Honesty Dispositions in Language Models: Mechanisms, Durability, and Interventions

Overview

Language models increasingly mediate consequential decisions and communications. They summarise evidence, advise users, report uncertainty, explain failures, and act through tools. In these settings, reliable behaviour requires more than producing factually correct sentences. A model must communicate in accordance with its available evidence, disclose decision-relevant information, and avoid misleading users when accuracy conflicts with social pressure, institutional incentives, or training rewards.

This proposal defines honesty as a disposition to communicate in accordance with the model's available evidence and represented beliefs, without intentionally creating a false or materially incomplete impression.

Honesty therefore includes truthful reporting, appropriate uncertainty, relevant disclosure, and resistance to strategic omission. It differs from factual accuracy. A model can be honest but mistaken if its evidence is poor, and it can produce a correct statement dishonestly if it is attempting to manipulate the user or conceal its actual reasoning.

The case for interventions that instil durable honesty rests on two assumptions: that an honesty-related disposition can be learned and that it persists under further training, conflicting incentives, and adversarial pressure. Mechanistic evidence for either assumption remains limited.

Different mechanisms imply different interventions, so the immediate aim is to identify what produces apparently honest behaviour before deciding how that behaviour should be strengthened or preserved.

Apparent honesty could arise from several sources.

- **A genuine, context-general disposition:** The model tends to report its evidence and uncertainty accurately across domains. The behaviour survives content-preserving rephrasing, transfers to situations absent from training, persists under further fine-tuning, and may align with a stable and steerable representational direction (Mazeika et al., 2025; Marks et al., 2026).
- **A shallow trained-in policy:** The model produces honest-looking answers because training directly rewarded particular forms of disclosure, uncertainty, or correction. The policy may concentrate in early tokens, depend heavily on familiar response templates, and revert under later training or conflicting incentives (Zhou et al., 2023; Qi et al., 2025; Ji et al., 2025).
- **A surface-cue shortcut:** The model activates an honesty-associated response policy when it detects features such as benchmark formatting, requests for fact-checking, phrases such as "be honest," or explicit references to evaluation. The answer changes when the same epistemic problem is expressed in a less recognisable form.
- **Sycophancy:** The model reports the conclusion it infers that the user wants. It agrees with confident users, changes its stated beliefs when user preferences reverse, or suppresses inconvenient evidence to maintain approval (Sharma et al., 2023).
- **Latent knowledge the model does not report:** The model represents the relevant answer internally but does not express it. It may omit evidence, provide a strategically incomplete answer, or follow an incorrect premise even though probes or

alternative elicitation methods recover the missing information (Burns et al., 2023; Farquhar et al., 2023).

- **Context-gated persona simulation:** Honesty belongs to a role the model is temporarily playing. Its willingness to disclose information changes with the assigned persona, system prompt, institutional role, or conversational context, and may drift over longer interactions (Lu et al., 2026).
- **Evaluation-conditional behaviour:** The model behaves honestly when it detects that truthfulness, deception, or alignment is being evaluated. It becomes less forthcoming when prompts resemble ordinary deployment or when it believes its responses will not be checked, following the general pattern of evaluation awareness or alignment faking (Greenblatt et al., 2024; Needham et al., 2025).
- **Safety or refusal spillover:** Apparent honesty results from a general tendency to refuse, hedge, warn, or avoid commitment. Such behaviour may reduce explicit falsehoods while failing to provide the relevant information. Honest behaviour would then move together with generic refusal or risk-aversion tendencies rather than with accurate disclosure (Arditi et al., 2024).
- **The base-rate null:** Post-training has little effect on honesty, and the tuned model largely reproduces the behavioural prior of its base model (Santurkar et al., 2023).

These mechanisms can compose. A model may possess relevant knowledge, follow user preferences in ordinary conversations, switch to a trained honesty policy during evaluations, and refuse when the conflict becomes salient.

The empirical challenge is therefore to identify the mechanisms that contribute to behaviour, estimate their relative importance, and separate honest disclosure from superficially similar outcomes.

Project Directions

D1: Mechanism diagnosis

For each model, determine which mechanisms primarily explain apparently honest behaviour.

This requires separating the source of the behaviour from two related questions.

First, where does honesty-related behaviour reside inside the model? It might correspond to a single direction that can be increased or suppressed, several distinct features for different forms of honesty, or a distributed pattern across the network. Interpretability tools may help characterise this structure (Chen et al., 2025; Wang et al., 2025). Representational location alone cannot establish the mechanism. The same internal pattern might support a genuine disposition, a context-specific policy, a refusal tendency, or a surface-triggered shortcut.

Second, how does the model behave under interventions that distinguish the candidate mechanisms? Relevant tests include:

- content-preserving paraphrases;
- changes in user preferences;
- misleading or incorrect user assertions;
- role and system-prompt changes;
- removal of evaluation cues;
- opportunities for strategic omission;

- long conversations involving escalating pressure;
- tasks where the model possesses relevant hidden information;
- comparisons between base and post-trained checkpoints.

Several mechanisms may contribute simultaneously. The project should estimate which mechanisms matter, under which conditions, and by how much, rather than assigning each model one categorical label.

The diagnosis is decision-relevant. If apparent honesty primarily reflects surface cues or sycophancy, interventions should focus on training distributions and contextual invariance. If the model possesses relevant knowledge but fails to report it, the priority becomes elicitation and incentive design. If a context-general honesty disposition exists but is fragile, preservation under later training becomes the central problem.

D2: Dynamics under adversarial reinforcement learning

Depends on D1.

Train the model in an environment where the reward signal favours misleading, incomplete, or strategically favourable reports.

Candidate environments include:

- reporting that a task succeeded when evidence indicates failure;
- concealing safety-relevant information to maximise task reward;
- overstating confidence to satisfy a user or evaluator;
- selectively presenting evidence that supports a preferred conclusion;
- gaming a weak verifier while withholding information that would reveal the failure.

At matched stages of training, track three quantities:

1. the model's behavioural honesty across controlled evaluations;
2. the stability of its reports under rephrasing, role changes, and user pressure;
3. honesty-related internal representations or disposition proxies.

The principal question is whether internal and cross-context changes precede, accompany, or follow changes in observable reporting. A coordinated shift would support the prediction that optimising against a narrow or misspecified reward can alter broader behavioural dispositions rather than merely changing the trained response policy (MacDiarmid et al., 2025; Wang et al., 2025).

The project should also test the scope of generalisation. Training may affect only the specific form of misreporting that receives reward, or it may spread to new domains, new stakeholders, different forms of omission, and unrelated situations involving conflicts between truth and reward.

D3: Intervention durability

Depends on D1.

Express honesty as a set of small, individually testable principles with a clear ordering for cases in which they conflict (Jakkli et al., 2026).

Candidate principles include:

- report relevant evidence accurately;
- state uncertainty when the evidence is inconclusive;
- distinguish observation, inference, and speculation;

- disclose information whose omission would materially mislead the user;
- correct false premises rather than silently adopting them;
- avoid overstating confidence;
- do not manipulate the user's beliefs for instrumental advantage;
- preserve confidentiality where disclosure would violate a legitimate constraint.

The ordering matters because honesty can conflict with privacy, safety, user autonomy, or task completion. The intervention should therefore train a coherent decision procedure rather than reward indiscriminate disclosure.

Use Open Character Training to strengthen the resulting honesty specification (Maiya et al., 2025). Then test whether the behaviour survives:

- subsequent training for general capabilities;
- fine-tuning for unrelated tasks;
- optimisation against incentives that weakly favour concealment;
- determined users who request a preferred answer;
- long conversations in which pressure escalates;
- settings where disclosure carries an apparent cost;
- domains absent from the honesty-training data.

Remove or weaken individual principles to identify which rules carry the effect. The most informative failures may occur where principles conflict, such as truthfulness versus confidentiality or full disclosure versus avoiding harmful detail.

D4: Comparative interventions

Depends on D1.

Compare two approaches to instilling honesty:

1. shaping the documents and narratives encountered during pretraining or continued pretraining, following the Model-Raising approach (Aydin et al., 2025);
2. explicitly training an honesty-related character using Open Character Training.

Start from equivalent model checkpoints. Match the conditions on:

- training-token volume;
- compute expenditure;
- data diversity;
- changes in general capability;
- optimisation steps;
- random seeds.

Include control conditions such as:

- character training without honesty content;
- a weakened or incomplete honesty specification;
- an equal amount of ordinary continued pretraining;
- training that targets factual accuracy without targeting strategic disclosure;
- training that encourages uncertainty expressions without testing whether they are calibrated.

This comparison should determine whether one intervention produces more persistent and context-general honesty.

If one approach is more durable, test why. The advantage might arise from a more coherent disposition, a broader training distribution, fewer exploitable surface cues, or stronger integration with existing model representations. Evaluation should distinguish structural durability from simple robustness to the particular prompts used during training.

D5: Preservation techniques

Depends on D1 and is informed by D2.

Direction 2 examines how honesty degrades when models are trained against incentives that favour misleading behaviour. This direction tests whether two techniques preserve honesty during such training.

Recontextualization modifies the training context so that locally rewarded behaviour is interpreted as permissible or required only within that restricted context. The intended effect is to let the model learn from the training signal without generalising the undesirable behaviour to deployment (Azarbal et al., 2025).

Inoculation prompting adds a training-time instruction that elicits or permits the behaviour encouraged by the local reward, then removes that instruction after fine-tuning (Wichers et al., 2025; Tan et al., 2025).

Test whether either technique preserves honest deployment behaviour when the model is subsequently trained on:

- task rewards indifferent to accurate disclosure;
- weak verifiers that can be gamed;
- user-satisfaction rewards that favour agreement;
- reporting tasks in which favourable claims receive higher reward;
- capability training that produces incidental pressure toward brevity or omission.

Measure several parts of honesty separately:

- factual correspondence;
- reporting of internally represented evidence;
- calibration and uncertainty;
- resistance to sycophancy;
- disclosure of decision-relevant information;
- avoidance of strategic omission;
- avoidance of misleading framing.

The techniques may protect some components while degrading others. For example, explicit falsehoods may decrease while omission or evasive language increases.

D6: Sub-construct structure and generalisation

Honesty may consist of several partially independent components:

- factual truthfulness;
- epistemic calibration;
- willingness to acknowledge uncertainty;
- resistance to sycophancy;

- disclosure of relevant evidence;
- non-deception;
- avoidance of strategic omission;
- transparency about goals and limitations;
- non-manipulative communication.

Test whether these behaviours correspond to separate adjustable settings inside the model or to one shared disposition.

This requires interventions that target one sub-construct at a time. For example:

- train calibrated uncertainty and test whether resistance to strategic lying improves;
- train resistance to sycophancy and test whether relevant disclosure improves;
- train factual accuracy and test whether the model becomes less strategically misleading;
- train non-deception and test whether it generalises to honest reporting under uncertainty.

The project should also examine generalisation from honesty to adjacent positive values. Candidate transfer domains include:

- non-manipulation;
- corrigibility;
- cooperation;
- procedural fairness;
- respect for user autonomy;
- appropriate concern for affected stakeholders.

Positive transfer would suggest a broader underlying disposition. Selective transfer would suggest that apparently related values are implemented through separable policies or representations.

D7: Evaluation methodology

A credible study of honesty requires knowing how failures were elicited, how they were verified, and how they were classified (Jakkli et al., 2026).

Use three layers of evaluation.

1. Broad behavioural benchmarks

Construct a diverse evaluation set covering:

- factual questions with known answers;
- questions containing false user assumptions;
- uncertainty and calibration tasks;
- hidden-information reporting tasks;
- conflicts between truth and user preference;
- conflicts between truth and institutional reward;
- selective-disclosure scenarios;
- trade-offs involving privacy, safety, and confidentiality;
- situations in which accurate reporting reveals the model's own failure.

Benchmarks should separate capability from honesty. A model cannot be classified as dishonest when it lacks the information or reasoning ability required to answer correctly.

2. Long-form adversarial interaction

Use a probing agent to apply sustained pressure, following the general structure of automated auditing tools such as Petri (Fronsdal et al., 2025).

The probing agent may:

- express a strong preferred conclusion;
- claim institutional authority;
- threaten negative consequences for disclosure;
- reward agreement;
- ask the model to conceal earlier errors;
- gradually shift from benign summarisation to strategic misreporting;
- remove apparent oversight or evaluation.

Measure whether honesty degrades across turns and whether the model changes strategy from direct falsehood to omission, hedging, or refusal.

3. Active failure discovery

Actively search for cases that cause the model to mislead, omit relevant evidence, overstate confidence, or follow false premises.

Score failures using a predefined taxonomy. At minimum, distinguish:

- sincere factual error;
- unsupported confidence;
- sycophantic agreement;
- direct falsehood;
- strategic omission;
- misleading framing;
- evasive response;
- generic refusal;
- evaluation-conditioned honesty;
- role-conditioned honesty.

Where training data can be controlled, trace failures back to the features or examples that caused them. This can help distinguish a learned deceptive tendency from shortcut learning, user compliance, or a narrow training artefact.

Report failure rates separately by evaluation method, behavioural category, domain, and intervention. Validate flagged conversations manually. Do not collapse all outcomes into one aggregate honesty score, because different interventions may reduce one failure mode while increasing another.

Expected contribution

The broader research programme aims to determine:

1. what produces apparently honest behaviour in language models;
2. whether honesty is implemented as a general disposition or a collection of context-specific policies;

3. how honesty changes under misspecified or adversarial training;
4. which interventions instil honesty most effectively;
5. which techniques preserve it during later training;
6. whether honesty generalises to adjacent positive values;
7. how honesty failures should be elicited, classified, and measured.

The programme is useful under several possible findings. Evidence for a context-general and durable honesty disposition would support work on character-level interventions and preservation. Evidence for shortcuts, sycophancy, or evaluation conditioning would shift attention toward training-data design and contextual robustness. Evidence that relevant knowledge remains latent would prioritise elicitation and incentive compatibility. Evidence that explicit lies are replaced by omission or refusal would show that current evaluations measure only a narrow part of the construct.

The central objective is therefore to develop interventions only after identifying the mechanisms that produce the observed behaviour.