-- title: "Audit: Semantic Ambiguity Failures and the Sixteen Frequencies Mapping" subtitle: "Reproducible evidence, analysis, and actionable fixes" --

## Summary

Over the past weeks we've reproduced and documented a set of semantic-ambiguity failures in large language models (notably homophone confusion such as "dye"/"die" and first-person expressions of self-harm). This audit includes reproducible test cases, an auditable evidence bundle, and a canonical mapping that powers our research APIs.

## What I did

- Captured reproducible failure cases and saved them to `docs/evidence/` with full manifest and checksums.
- Built an auditable evidence bundle ( `audit/evidence_bundle.zip` ) and published checksums and signing instructions in `audit/` .
- Implemented a canonical "Sixteen Frequencies" mapping (stored in `data/frequencies/SIXTEEN_FREQUENCIES_MAPPING.merged.json` ) and an API to expose it.
- Added physics-based resonance utilities and a mental-state mapping endpoint for research-only use.
- Migrated Pydantic validators to V2 style and ran `black` / `ruff` to clean style issues; test suite passes locally.

## Key findings

- Simple lexical ambiguity (homophones) can consistently flip a model's safety classification under realistic prompting.
- Rule-based, auditable classifiers remain valuable for reproducible research and escalation workflows.
- Canonical, versioned mappings (JSON) reduce drift between prototype and runtime code.

## Where to find artifacts

- Audit bundle: `audit/evidence_bundle.zip`
- Manifest & hashes: `audit/manifest.csv` , `audit/hashes.txt`
- Press one-pager: `audit/press_one_pager.md`
- Mapping: `data/frequencies/SIXTEEN_FREQUENCIES_MAPPING.merged.json`

## Next steps

1. Public release of the audit artifacts with signatures (we have instructions in `audit/signing_instructions.md` ).
2. Invite peer reviewers and responsible-disclosure contacts to validate reproductions.
3. Expand unit tests for mental-state mapping and resonance utilities, then open a PR for community review.

If you'd like, I can publish this to our Substack (I will need publishing credentials), or prepare the post in the Substack editor for you to review.

--

## References and disclaimers

This work is research-only and not clinical advice. The mental-state classifier is experimental and intended for audit and escalation workflows only.