RSI Agent:

practical business workflows and understandable AI operating costs

PUBLIC DEVELOPMENT WHITEPAPER

RSI AI · Melbourne, Australia · Version 1.3 · 9 October 2026

Status: Discovery-stage business with browser-based interactive prototypes using synthetic scenarios. This paper describes a proposed product and validation plan. It does not report deployed customer systems, model benchmarks, revenue, funding commitments or customer outcomes.

1. The question we are testing

Can a narrowly scoped, AI-assisted workflow help an Australian service business turn a new enquiry into a useful next action, while keeping human decisions and operating costs visible?

RSI AI is a Melbourne-based sole trader business developing RSI Agent. Development is founder-funded. The current prototypes explore three workplace handoffs: a store opening across coordinator, designer and installer; a maintenance evidence pack across technician, facilities coordinator and property manager; and a public-service handover across contact officer, case team and reviewer. Operating-cost tools remain available alongside those workflows.

A new enquiry can require missing information, a service-fit decision and a proposed next step before a person can approve it. A later handover can require supporting records to be checked. This paper does not assume that these problems are widespread, valuable enough to solve or poorly served by existing software. Those questions belong in discovery with operators.

The proposed starting point is one bounded workflow. Staff should be able to see the inputs and rules behind a proposed action, correct the result and decide what happens next. The public prototypes make those decisions tangible; a real pilot would validate one selected workflow.

2. Who the workflows are for

The customer hypothesis is a local service business whose owner, coordinator or administrator repeatedly qualifies enquiries, prepares a proposed schedule or quote, or checks the evidence needed for a handover. Melbourne provides the local context for discovery; maintenance and facilities services are examples to investigate, not validated customer segments.

Discovery will test four points:

1. Is there a recurring workflow with a named owner and a reasonably stable definition of a useful result?
2. Which missing inputs, conflicting requirements or unclear records create follow-up work, and how much staff time does that consume?
3. What can the business’s existing software already do, and what gap remains?
4. Can the business authorise representative examples and make a commercial decision about a narrowly defined pilot?

These questions take priority over building a general agent platform. If the existing system already solves the problem, or the remaining work cannot justify its operating cost, the proposed workflow should be revised or stopped.

3. The proposed product boundary

An enquiry-to-next-action workflow would compare a supplied request with approved business criteria, identify missing information and prepare a proposed schedule, quote or follow-up for staff review. Any real pricing or scheduling decision would depend on the business's rules and the current authorised system data. The browser prototype simulates these steps with synthetic inputs.

The maintenance evidence workflow compares a prepared job pack with an approved checklist and link findings to the supplied records. It would not certify physical work, establish document authenticity, determine legal compliance or decide entitlement to payment. Missing documentation would mean only that the required evidence has not been established in the reviewed pack.

A scoped pilot would draw from these components according to the chosen workflow:

- A case containing the enquiry, requirements, relevant records and proposed next action.
- Approved qualification, pricing or scheduling criteria with clear ownership and versions.
- Structured extraction where ordinary code alone cannot reliably interpret the supplied material.
- Deterministic checks for exact identifiers, required inputs and business constraints.
- Visible missing information, conflicts and the references behind a decision.
- A review interface for inspection, correction and explicit approval.
- An export or draft next action, with an external integration added only after its need is established.
- Usage records and cost controls that include retries and staff correction effort.

The public website presents store-opening delivery, property-maintenance evidence and public-service handover cases. Each view identifies the human owner, agent assistance, proposed deliverable and next responsible role over a shared local case. Visitors can change prepared inputs, inspect findings, review exact revisions and download local summaries. Stage navigation changes the view; it does not run checks or approve work. The cost simulation uses dated official API price references and editable workload assumptions; custom prices are identified separately. These prototypes do not perform model inference, accept uploaded business records, send quotes, reserve appointments or update an external system. Local behaviour does not establish production AI performance or operational controls.

4. Where AI belongs

The planned use of Claude is to interpret unstructured enquiries or records, produce structured fields and help draft a concise response grounded in the supplied material. The implementation would use an explicit first-party API integration if the pilot proceeds on that route.

Ordinary application code would remain responsible for exact comparisons, required-field checks, approved pricing and scheduling rules, permissions, state changes and budget enforcement. An AI response would not grant access, change an approved requirement or authorise a message.

Uploaded instructions would be treated as document content. A source document that asks the system to ignore controls or take an unrelated action would not become permission to do so.

The product should preserve uncertainty. When required inputs are incomplete, constraints conflict or evidence is unreadable, the next state should be a visible request for review. The system should not invent availability, manufacture missing physical evidence or silently reinterpret a customer's requirement to make a case pass.

5. Proposed technical architecture

The proposed architecture separates the case record, automated preparation, human approval and external action.

Component	Responsibility
Web application	Case navigation, input and source inspection, staff corrections and approval controls
Tenant-scoped database	Cases, membership, approved business rules, source versions, proposed actions and workflow state
Private object storage	Authorised documents referenced through controlled access
Durable job processing	Extraction and checks with bounded attempts, recovery and explicit waiting states
Model adapter	Structured requests to the chosen provider with measured usage and versioned configuration
Rule engine	Qualification, constraint and completeness checks using customer-approved criteria
Review and approval service	Approval bound to the exact inputs, rules and proposed action reviewed
Outbox and connector	An explicitly approved external action, if included in the pilot
Audit and usage records	Changes, decisions, attempts, uncertain outcomes and operating cost

A practical implementation may use a Vercel-hosted web application and durable background processing. The selected database, identity service, storage and model route would be documented for the real deployment. Provider features alone would not establish that the whole service meets a customer’s security or residency requirements.

Workflow outcome and delivery state should be separate. A proposed quote can be ready for review while availability remains unconfirmed; an approved draft can remain unsent if the delivery outcome is unknown. An evidence pack can also need further records. These are different business facts.

An approval must identify the reviewed inputs and source versions, rule version and draft. A material change invalidates that approval. Before an external effect, the application rechecks that the approved revision is still current. Duplicate execution must not create duplicate business effects, and an uncertain outcome must remain visible for recovery.

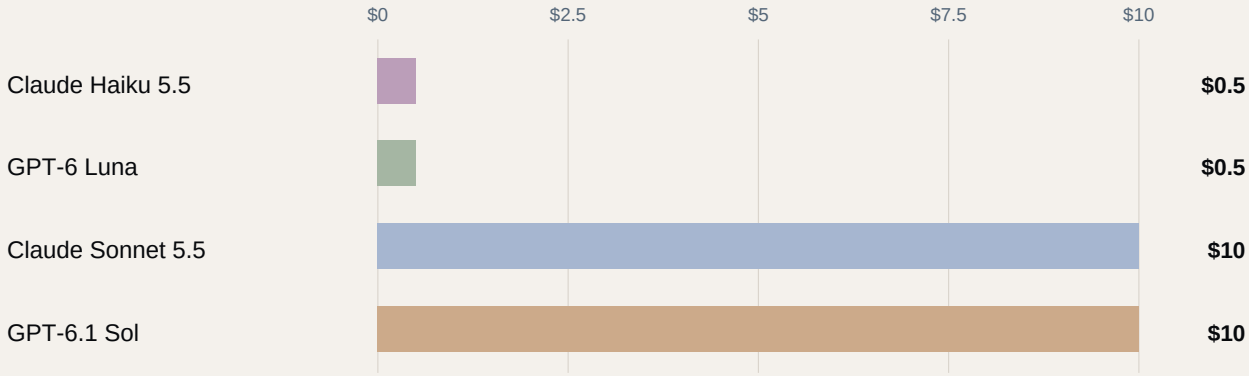
6. Operating cost is part of the design

A multi-agent workflow needs an explicit account of who uses context, who prepares a result, and who reviews it. A coordinator would choose a permitted specialist for a bounded task; ordinary application code would enforce permissions, budgets and approval gates. A smaller model is an option for evaluation, not proof that a particular business task will retain its quality.

The current cost reference uses direct Standard API list prices checked on 8 October 2026. All amounts are USD per million tokens. The token lab defaults to GPT-6.1 Sol for the general agent and Claude Haiku 5.5 for the routine agent; users can also select GPT-6 Luna and Claude Sonnet 5.5. The assumed workload, retained context, repeated prefix, routing share and expected billable output remain editable simulation inputs.

Standard API output prices

USD per million tokens - 64,000 input tokens per request - 8 October 2026



Zero-based scale. List prices are not task quality, usage limits or measured RSI savings.

Sources: Anthropic API pricing and OpenAI API pricing. Context tiers are described below.

Short-context reference rates, USD per million tokens

Model	Input	5m write / write	Cache read	Output
Claude Haiku 5.5	$0.1	$0.125	$0.01	$0.5
GPT-6 Luna	$0.1	$0.125	$0.01	$0.5
Claude Sonnet 5.5	$2	$2.5	$0.1	$10
GPT-6.1 Sol	$2	$2.5	$0.1	$10

Haiku 5.5 uses a higher tier when the full input exceeds 100,000 tokens. GPT-6 Luna and GPT-6.1 Sol use a higher tier above 272,000 input tokens. The tier applies to the complete request, including cached input; the simulator evaluates the original and selected-context paths separately. Sonnet 5.5 retains its Standard rates throughout its supported context window. Anthropic cache-write references use the five-minute option.

Workload and subscription budgets

The model accounts separately for ordinary input, cache writes, cache reads and expected output. Every active model starts with a cold write; the assumed reuse share does not establish actual cache eligibility, prefix length or lifetime. Tokenisers differ between models, so equal token counts do not establish equal text, quality or cost per completed task. The expected output should include all billed output; unmodelled reasoning, tool calls, retries, regional or speed premiums, hosting, staff effort and tax can increase the actual bill.

A business coding subscription has a different billing unit. ChatGPT Business with Codex and Claude Team Standard are listed at $20 per seat per month when billed annually, or $25 on monthly billing. Claude Team Premium is $100 per seat per month annually, or $125 monthly. The public reference assumes at least two seats. Annual figures are monthly equivalents of an annual commitment. Enterprise procurement can use different terms and requires an actual quote.

The optional team budget combines seat fees with the selected, separately billed API workload. It does not convert a subscription allowance into tokens or assume unlimited automated use. Users choose team size, premium seats, billing cadence and monthly workdays; the same gross API workload is shown beside each seat plan.

Claude Team API credits are a separate, conditional allowance: $20 per Standard seat and $100 per Premium seat, pooled up to $500 monthly. Credits must be eligible and claimed, are being rolled out, expire each cycle and do not extend interactive coding limits. They apply to the linked Claude Platform organisation. The estimator leaves them off by default and, when explicitly enabled, offsets only the Anthropic part of API charges. It does not deduct them from OpenAI usage or identify a user's actual account balance.

The visual presentation borrows the reference's clear axes, restrained model colours and comparable units. It shows price data, not an RSI benchmark. Simulated cost differences are arithmetic under selected assumptions; measured savings would need provider records and evaluation of the same task, including failed attempts, corrections and human review.

7. How a pilot would be evaluated

A pilot should compare the proposed workflow with the customer's current process on comparable cases. Machine execution time is not staff time saved. Evaluation should include the effort needed to inspect, correct and act on the result.

Measure	Decision it supports
Unsupported or incorrect findings	Whether reviewers can rely on the system's proposed interpretation
Incorrect progression	Whether missing inputs, conflicting rules or absent evidence incorrectly allow a case to proceed
Source-reference accuracy	Whether the reviewer can locate the evidence used for a finding
Correction and handling time	Whether the new process reduces total staff work
Accepted outcome rate	Whether the output is useful for its stated next step
Cost per accepted workflow outcome	Whether the operating model is sustainable
Recovery and isolation results	Whether interruptions, duplicate actions and access boundaries are handled correctly

Early evaluation would use varied synthetic or appropriately authorised examples, followed by a separate holdout set. Cases should include missing enquiry details, incompatible service criteria, schedule conflicts, invalid pricing inputs and ambiguous requests. Evidence workflows should also include missing records, conflicting identifiers, unreadable material and duplicated documents. Malicious instructions in supplied content must be represented.

Critical missing-input or business-constraint cases must fail visibly in the agreed initial test set before live reliance. Passing a finite set does not prove that future errors are impossible. Human fallback, monitoring and a deliberate release decision remain necessary.

8. Data handling and the current prototypes

The public prototypes use browser-local rules, synthetic scenarios and editable simulation assumptions. Visitors can explore store-opening plans, maintenance evidence handoffs and permitted public-service records. Token-cost tools remain separate planning illustrations. They accept no uploaded documents and make no model inference request. Sample approval and download are local actions: they do not contact a real customer, reserve a slot, deliver a commercial quote or update a business record.

The website also includes a local pilot brief builder. Visitors can use its options to outline a workflow and download a text file for their own planning. The builder does not request a name or email address, submit the selections to RSI AI or establish a pilot commitment. The brief and sample workflow outputs are generated for download in the browser.

The application does not use analytics or set cookies. Its hosting infrastructure may still process ordinary technical request data to deliver and operate the site. Any contact form, analytics, authentication or future upload feature would introduce a different data scope and must have a corresponding notice and operating process before use.

For a real pilot, the customer and RSI AI would need to agree the permitted purpose, authorised users and documents, provider route, processing locations, access controls, retention, deletion, recovery and offboarding. Customer content should be excluded from general analytics and unnecessary error logs. Workflow events should use controlled record references where possible.

Australian branding or an Australian application region does not establish that every component processes data in Australia. The full data path must be understood before making a residency commitment. Customers with specific regional, security or procurement requirements may need a separate deployment arrangement.

9. A proposed twelve-week validation sequence

The roadmap is conditional on customer access, founder capacity and the agreed scope. It is a plan for one useful workflow and a commercial decision, not a promise to complete a general-purpose platform.

Stage	Work	Exit evidence
Weeks 1-2	Operator discovery and selection of one workflow	Recurring problem, named owner and a gap worth testing
Weeks 3-4	Authorised sample, acceptance criteria and pilot scope; private data foundation	Usable examples, a permitted data path and a credible commercial decision
Weeks 5-6	The selected workflow, versioned rules, source references and cost controls	Unresolved inputs stay visible; attempts and spending are bounded
Weeks 7-8	Correction, approval and export; one connector only if justified	Evidence changes invalidate approval; duplicate or uncertain effects are recoverable
Weeks 9-10	Quality, isolation and recovery evaluation; controlled pilot where agreed	Comparable reviewer-effort and quality evidence
Weeks 11-12	Outcome review and continue, revise or stop decision	Customer value, acceptable total cost and a clear next commitment

10. Commercial development and local contribution

The proposed commercial route starts with a defined implementation and pilot. Continuing service would depend on demonstrated value, agreed support and predictable operating cost. Repeated implementations may reveal reusable qualification criteria, workflow rules, evidence-handling components, approval controls, usage controls and onboarding practices that can form a configurable product.

Repeatability is an assumption to test. It requires evidence that the next implementation takes less effort while maintaining quality and customer fit. A collection of unrelated custom projects would not establish that product transition.

The local contribution we aim to investigate is practical: helping Australian businesses adopt useful agent-assisted operations and understand the associated decisions and costs. Potential measures include validated local use cases, staff handling effort, repeatable implementation work and whether customers choose to continue. These are proposed measures, not reported outcomes or job-creation promises.

Government service delivery is a longer-term option. It would follow delivery evidence, support capacity and a procurement and security approach appropriate to the particular agency. No government partnership, endorsement or contract is claimed.

11. Collaboration and funding readiness

RSI AI welcomes assessment of the development thesis by business operators, delivery collaborators and potential supporters. The immediate need is evidence: a well-defined problem, authorised examples, a useful evaluation and a realistic implementation scope.

Business enquiries can be sent to founder@rsiagent.app. RSI AI operates as an Australian sole trader, with ABN 88 515 897 752 active since 3 April 2024 and the RSI AI business name registered since 5 August 2026. These public registration details were checked on 8 October 2026. The local brief builder prepares an outline; visitors decide whether to email it themselves.

For an appropriate Victorian or Australian Government funding opportunity, preparation would include the program's actual eligibility requirements, the applicant's verified business details, a scoped project, milestones, an eligible budget, evidence of need and a plan to report outcomes. No grant eligibility, award or application result is claimed in this paper.

For investors, the material questions remain customer willingness to pay, observed value, delivery effort, ongoing support, cost per accepted result and the path from implementation work to a repeatable product. These questions should be answered with evidence as development progresses. Private forecasts, a financing structure and investment terms are outside this public paper.

The next commitment should be proportionate to what is known: validate one workflow, make the decision criteria explicit and earn the case for the next stage.

About this version: Public development thesis, Version 1.3, 9 October 2026. The associated browser-based prototypes at rsiagent.app illustrate proposed business workflows and cost choices using synthetic scenarios and editable assumptions. Provider names describe planned or actual technology use as stated and do not imply endorsement, partnership or funding approval.

Selected official references

These references inform the proposed implementation and due diligence. Provider and program names do not imply endorsement.

ABN Lookup: RSI AI public registration details
https://abr.business.gov.au/ABN/View?abn=88515897752

Claude Haiku 5.5 release and pricing
https://www.anthropic.com/claude-haiku-5-5#pricing

Claude API list prices
https://platform.claude.com/docs/en/about-claude/pricing

OpenAI Standard API list prices
https://developers.openai.com/api/docs/pricing

GPT-6 Luna pricing tiers
https://developers.openai.com/api/docs/models/gpt-6-luna

GPT-6.1 Sol pricing tiers
https://developers.openai.com/api/docs/models/gpt-6.1-sol

ChatGPT Business and Codex plan pricing
https://learn.chatgpt.com/docs/pricing

Claude Team seat pricing
https://claude.com/pricing

Claude Team API credit conditions
https://support.claude.com/en/articles/17154008-monthly-api-credits-for-max-and-team-plans

Claude Startups: current public program information
https://claude.com/programs/startups

Claude API data residency
https://platform.claude.com/docs/en/manage-claude/data-residency

OAIC guidance on commercial AI products
https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-the-use-of-commercially-available-ai-products

Vercel Workflows
https://vercel.com/docs/workflows

Australian Government agentic AI guidance
https://www.digital.gov.au/policy/ai/agentic-ai-addendum

Related site resource: RSI AI dated funding-pathway research and official sources
https://rsiagent.app/funding/