Transactions on Computing Science
WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.519 Volume 2 , Issue 2

Research Article

Autonomous Agents Based on Large Language Models: Reasoning Strategies and Multi-Agent Interaction

Zhengyuan Wang[1,*], Yuyao Wang[2], Yixin Deng[3]
[1]Johns Hopkins University, Baltimore 98005, United States
[2]New York University, New York 10044, United States
[3]College of Engineering, University of California, Berkeley, Berkeley 94720, United States
[*]**Correspondence:** Zhengyuan Wang, harleycodywzy94@gmail.com

© The Author(s) 2026.
This article is published by Wisdom Academic Press Ltd. under a Creative Commons Attribution 4.0 International License. The license permits use, sharing, adaptation, distribution, and reproduction in any medium or format, provided that appropriate credit is given to the original author(s) and the source, a link to the license is supplied, and any modifications are indicated. Unless otherwise stated, all images and third-party materials included in this article are also covered by the same license. For any material not covered by this license, if the intended use is neither permitted by applicable law nor allowed under the license terms, direct permission must be obtained from the copyright holder. To view a copy of this license, please visit http://creativecommons.org/licenses/by/4.0/.

Abstract

Language-model agents acquire capability along two routes that published work almost always varies together, one enlarging the computation a single agent performs and the other enlarging the number of agents and the channels along which they pass partial results, which leaves the contribution of each route indistinguishable. This paper separates an agent system into a reasoning operator and an interaction topology, holding the number of agents fixed at three across all multi-agent topologies, together with the number of rounds, the message cap, the prompt skeleton, and the memory settings for all combinations, and evaluates a full factorial design of four reasoning levels and four topologies on three public test sets totalling 484 instances under five random seeds, giving 38,720 task executions. Comparisons are paired at the instance level, and effects are estimated from logistic mixed-effects models fitted to each test set separately. The reasoning factor spans 18.0 points on the static suite whereas the topology factor spans 7.8 points on the interactive one, a reversal indicating that the two mechanisms act on different failure modes. The topology span is taken across all four topology levels and therefore includes the step from one agent to three, against a structure-only span of 2.8 points among the multi-agent topologies. The two factors also depart from additivity in a localised manner. A mechanism-motivated single-degree-of-freedom contrast on the only cell in which both factors contribute a corrective pass is significant at $p = 0.001$ on the static suite, whereas the omnibus nine-degree-of-freedom interaction test is not ($p = 0.094$), and in that cell the observed 50.0 percent lies 5.8 points below the additive expectation. After tokens are counted, seven of the sixteen configurations remain undominated on the static suite and five on the operating-system suite, and returns saturate at 12,400 tokens per instance on the static suite under the present design.

Transactions on Computing Science
WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.519 Volume 2 , Issue 2

Keywords

Large language model, Autonomous agent, Reasoning strategy, Multi-agent interaction, Cost-aware evaluation

1. Introduction

Language models have moved from producing text to acting as agents that plan, call tools (Schick et al., 2023), maintain state over long horizons (Park et al., 2023; G. Wang et al., 2024), and revise their own outputs, a shift that recent surveys treat as the defining development in the field (L. Wang et al., 2024; Xi et al., 2025). Two developments account for most of the capability gained along the way. The first enlarges the computation an individual agent performs before it commits to a response, a practice that grew out of the observation that prompting for intermediate steps improves accuracy on multi-step problems (Wei et al., 2022). The second enlarges the number of agents and the pathways along which they communicate partial solutions. Both are widely used, both are known to help, and both are routinely varied at the same time, so how they combine has remained an open question.

Work on the first mechanism has enriched the intermediate state in several ways. Interleaving reasoning with tool calls lets an agent condition later steps on information gathered en route (Yao, Zhao, et al., 2023). Sampling several trajectories from one request and returning the most frequent answer removes errors confined to a single trajectory (X. Wang et al., 2023). Allowing an agent to inspect and correct its own answer adds a further pass over what it has already produced (Shinn et al., 2023). Search over trees or graphs of intermediate thoughts extends the single trajectory to deliberate exploration of alternatives (Besta et al., 2024; Yao, Yu, et al., 2023). Work on the second mechanism has instead varied the structure of interaction. Symmetric debate places agents opposite one another and settles disagreement by consensus (Du et al., 2024). Role assignment arranges agents in a pipeline in which each performs one task and hands its output to the next (Hong et al., 2024). Treating agents as nodes of a computational graph turns connectivity itself into a design parameter (Zhuge et al., 2024). General conversational frameworks and role-played agent societies supply reusable infrastructure for such interaction (Li et al., 2023; Qian et al., 2024; Wu et al., 2024).

Three difficulties run through both literatures. Reported gains are usually obtained by changing the reasoning procedure, the number of agents, and the communication scheme at once, which leaves no way to attribute the gain to any one of them. Accuracy is rarely reported against the tokens spent to reach it, so a genuine improvement in reasoning cannot be told apart from a simple increase in inference compute, a point made recently by a systematic comparison in which several debate frameworks failed to beat single-agent baselines at matched compute (H. Zhang et al., 2025). Results also tend to come from one sampling run, so the seed-to-seed variation against which a reported

difference would have to be judged is simply not available.

This study makes four contributions. It formalises an agent system as a reasoning operator applied inside each agent and an interaction topology along which messages pass between agents, with every remaining element held fixed across all combinations. It runs a full factorial experiment over four reasoning levels and four interaction topologies on three public suites under five random seeds, pairing comparisons at the instance level and estimating the joint behaviour of the two factors by mixed-effects analysis. It reports the accuracy of each configuration against the tokens that configuration consumes and identifies the undominated subset. It attributes failures to categories and shows that message passing reduces three kinds of error while introducing two.

2. Methodology

2.1. Overall Architecture of the RSMI Framework

The behaviour of an agent system rests on two design decisions that earlier work takes together, one fixing how much intermediate computation a single agent performs and the other fixing how many agents take part and along which paths they exchange information. In the Reasoning Strategy and Multi-agent Interaction (RSMI) framework, the two are isolated and controlled independently, and the cost of every combination is measured directly.

The framework has three layers. The dispatch layer instantiates the reasoning operator and the topology named in the configuration descriptor, and it holds the only reference to the model endpoint, so no agent calls the backbone directly. The agent layer holds agents that share a backbone but differ in the system prompt defining their role, and they communicate through messages carrying an originator, a recipient, and a payload. The accounting layer runs the sandbox for interactive actions and records prompt tokens, response tokens, elapsed time, and the number of rounds. The system is written as the ordered five-tuple of Equation (1).

$$S=\langle M, A, R, T, \Phi \rangle \quad (1)$$

where M is the backbone model, A the agent set, R the reasoning operator, T the topology governing message flow, and Φ the aggregation rule, which the topology determines and which is listed separately to make the point of adjudication explicit. The five components appear in the order backbone model, agent set, reasoning operator, topology, and aggregation rule, and this order is preserved wherever the tuple is restated.

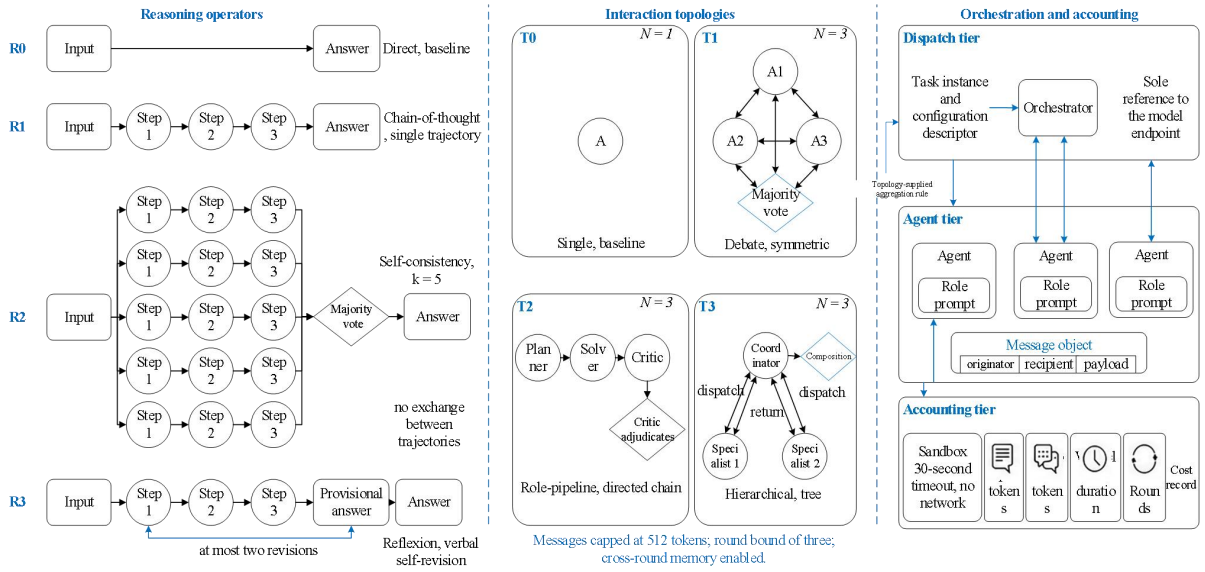
The two factors act on different regions of the space of intermediate states. One sets how far the operator pushes along a single trajectory, the other sets how many hypotheses are entertained in parallel and along which paths they are exchanged. Where the regions opened by the two overlap, the benefit of combining them falls short of the sum of their separate benefits; where they do not overlap, the combined benefit can exceed that sum. Figure 1 shows the three layers together with the flow of

Transactions on Computing Science
WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.519 Volume 2 , Issue 2

information among them.

Figure 1

Architecture of the RSMI framework



2.2. Reasoning Strategy Operators

The reasoning factor has four levels, all exposed through one interface whose methods propose an answer, absorb an observation, and produce a revision; a level that does not use a method receives a null implementation, so orchestration is identical at every level. Level R0 answers directly with no intermediate state and serves as the control condition. Level R1 writes an explicit sequence of steps and extends a single trajectory. Level R2 draws k = 5 trajectories and returns the most frequent answer, discarding errors confined to any one of them. Level R3 revises its previous answer in words, with a budget of two revisions per round.

R2 resembles five agents that vote but cannot talk to one another, and it shares with debate the aim of removing errors that a single trajectory makes, drawing on diversity within a population of candidates. The two differ in where that diversity comes from. R2 generates it inside one agent with no transfer of partial knowledge, whereas debate generates it across agents, each of which conditions on the positions the others hold.

2.3. Interaction Topologies

The interaction factor has four levels, distinguished by the communication structure among nodes and by the rule that aggregates their answers. Level T0 is a single agent with no edges and serves as the baseline. Level T1 places three agents on a fully connected symmetric graph and adjudicates by majority vote. At level T2, three agents form a directed chain of planner, solver, and critic, each node taking input only from its predecessor. At level T3, a coordinator sits above two specialists in a tree,

decomposing the task and composing the returned fragments.

Five constraints enforce orthogonality. The number of agents is fixed at N = 3 for all non-baseline topology configurations, which ensures that comparisons among T1, T2, and T3 isolate the communication structure at a constant number of model calls. T0 anchors both factors as the single-agent baseline; a comparison between T0 and any multi-agent topology therefore bundles the effect of adding agents with the effect of the structure connecting them, and the two contributions are reported separately below and probed independently by the agent-count ablation. The round bound is three throughout. Message size is capped at 512 tokens, so a topology with more edges cannot gain from longer text. Every role prompt follows one skeleton comprising a role definition, an input format, an output format, and a list of forbidden constructs. Cross-round memory is identical for all agents, so no topology enjoys a context advantage. These bounds are set for experimental control rather than as claims about scale, so quantities estimated under them, including the 12,400-token level at which returns level off under the present design, are properties of this design and not invariants of the framework. Table 1 records both factors with the mechanism and controlled parameter attaching to each level.

Table 1

Factor levels of the RSMI design

Factor	Level	Mechanism	Parameter
Reasoning	R0	Direct answer, no intermediate state	Baseline
Reasoning	R1	Chain-of-thought, single trajectory	None
Reasoning	R2	Self-consistency, majority over samples	$k = 5$
Reasoning	R3	Reflexion, verbal self-revision	≤ 2 per round
Interaction	T0	Single agent, no edges	Baseline
Interaction	T1	Debate, symmetric graph, majority vote	$N = 3$
Interaction	T2	Role-pipeline, planner to solver to critic	$N = 3$
Interaction	T3	Hierarchical, coordinator over specialists	$N = 3$

Note. Non-baseline topologies share three agents, a round bound of three, a 512-token message cap, one prompt skeleton, and cross-round memory. The operator applies identically at every node, so the factors vary independently by construction.

2.4. Unified Orchestration Procedure

Transactions on Computing Science
WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.519 Volume 2 , Issue 2

One procedure drives all sixteen configurations, removing the possibility that a difference in control flow is mistaken for a difference in factor level. Termination is governed by Equation (2).

$$\mathrm{Stop}(t)=1[\rho_t \geq \tau] \vee 1[t=T\max] \quad (2)$$

where ρt is the proportion of agents holding the modal answer after round t, τ the consensus threshold set to unity, Tmax the round bound set to three, with the indicator and the disjunction carrying their usual sense. Under the single-agent topology the first term is satisfied at the end of round one, so the procedure runs for a single round and the baseline incurs no coordination cost. Algorithm 1 states the procedure, which takes the operator and the topology as parameters rather than branching on them. All sixteen configurations are therefore parameterisations of a single procedure.

Algorithm 1

Unified RSMI orchestration

Line	Statement	Comment
	Input: task instance x, operator R, topology T, bound T_max	
	Output: answer y, token cost c, terminal round t	
1	c ← 0; M ← ∅	accumulator and message store
2	A ← Instantiate(T)	role prompts fixed by T
3	for t = 1 to T_max do	
4	for each a ∈ A do (y_a, c_a) ← R.propose(a, x, M)	invocation under R
5	c ← c + Σ c_a	counters updated
6	M ← Route(T, {y_a})	edges prescribed by T
7	for each a ∈ A do (y_a, c_a) ← R.revise(a, x, M)	identity pass unless R3
8	c ← c + Σ c_a	
9	ρ ← ModalShare({y_a})	agreement proportion
10	if ρ ≥ τ or t = T_max then	criterion, Equation (2)
11	return (Φ({y_a}), c, t)	aggregation applied
12	end if	
13	end for	

Note. Line 6 is the only point at which the topology affects routing and line 4 the only point at which

the reasoning level affects node computation. Line 7 is instantiated only under R3. Under R0, R1, and R2 the call to R.revise evaluates to an identity pass that returns the incoming answer unchanged, issues no invocation to the backbone, and adds zero to the counter in line 8, so no execution cost arises where the operator defines no revision. Under T0 line 6 is null for want of edges, and line 7 acts on the trajectory of the single agent alone, so the baseline carries no coordination cost.

3. Experimental Evaluation and Results

3.1. Experimental Setup

Three research questions (RQs) structure the analysis. RQ1 asks how large the main effect of each factor is and whether it varies by task type. RQ2 asks whether the two factors interact and, if so, in which direction. RQ3 asks which configurations remain undominated once token use is taken into account.

Three test suites are used. BBH-8 consists of eight exact-match tasks drawn from BIG-Bench Hard (Suzgun et al., 2023), itself distilled from BIG-Bench (Srivastava et al., 2023), namely boolean expressions, date understanding, five-object logical deduction, depth-two multi-step arithmetic, navigation, object counting, tracking five shuffled objects, and word sorting, each run on 30 instances drawn with a fixed seed, giving 240 instances in total. AB-OS reproduces the operating-system suite of AgentBench (Liu et al., 2024) in full, comprising 144 instances in which an agent issues bash commands inside an Ubuntu container. AB-WS is a fixed-seed subset of 100 instances from the web shopping test set of the same benchmark, which derives from the WebShop environment (Yao et al., 2022). Exact match scores BBH-8, and success rate scores the two interactive suites. In total, there are 484 instances reused without change across configurations and 38,720 task executions.

All agents share one backbone. The main grid runs on glm-4-flash-250414, using the provider application programming interface (API) at temperature 0.7, top-p 0.9, and 1,024 new tokens per call. A cross-model check repeats the same procedure on a reduced grid of four configurations with Qwen2.5-7B-Instruct under the same interface and parameters. Interactive actions run inside a subprocess sandbox with a 30-second command timeout and no network access. Each configuration is executed five times under distinct seeds; a run that fails for external reasons is retried at most twice, and a third failure would be recorded as a failed run. Forty-seven retries were needed in total and no run ended unresolved. The unit of analysis is the binary outcome of an instance–seed pair; every configuration is evaluated on the identical $484 \times 5 = 2{,}420$ pairs, so all comparisons are fully paired. Pairwise differences between configurations are tested with the exact McNemar test on these paired outcomes; a paired bootstrap with 10,000 replicates, resampling instances as clusters so that the five seed replicates of an instance remain together, supplies interval estimates of the accuracy differences and complements rather than replaces the McNemar test; Holm adjustment at a family level of 0.05

covers the 120 pairwise comparisons within each suite. Effect sizes are estimated from the logistic mixed-effects model of Equation (3), in which the seed replicates enter as repeated observations of each instance and the instance random intercept absorbs between-instance difficulty. Refitting the model with an additional crossed random intercept for seed changes no fixed effect by more than 0.02 on the log-odds scale and estimates the seed variance below 0.01, so the simpler specification is retained. Likelihood-ratio (LR) statistics, odds ratios (ORs), and 95 percent confidence intervals (CIs) are reported throughout.

$$\text{logit}\left(\Pr\left[y_{ijk}=1\right]\right)=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+u_k \quad (3)$$

where αi denotes the reasoning level, βj the topology level, (αβij) their interaction, and uk a random intercept for instance k, which absorbs between-instance difficulty that a configuration-level analysis cannot. Each suite is modelled separately, since pooling would weight the estimates toward the largest suite. Task dependence is assessed by comparing the three fitted models. The additivity of the two factors is assessed by the likelihood-ratio tests specified below, and estimates are not presented where the baseline rate is close to zero. Wilson intervals accompany all rates, with half-widths of 6.3 points on the largest suite and 8.1 points on the smallest at the rates observed in each.

3.2. Main Effects of the Two Factors

Table 2 reports accuracy for all sixteen combinations. Reading across a row holds the topology fixed and isolates reasoning, and reading down a column does the reverse, which is what RQ1 requires.

Table 2

Accuracy across the sixteen RSMI configurations on three suites

Configuration	BBH-8	AB-OS	AB-WS
R0-T0	34.2 (1.3)	9.0 (1.1)	11.0 (2.6)
R0-T1	38.6 (1.6)	16.0 (1.5)	17.0 (3.4)
R0-T2	40.1 (1.4)	17.4 (1.4)	19.0 (3.1)
R0-T3	37.5 (1.5)	14.6 (1.3)	16.0 (3.0)
R1-T0	48.6 (1.2)	14.6 (1.2)	15.0 (2.9)
R1-T1	52.3 (1.7)	20.8 (1.6)	20.0 (3.6)
R1-T2	53.8 (1.4)	22.2 (1.5)	23.0 (3.3)
R1-T3	51.2 (1.6)	19.4 (1.4)	19.0 (3.2)

R2-T0	53.1 (1.5)	16.0 (1.3)	16.0 (3.0)
R2-T1	55.8 (2.1)	21.5 (1.8)	21.0 (3.9)
R2-T2	58.1 (1.6)	23.6 (1.6)	24.0 (3.5)
R2-T3	55.5 (1.8)	20.8 (1.7)	20.0 (3.6)
R3-T0	51.4 (1.4)	17.4 (1.4)	17.0 (3.1)
R3-T1	50.0 (1.9)	20.1 (1.6)	19.0 (3.8)
R3-T2	56.2 (1.5)	25.0 (1.5)	25.0 (3.4)
R3-T3	53.8 (1.7)	22.2 (1.6)	22.0 (3.5)

Note. Mean over five seeds with the standard deviation in parentheses, expressed as exact match on BBH-8 and success rate on the two interactive suites. Wilson 95 percent intervals, token consumption, latency, and terminal rounds for every configuration appear in the supplementary material.

Spans denote the difference, in accuracy percentage points, between the marginal means of the highest- and lowest-performing level of one factor, averaging over the levels of the other. Row margins are widest on BBH-8, where the highest and lowest reasoning levels differ by 18.0 points against 6.9 on AB-OS and 5.0 on AB-WS. Failures on the static suite arise mainly from skipped intermediate steps rather than from early commitment to a wrong trajectory, so deeper reasoning inside a single agent addresses the dominant failure. Column margins run the other way, with topology spanning 7.8 points on AB-OS and 8.0 on AB-WS against 5.2 on BBH-8. Because T0 differs from the other levels in agent count as well as in structure, the column span bundles the two contributions; restricted to the multi-agent topologies T1–T3, the structure-only span is 2.8 points on AB-OS, 3.5 on AB-WS, and 2.9 on BBH-8. The larger part of the topology margin on the interactive suites therefore comes from the transition from one agent to several, while the choice among debate, pipeline, and hierarchy contributes a smaller but directionally consistent effect, with the pipeline attaining the highest point estimate in every suite. In interactive tasks an action is committed before its consequence can be observed, and a trajectory that commits to the wrong action cannot undo it, which is why several trajectories with exchange between them help. The ordering of topologies is stable across seeds on AB-OS, whereas on AB-WS the two best topologies differ by 3.5 points against a seed-to-seed standard deviation of 3.8, so these data do not separate them. The direct operator on the single-agent topology is the weakest configuration on every suite, at 34.2, 9.0, and 11.0 percent.

3.3. Interaction Between the Two Factors

The marginal effects above establish that each factor matters on its own, which raises the question

RQ2 poses, namely whether the two effects are additive. Both main effects hold in all three suites; on BBH-8 the strongest reasoning level and the strongest topology carry ORs of 2.08 [1.74, 2.49] and 1.24 [1.05, 1.46] against their respective baselines. Two nested tests address additivity. The omnibus test compares the additive model against the model carrying the full interaction term, which adds nine parameters for two four-level factors; the LR statistics are 14.9 on BBH-8 and 10.3 on AB-OS on nine degrees of freedom ($p = 0.094$ and $p = 0.327$), so a global interaction is not established. A single-degree-of-freedom contrast then augments the additive model with the indicator $I(R = R3 \wedge T = T1)$, the only cell of the grid in which both factors contribute a corrective pass; the contrast is motivated by this mechanism but was specified after inspection of the grid and is reported as such. Its LR statistics are 11.42 and 7.86 on one degree of freedom ($p = 0.001$ and $p = 0.005$), with coefficients of -0.31 [-0.49, -0.13] and -0.38 [-0.65, -0.11] on the log-odds scale. Because the cell was selected after inspection, a conservative Bonferroni correction over the sixteen cells of the grid is also reported, and it leaves the contrast below the 0.05 level on BBH-8, at an adjusted p of 0.012, while on AB-OS the adjusted p of 0.081 does not reach that level, so the departure is treated as established on the static suite and as directionally consistent on the interactive one. On AB-WS the 11.0 percent baseline rate is too close to the lower identifiability limit at 100 instances for either test.

The departure from additivity sits in one region of the grid. The reflexion operator paired with the debate topology reaches 50.0 percent on BBH-8 against 55.8 percent under the additive model, the marginal predicted probability for the R3-T1 cell given by the inverse logit of $\mu + \alpha_{R3} + \beta_{T1}$ marginalised over the fitted distribution of the instance intercept, a shortfall of 5.8 points and 3.1 times the standard deviation of that cell; the same cell falls 4.3 points below prediction on AB-OS, while every other cell departs from its additive prediction by less than one standard deviation. Two accounts combine at this cell. Each mechanism gives a candidate solution a corrective pass, one from the agent itself and one from the positions held by its peers, and because the two passes act on partially overlapping sets of states the second finds fewer errors to remove, which on its own would make the pairing sub-additive. The larger part of the shortfall comes from the order in which the two passes occur. Debate drives the modal share upward within a round, and once unanimity is reached the termination rule of Equation (2) closes the run at the end of that round. Within a round the revision call of line 7 precedes the consensus check of line 10, so termination cannot pre-empt the revision of the same round; what it forecloses are the subsequent rounds, in which the reflexion operator would have spent its remaining per-round revision budget on the peer positions routed at line 6. An incorrect modal answer reached early is therefore locked in before those later corrective passes can occur. Because the revision call acts after routing, it tends to move agents toward the modal position, so the revision pass itself accelerates consensus. The execution logs support this account, in that 58 percent of R3-T1 failures on BBH-8 terminate at the end of round one with a unanimous incorrect answer,

against 34 percent for R1-T1, where no revision follows the routing, and 22 percent for R3-T2; a round-by-round trace of one such run is provided in the supplementary material. Premature consensus on an incorrect answer is one of the two failure categories that arise only where messages are exchanged, which is why the shortfall concentrates in the pairing of reflexion with debate rather than spreading across the grid. These data confirm the sign and the location of the departure from additivity rather than a global interaction. The magnitude of 5.8 carries an uncertainty of about 2.4 points, obtained by parametric bootstrap of the fitted additive model, so no claim is made about the functional form of the discount.

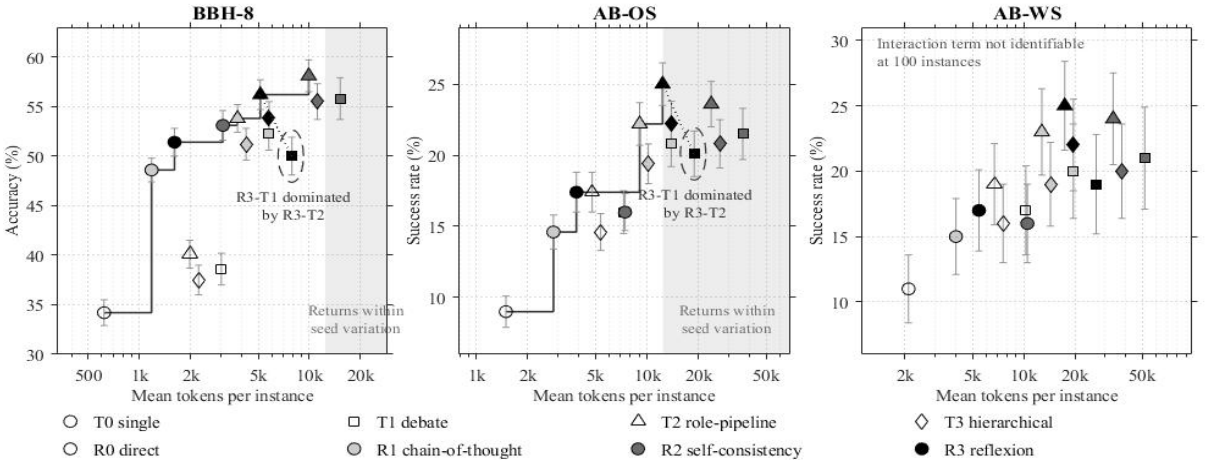
3.4. Accuracy-Cost Trade-off and Ablations

Every gain in the table above is bought with tokens, and the price decides which configurations are worth building, which is what RQ3 asks. Figure 2 plots accuracy against mean total token consumption per instance on a logarithmic horizontal axis. The frontier is sparse, with seven of the sixteen configurations undominated on BBH-8 and five on AB-OS, and although the reflexion operator paired with the debate topology ranks among the more accurate cells, it is dominated on both suites. On BBH-8 it consumes 7,890 tokens per instance against the 620 of the direct single-agent baseline, that is 12.7 times the reference cost, for an accuracy of 50.0 percent that the role-pipeline pairing of the same operator surpasses at 5,150 tokens and 56.2 percent. The role pipeline dominates debate because a directed chain reads each output once, whereas debate has every agent read every position in every round, so its communication volume grows with the square of the agent count. Averaged over the four reasoning levels, the pipeline exceeds debate by 2.9 points on BBH-8 and 2.5 on AB-OS while consuming fewer tokens at every level; under the reflexion operator specifically the gap widens to 6.2 points on BBH-8 and 4.9 on AB-OS, and the pipeline reaches this higher accuracy at two-thirds of the token cost, a factor of 1.5, so the chain dominates on accuracy and cost alike and the cost asymmetry compounds rather than offsets the accuracy difference. Returns level off after 12,400 tokens per instance under the present design of three agents, a three-round bound, and 512-token messages, a threshold equal to twenty times the 620-token direct baseline and seven times the 1,760 tokens per instance of the chain-of-thought single agent R1-T0 on the same suite, with most undominated configurations lying below it, as further consumption makes differences within the seed-to-seed standard deviation.

Figure 2

Accuracy against token consumption across the sixteen configurations, by suite

Transactions on Computing Science
WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.519 Volume 2 , Issue 2



Relaxing the constraints that keep the two factors orthogonal shows how far the topology effect depends on them. The ablations run on the two suites where the interaction term is estimable, under the same seeds and computational budget as the main grid. Raising the agent count from N = 2 to N = 3 gains 3.1 points on BBH-8 and 4.2 on AB-OS. Raising it further to N = 5 adds 0.6 and 0.9 points while consumption grows by 68 percent, and both increments lie inside the seed-to-seed standard deviations of 1.4 and 1.6, so neither supports a claim of improvement. With terminal rounds averaging 1.62, the third round adds 14 solved instances across the 38,720 executions of the grid, or 0.04 points on average, which lies below the smallest seed-to-seed standard deviation of 1.1 points and is therefore not statistically distinguishable from zero. Removing role diversity costs 4.7 points, and removing cross-round memory costs 1.1, which lies within one standard deviation on BBH-8 and does not establish that the memory component is necessary there.

3.5. Failure-Mode Attribution and Cross-Model Check

Each failure is assigned to exactly one category at the earliest point where the trajectory leaves a correct solution path. The operational definitions, the fixed order in which categories are tested, the annotation procedure, and the agreement statistics are given in the supplementary material; two authors annotated a stratified 20 percent sample independently and blinded to the configuration label, reaching Cohen's $\kappa = 0.81$, disagreements were resolved by discussion, and the remaining trajectories were annotated by one author under the agreed rules. Three categories occur under single-agent and multi-agent execution alike, namely per-step computation errors, errors in the decomposition or sequencing of subgoals, and false factual statements. Taken together these fall by 12.6 points on BBH-8 when the chain-of-thought single agent gives way to the three-agent debate topology, that is from R1-T0 to R1-T1, the pair on which the attribution was carried out. Two further categories arise only where messages are exchanged, namely distortion of a peer position on restatement and premature consensus on an incorrect answer. These account for 8.9 points between them, leaving a net improvement of 3.7 points, the difference Table 2 records for that pair. A trajectory whose earliest

deviation matches no category is left unresolved, which happened in 6.2 percent of failed trajectories, so the figures above are lower bounds. The connection with the departure from additivity reported earlier is direct, since premature consensus forecloses the subsequent rounds in which the reflexion operator would apply its remaining corrective passes. The reduced grid on the second backbone crosses both baselines with the strongest levels, {R0, R3} × {T0, T1}, and its complete results appear in Supplementary Table A1. Both main effects and the sub-additive shortfall reproduce in direction on BBH-8 and AB-OS. On BBH-8, Qwen2.5-7B-Instruct yields 41.3, 45.9, and 55.4 percent for R0-T0, R0-T1, and R3-T0, while R3-T1 reaches 55.0 percent against an additive expectation of 60.0, a shortfall of 5.0 points of the same sign and similar size as on the main backbone. The single non-reproducing configuration is R0-T1 on AB-WS, where the debate topology fails to improve on the single agent, 8.0 against 9.0 percent; the baseline rate sits at the identifiability floor already noted for this suite, so the deviation is attributed to a floor effect of the weaker backbone rather than to the framework.

4. Discussion

Once the system is decomposed, the estimates fall into a pattern that can be read off directly. Both mechanisms move outcomes, but they move different failures, so the choice between them turns on which failure dominates a given task. Depth within one agent removes the failure of skipping an intermediate step, while breadth across agents removes the failure of committing early to an action that cannot be undone, and the suites differ in how often each occurs. The localised departure from additivity sharpens the same reading, since the two mechanisms overlap precisely where a candidate answer is offered for a second corrective pass, and applying both there amounts to correcting the same thing twice. Cost behaves in a parallel way. The topology that survives most often in these comparisons is the one whose message volume grows linearly rather than quadratically in the number of agents. This is consistent with evidence that pruning redundant messages preserves accuracy at a fraction of the token expenditure (G. Zhang et al., 2025), although that work optimises an existing pipeline whereas the topologies compared here remain unoptimised baselines.

Six limitations bound these statements. The suites contain 484 instances in total, which limits the precision of absolute rates and leaves the conclusions resting on comparisons between configurations. The suites are English and range from single-turn to moderate horizons, and none involves code generation, so behaviour on repository-scale work is not constrained by this evidence. In long-horizon settings the localisation found here need not persist, since per-step errors compound over many more decision points, memory and context demands grow with the trajectory, and the balance between the two mechanisms may shift, with breadth across agents gaining value where early irreversible commitments become more frequent and losing it where accumulated context favours a single deep

trajectory. The number of agents and the round bound were constrained to three by the available computational resources, so the shape of interaction at higher numbers remains unknown. The ablation at five agents places the accuracy increment inside the seed-to-seed standard deviation while consumption grows by 68 percent, which indicates that the qualitative pattern of saturation persists, but the point at which returns level off is estimated under the three-agent bound and would shift upward in absolute terms on larger networks or deeper orchestration graphs, so the figure of 12,400 tokens per instance should be read as specific to this design. The main grid uses one backbone and the cross-model check covers only a reduced grid, so framework effects and model effects remain partly entangled. Failure modes are defined by the first deviation from a correct path, which leaves composite failures unresolved and makes every category count an underestimate.

Read as operational quantities, the token figures are lower bounds, since consumption depends on context beyond what these suites supply. The practical implication is that a configuration should be selected on accuracy per token rather than on accuracy alone, and that the round bound is better treated as a budgeting device, with no statistically distinguishable contribution of the third round detected.

5. Conclusion

This study decomposed a language-model agent system into a reasoning operator inside each agent and an interaction topology between agents, holding the number of agents constant at three across all multi-agent topologies, together with the round bound, the message cap, the prompt skeleton, and the memory setting so that the two factors vary independently. A full factorial grid of four reasoning levels and four topologies was evaluated on three public collections under five random seeds, covering 484 instances and 38,720 task executions, with comparisons paired by instance and effects estimated from a logistic mixed-effects model fitted to each collection separately. Both factors carry main effects, but their relative strength differs between the static and the interactive case, reasoning spanning 18.0 points on the former against topology spanning 7.8 points on the latter, a span measured across all four topology levels and therefore inclusive of the step from one agent to three, against a structure-only span of 2.8 points among the multi-agent topologies. The two effects depart from additivity in one localised cell. A mechanism-motivated single-degree-of-freedom contrast on the only cell in which both components have a correction pass is significant at $p = 0.001$ on the static suite, whereas the omnibus interaction test across the grid is not, and in that cell the observed value of 50.0 percent falls 5.8 points short of the additive prediction. Once tokens are counted, seven of the sixteen configurations are undominated on the static suite and five on the operating-system suite, and improvements become marginal above 12,400 tokens per instance on the static suite under the present design. These results recommend reporting accuracy together with the consumption that produced it and varying one factor at a time. Three directions follow. Decentralised and asynchronous protocols

would remove the centralised coordination assumed throughout. Treating the topology as a learnable variable rather than fixing it at a level would let connectivity adapt to the task instance, interpolating between the discrete levels enumerated here and placing message edges where they remove specific failure modes, so that the accuracy–cost frontier is traced directly rather than sampled at fixed design points (Zhuge et al., 2024). Re-estimating the interaction on long-horizon, repository-scale problems would show whether the localisation observed here survives at that scale.

Acknowledgments

Not applicable.

Author contributions

Conceptualization, Z.W.; methodology, Z.W. and Y.W.; software, Z.W. and Y.W.; validation, Y.W. and Y.D.; formal analysis, Z.W.; investigation, Y.W. and Y.D.; data curation, Y.W. and Y.D.; writing—original draft preparation, Z.W.; writing—review and editing, Z.W., Y.W. and Y.D.; visualization, Y.D.; supervision, Z.W.; project administration, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data availability

The data presented in this study are available on request from the corresponding author.

Declarations

Conflict of interest

The authors declare that there is no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Additional information

Received: 2 September 2026

Accepted: 20 September 2026

Transactions on Computing Science
WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.519 Volume 2 , Issue 2
Published online: 23 October 2026

Publisher's Note

Wisdom Academic Press Ltd. remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

[1] Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., & Hoefler, T. (2024). Graph of Thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17682–17690. https://doi.org/10.1609/aaai.v38i16.29720

[2] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 11733–11763). PMLR. https://proceedings.mlr.press/v235/du24e.html

[3] Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2024). MetaGPT: Meta programming for a multi-agent collaborative framework. In *Proceedings of the 12th International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=VtmBAGCN7o

[4] Li, G., Hammoud, H., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, *36*, 51991–52008. https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html

[5] Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., & Tang, J. (2024). AgentBench: Evaluating LLMs as agents. In *Proceedings of the 12th International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=zAdUB0aCTQ

[6] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (Article 2). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

[7] Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., & Sun, M. (2024). ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 15174–15186). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.810

[8] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, *36*, 68539–68551. https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html

[9] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, *36*, 8634–8652. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html

[10] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=uyTL5Bvosj

[11] Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., & Wei, J. (2023). Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13003–13051). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.824

[12] Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2024). Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2305.16291

[13] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), Article 186345. https://doi.org/10.1007/s11704-024-40231-1

[14] Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=1PL1NIMMrw

[15] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[16] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2024). AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *Proceedings of the First Conference on Language Modeling*. OpenReview. https://openreview.net/forum?id=tEAF9LBdgu

[17] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., … Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, *68*(2), Article 121101. https://doi.org/10.1007/s11432-024-4222-0

[18] Yao, S., Chen, H., Yang, J., & Narasimhan, K. (2022). WebShop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, *35*, 20744–20757. https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html

[19] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, *36*, 11809–11822. https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html

[20] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=WE_vluYUL-X

[21] Zhang, G., Yue, Y., Li, Z., Yun, S., Wan, G., Wang, K., Cheng, D., Yu, J. X., & Chen, T. (2025). Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *Proceedings of the 13th International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbc461518c59a2a8d64e70e2c38c4a0e-Abstract-Conference.html

[22] Zhang, H., Cui, Z., Chen, J., Wang, X., Zhang, Q., Wang, Z., Wu, D., & Hu, S. (2025). Stop overvaluing multi-agent debate: We must rethink evaluation and embrace model heterogeneity (No. arXiv:2502.08788). arXiv. https://doi.org/10.48550/arXiv.2502.08788

[23] Zhuge, M., Wang, W., Kirsch, L., Faccio, F., Khizbullin, D., & Schmidhuber, J. (2024). GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 62743–62767). PMLR. https://proceedings.mlr.press/v235/zhuge24a.html

Appendix A

Failure attribution protocol

Each failed trajectory is assigned to exactly one category at the earliest step whose output is inconsistent with every correct continuation of the task. Categories are tested in a fixed order, namely per-step computation error, decomposition or sequencing error, false factual statement, distortion of a peer position on restatement, and premature consensus on an incorrect answer, and the first matching category is assigned; a trajectory matching none is labelled unresolved, which occurred in 6.2 percent of failures, so all category counts are lower bounds. Two authors annotated a stratified 20 percent sample independently and blinded to the configuration label, reaching Cohen's $\kappa = 0.81$; disagreements were resolved by discussion, and the remaining trajectories were annotated by one author under the agreed rules.

Table A1

Cross-model validation on Qwen2.5-7B-Instruct, reduced grid {R0, R3} × {T0, T1}. Mean over five seeds with standard deviation in parentheses

Configuration	BBH-8	AB-OS	AB-WS
R0-T0	41.3 (1.5)	6.9 (1.0)	9.0 (2.4)
R0-T1	45.9 (1.8)	11.1 (1.4)	8.0 (2.8)
R3-T0	55.4 (1.6)	13.2 (1.3)	14.0 (2.9)
R3-T1	55.0 (2.0)	14.6 (1.5)	13.0 (3.1)

Note. Additive expectations for R3-T1 are 60.0 on BBH-8 and 17.4 on AB-OS, giving observed shortfalls of 5.0 and 2.8 points, of the same sign as on the main backbone; on AB-WS the base rate lies at the identifiability floor and R0-T1 constitutes the single non-reproducing configuration discussed in the main text.

Table A2

Round-by-round execution trace of a failed R3-T1 run on BBH-8 (correct answer B)

WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.519 Volume 2 , Issue 2

Round / line	Event	Content
t = 1, line 4	Proposals under R3	Agents a1, a2, a3 answer (A, A, B)
t = 1, line 6	Routing under T1	Every agent receives both peer positions
t = 1, line 7	Reflexion revision (1 of ≤ 2)	a3 abandons B for the majority answer A; a1 and a2 confirm A
t = 1, lines 9–11	$\rho = 1 \geq \tau$, terminate	Φ returns A (incorrect); the run ends at round one
t = 2 (foregone)	Remaining reflexion budget	The corrective passes that would re-examine A against the routed reasoning never execute

The trace shows that within round one revision precedes the consensus check, and that what unanimity forecloses is the second round in which the reflexion operator would have spent its remaining budget.