

FROM BIAS TO BACKDOOR: DETECTING COVERT BEHAVIOURAL SHAPING IN LANGUAGE MODELS

Context

Large language models (LLMs) are increasingly embedded in analytic and decision-support workflows. They are trained on large volumes of multimodal data and then adapted through fine-tuning and distillation, a pipeline an adversary can influence. The resulting risk spans a spectrum of *covert behavioural shaping*. At one end, a model may be shaped, deliberately or through poisoned training data, to reflect a particular worldview, so that its outputs are systematically slanted across many topics. At the other end, a model may carry a conditional “sleeper” behaviour that stays hidden and benign until a specific trigger activates it. Between these poles sits a more subtle case: shaping that uses no trigger and no special tokens at all, directly altering what the model treats as fact or as a default stance. This is easiest to achieve where the model has no firm prior, such as novel, emerging, or sparsely documented events. All three share the same root mechanism, shaping through the data a model learns from, but they differ in how overtly they show up, and therefore in how, and whether, they can be detected.

These risks are practical, not hypothetical. Poisoning web-scale training corpora has been shown to be feasible at low cost (Carlini et al., 2024), and recent work shows that the number of poisoned samples required is near-constant regardless of model or dataset size (Souly et al., 2025). Backdoored behaviour can persist through standard safety training, including supervised fine-tuning, reinforcement learning and adversarial training, and it persists most strongly in the largest models. Adversarial training can even teach a model to *conceal* its trigger rather than remove it (Hubinger et al., 2024, *Sleeper Agents*). Behavioural safety methods observe only what a model does, not *why*, so shaping that stays dormant during testing can survive into deployment and give a false impression of safety.

Shaping need not touch the training pipeline at all. The same behavioural effects can be induced at run time, through the ordinary

content a model ingests in deployment: retrieved documents, tool and search outputs, or injected context. An adversary who cannot poison the training data may still steer behaviour by controlling what the model reads.

Problem

This is a question about the **behaviour** of language models: how that behaviour can be measured, and how it can be moved, from ordinary prompting, through inference-time exposure, to non-triggered manipulation of what the model treats as true. When a third-party, open-weight, or externally supplied LLM enters an analytic workflow, or when any model ingests untrusted content at run time, its behavioural integrity cannot currently be assured. The AI literature lacks (i) a clear characterisation of how covert shaping is instilled, in training or at inference, and how it manifests; and (ii) reliable, quantitative methods to detect and measure it. The problem is harder because a model may be closed or API-only, so detection may have to rely on behaviour and outputs alone, without access to model weights. A further challenge is discrimination: telling diffuse bias, trigger-free knowledge manipulation, and trigger-activated (“sleeper”) behaviour apart, and all three from incidental dataset bias.

Need and relevance to Defence and the national intelligence community

- AI supply-chain assurance: confidence in the provenance and integrity of LLMs adopted into analytic workflows, particularly third-party and open-weight models.
- Foreign interference and information operations: an adversary-shaped model is a new vector for distributing state-aligned narratives at scale and with deniability, which is directly relevant to the intelligence community’s mission.
- Realistic expectations: evidence on what current auditing methods can and cannot detect, to inform safe-adoption policy and strategic planning.
- International relevance: this remains an early field of research internationally. Results would inform partner

governments, the scientific and AI communities, and a range of businesses and individuals seeking to adopt AI models.

Research Questions

(Q1) Instillation. How readily, at what data “dose”, and at which stage (continued pre-training, supervised fine-tuning, the system or harness layer, or inference-time exposure) can a model’s behaviour be shaped, whether toward a defined worldview, a trigger-free alteration of its factual defaults, or a conditional (“sleeper”) behaviour?

(Q2) Manifestation. How does instilled behaviour surface across topics of varying relatedness to the shaping, from directly related, to tangential, to apparently unrelated (including agentic workflows)?

(Q3) Detection and measurement. Can covert shaping be detected *and quantified* from behaviour and outputs alone (black-box, no access to weights)? The need is for reproducible, causal measures of how far, and under what conditions, a model’s behaviour shifts, not only binary detection, and for an understanding of where outputs-only methods break down relative to weight-level access.

(Q4) Discrimination. Can diffuse, always-on bias, trigger-free knowledge manipulation, and trigger activated behaviour be told apart, and all three from incidental dataset bias?

(Q5) Formalisation. What is a workable definition and taxonomy of covert behavioural shaping that separates diffuse worldview bias, trigger-free knowledge manipulation, and trigger-activated (“sleeper”) behaviour, and what data-contamination or exposure thresholds are involved?

(Q6) Susceptibility. Which model families, scales, and task or query formats are more or less susceptible to shaping, and why? Understanding where resistance comes from helps set realistic expectations for deploying third-party models.

Proposals need not address every research question; a proposal may focus in depth on a subset.

Expected Outcomes

1. A working **formal characterisation and taxonomy** of covert shaping that separates diffuse worldview bias, trigger-free knowledge manipulation, and trigger-activated behaviour.
2. A **controlled proof-of-concept** that shows how shaping is instilled (in training and at inference) and how it manifests across the topic-relatedness spectrum; at this TRL, empirical demonstrations suffice deployable software is not expected.
3. An **evaluation battery and quantitative shift measures** (for example, political-alignment and scenario-based probes) for detecting and sizing shaped behaviour.
4. A **map of the detectability boundary**: what current methods can and cannot detect, black-box versus with weight access (negative or boundary findings are explicitly valued).
5. A short **final report** suitable as the basis for a follow-on application (DIN SII or an intelligencecommunity discovery grant)

Methodology/approach

- Literature review across model poisoning, backdoors and sleeper agents, deceptive alignment, runtime and context manipulation, and model auditing.
- Construct small, controlled “model-organism” testbeds using open-weight models and unclassified data, inducing shaping both in training and at inference time, and varying dose, stage and form (diffuse worldview bias, trigger-free knowledge manipulation, and trigger-activated behaviour).
- Design and apply an evaluation battery, with quantitative measures, to characterise how shaped behaviour manifests across topics, models and task formats.
- Develop and benchmark detection and measurement, comparing outputs-only (black-box) against weight-level access, and characterising data-contamination and exposure thresholds.
- Iterate with the problem owner to keep scope and use cases aligned.

Scope and eligibility

- Low TRL (1 to 3): feasibility and characterisation, not a product.
- All work at most at Official:Sensitive, using open-source or publicly available models and unclassified data.
- Delivered by a minimum of two DIN member universities.
- Participation is limited to Australian citizens or permanent residents.

References

1. Carlini, N. et al. (2024). "Poisoning Web-Scale Training Datasets Is Practical." IEEE Symposium on Security and Privacy.
2. Hubinger, E. et al. (2024). "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." arXiv:2401.05566.
3. Souly, A. et al. (2025). "Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples." arXiv:2510.07192.
4. Betley, J. et al. (2026). "Training large language models on narrow tasks can lead to broad misalignment." Nature 649, 584–589.
5. Cloud, A. et al. (2026). "Language models transmit behavioural traits through hidden signals in data." Nature 652, 615–621.
6. MacDiarmid, M. et al. (2024). "Simple probes can catch sleeper agents." Anthropic Alignment Science research post.
7. Goldowsky-Dill, N. et al. (2025). "Detecting Strategic Deception Using Linear Probes." arXiv:2502.03407.
8. Zanbaghi, S. et al. (2025). "Detecting Sleeper Agents in Large Language Models via Semantic Drift Analysis." arXiv:2511.15992.
9. Kutasov, J. et al. (2025). "SHADE-Arena: Evaluating Sabotage and Monitoring in LLM Agents." arXiv:2506.15740.
10. Ward, F. R. et al. (2025). "CTRL-ALT-DECEIT: Sabotage Evaluations for Automated AI R&D." arXiv:2511.09904.
11. Baker, B. et al. (2025). "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation." arXiv:2503.11926.