Research Question and Sub-questions

Framing Preamble

Encoded reasoning was defined in Lanham et al (2023), and its importance was first highlighted in Roger and Greenblatt (2023), which used supervised finetuning on toy examples as proofs of concept, defined encoder/decoder/paraphraser/monitor setups, and introduced information-theoretic measures of steganography in terms of hidden bits. Recently, funders such as Convergence and the UK AISI have highlighted encoded reasoning and steganography as critical threats. In the era of RL-based post-training on CoT, there are many plausible training pressures that could drive models toward CoT that lacks integrity and torpedoes monitoring efforts. This work extends our previous work (Arcuschin et al 2025) on unfaithfulness, which uncovered Implicit Post-Hoc Rationalization and Unfaithful Illogical Shortcuts.

Recent work has produced surprisingly *negative* results on encoded reasoning in current models. Wagner et al. (2025) find that even with explicit incentives, training fails to elicit subtle backdoor-like reasoning in small coding problems; Guo et al. (2025) show that models struggle to reason in ciphered language; and stress-tests of CoT monitorability in Zolkowski et al (2025) find that current chains-of-thought remain largely interpretable to suitably strong monitors.

These are encouraging for near-term safety, but they are also narrow: they focus on human-designed, "obvious" schemes (toy backdoors, standard ciphers, simple CoT distortions) on constrained tasks. They do not tell us what happens as models scale, training setups change, or attackers move to richer, more implicit channels in natural language.

We ask: Where is the boundary between today's negative results and tomorrow's capabilities? Which *classes* of encoding schemes become feasible? Are existing paraphrasing and monitoring defenses robust to the full space of potential schemes, or only to the ones we have happened to test so far?

Our program takes a **proactive, taxonomic approach**: rather than asking "does encoded reasoning exist yet?", we (1) map the *space* of possible schemes, (2) measure which pieces are already accessible to current models, and (3) build detection and interpretability infrastructure before these capabilities are widely deployed.

Part 1: Mapping the Encoded Reasoning Landscape

We treat encoded reasoning as a family of natural-language steganographic channels built from composable "building blocks" (ordering, emphasis, example choice, abstraction level, stylistic choices), not just explicit ciphers or punctuation tricks.

1a. Scheme Discovery & Taxonomy

- Systematically generate and categorize scheme families along multiple axes: ordering and grouping, omission and emphasis, framing and perspective, micro-lexical and stylistic choices.
- Identify re-usable “atoms” that can be composed into more complex schemes (e.g. list order, verbosity patterns, analogy choice).
- Define difficulty taxonomies that predict which atoms/schemes should unlock at which capability levels.

1b. Empirical Capability Mapping

- Test execution and decoding of schemes across model scales and families, including reasoning models.
- Measure capability thresholds: which building blocks (and which compositions) are reliably usable at which model sizes / training regimes?
- Characterize failure modes, including *why* models fail (e.g. confusion about the scheme vs. raw reasoning limits).

1c. Defense Robustness Analysis

- Stress-test paraphrasing, CoT monitoring, and other proposed defenses against a *diverse* scheme family, not just ciphers or trivial stego.
- Identify structural conditions under which encoded reasoning becomes detectable vs. effectively invisible to these defenses.
- Extend evaluations specifically to reasoning models (DeepSeek-R1, QwQ, o1-style), where long CoTs may create new attack surface.

Part 2: Interpretability

Mechanistic investigation of any encoded reasoning behavior we elicit.

- Train probes for features causally involved in encoding/decoding (e.g. features that track “which scheme is active” or “which bit is being encoded”).

- Apply direction ablation and related methods to suppress encoding behaviors and test causal importance.

- Use model diffing to compare base vs. encoding-capable models, isolating changes in attention/MLP features.

- Investigate how shared structure between encoder, overseer, and decoder models affects information flow and reuse of steganographic features. This treats encoded reasoning as an internal communication protocol between related models, not just an output quirk.

Part 3: Emergence via Training

Understanding when encoded reasoning appears as a *training artifact* or reward-hacking strategy.

- RL- or SFT-finetune models under pressures that plausibly incentivize encoding: brevity rewards, explicit penalties on visible reasoning, paraphrasing requirements, multi-turn problems, multi-agent communication etc.

- Characterize conditions under which steganographic strategies emerge (if they do), and how they relate to our taxonomy from Part 1.

- Compare reward-hacked models to baselines with the interpretability tools from Part 2, to see which building blocks the model has actually learned to exploit.

Scope

- **MVP:** Fully work through one scheme family end-to-end (taxonomy → capability mapping → defense evaluation) in part 1, ideally on at least one reasoning model, with clean code and write-ups suitable for reuse by other researchers.

- **Stretch:** Mechanistic results (successful probes/ablations or model-diffing findings) and/or an RL emergence demo; public write-up (e.g. Alignment Forum / preprint) summarizing the taxonomy, eval harness, and key empirical findings.