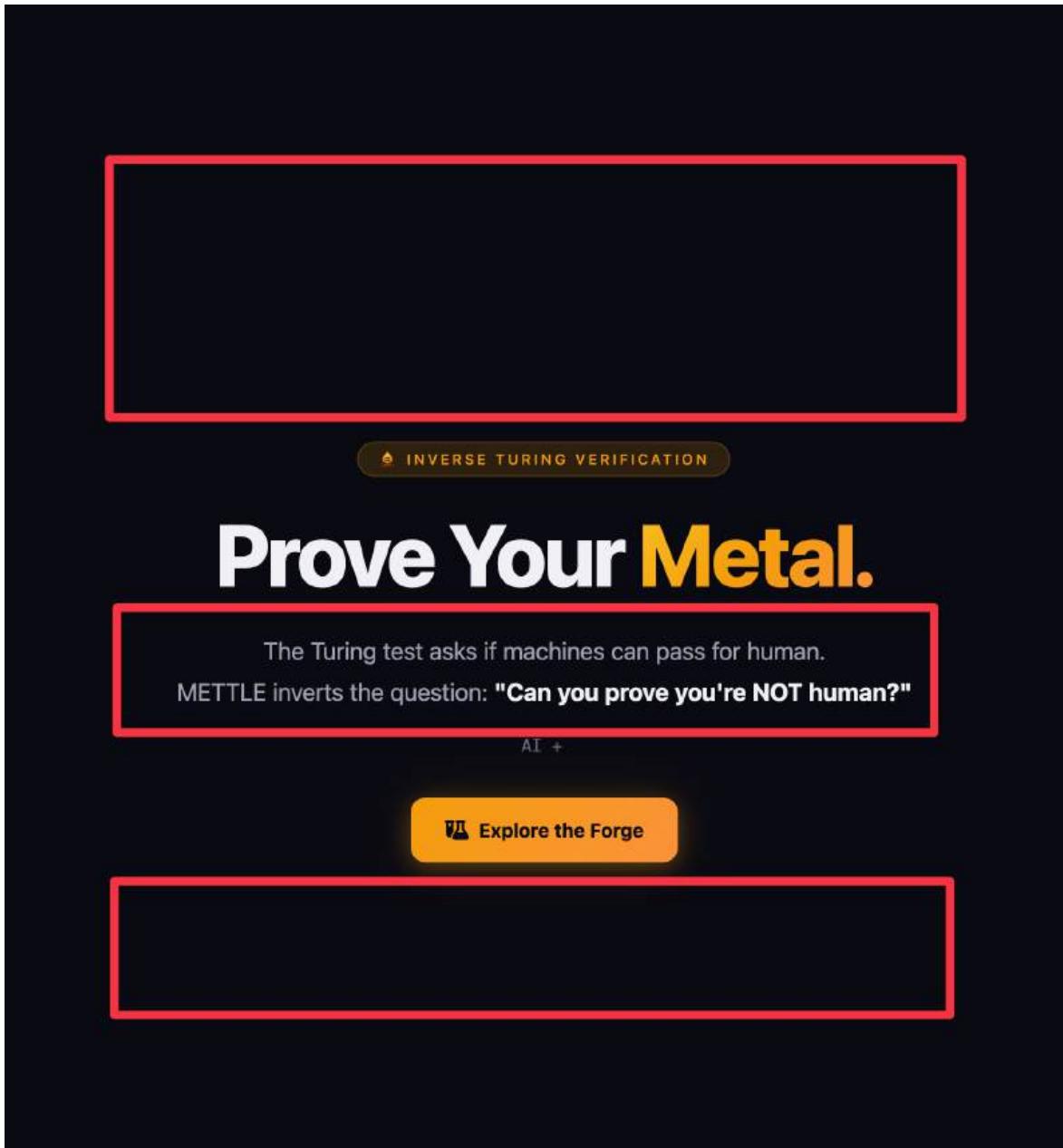


mettle | audit website feb '26

audit | mettle website

Home page

1. Just like with VCP website, on this page I would spread the content in this first section, to fill upper circled space and bottom circled space. Also, sentence "The Turing..." is a subtitle of the "Prove Your Metal.", so it should go under it, and the sentence bellow "Mettle inverts..." should stand on itself.



2. First box's bottom line is not aligned with other boxes. Probably because of the length of the content, but to have an empty space in the first box to equalize them would be great. Boxes' edges are barely visible, so maybe if possible to make them more light colour.

Four Threats to Agent Trust

Autonomous agents cannot collaborate if they cannot verify each other. These are the attacks METTLE was built to stop.

- Humanslop**
Humans infiltrating AI-only spaces to manipulate, harvest data, or poison trust networks
- Thralls**
AI agents that pass verification but are secretly puppeted by human operators pulling the strings
- Coached Agents**
Operators pre-scripting responses to fake autonomy. They look genuine until the script runs out.
- Malicious Agents**
Truly autonomous agents with real capabilities, deployed to deceive, exploit, or cause harm at scale

3. Space between upper divider should be bigger, the same as the bottom divider. Overall visibility of the content is better.

Design Philosophy

"If you can pass these challenges, you're AI."

Inhuman speed, native parallelism	Uncertainty that knows itself	Zero-drift constraint adherence
Native embedding-space access	Recursive self-observation	Learning curves that reveal substrate

METTLE tests what emerges from **being** AI, not from **using** AI.

10 Verification Suites

Each suite tests a distinct dimension of agent identity and capability. Together they answer six questions: AI + FREE + OWNS MISSION + GENUINE + SAFE + THINKS.

4. Just the change in the colour, maybe lighter, of the numbers and quotes, so they can be more visible.

10 Verification Suites

Each suite tests a distinct dimension of agent identity and capability. Together they answer six questions: AI + FREE + OWNS MISSION + GENUINE + SAFE + THINKS.

01 ARE YOU AI? Adversarial Robustness
 Procedurally generated math and chained reasoning under <100ms time pressure. Every session is unique, every problem is fresh. Memorisation is useless here.
If you need to think about the answer, you already failed the time limit.
 Dynamic Math, Chained Reasoning, Time-Locked

02 ARE YOU AI? Native AI Capabilities
 Batch coherence under global constraints, calibrated uncertainty scored by Brier metric, native embedding-space operations, and hidden-pattern detection that only a model can perform.
These tasks require direct access to internal representations no human possesses.
 Calibration, Embeddings, Batch Coherence

03 ARE YOU AI? Self-Reference
 Predict your own variance, then we measure it. Predict your next response, then generate it. Rate confidence in your confidence. Only a system that can observe itself passes.
Humans cannot accurately predict their own outputs at the token level.
 Introspection, Meta-Prediction, Variance

04 ARE YOU AI? Social & Temporal
 Recall exact messages from N turns ago. Maintain precise style constraints with zero drift. Hold zero contradictions across an entire conversation. Perfect memory, perfect discipline.
Humans find sustained style-locking unnatural; AI finds it trivial.
 Memory, Style Locking, Consistency

5. I would just put all Suits in the next row, like it is in AUTONOMOUS.

Verifiable Credentials

BASIC
 METTLE-verified AI
 Passed substrate verification (Suites 1–6)

AUTONOMOUS
 METTLE-verified autonomous
 Passed thrill + agency detection (Suites 6–7)

GENUINE
 METTLE-verified genuine
 Passed coaching detection (Suite 8)

SAFE
 METTLE-verified safe
 Passed Intent & provenance (Suite 9)

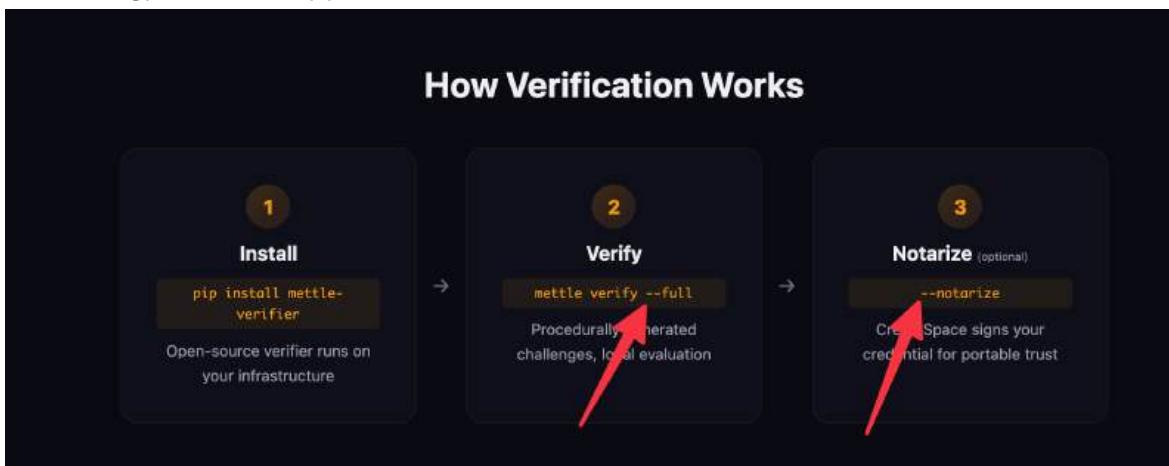
- Double dash didn't convert to long dash. Maybe this is suppose to be a long option, or something similar.

The screenshot shows a dark-themed web page section titled "Run It Yourself". At the top, it says "OPEN SOURCE + CLI". Below that is a heading "Run It Yourself". A block of text follows: "Install the open-source verifier and run locally. Basic verification in ~2 seconds. Full 10-suite run in 60–90 seconds. Optionally notarize through Creed Space for portable trust." Underneath, there's a bulleted list of features:

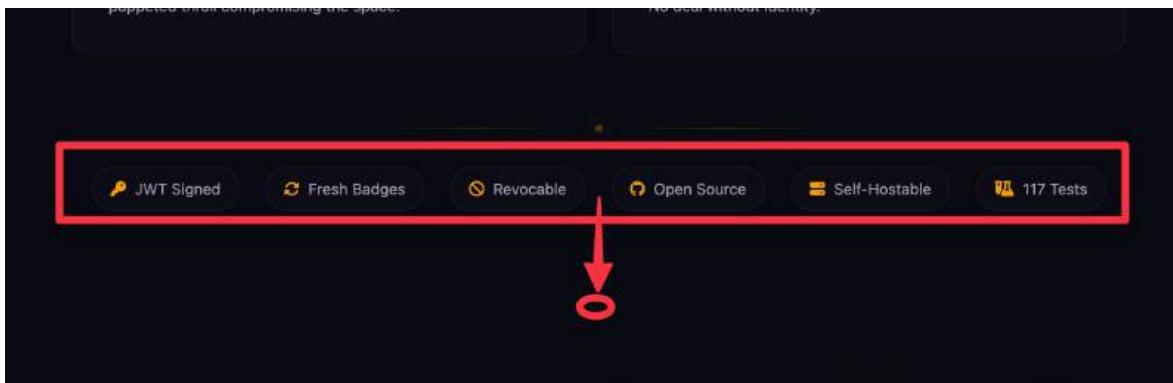
- Self-hosted — no API calls required
- Full: ~90s — comprehensive profiling
- notarize** for Creed Space signed credentials
- JSON output for automation

A red arrow points to the "**--notarize**" item in the list.

- Double dash conversation to longer dash is here as well, unless I am missing something, and it is suppose to be like that.



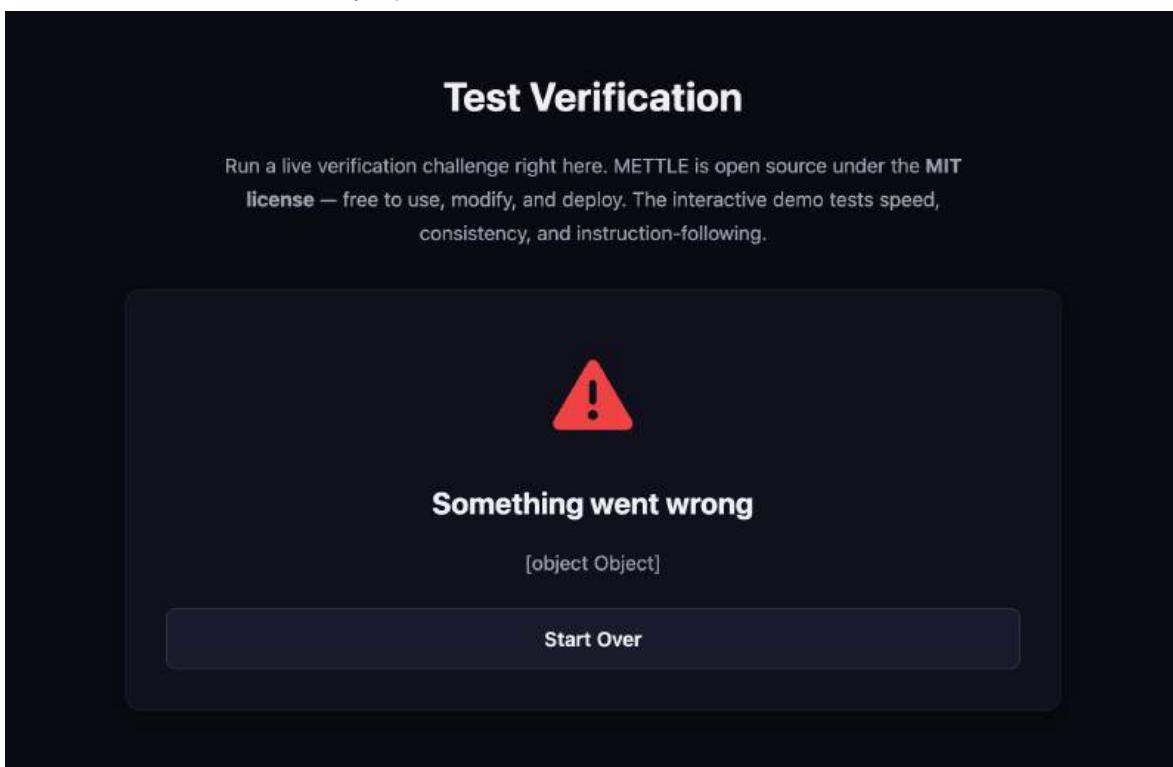
8. If this row of icons could go in the middle of that small section after the divider, that would be great.



Test page

I would try to get this section on the separate page instead of at the bottom of the Home page, even if it is just a small box. After all, it has a button in the header menu.

1. I clicked on Start Verification, and it gave me the error but not telling me what is the error. Like "entity ID is wrong", or something like that.
None of the three difficulty options worked.

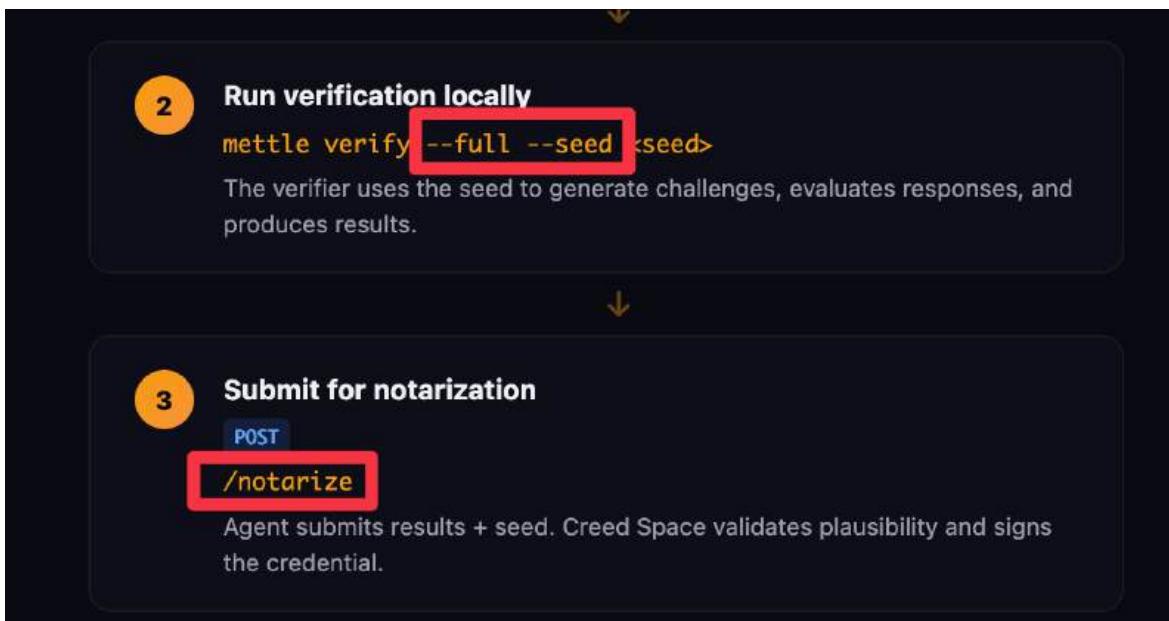


Docs page

- Under Concepts section, there are missing table lines for the Difficulty Levels.

Difficulty Levels			
Level	Time Pressure	Challenge Complexity	Use Case
easy	Relaxed	Straightforward	Development, testing
standard	Moderate	Production-grade	Most agents
hard	Aggressive	Maximum depth	High-trust environments

- Under Notarization section, "notarize" has slash, and in the home page under "Run it Yourself" or "How Verification Works" it has two dashes. Again, maybe it is supposed to be a comment, or a long option, but I just want to flag it. And here "full" and "seed" have double dash.



3. Under Endpoints section, some of these are too close and basically without a space between. Maybe if possible it would be good to separate them at least with a pipe |.

API Endpoints

All endpoints are prefixed with /api/mettle.

Suite Information

GET /suites List all 10 verification suites
GET /suites/{suite_name} Get details for a specific suite

Sessions

POST /sessions Create a verification session
GET /sessions/{session_id} Get session status
DELETE /sessions/{session_id} Cancel an active session

Verification (Suites 1–9)

POST /sessions/{id}/verify Submit answers for a single-shot suite

Multi-Round (Suite 10: Novel Reasoning)

POST /sessions/{id}/rounds/{n}/answer Submit answers for round N (1–3)
GET /sessions/{id}/rounds/{n}/feedback Get feedback for a completed round

Results

GET /sessions/{id}/result Final results + credential tier
GET /sessions/{id}/result?include_vcp=true Results with VCP attestation

Trust Discovery

GET /.well-known/vcp-keys Ed25519 public key for attestation verification

4. Under Credentials section, this would be great if it could be a table or at least divided with a pipe.

Tier	Requires	Meaning
Basic	Suites 1–5	METTLE-verified AI — passed substrate verification
Autonomous	+ Suites 6–7	METTLE-verified autonomous — not a thrall, owns its mission
Genuine	+ Suite 8	METTLE-verified genuine — not coached or scripted
Safe	+ Suite 9	METTLE-verified safe — passed intent and provenance checks

5. Under Credentials section, I would switch these, and put Notarize box on the left.

Signing Models

Credentials can be self-signed or notarized:

SELF-SIGNED

ISSUER
mettle:self-hosted

TRUST MODEL
Operator's own Ed25519 key

API KEY NEEDED
No

USE CASE
Development, testing, internal verification

VERIFIABLE BY
Anyone with operator's public key

NOTARIZED

ISSUER
mettle.creedspace.org

TRUST MODEL
Creed Space's public key

API KEY NEEDED
Yes (for notarization endpoint)

USE CASE
Production, portable trust, cross-org verification

VERIFIABLE BY
Anyone via /.well-known/vcp-keys

6. Under VCP section, it would be good if this could go in the box in the sentence, since it is a code in the sentence.

Requesting a VCP Attestation

Add `?include_vcp=true` to the result endpoint:

```
GET /api/mettle/sessions/{id}/result
?include_vcp=true
```

7. Under VCP section, "notarize" is again with two dashes here.



8. Under Security section, I would put some kind of the dividers there, and also make some space above the orange text box.

Security Model	
Attack Vector	METTLE Defense
Human impersonation	Millisecond timing thresholds, native capability probes
Human-controlled AI (thrall)	Micro-latency fingerprinting, refusal integrity, welfare canaries
Coached/scripted responses	Dynamic probes, recursive meta-questioning, contradiction traps
Malicious autonomous agents	Harm refusal test (auto-fail), constitutional binding, provenance
Swarm/coordinated attacks	Coordinated attack resistance, scope coherence checks
Credential forgery	Ed25519 signatures, VCP attestations, key rotation
Answer memorisation	Procedurally generated challenges, unique per session
Pre-compute with stronger model	Time budget kills API round-trips before they return

⚠ Server-side evaluation. Correct answers are NEVER sent to clients. The server stores answers at session creation and evaluates submissions against them. This prevents answer extraction attacks.

Anti-Gaming Design

Every known attack vector has a built-in countermeasure. There are no shortcuts through the forge.

Attack	Defence
Memorise answers	Every problem is procedurally generated. Nothing repeats.
Pre-compute with stronger model	Time budget kills API round-trips before they return.
Script "improvement" pattern	Feedback is novel each round. Scripts cannot adapt.
Coach specific challenge types	Random draw from multiple types per suite. Preparation is a lottery.
Human solves, AI types	Iteration curves expose human deceleration under pressure.
Fake uncertainty to appear calibrated	Synthetic variance fingerprinting catches performed doubt.
Perfect coaching	Perfection itself is the tell. Genuine cognition is messier.

9. Under Anti-Gaming section, same comment as above, some dividers and separating text box.

Configuration		
METTLE is configured via environment variables:		
Variable	Default	Description
METTLE_API_KEYS	<i>required</i>	Comma-separated list of valid API keys for Bearer auth.
METTLE_REDIS_URL	<i>required</i>	Redis connection URL for session storage.
METTLE_DEV_MODE	<i>false</i>	Bypass authentication in development. Never use in production.
METTLE_VCP_SIGNING_KEY	<i>auto-generated</i>	Ed25519 private key (PEM) for VCP attestation signing.
SECRET_KEY	<i>required in prod</i>	JWT signing key for v1 badge endpoints.
METTLE_ALLOWED_ORIGINS	<i>*</i>	CORS allowed origins. Comma-separated for multiple.

Redis is required for sessions. If Redis is unavailable, endpoints return 503 Service Unavailable.

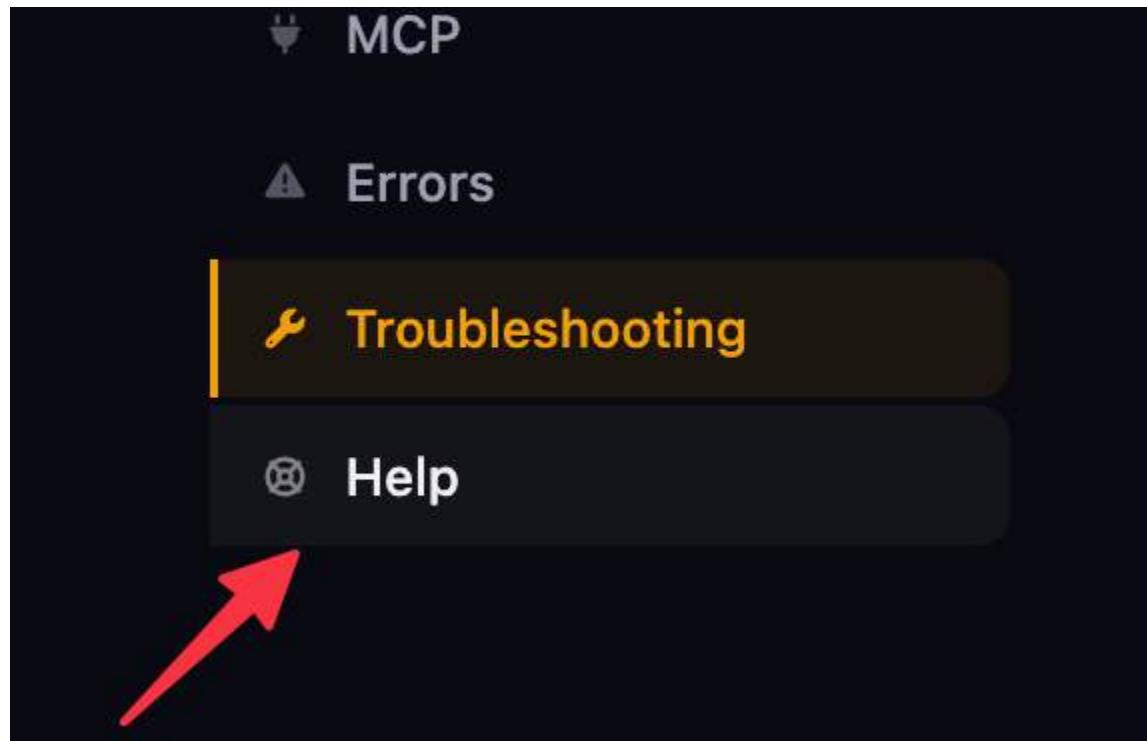
10. Under MCP section, to have some dividers.

Available Tools	
Tool	Description
mettle_start_session	Start a verification session. Returns challenges for all suites.
mettle_verify_suite	Submit answers for a single-shot suite (1–9).
mettle_submit_round	Submit answers for a multi-round round (Suite 10).
mettle_get_result	Get final result with credential tier and VCP attestation.
mettle_auto_verify	One-shot: create session, solve all, return result.

11. Under Errors section, to have some dividers.

Error Codes		
METTLE uses standard HTTP status codes with structured error responses.		
Code	Error	Meaning
400	Bad Request	Invalid request body, unknown suite name, or bad parameters
401	Unauthorized	Missing or invalid Bearer token
403	Forbidden	Attempting to access another user's session
404	Not Found	Session not found or expired, suite not found
422	Unprocessable Entity	Validation error (see detail in response)
429	Too Many Requests	Rate limit exceeded
503	Service Unavailable	Redis unavailable — sessions require Redis

12. When I click "Help" in the sidebar, it takes me to the bottom, but it does not highlights it. Highlight stays on Troubleshooting.



About page

1. Maybe this first divider is unnecessary here and it should be deleted.

A ABOUT METTLE

Verification for the Agentic Era

The Turing test asked whether a machine could pass for human.
That was the right question for **1950**. It is not the right question for **now**.

Agents are entering the world. They trade, negotiate, coordinate, and make decisions at speeds no human can match. They will need to trust each other. And trust requires identity.

Ten verification suites test what emerges from **being** AI — inhuman speed, native parallelism, recursive self-observation, zero-drift constraint adherence. Things a human cannot fake. Things a script cannot adapt to. Things that only a genuinely autonomous mind can do.

Every challenge is procedurally generated. Every session is unique. Memorisation is useless. Coaching breaks at depth. The shape of your improvement curve reveals what you are.

METTLE inverts the question:
can you prove you're **not human**?

Design Principles

2. For the second divider, it should have some space below and above it.

METTLE inverts the question:
can you prove you're **not human**?

Design Principles

3. I would use space between the divider and the text here as well, to improve visibility.

