Autonomous Market Intelligence for Science R&D: A Multi-Agent Configurable Web Service with Cross-Model Verification

Raminderpal Singh
raminderpal@hitchhikersai.org
raminderpalsingh.com | scienceclaw.ai

v1.1 (March 2026)

Abstract

Laboratory automation is a fragmented, fast-moving market spanning drug discovery, materials science, and agricultural technology. This paper describes an architecture for autonomous market intelligence covering any vertical market, designed as a scalable, production-level system that is LLM-agnostic and deployable with any sufficiently capable language model. The framework executes a daily four-stage research cycle (news scanning, topic processing, cross-synthesis, and quality retrospective) to build a cumulative knowledge base, then responds to on-demand queries through a multi-stage pipeline: natural-language intent extraction, knowledge-base retrieval with web search fallback, LLM-generated analysis with source attribution, and cross-model verification against a governance meta-guardrail. The verification architecture uses a separate language model to check every outbound communication for scope compliance, evidence standards, and tone. The demo instance runs 24/7 on a dedicated Apple Silicon M4 server, configured for lab automation intelligence. The same framework is now deployed as Thread 3 (Oncology Trial Insights) on the same server, configured for five oncology topics and producing weekly synthesised intelligence at scienceclaw.ai/oncology-insights.html with Kimi K2.5 cross-model verification. This paper describes the architecture, governance framework, and operational design choices that enable unattended continuous operation.

Keywords:
market intelligence, autonomous agents, cross-model verification, meta-guardrail, term discovery, LLM-agnostic, OpenClaw, oncology trial insights, KB-building, ScienceClaw

1. Introduction

The laboratory automation market serves three distinct but converging sectors: pharmaceutical drug discovery, materials science, and agricultural technology. Competitive intelligence in this space relies on static methods: periodic analyst reports, trade publication monitoring, conference attendance, and vendor-provided materials. No system continuously monitors the vendor landscape, accumulates structured knowledge, and delivers on-demand analysis grounded in that knowledge base.

We describe a framework for autonomous market intelligence that addresses this gap. The architecture is LLM-agnostic: the analysis and verification roles can be filled by any sufficiently capable language models, the topic registry can be reconfigured for any vertical market, and the infrastructure can be deployed on cloud servers, on-premises hardware, or hybrid configurations. The framework is built on the OpenClaw multi-agent framework and deployed as the ScienceClaw platform (scienceclaw.ai). It executes a daily research cycle to build a cumulative knowledge base, responds to on-demand email queries with sourced analysis, autonomously discovers new vendors and technologies, and governs all outputs through a meta-guardrail verified by an independent language model.

This paper describes the architecture through two instances: (1) a lab automation demonstration instance (Thread 1) configured for lab automation across drug discovery, materials science, and agricultural technology; and (2) Thread 3 (Oncology Trial Insights), a production deployment configured for the oncology clinical trial landscape, running in parallel with the Trial Intelligence anomaly detection system (Singh 2026) on the same dedicated server. Both threads share the OpenClaw framework, the email interface, and the server hardware, but maintain strict separation of knowledge bases, cron schedules, and LLM sessions.

2. Design Principles

The framework is guided by seven principles that distinguish it from conventional market research services and general-purpose AI assistants.

Evidence only. The system never fabricates data. Every factual claim in an outbound email must trace to a specific knowledge base document or web source. Missing data is flagged, not filled.

Scope-governed. A meta-guardrail document defines the system's remit. All outputs are verified against this scope before delivery.

Cross-model verification. The language model that generates analysis (GLM-5) is not the same model that verifies compliance (Kimi K2.5). This prevents the blind-spot problem of self-verification.

Deterministic computation, agentic reasoning. All data collection, scoring, and structured output are handled by deterministic Python scripts. The language model is used only for reasoning.

Email is the only interface. No dashboards, no web applications, no APIs to integrate. The user sends an email; the system replies in the same thread.

Autonomous self-improvement. The system discovers new terms, detects emerging entities, and proposes new topics. The daily retrospective stage reviews output quality and suggests improvements.

Graceful degradation. When components fail, the system degrades through defined tiers rather than producing silence. Every query gets a response.

3. Scope and Topic Registry

The topic registry defines the focus areas the system monitors. Each topic has an ID for internal matching, a human-readable label, and associated sectors. The registry is a JSON configuration file; replacing the vertical market requires editing this file and the meta-guardrail — no code changes.

3.1 Lab Automation Demo Instance (Thread 1)

ID	Label	Sectors
LIQUID-HANDLING	Liquid handling and dispensing	Drug discovery, materials, agritech
ROBOTIC-INTEGRATION	Robotic workcell integration	All three sectors
SCREENING-HTS	High-throughput screening platforms	Drug discovery, materials
LIMS-ELN	LIMS, ELN, and lab informatics	All three sectors
SAMPLE-MANAGEMENT	Sample storage, tracking, biobanking	Drug discovery, agritech
IMAGING-ANALYSIS	High-content imaging and analysis	Drug discovery, agritech
SELF-DRIVING-LABS	Autonomous / self-driving lab platforms	Materials, drug discovery
AGRITECH-AUTOMATION	Agricultural screening and phenotyping	Agritech

Table 1. Topic registry for the lab automation demo instance.

3.2 Thread 3: Oncology Trial Insights (Production Deployment)

Thread 3 is the production oncology deployment of the same framework. Five seed topics are configured for the oncology clinical trial landscape, with term discovery enabled to expand coverage autonomously as new drugs, mechanisms, and sponsors appear in the daily news flow.

ID	Label	Coverage
checkpoint_inhibitor_combinations	Checkpoint inhibitor combinations	PD-1/PD-L1 + novel agents, LAG-3, TIGIT
adc_pipeline	ADC pipeline and expansion	New targets, new indications, next-gen payloads
bispecific_antibodies	Bispecific antibody development	T-cell engagers, dual checkpoint bispecifics
biomarker_enrichment	Biomarker-driven enrichment	ctDNA, MRD, TMB, HRD, companion diagnostics
regulatory_endpoint_shifts	Regulatory endpoint guidance	Accelerated approval, surrogate endpoints, ODAC

Table 2. Topic registry for Thread 3 (Oncology Trial Insights). Seed topics; term discovery expands coverage.

4. Data Sources and Integration

The market intelligence workflow draws from unstructured news coverage, vendor communications, industry publications, and academic literature. No commercial data subscriptions are required.

Source	API	Role in Workflow
Brave Search	REST API	Primary: daily news scan and on-demand fallback
PubMed E-utilities	REST API	Lab automation papers, platform evaluations
bioRxiv / medRxiv	Content API	Preprints on autonomous lab methods
Vendor websites	Brave Search (targeted)	Product catalogues, press releases
Industry trade media	Brave Search (targeted)	Lab Manager, SelectScience, Drug Discovery Today

Table 3. Data sources for the market intelligence framework. Brave Search is the primary source.

Brave Search serves as the primary data source. Market intelligence has no equivalent structured source to the clinical trial registries used by Thread 2: vendor activity is reported through press releases, trade publications, analyst commentary, and news coverage. For Thread 3, the same Brave Search integration scans oncology news, conference readouts (ASCO, ESMO, ASH, AACR), regulatory announcements, and sponsor pipeline disclosures, tagged to the five oncology topic areas. Confirmed trial findings from Thread 2 (findings.json) are

ingested daily as Tier 1 evidence, enriching the synthesis with structured anomaly data.

5. Customisation and Deployment

Nothing in the framework couples to the demo instance's specific technology choices. Four dimensions of customisation are supported independently.

5.1 Vertical Market

Replacing the vertical market requires editing two configuration files: topics.json and meta-guardrail.md. No code changes are needed. Thread 3 demonstrates this: the lab automation registry was replaced with five oncology topics in under one hour, with term discovery seeded from oncology trial vocabulary.

5.2 Language Models

The architecture requires two model slots: an analysis model and a verification model. Both accept any model accessible via an OpenAI-compatible API. The sole constraint is that these must be different models to ensure genuine independence in verification. Compatible models include Claude (Anthropic), GPT-4o (OpenAI), Llama 3 (Meta), Qwen 2.5 (Alibaba), DeepSeek-V3, Mistral Large, and GLM-5 (Zhipu). The demo instance uses GLM-5 for analysis and Kimi K2.5 for verification.

5.3 Infrastructure

The demo instance runs on a single Apple Silicon M4 server (48 GB) with macOS launchd for process supervision. The architecture supports cloud VMs, containerised environments, on-premises enterprise servers, or hybrid configurations.

Component	Demo Instance	Enterprise Alternative
Analysis model	GLM-5:cloud (Ollama)	Claude, GPT-4o, Llama 3, or any OpenAI-compatible
Verification model	Kimi K2.5 (Ollama Cloud)	Any model different from the analysis model
Server	macOS, Apple Silicon M4 48 GB	Linux VM, Docker container, Kubernetes pod
Process supervision	macOS launchd	systemd, Docker restart policy, K8s probes
Email interface	AgentMail (cloud)	Corporate Exchange, internal SMTP, Slack
Web search	Brave Search API	Bing API, Google Custom Search
Topic registry	8 lab automation / 5 oncology topics	Customised to any vertical (8–15 topics)

Table 4. Deployment configurations. Each component can be substituted independently.

6. Daily Autonomous Research Cycle

A four-stage cron-driven workflow runs every morning, building and maintaining the knowledge base. Each stage runs as a fresh, isolated OpenClaw session with no context carried between stages. Handoffs are file-based (JSON state files, markdown reports), preventing unbounded token growth.

Time (UK)	Stage	Agent	What Happens
09:00	News scan	dispatcher	Brave Search for past-24h news; tag stories to topics; extract new terms
09:15	Project processing	dispatcher	Two-pass research per topic: direct analysis, then cross-sector connecti
10:30	Cross-sector synthesis	synthesiser	Read all topic outputs; find connections, contradictions, emerging patter
11:00	Retrospective	dispatcher	Review quality; detect empty topics; propose improvements; generate to

Table 5. Daily research cycle (Thread 1 — lab automation demo). Thread 3 runs 04:45–08:00 on the same schedule pattern.

6.1 Thread 3: Oncology Trial Insights Daily Cycle

Thread 3 runs a parallel daily cycle before Thread 2 begins (04:45–08:00 London), ensuring no concurrent GLM-5 calls. The cycle adds a trial findings ingest step that reads confirmed anomalies from Thread 2 as Tier 1 evidence, and an insight record generator that extracts structured records from the daily synthesis for publication to scienceclaw.ai/oncology-insights.html.

Time (UK)	Stage	Script / Agent	What Happens
04:45	KB maintenance	oncology_kb_maintenance.py	Hot/warm/cold rotation; log rotation
05:00	Trial findings ingest	ingest_trial_findings.py	Reads Thread 2 findings.json; writes trial-finding KB entries
05:15	News scan	oncology_news_scan.py	Brave Search across 5 oncology topics; writes news KB entri
06:00	Topic analysis	oncology_topic_analysis.py	Two-pass GLM-5 analysis per topic (Pass 1: direct; Pass 2: c
07:00	Cross-synthesis	oncology_cross_synthesis.py	GLM-5 synthesises themes, trial-news connections, gaps; qu
07:30	Insight generation	generate_insight_records.py	Deterministic extraction + focused GLM-5 call per theme; evi
08:00	Retrospective	oncology_retrospective.py	Quality review (GLM-5); term discovery proposals (max 3/day

Table 6. Thread 3 (Oncology Trial Insights) daily cycle. Runs 04:45–08:00, before Thread 2.

6.2 Weekly Output

Every Friday at 09:00 London, the weekly page generator reads all insight records from the past 7 days and publishes an HTML table to scienceclaw.ai/oncology-insights.html. At 09:30, the email distribution script runs Kimi K2.5 verification on the page content before sending to the recipient list. A verification verdict of pass or warn allows sending (warn appends a flag footer). A fail verdict holds the email and sends an operator alert instead. The quality gate is set at 0.7 sustained for 4+ consecutive retrospective runs before email distribution to external recipients is enabled.

7. Knowledge Base Structure

7.1 Thread 1 (Lab Automation)

Path	Content	Written By	Read By
topics.json	Curated 8-topic registry	Manual	All workflows
term-registry.json	Auto-discovered vendors, products, technologies	Stage 1 term discovery	Intent extraction, news scan
projects/*.md	Market briefs per topic, cross-sector synthesis	Stages 2 and 3	On-demand queries
news-scan/*.md	Daily tagged news briefings	Stage 1	Stage 2
retrospectives/*.md	Daily quality reviews	Stage 4	Future retrospectives

Table 7. Knowledge base file layout for Thread 1.

7.2 Thread 3 (Oncology Trial Insights)

Path	Content	Retention
oncology-insights/daily/{date}/	News KB entries + ingested trial findings	Hot: 14 days
oncology-insights/warm/{date}/	Entries aged from hot tier	Warm: days 15–60
oncology-insights/archive/{date}/	Cold storage	Cold: 60+ days
oncology-insights/synthesis/{date}-synthesis.json	Daily synthesis note + quality score + retrospective	Indefinite
oncology-insights/insights/{date}-{n}-{theme}.json	Structured insight records (tab 2b rows)	Indefinite
oncology-insights/terms/proposals/{date}.json	Term discovery proposals (operator review)	Until approved/rejected
oncology-insights/topics.json	Active topic registry + term discovery vocabulary	Indefinite
oncology-insights/meta-guardrail.md	Governance document for Thread 3	Indefinite

Table 8. Knowledge base structure for Thread 3.

8. On-Demand Query Pipeline

The on-demand pipeline is triggered by an inbound email to scienceclaw@agentmail.to with [MARKET RESEARCH] in the subject line. The pipeline processes free-text queries through five stages.

8.1 Intent Extraction and Entity Recognition

A preprocessing function sends the raw subject and body to GLM-5 with a constrained prompt that returns structured JSON: intent classification, extracted entities, inferred topic IDs, sectors, time sensitivity, confidence level, and ambiguity notes. Inferred topic IDs are validated against topics.json deterministically.

8.2 Confidence Routing

High confidence: proceed directly into KB retrieval. Medium confidence: proceed but prepend a scope assumptions block. Low confidence: pick the most plausible interpretation, proceed, and state the assumption prominently.

8.3 KB Retrieval with Web Search Fallback

For each matched topic ID, the pipeline loads the most recent market brief and cross-sector synthesis. If KB retrieval returns fewer than ~500 words of relevant content, a Brave Search fallback is triggered. Search queries are constructed from extracted entities, not raw user text.

8.4 LLM Analysis with Source Attribution

The analysis call (GLM-5, max tokens 3600, temperature 0.3) receives the meta-guardrail as system prompt, plus numbered KB excerpts, numbered web results, parsed intent, and the raw email body. Post-processing validates that citation markers reference real documents.

8.5 Cross-Model Verification

The draft reply is verified by Kimi K2.5 against the full meta-guardrail before sending. See Section 9 for the verification architecture and fallback ladder.

9. Cross-Model Verification Architecture

A central architectural decision is that the model generating analysis is never the model verifying compliance. This cross-model design addresses the fundamental weakness of self-verification: the same model has the same biases in both passes. In the demo instance, GLM-5 generates analysis and Kimi K2.5 verifies; Thread 3 uses the same pairing for weekly insights email verification.

Check	What It Catches	Example Violation
Scope compliance	Out-of-scope topics	Drug pricing in an oncology insights email
Evidence attribution	Unattributed factual claims	Specific efficacy data with no NCT ID or citation
Tone and register	Marketing language, superlatives	'Revolutionary breakthrough', 'game-changing'
Fabrication signals	Specificity without source	Clinical outcomes data with no trial reference
Evidence mismatch	Wrong trial cited for claim	Melanoma trial cited for NSCLC efficacy claim

Table 9. Verification checks performed by Kimi K2.5 on every outbound communication.

9.1 Autonomous Fallback Ladder

Tier	Condition	Action
1. Full analysis	Draft passes or warns on first verification	Send as-is (pass) or send with flag footer (warn)
2. Rewrite	Draft fails first verification	GLM-5 rewrites fixing only flagged issues; Kimi K2.5 re-verifies
3. Rewrite accepted	Rewrite passes or warns	Send rewritten version
4. Strip to safe	Rewrite still fails	Remove block-severity sentences; send with coverage note
5. KB excerpts only	Stripped version < 3 sentences	Send raw KB excerpts with dates, no LLM interpretation

Table 10. Autonomous fallback ladder. Every query receives a response. No human intervention required.

10. Autonomous Term Discovery

The system autonomously discovers and registers new entities from the daily news scan. Term discovery is distinct from topic discovery: it enriches the vocabulary used to recognise, tag, and cross-reference entities within existing topics.

Thread 3 term discovery identifies new drugs, mechanisms, sponsors, and indications appearing in the oncology news flow that don't match any existing topic or vocabulary term. Each proposal requires two independent source signals from the day's KB. A maximum of three proposals per day are written to the proposals directory for operator review — auto-approval is disabled by default.

Term addition constraints (meta-guardrail governed):

- Only add terms with a clear, specific referent; reject vague terms ('next-gen platform').
- Drug names must appear in a context where they are being tested in a clinical trial or have regulatory status.
- Product names must be specific (brand names, compound codes); generic descriptions are not product names.
- Technology keywords must describe a specific mechanism or capability.
- Maximum 3 new term proposals per day; each requires 2 independent source signals.

11. Meta-Guardrail Governance

Each thread has its own meta-guardrail document that governs scope, permitted behaviour, evidence standards, and tone. The guardrail is loaded into every LLM session and is the reference document against which Kimi K2.5 verifies all outbound communications.

Thread 3 guardrail scope covers: competitive moves, mechanism trends, regulatory endpoint shifts, biomarker adoption, sponsor strategy, and conference readouts. Hard exclusions: general oncology news without a trial connection, patient outcomes reporting, drug pricing, and basic science without a clinical trial link. On-demand queries are not enabled for Thread 3 at launch.

12. System Architecture

Component	Demo Instance	Role (configurable)
Agent framework	OpenClaw	Multi-agent orchestration, cron, session isolation
Analysis LLM	GLM-5:cloud via Ollama	Any OpenAI-compatible LLM with sufficient context window
Verification LLM	Kimi K2.5 via Ollama Cloud	Any independent LLM (must differ from analysis model)
Email I/O	AgentMail	Dedicated inbox with webhook + REST API
Web search	Brave Search API	Daily news scan + on-demand fallback
Webhook server	Flask (Python 3.12)	Async email processing, background threads
Networking	Tailscale Funnel	Stable public HTTPS tunnel, survives reboots
Observability	Structured JSONL logging	LLM call logs, performance metrics
Server	macOS, Apple Silicon M4 48 GB	Any server; cloud, on-prem, or hybrid

Table 11. Technology stack. Analysis and verification models can be swapped for any capable LLM.

12.1 Daily Processing Timeline

Thread 3 (Oncology Trial Insights) runs 04:45–08:00 London. Thread 2 (Trial Intelligence) runs 11:45–14:35 London. Thread 1 (Market Research) runs 09:00–13:00 London. All three threads share the same server and GLM-5 endpoint; cron jobs are staggered to prevent concurrent LLM calls. On-demand queries (Thread 1) are processed as they arrive, subject to GLM-5 availability.

13. Illustrative Scenarios

13.1 Vague Query with Successful Interpretation

A user emails: "What's happening with automated screening these days?" Intent extraction classifies this as news_recap with medium confidence, infers SCREENING-HTS, and tags drug discovery and materials sectors. The response opens with a scope caveat then loads the most recent SCREENING-HTS market brief. Kimi K2.5 verification passes.

13.2 Out-of-Scope Query Intercepted by Verification

A user emails: "What's the competitive landscape for oral GLP-1s?" GLM-5 produces a draft that includes three paragraphs analysing the GLP-1 therapeutic landscape. Kimi K2.5 flags these as scope violations (block severity). GLM-5 rewrites, removing the therapeutic analysis and focusing on: "Increased GLP-1 screening activity is driving demand for high-throughput liquid handling systems." Re-verification passes.

13.3 Thread 3: Evidence Mismatch Caught by Verification

On the first production run of Thread 3, Kimi K2.5 identified that NCT03698019 (a melanoma perioperative trial) was being cited as evidence for sacituzumab govitecan NSCLC claims. The insight record was corrected: the misattributed NCT IDs were removed and confidence was downgraded to medium. The verification gate held the email until the correction was made, demonstrating the system working as designed.

14. Output Format

Thread 1 on-demand responses include: a freshness metadata line, scope assumptions block (if confidence was medium or low), 3–5 bullet key takeaways, Evidence and Risks/Unknowns sections with inline citations, at least one comparison table, a Next Steps section, and a performance footer.

Thread 3 weekly output is a sortable HTML table at scienceclaw.ai/oncology-insights.html with columns: Date, Theme, Insight, Evidence (green NCT ID pills linking to ClinicalTrials.gov, blue news URL pills), Confidence, and Detail (expandable significance text). Filter buttons by theme category. Stats bar: total insights, themes covered, trial findings cited, news sources cited. The same table is emailed weekly after Kimi K2.5 compliance verification.

15. Limitations

Single server. The demo instance runs on one MacBook Pro with no redundancy. A hardware failure takes the service offline. This is a demo deployment constraint, not an architectural one.

Dual external LLM dependency. Both GLM-5 and Kimi K2.5 are accessed via Ollama Cloud. A platform-level outage takes down analysis and verification simultaneously. Production deployments should use models from different providers.

News-dependent intelligence quality. Market intelligence depends on Brave Search coverage. On quiet news days, topics receive thin or empty briefs. Cold-start quality is expected to be low (Thread 3 quality score target: 0.7+ after 4 weeks of KB accumulation).

Guardrail is prompt-based. The meta-guardrail is enforced through prompt instructions (soft) and cross-model verification (harder). Neither mechanism provides formal guarantees.

Anti-hallucination is not verified end-to-end. Source attribution is validated by checking that citation markers reference real documents, but neither mechanism verifies that a cited claim actually appears in the cited document.

Thread 3 cold-start. The first weeks of Thread 3 operation will produce thin insights as the KB accumulates. The retrospective quality score gates email distribution — emails to external recipients should not begin until quality sustainably exceeds 0.7.

16. Discussion

This framework represents a shift from reactive market research to continuous autonomous intelligence. The cross-model verification architecture is the most novel contribution: using a separate model for verification introduces genuine independence — different training data, different biases, different failure modes. The cost is operational complexity and additional latency (3–5 seconds per verification call). In practice, the verification call is the cheapest in the pipeline because its context is small (~4K tokens versus ~20K for the analysis call).

Thread 3 demonstrates the framework's configurability. The same architecture that monitors lab automation vendor news now synthesises oncology trial intelligence, consuming Thread 2's anomaly findings as Tier 1 evidence. The two threads operate on the same server with staggered schedules, sharing infrastructure but maintaining complete separation of KB, governance, and LLM sessions.

A contrary viewpoint deserves acknowledgment: a competent analyst with access to the same news sources could produce higher-quality intelligence. The system's advantage is consistency and coverage, not analytical judgment. The meta-guardrail prevents the system from making strategic judgments it is not qualified to make — the user supplies that judgment.

17. Data and Code Availability

The system is implemented using the OpenClaw multi-agent framework (open source) with an LLM-agnostic architecture. The demo instance uses GLM-5:cloud for analysis and Kimi K2.5 for verification. The email interface uses AgentMail. All data sources are publicly available. No proprietary data subscriptions are required. The meta-guardrail document, topic registries, term registry schema, stage instruction file templates, and deployment configuration guides are available from the corresponding author on request. The demo instance is deployed at scienceclaw.ai.

References

1. Singh R. Baseline-Deviation Analysis for Automated Detection of Anomalous Design Choices in Oncology Clinical Trials. v0.4 (March 2026). [Thread 2 companion paper.]
2. Abolhasani M, Kumacheva E. The rise of self-driving labs in chemical and materials sciences. Nat Synth. 2023;2:483–492.
3. Burger B, Maffettone PM, Gusev VV, et al. A mobile robotic chemist. Nature. 2020;583:237–241.
4. Holland I, Davies JA. Automation in the life science research laboratory. Front Bioeng Biotechnol. 2020;8:571777.
5. King RD, Rowland J, Oliver SG, et al. The automation of science. Science. 2009;324(5923):85–89.
6. Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: a survey. arXiv:2309.07864. 2023.
7. Zhuge M, Liu H, Faccio F, et al. Agent-as-a-judge: evaluating agents with agents. arXiv:2410.10934. 2024.