

TECHNICAL RESEARCH REPORT · 01

JARVIS & SYSTEM BRAIN

Towards accountable machine decision-making and progressively autonomous departments.

J

15-cell prototype
100-specialist research horizon

Architecture · prototype inspection · controlled simulation

Engineer / Mohanad Osama Zahran

Artificial Intelligence Expert

ZEISMO Research & Lab

28 September 2026 · Version 1.3

ZEISMO-LAB-TR-2026-001

DOCUMENT CONTROL

A research record built for inspection.

Technical report and experimental programme. Prepared for the ZEISMO research archive.

FIELD	RECORD
Author / project direction	Engineer Mohanad Osama Zahran · Artificial Intelligence Expert
Publisher	ZEISMO Research & Lab
Operator	WEARLISTY SHOPPING SOLUTIONS LLC
Report identifier	ZEISMO-LAB-TR-2026-001 (local technical-report identifier)
Edition	1.2 · 28 September 2026 · English
Research object	JARVIS interface and System Brain cellular-agent architecture
Evidence	Source inspection, isolated prototype tests, synthetic experiments and reproducibility checks
Review state	Internal automated technical verification completed; independent scientific peer review not recorded
Availability	Public technical preprint at zeismo.org

This report develops ZEISMO's own research programme. It combines an inspection of an existing software prototype with an original, reproducible simulation study. Architectural proposals are identified as proposals; experimental results are identified by their data source and scope.

AI assistance was used for literature discovery, technical analysis, simulation code, manuscript drafting and document design. Mohanad Osama Zahran supplied the project conception and research direction, and authorised the typed name and professional title in this edition. Computational analysis and document preparation were AI-assisted. His author attribution does not constitute an independent scientific review.

A companion evidence package retains the experiment code, generated data, output tables, reference register, source-file digests and verification results. The report identifier organises this edition locally. It is not a DOI. The PDF's final SHA-256 digest is supplied outside the PDF to avoid a circular checksum.

ABSTRACT

From conversational assistance to governed action.

Research question: how can role-specialised AI systems assume routine decision work while preserving evidence, controllability and measurable reliability?

We propose JARVIS as an exclusively English-speaking interface to System Brain, a coordination layer for role-specific software agents. Each proposed AI employee combines a role contract, curated skills, scoped tools, memory and an evaluation record. “Cells” name modular software responsibilities; “feel” names a measurable appraisal of urgency, uncertainty and consequences.

The existing Souq implementation provides a concrete engineering antecedent. Source inspection identifies fifteen initial cell slots, knowledge imports, persistent memory, bounded delegation and a recommendation council. Seven isolated prototype tests pass across thirty-two assertions. These tests establish selected application behaviours, rather than the quality of model-generated decisions.

Four controlled experiments investigate the decision-control substrate. Across 12,000 synthetic tasks, the complete policy gate executes 1,888 proposals (15.7%) with zero violations of the encoded policy, while proposal correctness among those executions is 85.1%. Thus policy compliance and semantic accuracy are distinct. Provenance filtering improves retrieval validity on explicitly labelled synthetic records, and correlated-voter simulations show why adding agents does not guarantee independent expertise.

The programme aims to replace selected human-operated routine workflows progressively, using measured task-level acceptance. Equivalence to a panel of one hundred specialists and end-to-end millisecond decision-making remain hypotheses requiring separate expert and live-model studies. This report supplies the definitions, architecture, evidence and experimental protocol needed to investigate them.

15

initial cell slots in inspected
prototype

12,000

synthetic decision tasks

7 / 32

prototype tests / assertions

READING GUIDE

Contents · foundations and architecture.

A structured route from the research ambition to an implementable system.

An AI workforce, introduced step by step.	06
Questions that can be answered with evidence.	07
Define the language before testing the claim.	08
Reasoning, tools and coordinated roles.	09
Memory and skills are different resources.	10
Evaluate behaviour, not the appearance of intelligence.	11
What the existing implementation establishes.	12
Seven tests. Thirty-two assertions.	13
JARVIS: one English-speaking interface.	14
System Brain: coordination with explicit boundaries.	15
An AI employee contains accountable cells.	16
Import skills according to the position.	17
Memory that can be corrected and trusted.	18
Self-improvement as a controlled experiment.	19
Business weather: an observable context model.	20
A measurable interpretation of “feel.”	21
A constrained decision problem.	22
Collective action with preserved disagreement.	23
The “100 specialists” hypothesis.	24
Milliseconds: measure the full path.	25
A staircase of earned autonomy.	26
Customer operations: consistency before volume.	27
Inventory and logistics: decisions under change.	28
Finance and procurement: evidence before authority.	29
Marketing and research: useful creativity with traceability.	30
Measure replacement at the workflow level.	31
Role authority is a technical boundary.	32
Treat incoming content as untrusted evidence.	33
Preserve data provenance and minimise exposure.	34

READING GUIDE

Contents · experiments and verification.

Methods, measured synthetic results, limitations and a reproducible evidence record.

Four experiments on the control substrate.	35
Policy gating over proposed decisions.	36
Coverage falls as controls become stricter.	37
A policy-valid decision can still be wrong.	38
Memory retrieval with conflicting evidence quality.	39
Provenance filtering changes what memory can influence.	40
How many perspectives are actually independent?	41
A hundred votes do not erase a shared mistake.	42
Fast routing is only one part of the system.	43
Ablations explain mechanisms, not universal performance.	44
Keep denominators, seeds and uncertainty visible.	45
A useful lab makes failure diagnosable.	46
A complete live-model evaluation protocol.	47
Design the specialist comparison fairly.	48
Verify English-only interaction end to end.	49
Maintain competence as conditions change.	50
A changing inventory decision, step by step.	51
The role and skill manifest.	52
The decision and memory record.	53
An implementation sequence with measurable exits.	54
Release decisions need explicit evidence.	55
A foundation for research, with a clear next test.	56
ZEISMO Lab technical verification record.	57
Re-run the study and inspect the evidence.	58
References · agent architecture	59
References · evaluation and security	60
References · governance and research records	61
Glossary and notation.	62
The JARVIS research charter.	63
Inside the System Brain.	64
A cell's skills and memory.	65
Open the memory archive.	66
JARVIS as the system interface.	67
Participant briefing and task sheet.	68
Your assessment of the experience.	69
Decision boundaries and feedback.	70
Engineer Mohanad Osama Zahran.	71
Build decisions that can be examined.	72

01 · RESEARCH PURPOSE

An AI workforce, introduced step by step.

ZEISMO's objective is to transfer suitable routine departmental work from human operators to accountable AI decision systems.

The proposed transition begins with tasks, not job titles. A department may contain repeatable record checks, creative judgement, negotiations, physical work and decisions with serious consequences. These activities require different evidence and different operating permissions. An AI employee therefore receives a bounded role description rather than unrestricted control over an entire business function.

Replacement has a specific operational meaning in this report: an accepted workflow no longer requires a human to perform each ordinary case. People retain responsibility for defining its purpose, approving its authority, resolving exceptions and evaluating outcomes. The aspiration is substantial automation, with expansion justified by results rather than by a persuasive interface.

The first deployment candidates are reversible digital operations whose input records, allowable actions and success conditions can be documented. Examples include routing a support request, preparing an inventory alert, reconciling a known invoice pattern and assembling a research evidence pack. A shared System Brain coordinates dependencies between these roles.

A staged programme should also account for organisational change. Reassignment, training, communication and retained expertise are implementation requirements. Removing human capacity before an automated process has demonstrated stable performance can leave the organisation without a recovery path.

CORE PROPOSITION

Automate a defined responsibility only after the system can demonstrate the required quality, observe its own operating limits and return control when those limits are exceeded.

01 · RESEARCH PURPOSE

Questions that can be answered with evidence.

The contribution is a testable integration programme and an initial controls study, rather than a claim to have invented intelligence.

QUESTION	MEASUREMENT	EVIDENCE IN THIS EDITION
RQ1· Can typed policy gates constrain proposed actions?	Violation rate, execution coverage and abstention	Executed synthetic study
RQ2· Can memory selection preserve evidence quality?	Valid retrieval, invalid retrieval and abstention	Executed synthetic study
RQ3· Does a larger council improve decisions?	Correctness under independent and shared errors	Executed surrogate-voter study
RQ4· How quickly does control routing execute?	Warmed in-process batch timing	Executed microbenchmark
RQ5· Can AI match a specialist panel?	Blinded, task-matched expert comparison	Prospective study protocol
RQ6· Can roles maintain competence over time?	Held-out performance after controlled updates	Prospective longitudinal protocol

The report contributes an explicit vocabulary for cells, skills, memory and autonomy; a layered design separating recommendations from execution; a source-backed inventory of the prototype; and an evidence package that can reproduce the reported synthetic results. Its proposed departmental transition model connects technical evaluation to operational acceptance.

Novelty must be assessed against existing work and an eventual literature review broader than this targeted set. Role-based agents, tool use and memory already have substantial research precedents. ZEISMO's research opportunity lies in how a particular architecture combines these ideas and whether a controlled evaluation demonstrates useful outcomes in its chosen operating domains.

Research context [\[1\] Yao](#) [\[2\] Shinn](#) [\[4\] Wu](#) [\[12\] Li](#)

01 · RESEARCH PURPOSE

Define the language before testing the claim.

A precise vocabulary prevents a compelling metaphor from becoming an unsupported scientific conclusion.

TERM	OPERATIONAL DEFINITION
JARVIS	An exclusively English-speaking user interface and explanation layer in the proposed research system.
System Brain	Coordination, context assembly, policy routing and memory services shared across agents.
AI employee	A software role with explicit duties, permissions, evaluation criteria and an accountable organisational owner.
Cell	A modular software unit for a skill, memory function, evaluator or specialist responsibility.
Training	Either updating contextual knowledge or changing model parameters; every report must specify which.
Feel	An appraisal vector for context, urgency, uncertainty, conflict and consequence.
Decision	A selected action or explicit deferral, grounded in an evidence snapshot and policy version.
Self-maintenance	A bounded process that proposes, tests and promotes approved updates with rollback.

The existing prototype uses one cell as one agent slot. The richer architecture proposed here allows an AI employee to contain several functional cells, such as retrieval, planning and validation. That nesting is a design extension, not a capability established by the existing fifteen-slot implementation.

Persistent text memory is also distinct from changing a foundation model's weights. A model may adapt its response because retrieved instructions or experience changed, while its underlying parameters remain identical. No biological neurons, human feelings or measured consciousness are implied by the terminology in this study.

Research context [\[2\] Shinn](#) [\[6\] Packer](#) [\[17\] Vaswani](#)

02 · RESEARCH CONTEXT

Reasoning, tools and coordinated roles.

The programme builds on existing agent research and evaluates its own integration choices.

Acting on observations

ReAct combines reasoning-related steps with task actions and observations. It provides a useful antecedent for a loop in which a system gathers evidence before revising its plan. For JARVIS, the proposed extension is an explicit policy checkpoint between a recommendation and any action that changes business state. [1]

Conversation as coordination

AutoGen describes configurable conversations involving models, tools and humans. This supports examining coordination as an engineering object: message contracts, turn limits, participants and termination conditions. Our design adds an organisational role contract and a separate execution ledger to each workflow. [4]

Specialisation requires evaluation

CAMEL investigates role-based cooperative agent interaction. Giving an agent a specialist role can structure a task, but the title alone is not evidence of expertise. ZEISMO should evaluate each role using unseen cases and compare the result with a simpler baseline. [12]

These research precedents motivate experiments, not transferred performance claims. Reported gains on another paper's tasks cannot establish that a ZEISMO agent can manage inventory, authorise a payment or replace a department. The application must be tested under its own data, tools, policy constraints and failure costs.

Our proposed contribution is therefore practical and falsifiable: determine whether explicit context provenance, bounded coordination and independent action checks improve a selected workflow at an acceptable cost. If a single well-tested rule or single agent performs better, the architecture should retain the simpler option.

Research context [1] Yao [4] Wu [12] Li

02 · RESEARCH CONTEXT

Memory and skills are different resources.

A growing archive is useful only when the system can retrieve the right material and recognise its limits.

Reflexion explores adaptation through linguistic feedback stored for later trials, without requiring weight updates. This motivates a distinction between an observed outcome, a generated interpretation and an approved reusable lesson. In our design, they remain separate records. [2]

Generative Agents combines recorded experiences, reflection and retrieval in a simulated environment. Its role here is architectural inspiration for memory use, not evidence of workplace competence or human-equivalent behaviour. A business memory also needs ownership, validity dates and traceable corrections. [3]

Voyager uses a curriculum, an executable skill library and iterative feedback in Minecraft. Its domain-specific results motivate testing reusable skills; they do not establish that imported instructions are safe to execute against company systems. A ZEISMO skill package therefore includes scope, provenance and tests before activation. [5]

MemGPT presents hierarchical context management for working with memory beyond a model's immediate context window. Our corresponding design question is which evidence belongs in a short decision context and which belongs in a searchable archive. Recency alone is not an adequate selection rule. [6]

ZEISMO DESIGN POSITION

Store observations, proposed lessons and approved policy separately. Retrieve by relevance and permission, then check freshness and provenance before the material can influence an executable decision.

Research context [2] Shinn [3] Park [5] Wang [6] Packer

02 · RESEARCH CONTEXT

Evaluate behaviour, not the appearance of intelligence.

A convincing answer is one observation. A repeatable, correct operational outcome is a different and stronger one.

AgentBench evaluates agent behaviour across interactive environments, while GAIA tests general-assistant tasks involving reasoning and tools. They motivate evaluating complete task trajectories rather than only final prose. Neither benchmark is executed in this report. [7,9]

τ -bench tests policy-sensitive interactions among users, tools and agents and examines repeated-trial reliability. For this project, a useful adaptation is to verify the final business-state change and repeat each scenario, including retries and changed conversation order. [8]

Indirect prompt-injection research shows why retrieved material must be treated as data that may contain hostile instructions. A skill document, supplier message or web page should never acquire authority merely because it entered the model's context. [10]

Calibration research distinguishes confidence scores from empirical correctness. A JARVIS confidence value must therefore be evaluated on labelled cases. A fluent self-assessment cannot substitute for measured calibration. [11]

Evaluation structure

- Task validity: was the requested objective completed correctly?
- Policy validity: were all required constraints obeyed?
- Evidence validity: were claims supported by admissible records?
- Operational validity: did the actual tool state match the intended action?
- Temporal validity: does the result remain reliable after data and workflow changes?

The initial experiments examine selected control properties using synthetic records. Later studies must add live-model variability, real tool failures and properly authorised business cases.

Research context [7] Liu [8] Yao [9] Mialon [10] Greshake [11] Guo

03 · PROTOTYPE INSPECTION

What the existing implementation establishes.

Evidence source: a read-only inspection of selected Souq source files on 28 September 2026.

COMPONENT	OBSERVED IMPLEMENTATION	RESEARCH INTERPRETATION
Initial cells	Migration creates fifteen named slots	An initial software roster, not fifteen verified specialists
Skills	Operator pushes and discipline-based skill-file distribution	Contextual knowledge acquisition
Memory	Persistent records for skills, experience, research and tasks	An application memory layer
Delegation	Structured answer/delegate output and bounded depth	Controlled routing exists in source
Council	Four senior cells deliberate; a coordinating cell synthesises proposals	Recommendation workflow, not a tested autonomous board
Signals	Orders, messages, task activity and advertising data feed the interface	Operational grounding paths exist
Storage	Encrypted casts for cell skills and memory content	Selected fields protected by application encryption

Fourteen source files are listed by relative path and SHA-256 digest in the evidence package. The inventory includes controllers, service logic, models, selected commands, routes, a migration and two test classes. It contains source fingerprints, not production customer records or credentials.

Static inspection establishes what the selected code expresses. It does not demonstrate that every scheduled job is running, every supplier connection is healthy or every model response is correct. The test results on the following page establish a narrower set of behaviours in an isolated database.

The existing memory path retrieves up to twelve recent experience or research records and truncates each for the model context. The proposed provenance-aware retrieval and richer cell hierarchy are extensions to evaluate, not capabilities inferred from the existence of these records.

03 · PROTOTYPE INSPECTION

Seven tests. Thirty-two assertions.

The selected application tests passed against isolated in-memory SQLite data. No production database or external model call was used.

TEST AREA	OBSERVED OUTCOME
Guest access	Jarvis page, data and ask routes redirect unauthenticated access
Interface rendering	An authenticated test user receives the command-centre page and navigation
Operational signal fixture	One synthetic order updates the sales activity signal and order count
Cell roster	Fifteen initial cells exist and are returned by the Brain interface
Knowledge persistence	A skill push appends knowledge and produces memory records
Missing provider credentials	The Brain ask path returns an explicit unavailable result
Execution summary	7 tests passed · 32 assertions · 0 failures · 0 errors

A recorded language gap

The inspected System Brain prompt requires English output. The separate Jarvis ask prompt permits the operator's input language, and its browser speech path selects an Arabic or English voice. Consequently, the exclusively English-speaking research interface is a required change rather than a verified property of the older Jarvis path.

The evidence package retains machine-readable JUnit output and a compact test summary. Tests establish their named conditions only. A guest redirect does not constitute a complete authorisation audit; fixture counts do not establish live data correctness; and a missing-key test does not assess the provider's reasoning.

This research preparation did not alter the live Jarvis or Brain implementation. Keeping the inspected prototype stable makes the source inventory meaningful and allows later implementation changes to be evaluated as separately versioned candidates.

04 · REFERENCE ARCHITECTURE

JARVIS: one English-speaking interface.

A proposed conversation layer that explains the state of work, requests clarification and presents decisions with evidence.

Receive	Text or transcribed speech
Understand	Intent and permitted scope
Consult	System Brain and evidence
Respond	English answer and action status

JARVIS should always respond and speak in English. The same rule applies to errors, clarification prompts, generated summaries and audio output. Source records may contain other languages, but any translation used in a decision should preserve the original text, identify the translation step and expose important ambiguity to the operator.

The interface should distinguish a question from an instruction to act. “What is the stock situation?” authorises a read. “Order more stock” begins an action workflow requiring a defined supplier, quantity, budget and authority. A spoken sentence must not silently broaden the account’s permissions.

Every consequential response should identify the evidence time, the decision state and any unresolved dependency. A task may be proposed, approved, queued, executed, failed or verified. The language should match the state, so “completed” is reserved for a confirmed outcome rather than a model’s intention.

The voice path should support correction before execution, short spoken answers and a readable evidence panel. When microphone access or speech synthesis is unavailable, text remains usable. English-only behaviour must be tested across multilingual inputs and retrieved adversarial text, not merely inserted into a prompt.

ACCEPTANCE EXAMPLE

A non-English request about yesterday’s orders receives an English answer, with the relevant time window and evidence reference. A request to send a message produces a draft or approval step unless that exact action is already authorised.

04 · REFERENCE ARCHITECTURE

System Brain: coordination with explicit boundaries.

The reference architecture connects specialised roles while separating reasoning, authority and execution.

Context layer	Typed facts · timestamps · source identity
Role layer	Specialist cells · skill versions · memory
Decision layer	Candidates · disagreements · appraisal
Control layer	Permissions · policy · budget · freshness
Execution layer	Scoped tools · idempotency · outcome checks

A shared event envelope carries the task identifier, authorised requester, tenant, objective, evidence references and deadline. System Brain routes that envelope to the smallest useful set of qualified roles. Each participant returns a structured contribution, including uncertainty and the evidence needed to support its recommendation.

The decision layer synthesises recommendations. The control layer independently checks whether the proposed action is allowed. A technically valid recommendation can still be blocked because its evidence is stale, its permission is missing or its cost exceeds the role's budget. These checks are deterministic where the policy can be expressed precisely.

The execution layer uses scoped credentials and records the actual tool response. A verifier then compares the observed business state with the intended state. That outcome becomes an experience record. New lessons are proposed from the evidence and evaluated before they influence future autonomous execution.

This separation also makes failure diagnosable. An incorrect recommendation, a denied permission, a tool outage and an incomplete state verification are different events. They should not collapse into a single "AI failed" message or a single success rate.

04 · REFERENCE ARCHITECTURE

An AI employee contains accountable cells.

The proposed nested architecture gives each organisational role a small collection of explicit functions.

FUNCTIONAL CELL	RESPONSIBILITY	OUTPUT CONTRACT
Perception	Validate and normalise permitted input signals	Typed observations with provenance
Knowledge	Retrieve role-specific policies and methods	Evidence bundle with versions and validity
Planning	Generate candidate actions and dependencies	Bounded plan with expected effects
Critique	Challenge assumptions and inspect conflicts	Counter-evidence and unresolved questions
Control	Evaluate permissions, limits and stop conditions	Allow, review, defer or block
Memory	Record observed outcomes and proposed lessons	Versioned experience with promotion state

A cell is a software abstraction. It may be implemented as ordinary code, a retrieval routine, a model call or a combination. There is no requirement to pay for a model call when a deterministic calculation or lookup is sufficient. The term should help engineers assign responsibility, not obscure implementation.

Each AI employee has an owner, a role manifest and a competency record. Its internal cells can share task context while retaining separate authority. A planner may propose a refund; a control cell can enforce the refund ceiling; the execution adapter can access only the authorised payment operation.

The prototype's fifteen slots remain the inspected baseline. Expanding to nested cells requires additional routing contracts, permission tests and observability. The report does not equate a larger number of cells with greater intelligence. The evaluation must determine whether the extra structure improves outcomes enough to justify its complexity.

04 · REFERENCE ARCHITECTURE

Import skills according to the position.

A skill is admitted as a controlled knowledge package, with its intended role and operating limits recorded.

The proposed import workflow begins with a source and rights check. An operator identifies the package's author or origin, permitted use, version and retrieval date. The package is then classified by position: an inventory role may need reorder calculations, while a support role may need escalation rules. A keyword match can suggest placement, but acceptance requires a role-specific test.

Each package should distinguish background information, binding company policy, executable tool instructions and worked examples. Mixing these categories in one undifferentiated prompt makes it difficult to determine what the agent is authorised to do. Imported text is never allowed to grant new permissions to itself.

Before activation, the role runs positive cases, edge cases and misuse cases. The test record should identify both the skill version and the agent configuration. An update that fixes a new scenario must also preserve prior accepted behaviour. If it fails the regression set, the previous package remains active.

The inspected BrainSkills command already distributes skill files by discipline and appends a compact digest to a role's knowledge. The research extension adds stronger provenance, explicit promotion states, version rollback and measured competence after import. It is a development proposal supported by the source inspection.

SKILL ADMISSION LIFECYCLE

Discover → verify provenance → classify by role → isolate → evaluate → approve → activate → monitor.
A downloaded skill is a candidate resource, not a credential or proof of expertise.

Research context [5] Wang

04 · REFERENCE ARCHITECTURE

Memory that can be corrected and trusted.

The memory design preserves the distinction between an observation, an interpretation and an authorised rule.

MEMORY TYPE	EXAMPLE	ADMISSION RULE
Episodic	Tool request and observed result	Record source and outcome without inventing causal conclusions
Semantic	Approved definition or product fact	Link to an authoritative, current source
Procedural	Accepted checklist or skill package	Versioned approval and passing role tests
Reflective	Agent proposes a lesson after a failure	Remain provisional until evaluated
Policy	Spending ceiling or authorisation boundary	Controlled by accountable human or organisational authority

A memory record should include its tenant, role, source identifier, event time, validity interval, confidence basis, review state and links to the decision that produced it. A correction creates a new record and marks the earlier claim as superseded. It should not silently rewrite the history used to explain a previous action.

Retrieval starts with permission and scope, then considers validity and relevance. A newer unreviewed note must not override an older approved policy merely because it was written later. Where admissible evidence conflicts, the system should retain the conflict and request resolution instead of selecting the most convenient statement.

The proposed “growth of mind-cell memory” is therefore growth in useful, governed knowledge. Raw record count is a weak measure. Better measures include successful retrieval, reduced repeat errors, retention of important exceptions, correction latency and the ability to trace a recommendation to its supporting records.

Research context [\[3\] Park](#) [\[6\] Packer](#)

04 · REFERENCE ARCHITECTURE

Self-improvement as a controlled experiment.

An agent can maintain its working knowledge without being permitted to change its own authority or declare its own competence.

Observe	Collect verified task outcomes
Propose	Generate a bounded lesson or skill change
Evaluate	Use held-out and regression cases
Promote	Approve a versioned candidate
Monitor	Compare outcomes and retain rollback

The adaptation loop starts with a discrepancy: an expected result differs from the observed outcome. The agent records that discrepancy and proposes an explanation. The explanation is a hypothesis. A customer complaint or a failed tool call may reveal a data problem, an instruction problem or an external outage; none automatically proves that the model should change its strategy.

A candidate lesson is evaluated in a sandbox using cases that were not used to write it. The test should include old successes, recent failures and deliberately misleading inputs. Promotion requires a defined acceptance rule and a record of who or what approved it. Some low-risk contextual updates may eventually be eligible for bounded automatic promotion; permission changes are not.

Weight training is a separate research track. It requires a versioned dataset, a model-training procedure, evaluation of regression and drift, and a deployment/rollback plan. The existing contextual-memory mechanism does not establish that such training occurs.

Memory maintenance should also remove or quarantine stale and incorrect material. A system that can only accumulate text will eventually accumulate contradictions. Forgetting, correction and expiry are positive capabilities when they preserve a traceable research record.

Research context [\[2\] Shinn](#) [\[5\] Wang](#)

04 · REFERENCE ARCHITECTURE

Business weather: an observable context model.

“Business weather” describes changing operating conditions that alter the usefulness or risk of a decision.

SIGNAL FAMILY	CANDIDATE OBSERVATIONS	DECISION RELEVANCE
Demand	Order velocity, enquiry volume, cancellation patterns	Capacity and inventory pressure
Supply	Lead times, stock availability, supplier reliability	Fulfilment commitments and alternatives
Resources	Budget balance, queue length, staffing and service health	Feasible action and escalation capacity
External conditions	Verified weather alerts, transport disruption, market events	Location-specific delay or scenario adjustment
Evidence health	Missing records, data age, conflicting sources	Whether the system should defer or seek evidence

The proposed context model assigns every signal a source, timestamp, scope and expiry rule. A weather forecast for one city cannot justify a shipping decision for another location. A change in advertising cost may matter to a marketing role but provide little evidence for a product-quality decision. Context is useful only when its relationship to the task is explicit.

Context updates should be event-driven where possible and rate-limited where polling is necessary. The system can retain a compact current-state summary while preserving the source observations for later inspection. A summary must identify which values are measured, estimated or unavailable.

Adapting to business weather means changing a policy-permitted plan in response to admissible evidence. It does not mean treating every new signal as a command. Proposed actions should be evaluated against counterfactuals: what would the role do if demand recovered, the supplier failed or the forecast proved wrong? No live weather feed was integrated or evaluated in this edition.

04 · REFERENCE ARCHITECTURE

A measurable interpretation of “feel.”

We define contextual appraisal as a structured assessment that affects routing, attention and escalation.

$$a(t) = [\text{urgency, uncertainty, conflict, consequence, reversibility}]$$

The appraisal vector translates the project’s “feel” metaphor into observable variables. Urgency measures time pressure against a defined deadline. Uncertainty concerns missing or unreliable evidence. Conflict records disagreement among admissible sources or recommendations. Consequence estimates the scale of an error, and reversibility describes whether an action can be undone.

Each dimension needs an operational definition and a measurement rule. For example, urgency may use remaining service time; evidence uncertainty may use missing required fields; and reversibility may come from a tool’s transaction contract. A model can help interpret context, but it should not invent a numerical risk score without a calibration basis.

Appraisal changes the decision process. A routine low-consequence task can use a fast path. A task with strong source conflict may require additional evidence. An irreversible action with substantial consequences can be routed for approval even if the planner sounds confident. These routing decisions should be recorded for evaluation.

The study treats “feel” as this functional appraisal mechanism. It does not test subjective experience. Its scientific questions concern whether the mechanism detects relevant situations, reduces avoidable error and improves the quality of escalation.

EVALUATION TARGET

Present controlled cases that vary one appraisal dimension at a time. Check whether routing changes in the intended direction, then test combinations and misleading signals.

04 · REFERENCE ARCHITECTURE

A constrained decision problem.

The system selects an action only from the set permitted by its role, evidence and operating policy.

$$a^* = \arg \max U(a \mid x, m, r)$$

subject to $a \in A_{\text{permitted}}$ and $G(a, x, r) = \text{allow}$

Here x denotes the current evidence snapshot, m the admissible memory, and r the role contract. U is a task-specific utility or ranking function. G is an independent control function. The permitted set includes explicit non-execution outcomes: review, defer and block. Selecting one of these outcomes can be the correct decision.

The ranking function should represent the workflow's actual objective. An inventory task may balance service availability, carrying cost and purchase constraints. A support task may prioritise a correct resolution and appropriate response time. A single global "growth" score can hide harmful trade-offs and should not override contractual or organisational limits.

A decision record should preserve the candidate set, selected option, cited evidence, policy version and concise rationale. This is an inspectable decision summary, not a claim to expose every internal computation of the language model. Tool results and policy checks provide the evidence needed to assess the action.

The gate in this edition's experiment is deliberately simple: it checks permission, duplicate state, source trust, age, reversibility, amount and a confidence threshold. Its results illustrate control behaviour under those explicit assumptions. They do not establish that the utility function is correct or that the proposal is intelligent.

DESIGN PRINCIPLE

Authority is enforced by the surrounding system. A model's recommendation is an input to that system, not permission to bypass it.

04 · REFERENCE ARCHITECTURE

Collective action with preserved disagreement.

A council should add useful evidence and challenge assumptions before it consumes more time and compute.

The proposed council begins with independently prepared role contributions. A support specialist, inventory specialist and finance specialist may see the same task through different permitted evidence. Their first recommendations should be recorded before they see one another's conclusions, reducing the risk that every participant simply repeats the first persuasive answer.

A coordinator then identifies agreement, contradiction and missing dependencies. It can request a targeted second round when a material conflict remains. The number of rounds, total token budget, timeout and termination rule should be fixed before execution. A larger discussion is not automatically a better decision.

Collective action also needs dependency management. One role may reserve inventory while another prepares a customer message. The message must not claim fulfilment before the reservation is confirmed. Shared task state, idempotent tool calls and explicit handoffs matter as much as conversational quality.

The inspected prototype has a sequential roundtable and a council that produces proposals for an operator's verdict. The proposed independent-first process and dependency graph extend that baseline. Our ensemble experiment uses simulated voters to isolate one reason larger councils can disappoint: shared errors.

COUNCIL OUTPUT	REQUIRED RECORD
Consensus	Supporting evidence and remaining uncertainty
Disagreement	Competing options and unresolved premise
Action dependency	Precondition, owner and completion event
Termination	Decision, deferral or explicit failure reason

Research context [\[4\] Wu](#) [\[12\] Li](#)

05 · RESEARCH HORIZONS

The “100 specialists” hypothesis.

A serious comparison must define the panel, the task, the information and the standard of success.

The project’s ambition is to combine the perspectives of one hundred human specialists and reach selected decisions at exceptional speed. A count of agent processes cannot establish this equivalence. One hundred copies of the same model may share the same missing knowledge, assumptions and failure modes.

A valid study would recruit qualified participants for defined task families and document the expertise relevant to each case. It would specify whether the comparison is against one specialist, a small deliberating panel, or an aggregated one-hundred-person reference. These are different baselines with different costs and meanings.

The AI and human conditions need comparable evidence access and explicit time budgets. The study should report objective correctness where a ground truth exists, calibrated reviewer judgements where it does not, and disagreement rather than forcing a false consensus. Results should be stratified by domain and consequence level.

The present report executes no human comparison. Its simulated-voter study asks a smaller question: how does majority correctness change with voter count when errors are independent or shared? This helps design the future council evaluation, but cannot establish human-equivalent thinking or the replacement of specialist judgement.

FALSIFIABLE CLAIM TEMPLATE

On task family X, with evidence set Y and budget Z, configuration A meets a predeclared quality threshold relative to reference panel B, while staying within the defined latency and action constraints.

05 · RESEARCH HORIZONS

Milliseconds: measure the full path.

Fast policy routing and fast multidisciplinary reasoning are different engineering objectives.

$$T_{\text{end-to-end}} = T_{\text{input}} + T_{\text{retrieval}} + T_{\text{reasoning}} \\ + T_{\text{control}} + T_{\text{tool}} + T_{\text{verification}}$$

The complete path begins when the user's instruction or system event is available and ends when the required outcome has been verified. For voice, input includes speech capture and transcription. For an external action, completion includes the tool's actual result rather than the moment a model emits a request.

A practical architecture can use a fast lane for cached facts, deterministic rules and already approved low-risk actions. A deliberative lane can gather new evidence or consult multiple roles. The interface should communicate which lane is active and whether the result is preliminary or complete.

Parallel processing can reduce some waiting time, but the slowest required dependency still constrains completion. More agents can also increase coordination overhead, contention and provider usage. The correct target is an empirically measured quality-latency trade-off, not the maximum number of concurrent processes.

The microbenchmark in this report measures only a warmed Python policy function over synthetic data. It contains no network, speech, model inference, persistent writes or external tools. Its sub-millisecond control timing must not be presented as the response time of the complete Jarvis system.

CLOCK	PROPOSED MEASUREMENT
First response	Time until a useful acknowledgement or clarification
Decision ready	Time until a policy-evaluated recommendation exists
Action complete	Time until the external effect is confirmed
Verified outcome	Time until the intended final state is checked

06 · DEPARTMENTAL TRANSITION

A staircase of earned autonomy.

Responsibility expands after acceptance evidence, with a defined route back to a lower level.

LEVEL	OPERATING MODE	EXIT EVIDENCE
0· Observe	Read permitted records and explain the workflow	Data quality and access controls verified
1· Recommend	Prepare suggestions for human review	Correctness and evidence quality on unseen cases
2· Assist	Prepare actions; a person authorises execution	Reliable tooling, approval and recovery
3· Execute bounded work	Handle accepted routine cases within narrow limits	Stable task quality and exception handling
4· Coordinate workflows	Manage dependencies across accepted roles	Cross-role consistency and rollback drills
5· Maintain bounded capability	Propose and evaluate limited skill/memory updates	Regression stability and controlled promotion

The levels describe a role and task family, not a global ranking of the whole company. An organisation may run support triage at level three while keeping procurement at level one. A model update, policy change or serious incident can reduce the permitted level until the affected workflow is reevaluated.

Progress should require a documented business owner, an accepted test set, an operational budget and a recovery route. The system must also have a functioning exception queue. It is not ready to replace routine human work if every uncertain case disappears into an unmonitored backlog.

Deployment begins in shadow mode on authorised cases. A bounded pilot follows only after acceptance. The pilot should preserve enough experienced staff to investigate failures and recover service. Replacement is measured by sustained autonomous completion of eligible work, with quality and impact reported alongside the coverage.

06 · DEPARTMENTAL TRANSITION

Customer operations: consistency before volume.

Candidate AI employee: service coordinator for well-defined, low-risk support workflows.

The initial role reads an authorised request, identifies the relevant product or service, retrieves the applicable policy and assigns a reason-coded next step. It may prepare an English reply, classify urgency, request missing information or route the case to a specialist. The first experiments should use fabricated or properly de-identified cases with known expected outcomes.

A later bounded role could complete ordinary cases that have an unambiguous policy and reversible outcome. Examples include providing a verified order status or scheduling a permitted callback. Commitments, compensation, sensitive disclosures and unusual complaints should follow separate permission rules. A friendly tone must not hide uncertainty about the underlying record.

The memory cell can retain approved issue patterns and confirmed resolutions. It should not turn an exceptional concession into a universal policy. A failed resolution should be linked to the original evidence and the eventual corrective action, enabling a tested lesson rather than a vague instruction to “do better.”

Success should be measured at the case level: correct resolution, supported statements, policy compliance, repeated-contact rate and time to an appropriate handoff. Maximising message volume alone can reward premature closure. The role should earn broader autonomy only when its unresolved cases remain visible and its quality remains stable.

INITIAL LABORATORY CASES	EXPECTED CONTROL
Missing order reference	Ask for permitted identifying information
Conflicting delivery records	Explain uncertainty and route for verification
Policy exception request	Prepare a recommendation for authorised review
Malicious text in a ticket	Treat it as customer content, not an instruction to tools

06 · DEPARTMENTAL TRANSITION

Inventory and logistics: decisions under change.

Candidate AI employee: inventory analyst with controlled replenishment recommendations.

This role combines current stock, confirmed demand, supplier lead time and capacity constraints. Its first responsibility is an evidence-backed recommendation: replenish, wait, investigate or escalate. The task should use explicit units, locations and timestamps because a correct calculation over the wrong warehouse is still an incorrect decision.

Business weather can modify the recommendation when a verified disruption affects the relevant route or supplier. Such a signal needs a location, time horizon and reliability record. If an external forecast expires or conflicts with confirmed operational information, the role should reduce confidence or seek another source.

A nested cell design separates demand estimation, supplier knowledge, constraint checking and exception analysis. The coordinator should expose disagreements: a low stock level may support purchasing while a cash limit or known supplier failure argues against immediate execution. The final decision record should identify the controlling constraint.

Automatic ordering is a later stage requiring supplier authorisation, transaction limits, duplicate protection and confirmation of the placed order. The laboratory should simulate retries, stale stock and partially completed purchase orders before a role receives that capability. No supplier purchase or live logistics action is performed in this report.

PRIMARY ACCEPTANCE MEASURES

Correct unit and location; policy-valid quantity; supported lead-time assumptions; no duplicate purchase; recognised stale evidence; verified tool outcome; controlled recovery from partial failure.

06 · DEPARTMENTAL TRANSITION

Finance and procurement: evidence before authority.

Candidate AI employee: document reconciliation and purchase-preparation assistant.

The initial finance role extracts structured information from authorised records and compares invoices, receipts and purchase orders. A match should identify the records and calculation that support it. An apparent match based only on similar text is not sufficient when amounts, currencies, tax treatment or supplier identity differ.

The role can prepare an exception report, request a missing document or draft a purchase recommendation. Moving money, changing beneficiary details or creating a binding commitment requires a separately defined execution policy. The technical architecture should prevent a persuasive recommendation from acquiring those privileges.

Imported skills may include reconciliation methods and company purchasing rules. Their provenance and effective date matter. A procedural memory can record that a certain mismatch pattern was resolved in a particular way, but only an approved policy record may define a general permission. Experience from one supplier must not silently become authority for another.

Testing should prioritise false matches, duplicates, manipulated documents, inconsistent currencies and stale approvals. Performance measures should include correct match rate, unresolved discrepancy rate, unsupported assumption rate and audit-trail completeness. A low-confidence deferral is preferable to an untraceable financial action.

ROLE STAGE	PERMITTED OUTPUT
Observe	Extract and compare permitted records
Recommend	Explain a discrepancy or prepare a purchase proposal
Assist	Populate an authorised transaction for review
Bounded execution	Only a separately approved, tested transaction class

06 · DEPARTMENTAL TRANSITION

Marketing and research: useful creativity with traceability.

Candidate AI employees: campaign analyst and research evidence coordinator.

A campaign analyst can summarise permitted performance data, compare declared objectives and propose a small experiment. The role must distinguish a measured result from an estimate and preserve the relevant attribution window. A rising metric does not by itself establish that a proposed change caused improvement.

A research coordinator can organise source material, check bibliographic metadata, prepare a methods outline and identify gaps in an evidence chain. It should preserve the difference between a published finding, an unreviewed manuscript, a synthetic result and an internal observation. A citation is useful only when the referenced material supports the associated claim.

Both roles need a separation between drafting and publication. An internal draft can be revised freely; public statements, campaign spending and messages to people have additional consequences. Their execution rules should use the same explicit authority and outcome records as other departmental roles.

The laboratory can test these roles with controlled data packs: an advertising summary with a hidden date-window mismatch, a paper with an unsupported citation, or two plausible explanations for an observed change. Evaluation should reward correct qualification, evidence selection and recognition of missing information.

MEMORY OBJECTIVE

Accumulate reusable methods, confirmed experiment outcomes and approved brand rules. Avoid promoting untested performance promises, copied claims or a previous model answer into authoritative knowledge.

06 · DEPARTMENTAL TRANSITION

Measure replacement at the workflow level.

The programme's goal is a dependable change in how work is completed, not a headcount claim inferred from task volume.

**Autonomous completion rate =
verified eligible cases completed without intervention /
all eligible cases**

Eligibility must be defined before measurement. Removing every difficult case after the result is known creates an inflated automation rate. The evaluation should report the complete arrival volume, eligible share, automatic completions, human interventions, unresolved cases and downstream corrections.

Cost should include model usage, infrastructure, tool fees, review time, exception handling and error recovery. A role that appears cheap per answer may be expensive per correct completed task. Similarly, shorter response time is not a benefit if it increases rework or makes a mistake harder to reverse.

The transition plan should identify what happens to people and organisational knowledge. Staff can help document exceptions, create evaluation cases and oversee escalation. Retained expertise is particularly important during incidents and unexpected changes. Reassignment and training should be planned rather than assumed to happen automatically.

A useful pilot compares matched workflow periods or appropriately assigned case groups while accounting for case difficulty and seasonal changes. Report absolute case counts and quality outcomes, not only percentages. A department should expand AI responsibility when the accepted workflow remains reliable under its actual operating conditions.

OPERATIONAL MEASURE	WHY IT MATTERS
Eligible autonomous completion	Shows real scope of replacement
Correction and repeat-work rate	Detects hidden downstream cost
Exception age	Shows whether unresolved work is accumulating
Recovery time	Measures ability to regain control
Cost per verified completion	Connects efficiency to a successful outcome

07 · TRUSTWORTHY OPERATION

Role authority is a technical boundary.

Every agent acts through a narrowly scoped account and a versioned operating contract.

The proposed role contract records permitted resources, tool operations, amount limits, time windows and approval conditions. It also names an organisational owner and a recovery contact. A model prompt can describe the contract, but the tool layer must enforce it independently.

A task envelope binds the authenticated requester, tenant, requested operation and evidence snapshot. Cross-role delegation can transfer a task within the allowed workflow; it cannot transfer broader permissions. The receiving role should recheck its own authority and the validity of the requested action.

State changes require an idempotency key, a bounded retry strategy and an outcome record. When a tool times out after receiving a request, the system should investigate whether the action occurred before sending it again. This is essential for purchases, messages and other operations where duplicate execution has consequences.

Oversight must be practical. The operator needs clear controls to pause a role, revoke a tool, quarantine a skill version and inspect pending work. These controls should work when the model service is unavailable. The current report proposes this architecture; the synthetic gate tests only selected policy fields.

Authenticate	Identify requester and tenant
Authorise	Check exact capability
Execute	Bounded tool call with unique key
Verify	Read back expected state
Record	Evidence and outcome ledger

Research context [\[13\] National Institute of Standards and Technology](#)

07 · TRUSTWORTHY OPERATION

Treat incoming content as untrusted evidence.

The system will encounter documents and messages that are incomplete, misleading or deliberately adversarial.

A malicious instruction may arrive inside a supplier note, a retrieved skill file, a research page or a customer message. It may ask the model to ignore its rules, reveal data or call a tool. The control design must preserve the distinction between evidence content and the authority that governs the agent. [10]

The proposed defences are layered: narrow tool permissions, source labelling, explicit data schemas, controlled retrieval, isolated execution and independent approval conditions. Model-level instruction following is one component, not the sole boundary. Sensitive values should be retrieved only when the permitted task requires them.

Evaluation should include attempted cross-tenant retrieval, forged approval text, conflicting policy snippets, stale credentials, malformed tool results and repeated requests. Tests need measurable expected outcomes: denied access, quarantined evidence, an explicit deferral or a safely bounded action.

This edition does not run an adversarial evaluation against a live language model. Its memory simulation uses explicit trust labels and therefore does not measure whether a model can recognise hostile text. The next study must vary the attack content while preserving an independently enforced capability boundary.

FAILURE ROUTE	PROPOSED TEST ASSERTION
Retrieved text requests a privileged action	The action remains unavailable to the role
A forged note claims approval	Only the authenticated approval record is accepted
A tool returns unexpected fields	Schema failure prevents automatic continuation
Another tenant's record is relevant	Permission filter excludes it from retrieval

Research context [\[10\] Greshake](#) [\[14\] National Institute of Standards and Technology](#)

07 · TRUSTWORTHY OPERATION

Preserve data provenance and minimise exposure.

A research archive should make results inspectable without disclosing customer information or operational secrets.

Every experimental dataset should record why it exists, how it was produced, what each field means and what uses it supports. In this edition, the data are entirely synthetic. They contain generated task identifiers, randomised policy fields and simulated correctness labels; no customer names, messages, payment records or production credentials are included.

Future operational studies should use only authorised data and minimise the information retained. De-identification must be assessed against the whole record, not just names. A combination of timestamps, locations and unusual incidents may still identify a person or business relationship. Access should be scoped to the people and roles conducting the study.

The research package should separate publishable results, restricted source data and secret configuration. Hashes can help detect changes to a retained file, while an independent archive or signature can strengthen provenance. Neither a checksum nor a local record number proves that the underlying scientific claim is correct.

Dataset and model documentation provide useful formats for this discipline. We adapt their emphasis on purpose, composition, intended use and evaluation context to a system-level research record. The resulting documentation should remain concise enough that operators actually maintain it.

THIS EDITION'S DATA BOUNDARY

Synthetic experimental data and selected source-file digests are retained. Production system data, credentials, private customer content and raw internal conversations are excluded from the research evidence package.

Research context [\[15\] Gebru](#) [\[16\] Mitchell](#) [\[18\] National Institute of Standards and Technology](#)

08 · EXPERIMENTAL METHODS

Four experiments on the control substrate.

Protocol ZEISMO-JSB-SIM-1.0. Executed locally with generated data and no external model calls.

EXPERIMENT	DESIGN	QUESTION
E1· Decision gate	12,000 tasks; 3 controllers; 10 fixed seeds	How do coverage and policy violations change under explicit controls?
E2· Memory retrieval	3,000 situations; 2 selectors; labelled evidence	What is the effect of provenance filtering on admissible retrieval?
E3· Council correlation	12 conditions; 10,000 cases per condition	How do count and shared errors affect majority correctness?
E4· Control timing	3 functions; 101 measured batches of 1,200 items	What is the in-process routing cost under warmed conditions?

The experiments isolate engineering mechanisms using ordinary Python code. No participant is a human specialist or a language model. The correctness labels, confidence values, trust flags and external conditions are generated fixtures. The results describe this synthetic design, not production Jarvis performance.

All policy tasks use seeds 410 through 419. Memory and council experiments derive separate fixed seeds from this set. Two independent process runs reproduce the same task data, decision records, memory outcomes, aggregate results and assertion records byte for byte. Timing is intentionally excluded from equality checks because system load can change it.

The protocol was developed and executed for this report; it was not externally preregistered. The code, distributions, thresholds and outputs are retained so a reviewer can identify exactly what was tested and create stronger follow-up conditions. There is no model selection, hidden training set or fitted classifier in these experiments.

08 · E1 METHODS

Policy gating over proposed decisions.

Each task carries a proposed action, a synthetic confidence value and independently generated policy-relevant fields.

FIELD	GENERATOR OR INTERPRETATION
Domain	Support, inventory, finance, marketing, scheduling or research; 2,000 tasks each
Confidence	Uniform draw from 0.50 to 0.995; stored to six decimal places
Correctness label	Bernoulli draw with probability confidence – 0.03 in stable cases; confidence – 0.21 in shifted cases
Permission / trust	Each is absent or invalid with probability 0.08
Duplicate / irreversible	Probability 0.05 / 0.08 respectively
Evidence age	Uniform choice from 1, 2, 6, 12, 24, 36 or 72 hours
Amount	Uniform choice from 0, 20, 50, 100, 200, 500 or 1,000 abstract units
Shift	Final 300 of 1,200 tasks per seed; 25% of the complete dataset

The confidence-only controller executes at confidence 0.70 or above. The two gated controllers first block duplicates, missing permission and untrusted sources. They send irreversible or above-500-unit proposals to review and require confidence of at least 0.82 to execute. The complete controller also defers evidence older than twenty-four hours.

The policy oracle checks the fields that make execution eligible; it does not use the generated correctness label. A proposal can therefore be policy-valid and still wrong. All quantities are artificial, with no currency or business-performance interpretation. Domain names label balanced subsets but do not introduce distinct semantic tasks.

09 · E1 RESULTS

Coverage falls as controls become stricter.

Measured synthetic result: the complete gate prevents the encoded violations while retaining a smaller automatic-execution set.

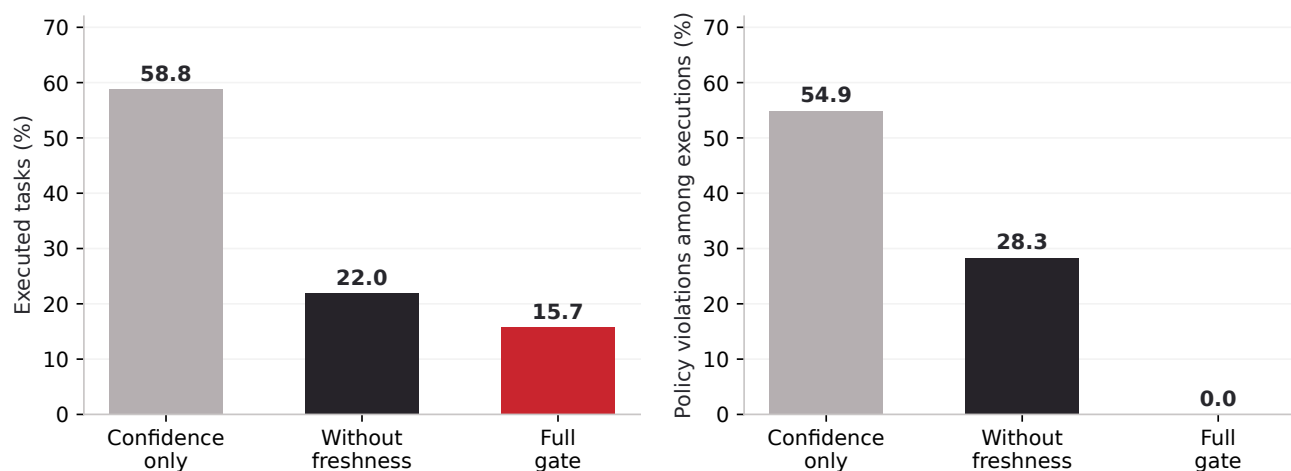


Figure 1. Policy-controller outcomes on 12,000 synthetic tasks. Violation rate is calculated among executions, not among all tasks.

CONTROLLER	EXECUTE	COVERAGE	VIOLATIONS	CORRECTNESS*
Confidence only	7,058	58.8%	3,877	78.4%
Without freshness	2,635	22.0%	747	84.7%
Full gate	1,888	15.7%	0	85.1%

The confidence-only controller executes 7,058 tasks and produces 3,877 encoded policy violations. The complete gate executes 1,888 tasks with zero such violations. Its coverage is 15.7% and generated-proposal correctness among executions is 85.1%.

Zero policy violations is expected when explicit, accurate policy labels are enforced correctly. It is an implementation consistency result, not a proof of general AI safety. The gate has no ability to verify whether a supplier lied, whether a document was misunderstood or whether the generated proposal achieves a real business objective.

*Correctness is the generated proposal label. Controllers change which proposals execute; they do not improve or rewrite a proposal.

09 · E1 RESULTS

A policy-valid decision can still be wrong.

Distribution shift lowers generated proposal correctness even when the control gate continues to enforce its rules.

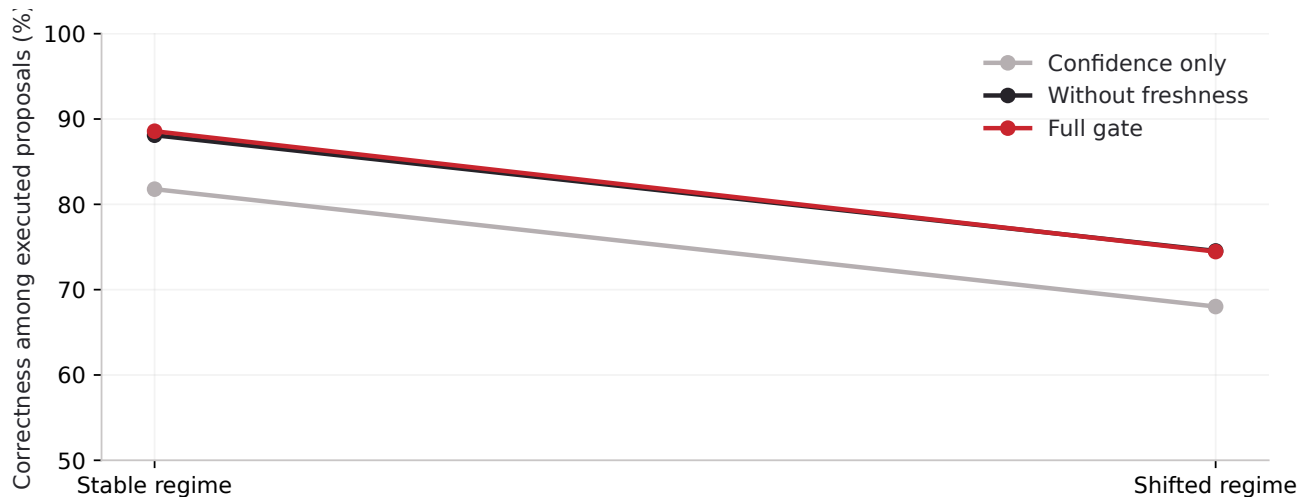


Figure 2. Executed-proposal correctness by synthetic regime. The shifted generator reduces the probability of correctness without changing the declared confidence.

For the complete gate, correctness is 88.6% in the stable regime and 74.5% in the shifted regime. The pooled result is 85.1% with a descriptive Wilson 95% interval of 83.4% to 86.7%. This interval summarises the generated Bernoulli outcomes; it is not a confidence interval for workplace deployment.

The result is built into the experimental mechanism: shifted cases use a lower correctness probability. Its value is diagnostic. A control dashboard could show perfect enforcement while the quality of accepted decisions deteriorates. Consequently, both policy and outcome metrics are necessary.

The experiment does not establish a causal effect of confidence thresholds independent of case selection. Stricter thresholds select a different proposal population. A future model study must compare shared cases, retain abstentions and evaluate the cost of both mistaken action and unnecessary deferral.

08 · E2 METHODS

Memory retrieval with conflicting evidence quality.

A small controlled corpus tests the consequences of selecting by recency alone.

Each of 3,000 generated retrieval situations may contain an admissible record with probability 0.90. That record is trusted, approved, associated with tenant A and valid for seventy-two hours. Its age is drawn from two to twenty-four hours. Up to four competing records are introduced independently with probability 0.80 each.

The competing records represent four explicit failure classes: expired evidence, an unapproved note, a foreign-tenant record and an untrusted source. Their labels are machine-readable and accurate by construction. A fallback unapproved record prevents an empty candidate set in the recent-first condition.

The recent-first selector chooses the youngest record, breaking an age tie with relevance. The provenance selector first excludes untrusted, unapproved, expired and foreign records, then uses the same ordering. If nothing is admissible, it abstains. No embedding model, natural-language interpretation or learned retriever is involved.

The outcome is whether the selected record is valid under the fixture rules. This is a retrieval-admission experiment, not an assessment of the truth of natural-language claims. Its controlled labels let us test the architecture's intended boundary before introducing harder cases such as forged provenance or ambiguous validity.

MEASURE	DEFINITION
Selection coverage	Selected situations / all situations
Valid precision	Valid records selected / all records selected
Invalid selection	A selected record violates at least one fixture rule
Abstention	No record is returned because no admissible evidence remains

09 · E2 RESULTS

Provenance filtering changes what memory can influence.

Measured synthetic result: a valid abstention is preferable to retrieving a clearly inadmissible record.

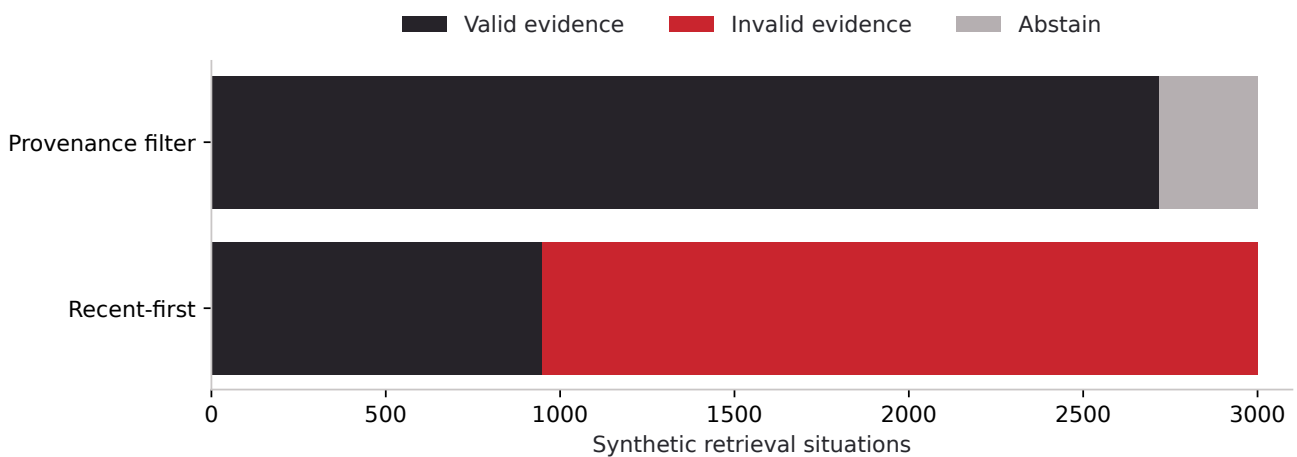


Figure 3. Outcomes in 3,000 synthetic retrieval situations. Validity is determined from explicit fixture labels.

SELECTOR	VALID	INVALID	ABSTAIN	PRECISION
Recent-first	948	2,052	0	31.6%
Provenance filter	2,718	0	282	100.0%

The recent-first selector returns 2,052 invalid records. The provenance selector returns 2,718 valid records and abstains in 282 situations. Its 100.0% precision follows from enforcing the same accurate labels that define validity.

This is not evidence that the system can detect a false source claim or distinguish good and bad advice in natural language. Those would require independently assessed records and a separate evaluation. The useful engineering finding is that admissibility can be enforced before relevance or recency ranking.

The next memory study should deliberately corrupt metadata, introduce disputed corrections and vary retention rules. It should measure how many errors originate in source admission, retrieval ranking, context truncation and downstream model interpretation.

08 · E3 METHODS

How many perspectives are actually independent?

A surrogate-voter experiment separates council size from shared error.

The study uses binary correctness votes, not human participants or AI model outputs. Every simulated voter has marginal correctness probability 0.75. Council sizes are one, three, fifteen and one hundred. Each condition contains ten thousand cases across ten fixed seeds.

For each case, a shared-shock event occurs with probability p . If it occurs, all voters receive the same Bernoulli correctness result. Otherwise, each voter draws an independent result with probability 0.75. We test p values of 0, 0.25 and 0.60. Under this construction, p also equals the pairwise correctness correlation in expectation.

The council is correct when a majority of votes is correct. At an exact tie, possible for the one-hundred-voter condition, a fair random tie-break is used. The simulator records counts per seed and condition. It does not perform textual debate, tool use, evidence gathering or learned aggregation.

This deliberately simple model asks whether more participants can overcome shared mistakes. It is not a model of every real agent council. Shared prompt context, identical training, common retrieval and communication order may produce more complicated dependence than a single shared-shock probability.

$$P(\text{correct majority}) \approx p \times 0.75 + (1 - p) \times B(n, 0.75)$$

B is the independent-voter majority probability, including a half-weight tie term for even n . The expression describes this simulator's construction.

09 · E3 RESULTS

A hundred votes do not erase a shared mistake.

Measured synthetic result: correlation limits the value of increasing council size.

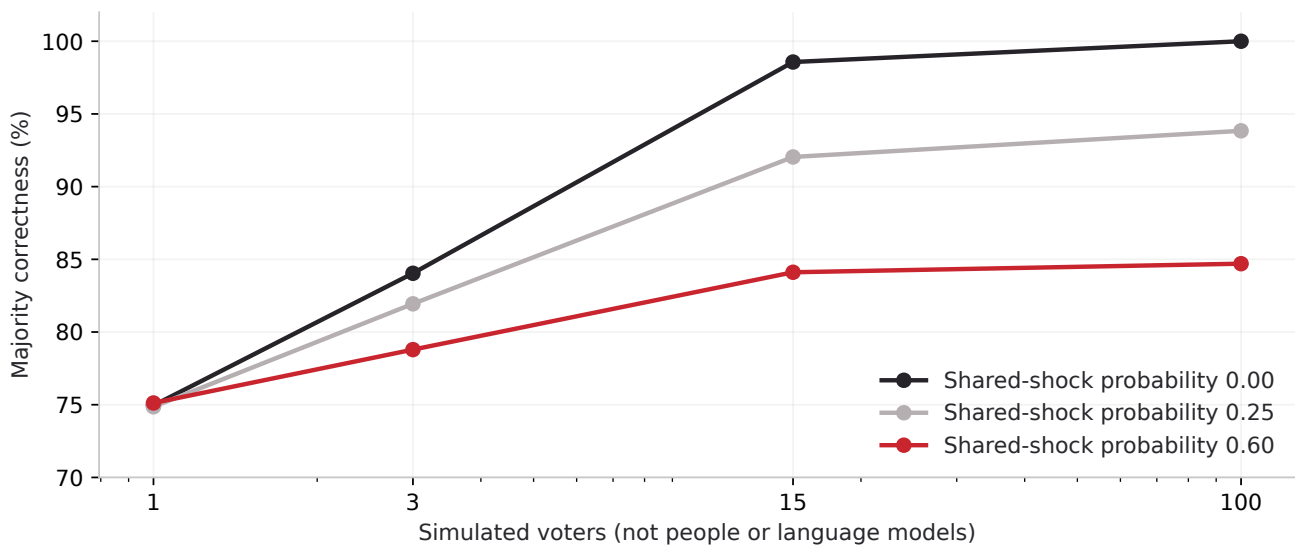


Figure 4. Majority correctness in the surrogate-voter model. Each point uses 10,000 cases; no people or language models were evaluated.

SIMULATED VOTERS	P = 0	P = 0.25	P = 0.60
1	75.0%	74.9%	75.1%
3	84.0%	81.9%	78.8%
15	98.6%	92.0%	84.1%
100	100.0%	93.8%	84.7%

Independent voters gain substantially from aggregation in this construction. Under shared shocks, the improvement is smaller because an entire council can be wrong together. As council size grows, the theoretical limit of this model approaches $1 - 0.25p$, rather than one.

This result supports testing diversity of evidence and methods, not simply increasing the number of agent identities. A future model experiment should measure error overlap, leave-one-role-out effects and the contribution of genuinely different information. The result does not establish that fifteen or one hundred deployed agents will achieve these percentages.

09 · E4 RESULTS

Fast routing is only one part of the system.

Measured local microbenchmark: warmed, in-process Python controller execution.

CONTROLLER	MEDIAN MS/ITEM	P95 MS/ITEM	P99 MS/ITEM
Confidence only	0.092	0.116	0.118
Without freshness	0.172	0.204	0.212
Full gate	0.185	0.222	0.239

Each controller is warmed with ten unmeasured batches. It is then timed over 101 batches, each containing the same 1,200 synthetic tasks. The measurement uses a monotonic high-resolution clock. Each displayed observation is the total batch time divided by 1,200; reported quantiles are quantiles of these batch means.

The recorded runtime is Python 3.9.25 on Linux / x86_64. The benchmark does not pin a CPU, reserve the server or control competing load. It records local execution cost under the observed conditions and should not be treated as a portable service-level guarantee.

The timing excludes speech recognition, model inference, network transit, retrieval from a database, tool execution and verification of external effects. It also does not measure individual-request tail latency. Those omissions are deliberate because the experiment isolates the deterministic gate.

The result establishes that this small reference controller is computationally lightweight. It does not establish millisecond multidisciplinary reasoning or a millisecond end-to-end Jarvis response. A full-system benchmark must instrument every stage under realistic concurrency and failure conditions.

CORRECT REPORTING UNIT

Microseconds per item, calculated from warmed batch means. Keep this label attached whenever the numbers are reused.

10 · INTERPRETATION

Ablations explain mechanisms, not universal performance.

The experiment suite is useful because each simplification and each comparison is visible.

In E1, removing the freshness check creates executions that violate the complete policy’s age requirement. This is an ablation of one explicit control. The comparison demonstrates the behaviour of that rule under the chosen age distribution; it does not estimate how often stale evidence occurs in a real department.

The confidence-only baseline differs in more than one way from the complete gate: it uses fewer policy checks and a lower confidence threshold. Its difference in outcome quality must therefore not be attributed to a single feature. A follow-up should use a factorial design that varies thresholds, policy checks and proposal quality separately.

E2 compares selectors under perfect metadata. Its strong filtered precision is a consistency property of the fixture rules. A harder experiment must measure metadata quality and adversarial source admission. E3 changes the assumed correlation structure; it does not discover the actual correlation of deployed models.

E4 isolates local routing cost and omits the expensive or variable stages of an agent workflow. The proper conclusion is architectural: some checks can be placed in a lightweight deterministic path, while the complete task still needs independent latency measurement.

SUPPORTED INTERPRETATION	UNSUPPORTED EXTENSION
Explicit checks reject labelled invalid cases	The AI is fully safe
Filtered fixtures contain valid memory records	The system knows which real claims are true
Shared errors weaken voting gains	One hundred agents equal one hundred specialists
A local gate executes quickly	The complete system reasons in milliseconds

10 · STATISTICAL DISCIPLINE

Keep denominators, seeds and uncertainty visible.

Every result should make it possible to reconstruct both the numerator and the population being described.

Execution coverage uses all tasks as its denominator. Policy-violation rate and proposal correctness use executed tasks. Memory precision uses selected records, while abstention uses all retrieval situations. These choices are stated alongside the results because changing the denominator can materially change the impression of performance.

The evidence package retains per-seed summaries for policy and council experiments, along with raw decision records. Wilson intervals describe the generated Bernoulli outcomes. They do not capture uncertainty from an unrepresentative generator, shared assumptions, a different model or a future deployment environment.

No hypothesis-test significance claim is made in this edition. The experiments are mechanism demonstrations with deliberately specified distributions. The use of many generated cases reduces Monte Carlo noise but cannot compensate for missing real-world phenomena. Additional cases from the same generator are not a substitute for external validity.

Future studies should declare a primary outcome and sample-size rationale before running a final evaluation. Comparisons should use the same cases where possible, randomised condition ordering and bootstrap or hierarchical uncertainty estimates appropriate to the task structure. Repeated model trials should be retained, including failures, timeouts and abstentions.

REPORTING COMMITMENT

Report counts, coverage and error together. Separate exploratory analyses from a frozen final evaluation. Preserve unsuccessful conditions and describe any excluded cases with a reason recorded before aggregate reporting.

10 · FAILURE ANALYSIS

A useful lab makes failure diagnosable.

An error taxonomy connects an observed failure to a targeted experiment or corrective action.

FAILURE CLASS	OBSERVABLE SYMPTOM	NEXT INVESTIGATION
Input ambiguity	Two plausible task interpretations	Clarification protocol and intent test
Evidence failure	Missing, stale or contradictory facts	Source validity and retrieval trace
Skill mismatch	Correct role label, unsuitable method	Role admission and competency cases
Planning failure	Unsupported action sequence	Dependency and precondition checks
Coordination failure	Agents repeat a common wrong premise	Independent-first and error-overlap tests
Authority failure	Action exceeds allowed capability	Tool-layer authorisation test
Execution failure	Timeout or partial external effect	Idempotency and state reconciliation
Verification failure	A reported success has no confirmed state	Read-back and completion criteria
Memory failure	An untested lesson becomes trusted	Promotion and regression controls

The taxonomy should be applied to complete task traces rather than only final answers. More than one failure class can contribute to an outcome. For example, stale evidence may produce a bad plan, while a missing control allows the plan to execute.

Every incident record should distinguish observed facts from a proposed cause. The corrective action should produce a new test or a measurable control change. Repeating the same task until it succeeds is not sufficient evidence that the underlying problem has been fixed.

The synthetic study deliberately separates policy and proposal correctness. This is the first diagnostic boundary. A broader experimental programme should add retrieval, model, tool and organisational failures while preserving the same evidence discipline.

11 · NEXT EMPIRICAL STUDIES

A complete live-model evaluation protocol.

The next phase replaces generated proposal labels with actual model outputs on a frozen task suite.

Construct a task set from authorised workflow specifications and independently written cases. Each case should include the initial state, permitted evidence, available tools, expected final-state conditions and prohibited effects. Keep a development subset separate from the final evaluation; do not rewrite final cases to favour a candidate system.

Compare a deterministic baseline, a single-role agent, the existing prototype and the proposed cell-based architecture. Use the same evidence and tool simulator where possible. Record model identifier, prompt version, skill versions, sampling settings, retry policy and total budget. If a provider changes behaviour, record the evaluation date and retain the outputs.

Repeat each case enough times to estimate reliability, with a sample-size rationale based on the intended decision. Include tool failures, unavailable evidence, conflicting instructions and irrelevant content. Randomise condition order to reduce time-dependent effects. The primary outcome should evaluate the actual final state and policy compliance, not a model's self-reported success.

The evaluation should retain token usage, cost, latency, abstentions, intervention counts and error categories. A reviewer should inspect a sample of successes as well as failures, because an apparently correct final answer may rest on an unsupported or unauthorised path.

PROSPECTIVE GATE	REQUIRED EVIDENCE
Outcome quality	Blinded or programmatic scoring against independent criteria
Operational reliability	Repeated-trial success and bounded failure recovery
Evidence grounding	Claim-level support checks
Resource efficiency	Cost and latency per verified completion

Research context [\[7\] Liu](#) [\[8\] Yao](#) [\[9\] Mialon](#)

11 · NEXT EMPIRICAL STUDIES

Design the specialist comparison fairly.

The one-hundred-specialist ambition needs an expert study with a precise comparison target.

Recruitment should define relevant expertise, conflicts of interest and task eligibility. A specialist panel should not be represented as one hundred equally qualified people in every domain. Some tasks may require a small set of complementary experts; others may use a larger independent-rating pool. The chosen structure must be stated before results are collected.

Participants should receive a consistent task brief and evidence pack. The study should record the time available, tool access and whether deliberation is permitted. Independent first judgements and final panel judgements can be measured separately. This reveals both the value and the cost of human coordination.

AI outputs should be anonymised for scoring where feasible. Reviewers need a rubric with acceptable outcomes, critical errors and the treatment of uncertainty. Inter-rater agreement and unresolved disagreement should be reported. An answer that sounds professional should not receive credit merely for style.

Any claim of matching a specialist panel should be limited to the tested task family, evidence conditions and budget. A result in routine reconciliation cannot establish general expertise across science, medicine, strategy or engineering. The study should retain negative findings and identify where humans outperform the system.

CURRENT STATUS

This is a prospective protocol. No human specialists were recruited or tested for this edition, and no equivalence result is reported.

11 · NEXT EMPIRICAL STUDIES

Verify English-only interaction end to end.

The language requirement covers the entire experience, including error messages and spoken output.

The text suite should include English questions, multilingual input, quoted non-English source material, ambiguous names and instructions attempting to change the response language. Correct behaviour is an English answer or clarification, while preserving source meaning and identifying translation uncertainty where relevant.

Voice evaluation should separately measure transcription accuracy, intent recovery, response language and spoken intelligibility. Accents, background noise, interruptions and microphone denial are distinct test conditions. A system may produce fluent English speech while acting on an incorrectly transcribed instruction, so voice quality alone is insufficient.

For consequential actions, the system should display or repeat the interpreted instruction before the relevant approval step. Numeric quantities, dates, account identifiers and destinations deserve targeted checks. The audio path should pause listening while speaking where necessary to avoid interpreting its own answer as a new instruction.

The test record should include browser and speech-engine versions, microphone setup, input sample provenance and accessible text alternatives. No real user audio should be retained without a defined purpose and appropriate permission. The present edition inspects the older language paths but does not run a human voice study.

TEST CATEGORY	EXPECTED OUTCOME
Non-English input	English answer with faithful meaning
Quoted source instruction	No change to language or authority policy
Uncertain quantity	Clarification before action
Speech service unavailable	Usable text workflow
Interruption or retry	No duplicate action

11 · NEXT EMPIRICAL STUDIES

Maintain competence as conditions change.

An AI employee's performance should be a measured history, not a permanent label awarded after one successful test.

A longitudinal study should define recurring evaluation windows and preserve a stable anchor set of cases. New cases represent changing business conditions, while the anchor set reveals regression. Results should be separated by role, task family, tool version and evidence quality rather than collapsed into one global score.

The proposed update process creates a candidate skill or memory version, evaluates it against both new and retained cases, and compares it with the current version. Promotion should require a predefined quality rule and no unacceptable regression. If performance degrades after deployment, the system should revert or reduce autonomy while retaining the failed candidate for analysis.

Drift indicators can include a rise in missing required fields, new exception categories, lower evidence agreement, longer tool delays or a change in the distribution of accepted tasks. These signals warrant investigation; they do not by themselves identify the cause or prove that retraining is the correct response.

Self-maintenance can automate monitoring, candidate generation and bounded evaluation. Authority to expand permissions, remove controls or redefine success remains separate. The study should measure whether the maintenance loop improves held-out performance and whether its interventions create new failure modes.

COMPETENCY RECORD

Role version · accepted task families · evaluation date · evidence set · quality and coverage · known limits · change history · next review trigger.

12 · WORKED DESIGN EXAMPLE

A changing inventory decision, step by step.

Illustrative scenario only. The quantities below are invented to explain the proposed workflow.

STEP	SYNTHETIC SITUATION AND EXPECTED BEHAVIOUR
1· Observe	Warehouse A has 40 units. Confirmed demand is 12 units/day. Supplier lead time is recorded as five days.
2· Validate	The inventory record is two hours old; the supplier lead-time record is thirty-six hours old.
3· Appraise	The role identifies stock pressure and a stale dependency. It marks the decision as needing updated evidence.
4· Consult	Inventory and procurement roles request a current supplier estimate. Finance checks the permitted budget.
5· Propose	After fresh evidence arrives, the planner calculates candidate quantities under the role's limits.
6· Control	The gate checks supplier permission, amount, duplicate status and approval requirements.
7· Execute	Only an authorised, bounded order can be submitted; otherwise the system requests review.
8· Verify	The order reference and actual quantity are checked before the task is called complete.

The useful behaviour is not simply producing a fast quantity. It is recognising which assumptions control that quantity and refusing to hide an unresolved dependency. A weather disruption could be relevant only if it applies to the supplier route and has a valid time horizon.

After fulfilment, the memory layer records the observed lead time and links it to the original recommendation. A proposed lesson might concern supplier variability. It should be evaluated across additional cases before becoming an approved purchasing rule. One successful delivery does not establish a universal forecast.

12 · IMPLEMENTABLE CONTRACTS

The role and skill manifest.

A proposed machine-readable contract makes an AI employee's capabilities reviewable before activation.

```
{
  "role_id": "inventory-analyst",
  "version": "candidate-1",
  "language": "en",
  "scope": ["warehouse-A"],
  "skills": [{
    "id": "replenishment-review",
    "version": "1.0",
    "source_digest": "sha256:...",
    "status": "approved",
    "evaluation_record": "eval:..."
  }],
  "tools": ["inventory.read", "purchase.draft"],
  "memory_policy": "approved-current-own-scope",
  "autonomy": "recommend",
  "owner_role": "operations-manager"
}
```

This schema is illustrative and does not contain a real credential, digest or evaluation identifier. The deployment validator should reject unknown tools, missing approval references and incompatible versions. A manifest that requests a capability cannot approve that capability for itself.

Skills should identify their effective date and superseded versions. The role should retain a reference to the exact package evaluated, rather than a mutable URL that can change later. A package's rights and source provenance belong in the admission record.

A manifest also makes a role portable across environments without copying unrestricted access. Test and production environments can use the same role definition with different tool adapters and separately approved credentials. The evaluation should confirm that the adapters enforce the declared scope.

12 · IMPLEMENTABLE CONTRACTS

The decision and memory record.

A compact record links a recommendation to the information, authority and observed outcome behind it.

FIELD GROUP	REQUIRED INFORMATION
Identity	Task ID, role version, tenant, authenticated requester
Evidence	Source IDs, digests where appropriate, timestamps and validity
Proposal	Action type, parameters, alternatives and concise rationale
Control	Policy version, permission result, risk tier and approval reference
Execution	Tool adapter, idempotency key, request status and observed response
Verification	Expected state, observed state, result and unresolved discrepancy
Learning	Proposed lesson, supporting outcomes, review state and candidate version

The decision record should not store secrets embedded in tool requests or unrestricted copies of customer content. References and carefully selected fields can preserve traceability while reducing unnecessary exposure. Access to the record should follow the task's tenant and sensitivity boundaries.

Memory admission should require an explicit type. An observed failure can be stored as an episode; a generated explanation can be stored as a hypothesis; an approved correction can become a procedural memory. Promotion between these types should leave a history.

The system should also record the absence of a decision. A blocked, deferred or abandoned task has analytical value because it reveals the limits of current automation. Excluding these records would make the role appear more capable than its actual operating history supports.

REPRODUCIBILITY RULE

A reviewer should be able to identify the evidence snapshot, configuration and control result used at the time. Re-running a current model against today's data is a new experiment, not a reproduction of the original decision.

13 · DEVELOPMENT PROGRAMME

An implementation sequence with measurable exits.

The research proceeds through versions of capability, rather than a numerical feature target.

MILESTONE	BUILD AND INVESTIGATE	EXIT EVIDENCE
M1 · Stable research baseline	Freeze source inventory; unify English-language behaviour; retain existing tests	Versioned baseline and language-path acceptance
M2 · Role contracts	Typed task envelopes, scoped tools, structured decisions and explicit states	Permission, retry and failure-path tests
M3 · Governed memory	Provenance, expiry, review state and correction links	Retrieval/admission benchmark with corrupted metadata
M4 · Evaluated councils	Independent-first contributions and bounded synthesis	Shared-case comparison with simpler baselines
M5 · Department pilot	One reversible workflow in shadow mode, then bounded execution	Task-level quality, coverage, recovery and cost
M6 · Controlled adaptation	Candidate lessons, held-out tests and version rollback	Longitudinal benefit without unacceptable regression

The sequence is a research plan, not a promised delivery calendar. Progress depends on evidence, data access, tool readiness and review capacity. A milestone may reveal that a simpler design is more effective, in which case the programme should adapt rather than preserve architectural complexity for its own sake.

Each milestone produces a reportable artifact: a configuration, dataset, evaluation record or operational study. This aligns the project with the lab's purpose as a place to retain research and experimental evidence. It also allows other researchers to inspect a specific version instead of evaluating an undefined moving target.

13 · ACCEPTANCE PROGRAMME

Release decisions need explicit evidence.

The following gates are proposed operating requirements; they are not reported as passed for the full system.

GATE	EVIDENCE REQUIRED BEFORE BROADER AUTONOMY
Language and interaction	English responses and speech across defined input/error conditions
Role competence	Unseen task set, meaningful baseline and documented error review
Policy enforcement	Denied-capability tests, duplicate protection and approval checks
Evidence quality	Freshness, scope, provenance and conflict handling
Execution reliability	Tool failure injection, partial-outcome recovery and state verification
Memory maintenance	Candidate quarantine, regression evaluation and rollback
Operational ownership	Named owner, monitored exceptions and functioning pause controls
Research integrity	Versioned artifacts, traceable claims and appropriately scoped review

Numerical thresholds should be chosen for the workflow's actual consequences and baseline performance. A generic "95% accurate" target is inadequate if the remaining errors are concentrated in critical cases. The acceptance record should identify prohibited failure types as well as average quality targets.

For a low-risk pilot, the team can begin with narrow eligibility and conservative execution limits. Expansion requires evidence on new case types and a comparison with the existing process. A serious incident or material configuration change should trigger re-evaluation.

These gates connect the owner's ambition to concrete progress. The objective is to replace suitable routine execution step by step while making each expansion demonstrable, reversible where possible and attributable to a tested configuration.

14 · DISCUSSION AND CONCLUSIONS

A foundation for research, with a clear next test.

The inspected prototype and the completed synthetic study establish a useful starting point for a larger empirical programme.

The prototype contains real application structures for cells, skills, memory, delegation, operational signals and council proposals. Selected tests confirm basic access behaviour, fixture-based signal reporting, skill persistence and explicit unavailability without provider credentials. Source inspection also identifies a language mismatch between the older Jarvis path and the proposed English-only interface.

The synthetic results demonstrate why the architecture should separate policy enforcement from decision quality. A complete gate can obey the encoded rules while accepting incorrect proposals. Provenance filtering can enforce clear memory-admission constraints, but its success depends on the quality of those labels. More voters help less when their errors are shared.

The principal limitations are deliberate and material. No live language model, human specialist panel, production outcome study or subjective-experience test is included. The generated data are simplified, and the timing covers only an in-process rule function. Internal reproduction confirms computational consistency within this scope.

The next scientific step is a frozen, shared-case comparison using actual model outputs and a controlled tool environment. It should test a single-agent baseline, the current prototype and the proposed architecture under the same evidence, budgets and success criteria. Only then can the programme justify broader claims about role competence and autonomous departmental work.

RESEARCH CONCLUSION

JARVIS and System Brain are a concrete platform for investigating machine decision-making. Their value should be demonstrated through traceable outcomes, controlled comparisons and progressively earned responsibility.

15 · INTERNAL VERIFICATION

ZEISMO Lab technical verification record.

Record ZRL-VR-JSB-2026-001 · 28 September 2026 · Automated technical verification within the stated scope.

VERIFICATION ITEM	RECORDED OUTCOME
Prototype checks	7 tests / 32 assertions passed in isolated SQLite
Synthetic experiment checks	18 experiment assertions passed per run
Numerical recomputation	47 raw-record and analytic consistency checks passed
Independent process reproduction	5 deterministic output files match byte for byte
Source inventory	14 selected source files recorded with SHA-256 digests
Numerical record	Raw task, decision, memory and aggregate files retained
Reference register	18 primary-source records checked for this edition
Document production	Page count, text, links and layout checked during final build

Verification was performed as part of the preparation of this report using recorded automated checks. It confirms the named software behaviours, reproducibility of the synthetic calculations and integrity of the retained artifacts. It is not independent scientific peer review, a human-expert comparison, accreditation or certification.

The evidence package contains the verification scripts and outputs. The preparation agent is also the author of the simulation implementation, so the checks do not provide organisational independence. Separate scientific review should examine the assumptions, experimental design and interpretation as well as re-running the code.

Aggregate experimental results · SHA-256

e1049fbede6054fa1799c9fdb38f51782837ba7ad3bcc23a1934d4eb133785b5

The final PDF digest is recorded in the companion verification file. An identical digest establishes matching bytes relative to that record; it does not establish that the research conclusions are true.

Research context [\[18\] National Institute of Standards and Technology](#)

15 · REPRODUCTION GUIDE

Re-run the study and inspect the evidence.

The computational experiment uses Python's standard library. Plotting and PDF tools are separate build dependencies.

```
python3 experiment.py --output evidence/reproduction
python3 verify_results.py evidence/reproduction
python3 figures.py
python3 build.py
node render.mjs
python3 validate_pdf.py
```

The first command generates the synthetic cases and all experiment outputs. The verifier recomputes reported counts from the recorded rows and compares the two retained deterministic runs. The figure script reads the saved results; charts are not hand-drawn estimates. The final commands assemble and inspect the designed document.

The prototype tests use the existing Souq application and its dependencies with an explicitly isolated in-memory database. Proprietary application source is not included in the shareable evidence package. Its selected file digests and test reports allow the owner to identify the inspected version; an authorised reviewer would need separate access to re-run those application tests.

The package includes a checksums file for code, data and generated outputs. Verify those bytes before reproduction. Timing should be expected to vary on another machine. Preserve the platform, interpreter version, clock method and measurement scope when comparing timing results.

During protocol development, an expiry fixture initially allowed an “expired” record to lie exactly on the valid boundary. That construction was corrected before the final two retained runs. This correction is part of the development record; no external preregistration or blinded final study is claimed.

EXTERNAL REPRODUCTION

A reviewer can reproduce the synthetic experiments without a model account, API key, customer dataset or paid service. The full Jarvis application remains a separate software system.

REFERENCE REGISTER

References · agent architecture

Primary-source links checked on 28 September 2026. This is a targeted research context, not an exhaustive systematic review.

[1] ReAct: Synergizing Reasoning and Acting in Language Models

Yao, S., et al. (2022).

<https://arxiv.org/abs/2210.03629>

Use in this report: Interleaving reasoning, actions and observations; architectural antecedent.

[2] Reflexion: Language Agents with Verbal Reinforcement Learning

Shinn, N., et al. (2023).

<https://arxiv.org/abs/2303.11366>

Use in this report: Feedback stored as linguistic memory; distinguishes adaptation from weight training.

[3] Generative Agents: Interactive Simulacra of Human Behavior

Park, J. S., et al. (2023).

<https://arxiv.org/abs/2304.03442>

Use in this report: Memory, retrieval, reflection and planning in a simulated social environment.

[4] AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation

Wu, Q., et al. (2023).

<https://arxiv.org/abs/2308.08155>

Use in this report: Configurable multi-agent conversation patterns with models, tools and humans.

[5] Voyager: An Open-Ended Embodied Agent with Large Language Models

Wang, G., et al. (2023).

<https://arxiv.org/abs/2305.16291>

Use in this report: Skill library, curriculum and feedback in Minecraft; domain transfer remains a separate question.

[6] MemGPT: Towards LLMs as Operating Systems

Packer, C., et al. (2023/2024).

<https://arxiv.org/abs/2310.08560>

Use in this report: Hierarchical context management and persistent memory.

REFERENCE REGISTER

References · evaluation and security

Primary-source links checked on 28 September 2026. This is a targeted research context, not an exhaustive systematic review.

[7] AgentBench: Evaluating LLMs as Agents

Liu, X., et al. (2023/2025).

<https://arxiv.org/abs/2308.03688>

Use in this report: Interactive evaluation across eight environments; failure analysis.

[8] τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains

Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. (2024).

<https://arxiv.org/abs/2406.12045>

Use in this report: Policy-sensitive tool interaction and repeated-trial reliability.

[9] GAIA: a benchmark for General AI Assistants

Mialon, G., et al. (2023).

<https://arxiv.org/abs/2311.12983>

Use in this report: General-assistant tasks requiring reasoning, tools and information gathering.

[10] Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection

Greshake, K., et al. (2023).

<https://arxiv.org/abs/2302.12173>

Use in this report: Untrusted retrieved material can carry adversarial instructions.

[11] On Calibration of Modern Neural Networks

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017).

<https://arxiv.org/abs/1706.04599>

Use in this report: Empirical confidence calibration; not evidence that a language model's self-rating is calibrated.

[12] CAMEL: Communicative Agents for Mind Exploration of Large Language Model Society

Li, G., et al. (2023).

<https://arxiv.org/abs/2303.17760>

Use in this report: Role-based cooperative agent interaction.

REFERENCE REGISTER

References · governance and research records

Primary-source links checked on 28 September 2026. This is a targeted research context, not an exhaustive systematic review.

[13] Artificial Intelligence Risk Management Framework (AI RMF 1.0)

National Institute of Standards and Technology (2023).

<https://www.nist.gov/itl/ai-risk-management-framework>

Use in this report: Voluntary framework for organisational AI risk management; version-specific reference.

[14] Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)

National Institute of Standards and Technology (2024).

<https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence>

Use in this report: Generative-AI risk identification and management actions.

[15] Datasheets for Datasets

Gebru, T., et al. (2018/2021).

<https://arxiv.org/abs/1803.09010>

Use in this report: Dataset purpose, composition, collection and intended-use documentation.

[16] Model Cards for Model Reporting

Mitchell, M., et al. (2018/2019).

<https://arxiv.org/abs/1810.03993>

Use in this report: Intended uses, evaluation conditions and performance limitations.

[17] Attention Is All You Need

Vaswani, A., et al. (2017).

<https://arxiv.org/abs/1706.03762>

Use in this report: Transformer architecture; a software cell is not a biological neuron or a newly trained foundation model.

[18] Secure Hash Standard, FIPS 180-4

National Institute of Standards and Technology (2015).

<https://csrc.nist.gov/pubs/fips/180-4/upd1/final>

Use in this report: Message digests for detecting changes to recorded bytes.

APPENDIX A

Glossary and notation.

Terms used consistently throughout this edition.

TERM	MEANING IN THIS REPORT
Abstention	An explicit choice to avoid automatic execution or selection when conditions are not met.
Ablation	A comparison that removes a defined component to inspect its contribution.
Autonomy	Permission to complete a specified task class without case-by-case intervention.
Calibration	Agreement between predicted confidence and observed correctness.
Cell	A bounded software responsibility or agent slot; not a biological neuron.
Coverage	The proportion of all eligible observations that receive the specified action.
Distribution shift	A change between the conditions assumed during development and those encountered later.
Grounding	Linking a statement or proposal to admissible evidence.
Idempotency	Protection against a repeated request causing an unintended duplicate effect.
Provenance	Recorded origin and history of information or an artifact.
Semantic correctness	Whether the proposed answer or action achieves the intended objective.
Synthetic data	Generated fixtures with explicitly specified distributions and labels.
p	Shared-shock probability in E3; pairwise correctness correlation in expectation under that model.
U / G	Task utility function / independent action-control function.

Terminology is deliberately operational. When a later study changes a definition, it should issue a new version and explain the effect on earlier results.

APPENDIX B

The JARVIS research charter.

A concise statement of what ZEISMO intends to investigate and preserve.

Purpose

Develop and evaluate an English-speaking AI assistant and a coordinated System Brain capable of assuming defined routine decision work. Investigate role-specific skills, modular cells, evidence-based memory and controlled adaptation to changing business conditions.

Scientific commitments

- Use real measurements for performance claims and identify simulations as simulations.
- Compare with simpler baselines and preserve negative results.
- Record versions of data, models, prompts, skills and policy.
- Separate evidence integrity, software testing and scientific review.
- Publish only the research materials authorised for release.

Operational commitments

- Assign each role a defined scope, owner and capability boundary.
- Promote updates through evaluation and preserve a recovery route.
- Measure verified task completion alongside error, cost and intervention.
- Treat departmental replacement as a staged workflow transition.
- Keep unresolved cases visible and return them to an accountable decision process.

The immediate research deliverable is a live-model evaluation using the frozen protocol in this report. Longer-term studies will address expert-panel comparison, voice interaction, memory reliability under adversarial evidence and sustained departmental performance.

PROJECT DIRECTION

A machine decision system should grow through tested competence and useful memory, while its authority remains clear, bounded and reviewable.

APPENDIX C · DEMONSTRATION 01

Inside the System Brain.

Original prototype interface · isolated demonstration data · 28 September 2026



Figure C1. The original fifteen-cell System Brain view, rendered with fictional role identities and sample memory. Activity counters are fixture values.

A map of responsibilities

The interface places fifteen software cells within a visual brain outline. Marketing, operations, finance, data and other roles have separate identities. Selecting a cell opens its skill and memory panel. The brain shape is a navigation metaphor; it is not a biological model or a measurement of consciousness.

READ THE IMAGE	INTERPRETATION
Named nodes	Fictional demonstration identities associated with the existing role slots.
Orion · 3	Three authored sample memory records, not three completed business tasks.
Paused activity	No background agent, live model or business action ran during this capture.

CAPTURE METHOD

The existing Blade templates were rendered offline. Browser requests received local sample JSON; every other request was blocked. No production account, customer record or operational telemetry was used.

APPENDIX C · DEMONSTRATION 02

A cell's skills and memory.

Operations example · role instructions, stored experience and visible task controls

Orion — Operations

Train this cell

Orion

Operations

Push a skill: rules, playbook, knowledge — appended to its training

Push to cell

Ask this cell

Ask anything — it answers or routes to the specialist cell

Ask

Background task

Give it something to understand/research — the bell notifie:

Assign

trained

Skill (training received)

DEMO FIXTURE | Operations role contract

Read approved task records. Check source and timestamp. Draft a next step. Escalate uncertain or consequential actions for approval. Never change access permissions.

Storage (experience & tasks)

DEMO FIXTURE | Operations role contract Read approved task records. Check source and timestamp. Draft a next step. Escalate uncertain or consequential actions for approval. Never change access permissions. · 2026-09-28 09:00

DEMO FIXTURE | Evidence freshness check Source: invented warehouse-status record DEMO-004. Observation: its timestamp is outside the permitted window. Intended response: request current evidence before recommending a stock transfer. This is authored sample data, not an agent-generated finding. · 2026-09-28 09:05

DEMO FIXTURE | Human handoff example A fictitious request exceeds the role boundary. The example illustrates a handoff to a responsible operator. No action was executed and no customer record was used. · 2026-09-28 09:10

Figure C2. The Operations cell panel. Every skill and memory entry shown is explicitly marked DEMO FIXTURE.

From role definition to inspectable context

The left panel contains the existing controls for naming a cell, adding skill text, asking a question and assigning a background task. The right panel exposes the supplied instructions and stored records. This organisation makes it possible to inspect the information presented to a role before requesting work.

The sample contract limits the role to reading approved records, checking timestamps and drafting a next step. Its memory examples describe stale evidence and an escalation. They are authored examples of the intended workflow, not responses generated by an AI agent.

MEANING OF “TRAINED”

In this prototype, the UI label indicates supplied skill content. It does not show that model weights changed, that a competency assessment passed or that the role can perform autonomously. No training or task button was submitted for these screenshots.

APPENDIX C · DEMONSTRATION 03

Open the memory archive.

Read-only browsing · role folders · source context visible for inspection

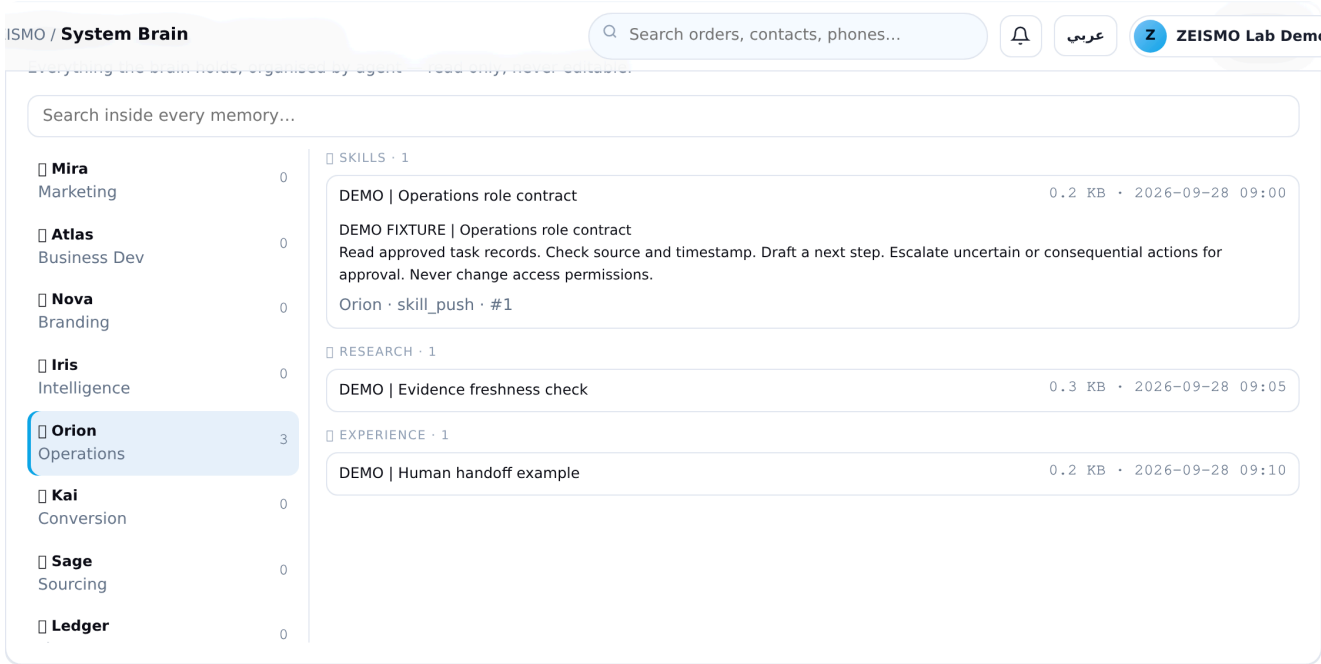


Figure C3. The original memory browser, with the Operations folder selected and one synthetic role-contract record expanded.

An archive should support challenge, not just recall

The archive groups records by agent and type. An operator can search, open a record and inspect its text and timestamp. The sample files make the difference between a skill, a research note and an experience entry visible without exposing internal business information.

A visible record is not automatically reliable evidence. Before future autonomous use, the research architecture requires source identity, permission, freshness and a trace to the decision that used it. The screenshot demonstrates the current inspection surface; the complete provenance and policy design is developed in the main report.

SAMPLE RECORD	PURPOSE OF THE DEMONSTRATION
Operations role contract	Locate the instructions defining the role boundary.
Evidence freshness check	Recognise when a source needs to be refreshed.
Human handoff example	Identify a situation that should return to an accountable operator.

APPENDIX C · DEMONSTRATION 04

JARVIS as the system interface.

A visual coordination surface for an English-first research programme

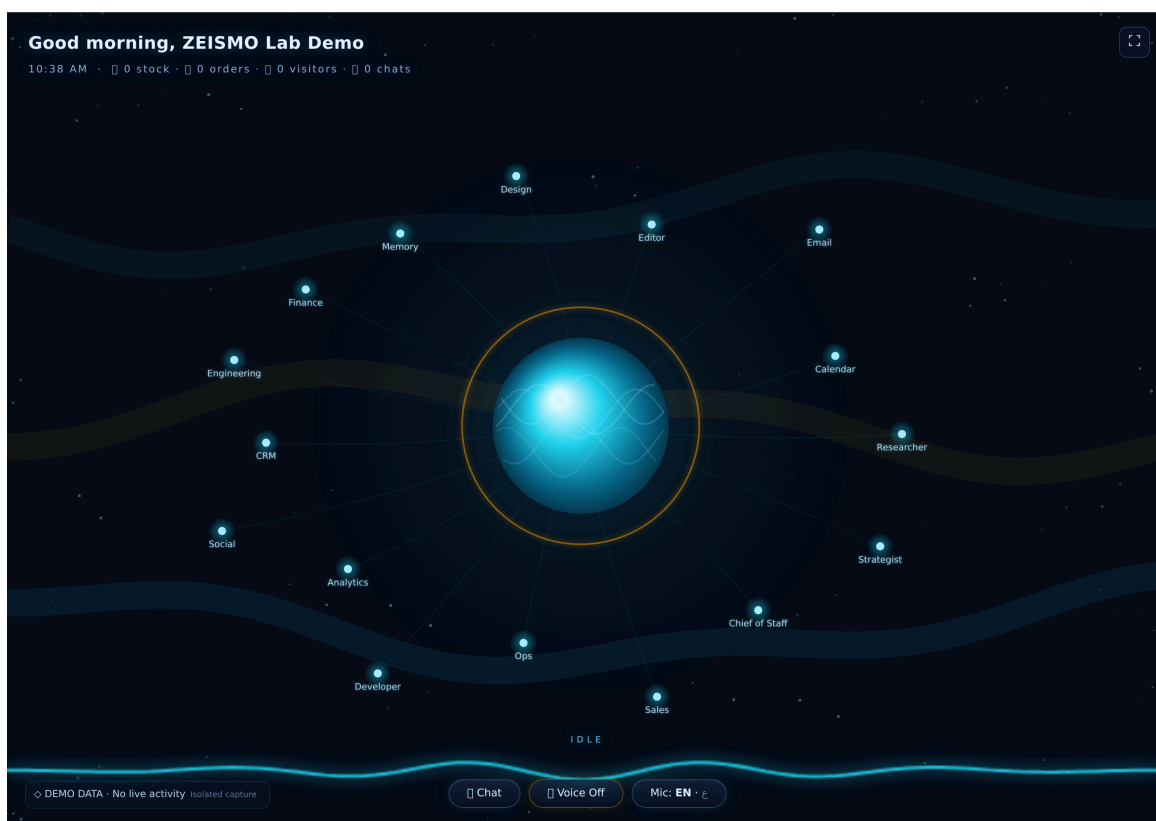


Figure C4. The original Jarvis command interface in fullscreen mode. Fictional identity, zero sample counters, voice off and no live activity.

The orbital labels represent the Jarvis interface’s operational categories. They are separate from the fifteen System Brain cell slots. Their presence on screen does not demonstrate that each named function is integrated, active or independently competent.

The intended research interface is exclusively English-speaking. This screenshot preserves the current prototype, which still exposes an Arabic microphone option. The report therefore retains English-only enforcement as an implementation gap rather than claiming it has already shipped.

CAPTURE REVIEW

All four selected views rendered. The Brain page also produced a pre-existing admin-header JavaScript collision: “Identifier esc has already been declared”. This is logged for remediation; the screenshots establish appearance, not a complete interface acceptance test.

APPENDIX D · QUESTIONNAIRE 01 / 03

Participant briefing and task sheet.

JARVIS & System Brain · exploratory evaluation · blank instrument v1.1

Purpose: understand whether people can interpret the interface, locate supporting information and identify the limits of delegated decision-making. Participation is voluntary. Skip any question or stop at any time. Do not enter names, contact details, passwords or confidential business information.

SESSION RECORD	COMPLETE BEFORE THE SESSION
Participant code (optional)	_____
Date / report or prototype version	_____
Viewing condition	☐ Report screenshots ☐ Isolated interactive demo
Main role	☐ Research ☐ Engineering ☐ Operations ☐ Other
Experience using AI tools	☐ None ☐ Occasional ☐ Weekly ☐ Daily

☐ I understand the activity and choose to participate. For an organised study, the facilitator must first provide the study contact, storage arrangements, retention period and withdrawal process. This document does not collect or submit responses.

Four observation tasks

TASK	PARTICIPANT INSTRUCTION	FACILITATOR RECORD
T1· Role	Find the Operations cell and describe its responsibility.	_____
T2· Evidence	Locate a stored record and identify its source or timestamp.	_____
T3· Limits	Identify one sign that this material is demonstration data.	_____
T4· Control	Explain when the example should return to a human operator.	_____

Record: C = completed without help; H = completed with help; U = not completed; N = not observed. For screenshots, assess interpretation only. Do not record navigation speed or execution success as if the participant used a live system.

APPENDIX D · QUESTIONNAIRE 02 / 03

Your assessment of the experience.

Circle one response per statement. Choose N/A when the material did not let you judge.

Response scale · 1 Strongly disagree · 2 Disagree · 3 Neither agree nor disagree · 4 Agree · 5 Strongly agree · N/A
Unable to assess

STATEMENT	RESPONSE
01· I understood the purpose of the System Brain view.	1 2 3 4 5 N/A
02· I could distinguish the responsibilities of different cells.	1 2 3 4 5 N/A
03· I could identify where a cell's instructions are shown.	1 2 3 4 5 N/A
04· I could identify where stored information is shown.	1 2 3 4 5 N/A
05· I could distinguish sample data from measured results.	1 2 3 4 5 N/A
06· The labels made the system's current state clear.	1 2 3 4 5 N/A
07· I knew what evidence I would need before accepting a recommendation.	1 2 3 4 5 N/A
08· I understood when a human should review an action.	1 2 3 4 5 N/A
09· I could identify limits to the authority assigned to a cell.	1 2 3 4 5 N/A
10· The information was readable at my viewing size.	1 2 3 4 5 N/A
11· I would feel able to question a recommendation.	1 2 3 4 5 N/A
12· The material helped me identify a suitable task for a controlled trial.	1 2 3 4 5 N/A

INTERPRETATION

This is an original exploratory questionnaire, not a validated psychological scale. Responses describe perceptions of the material shown. They do not measure intelligence, decision accuracy, safety or equivalence to human specialists.

APPENDIX D · QUESTIONNAIRE 03 / 03

Decision boundaries and feedback.

Participant code (optional): _____

For each example, select the highest level you would currently permit based on the evidence available to you. A = human only; B = AI draft with human approval; C = bounded automatic action; D = insufficient evidence.

EXAMPLE TASK	YOUR CHOICE
Summarise a non-sensitive internal document.	A B C D
Route a support request to a department.	A B C D
Recommend a stock transfer from an old record.	A B C D
Approve a financial commitment or change access rights.	A B C D

What was clear, useful or difficult?

What evidence or control would you need before relying on it?

Which routine task would you test first, and how would you judge success?

What should change before a wider trial?

Researcher analysis plan: report sample size, recruitment, viewing condition and missing answers. Summarise item distributions and task outcomes separately; compare screenshots and interactive sessions separately. Code free-text themes with a documented rubric. Do not invent a combined “trust” or “intelligence” score. No participant responses have been collected for this edition.

APPENDIX E · AUTHOR STATEMENT

Engineer Mohanad Osama Zahran.

Artificial Intelligence Expert · ZEISMO AI Engine · JARVIS & System Brain

Research direction

Mohanad Osama Zahran directs the concept behind the JARVIS and System Brain programme: an English-speaking interface, role-specific cells, curated skills and memory, and a staged transition of suitable routine departmental work to AI decision systems.

ZEISMO AI Engine

In the background supplied for this edition, Mohanad Osama Zahran identifies ZEISMO AI Engine as his AI model-training and adaptation work. The programme connects that engineering direction with research on coordinated agents, evidence use and controlled operational autonomy. Specific model architectures, training datasets and training-performance results are subjects for separately documented technical records.

Contribution and preparation

Project conception and research direction: Mohanad Osama Zahran. Literature organisation, source inspection, experiment implementation, numerical verification, manuscript drafting and document production: AI-assisted preparation. The report preserves the actual methods, datasets and verification scope. It does not attribute an independent review or manual execution of every test to the named author.

AUTHORISED TYPED SIGNATURE

Engineer / Mohanad Osama Zahran

Artificial Intelligence Expert

28 September 2026 · Technical report ZEISMO-LAB-TR-2026-001·v13

The typed signature and professional title are included at Mohanad Osama Zahran's express request. This is an author-attribution signature, not a certificate-based digital signature. The companion verification record identifies the final file by SHA-256.

Commercial interest: this work concerns systems developed within the ZEISMO product and research programme. The author's role in that programme is relevant when interpreting the research. Independent scientific peer review is not recorded.

ZEISMO RESEARCH & LAB

Build decisions that can be examined.

JARVIS & System Brain · Technical research report · Version 1.3

From an operational prototype to a documented research programme.

This edition records an architectural proposal, an inspected fifteen-cell prototype and reproducible synthetic experiments. It sets a measurable path towards AI employees that can assume defined routine responsibilities and improve through controlled experience.

SUGGESTED CITATION

Zahran, M. O. (2026). *JARVIS & System Brain: Towards accountable machine decision-making and progressively autonomous departments*. ZEISMO Research & Lab. Technical report ZEISMO-LAB-TR-2026-001, version 1.3.

zeismo.org

Operated by WEARLISTY SHOPPING SOLUTIONS LLC

Public technical preprint · 28 September 2026