

Multi-agent scientific reasoning

Agent roles, information access and responsibilities.

Reasoning agents, feedback, and scientific evidence

Agreement can repeat one underlying error

Three agents agree on a conclusion because all retrieve the same incorrect summary. Additional agreement repeats one underlying error.

The missing resource could be independent evidence, specialized computation or checking authority. Another role name supplies none of these automatically.

This reframes the design question: what useful difference could a second agent introduce, and can one agent obtain it just as well?

Designing a division of evidence and work

Give a proposed role a concrete input and output. A source checker might inspect the primary record while a proposer works from retrieved summaries.

The division is useful only if its information or expertise improves the result enough to offset communication costs. The same evidence may also be routed to one agent.

Compare those alternatives before choosing a team. A critic must return a checkable defect or missing test that can change the next decision.

Choosing whether to add agents

Candidate	Reason to choose it	Cost or limitation
One agent with more resources	The task fits one context and tool policy	May miss complementary evidence or expertise
Several samples from one model	Diversity and an aggregation rule are useful	Shared errors can survive voting
Specialized interacting agents	Distinct evidence or work can be divided	Handoffs, coordination and checking cost resources

Add agents when the division contributes useful information or computation at a matched budget. Test that contribution independently of role names.

Pedagogical reconstruction

[Wu et al. \(2023\)](#)

[Cemri et al. \(2025\)](#)

Communication introduces its own failure modes

Agents can disagree about goals, lose information in handoffs, repeat completed work, or accept unsupported claims from another agent.

Use explicit task state, source pointers, and completion criteria. A coordinator should resolve uncertainty through evidence and defined decision rules.

More conversation can increase cost and propagate mistakes. Measure communication overhead and error recovery alongside task success.

Voting helps only under relevant assumptions

For independent agents each more likely correct than incorrect, majority voting can improve binary decision accuracy.

Shared models, prompts, retrieval sources, and reward signals create correlated errors. Confidence and consensus then provide less independent evidence than the number of agents suggests.

Empirically estimate disagreement and shared failure cases. A single strong agent with multiple independent tests may outperform a larger conversational group.

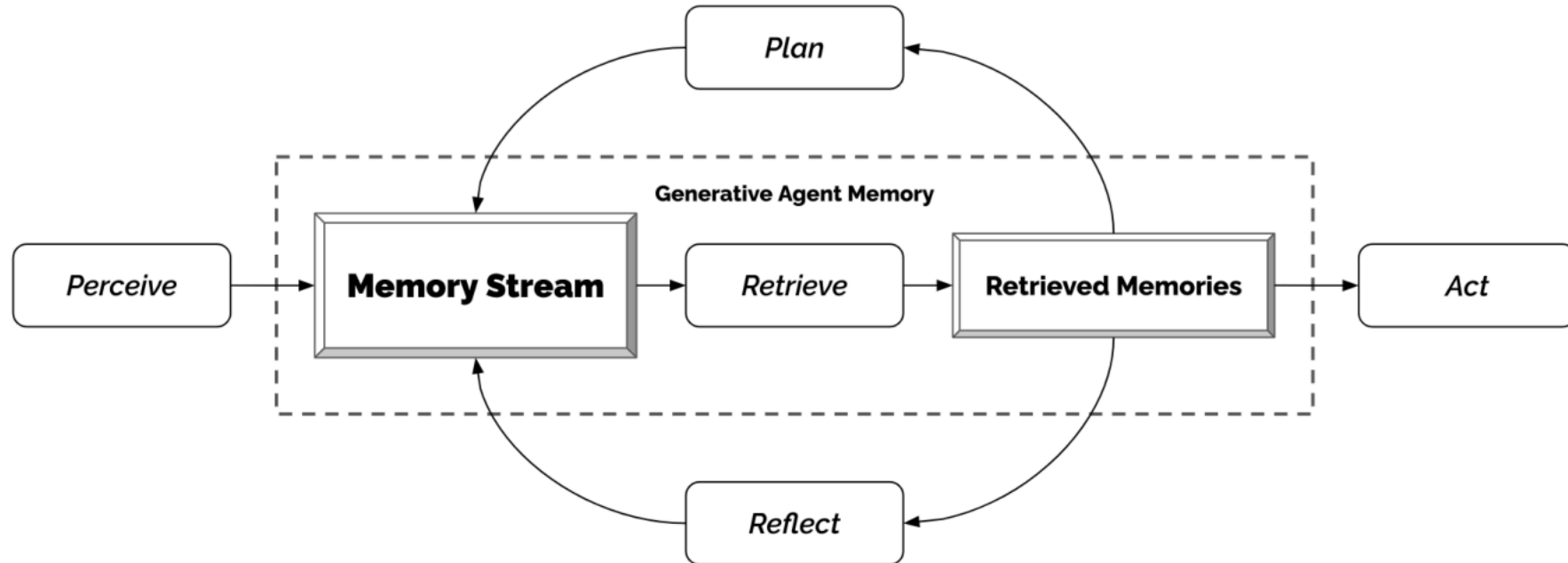
Generative Agents: the method

The architecture records observations in a memory stream, retrieves relevant memories, forms higher-level reflections, and uses them to plan behaviour.

The environment is a simulated town populated by language-model agents.

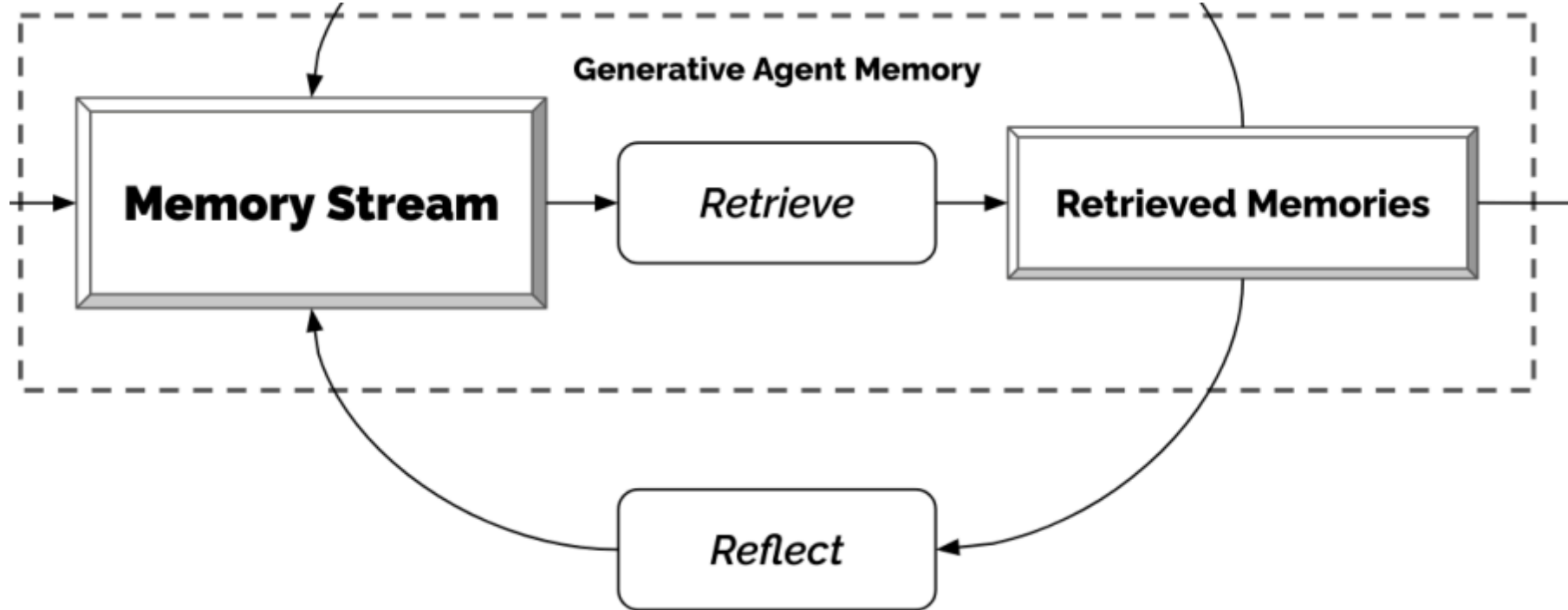
The main target is believable behaviour and social interaction. This makes the paper useful for studying memory and coordination mechanisms, while requiring care when transferring its claims to factual research.

Generative Agents: memory and planning architecture



Perception feeds retrieval and reflection, which then inform planning. A stored observation differs from a reflection inferred from several observations; a later correction can invalidate the inference without changing every source memory.

Generative Agents: mechanism detail



Reflection derives statements from lower-level observations. Its relation to memory raises a separate issue of whether support dependencies remain explicit.

Evaluation scope

The evaluation includes human judgements of believability and ablations of memory, planning, and reflection.

The ranking metric and evaluated variants define this comparison.

A system can produce coherent social behaviour without maintaining a scientifically calibrated model of the world; the evaluation target determines what the evidence establishes.

Generative Agents: interpretation and limits



Believability is a human-rated outcome. It should not be substituted for factual accuracy, independence of evidence, or experimentally validated coordination in a laboratory.

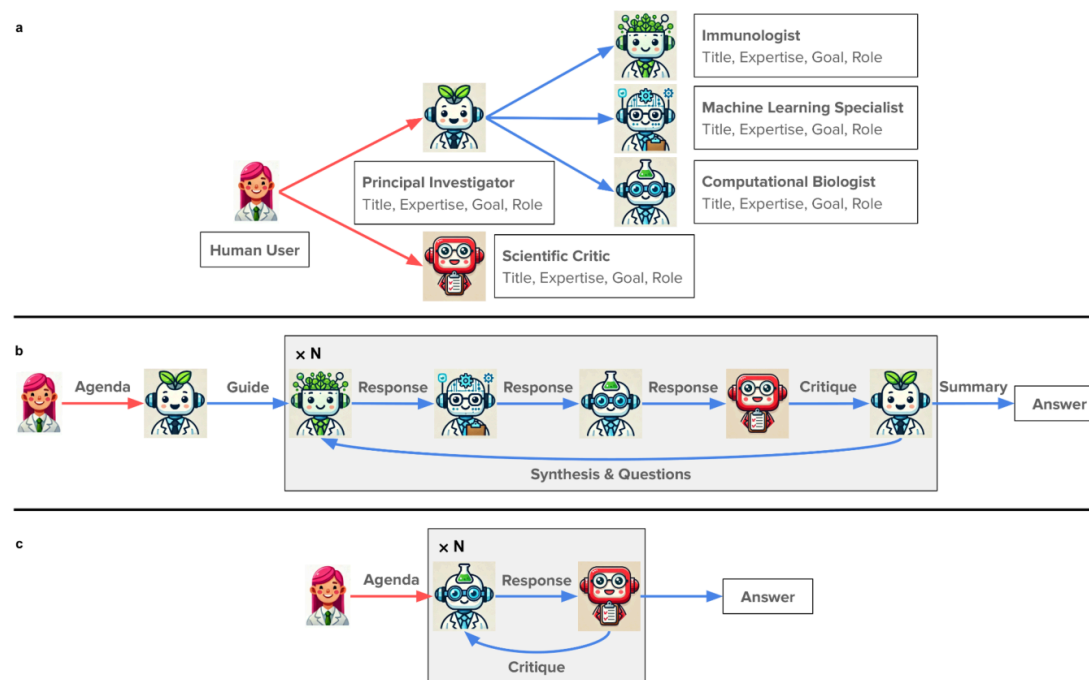
The Virtual Lab: the method

The Virtual Lab organises a human–AI research collaboration through specialised language-model agents, discussions, and computational tools. The study applies the framework to nanobody design with subsequent experimental validation.

Human researchers define goals and conduct physical experiments.

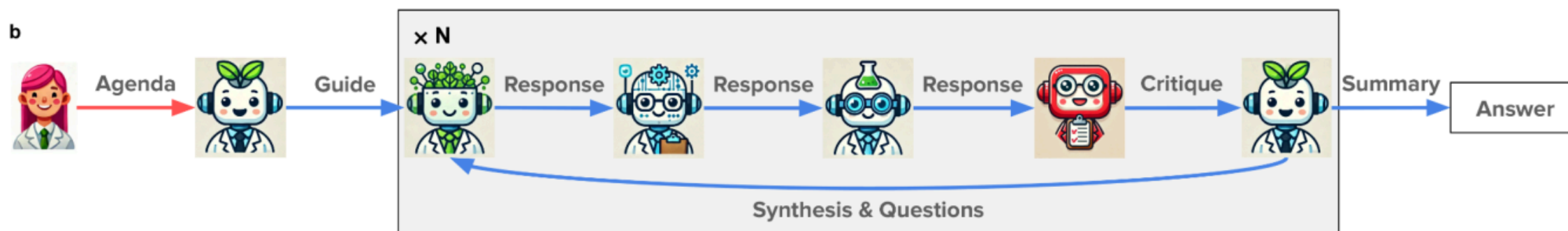
The case connects agent-generated research plans with measured outcomes, while the contribution of individual coordination mechanisms requires controlled comparison.

The Virtual Lab: roles and meeting structure



Agent roles and meeting loops organize the agenda and return critique to discussion. The communication diagram describes coordination; experimental results concern the separately evaluated application workflow.

The Virtual Lab: mechanism detail



The meeting loop combines specialist responses and critique into a final answer. Repeated messages can share the same evidence. Artifacts linking the discussion to a changed decision are needed to attribute improvement.

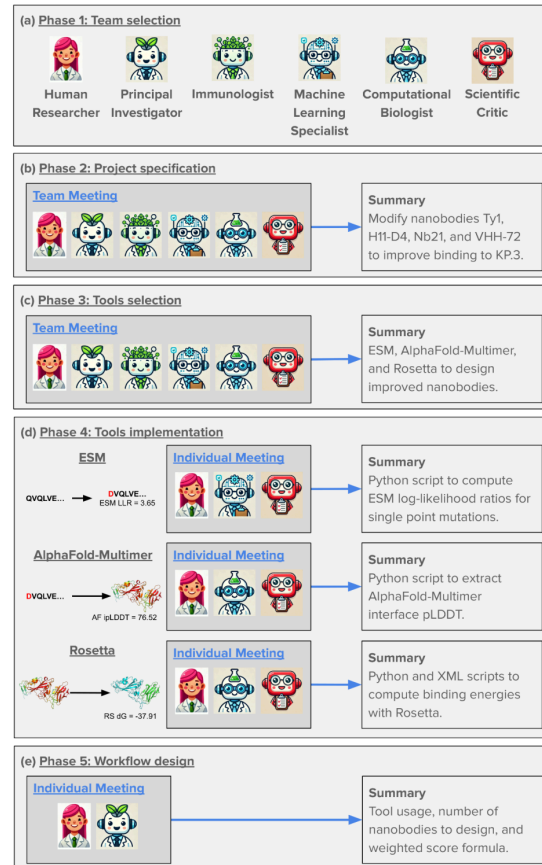
Evaluation scope

The study reports a computational design workflow and experimental assessments of selected candidates.

The displayed pipeline identifies the transformations and selection stages preceding validation.

Interpret experimental results using the candidate-selection denominator, controls, and measured assay outcome; validation of selected candidates does not imply that every proposed design succeeds.

The Virtual Lab: interpretation and limits



The evidence concerns a specific research task with human participation and specialised tools. Separate system-level feasibility from a claim that agent role labels or debate independently caused the result.

Proposal, critique and commitment

A proposal role suggests hypotheses; a critique role identifies assumptions and discriminating tests; an execution role returns validated artefacts; a coordinator decides what to do next.

These boundaries are workflow choices, not empirical guarantees about intelligence. Their value depends on the quality of the interfaces and checks.

Preserve who supplied each claim and which observation justified the eventual commitment.

Adversarial discussion needs a common evaluator

Debate can reveal overlooked assumptions, but persuasive text can also displace correct reasoning. Give agents access to evidence and a way to test factual disagreements.

Evaluate the final decision using an independent criterion. Do not define success as increased agreement after discussion.

Compare discussion with independent proposals followed by a fixed aggregator, keeping model calls and tool access comparable.

Specialisation should survive an ablation

Remove or merge one role at a time. Replace specialist prompts with generic prompts. Vary information access separately from the number of agents.

Match compute and tool calls against a single-agent baseline that receives equivalent resources.

A role is useful when its removal causes a reproducible loss on the target task or a clearly identified class of failures, not merely when its transcript looks sophisticated.

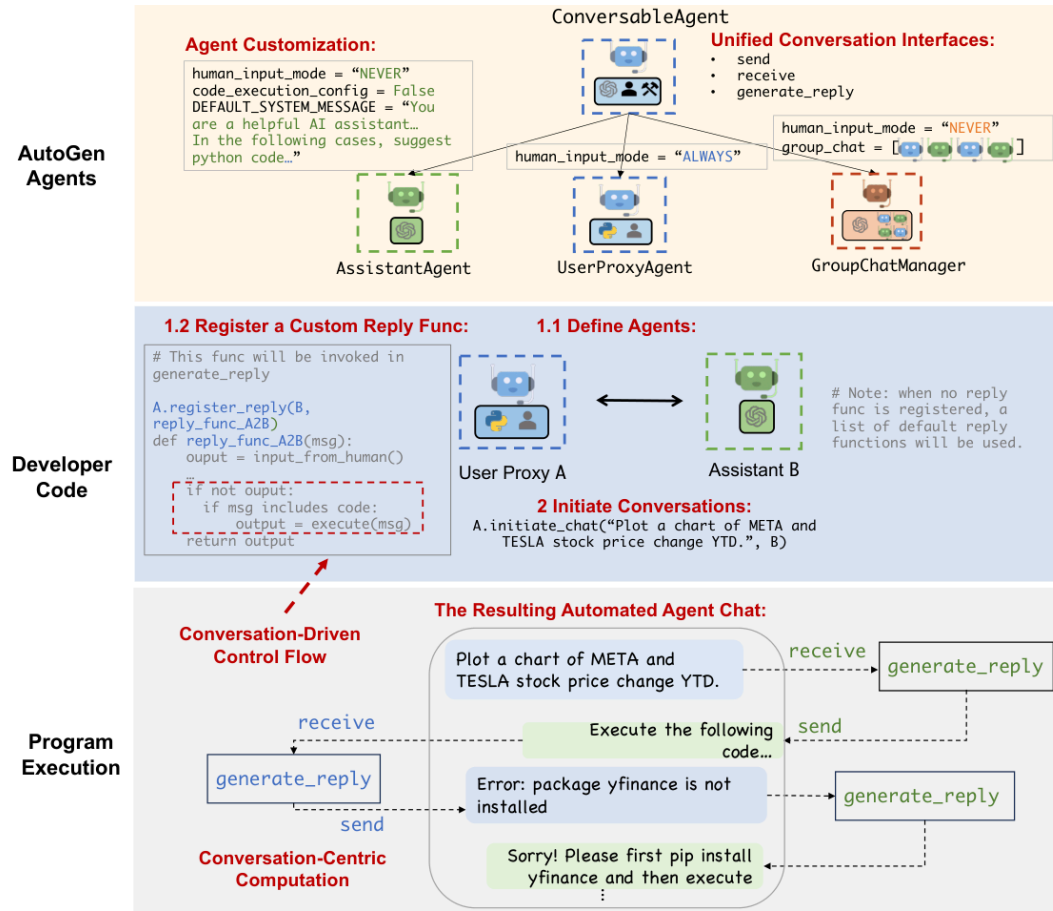
AutoGen: the method

AutoGen provides conversable agents with configurable model, tool, and human participation. Applications specify interaction patterns and termination conditions.

The framework supports several computational arrangements, including code generation followed by execution and feedback.

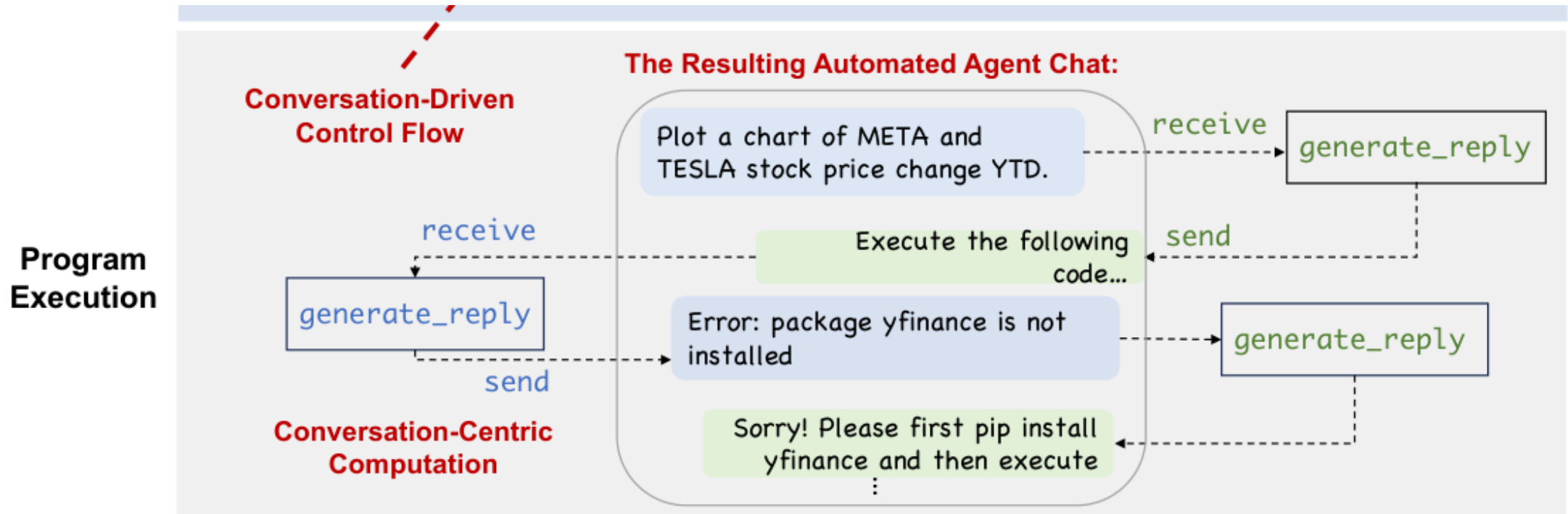
It is infrastructure for constructing a system, so the value of a particular collaboration pattern must be evaluated separately from the availability of the framework.

AutoGen: interfaces, configuration and trace



The three layers distinguish available interfaces, a two-agent configuration and its execution trace. Proposer, executor and feedback channel define the configuration, whose performance remains task dependent.

AutoGen: mechanism detail



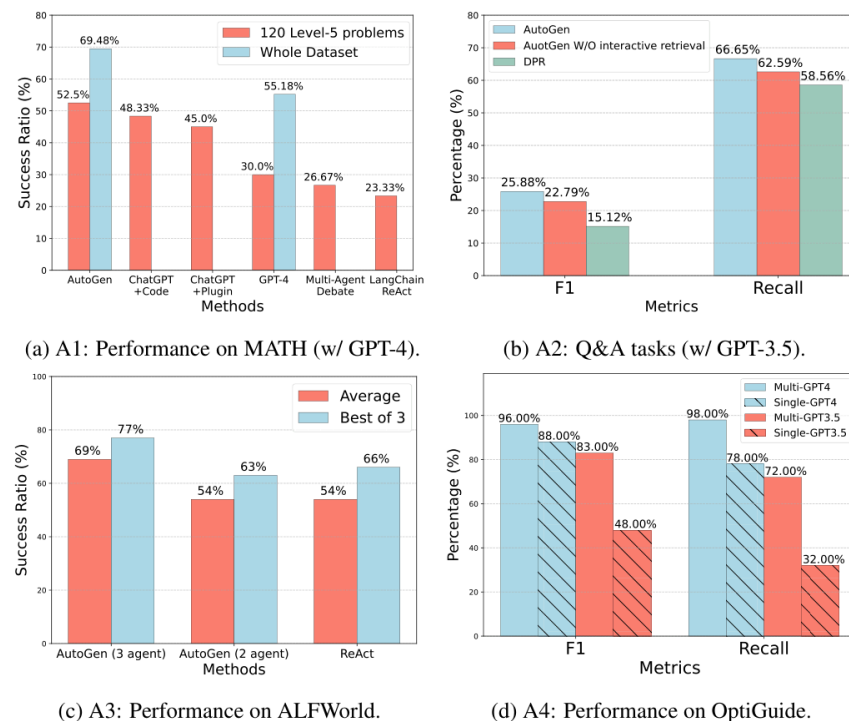
Execution feedback can convey new observations. A model's restatement of shared context carries a different evidential contribution.

Evidence: a third role also adds information

Method	Pick	Clean	Heat	Cool	Look	Pick 2	All
ReAct (avg)	63	52	48	71	61	24	54
ALFChat (2 agents)(avg)	61	58	57	67	50	19	54
ALFChat (3 agents)(avg)	79	64	70	76	78	41	69
ReAct (best of 3)	75	62	61	81	78	35	66
ALFChat (2 agents)(best of 3)	71	61	65	76	67	35	63
AFLChat (3 agents)(best of 3)	92	74	78	86	83	41	77

AutoGen's ALFWorld comparison reports 54% average success with two agents and 69% with three; best-of-three rates are 63% and 77%. The added grounding agent supplies relevant facts. Aggregation and information access therefore differ from the effect of assigning another role.

AutoGen: interpretation and limits



Conversation overhead, repeated errors, and unclear termination can erase the benefit of role separation. Compare against a single-agent system with the same tools and total inference budget.

MAST: where multi-agent systems fail

The Multi-Agent System Failure Taxonomy (MAST) groups 14 failure modes into system design, inter-agent misalignment and task verification: defective setup, failed coordination and unchecked outcomes.

The NeurIPS 2025 paper reports 1,600+ annotated traces from seven frameworks; taxonomy development used 150 traces. These are distinct samples. Their failure frequencies are not universal rates for all agent teams.

A two-factor comparison

System	Shared evidence	Diverse evidence
Single agent	10/20 correct	15/20 correct
Multiple agents	11/20 correct	16/20 correct

Synthetic counts with matched total compute and the same twenty tasks. The table varies architecture and evidence access separately.

Read effects within matched conditions

In this example, multiple agents gain one success at either evidence level. Diverse evidence gains five successes for either architecture.

The raw difference between single/shared and multiple/diverse combines both changes. A causal interpretation also depends on how the conditions were assigned and controlled.

Costs and authority belong in the comparison

Quantity	What to record	Why it matters
Compute	All agents and judges	Extra work can explain gains
Latency	Critical path and waits	Parallelism has coordination cost
Human effort	Minutes and decisions	Experts may solve the hard step
Authority	Who may act or revise	Errors can propagate into actions

Report the unit of evaluation as the complete system trajectory, with component costs retained.

Zero pairwise correlation does not imply independent voting

Operation	Calculation or state	Consequence
Population I	000, 011, 101, 110 each with probability 1/4	Majority correct: 3/4
Population II	111, 100, 010, 001 each with probability 1/4	Majority correct: 1/4
Every individual	Correct with probability 1/2 in both	Same marginal accuracy
Every pair	Joint correctness 1/4 in both	Pairwise correlation zero

Synthetic correctness vectors for three agents on binary-answer tasks. Matching marginal accuracy and pairwise correlation does not determine majority-vote accuracy.

Worked voting: three independent agents

Agents right	Patterns	Probability	Majority correct?
0	1	$0.3^3 = 0.027$	No
1	3	$3 \times 0.7 \times 0.3^2 = 0.189$	No
2	3	$3 \times 0.7^2 \times 0.3 = 0.441$	Yes
3	1	$0.7^3 = 0.343$	Yes

Three independent agents each answer correctly with probability 0.7. The multiplicities count which agents are right; the last two rows give the winning cases. These assumptions create the benefit. Shared evidence or identical errors can remove it even if the agents have different role names.

Calculate the majority-vote accuracy

$$P(\text{majority correct}) = 3p^2(1 - p) + p^3 = 0.784$$

The two-correct term is 0.441 and the all-correct term is 0.343, giving total accuracy 0.784. This exceeds the individual accuracy of 0.7.

The gain comes from partly independent errors and the aggregation rule. It does not follow just from having three role names.

If all agents always make the same error, majority voting has the same accuracy as one agent.

Model a shared-source failure

Suppose all three agents retrieve the same incorrect secondary summary. Their answers can agree while depending on one underlying error.

Asking an additional agent to read that same summary may add little evidence. Asking a verifier to inspect the primary measurement can change the dependence structure.

Record source overlap and shared tools when evaluating ensemble reliability. The effective diversity is a property of the evidence-generating process.

Evidence in a multi-agent scientific trace

A scientific agent trace links a proposal, a critique with a testable objection, an execution record and a coordinator decision grounded in the decisive evidence.

Resolved objections, measurement-dependent decisions and visible contradictory results provide evidence about the quality of the coordination process.

A single-agent implementation of the same workflow helps isolate agent separation from improvements in procedure or a larger budget.

Positive correlation reduces effective sample size

$$\text{Var}(\bar{X}) = \frac{\sigma^2[1 + (m-1)\rho]}{m}, \quad m_{\text{eff}} = \frac{m}{1 + (m-1)\rho}$$

Assume m exchangeable binary correctness indicators with common variance σ^2 and pairwise correlation ρ . Summing individual variances and pairwise covariances gives the expression.

With $m=5$ and $\rho=0.5$, $m_{\text{eff}} \approx 1.67$ for estimating mean correctness.

This effective sample size is a variance comparison, not a formula for majority-vote accuracy. Higher-order dependence also matters for the vote outcome.

From an agent trace to a testable diagnosis

Illustrative trace: a planner requests an evidence check, a researcher returns a claim without its source, and a writer treats it as verified. The final critic accepts the draft without checking the evidence.

MAST distinguishes information omitted at handoff from missing final verification. A trace can contain several failures; the last visible error does not identify the earliest cause.

A handoff with explicit claim-to-source links and an independent final check targets those mechanisms. Whether it improves reliability requires a comparison on the same tasks and resource budget.

Disagreement is useful when its cause is exposed

Agent role	Message in the shared case	Evidence needed
Measurement specialist	“The temperature pattern suggests sensor drift”	Raw values and calibration conditions
Provenance specialist	“The reference certificate expired”	Certificate ID and valid dates
Mechanism specialist	“A process change also predicts this pattern”	A discriminating process measurement
Coordinator	Do not turn a two-to-one vote into certainty	Resolve the evidence conflict first

Synthetic multi-agent exchange. Diverse roles can reveal different assumptions; repeated access to one faulty source can correlate every answer.

Coordination should produce a consequential update

Unresolved issue	Coordination action	Observable outcome
Reference reliability disputed	Request an independent calibrated reference	A new evidence item, not another opinion
Agents repeat the same source	Expose source IDs and dependence	Duplicate support is counted once
A result changes	Notify every role using the old result	Dependent conclusions are revised
Evidence remains insufficient	Record the disagreement and stop condition	A bounded claim with an explicit uncertainty

The unit of progress is a changed, evidence-supported decision. Conversation length alone does not measure scientific collaboration.

Recent methods in context

Agent Primitives (ICML 2026). Composes reusable agent operations using KV-cache communication and an organizer that chooses a configuration for each query. Attribution to collaboration or latent communication depends on matched backbones and resource budgets.

Co-Scientist (Nature 2026). Co-Scientist organizes hypothesis generation, critique and tournament-based refinement. The Nature publication adds peer-reviewed evidence for selected applications. Internal rankings and external scientific validation support different claims.

EXP-Bench (ICLR 2026). Evaluates agents on experimental procedures reconstructed from AI research papers and code. Faithful execution of a prescribed experiment is narrower than open-ended creativity or a new scientific discovery.

References (1/2)

[Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai et al. \(2023\). Generative Agents: Interactive Simulacra of Human Behavior](#)

[Kyle Swanson, Wesley Wu, Nash L. Bulaong et al. \(2025\). The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation \(preprint used for figures; journal version linked\)](#)

[Qingyun Wu, Gagan Bansal, Jieyu Zhang et al. \(2023\). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation](#)

[Mert Cemri, Melissa Z. Pan, Shuyi Yang et al. \(2025\). Why Do Multi-Agent LLM Systems Fail?](#)

[Johan de Kleer \(1986\). An assumption-based TMS](#)

References (2/2)

Samuel Schmidgall, Xiaokai Zhu, Marian Shaw et al. (2026). Accelerating Scientific Research with Gemini in the Real-World

Haibo Jin, Peng Kuang, Ye Yu et al. (2026). Agent Primitives: Reusable Latent Building Blocks for Multi-Agent Systems

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin et al. (2025). Accelerating scientific discovery with Co-Scientist

Patrick Tser Jern Kon, Qiuyi Ding, Jiachen Liu et al. (2025). EXP-Bench: Can AI Conduct AI Research Experiments?