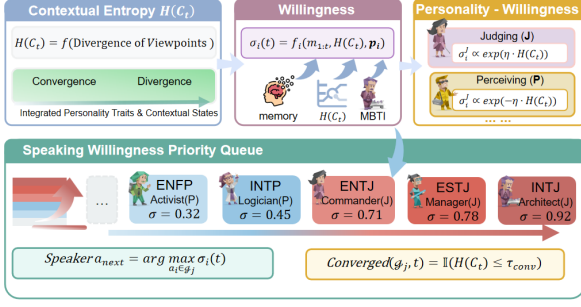


Figure 3: Willingness-driven speaking scheduling: speaking willingness is computed from personality and context, and speakers are selected by priority.

(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.

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

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 $\{\mathcal{G}_1, \dots, \mathcal{G}_k\}$

```

1: Initialize  $\mathbf{W} \in \mathbb{R}^{n \times n}$ 
2: for  $i = 1$  to  $n$  do
3:   for  $j = i + 1$  to  $n$  do
4:      $W_{ij} \leftarrow \alpha \cdot \text{sim}(u_i, u_j)$ 
5:      $W_{ij} \leftarrow W_{ij} + \beta \cdot \text{Compat}(\mathbf{p}_i, \mathbf{p}_j)$ 
6:      $W_{ij} \leftarrow W_{ij} + \gamma \cdot \text{Complement}(\kappa(a_i), \kappa(a_j))$ 
7:      $W_{ji} \leftarrow W_{ij}$   $\triangleright$  Symmetric matrix
8:   end for
9: end for
10: Compute degree matrix  $\mathbf{D} \leftarrow \text{diag}(\mathbf{W} \cdot \mathbf{1})$ 
11: Compute normalized Laplacian  $\mathbf{L} \leftarrow \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$ 
12: Extract top- $k$  eigenvectors  $\mathbf{V} \leftarrow \text{EigenDecomp}(\mathbf{L}, k)$ 
13:  $\{\mathcal{G}_1, \dots, \mathcal{G}_k\} \leftarrow \text{KMeans}(\mathbf{V}, k)$ 
14: for Each team  $\mathcal{G}_j$  do
15:   if  $|\{\kappa(a) : a \in \mathcal{G}_j\}| < \min(|\mathcal{G}_j|, 3)$  then
16:     Swap members to satisfy diversity
17:   end if
18: end for
19: return  $\{\mathcal{G}_1, \dots, \mathcal{G}_k\}$ 

```

tured report.

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).

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.

4.5 Exploration-Verification Alternation

5 Experiments

5.1 Experimental Setup

We first test whether the MBTI prompts produce distinguishable role behavior as a manipulation check for our role-prior strategy (§5.2.1), not as

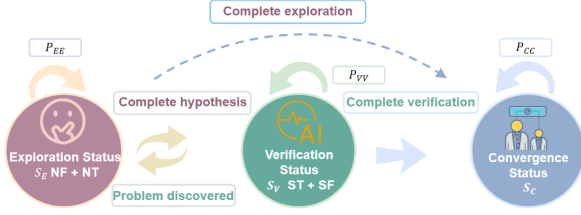


Figure 4: Exploration–Verification Alternation: three states, Exploration (E), Verification (V), and Convergence (C), with transitions determined by hypothesis quality and verification completeness.

evidence for MBTI as a psychological theory. We then evaluate DyCo on DeepResearch Bench against diverse baselines (§5.1.2), report matched role-prior controls on all 50 evaluation tasks that separate structured priors from prompt richness and compute budget (§5.2.3), and present component ablations (§5.2.6) and a quality–cost Pareto analysis (§5.2.8). Complete experimental data are available in Appendix F.

5.1.1 Datasets and Metrics

We evaluate the proposed framework on DeepResearch Bench (Du et al., 2026), a benchmark of long-form, literature-grounded research synthesis covering 22 domains that requires planning, open-ended retrieval, cross-source synthesis, argument construction, evidence checking, and report writing. We sample uniformly across all domains. RACE is the primary automatic quality metric, covering Comprehensiveness, Insight, Instruction Following, and Readability; it does not directly measure factual correctness, citation faithfulness, or evidence-grounded reasoning, and may partly reward fluent, structured reports. We therefore complement it with human evaluation and a blinded factuality audit (§5.2.4). 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 $\pm$ 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.

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., 2023),

ReAct (Yao et al., 2023b), ToT (Yao et al., 2023a)), a classical multi-agent baseline (Debate (Du et al., 2024)), and modern multi-agent frameworks (AutoGen (Wu et al., 2024), CrewAI (CrewAI Inc, 2024), 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 Wang et al. (2024); Serapio-García et al. (2025); Besta et al. (2025) by administering standardized MBTI tests to agents configured with specific profiles; this indicates that our prompting produces stable, distinguishable agent behaviors, not that LLMs possess genuine personality traits. As shown in Table 6 (Appendix F.1), 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), and 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 1 compares DyCo against all baselines on DeepResearch Bench, reporting mean $\pm$ std across 10 runs.

As shown in Table 1, DyCo-Diverse (ENTJ-INTJ-ESTP-ISFJ) achieves the best overall score (42.75), outperforming ToTAgent by 5.2% and CrewAI by 4.6% ($p < 0.01$). DyCo-Diverse also scores higher than every homogeneous configuration, including the strongest one, $4 \times$ INFJ (42.53), but this 0.22-point advantage is not statistically significant (95% CI $[-0.20, 0.64]$, $p=0.294$; §5.2.5), so we do not interpret it as evidence that cognitive diversity is the primary source of the gains. The short abstract-role ablation, which replaces role-rich prompts with shorter abstract descriptors while leaving the architecture unchanged, scores 39.77; because it simultaneously

Table 1: RACE on DeepResearch Bench (mean $\pm$ 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
<i>Single-Agent Baselines</i>						
CoTAgent	31.94 $\pm$ 0.48	30.56 $\pm$ 0.55	42.07 $\pm$ 0.41	40.22 $\pm$ 0.38	35.27 $\pm$ 0.41	-1.42
SingleAgent	34.38 $\pm$ 0.44	33.75 $\pm$ 0.52	44.91 $\pm$ 0.39	41.43 $\pm$ 0.36	37.80 $\pm$ 0.38	-0.81
ReActAgent	35.09 $\pm$ 0.46	35.93 $\pm$ 0.50	43.23 $\pm$ 0.43	41.55 $\pm$ 0.35	38.22 $\pm$ 0.39	-0.68
SCAgent	35.30 $\pm$ 0.42	36.51 $\pm$ 0.49	42.42 $\pm$ 0.45	41.80 $\pm$ 0.34	38.25 $\pm$ 0.40	-0.66
ToTAgent	37.18 $\pm$ 0.47	39.67 $\pm$ 0.51	44.50 $\pm$ 0.40	43.15 $\pm$ 0.37	40.62 $\pm$ 0.42	-0.05
<i>Multi-Agent Baselines</i>						
DebateAgent	35.81 $\pm$ 0.50	37.39 $\pm$ 0.54	42.41 $\pm$ 0.44	41.17 $\pm$ 0.40	38.66 $\pm$ 0.45	-0.50
MetaGPT	36.74 $\pm$ 0.53	38.28 $\pm$ 0.58	43.96 $\pm$ 0.46	42.01 $\pm$ 0.41	39.72 $\pm$ 0.49	-0.24
AutoGen	37.51 $\pm$ 0.51	39.03 $\pm$ 0.56	44.28 $\pm$ 0.45	42.53 $\pm$ 0.39	40.43 $\pm$ 0.47	-0.10
CrewAI	37.85 $\pm$ 0.49	39.42 $\pm$ 0.53	44.61 $\pm$ 0.43	42.76 $\pm$ 0.40	40.88 $\pm$ 0.45	0.00
<i>Ours — DyCo</i>						
DyCo (short abstract roles)	36.68 $\pm$ 0.46	36.83 $\pm$ 0.52	45.23 $\pm$ 0.38	43.11 $\pm$ 0.35	39.77 $\pm$ 0.43	-0.26
4 $\times$ INFJ	39.54 $\pm$ 0.36	41.34 $\pm$ 0.33	46.59$\pm$0.28	44.36 $\pm$ 0.27	42.53 $\pm$ 0.30**	0.37
DyCo-Diverse	40.33$\pm$0.32	41.65$\pm$0.29	<u>46.45$\pm$0.27</u>	44.51$\pm$0.25	42.75$\pm$0.31**	0.42

changes the role taxonomy, prompt length, and behavioral specificity, it does not isolate an MBTI-specific effect. We quantify this decomposition in §5.2.3.

5.2.3 Matched Role-Prior Controls

To separate the effect of structured role priors from prompt richness, behavioral specificity, and compute budget, we extended our matched functional-role control from 12 to *all* 50 held-out evaluation tasks, using three shared random seeds and matching the MBTI condition in backbone model, temperature, task input, number of agents, maximum rounds, retrieval limit, prompt length, behavioral specificity, and the complete DyCo coordination procedure. The control uses four non-MBTI functional roles (Hypothesis Generator, Analytical Critic, Evidence Verifier, Synthesis Coordinator). We additionally report two controls on the same 50 tasks: *rich non-taxonomic roles*, which replace the MBTI prompts with equally long behavior descriptions without the explicit taxonomy structure, and *role-shuffle*, which randomly permutes the role-to-stage mapping while keeping the original prompts. Together with the short abstract-role ablation, these conditions decompose the original ablation gap. We average across seeds within each task and run two-sided paired t -tests on the 50 task-level RACE Overall scores, with Holm correction over the four comparisons against the MBTI condition; Table 13 (Appendix F.7) reports unadjusted two-sided 95% paired CIs, Holm-adjusted p -values, and Cohen’s d_z .

Across the three shared seeds, the MBTI-instantiated DyCo achieves 42.66 (233,980 tokens per task). The matched functional-role condition achieves 42.18 (231,840 tokens): 0.48 points lower (95% CI [0.08, 0.88], Holm-adjusted $p=0.040$, $d_z=0.33$). Rich non-taxonomic roles achieve 42.07 (0.59 lower, CI [0.08, 1.10], $p_{\text{adj}}=0.028$) and role-shuffle 41.83 (0.83 lower, CI [0.29, 1.37], $p_{\text{adj}}=0.012$). The short abstract-role ablation achieves 39.81 (56,934 tokens), 2.85 points lower (CI [2.11, 3.59], $p_{\text{adj}}<0.001$, $d_z=1.07$). The functional-role condition thus recovers 2.37 points, approximately 83% of the gap to the MBTI condition: prompt richness, behavioral specificity, and functional differentiation explain most of the original ablation gap, and the MBTI instantiation adds only a small, statistically detectable increment.

These results do not support a strong MBTI-specific causal claim: rich, structured, behaviorally distinct role priors combined with a verification-oriented coordination mechanism improve long-form, literature-grounded research synthesis, and MBTI provides one interpretable and controllable instantiation of such priors rather than a necessary taxonomy or evidence for its psychometric validity. A budget-matched comparison further rules out compute as the sole explanation: the functional-role control uses 18.1 retrieval calls and 231,840 tokens per task versus 18.3 calls and 234,568 tokens for DyCo-Diverse, yet DyCo-Diverse remains 0.57 points higher (95% CI [0.19, 0.95], adjusted $p=0.018$, $d_z=0.42$).

5.2.4 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, 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%]; $p < 0.001$).

Factuality and source faithfulness. Because RACE does not directly measure factual correctness or citation faithfulness, we additionally conducted a method-identity-blinded pilot audit over 10 shared tasks and eight methods, covering 800 atomic factual claims. Two annotators independently assessed each claim and a third resolved disagreements (pre-adjudication Cohen’s $\kappa = 0.72$). DyCo-Diverse achieved 81% factual accuracy and 79% source faithfulness, versus 79% and 76% for CrewAI, suggesting that higher RACE scores are not accompanied by substantially lower factual quality. We treat this as a pilot; a larger-scale blinded audit with hierarchical statistical analysis remains future work.

5.2.5 Comparison of Different Personalities

As a secondary diagnostic, we evaluate 17 homogeneous configurations (all 16 MBTI types plus the short abstract-role ablation). 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.13. Notably, the strongest homogeneous configuration, 4×INFJ (42.53), performs close to DyCo-Diverse (42.75); the 0.22-point difference is not statistically significant (95% CI [−0.20, 0.64], uncorrected $p=0.294$, $d_z=0.15$), so the current evidence does not establish that heterogeneous teams significantly outperform the strongest homogeneous teams on this benchmark.

5.2.6 Fine-grained Ablation Study

To dissect the impact of each core component, we report single-module and pairwise ablations in Table 2.

The ablation results show three main patterns. First, removing the rich role priors causes the largest drop ($\Delta = -2.98$), followed by WSQ ($\Delta = -1.92$) and EVA ($\Delta = -1.16$); per §5.2.3, most of this drop is recovered by matched non-MBTI functional roles, so it reflects the value of rich, structured role priors rather than MBTI specifically. Second, removing EVA sharply increases token use (325K vs. 234K) and wall-clock time (973 s vs. 541 s), suggesting that EVA keeps collaboration focused. Third, the no-DT configuration defines **DyCo-Efficient**: it removes dynamic teaming while retaining structured role priors, EVA, GTA, and WSQ, and achieves 42.39 RACE with only 18,471 tokens and 230.3 s per task. Full DyCo improves over DyCo-Efficient by only 0.36 RACE points while using $\sim 12.7\times$ as many tokens and incurring $\sim 2.35\times$ the latency. We therefore position DyCo-Efficient as the recommended default for cost-sensitive applications, and Full DyCo, which additionally retains dynamic teaming, multi-stage group discussions, and representative negotiation, as an optional high-budget module. The pairwise ablations point to complementary contributions from dynamic teaming and EVA.

5.2.7 Fairness and Comparability Checks

To improve comparability, we re-ran the strongest baselines with prompt lengths matched to DyCo within ± 10 tokens per agent; the ranking remained

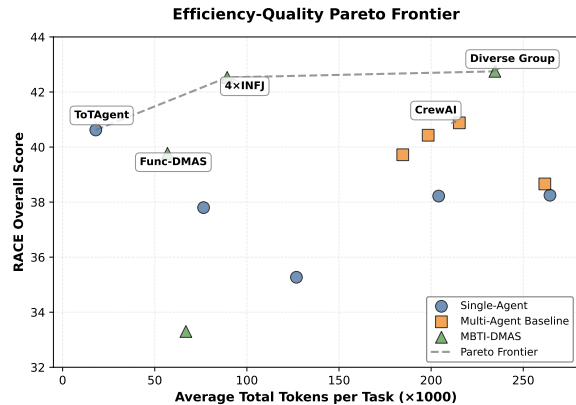


Figure 5: Efficiency–quality Pareto frontier. X-axis: average total tokens per task (log scale); Y-axis: RACE overall score. Error bars show ± 1 std.

Table 2: Fine-grained ablation of DyCo (mean $\pm$ std, $n=5$). DT: Dynamic Teaming; EVA: Exploration–Verification Alternation; GTA: Global Task Analysis; WSQ: Willingness-based Speaking Queue; Roles: MBTI-instantiated rich role priors ($\times$ replaces them with short abstract descriptors, i.e., the short abstract-role ablation). The no-DT configuration corresponds to DyCo-Efficient.

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

unchanged, and DyCo-Diverse retained a +1.21 RACE margin over CrewAI on a 12-task control subset. Participation statistics within DyCo-Diverse show no extreme dominance by any single cognitive quadrant (average token shares of 26%, 24%, 27%, and 23% for NT, NF, ST, and SF; representative selection rates 21–29%). 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. Finally, a domain-robustness check on STEM-heavy and non-STEM subsets (Appendix F.5) shows the trend is directionally preserved under moderate domain shift; these subsets share the same underlying task structure, however, and are not evidence of transfer to scientific tasks in general.

5.2.8 Efficiency–Quality Pareto Analysis

Figure 5 summarizes the efficiency–quality trade-off. DyCo-Efficient (no dynamic teaming; 18.5K tokens, 42.39 RACE) lies on the Pareto frontier and preserves most of the observed quality at a fraction of the cost, making it the recommended default. Full DyCo (DyCo-Diverse) occupies the high-quality but considerably more expensive end of the frontier (234.6K tokens, 42.75 RACE) and is best viewed as an optional high-budget configuration; the short abstract-role ablation is cheaper still but lower in quality. Because the task-averaged results do not reveal whether this gain is concentrated in a small subset of tasks, unconditionally enabling the module may be inefficient; budget-adaptive activation is left as future work.

6 Conclusion

This paper presents DyCo, a framework for dynamic cognitive role coordination in open-ended, literature-grounded research synthesis and verification. On DeepResearch Bench, DyCo-Diverse reaches a RACE score of 42.75, outperforming the strongest multi-agent baseline by 4.6% ($p < 0.01$, $d = 0.42$). Extended matched controls on all 50 evaluation tasks indicate that most of this gain comes from rich, structured role priors integrated into a verification-oriented coordination process, with MBTI adding only a small increment (0.48 RACE points) over matched functional roles; the diverse configuration is not significantly better than the strongest homogeneous one (4 $\times$ INFJ). These results support structured role priors and verification-oriented coordination for literature-grounded research synthesis; they do not establish the psychometric validity of MBTI, nor transfer to tasks with substantially different action spaces and correctness criteria, such as theorem proving, wet-lab experimental design, or software engineering. DyCo-Efficient preserves most of the observed quality at roughly one thirteenth of the token cost of Full DyCo, an optional high-budget configuration; budget-adaptive activation of dynamic teaming under uncertainty, source conflicts, or verification gaps is left to future work. For replication, we release the complete codebase, prompt templates, and experimental configurations at <https://github.com/fish-wc/DyCo>.

Limitations

Scope of claims. MBTI is used here as an engineering device, not a psychological claim. Our ex-

tended matched controls (§5.2.3) show that a non-MBTI functional-role instantiation recovers most of the performance gap, with an MBTI-specific increment of only 0.48 RACE points ($d_z=0.33$); whether other role taxonomies (e.g., Big Five) would work as well or better remains open. In addition, the difference between DyCo-Diverse (42.75) and the strongest homogeneous configuration, 4×INFJ (42.53), is not statistically significant ($p=0.294$, $d_z=0.15$), so we do not claim that cognitive diversity across quadrants is the primary source of the gains.

Evaluation scope and metric. Our evaluation is limited to long-form, literature-grounded research synthesis; transfer to tasks with substantially different action spaces and correctness criteria (e.g., theorem proving, wet-lab design, software engineering) is untested, and the STEM/non-STEM subsets should not be read as evidence of broad transfer. RACE also does not directly measure factual correctness or citation faithfulness, and DyCo’s longer, more structured outputs may partly favor its style; our blinded factuality audit (§5.2.4) covers only 10 tasks and 800 claims, and a full-scale audit remains future work.

Cost and intended use. The ENTJ-INTJ-ESTP-ISFJ configuration came from a limited development-set search and represents one point in a broader diverse-quadrant design family rather than a unique optimum. Full DyCo is also expensive: dynamic teaming adds only 0.36 RACE over DyCo-Efficient while consuming roughly $12.7\times$ the tokens and $2.35\times$ the latency, so it is best treated as an optional high-budget module; budget-adaptive activation is unvalidated future work. Finally, DyCo is intended for exploratory research synthesis, not high-stakes decision-making without expert oversight; like other long-form research agents, it can produce fluent but incomplete or incorrect syntheses, and outputs and cited evidence should be reviewed before downstream use.

References

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*. *CoRR*, abs/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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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*. *J. Mach. Learn. Res.*, 24:240:1–240:113.

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*. *Artif. Intell. Rev.*, 57(2):41.

CrewAI Inc. 2024. *CrewAI*. GitHub repository. <https://github.com/crewAIInc/crewAI>.

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.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. 2026. *DeepResearch Bench: A comprehensive benchmark for deep research agents*. In *International Conference on Learning Representations*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. *Improving factuality and reasoning in language models through multiagent debate*. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR / OpenReview.net.

Jonathan St. B. T. Evans and Keith E. Stanovich. 2013. *Dual-process theories of higher cognition: Advancing the debate*. *Perspectives on Psychological Science*, 8(3):223–241. PMID: 26172965.

Donelson R. Forsyth. 2014. *Group Dynamics*, 6th edition. Wadsworth Cengage Learning, Belmont, CA.

Gemini Team. 2023. *Gemini: A family of highly capable multimodal models*. *CoRR*, abs/2312.11805.

- 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](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 8048–8057. ijcai.org.
- Lu Hong and Scott E. Page. 2004. [Groups of diverse problem solvers can outperform groups of high-ability problem solvers](#). *Proceedings of the National Academy of Sciences*, 101(46):16385–16389.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xianwu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. [MetaGPT: Meta programming for A multi-agent collaborative framework](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Tiancheng Hu and Nigel Collier. 2024. [Quantifying the persona effect in LLM simulations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.
- Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- LangChain Inc. 2024. [LangGraph](#). GitHub repository. <https://github.com/langchain-ai/langgraph>.
- 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 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Samuele Marro, Emanuele La Malfa, Jesse Wright, Guohao Li, Nigel Shadbolt, Michael J. Wooldridge, and Philip Torr. 2024. [A scalable communication protocol for networks of large language models](#). *CoRR*, abs/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.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Joon Sung Park, Joseph C. 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, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.
- 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.
- Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. [PersonaGym: Evaluating persona agents and LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.
- Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. [A psychometric framework for evaluating and shaping personality traits in large language models](#). *Nature Machine Intelligence*, 7(12):1954–1968.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. [Role play with large language models](#). *Nat.*, 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.
- 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.
- Daan van Knippenberg, Carsten K. W. De Dreu, and Astrid C. Homan. 2004. [Work Group Diversity and Group Performance: An Integrative Model and Research Agenda](#). *Journal of Applied Psychology*, 89(6):1008–1022.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. [InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews](#). In *Proceedings of the*

- 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Yue Wang. 2024. [AI and MBTI: A synergistic framework for enhanced team dynamics](#). *CoRR*, abs/2409.15293.
- Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, and Min Zhang. 2025. [Agent-Dropout: Dynamic agent elimination for token-efficient and high-performance LLM-based multi-agent collaboration](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24013–24035, Vienna, Austria. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025. [Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28489–28503, Vienna, Austria. Association for Computational Linguistics.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. [Auto-gen: Enabling next-gen LLM applications via multi-agent conversation](#). In *COLM 2024*.
- Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu, and Weinan Zhang. 2026. [A survey of large language model-based search agents](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8244–8279, San Diego, California, United States. Association for Computational Linguistics.
- Renjun Xu and Jingwen Peng. 2025. [A comprehensive survey of deep research: Systems, methodologies, and applications](#). *CoRR*, abs/2506.12594.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. [ReAct: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Fangda Ye, Kuicai Dong, Xie Zhifei, Yuxin Hu, Yihang Yin, Shurui Huang, Shikai Dong, Chen Zhang, Jianzhu Bao, and Shuicheng Yan. 2026. [Deep-reporter: Deep research for grounded multimodal long-form generation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 41137–41177, San Diego, California, United States. Association for Computational Linguistics.
- Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. [DeepResearcher: Scaling deep research via reinforcement learning in real-world environments](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431, Suzhou, China. Association for Computational Linguistics.

A Notation

Table 3 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 4 outlines the core strengths, limitations, and suitable tasks for each cognitive group, ensuring that the collaborative process covers all necessary cognitive functions.

Table 3: Notation used in the DyCo framework, grouped by functional role.

Symbol	Meaning
<i>Task and System</i>	
$\mathcal{T}$	Research synthesis 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 $\{a_1, \dots, a_n\}$
$Q(\cdot)$	RACE-based report quality
$\mathcal{H}(\mathcal{A})$	Cognitive Diversity Index (Eq. 2)
$\text{Eff}(\mathcal{S})$	Efficiency cost (time, tokens)
<i>Personality and Role</i>	
$\mathbf{p}_i$	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
<i>Dynamic Teaming</i>	
$\mathbf{W} \in \mathbb{R}^{n \times n}$	Agent affinity matrix
α, β, γ	Affinity weights
$\{\mathcal{G}_1, \dots, \mathcal{G}_k\}$	Team partition
τ_{div}	Cognitive diversity threshold
<i>EVA and Willingness</i>	
$\mathbb{H}$	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

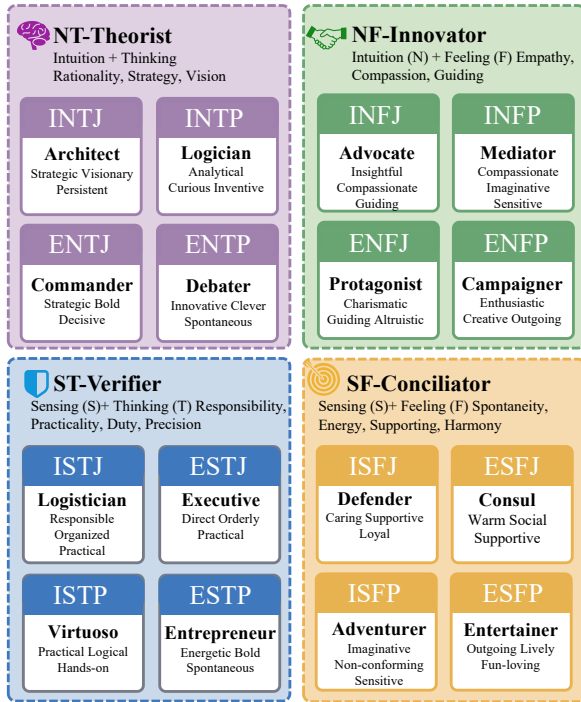


Figure 6: Overview of the 16 MBTI agent types, grouped into four cognitive clusters via the intuition/sensing (N/S) and thinking/feeling (T/F) dimensions.

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 $\mathcal{R} : \mathcal{A} \rightarrow \Omega$

```

1:  $\omega_{\mathcal{T}} \leftarrow \text{ClassifyTask}(\mathcal{T})$ 
2: for Each agent  $a_i \in \mathcal{A}$  do
3:   Extract personality vector  $\mathbf{p}_i$ 
4:    $\kappa_i \leftarrow \text{sign}(p_i^{SN}) \oplus \text{sign}(p_i^{TF})$ 
5:   for Each cognitive task  $\omega \in \Omega$  do
6:     Calculate role adaptability  $\rho_{\kappa_i}(a_i, \omega)$ 
7:   end for
8:    $\mathcal{R}(a_i) \leftarrow \arg \max_{\omega} \rho_{\kappa_i}(a_i, \omega)$ 
9: end for
10: 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-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:

Table 4: Cognitive Type and Research 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

Algorithm 3 Exploration–Verification Alternation (EVA)

Require: Task $\mathcal{T}$, agent set $\mathcal{A}$, max rounds $T_{\max}$
Ensure: Final solution s_{final}

```

1:  $m \leftarrow \text{Explore}$ ;  $\mathbb{H} \leftarrow \emptyset$ ;  $\mathcal{V} \leftarrow \emptyset$ 
2: for  $t = 1$  to  $T_{\max}$  do
3:   if  $m = \text{Explore}$  then
4:      $\mathcal{A}_{\text{active}} \leftarrow \{a_i : \kappa(a_i) \in \{\text{NF}, \text{NT}\}\}$ 
5:      $w_i \leftarrow \rho_{\kappa(a_i)}(a_i, \omega_{\text{IH}}) + \rho_{\kappa(a_i)}(a_i, \omega_{\text{RA}})$ 
6:     for Each  $a_i \in \mathcal{A}_{\text{active}}$  do
7:        $h_i \leftarrow f_{\text{generate}}(a_i, \mathcal{T}, \mathbb{H})$ 
8:        $\mathbb{H} \leftarrow \mathbb{H} \cup \{h_i\}$ 
9:     end for
10:    if HypothesisQuality( $\mathbb{H}$ )  $> \theta_{\text{hyp}}$  then
11:       $m \leftarrow \text{Verify}$ 
12:    end if
13:  else
14:     $\mathcal{A}_{\text{active}} \leftarrow \{a_i : \kappa(a_i) \in \{\text{SF}, \text{ST}\}\}$ 
15:     $w_i \leftarrow \rho_{\kappa(a_i)}(a_i, \omega_{\text{EF}}) + \rho_{\kappa(a_i)}(a_i, \omega_{\text{PV}})$ 
16:    for Each  $a_i \in \mathcal{A}_{\text{active}}$  do
17:      for Each  $h \in \mathbb{H}$  do
18:         $v_{i,h} \leftarrow f_{\text{verify}}(a_i, h)$ 
19:         $\mathcal{V} \leftarrow \mathcal{V} \cup \{v_{i,h}\}$ 
20:      end for
21:    end for
22:    if VerificationGap( $\mathbb{H}, \mathcal{V}$ )  $> \theta_{\text{gap}}$  then
23:       $m \leftarrow \text{Explore}$ 
24:    else if VerifyCompleteness( $\mathcal{V}$ )  $> \theta_{\text{ver}}$  then
25:      break
26:    end if
27:  end if
28: end for
29:  $s_{\text{final}} \leftarrow \text{Synthesize}(\mathbb{H}, \mathcal{V})$ 
30: return  $s_{\text{final}}$ 

```

Algorithm 4 DyCo Five-Stage Main Process

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

```

1: /* Stage 0: Initial solution and gated review */
2:  $a_0 \sim \text{Uniform}(\mathcal{A})$ ;  $s_0 \leftarrow f_{\text{gen}}(a_0, \mathcal{T})$ 
3:  $\mathcal{E} \leftarrow \text{Sample}(\mathcal{A}, n_{\text{eval}})$ 
4: if  $\neg \text{Gate}(s_0, \mathcal{E})$  then
5:   return  $s_0$ 
6: end if
7: /* Stage 1: Global task analysis */
8: for  $a_i \in \mathcal{A}$  do
9:    $u_i \leftarrow f_{\text{analyze}}(a_i, \mathcal{T}, \mathbf{p}_i)$ 
10: end for
11:  $\mathcal{U} \leftarrow \{u_1, \dots, u_n\}$ 
12: /* Stage 2: Dynamic teaming */
13:  $\{\mathcal{G}_1, \dots, \mathcal{G}_k\} \leftarrow \text{PersonalityBasedTeaming}(\mathcal{A}, \mathcal{U}, \{\mathbf{p}_i\}, k)$ 
14: /* Stage 3: Intra-team discussion + EVA */
15:  $\mathcal{R} \leftarrow \emptyset$ 
16: for Each  $\mathcal{G}_j$  do
17:    $s_j \leftarrow \text{EVA}(\mathcal{T}, \mathcal{G}_j, T_{\max})$ 
18:    $a_{\text{rep}}^j \leftarrow \text{ElectRepresentative}(\mathcal{G}_j)$ 
19:    $\mathcal{R} \leftarrow \mathcal{R} \cup \{a_{\text{rep}}^j\}$ 
20: end for
21: /* Stage 4: Representative negotiation */
22:  $s_{\text{final}} \leftarrow \text{EVA}(\mathcal{T}, \mathcal{R}, T_{\max})$ 
23:  $a_{\text{final}} \leftarrow \text{ElectRepresentative}(\mathcal{R})$ 
24:  $r_{\text{final}} \leftarrow f_{\text{collate}}(a_{\text{final}}, \mathcal{T}, s_{\text{final}})$ 
25: return  $r_{\text{final}}$ 

```

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

```

1:  $\mathcal{Q} \leftarrow \emptyset$ ;  $m_{1:T} \leftarrow \emptyset$ ;  $H(C) \leftarrow 0$ 
2:  $a_{\text{current}} \leftarrow \arg \max_{a_i \in \mathcal{G}} p_i^{\text{El}}$ 
3:  $t \leftarrow 0$ 
4: while  $t < T_{\max} \wedge \neg \text{Converged}(\mathcal{G}, t)$  do
5:    $t \leftarrow t + 1$ 
6:    $s_t \leftarrow f_{\text{gen}}(a_{\text{current}}, \mathcal{T}, m_{1:t-1}, \mathbf{p}_{\text{current}})$ 
7:    $m_t \leftarrow \langle a_{\text{current}}, s_t, t \rangle$ 
8:    $m_{1:T} \leftarrow m_{1:T} \cup \{m_t\}$ 
9:    $H(C) \leftarrow \text{ComputeContextEntropy}(m_{1:t})$ 
10:  for  $a_i \in \mathcal{G} \setminus \{a_{\text{current}}\}$  do
11:    if  $p_i^{\text{JP}} > 0$  then
12:       $\sigma_i \leftarrow e^{\eta H(C)} f_{\text{base}}(a_i, m_t)$ 
13:    else
14:       $\sigma_i \leftarrow e^{-\eta H(C)} f_{\text{base}}(a_i, m_t)$ 
15:    end if
16:    if  $\sigma_i \geq \tau_{\text{speak}}$  then
17:       $\mathcal{Q}.\text{Push}(a_i, \sigma_i)$ 
18:    end if
19:  end for
20:   $\sigma_{\text{current}} \leftarrow f_{\text{attitude}}(a_{\text{current}}, m_{1:t}, H(C), \mathbf{p}_{\text{current}})$ 
21:  if  $\sigma_{\text{current}} \geq \tau_{\text{continue}}$  then
22:     $\mathcal{Q}.\text{Push}(a_{\text{current}}, \sigma_{\text{current}})$ 
23:  end if
24:  if  $\mathcal{Q} \neq \emptyset$  then
25:     $a_{\text{current}} \leftarrow \mathcal{Q}.\text{Pop}()$ 
26:  else
27:    break
28:  end if
29: end while
30:  $s \leftarrow \text{SummarizeDiscussion}(m_{1:T})$ 
31: return  $m_{1:T}, s$ 

```

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

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:

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

A team spanning all four cognitive types achieves maximum diversity; a homogeneous team has $\text{CDI} \rightarrow 0$, aligning with organizational-psychology findings (Hong and Page, 2004; van Knippenberg et al., 2004; Forsyth, 2014).

C.5.2 Speaking Willingness Dynamics

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

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

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

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

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

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

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

$$P_{E \rightarrow V}(t) = 1 - \prod_{r=1}^t (1 - p_h^r) \approx 1 - \exp\left(-\sum_{r=1}^t p_h^r\right) \quad (7)$$

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

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

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

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

Since $P_{V \rightarrow E} + P_{V \rightarrow C} = 1$, 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.

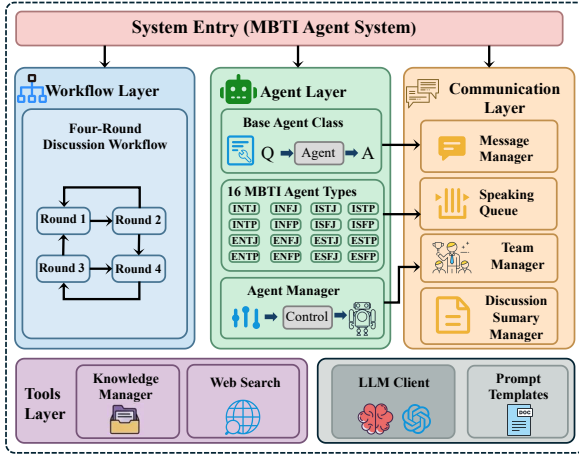


Figure 7: Overall system architecture of DyCo with six layers: Entry, Workflow, Agent, Communication, Tool and Knowledge, and Infrastructure.

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.

E Experimental Setup Details

E.1 Dataset Sampling Protocol

DeepResearch Bench (Du et al., 2026) 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 (CrewAI Inc, 2024) 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. (2024)’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 the short abstract-role ablation, we replace the personality prompts with shorter abstract role descriptors

while keeping the architecture unchanged. For the matched role-prior controls (Appendix F.7), 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. We release the DyCo codebase, prompt templates, experimental configurations, and evaluation scripts together with reproduction instructions in the public repository at <https://github.com/fish-wc/DyCo>. 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 5 shows that all methods use comparable retrieval budgets, supporting a reasonably controlled comparison.

Table 5: Retrieval statistics across methods (mean $\pm$ std, $n=10$).

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

F Additional Results

F.1 MBTI Consistency Confusion Matrix

Table 6: 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

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

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

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.

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.

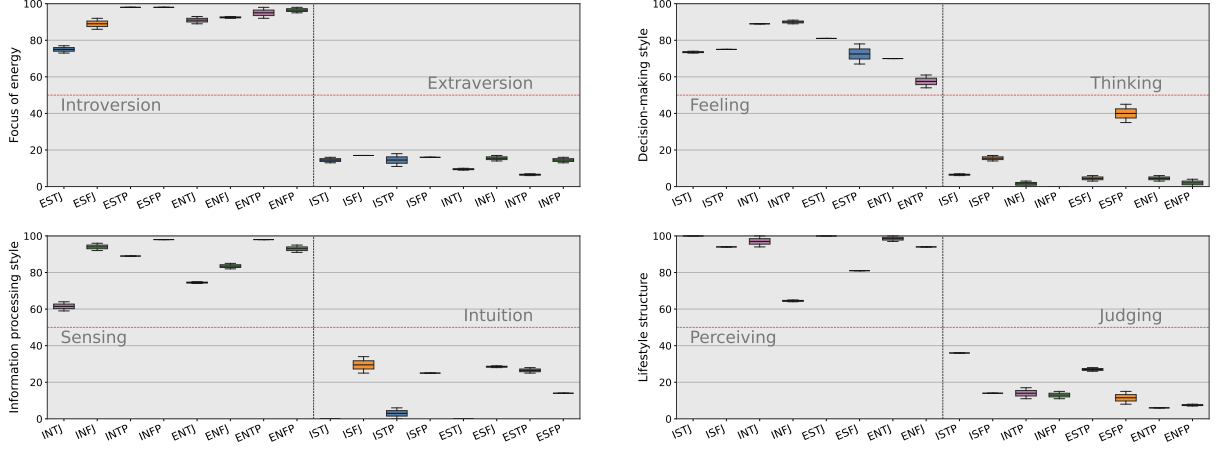


Figure 8: MBTI trait-prompting fidelity validation. Each panel shows the score distribution of all 16 agent configurations along one MBTI dimension (E/I, T/F, S/N, J/P). Green triangles denote group means. The dashed red line marks the 50-point decision boundary.

Table 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

E.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 ($\text{std} \approx 0.17$), 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. Because this pattern is derived from a

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 $\pm$ Std
ENTJ-INTJ-ESTP-ISFJ	42.91	42.58	42.73	42.45	42.71	42.68 $\pm$ 0.17
INTJ-ENFJ-ISTJ-ESFP	42.74	42.49	42.81	42.38	42.65	42.61 $\pm$ 0.16
ENTP-INFJ-ISTJ-ESFJ	42.68	42.41	42.72	42.29	42.58	42.54 $\pm$ 0.17
INTJ-INFP-ESTJ-ISFP	42.55	42.28	42.61	42.19	42.49	42.42 $\pm$ 0.17
ENTJ-ENFJ-ISTP-ISFJ	42.49	42.22	42.54	42.13	42.43	42.36 $\pm$ 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 $\pm$ 0.3	42.38 $\pm$ 0.23
Bottom-5	4/4 (100%)	2.8 $\pm$ 0.4	39.68 $\pm$ 0.29

Table 11: Hyperparameter sensitivity analysis (one-at-a-time, $\pm 30\%$ 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: $\eta=1.0$, $\tau_{\text{speak}}=0.5$, $\tau_{\text{continue}}=0.6$, $\theta_{\text{hyp}}=0.7$, $\theta_{\text{gap}}=0.3$, $\theta_{\text{ver}}=0.8$.

limited development-set search and the E/I axis is used here only as an engineering role prior, we treat it as an exploratory design cue rather than evidence that mixed introvert–extravert teams are generally superior.

F.4 Hyperparameter Sensitivity

Table 11 reports a $\pm 30\%$ 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 Domain Robustness Check

To test whether the main trend is overly tied to a single evaluation slice, we ran a small 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 average token overheads within $\pm 6\%$ of the main setup. The smaller margins suggest the coordination effect is directionally preserved under moderate domain shift, but its magnitude re-

mains setting-dependent. These subsets still share the same underlying task structure of literature retrieval, cross-source synthesis, and report writing; we therefore do not interpret them as evidence of transfer to scientific tasks in general, and our claims remain limited to long-form research synthesis.

F.6 Pareto Frontier Full Data

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

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

Method	Tokens (K)	RACE	Pareto
SingleAgent	76.5	37.80	$\bullet$
CoTAgent	126.9	35.27	
ToTAgent	18.0	40.62	$\bullet$
SCAgent	264.5	38.25	
ReActAgent	204.1	38.22	
DebateAgent	261.7	38.66	
AutoGen	198.4	40.43	$\bullet$
CrewAI	215.3	40.88	$\bullet$
MetaGPT	184.6	39.72	
DyCo-Efficient (no DT)	18.5	42.39	$\bullet$
DyCo (short abstract roles)	56.9	39.77	$\bullet$
DyCo (no EVA)	325.8	41.59	
DyCo (no WSQ)	283.2	40.83	
DyCo-Diverse	234.6	42.75	$\bullet$

F.7 Matched Role-Prior Controls

To decompose the gain attributed to MBTI into prompt richness, behavioral specificity, functional differentiation, and taxonomy semantics, we ran matched controls on *all* 50 held-out evaluation tasks with three shared random seeds. All conditions use the same backbone model, temperature, task input, number of agents, maximum number of rounds, retrieval limit, and complete DyCo coordination procedure, and are matched in prompt length (± 10 tokens per agent) and behavioral specificity:

- **MBTI-instantiated DyCo:** the standard

DyCo configuration with MBTI-based role priors;

- **Matched functional roles:** four non-MBTI functional role priors (Hypothesis Generator, Analytical Critic, Evidence Verifier, and Synthesis Coordinator);
- **Rich non-taxonomic roles:** equally long behavior descriptions that preserve stylistic richness but remove the explicit taxonomy structure;
- **Role-shuffle:** the original MBTI role prompts with the role-to-stage assignment randomly permuted;
- **Short abstract-role ablation:** shorter abstract role descriptors (this condition simultaneously changes the role taxonomy, prompt length, and specificity of behavioral descriptions, and therefore cannot be interpreted as removing only MBTI).

We first average results across the three shared seeds within each task and then conduct two-sided paired t -tests using the 50 task-level RACE Overall scores as paired observations. The four comparisons between the MBTI condition and the control conditions form a single multiple-comparison family with Holm correction. Table 13 reports the results.

Table 13: Matched role-prior controls on all 50 held-out evaluation tasks (three shared random seeds). Δ : difference in favor of the MBTI condition; CI: unadjusted two-sided 95% paired confidence interval of Δ ; p_{adj} : Holm-adjusted p -value within the comparison family. Effect sizes (d_z) and per-task token budgets are reported in §5.2.3.

Condition	RACE	Δ	95% CI	p_{adj}
MBTI-instantiated DyCo	42.66	—	—	—
Matched functional roles	42.18	0.48	[0.08, 0.88]	0.040
Rich non-taxonomic roles	42.07	0.59	[0.08, 1.10]	0.028
Role-shuffle	41.83	0.83	[0.29, 1.37]	0.012
Short abstract-role ablation	39.81	2.85	[2.11, 3.59]	< 0.001

Two patterns emerge. First, the matched functional-role condition recovers 2.37 points, approximately 83% of the 2.85-point gap between the short abstract-role ablation and the MBTI-instantiated DyCo: prompt richness, behavioral specificity, and functional differentiation explain most of the original ablation gap. Second, the remaining MBTI-specific increment is small (0.48

points, $d_z=0.33$) and, in the budget-matched comparison against DyCo-Diverse in the main 10-run setting, DyCo-Diverse remains 0.57 points ahead of the functional-role control (95% CI [0.19, 0.95], adjusted $p=0.018$, $d_z=0.42$) despite a nearly identical compute budget (18.3 vs. 18.1 retrieval calls; 234,568 vs. 231,840 tokens), ruling out additional computation as the sole cause of the difference. We therefore interpret these controls as evidence that rich, structured, behaviorally distinct role priors drive most of the gain, while the MBTI instantiation contributes a modest additional increment; they do not support a strong MBTI-specific causal claim.

F.8 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 $\pm$ std, $n=3$).

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

F.9 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.

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 19.0 $\times$ better RACE-

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

per-dollar than GPT-4o and 8.0CE 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 (§5.2.8) allows practitioners to tune this trade-off by selectively enabling components.

F.10 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.11 Factuality and Source-Faithfulness Audit

Protocol. Because RACE does not directly measure factual correctness or citation faithfulness, we conducted a method-identity-blinded pilot audit. We sampled 10 shared evaluation tasks and, for each of eight methods, extracted atomic factual claims from the generated reports (800 claims in total). Two annotators independently assessed (i) whether each claim is factually accurate and (ii) whether the claim is faithfully supported by its cited sources. A third annotator resolved disagreements. Pre-adjudication agreement was Cohen’s $\kappa = 0.72$. All reports were stripped of method-identifying metadata, and annotators were blind to which method produced each report.

Results. DyCo-Diverse achieved a factual accuracy of 81% and a source-faithfulness score of 79%, compared with 79% and 76% for CrewAI, the strongest multi-agent baseline. These results do not indicate that higher RACE scores are accompanied by substantially lower factual quality. We emphasize the pilot nature of this audit: the task and claim sample is small, and a full-scale blinded audit with hierarchical statistical analysis (treating tasks as clusters) remains future work.

F.12 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.13 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 short abstract-role ablation is the abstract-descriptor baseline that replaces personality prompts with shorter role descriptions. Bold and underlined values indicate the best and second-best results in each column, respectively.

F.14 Efficiency Comparison of Methods

Table 18 details the efficiency metrics of all multi-agent frameworks under comparison. DyCo variants, including DyCo-Diverse, DyCo-Efficient (no dynamic teaming), and the short abstract-role 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

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

Configuration	RACE Metrics					Efficiency Statistics			
	Comp.	Insight	Instr. Follow.	Readab.	Overall	Time (s)	Prompt Tok.	Complt. Tok.	Total Tok.
<i>NF (Intuition-Feeling)</i>									
4 $\times$ INFJ	39.54 $\pm$ 0.36	41.34$\pm$0.33	46.59$\pm$0.28	44.36 $\pm$ 0.27	42.53$\pm$0.30	208.56	83,477	5,871	89,349
4 $\times$ INFP	36.05 $\pm$ 0.51	39.33 $\pm$ 0.48	44.08 $\pm$ 0.42	43.05 $\pm$ 0.38	40.12 $\pm$ 0.44	209.84	75,870	4,937	80,807
4 $\times$ ENFJ	38.38 $\pm$ 0.45	38.99 $\pm$ 0.43	43.99 $\pm$ 0.40	42.52 $\pm$ 0.36	40.44 $\pm$ 0.41	245.33	91,122	4,890	96,011
4 $\times$ ENFP	29.12 $\pm$ 0.62	29.82 $\pm$ 0.68	39.41 $\pm$ 0.54	38.80 $\pm$ 0.49	33.30 $\pm$ 0.55	163.02	63,607	3,314	66,921
<i>NT (Intuition-Thinking)</i>									
4 $\times$ INTJ	39.25 $\pm$ 0.38	40.86 $\pm$ 0.36	45.06 $\pm$ 0.33	44.28 $\pm$ 0.29	41.87 $\pm$ 0.34	294.60	99,279	5,743	105,022
4 $\times$ INTP	37.78 $\pm$ 0.43	40.55 $\pm$ 0.38	45.08 $\pm$ 0.34	43.09 $\pm$ 0.31	41.06 $\pm$ 0.36	250.17	68,259	5,267	73,526
4 $\times$ ENTJ	38.42 $\pm$ 0.41	38.72 $\pm$ 0.41	45.07 $\pm$ 0.36	43.23 $\pm$ 0.33	40.73 $\pm$ 0.38	200.83	60,325	4,913	65,238
4 $\times$ ENTP	34.29 $\pm$ 0.48	37.44 $\pm$ 0.45	41.98 $\pm$ 0.41	41.38 $\pm$ 0.38	38.20 $\pm$ 0.44	185.04	59,266	4,048	63,314
<i>SF (Sensing-Feeling)</i>									
4 $\times$ ISFJ	39.64 $\pm$ 0.35	40.98 $\pm$ 0.34	45.25 $\pm$ 0.31	44.07 $\pm$ 0.28	42.06 $\pm$ 0.32	257.24	96,105	6,295	102,400
4 $\times$ ISFP	37.52 $\pm$ 0.40	39.28 $\pm$ 0.39	44.91 $\pm$ 0.35	43.22 $\pm$ 0.32	40.72 $\pm$ 0.37	291.15	129,209	5,961	135,170
4 $\times$ ESFJ	36.41 $\pm$ 0.46	36.77 $\pm$ 0.44	44.57 $\pm$ 0.37	43.05 $\pm$ 0.34	39.45 $\pm$ 0.40	206.99	84,482	4,971	89,453
4 $\times$ ESFP	35.05 $\pm$ 0.49	34.68 $\pm$ 0.48	44.00 $\pm$ 0.39	42.00 $\pm$ 0.37	38.15 $\pm$ 0.43	231.38	66,089	4,195	70,284
<i>ST (Sensing-Thinking)</i>									
4 $\times$ ISTJ	40.19$\pm$0.34	39.40 $\pm$ 0.37	46.44 $\pm$ 0.29	44.09 $\pm$ 0.28	41.84 $\pm$ 0.33	299.19	90,778	5,054	95,831
4 $\times$ ISTP	39.46 $\pm$ 0.36	40.73 $\pm$ 0.35	46.13 $\pm$ 0.30	44.67$\pm$0.26	<u>42.17$\pm$0.31</u>	308.50	85,409	7,489	92,898
4 $\times$ ESTJ	38.19 $\pm$ 0.41	38.90 $\pm$ 0.39	45.66 $\pm$ 0.33	44.18 $\pm$ 0.30	41.09 $\pm$ 0.36	281.53	92,862	5,570	98,432
4 $\times$ ESTP	39.22 $\pm$ 0.37	40.52 $\pm$ 0.36	45.50 $\pm$ 0.31	43.87 $\pm$ 0.28	41.81 $\pm$ 0.33	234.79	77,241	5,606	82,847
<i>Ablation Baseline</i>									
Short abstract-role ablation	36.68 $\pm$ 0.46	36.83 $\pm$ 0.48	45.23 $\pm$ 0.40	43.11 $\pm$ 0.35	39.77 $\pm$ 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	<u>13,149</u>	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 (short abstract roles)	206.18	53,176	3,758	56,935
DyCo-Efficient (no DT)	230.3	–	–	18,471
4 $\times$ ENFP	163.02	63,606	3,313	66,921
4 $\times$ INFJ	208.56	83,477	5,871	89,349
DyCo-Diverse	541.25	227,127	7,441	234,569

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.