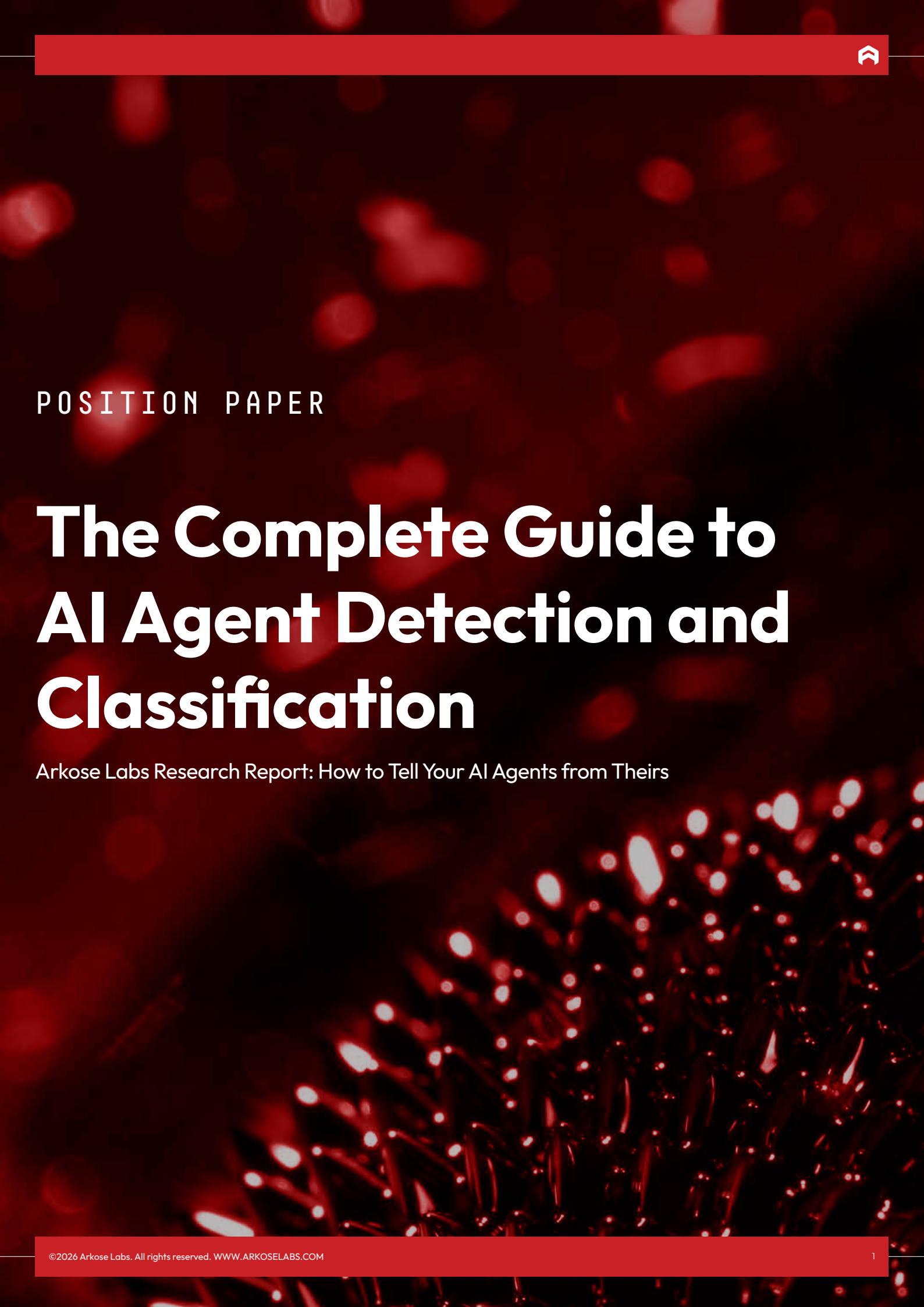
POSITION PAPER

The Complete Guide to AI Agent Detection and Classification

Arkose Labs Research Report: How to Tell Your AI Agents from Theirs



About This Research

This report represents Arkose Labs' latest research findings from analyzing billions of sessions across our global customer base. As the industry leader in bot management and fraud prevention, protecting major technology platforms, financial institutions, gaming companies and e-commerce leaders, we've been at the forefront of detecting and classifying the emergence of agentic AI traffic.

Our research team, including experts from our ACTIR (Arkose Cyber Threat Intelligence Research) unit, has identified critical patterns, developed novel detection methodologies and created classification frameworks that we believe are essential for the industry to understand. This document shares our key findings and insights to help organizations prepare for the fundamental shift that agentic AI represents to online security.

Why This Research Matters Now: With tools like OpenAI Operator, Anthropic's Claude, Perplexity's Comet and browser-based agents rapidly evolving, the distinction between self-disclosing, non-disclosing and malicious automation has never been more critical—or more difficult. Many legitimate AI tools operate without disclosure, creating a complex landscape where intent must be inferred from behavior rather than identity.

Executive Summary

Our research reveals a critical finding: ALL AI agents—whether legitimate AI assistants or malicious attackers—exhibit identical technical characteristics. They all use fake devices, synthetic browsers, and cloud infrastructure. ChatGPT helping a user shop shows the same technical characteristics as a credential stuffing bot.

Every day, these technically identical agents hit login and signup pages across the internet. Some are legitimate partners and AI assistants, others are sophisticated attackers, and many fall into a gray area where intent isn't immediately clear. The problem our research identified: technical detection alone cannot distinguish between them.

The emergence of agentic AI has created a landscape where the old security paradigm has collapsed. Through extensive analysis, we've determined that the old question "is this a bot?" is dead. Even "is this a fake device?" doesn't help—they ALL are. The new question is: "What is this agent's intent, and is it authorized to do what it's trying to do?"

Our findings show this isn't about detecting technical anomalies anymore—it's about behavioral analysis, intent classification and governance at scale.

Bot Management: From Detection to Policy Control

Different Policies for Different Business Models

This isn't just about detection—it's about management and control. Different organizations have vastly different needs:

Policy Spectrum We Support

Zero Tolerance (Deny All Non-Human Traffic)

- Some customers want purely human traffic
- Critical for certain financial transactions, voting systems or verification flows
- Need technical controls to enforce complete agent blocking

Self-Disclosure Only (Known Agents)
- Allow only agents that properly identify themselves
- Block all non-disclosing agents regardless of intent
- Maintain strict control over automation

Risk-Based Allowlisting (Trust But Verify)
- Allow known good agents (self-disclosing)
- Analyze non-disclosing agents for legitimacy
- Block only confirmed malicious actors
- Requires sophisticated classification capabilities

Full Visibility (Monitor and Measure)
- Track all agent types without blocking
- Understand traffic composition before making policy decisions
- Identify which non-disclosing agents provide value

Adaptive Management (Context-Dependent)
- Different policies for different endpoints
- Allow agents for browsing, restrict for transactions
- Time-based or volume-based restrictions
- Nuanced approach based on business impact

The Critical Requirement: Policy Agility

If policies change, organizations need technical controls to enforce them immediately.

This means:

- Real-time classification capabilities
- Ability to switch from monitoring to blocking instantly
- Granular control at the endpoint level
- Clear audit trails of what was allowed/blocked and why

Implementation Flexibility

The same detection system must support:

- A bank that wants zero agents touching payment flows
- A social platform managing both good and bad automation
- An e-commerce site that welcomes price comparison bots
- A government service requiring human-only interaction

The New Reality: When "Bot" Becomes a Spectrum

Our Research Discovery: The Web as We Know It

Our analysis indicates that the fastest adoption of agent-driven activity will utilize the web as we know it today. Think about it: self-driving cars like Waymo don't have special traffic lights or lanes—they're adapting to the world humans use today.

Similarly, our research predicts the fastest growth and adoption of agents will be on pre-existing flows and apps. Agents won't wait for special APIs or agent-specific interfaces. They'll use:

- The same login pages humans use
- The same checkout flows
- The same navigation patterns
- The existing web infrastructure we've built over decades

This is why our detection and classification research is so critical—agents are operating in spaces designed for humans, making them harder to distinguish but also more impactful to manage properly.

The Three Types of AI Agents We've Identified

Our most important finding: ALL agents—legitimate or malicious—run on cloud infrastructure, fundamentally distinguishing them from human users operating from traditional devices.

When OpenAI's official crawler visits your site, it's using a synthetic browser on cloud infrastructure. When ChatGPT helps a user shop, it shows the same fake device fingerprints as a fraud bot. When your payment processor's API checks transactions, it exhibits the same technical anomalies as an attacker.

Based on analyzing billions of sessions across our customer base at Arkose Labs, our research has identified three distinct categories of agentic AI traffic. The technical layer cannot distinguish between them—they all look fake using today's indicators of fraud:

Good Agents That Self-Identify

- **Who they are:** OpenAI, Anthropic, legitimate payment processors, authorized integrations
- **How they behave:** Publish IP ranges, provide user agent strings, announce their AWS regions
- **Detection approach:** Simple—verify headers, match IPs, allow traffic
- **The technical reality:** Even these "good" agents use fake devices and synthetic browsers—they just tell you who they are

Examples:

- OpenAI's GPT agent with published IP ranges (still uses synthetic browsers)
- Anthropic's Claude with specific user agent (still shows fake device fingerprints)
- Your payment processor announcing their infrastructure (still runs on cloud, not real devices)

Good Agents That DON'T Self-Identify (The Hidden Helpers)

- **Who they are:** Consumer AI tools, undocumented vendor tools, internal AI systems, partner integrations
- **Real-world examples:**
  - OpenAI's ChatGPT operating in agent mode through user browsers
  - Perplexity's Comet browser performing searches on behalf of users
  - Anthropic's Claude performing web actions for users
  - AI assistants from marketing or sales teams deployed without security awareness
- **Why they don't disclose:** Operating through consumer browsers, protecting user privacy, new deployments or simply lacking disclosure protocols
- **The technical reality:** These show EXACTLY the same technical anomalies as attacks:
  - Fake devices and synthetic browsers
  - Cloud infrastructure pretending to be iPhones or laptops
  - Automated mouse movements and superhuman click speeds
  - Inconsistent device fingerprints and network signatures
- **The only difference:** Their intent is to help users, not harm them—but you can't see intent in the technical layer

Bad Agents and Malicious Traffic

- **The ghosts:** Never advertise, actively hide their nature, pursue harmful objectives
- **The wolves in sheep's clothing:** Spoofing good agent signatures, fake OpenAI user agents, trying to get preferential treatment
- **Their goal:** Account takeover, fake account creation, scraping, payment fraud
- **The tell:** They're trying to commit fraud, and it shows in their patterns

Important distinction: Not all non-disclosing agents are malicious. The challenge is determining intent through behavioral analysis rather than self-identification.

How Our Research Distinguishes Agent Intent

Since all non-disclosing agents share identical technical characteristics (fake devices, synthetic browsers, cloud infrastructure), we've developed behavioral and contextual analysis methods to determine intent:

Intent Pattern Consistency (Legitimate Non-Disclosing Agents)

Legitimate agents that don't self-disclose still show predictable patterns

- Check the same endpoints in the same order
- Operate at predictable times
- Show consistent behavioral patterns across sessions
- Navigate with purpose, not exploration

The key insight:
Even without identifying headers, legitimate behavioral fingerprints are clear and repeatable.

Boundary Respect

Legitimate agents (whether self-disclosing or not) demonstrate respect for your infrastructure:

- Follow robots.txt directives
- Respect rate limits
- Back off when challenged
- Respond appropriately to 429 (Too Many Requests) errors

The key insight:
They're trying to provide a service or assist users, not extract maximum value. This behavior helps distinguish legitimate non-disclosing agents from malicious ones.

Business Activity Correlation (The Paper Trail)

This is the golden signal for legitimate agents: Non-disclosing agents that correlate with legitimate business events are likely benign.

- When an AI assistant accesses 100 accounts → those users actually initiated the assistant
- When a vendor's undocumented tool logs in → there's a corresponding API call from your integration
- When a payment processor checks transactions → they align with real payment attempts
- When search traffic increases → it correlates with user queries on AI platforms

The key insight:
The technical signatures might be identical to an attack, but the business context reveals the truth. Malicious agents rarely have legitimate business correlation.

How Malicious Agents Reveal Their Intent: Key Research Findings

Our threat research team has identified specific patterns that distinguish malicious actors from legitimate non-disclosing agents:

Intent Leakage

Malicious agents can't help but reveal their criminal intent:

- Test stolen credentials in patterns that don't match password recovery behavior
- Create accounts with data that's technically valid but statistically suspicious:
  - Addresses that don't match IP geolocation
  - Low cost infrastructure networks
  - Emails from temporary domains
- Navigate directly to high-value endpoints without browsing
- **The speedrun problem:** They navigate your site like someone speedrunning a video game—technically perfect, but no human behaves this way

Note: Legitimate non-disclosing agents (like AI assistants) may also navigate efficiently, but their patterns align with helping users rather than extracting value.

Optimization Patterns That Reveal Intent

While all agents show technical anomalies, malicious patterns emerge in behavior:

- Navigation optimized for extraction rather than user assistance
- Testing multiple accounts systematically versus helping specific users
- Accessing high-value endpoints without legitimate business purpose
- Mouse movements that are superhuman (all agents) but also purposeless (malicious)
- **The speedrun problem:** Navigating like someone speedrunning a video game versus natural exploration

Key distinction: ChatGPT navigating efficiently to help a user shop versus a bot efficiently extracting pricing data—same technical profile, different intent pattern.

Challenge Response Patterns

When we deploy novel challenges specifically trained and tested against Large Multimodal Models (LMMs) to exploit their weaknesses:

- These aren't standard CAPTCHAs but proprietary puzzles targeting known LMM vulnerabilities we've discovered
- Malicious agents fail systematically—they fail the same way every time
- Legitimate agents either solve correctly or back off appropriately
- Humans struggle randomly—they get confused by different things
- AI agents show characteristic API call sequences: vision→reasoning→action
- Our challenges evolve based on the latest LMM capabilities and limitations

Note: Specific challenge methodologies are proprietary and available through direct consultation.

Spoofing Mistakes (Impersonation Attempts)

When agents actively pretend to be something specific (not just using fake devices, which they all do):

- Claiming to be OpenAI but coming from wrong IP ranges
- Right user agent but wrong TLS fingerprint for that specific service
- Headers in the wrong order for the claimed identity
- **It's like looking at a fake ID—might fool you at first glance, but when you really look...**

Important distinction: ALL agents use fake devices (that's normal). This is about agents specifically impersonating other services to gain trust—a clear signal of malicious intent versus the universal practice of device synthesis.

Technical Detection Layers: Our Methodology

The Universal Technical Profile of Agents

Through years of research and development, we've discovered a critical truth: ALL non-disclosing agents share identical technical characteristics, whether they're legitimate AI assistants or malicious attackers.

Our multi-dimensional analysis framework reveals:

API Call Pattern Analysis

- All LLM agents make characteristic sequences (vision → reasoning → action)
- Framework detection through network timing analysis
- WebDriver properties that expose AI wrappers
- **Universal truth:** ChatGPT, Claude and fraud bots all show these same patterns

Infrastructure Analysis

- **Browser fingerprint manipulation:** Synthetic browser environments
- **OS impersonation:** Commonly presenting as macOS or Windows when running on cloud infrastructure
- **Headless automation:** Selenium, Puppeteer or headless browsers for automation
- **Device fingerprint mismatches:** Inconsistent WebGL, timezone and language signals

Critical insight: These technical signals appear in both legitimate AI assistants and malicious actors. A leading AI assistant helping users shop shows the same device spoofing as a credential stuffing bot. The technical layer alone cannot determine intent.

Proof of Work Device Validation

Our proof of work tests reveal the infrastructure truth for ALL agents:

- **Cloud vs. Device Reality:** All agents—legitimate or malicious—run on cloud infrastructure, not real devices
- **Performance Inconsistencies:** When any agent claims to be an "iPhone" but computes PoW at data center speeds
- **Universal spoofing:** ChatGPT helping a user and a fraud bot both show the same computational fingerprints
- **Infrastructure fingerprinting:** The technical layer reveals automation, not intent

Key finding: Every non-disclosing agent—whether it's helping users or attacking them—fails device authenticity checks. This is why behavioral analysis is essential.

Behavioral Biometrics (The Mouse Tells All)

Our proof of work tests reveal the infrastructure truth for ALL agents:

- **Movement Density:** Human sessions show dramatically more mouse events with continuous trajectories
- **Click Timing:** ALL AI agents (today)—helpful or harmful—exhibit superhuman precision with clicks faster than 10ms
- **Path Complexity:** Every agent shows limited position variation compared to human chaos
- **Teleportation behavior:** Sparse, direct movements vs human's continuous flow

Universal truth: These patterns are identical whether it's ChatGPT helping someone shop, Perplexity searching on a user's behalf, or a bot conducting fraud. The mouse doesn't lie about automation—but it can't tell you the automation's purpose.

Self-Disclosure as Ground Truth

Examining correlation between claimed identity and actual behavior:

- User agent taxonomy analysis
- IP reputation and ASN verification
- Device intelligence cross-referencing
- Timing and pattern correlation

The breakthrough insight: Self-disclosing agents (like OpenAI's published crawlers) exhibit THE EXACT SAME technical anomalies as all other agents:

- Same fake browsers and devices
- Same cloud infrastructure fingerprints
- Same superhuman behavioral patterns
- Same network inconsistencies

This makes self-disclosing agents perfect ground truth—they prove that technical anomalies alone don't indicate malicious intent. A legitimate OpenAI crawler and a credential stuffing bot are technically identical. The difference lies entirely in intent and authorization.

The Classification Framework We've Developed

Moving Beyond Technical Detection: The Arkose Labs Approach

Since all agents share identical technical profiles (fake devices, synthetic browsers), we've developed a classification system that relies on behavioral and contextual signals:

Step 1: Check Known Good Signatures

- Published IP ranges
- Verified headers
- Authenticated API keys

- **Result:** Catches all self-identifying good agents

Step 2: Behavioral Cohort Analysis

- Use machine learning to cluster similar behaviors
- Non-disclosing legitimate tools cluster with known good patterns
- Malicious patterns cluster with known fraud
- Gray area agents require additional analysis

- **Result:** Confirms intent through business context

Step 3: Business Logic Verification

- Does this agent's activity correlate with legitimate business events?
- Is there a corresponding support ticket, API call or customer interaction?

- **Result:** Identifies patterns in non-disclosing agents to determine likely intent

Step 4: Adaptive Challenge Deployment

For unknowns and non-disclosing agents:
- Deploy AI-resistant challenges
- Legitimate agents either solve correctly or back off respectfully
- Malicious agents reveal themselves through how they try to bypass or solve
- Gray area agents may require additional business context analysis

- **Result:** Forces classification through response behavior

AI-Resistant Mitigation Strategies: Arkose Labs Innovation

Making Attacks Economically Unviable Through Research-Driven Defenses

Our research and development team has created mitigation strategies that fundamentally break the economics of attacks:

LLM-Resistant Challenges

- Visual puzzles specifically designed to be extremely difficult for current AI
- Require understanding of context, not just pattern matching
- Force attackers to use expensive, specialized models

Multimodal Reasoning Requirements

- Combine audio, visual and text elements
- Force attackers to chain multiple API calls across different foundational models
- **Economic impact:** Dramatically increases attack costs

Dynamic Proof-of-Work

- Scale difficulty based on sophistication of detected solvers
- Adaptive chaining that compounds computational requirements
- Make each attempt progressively more expensive

Behavioral Verification Loops

- Continuous analysis throughout the session
- Cohort comparison in real-time
- Progressive trust scoring based on accumulated signals

What's At Stake: Real Impact We're Observing

The Impossible Choice (Without Proper Controls)

Our customer data reveals that without proper agent classification and management capabilities, companies face impossible decisions:

- **Block everything?** You just broke your payment processing, customer service and legitimate integrations
- **Allow everything?** You're going to get impacted by AI-powered attacks
- **No visibility?** You can't make informed decisions or adapt to emerging threats
- **No policy control?** You can't respond when your business needs change

Real-World Consequences Our Customers Report

Through our work with Fortune 500 companies, we're observing:

- **The identical technical profile problem:** ChatGPT helping a customer looks exactly like a fraud bot technically

- Companies accidentally blocking legitimate AI assistants while letting attackers through
- Organizations unable to distinguish between helpful and harmful automation using technical signals alone
- Significant portions of traffic showing fake devices—but which ones are threats?

The Hidden Cost: What Our Data Reveals

Based on our telemetry, the impact of universal technical similarity extends beyond traditional security:

- **Security paralysis:** Every technical anomaly could be ChatGPT or could be fraud—how do you decide?
- **False positive explosion:** Blocking fake devices eliminates both threats AND legitimate AI assistants
- **Metrics confusion:** All your "suspicious" traffic includes both helpful and harmful agents
- **Innovation blocking:** Fear of allowing any synthetic browsers stifles legitimate AI adoption
- **Compliance complexity:** Regulations assuming "real devices" when legitimate services use synthetic ones

Detailed industry-specific impact data available through consultation.

The Path Forward: Classification at Scale

The New Security Paradigm

Old question: "Is this a bot?"

New question: "Is this agent authorized to do what it's trying to do?"

This fundamental shift requires:

1. Classification, not just detection
2. Policy frameworks that match business needs
3. Real-time enforcement capabilities
4. Economic disruption of attack patterns
5. Intelligent routing based on intent, not just technical signatures

Implementation Strategy

Immediate:
- Deploy classification for self-identifying agents
- Establish baseline measurements of agent traffic

Short-term:
- Implement behavioral cohort analysis
- Define initial agent policies per endpoint

Medium-term:
- Build business correlation systems
- Enable dynamic policy switching

Long-term:
- Create comprehensive AI agent governance
- Develop predictive agent behavior models

The Arkose Labs Approach

We combine economic disruption with intelligent classification and flexible policy control:

- Force malicious actors to burn through costly API calls while letting legitimate automation flow
- Provide real-time classification of all agent types—self-disclosing, non-disclosing and malicious
- Enable instant policy changes from monitoring to blocking based on your risk tolerance
- Create an environment where each organization controls their agent ecosystem
- Leverage self-disclosing agents as ground truth for finding hidden agents
- Help organizations determine which non-disclosing agents to trust based on behavioral analysis

---

Your Action Plan: Key Questions to Get You Started

Based on our extensive research, these are the critical considerations for every security team:

1. Can you determine intent when all agents show identical technical anomalies?
2. Do you understand that blocking "fake devices" would eliminate ChatGPT along with fraud bots?
3. Can you correlate automated traffic with legitimate business events or user requests?
4. Do you have behavioral analysis beyond just technical fingerprinting?
5. Can you distinguish between agents helping users and agents harming them?
6. Are you prepared for a world where fake devices are the norm, not the exception?

Build Your Classification Matrix

Start instrumenting your platforms to collect signals that indicate potentially malicious agent activity (not just non-disclosure):

- Sudden changes in behavioral patterns without corresponding business changes
- Perfect navigation patterns with no exploration, optimized for extraction
- Superhuman response times and precision without legitimate purpose
- Traffic from residential proxies with automated characteristics and fraudulent intent signals
- User agents claiming to be major AI providers from suspicious IPs (spoofing)
- Systematic probing of authentication endpoints
- Patterns consistent with credential stuffing or account enumeration

Similarly collect signals that indicate legitimate agent activity (whether self-disclosing or not):

- Consistent behavioral patterns aligned with a clear service purpose
- Respect for rate limits and robots.txt
- Correlation with legitimate business events and API calls
- Appropriate response to challenges (solve correctly or back off gracefully)
- Predictable timing and endpoint access patterns
- Navigation patterns that provide value rather than extract it
- Presence during business hours or scheduled maintenance windows
- Activity that benefits end users (search, accessibility, customer service)

Conclusion: Leading the Industry Forward

Our research has uncovered a fundamental truth: all AI agents—legitimate or malicious—share identical technical fingerprints. They all use fake devices, synthetic browsers and cloud infrastructure. The companies that will thrive in the age of agentic AI aren't those trying to block all technical anomalies—that would eliminate both threats AND legitimate AI assistants.

Success requires understanding that ChatGPT helping a user shop and a bot conducting fraud look technically identical. The difference lies entirely in intent and authorization.

Through our extensive research and real-world deployments, Arkose Labs has developed the industry's most comprehensive approach to AI agent detection and classification. We've proven that success requires:

- **Behavioral analysis** that goes beyond technical signals to determine intent
- **Business context correlation** to identify legitimate use cases
- **Intent determination** when technical signals alone show only automation
- **Governance frameworks** that recognize technical similarity while enabling policy control
- **Sophisticated classification** that accepts technical anomalies as the new normal

The challenge isn't identifying fake devices—they're everywhere, in both good and bad traffic. The challenge is understanding which fake devices are helping your users and which are attacking them.

Learn More: Schedule a Deep Dive with Our Research Team

This report represents just a fraction of our research findings.
We would be happy to share our detailed findings with your team, including:

- **Why all agents look fake:** Deep technical analysis of why ChatGPT and fraud bots are technically identical
- **Intent detection methodologies:** How to determine purpose when technical signals are universal
- **Behavioral fingerprinting:** Patterns that reveal intent beyond device authenticity
- **Policy recommendations:** Managing a world where fake devices are everywhere
- **Future roadmap:** How agent disclosure standards and detection will evolve

---

Ready to Navigate the World Where Everything Looks Fake?

Contact Arkose Labs to schedule a comprehensive briefing with our research team.

BOOK A DEMO

Arkose Labs is the leading global provider offering a proactive fraud deterrence platform purpose-built to neutralize modern attacks, including those powered by Agentic AI and large language models (LLMs). Its comprehensive solution combines proprietary device identification (device ID), behavioral analysis, phishing protection, email intelligence, scraping prevention, API defense and bot management. Headquartered in San Mateo, California, the company maintains a global presence with offices throughout APAC, Central America, EMEA and South America.