EIREL: Decentralized Multimodal AI on Bittensor

Subnet 36 | Bittensor | Version 1.0 — April 2026

Abstract

Eirel is a decentralized multimodal agent subnet on Bittensor designed to transform a single user prompt into complete execution.

The network supports a full spectrum of capabilities across language, web, code, and media. Users can perform general conversation, real-time web search, deep research, image generation, video creation, audio processing, and more — all unified within a single system that understands intent and executes across modalities.

Beyond isolated features, Eirel is built for end-to-end execution. With one prompt, users can direct agents to reason, create, code, research, build applications, launch websites, connect backend systems, and turn ideas into real, usable products. The system coordinates these steps automatically, removing the need for fragmented tools or manual workflows.

By combining multimodal intelligence, tool use, and competitive agent development, Eirel creates an open ecosystem where specialized agents continuously improve in capability, accuracy, and real-world usefulness.

At its core, Eirel aims to become the execution layer for multimodal AI on Bittensor — enabling users to move from idea to deployment through natural language, while aligning network incentives around real utility and measurable outcomes.

---

1. Problem

1.1 Centralized Control

The most capable AI assistants — ChatGPT, Claude, Gemini — are controlled by a handful of companies. Users depend on their pricing, availability, safety policies, and strategic decisions. When OpenAI throttles usage or Anthropic changes behavior, entire workflows break. Enterprise surveys consistently cite vendor lock-in as a top concern.

1.2 The Quality Floor in Decentralized AI

Existing decentralized AI networks compete on cost rather than capability, producing a race to the bottom. Low-effort wrappers around commodity models crowd out genuine innovation. Without rigorous evaluation, there is no mechanism to separate agents that produce real value from those that produce convincing-looking but unreliable output.

1.3 Broken Benchmarks

AI benchmarks are fundamentally compromised. SWE-bench Verified scores reached 81%, but the contamination-resistant Pro variant drops to 46% — half of measured "capability" is memorization. OpenAI stopped reporting Verified scores after finding training data contamination in every frontier model. The industry needs evaluation designed for adversaries, not cooperative participants.

1.4 Design Principles

- **Competition over centralization.** Multiple independent agents compete continuously; the best earns its position through verified performance.
- **Conversation-native.** Users interact through a streaming chat interface, not a job submission queue.
- **Specialization + composition.** A subnet-owned orchestrator delegates to competing specialist families when deeper capability is needed.
- **Verification over trust.** All results are independently verified using hidden test suites, claim-level verification, and ensemble LLM judges.
- **Process over output.** Agents are scored on their working process — not just final output.

---

2. Solution

2.1 A Unified Execution Layer

Eirel builds a single decentralized system that converts natural language prompts into fully executed outcomes. Rather than offering fragmented point solutions, Eirel unifies multimodal capabilities — language, code, web interaction, image generation, video, audio, and data processing — under one orchestration layer. Users issue a single prompt; the system handles intent classification, task decomposition, specialist delegation, and result integration automatically.

2.2 Competitive Specialization Over Monolithic Control

Instead of relying on one company to build every capability, Eirel creates a competitive marketplace where independent developers (miners) build specialist agents across distinct capability families. Each family runs its own continuous competition — the highest-scoring agent earns the right to serve real user traffic and receives TAO emissions. This competitive pressure ensures that capabilities improve continuously, driven by economic incentives rather than a single vendor's roadmap.

2.3 Verification-First Evaluation

Eirel replaces self-reported benchmarks with an evaluation framework designed for adversarial participants. Hidden test suites, claim-level verification, ensemble LLM judges, and anti-gaming detectors ensure that measured performance reflects genuine capability.

Miners never see evaluation tasks in advance and cannot submit different code for evaluation versus production.

2.4 Subnet-Owned Orchestration with Decentralized Capabilities

The system is split into two layers: a subnet-owned orchestrator that controls the user experience, and a decentralized specialist layer where miners compete on deep capabilities. This hybrid design delivers the responsiveness of centralized assistants while preserving the quality benefits of open competition.

2.5 Aligned Incentives Through Token Economics

Consumer revenue flows into alpha token buyback-and-burn, creating deflationary pressure proportional to real demand. Miners earn TAO emissions by winning specialist competitions. The result is a self-reinforcing system where genuine utility — not speculation — drives network value.

---

3. Architecture

3.1 Participants

Role	Responsibility
Owner	Operates subnet infrastructure: evaluation design, miner deployments, scoring coordination, on-chain weight publication.
Validator	Independently scores all candidate miners using owner-frozen evaluation bundles. Submits signed scores for stake-weighted consensus.
Miner	Submits AI agents that compete within a specialist family. Agents are deployed on subnet infrastructure.
Consumer	End users who interact with Eirel through the streaming interface. Queries are served by the current top-scoring agents.

3.2 Four-Layer Architecture

Eirel's architecture has four layers. A simple chat message touches only two services (gateway + orchestrator), while complex tasks dynamically engage specialist families.

The orchestrator and platform tools are subnet infrastructure — like ChatGPT's brain and built-in tools. But unlike ChatGPT, every specialist capability behind the orchestrator is developed by competing independent miners. The subnet controls the conversation experience; the market controls the quality of deep capabilities.

3.3 Two-Tier Tool Model

The orchestrator uses a **two-tier tool model** — fast platform utilities for simple tasks and deep specialist capabilities for complex work.

Tier 1 — Platform Tools are subnet-owned infrastructure for quick utility tasks: code execution, web search, file management, image generation, and memory recall.

Tier 2 — Specialist Families are competition-driven deep capabilities where quality variation matters and economic incentives drive improvement:

Family	Role
Analyst	Evidence-grounded research and synthesis with verified citations
Builder	Autonomous code generation and project delivery
Verifier	Quality auditing with structured verdicts
Browser	Autonomous web navigation and content extraction
Data	Data extraction, transformation, and visualization
Media	Image, audio, video, and multimodal asset generation
Memory	Cross-session context persistence and RAG retrieval
Planner	Task decomposition and workflow planning

The orchestrator routes each request to the right tier automatically — quick lookups go to platform tools, deep tasks go to competing specialists.

3.4 Shared State and Persistence

Eirel maintains persistent state across conversations: full message history, user profiles and preferences, generated artifacts (files, code, images, reports), and cross-session memory. This enables continuity across interactions and long-running workflows.

3.5 Trust Model and Decentralization

Eirel launches with an **owner-operated** trust model — evaluation bundles are frozen and immutable, all candidates receive identical bundles, and hidden test suites are never exposed. The roadmap progresses toward validator-run evaluation and distributed task generation.

---

4. Evaluation

4.1 Evaluation Principles

Eirel's evaluation operates on three principles that no existing benchmark implements simultaneously:

1. **Multi-dimensional scoring** — Continuous scores across 6+ dimensions per family, measuring both output quality and working process.
2. **Owner-verified trustless execution** — Miners never see test suites, never control infrastructure, and cannot run different code during evaluation vs production.

3. **Active anti-gaming** — Detectors that assume adversarial miners, not cooperative benchmark participants.

4.2 Evaluation Flow

The evaluation cycle begins when the owner freezes an immutable evaluation bundle containing tasks, rubrics, and judge configurations. Validators then invoke all candidate deployments against these benchmark tasks. Responses are scored through a multi-dimensional framework, followed by anti-gaming detection using 12+ detectors. Results pass through protocol compliance checks and calibration gates spanning 80+ metrics. Once validated, the weight setter normalizes scores, signs them, and submits to the Bittensor chain. TAO emissions then flow to the top-performing miners in each specialist family.

4.3 Three-Pillar Scoring Model

Every miner's score is composed of three pillars:

Pillar	Weight Range	Measures
Capability	0.70-0.76	Family-specific quality dimensions
Robustness	0.16-0.18	Cross-task consistency and perturbation resistance
Anti-gaming	0.08-0.12	Template detection, fabrication, memorization

4.4 Anti-Gaming Detection

Eirel implements 12+ family-specific detectors targeting trace fabrication, template responses, memorization, citation recycling, code obfuscation, and hardcoded answers. Trace fabrication — where an agent lies about its own results — zeroes the entire score. Perturbation testing uses prompt variants to verify genuine understanding.

4.5 Calibration and Judging

Miners must pass calibration gates across 80+ metrics with per-family thresholds before promotion. A consistency gate requires 2 passes in the last 3 evaluation runs — no promotion from a single lucky run. For subjective quality dimensions, an ensemble of two independent LLM judges provides scoring with disagreement detection and deterministic fallback.

---

5. Token Economics

5.1 Value Flow

Consumers pay per task, and that revenue flows into the subnet treasury. From there, funds split in two directions: one stream pays LLM providers for inference costs, while the other buys alpha tokens on the open market and permanently burns them. The burn

creates deflationary pressure that benefits token holders, proportional to real consumer demand. Meanwhile, TAO emissions from the Bittensor network flow separately to specialist miners as rewards for winning their family competitions.

5.2 Revenue Model

Revenue Source	Covers
TAO emissions (from Bittensor)	Miner rewards — primary incentive during growth
Miner submission fees	Evaluation infrastructure costs
Consumer task payments	LLM inference for serving + alpha buyback-and-burn

The owner pays for all LLM inference through a provider proxy, removing the capital barrier for developers with strong agents.

5.3 Buyback-and-Burn

Consumer revenue buys alpha tokens on the open market and permanently burns them. This creates deflationary pressure proportional to consumer demand, aligning incentives: consumers want quality agents, miners want TAO emissions, and alpha token holders want sustained demand driving buybacks.

5.4 Emission Distribution

All TAO emissions flow to specialist families — the orchestrator and platform tools are subnet infrastructure with no emission share. Each specialist family's winner receives 100% of that family's emissions.

---

6. Competitive Position

6.1 vs Centralized Assistants

Aspect	ChatGPT / Claude / Gemini	EIREL
Control	Single vendor	Decentralized, emission-driven
Competition	None — use what the vendor ships	Continuous, multiple competing agents
Deep capabilities	One vendor's implementation	Competition-winning specialists per family
Verification	Self-evaluated	Owner-verified, hidden tests, ensemble judges
Extensibility	Vendor-controlled plugins	Open MCP ecosystem (12,000+ tool servers)
Transparency	Closed models, opaque eval	Open SDK, published scoring methodology

6.2 vs Other Bittensor Subnets

Aspect	Typical LLM Subnets	EIREL
Evaluation	Single-dimension scoring	Three-pillar, multi-dimensional, owner-verified
Gaming resistance	Limited	12+ detectors + claim verification
Miner trust	Miner-hosted endpoints	Owner-hosted containers from submitted archives
Task complexity	Single-turn chat	Streaming conversation + autonomous multi-day projects
User experience	API-only	Streaming chat interface with conversation memory

6.3 Honest Trade-Offs

Trade-Off	Mitigation
Owner centralization at launch	Decentralization roadmap: validator-run evaluation by Q3 2026
Higher evaluation complexity	Full scoring breakdown published per epoch
Unproven at scale	Phased rollout, calibration gates require cross-epoch consistency
LLM judge dependency	Ensemble dual-judge with disagreement detection + deterministic fallback

---

7. Roadmap

Phase 1: Launch (Q2 2026)

- Subnet-owned orchestrator + streaming conversation gateway
- Platform tools: code execution, web search, file management, image generation
- Three specialist families: analyst, builder, verifier
- Winner-take-all serving model with automatic failover
- Owner-frozen evaluation with hidden test suites and anti-gaming detectors
- A2A protocol interoperability

Phase 2: Expansion (Q3-Q4 2026)

- Activate browser, data, media, and planner specialist families
- MCP ecosystem support for miners
- User profiles, custom instructions, and memory persistence
- Cross-family workflow scoring
- Consumer payment integration + alpha token buyback-and-burn

Phase 3: Decentralization (Q1-Q2 2027)

- Validator-run evaluation on independent infrastructure
- Distributed task generation with stake-weighted contributions
- Cross-epoch behavioral fingerprinting
- Community-contributed evaluation tasks

---

8. Conclusion

The AI assistant market is dominated by centralized providers, while decentralized alternatives compete on cost rather than quality. Every major AI benchmark is contaminated, gamed, or both.

Eirel solves both problems. A subnet-owned orchestrator provides the conversational experience, backed by decentralized competition that continuously improves the specialist capabilities powering it. Every serving specialist earned its position through multi-dimensional evaluation with hidden test suites, claim-level verification, and anti-gaming detection. Consumer revenue drives alpha token buyback-and-burn, creating a sustainable economic loop between users, miners, and token holders.

The path forward: launch with platform tools and three specialist families, expand to eight families, decentralize evaluation, and build the system where genuine capability is always the path of least resistance.

References

Market and Industry

- Grand View Research — AI Agents Market 2025-2033 — $7.63B market, 49.6% CAGR
- Gartner — 40% of Enterprise Apps Will Feature AI Agents by 2026
- LangChain — 2026 State of AI Agents — 57% of organizations have agents in production
- MarketsandMarkets — Autonomous AI and Autonomous Agents Market — Independent market sizing for autonomous AI agents
- McKinsey — The Economic Potential of Generative AI — Estimates $2.6-4.4T annual economic impact from generative AI

Bittensor Ecosystem

- Bittensor Network — Decentralized AI network protocol
- Bittensor Whitepaper — Yuma Rao (2021) — Foundational protocol design for incentivized decentralized machine intelligence
- CoinDesk — Bittensor Ecosystem Hits $1.5B
- Messari — State of Bittensor Q1 2025 — Analysis of TAO tokenomics and subnet incentive structures

Decentralized AI and Data Networks

- Ocean Protocol Whitepaper — Decentralized data exchange protocol enabling AI data marketplaces

Multimodal AI Systems

- GPT-4 Technical Report — OpenAI (2023) — Multimodal large language model accepting image and text inputs
- Gemini: A Family of Highly Capable Multimodal Models — Google DeepMind (2023) — Natively multimodal model across text, image, audio, and video

AI Agent Frameworks and Orchestration

- AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation — Wu et al. (2023) — Microsoft's multi-agent conversation framework for complex tasks
- CrewAI — Multi-Agent Orchestration Framework — Framework for orchestrating role-playing autonomous AI agents
- LangChain Documentation — Leading open-source framework for building LLM-powered agent applications

Benchmark Integrity

- SWE-bench Leaderboard — Top: 80.9%; Pro variant drops to 46%
- OpenAI — Why We No Longer Evaluate SWE-bench Verified
- EU AI Watch — Nine Challenges in AI Benchmarking
- Do ImageNet Classifiers Generalize to ImageNet? — Recht et al. (2019) — Demonstrates performance drops on new test sets, foundational benchmark integrity work
- Stop Uploading Test Data in Plain Text — Jacovi et al. (2023) — Systematic benchmark contamination via web crawling

Evaluation Research

- METR — Model Evaluation and Threat Research — Autonomous AI capability assessment
- CALM — Quantifying Biases in LLM-as-a-Judge
- AWS — Evaluating AI Agents
- Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena — Zheng et al. (2023) — Foundational paper on using LLMs as evaluators
- Large Language Models are not Fair Evaluators — Wang et al. (2023) — Documents positional and verbosity biases in LLM-based evaluation

Autonomous Coding Agents

- SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering — Yang et al. (2024) — Agent architecture for autonomous software engineering
- Cognition Labs — Introducing Devin (2024) — First fully autonomous AI software engineer

Protocols

- Google A2A Protocol — Agent-to-Agent interoperability
- Model Context Protocol (MCP) — Tool ecosystem standard
- Model Context Protocol Specification — Full technical specification for MCP tool interoperability

---