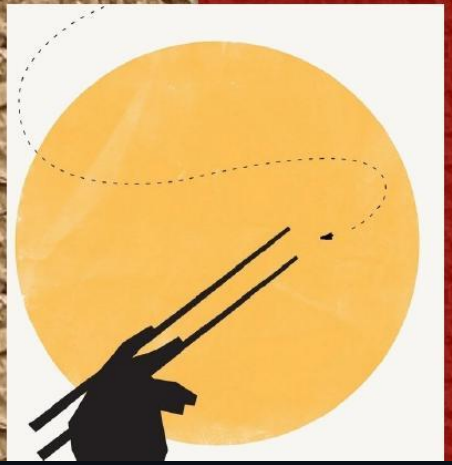
Welcome To The 11th Annual Hacking Conference



Know Your Enemy: AI Fraud Detection

Presented by **Santosh Vasudevan** · CPE Code: **WAXON**

The Institute of Internal Auditors Chicago

ISACA Chicago Chapter

The Fraud Investigation Problem

Hundreds of alerts daily

Most teams are buried — every alert takes 30-60 min to investigate manually

Data lives everywhere

Transaction systems, vendor records, policy docs, historical cases — all in different places

Speed creates risk

Slow reviews mean fraud continues. Fast reviews mean missed signals.

Pattern recognition at scale

Humans are great at single-case judgment. They struggle with 500 simultaneous patterns.

The bottleneck isn't judgment — it's the legwork before judgment.

Meet Balto

Named after the sled dog ***who delivered when it mattered most.***

What Balto does:

- 🔍 Receives a fraud alert
- ⚙️ Autonomously gathers evidence from 4 data sources
- 🔗 Cross-references findings and identifies patterns
- 📄 Delivers a structured investigation memo
- 🧑‍⚖️ Hands it to YOU to make the call

Balto doesn't make the call. Balto makes the call possible.

What We're Covering Today

01

The Fraud Scenarios

Three patterns Balto is trained to catch

02

How Balto Detects

Rule-based scanning — no AI, instant results

03

The Agent Loop

Plan → Act → Synthesize → Memo (ReAct pattern)

04

MCP Tools

How Balto talks to 4 independent data sources

05

RAG in Plain English

How Balto searches policy docs & past cases

06

Live Demo

Watch Balto investigate a real (synthetic) alert

The Threat Has Already Scaled.

The question is whether your defenses have.

$9.5T

Global cybercrime cost, 2024

Cybersecurity Ventures, 2024. Up from $3T in 2015 — projected $10.5T by 2025.

258

Days — avg. breach lifecycle

IBM Cost of a Data Breach 2024. 7-year low — 194 to identify, 64 to contain.

70%

Of fraud caught by tips, internal audit, or mgmt

ACFE 2024 Report to the Nations. Only 3% by external auditors.

960+

Alerts per day at an average SOC

Praetorian / Dropzone AI, 2025. 50–80% are false positives.

Cybersecurity Changed the Rules of Auditing

Traditional Fraud

- 📋 Paper trails and physical records
- 🏦 Fraud happened at human speed
- 📅 Quarterly or annual audit cycles
- 👥 Insider fraud: one person, one scheme
- 🗺️ Contained within one geography or system
- 🔍 Sample-based testing was sufficient

VS

Cyber-Era Fraud

- 💻 Digital footprints scattered across 100+ systems
- ⚡ Attacks happen in milliseconds and days
- 🔄 Continuous monitoring is the new baseline
- 🌐 Organized criminal networks, nation-state actors
- 🔗 Cross-border, cross-system, cross-entity
- 📊 100% transaction coverage is now expected

How Traditional Auditing Works

It works. For the problems it was designed to solve.

1

SELECT SAMPLE

Test 5-10% of transactions. Statistical sampling. Assume the rest behave similarly.

2

PULL RECORDS

Request supporting docs from business units. Days to weeks to receive.

3

MANUAL REVIEW

Auditor reviews each record against policy. 30-60 min per complex transaction.

4

INTERVIEW & TEST

Walk-throughs, control testing, corroboration from management.

5

WRITE FINDINGS

Document issues. Each finding requires evidence, risk rating, recommendation.

6

ISSUE REPORT

Final report delivered weeks or months after the period under review.

⚠ Typical cycle: 4–12 weeks. Fraud that occurred in Q1 may not surface until Q3.

Where Traditional Auditing Breaks Down

Not a criticism — an honest constraint analysis.

📊 Volume

A mid-size company runs 50,000+ transactions per month. A team of 5 auditors can meaningfully review maybe 2-3%.

→ 97%+ of transactions go unexamined — by design.

⏱️ Latency

Audit cycles are quarterly or annual. A vendor conducting threshold splitting in February may not be caught until an October audit.

→ The fraud window stays open for months.

🔗 Complexity

Modern fraud involves coordinated patterns across systems — the signal is in the aggregate, not any single transaction.

→ Humans struggle to hold 500 simultaneous patterns in memory.

😴 Alert Fatigue

Automated rule-based systems flag thousands of anomalies per day. Analysts triage by gut feel — and miss things.

→ A study found 45% of security alerts are never investigated.

The Adversary Automated First.

🔴 ATTACKERS — operating with automation, AI-generated phishing, scripted BEC, and ML-driven social engineering since ~2018

🟡 DEFENDERS — mostly still using manual processes, spreadsheets, and quarterly cycles that haven't changed since the 1990s

What attackers can now do that they couldn't before:

🤖 Scripted vendor fraud

Automated creation of shell vendor networks with generated documentation, synthetic identities, and AI-written invoices

📧 AI-generated BEC

Business email compromise at scale — AI mimics writing style, targets the right approver, adapts to responses in real time

🔄 Pattern evasion

Machine learning optimizes invoice amounts and timing to stay below detection thresholds — exactly like threshold splitting, but automated

⚡ Speed of execution

A coordinated fraud campaign can execute across hundreds of vendors and accounts in hours — before any audit cycle would catch it

You can't fight automated fraud with manual processes. Speed is not optional.

Traditional vs. AI-Augmented Auditing

DIMENSION	TRADITIONAL	AI-AUGMENTED
Coverage	2–10% sample	**100% of transactions**
Detection speed	Weeks to months	**Minutes to hours**
Consistency	Varies by analyst skill/fatigue	**Same logic, every time**
Pattern complexity	Single-case, linear	**Cross-system, aggregate**
Evidence gathering	30–60 min per alert, manual	**30–60 sec, automated**
Policy grounding	Auditor recalls from memory	**Cites exact clause, doc**
Final judgment	Human auditor	Still human auditor
Accountability	Human sign-off	Human sign-off

The last two rows are intentional. AI augments the analysis. Humans own the decisions.

The Modern Fraud Audit Stack

AI doesn't replace the auditor. It changes where in the stack auditors spend their time.

🗄️ LAYER 1 — DATA

Raw transaction, vendor & system data

Transactions, vendor registrations, ERP exports, email logs, access logs. The raw material. Volume: millions of records.

🔍 LAYER 2 — DETECTION (AI + Rules)

Automated anomaly detection

Rule engines + ML models scan 100% of data continuously. Flag threshold splits, ghost vendors, duplicates, BEC patterns. No human in this loop.

🤖 LAYER 3 — INVESTIGATION (AI Agent)

Autonomous evidence gathering

AI agent receives flagged alert, pulls evidence from 4+ data sources, cross-references policy and precedent, writes a structured memo. This is Balto.

🧑‍⚖️ LAYER 4 — JUDGMENT (Human)

Auditor reviews memo and decides

You apply context the AI can't know. You weigh relationships, business risk, regulatory exposure. You sign ESCALATE, MONITOR, or CLOSE.

🐺 Layer 3 — Investigation — is where Balto lives. The rest of this talk is about exactly how it works.

What Good AI Auditing Looks Like

⚠️ What bad AI auditing looks like: AI makes the call, human rubber-stamps it, no audit trail, no explainability. Regulators are already writing rules about this.

Transparent reasoning, not a black box

Every AI recommendation must come with an explanation: which evidence, which policy, which precedent led to that risk score. If you can't audit the AI, you can't trust it.

Human oversight built in by design

The AI generates recommendations. A qualified human must approve or reject every material action. This isn't a UX choice — it's a control requirement.

Consistency that survives staff turnover

The AI runs the same investigation logic on alert #1 and alert #10,000. No analyst fatigue, no 'I didn't have time to check the policy,' no inconsistency between shifts.

SECTION 01

The Fraud Scenarios

Three patterns hiding in plain sight

🔴 FRAUD PATTERN 1

Threshold Splitting

How it works: A vendor splits one big purchase into multiple smaller invoices, each just under the $10,000 auto-approval limit — so no human ever reviews them.

AUTO-APPROVED	AUTO-APPROVED	AUTO-APPROVED	AUTO-APPROVED
$9,800	**$9,750**	**$9,900**	**$9,850**
Feb 3 · MNO-0301	Feb 5 · MNO-0302	Feb 7 · MNO-0303	Feb 9 · MNO-0304
< $10K threshold	< $10K threshold	< $10K threshold	< $10K threshold

Combined total: $39,300 · All auto-approved · No human review · 6 days

Vendor: MNO IT Services (V-1005) · Detected in Balto's synthetic dataset

🔴 FRAUD PATTERN 2

Ghost Vendor

Two vendors registered 2 days apart, same building, adjacent units, sequential tax IDs — and the same person approved both.

V-1004

JKL Office Solutions

Registered:	Dec 18, 2025
Address:	4510 Market St, Unit 3B
Tax ID:	23-6710483
Total Spend:	$6,200

V-1007

STU Industrial Supply

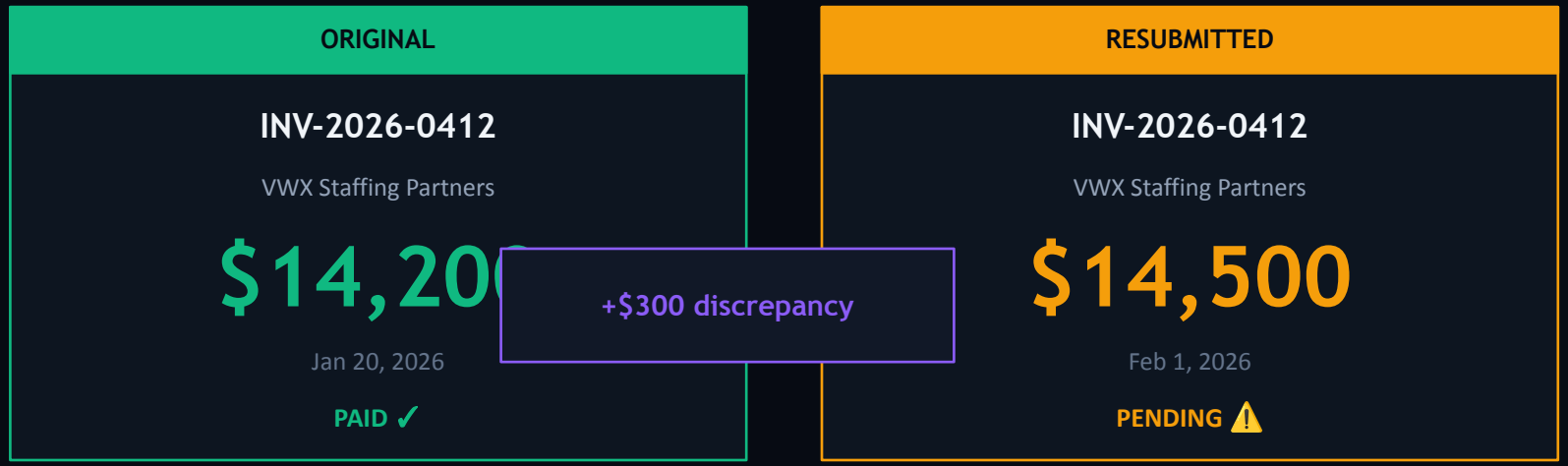
Registered:	Dec 20, 2025
Address:	4510 Market St, Unit 3A
Tax ID:	23-6710497
Total Spend:	$17,900

⚠️ Same building, adjacent units · Sequential tax IDs (14 numbers apart) · Both approved by: James Whitfield · No prior transaction history (V-1007)

FRAUD PATTERN 3

Duplicate Invoice

Same invoice number, submitted twice — with a slight amount difference. If both go through, the vendor gets paid twice.



Could be a clerical error — or a deliberate double-pay. Balto flags it either way and lets the auditor decide.

SECTION 02

How Balto Works

Detect → Investigate → Review

The Three-Phase Pipeline

Phase 1

DETECT

No AI · Milliseconds

Rule-based scanner reads raw transaction and vendor data. Flags threshold splitting, ghost vendors, duplicate invoices. No LLM involved.

Phase 2

INVESTIGATE

AI Agent · 30–60 seconds

Balto receives the alert, forms an investigation plan, and calls tools iteratively to gather evidence. Writes a structured memo.

Phase 3

REVIEW

You · Minutes

You read the memo: risk score, evidence, policy violations, similar cases. You apply judgment no AI can replace. You make the call.

The separation is intentional. Detection is cheap. Investigation requires reasoning. Review requires you.

What is an AI Agent?

(For those who haven't met one yet)

The Analogy

Think of Balto as a very smart research assistant.

You hand it a case file. It reads it, decides what to look up, goes and looks things up, reads what it finds, decides if it needs more, and then writes you a summary.

It loops through this process 2-3 times per investigation. That's the 'agent loop.'

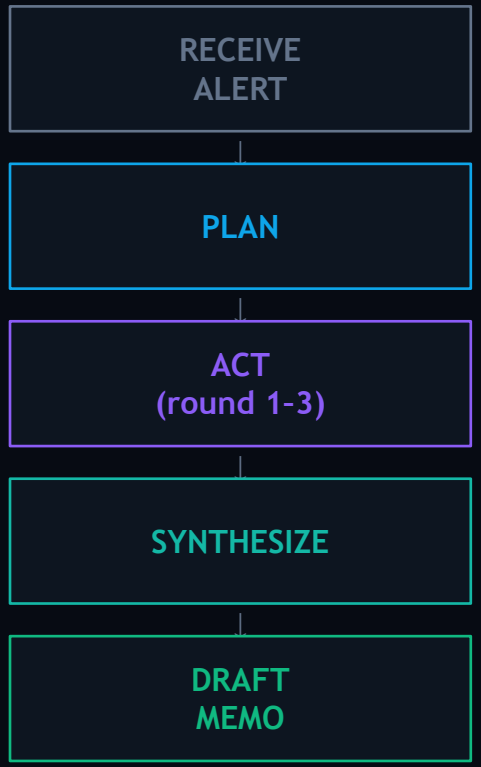
What Makes it an Agent

- 🎯 Makes decisions about what to do next
- 🔧 Calls external tools (not just text)
- 🔁 Loops: evaluates if it has enough info
- 📋 Produces structured, actionable output
- 🛑 Knows when to stop (hard cap: 3 rounds)

An AI Agent = AI + Tools + a loop that decides when it's done

The Agent Loop (ReAct Pattern)

ReAct = Reasoning + Acting. The agent thinks, then acts, then thinks again.



Receive Alert:

Vendor name, description, date range passed in as plain text.

Plan:

Claude reads the alert and writes an investigation plan: which tools to call, what to look for, what fraud pattern this might be.

Act (iterates 2-3 times):

Calls MCP tools to gather evidence. After each round, a quick check: 'Do I have enough?' If not, another round. Hard cap at 3.

Synthesize:

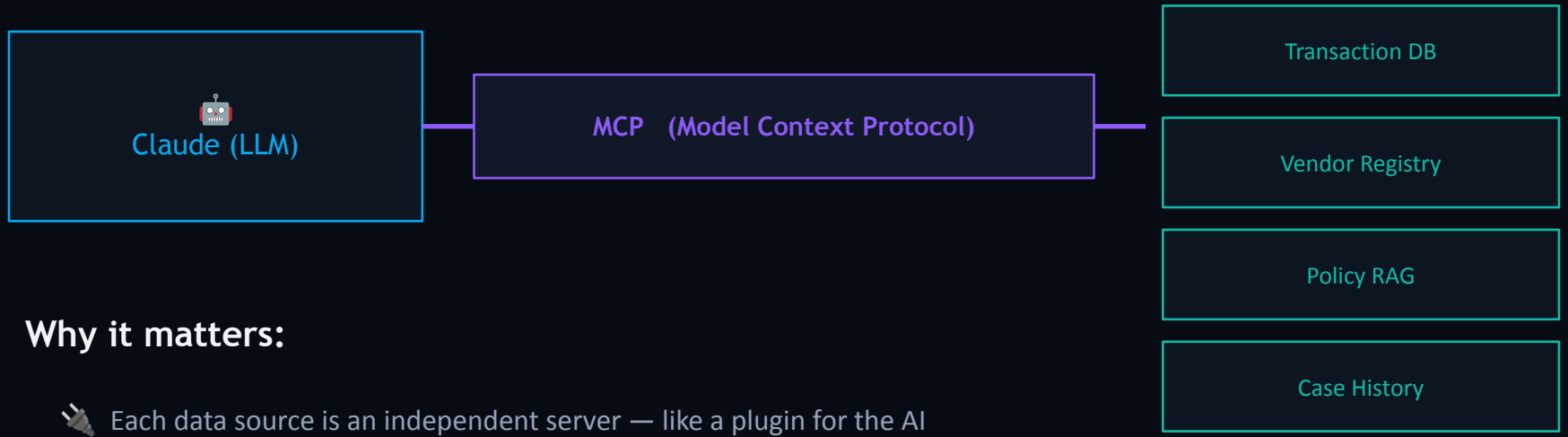
Cross-references all evidence: pattern match, supporting facts, mitigating factors, gaps, risk level.

Draft Memo:

Generates the final JSON memo — structured, consistent, always ending with 'Senior review recommended.'

What is MCP?

Model Context Protocol — how AI talks to your tools



Why it matters:

- 🔌 Each data source is an independent server — like a plugin for the AI
- 🔒 The AI can only see what the tool returns — not raw database access
- 🔄 Tools can be swapped, upgraded, or replaced without changing the AI
- 🏢 Mirrors how enterprise systems actually work — data lives in silos

MCP = a standard protocol so AI models and tools can speak the same language

Balto's 4 MCP Tools

💳 transaction_db

get_vendor_transactions · flag_threshold_patterns

Pulls vendor payment history. Flags splits near $10K threshold, duplicate invoice numbers, and unusual frequency spikes.

🚩 4 invoices at ~$9,800 in 6 days

🏢 vendor_registry

get_vendor_profile · flag_vendor_risks

Returns full vendor profile: registration date, address, related entities, tax ID, approver. Flags newly registered and shared-address vendors.

🚩 Registered 2 days before related vendor at same address

📋 policy_rag

query_policy

Semantic search over 3 compliance policy documents. Returns the most relevant policy sections for grounding the memo in actual language.

🚩 FIN-POL-2024-005 §3.1: 3+ invoices in 14 days near threshold

🗂️ case_history_rag

query_similar_cases

Finds the 3 most similar historical fraud cases. Calibrates risk score based on how those cases were resolved.

🚩 CASE-2025-006: TUV Maintenance → ESCALATED (HIGH)

SECTION 03

RAG in Plain English

Retrieval-Augmented Generation

RAG: Giving the AI a Filing Cabinet

Without RAG

The AI knows a lot *generally.*

But it doesn't know
your specific split payment policy.

It doesn't know
how your company resolved similar cases last year.

It would have to guess — or make things up.

VS

With RAG

Before answering, the AI searches a knowledge base.

For Balto, that's:

📄 3 policy documents
— chunked, indexed, and searchable by meaning

🗂️ 10 historical cases
— with outcomes, risk scores, and signals

Now the AI can cite chapter and verse from YOUR policies.

How RAG Works in Balto

1

INGEST

Documents are chunked & indexed

Policy markdown files are split into sections. Each section is converted into a 'vector' (a mathematical fingerprint of its meaning) and stored in ChromaDB.

2

QUERY

Agent asks a natural language question

e.g. 'split invoice approval policy' or '4 invoices under $10K in one week' — the question is also converted into a vector.

3

RETRIEVE

Most similar chunks are returned

ChromaDB finds the chunks whose vectors are closest to the query vector — semantic similarity, not just keyword matching.

4

GENERATE

AI uses the chunks to write the memo

The retrieved policy sections and case summaries become part of the AI's context — cited by name, quoted accurately.

Result: Balto cites FIN-POL-2024-005 §3.1 verbatim — not a hallucination, a retrieval.

The Investigation Memo

Every investigation produces the same structured output — consistent, auditable, never a verdict.

🔴 HIGH RISK

alert_summary:

4 near-threshold invoices from MNO IT Services in 6 days, totaling $39,300

key_evidence:

- 4 invoices $9,750–$9,900 in 6 days
- All auto-approved; no human review
- Matches FIN-POL-2024-005 §3.1 criteria

policy_violations:

Split Payment Policy FIN-POL-2024-005: 3+ invoices within 14-day window near threshold

similar_cases:

CASE-2025-006 (TUV Maintenance) → ESCALATED · CASE-2024-001 (EFG Tech) → ESCALATED

recommended_action:

ESCALATE — Senior review recommended

Why this schema matters:

risk_score

Consistent triage — LOW / MEDIUM / HIGH only. Forces prioritization.

key_evidence

Specific facts from tools — not AI guesses. Cites transaction IDs and dates.

policy_violations

Links to actual policy docs with section numbers. Auditable.

similar_cases

Grounds risk assessment in historical precedent, not pattern matching alone.

recommended_action

Always ends with 'Senior review recommended.' Never a verdict. Always escalates to you.

SECTION 04

Live Demo

Watch Balto investigate three alerts in real time

(All data is synthetic — no real PII)

DEMO 1 · THRESHOLD SPLITTING

MNO IT Services — V-1005

What Balto will do:

📋 PLAN	Identify as potential threshold-splitting pattern. Plan: transaction history + vendor profile + split payment policy + similar cases
💳 ACT round 1	get_vendor_transactions → 4 near-threshold invoices in 6 days flag_threshold_patterns → SPLIT_INVOICE_PATTERN confirmed
🔍 ACT round 1	get_vendor_profile → active vendor, no related entities flag_vendor_risks → THRESHOLD_PATTERN flag already on file
📖 ACT round 2	query_policy('split invoice threshold') → FIN-POL-2024-005 §3.1 query_similar_cases → CASE-2025-006 TUV Maintenance (HIGH · ESCALATED)
📄 MEMO	Risk: HIGH · Action: ESCALATE — Senior review recommended

DEMO 2 · GHOST VENDOR

JKL Office + STU Industrial

What Balto will do:

📋 PLAN	Identify as ghost vendor / related entity pattern. Will need vendor profiles on both V-1004 and V-1007, plus onboarding policy
🏢 ACT round 1	get_vendor_profile(V-1004) → registered Dec 18, related entity V-1007 flag_vendor_risks(V-1004) → RECENTLY_REGISTERED, SHARED_ADDRESS
🏢 ACT round 1	get_vendor_profile(V-1007) → registered Dec 20, NO_PRIOR_HISTORY flag_vendor_risks(V-1007) → sequential tax IDs, same building adjacent units
📖 ACT round 2	query_policy('new vendor monitoring') → FIN-POL-2024-007 §5 query_similar_cases('related entity same address') → CASE-2024-004 (HIGH · ESCALATED)
📄 MEMO	Risk: HIGH · Action: ESCALATE — Senior review recommended

DEMO 3 · DUPLICATE INVOICE (Watch for the MEDIUM / CLOSE result)

VWX Staffing Partners — V-1008

This one shows Balto recommending MONITOR (not always ESCALATE — that would be a broken demo)

📋 PLAN	Identify as duplicate invoice pattern. Need transaction history, duplicate flag analysis, and relevant policy + cases
💳 ACT round 1	get_vendor_transactions → INV-2026-0412 appears twice ($14,200 + $14,500) flag_threshold_patterns → DUPLICATE_INVOICE flag confirmed
🏢 ACT round 1	get_vendor_profile(V-1008) → active since 2022, existing DUPLICATE_INVOICE flag flag_vendor_risks → no registration/address issues found
📖 ACT round 2	query_similar_cases('same invoice submitted twice amount mismatch') → CASE-2024-003 (MEDIUM · MONITORED), CASE-2025-007 (LOW · CLOSED)
📄 MEMO	Risk: MEDIUM · Action: MONITOR — Senior review recommended

The Human in the Loop

This is the most important slide in the deck.

Balto is a junior investigator. **You are the senior auditor.**

🧠 Context Balto doesn't have

You know which vendors you've had informal conversations with. You know about the planned office move. You know your procurement manager is on leave.

⚖️ Judgment Balto can't apply

Is a $300 invoice discrepancy worth a relationship with a 3-year vendor? Balto can flag it. Only you can weigh it.

🛡️ Accountability that can't be automated

Every ESCALATE or CLOSE decision needs a human name behind it. The memo is evidence. The decision is yours.

Human judgment isn't being replaced. It's being freed up to actually judge.

Key Design Decisions

Rule-based detection first

No LLM for the alert scanner. Milliseconds, zero API cost. LLM only fires when a human says investigate.

MCP over direct function calls

Each data source is independent. Mirrors enterprise reality. Swap, upgrade, or replace any tool without touching the agent.

ReAct loop with hard cap

The agent loops up to 3 rounds but usually exits after 2. Hard cap prevents runaway costs during a live demo (or production).

Prompt-driven behavior

All 6 prompts control the agent — not hardcoded graph logic. Change a prompt, change the behavior. No code changes needed.

Memo schema enforcement

'Senior review recommended' is enforced in the prompt AND validated in code. Not a suggestion. It's a constraint.

What Balto Catches vs. What It Doesn't

✓ Balto Catches

- ✓ Threshold splitting patterns (3+ invoices near $10K in 14 days)
- ✓ Recently registered vendors with unusual first invoices
- ✓ Related entity clusters — shared addresses, sequential tax IDs
- ✓ Duplicate invoice numbers with amount discrepancies
- ✓ Pattern violations of specific policy thresholds
- ✓ Matches to historical fraud case signatures

✗ Beyond Balto's Scope

- ✗ Verbal agreements between employees and vendors
- ✗ Fraud that perfectly mimics legitimate patterns
- ✗ Novel schemes with no historical precedent
- ✗ Contextual factors only you know about
- ✗ The final judgment call — that's your job

Balto raises the floor, not the ceiling. Human judgment is what raises the ceiling.

AI + Human: The Right Division of Labor

TASK	BALTO (AI)	AUDITOR (YOU)
Pull transaction history across 10 vendors	✓ In seconds	30–45 min
Find relevant policy section from 3 docs	✓ Semantic search	10–20 min
Match current case to 10 historical cases	✓ Vector similarity	Requires memory or manual search
Structure findings in consistent format	✓ Always consistent	Varies by investigator
Weigh business context and relationships	✗ No visibility	✓ Essential
Apply judgment on borderline cases	✗ Not appropriate	✓ Required
Sign off on ESCALATE / CLOSE decision	✗ Never	✓ Always

Three Things to Take Home

1

The legwork is automatable. The judgment isn't.

AI agents are genuinely good at gathering, cross-referencing, and formatting evidence. They are not good at — and should not be used for — making final decisions on complex, contextual cases. Know which is which.

2

The tools are real and production-ready today.

MCP, RAG, and LangGraph aren't research projects. They are open standards and stable libraries you can deploy now. The hard part is the data curation and prompt engineering — not the infrastructure.

3

Your judgment is more valuable now, not less.

When Balto handles 200 routine alerts, you spend time on the 3 truly ambiguous ones that require context and experience. That's not job displacement. That's job elevation.

Questions?

Balto gathers the evidence. You make the call.

Things worth debating:

· How do you prevent the AI from becoming a rubber-stamp tool?

· What happens when Balto gets it wrong — and who's accountable?

· How do you handle novel fraud patterns the training data doesn't cover?

Cobra Kai Day · May 11, 2026 · Fraud Detection Case Study · 3:45–4:35 PM