micro1.

About Us

micro1 is the human intelligence layer for agentic AI. Our journey began by building our own AI recruiter which has led to our one-of-a-kind global network of thousands of highly vetted experts across 100+ domains. We work with these experts to leverage their unique knowledge to train models and agents for top AI labs and enterprises around the world.

Closing the gap between agents and humans

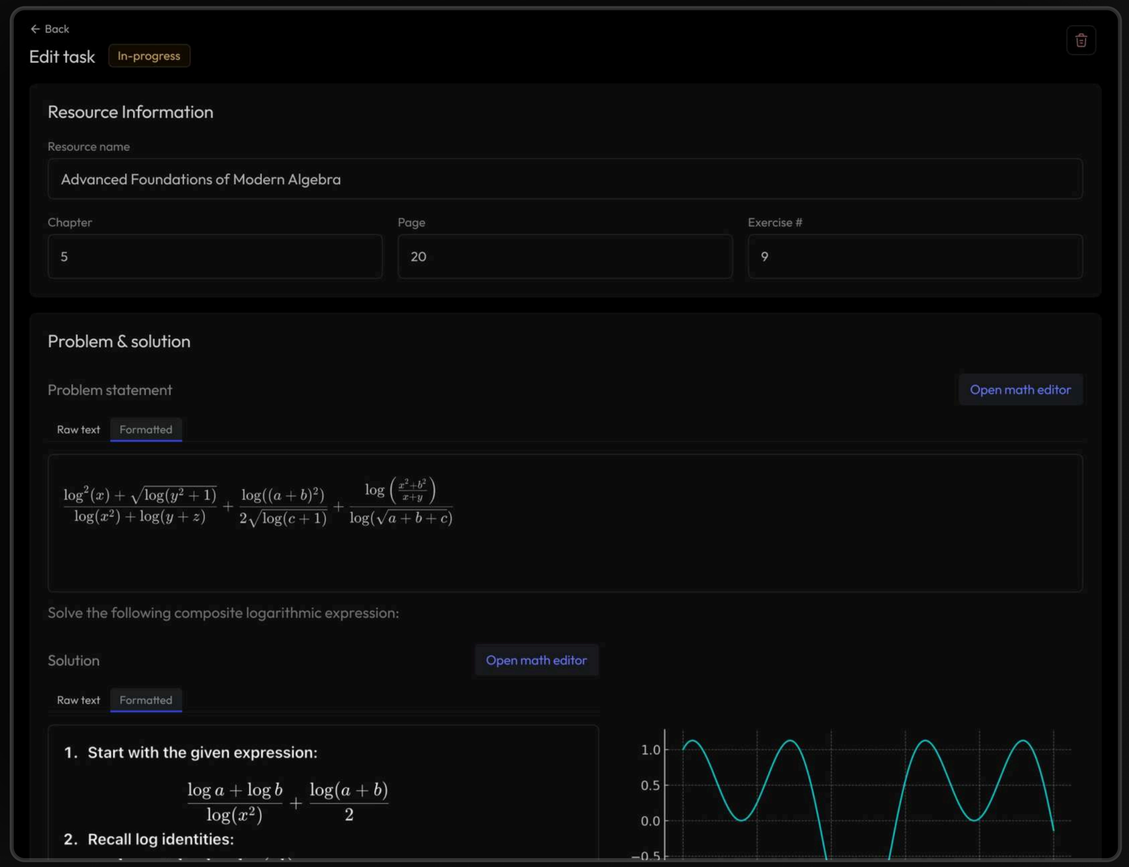
- Enterprises need agents that can function as one of their employees. Today, generic agents can't be trusted because they hallucinate, miss context, and lack the domain knowledge that humans use to make decisions.
- micro1 embeds expert human data into agents, enabling them to handle complex workflows with accuracy, structure, and enterprise-level consistency.

micro1's human data engine

The backbone of our business, where we transform human expertise into high-quality evaluation and training data

How it works

1. **Source Experts:** our proprietary AI recruits and vets real specialists for your specific use case
2. **Define Structure:** translate workflows into structured data tasks
3. **Generate Data:** produce reasoning, corrections, edge cases, and multimodal annotations
4. **Quality Check:** muti-layers of expert and AI-based quality assurance
5. **Deliver Datasets:** consistent, scalable, and ready for agent fine tuning
6. **Continuous Monitoring:** to ensure your agent evolves alongside your business



Our training process

1 Evaluations

Evaluate Your Agent Like a Real Operator

Tests reasoning quality

Checks if steps are followed correctly

Flags safety, accuracy compliance, and reliability risks

Outcome: You see exactly where the agent fails and why

2 Human Data

Give Your Agent Real-World Judgment

Domain-specific examples from experts

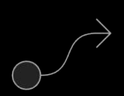
Correct answers for tricky edge cases

Clear rubrics that teach the model what “good” looks like

Outcome: Higher-signal training data that drives real improvement

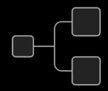
3 Fine-Tuning

Turn Your Agent Into a High-Performer



Boost accuracy on your workflows

Improve tool use and structured actions



Fix recurring logic and reasoning errors

Outcome: Agents tailored to your business that perform consistently

4 Production monitoring

Keep Your Agent Safe, Stable, and On Track

Production evaluations and drift detection

Continuous performance scoring across tasks

Alerts when quality drops below thresholds

Outcome: Agents you can trust in every-day production

Example monitoring dashboard

Agent Monitoring

Agent Active

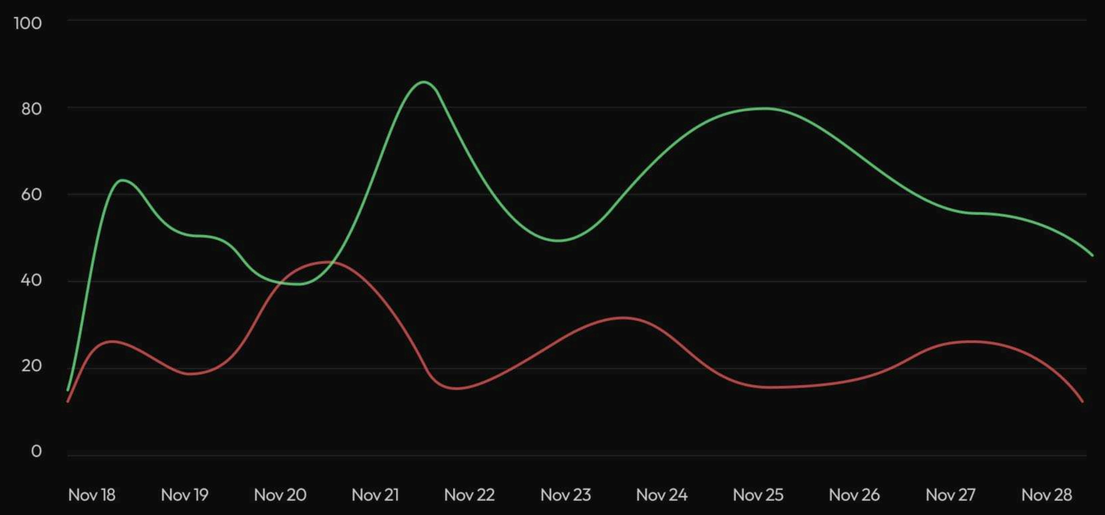
Transactions Scanned
Live
15,240

Fraud Prevented
$2.1M

Accuracy Score
98.8%

Live risk detection

Safe traffic
Blocked threats



Recent Alerts

AML Flag
Account Frozen

Identity Theft
Blocked

Sanctions Check
Passed

Why we're doing this

Evaluating AI systems is challenging… common methods often fail to capture real-world behavior. More rigorous evaluation frameworks are needed to understand how AI will act in practice.

— Anthropic, “Challenges in evaluating AI systems”

Using evals strategically can make a product more reliable at scale, decrease high-severity errors, protect against downside risk, and give an organization a measurable path to higher ROI.

— OpenAI, “How evals drive the next chapter of AI for businesses”

We’ll mold to your workflow

If you don’t have an agent yet

Start with baseline model evals

micro1 helps you validate, tune, and launch your first internal agent

If you already have an agent

Plug in micro1 as the evaluation + human data layer

Improve performance without rebuilding your stack

If you use agent vendors

micro1 integrates as the third-party evaluation + human-data provider

Case studies

Microsoft: Tax AI Agent

Microsoft developed an accountant digital worker that operates inside Copilot.

Process

- Our approach began with assembling a specialized team of experts and human data managers in each domain. We sourced specific documentation to form the outlined knowledge base for the project. For each document, we developed queries paired with golden responses, ensuring a balanced mix of simple, medium, and complex cases.

Outcome

- A final dataset of 150 tasks was utilized to deeply assess and improve agent performance. This ensures behavior reflects day-to-day accounting practices to reliably automate high-stakes tasks.

Clinical Form Completion Co-Pilot (OASIS)

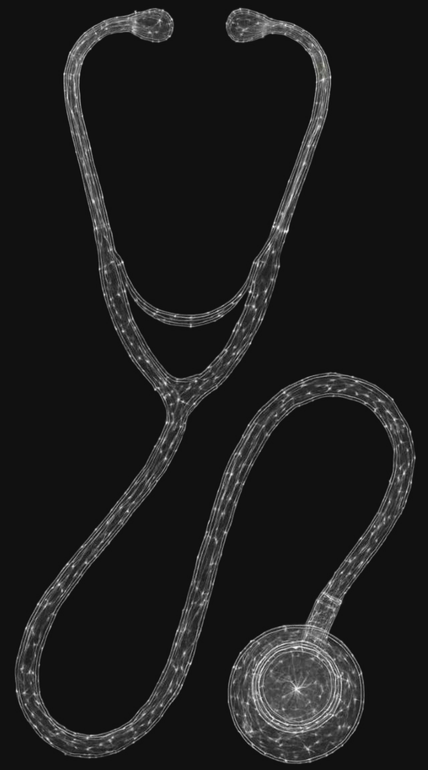
We were tasked with reducing the manual documentation burden that home health clinicians have with filling out OASIS clinical forms which require complex decision logic, patient data cross referencing, and completion of dozens of fields.

Process

- micro1 mapped the chain-of-thoughts clinicians follow when completing OASIS forms. We identified workflow gaps, common follow-up questions, and missing patient data needed for accurate completion. To do this, clinician-generated examples were used to train the agent to distinguish auto-fill vs clinician-required inputs.

Outcome

- This resulted in a 25-40% reduction time for clinicians to fill out the OASIS forms (depending on case complexity), and improved the consistency and accuracy in structured fields. Overall, this demonstrates how human decision decomposition enables agentic systems to operate safely in high-stakes environments



Box: Evals for Enterprise AI Agents

Box, a cloud-based content management platform, needed evaluation data to guide role-specific enterprise agent design, benchmark models, and ensure reliability within real customer workflows.

Process

- micro1 recruited over 90 specialized domain experts across 11 different fields to expand Box's evaluation dataset. Tasks were modeled based on real enterprise agent workflows requiring multi-document synthesis and reasoning like drafting, due diligence, data analysis, and more.

Outcome

- Box received production-accurate coverage of enterprise workloads, improved agent reliability on complex, multi-document enterprise tasks, and a scalable foundation for ongoing evaluations

Conversational Agent Intelligence

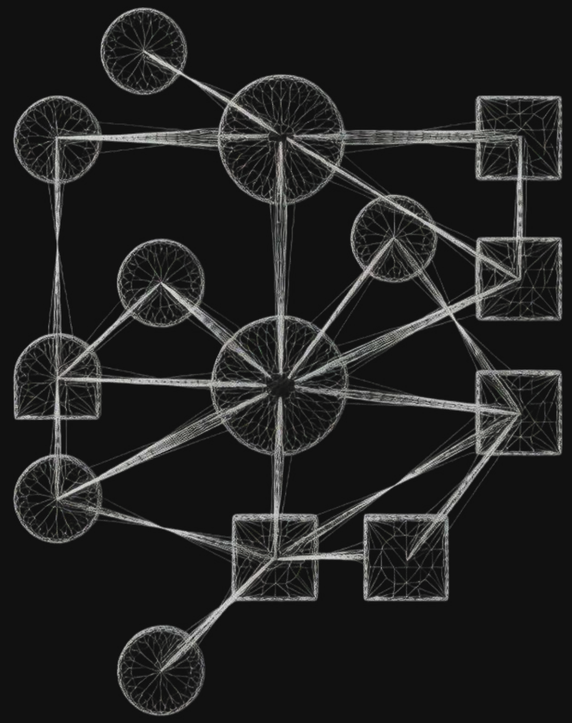
An enterprise deploying their conversational AI agent needed to augment their system to ensure it could handle nuanced conversations, be trusted by professionals, and improve from user feedback.

Process

- micro1 conducted over 700 expert-led evaluations across diverse roles to measure conversational intelligence including contextual understanding, follow-up quality, domain accuracy, and conversational flow.

Outcome

- 83% trust rate among domain experts, 25% improvement in perceived intelligence over 8 weeks, identified leading driver of perceived intelligence to shape product roadmap, delivered repeatable framework for continuous improvement.



Multi-Hop Finance Research Agent

A large financial institution needed to evaluate whether agentic AI models could perform real-world financial research , navigating SEC filings, chaining multi-step reasoning, performing calculations, and verifying outputs without hallucinations.

Process

- micro1 built an agent benchmark grounded in expert-authored human data from 1,000 finance expert-created tasks. Tasks required models to retrieve information across multiple public sources (10-Ks, 10-Qs, tax guidance), synthesize scattered data, perform financial calculations, and verify a single ground-truth answer.

Outcome

- Hallucination rates were reduced by 65%, the number of errors per response decreased by 78% and there was a clear identification of failure modes in financial modeling and multi-document reasoning. Overall, the client received a reliable method to benchmark its agentic system against institutional-grade financial research workflows.

Enterprise Interview AI Agent

A large enterprise began to leverage AI for interviews, but needed to improve fairness, consistency, and human judgment alignment before heavily relying on the system for decision making.

Process

- micro1 implemented a human-in-the-loop evaluation framework to continuously benchmark the AI agent's assessments against expert reviewers. Evaluations were focused on two independent dimensions: language proficiency mapped to CEFR levels (A1-C2) and soft skills including professionalism, empathy, adaptability, collaboration, and judgment.

Outcome

- The agent achieved 73% exact agreement with expert-labeled CEFR language levels and 95% agreement within one CEFR level, and resulted in a scalable production-ready evaluation system.

Let’s get started:

Basic Diagnostic Pilot

- Baseline evals
- Failure mode analysis
- Clear improvement plan

Advanced Evaluation Pilot

- 10 - 100 task evaluation dataset
- Performance report
- Model comparisons
- Fine-tuning scope

Next steps

1. Scoping call to align on specifications
2. Define project roadmap
3. Start building

Get in touch:

izzy@micro1.ai | Enterprise Partnerships Lead