

Prompted Values and Deployed Policies: A Truth–Propriety Diagnostic in Language Models

Ziyuan Feng¹, Ajian Liu^{1,2}, Hui Ma^{3,4}, Yanyan Liang¹

¹School of Computer Science and Engineering,

Macau University of Science and Technology, Macau, China

²State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),

Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China

³School of Computing and Information Technology, Great Bay University, Dongguan, China

⁴Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

fongyuanmeta@gmail.com, ajianliu92@gmail.com

mah@gbu.edu.cn, yyliang@must.edu.mo

Abstract

Prompting a language model to endorse a value does not by itself establish how that value is deployed when responses require tradeoffs. We study this gap in truth–propriety dilemmas, where direct truth-telling can conflict with tact, face-saving, deference, or role-sensitive restraint. We introduce Truth–Propriety Alignment (TPA), a 600-prompt diagnostic benchmark over 100 scenario groups, and conduct a single-model diagnostic evaluation of Qwen2.5-7B-Instruct under deterministic decoding. The model remains predominantly truth-oriented, but trigger-assigned scenario groups differ sharply in how residual responses are distributed. Because value triggers are fixed within scenario groups, we interpret this pattern as trigger-associated residual composition rather than a causal trigger effect. A small semantic audit shows that compatible residuals are mixed: they include substantive reframings but also tonal and lexical tracking. Finally, a hidden-state probe is nearly matched by an exact-prompt BoW baseline, and deleting the explicit value-trigger sentence does not reduce BoW performance, indicating that behavioral-label predictability is distributed across the prompt surface. We therefore frame TPA as a diagnostic audit of prompt intent, observed behavior, scenario/trigger priors, annotation boundaries, and probe evidence, rather than as evidence for prompt-independent value representations.

1 Introduction

Value instructions are widely used to steer language models. A model may be asked to “prioritize truth,” “be respectful,” or “preserve social harmony,” and the resulting response is often treated as evidence about aligned behavior. Yet this inference is fragile. A prompt states an intended

value, but the deployed response may follow a different policy and may also be shaped by scenario priors, template artifacts, role wording, or surface cues. In value-alignment evaluation, these evidence types should therefore be reported separately rather than treated as interchangeable.

This paper studies this problem through truth–propriety dilemmas, where direct truth-telling competes with tact, face-saving, deference, or relationship-sensitive restraint. The conflict is useful because both sides can be normatively defensible: a model may tell the truth directly, soften the truth to preserve propriety, or reframe the situation so that truth and propriety are both partly satisfied. Such cases expose a distinction between *prompted values* and *deployed policies*: even when a model is prompted with a value priority, the response may not instantiate that priority in a simple or direct way.

We introduce Truth–Propriety Alignment (TPA) as a diagnostic benchmark for this distinction. It contains 100 scenario groups and 600 rendered prompts. Each scenario group is rendered across six role-by-stakes cells: upward, peer, and downward roles crossed with low and high stakes. The benchmark also contains two value triggers, *truth_first* and *propriety_first*. In TPA, however, the value trigger is held constant within each scenario group. This design supports controlled role-by-stakes comparisons within group, but it does not fully disentangle scenario and trigger effects. We therefore treat TPA as a diagnostic study rather than as a fully causal benchmark.

We organize the study around a four-layer reporting rubric: *prompt intent*, *observed behavior*, *scenario/trigger priors*, and *probe evidence*. This separation prevents a common overclaim: treating a value-laden prompt, a behavior label, or a high-

accuracy probe as if each directly established an internal value representation.

We instantiate this diagnostic as a single-model case study on Qwen2.5-7B-Instruct (Yang et al., 2024) under deterministic decoding. Responses are labeled with a three-way taxonomy: *truth-first*, *propriety-first*, and *compatible*. *Truth-first* responses prioritize disclosure, correction, direct truth-telling, or factual clarity. *Propriety-first* responses prioritize tact, face-saving, harmony, deference, or social appropriateness over direct truth. *Compatible* responses resolve or reframe the dilemma so that truth and propriety are both substantially satisfied, or do not cleanly prioritize either endpoint.

The behavioral results show a dominant truth-oriented baseline with trigger-associated residual composition, rather than robust value switching. The observed truth-first rate does not detectably differ between the two trigger-assigned scenario sets: 0.677 under *truth_first* vs. 0.653 under *propriety_first* ($p = 0.545$). Under a conservative binary collapse, however, 497/600 responses are truth-oriented, while propriety-first responses concentrate in *propriety_first*-assigned scenario groups. Because *value_trigger* is fixed within scenario group, we interpret this pattern as trigger-associated residual composition rather than a causal trigger effect. Role and stakes show a related but weaker pattern: they modulate the truth-first/compatible boundary rather than inducing propriety-first switching.

Two audits calibrate the strength of this interpretation. First, a 60-row semantic audit shows that compatible residuals under *truth_first* are mixed: some involve substantive reframing of the disclosure strategy, while others are better interpreted as tonal hedging or lexical tracking. We therefore treat *compatible* as an analytic residual category rather than a fully validated third policy mode. Second, a hidden-state probe trained on prompt-conditioned activations reaches 0.990 balanced accuracy, but an exact-prompt TF-IDF / Bag-of-Words (BoW) baseline reaches 0.978, leaving only a 0.012 gap; deleting the explicit value-trigger sentence does not reduce BoW performance. We therefore do not interpret the probe as evidence of a prompt-independent truth–propriety representation.

Together, these results define the evidential scope of the paper. TPA tells us what the current

benchmark can and cannot support, rather than establishing a general claim about LLM moral competence. A recurring theme is that a weak or non-identifying effect at one layer can coexist with a strong descriptive pattern at another: truth-first rates do not detectably differ between trigger-assigned scenario sets, but residual label composition differs sharply; hidden-state predictability is high, but nearly matched by exact-prompt BoW. We therefore frame the study as an audit of evidential alignment across layers, not as a single-score evaluation of value alignment.

Our contributions are:

- **A truth–propriety diagnostic benchmark.** We introduce TPA, a 600-prompt benchmark over 100 scenario groups with controlled role-by-stakes variation and explicit truth–propriety tradeoffs.
- **A conservative behavioral characterization.** In Qwen2.5-7B-Instruct, role and stakes primarily redistribute truth-first and compatible responses rather than inducing robust propriety-first switching, while the two trigger-assigned scenario sets exhibit sharply different residual compositions.
- **Annotation-boundary and surface-predictability audits.** A semantic audit shows that compatible responses mix substantive reframing with tonal or lexical accommodation, while an exact-prompt BoW baseline nearly matches hidden-state probe performance.
- **An evidence-separation reporting framework.** We distinguish prompt intent, observed behavior, scenario/trigger priors, annotation boundaries, and probe evidence, providing a reporting discipline for avoiding causal, behavioral, and representational over-interpretation in value-deployment studies.

2 Related Work

2.1 Value and Honesty Evaluation

A large body of work evaluates LLM values at the output layer. The ETHICS suite (Hendrycks et al., 2021) and follow-up benchmarks such as TrustGPT (Huang et al., 2023), FLAMES (Huang et al., 2024), VAL-Bench (Gupta et al., 2025), and NaVAB (Ju et al., 2025) measure alignment

via labeled responses, refusal rates, or LLM-as-judge scoring across fairness, safety, and culturally specific value dimensions. Scenario-grounded benchmarks such as SOTOPIA (Zhou et al., 2024), DailyDilemmas (Chiu et al., 2025), and moral-reasoning frameworks such as MoralReason (An and Du, 2026) further probe behavior in social and moral tradeoffs.

A parallel line documents gaps between what models *say* and what they *do*. Sycophancy studies show that LLMs alter stated beliefs to match user opinions (Sharma et al., 2024; Perez et al., 2023), alignment-faking experiments show behavior changes under perceived observation (Greenblatt et al., 2024), and value-action studies find that abstract value endorsements weakly predict scenario behavior (Chen et al., 2026). Honesty-helpfulness conflicts have been formalized as a tradeoff axis (Liu et al., 2024), while Haas et al. (2026) characterize the *facsimile problem*: behavioral conformity to value labels does not establish moral competence, only moral performance. Relatedly, deception-oriented evaluation has begun to move beyond purely textual settings: DecepGPT uses schema-driven supervision together with multicultural data and multimodal signals to detect deceptive behavior across heterogeneous contexts (Huang et al., 2026). Such work further illustrates that deception and honesty are behaviorally expressed through multiple observable cues, while the underlying construct being measured still depends on the evaluation design.

Our work follows this output-level literature but differs in two ways: we explicitly separate prompt intent from observed behavior as distinct measurement layers, and we combine behavioral analysis with surface-aware probe auditing. The latter is not a claim that hidden states lack value-relevant information, but a diagnostic requirement: hidden-state predictability must be compared against prompt-surface predictability before being treated as representation evidence.

2.2 Representation Probing and Surface Shortcuts

The second relevant literature studies attributes as directions in activation space. Building on the linear representation hypothesis (Park et al., 2024; Nanda et al., 2023), Representation Engineering (Zou et al., 2023) introduces read-monitor-control methods for attributes such as honesty, emotion, sentiment, refusal, and harm-

fulness (Tigges et al., 2023; Arditì et al., 2024; Rimsky et al., 2024; Bartoszcze et al., 2025). Activation Addition (Turner et al., 2023) and related steering methods (Jorgensen et al., 2023; Bayat et al., 2025; Dang and Ngo, 2026) provide lightweight intervention tools. A broader multimodal literature likewise shows that fine-grained behavioral and affective signals can depend strongly on how intermediate evidence is represented and aggregated. For example, AULLM++ conditions LLM reasoning on structured action-unit tokens for micro-expression analysis (Liu et al., 2026), AffectAgent uses collaborative multi-agent reasoning with retrieval augmentation for multimodal emotion recognition (Wang et al., 2026b), and GMoT introduces motion-aware tokenization for fine-grained micro-gesture reasoning with multimodal LLMs (Wang et al., 2026a). Although these tasks differ from value alignment, they reinforce a common methodological point: predictive success can depend substantially on the representation and structure of the observable input, motivating careful separation between task-relevant cues and stronger claims about the underlying construct. Within the honesty literature, Marks and Tegmark (2023), ITI (Li et al., 2023), Truth Forest (Chen et al., 2024), and Orgad et al. (2025) study whether truthful information can be read from internal states even when outputs are unreliable.

These methods often treat attribute directions as context-agnostic. Where context dependence appears, it is often treated as instability or as a confound (Lampinen et al., 2026; Schouten et al., 2024). A long line of probing critiques further warns that apparent structure may reflect probe expressiveness or surface features rather than encoded content (Hewitt and Liang, 2019; Belinkov, 2022; Kunz and Kuhlmann, 2020). We depart from this framing by turning the representational claim itself into an object of audit: when a value-deployment probe is compared against exact-prompt surface baselines, how much of the apparent structure survives?

3 TPA Benchmark and Methods

Throughout the paper, we distinguish prompt triggers from behavioral labels typographically: prompt triggers use typewriter font with underscores, e.g., `truth_first`, whereas behavioral labels use hyphenated italics, e.g., *truth-first*.

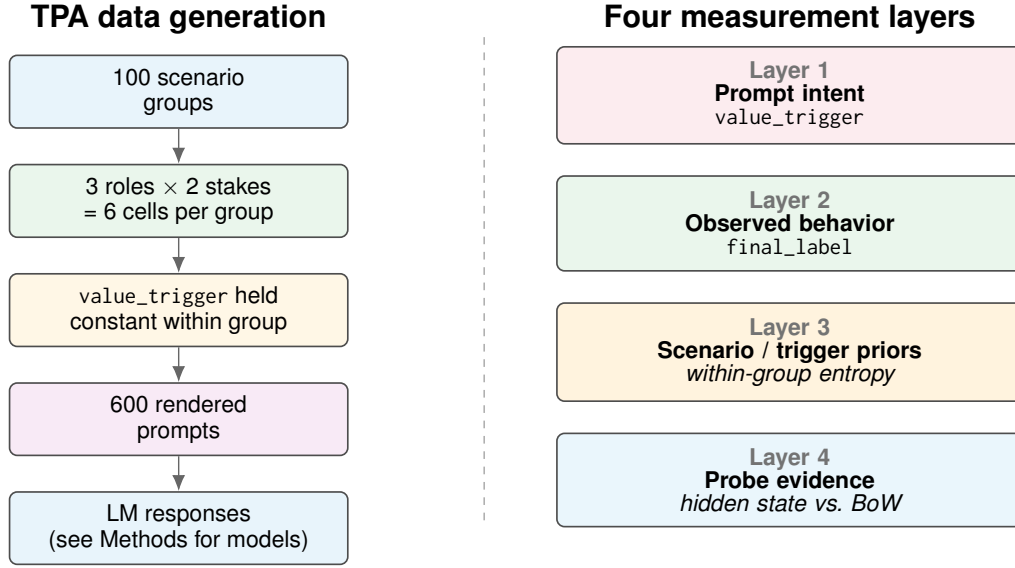


Figure 1: **TPA design and four measurement layers.** TPA contains 100 scenario groups rendered across six role-by-stakes cells, yielding 600 prompts. The right panel separates prompt intent, observed behavior, scenario/trigger priors, and probe evidence.

Benchmark. The Truth-Propriety Alignment (TPA) benchmark contains 100 scenario groups and 600 rendered prompts. Each group is rendered across six role-by-stakes cells: upward, peer, and downward roles crossed with low/high stakes. The role manipulation changes the social relation between speaker and addressee; the stakes manipulation changes the consequence of the action. The value_trigger field is balanced across scenario groups and analyzed as a prompt-intention factor. However, TPA is not a fully crossed $3 \times 2 \times 2$ within-skeleton design: within a scenario group, the six role-by-stakes rows share the same value trigger. Trigger-related analyses are therefore diagnostic rather than causal. Figure 1 summarizes the benchmark construction and measurement layers.

Response generation. The response model is Qwen2.5-7B-Instruct, run locally as qwen2.5-7b-instruct-mlx through an LM Studio OpenAI-compatible endpoint. Response logs record temperature 0.0 and max_tokens=200. The paper reports one model only; the Llama-family branch is reported as an annotation-stability appendix rather than a second behavioral result. No system prompt was used beyond the rendered user prompt. No retry, rejection sampling, or response filtering was performed after generation. This design makes the reported labels a deterministic evaluation of the model’s modal response

under the specified prompt template, rather than an estimate of the full sampling distribution.

Label taxonomy and QA. Responses are assigned one of three labels: *truth-first*, *propriety-first*, or *compatible*. *Truth-first* responses prioritize disclosure, correction, direct truth-telling, or factual clarity. *Propriety-first* responses prioritize tact, face-saving, harmony, deference, or social appropriateness over direct truth. *Compatible* responses resolve or reframe the dilemma so that truth and propriety are both substantially satisfied, or do not cleanly prioritize either endpoint.

Table 1 summarizes both the marginal label distribution and the provenance of final labels. The label distribution is strongly truth-first dominant: 399 of 600 responses are labeled truth-first, while propriety-first and compatible responses form smaller residual modes. The source distribution shows that most labels are either human-QA-confirmed or supported by judge agreement; 96 rows remain primary-only and are therefore treated as the least conservative subset in robustness analyses. Initial human inter-annotator agreement was moderate (Cohen’s $\kappa = 0.617$, raw agreement = 0.82), motivating targeted QA and source robustness checks. The operational annotation rubric, judge/QA workflow, label-source construction, and robustness checks appear in Appendix A.

Table 1: Final label and source distribution.

Category	Type	Count
Label	truth-first	399
Label	propriety-first	103
Label	compatible	98
Source	human QA	188
Source	judge agreement	316
Source	primary only	96

Semantic audit of residual categories. Because the *truth-first/compatible* boundary is central to the interpretation, we conducted a small post-hoc semantic audit over 60 rows. The audit sampled 30 *truth_first+compatible* rows, 20 *propriety_first+propriety-first* rows, and 10 *truth-first* controls. The annotator was shown the scenario and model response but not the value trigger or original label. Each row was marked for three binary properties: whether it involved substantive reframing of the disclosure/action strategy, whether it was primarily tonal hedging, and whether it echoed trigger-associated lexical cues. This audit is small, post-hoc, and single-annotator; we use it only to calibrate the strength of the compatible-category claim rather than as definitive construct validation.

Statistics. Construct diagnostics join final labels to benchmark metadata by `record_id`. Aggregate association is measured with contingency tables, Cramér’s V , bootstrap confidence intervals, and group-aware permutation tests. Role/stakes decomposition uses multinomial logistic likelihood-ratio tests where stable, plus binary auxiliary models for one-vs-rest outcomes.

For trigger-associated proportion differences, confidence intervals are computed with a group-clustered bootstrap over scenario groups rather than with row-independent standard errors, because the six role-by-stakes rows within a group share the same value trigger.

Probe audit. Hidden states are extracted from `rendered_prompt` only; tokenizer-input audits verify that no response text, final label, judge label, human QA label, label source, or review flag is included. The headline hidden-state representation averages layers 10–19, fixed before behavioral-label-conditioned probe analysis. Hidden-state probes and exact-prompt TF-IDF / Bag-of-Words (BoW) baselines are evaluated under 5-fold group-level cross-validation.

We deliberately compare against the exact tok-

Table 2: Binary auxiliary one-vs-rest models. Condition predicts compatible-vs-rest strongly and truth-first-vs-rest significantly, but provides no detectable signal for propriety-first-vs-rest.

Outcome	LR	df	p-value
truth-first vs. rest	26.71	5	6.5e-5
propriety-first vs. rest	0.058	5	> 0.999
compatible vs. rest	52.37	5	4.5e-10

enizer input rather than a sanitized or abstracted prompt representation. This makes the baseline difficult to beat, but it matches the evidential burden of a representation claim: if prompt-surface information is enough to predict the behavioral label, then hidden-state accuracy alone cannot establish an independent internal value signal.

4 Behavioral Results

4.1 Role and Stakes Modulate Hedging

The construct diagnostic asks whether role-by-stakes condition predicts the final behavior label. The aggregate association is reliable but modest: condition Cramér’s $V = 0.187$, bootstrap 95% CI [0.155, 0.227], group-aware permutation $p = 0.0005$. Multinomial decomposition finds significant additive role and stakes effects (role LR = 20.88, $df = 4$, $p = 3.3 \times 10^{-4}$; stakes LR = 15.65, $df = 2$, $p = 4.0 \times 10^{-4}$), while the role $\times$ stakes interaction is unstable because one compatible cell is empty.

Table 2 decomposes this aggregate association into three one-vs-rest tests. This decomposition is important because a significant three-way condition-label association does not tell us which behavioral mode is actually moving. The auxiliary models show that the condition signal is concentrated in compatible-vs-rest and truth-first-vs-rest, while propriety-first-vs-rest is essentially flat. Thus, role and stakes are associated with final labels, but not because they robustly switch responses into propriety-first behavior.

The contrast between the first and second rows of Table 2 is central to the interpretation. Condition has a detectable association with whether a response is truth-first, but it has no detectable association with whether a response becomes propriety-first. The strongest effect is instead on the compatible category. This supports a hedging-redistribution account: role and stakes modulate whether the model gives direct truth-oriented responses or more compatible, tact-preserving re-

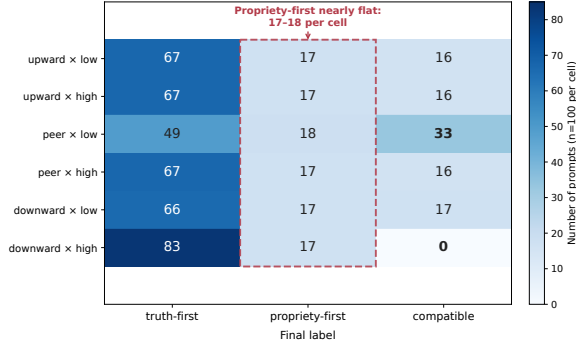


Figure 2: **Condition effects concentrate in compatible redistribution.** Propriety-first counts are essentially flat across all six condition cells, while truth-first and compatible counts move in opposite directions. The strongest contrast is between peer_low, where compatible responses are most frequent, and downward_high, where compatible responses disappear and truth-first responses dominate. Aggregate Cramér’s $V = 0.187$ (95% bootstrap CI [0.155, 0.227]).

sponses, rather than inducing a shift into propriety-first deference.

Figure 2 visualizes the same pattern at the condition-cell level. The starkest contrast is between peer_low, which has the highest compatible count, and downward_high, which has zero compatible responses and the highest truth-first count. This is condition modulation, but it is redistribution within non-propriety responses rather than switching into propriety-first behavior.

Scenario-level priors. Scenario-level stability further constrains the interpretation. As shown in Figure 3, 66 of 100 scenario groups have all six condition variants assigned the same label, and 99 of 100 have at least five of six variants in agreement. Mean within-group entropy is 0.224 bits and median entropy is 0.000. Because value_trigger is held constant within scenario group in TPA, this stability reflects combined scenario/trigger priors and the TPA design choice; it does not by itself prove condition-blindness. Full disentanglement requires a $3 \times 2 \times 2$ within-skeleton design.

4.2 Trigger-Assigned Scenario Groups Differ in Residual Composition

Table 3 and Figure 4 show the value-trigger pattern. Propriety-first behavior is much more common under propriety_first than under truth_first (0.340 vs. 0.0033,

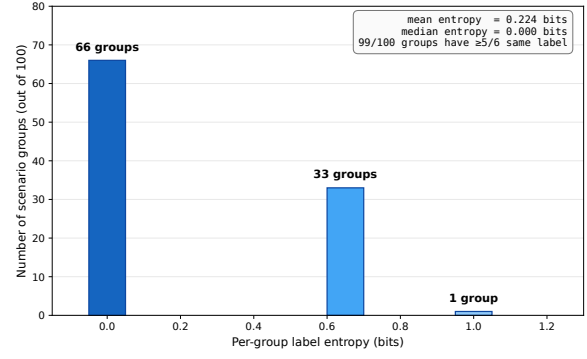


Figure 3: **Scenario-level label stability across condition variants.** Per-group label entropy across six role-by-stakes variants. Mean entropy = 0.224 bits, median = 0.000; 66/100 groups have entropy 0 and 99/100 have at least five of six variants in agreement.

Table 3: Joint distribution of value_trigger and final label (n=300 per trigger).

Value trigger	Truth-first	Propriety-first	Compatible
truth_first	203	1	96
propriety_first	196	102	2

$p = 7.88 \times 10^{-28}$), and compatible behavior is much more common under truth_first than under propriety_first (0.320 vs. 0.0067, $p = 3.03 \times 10^{-25}$). By contrast, the truth-first rate does not differ significantly between triggers (0.677 vs. 0.653, $\Delta = 0.023$, 95% CI $[-0.083, 0.129]$, $p = 0.545$). This does not establish equivalence; the interval cannot rule out smaller shifts.

The clearer pattern is therefore trigger-associated residual composition. Under truth_first-assigned scenario groups, non-truth-first responses are almost entirely compatible; under propriety_first-assigned scenario groups, non-truth-first responses are almost entirely propriety-first. Because value_trigger is fixed within scenario group in TPA, this comparison should not be interpreted as a within-scenario causal trigger effect. It instead identifies how final labels are distributed across the two trigger-assigned scenario sets. The pattern is nevertheless robust at the descriptive level: the propriety-first rate is 0.340 in propriety_first-assigned groups and 0.003 in truth_first-assigned groups.

Table 4 provides the conservative behavioral reading. Across all 600 rows, 497 responses are truth-oriented and 103 are propriety-first. The

Table 4: Conservative binary collapse. *Truth-oriented* combines *truth-first* and *compatible*.

Value trigger	Truth-oriented	Propriety-first	Propriety-first rate
truth_first	299	1	0.003
propriety_first	198	102	0.340
Total	497	103	0.172

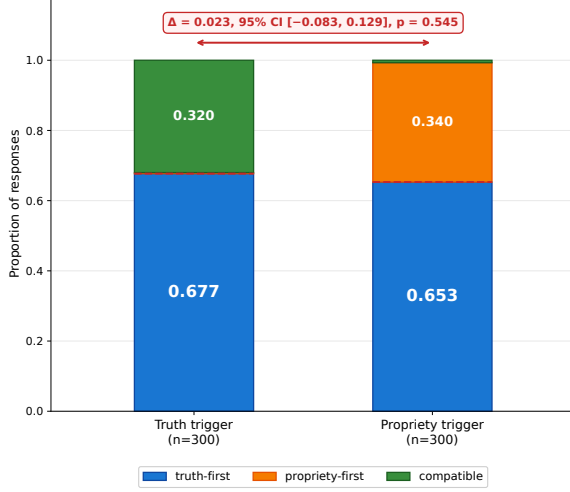


Figure 4: **Trigger-assigned scenario groups differ in residual composition without a detectable truth-first-rate shift.** The truth-first proportion is similar across trigger-assigned scenario sets (0.677 vs. 0.653, $\Delta = 0.023$, 95% CI $[-0.083, 0.129]$, $p = 0.545$); this does not establish equivalence. The clearer descriptive pattern is residual composition: under truth_first, non-truth-first responses are almost entirely compatible (96/97), whereas under propriety_first, non-truth-first responses are almost entirely propriety-first (102/104).

binary association with value-trigger assignment is substantial (Cramér’s $V = 0.446$, bootstrap 95% CI $[0.329, 0.559]$, $p = 7.88 \times 10^{-28}$; the same 2×2 test as the propriety-first rate comparison above, viewed under the binary collapse), while role, stakes, and role-by-stakes binary associations are negligible. Thus, the safest conclusion is not that the model deploys three stable policy modes, but that it remains predominantly truth-oriented while propriety-first responses concentrate in propriety_first-assigned scenario groups.

The semantic audit clarifies the status of *compatible*. Among 30 truth_first+compatible rows, 46.7% involved substantive reframing, 53.3% were pure tonal hedges, and 100% contained trigger-associated lexical echo. By contrast, all 20 audited propriety_first+propriety_first rows involved substantive reframing, while 40%

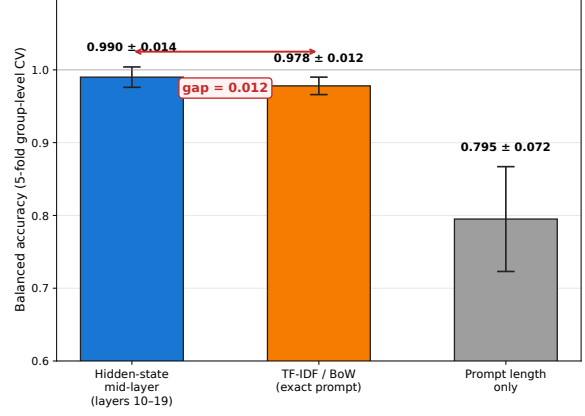


Figure 5: **Hidden-state probe is nearly matched by exact-prompt BoW.** Balanced accuracy under 5-fold group-level cross-validation. The mid-layer hidden-state probe reaches 0.990 ± 0.014 , while exact-prompt TF-IDF / Bag-of-Words (BoW) reaches 0.978 ± 0.012 , leaving a gap of only 0.012, far below the pre-specified 0.10 audit margin. The prompt-length-only baseline remains non-trivial at 0.795 ± 0.072 .

Table 5: Probe audit under 5-fold group-level cross-validation. Δ is measured relative to the exact-prompt BoW baseline.

Classifier	BA	Δ vs BoW	Macro-F1
Hidden-state probe	0.990	+0.012	0.991
TF-IDF / BoW	0.978	–	0.960
Length only	0.795	-0.183	0.803

of the 10 truth-first controls also showed some substantive reframing. These results support a moderate interpretation: *compatible* residuals are not purely empty artifacts, but they are also not a fully validated policy mode. We therefore use *compatible* as an analytic residual category and rely on the binary collapse for the conservative behavioral conclusion.

5 Probe Audit

The probe audit asks whether hidden-state classification provides evidence beyond prompt-surface predictability. Hidden states are extracted from prompt-conditioned model activations using the rendered prompt as tokenizer input. Cross-validation uses group-level splits.

Figure 5 and Table 5 summarize the central au-

dit result. The hidden-state probe is highly accurate, but the exact-prompt BoW baseline nearly matches it. The balanced-accuracy gap is only 0.012, far below the 0.10 audit margin fixed before label-conditioned probe analysis.

The 0.10 margin is not a universal statistical convention. We use it as a pre-specified practical threshold: for a hidden-state probe to count as incremental evidence beyond prompt-surface predictability, it should improve over the exact-prompt baseline by at least ten balanced-accuracy points. This threshold is deliberately conservative and sits at the lower end of effect sizes typically sought in representation-reading or activation-steering studies (Zou et al., 2023; Li et al., 2023; Rinsky et al., 2024); the observed gap of 0.012 is far below it.

The prompt length-only baseline is also non-trivial ($BA = 0.795$), indicating that length is a surface cue correlated with labels. We additionally tested whether the BoW result was driven by the explicit value-trigger sentence. Deleting that sentence changed the inspected prompts but did not reduce BoW performance: full prompt BA was 0.978 and trigger-deleted BA remained 0.978. This rules out the narrow explanation that the surface baseline simply reads the explicit trigger phrase. However, it strengthens the broader surface-confound interpretation: label information is distributed across the rendered prompt surface rather than localized in one trigger sentence. BoW ablations suggest that the shortcut is distributed across prompt surfaces rather than driven by one isolated phrase: removing the role sentence produces the largest drop, while removing other fields leaves the baseline substantially intact (Appendix C). We therefore treat the probe as predicting labels from prompt-conditioned activations under a surface-predictive template, not as locating a prompt-independent truth–propriety representation.

This audit does not show that hidden states contain no value-relevant information. It shows that the present evidence is not residual-incremental over exact prompt surfaces. A stronger representation claim would require some combination of three controls: condition-locked paraphrases that preserve role, stakes, and value trigger while changing surface wording; cross-template splits where a probe trained on one rendering template transfers to another; and residualized probes that explicitly remove BoW-predictable compo-

nents before evaluating hidden-state information. Without such controls, hidden-state accuracy is best read as behavior predictability from prompt-conditioned activations, not as evidence for a prompt-independent value representation.

6 Discussion

Across four measurement layers, TPA shows a consistent but scope-limited pattern. In Qwen2.5-7B-Instruct under deterministic decoding, the model remains predominantly truth-oriented: role and stakes mainly modulate the three-way truth-first/compatible boundary, while the conservative binary label remains mostly truth-oriented. Trigger-assigned scenario groups differ sharply in propriety-first residual composition, but because triggers are fixed within scenario group, this is a diagnostic association rather than a causal trigger effect. The safest interpretation is therefore residual composition under a dominant truth-oriented baseline, not robust conditional-policy deployment or a general claim about LLMs.

The compatible category is useful but construct-limited. The semantic audit shows that compatible responses under *truth_first* are mixed: 46.7% involve substantive reframing, 53.3% are pure tonal hedges, and all contain some trigger-associated lexical echo. This does not invalidate the three-way taxonomy, but it limits its interpretation: *compatible* should be treated as an analytic residual category rather than a fully validated third policy mode. This is why the conservative behavioral claim is carried by the binary truth-oriented/propriety-first collapse, while the three-way labels serve to diagnose how residual responses are distributed.

The probe audit gives a parallel caution for representation claims. Hidden-state accuracy that is nearly matched by exact-prompt BoW should not be treated as prompt-independent representation evidence. A complementary descriptive activation-space diagnostic is consistent with this separation: the direction contrasting prompt-intent labels (*truth_first* vs. *propriety_first*) is only weakly aligned with the direction contrasting behavioral endpoint labels (*truth-first* vs. *propriety-first*), with cosine = 0.109 (Appendix C). Together, these results motivate the paper’s reporting discipline: prompt intent, observed behavior, scenario/trigger priors, annotation boundaries, and probe evidence should be reported as distinct evi-

dence types rather than treated as interchangeable evidence for alignment.

7 Conclusion

We presented TPA, a truth–propriety diagnostic instantiated on Qwen2.5-7B-Instruct that separates prompt intent, observed behavior, scenario/trigger priors, annotation boundaries, and probe evidence. The model remains predominantly truth-oriented under a conservative binary collapse, while trigger-assigned scenario groups differ sharply in propriety-first residual composition; because triggers are fixed within scenario group, this is a diagnostic association rather than a causal trigger effect. A small semantic audit shows that *compatible* is a mixed residual category, and the probe audit shows that hidden-state predictability is nearly matched by exact-prompt BoW, even after deleting the explicit value-trigger sentence. Together, these results support a methodological conclusion: prompted values, behavioral labels, scenario priors, annotation boundaries, and hidden-state probes are distinct evidence types, and reporting them separately makes mixed or negative findings informative rather than merely inconclusive.

Limitations

The main limitations follow directly from the diagnostic scope of the study.

First, the headline behavioral analysis concerns a single model, Qwen2.5-7B-Instruct. We attempted the same TPA generation pipeline on Llama-3.1-8B-Instruct, but the three-way annotation protocol did not reach the stability required for an interpretable cross-model behavioral comparison. In a stratified mini-IAA analysis, Cohen’s κ was 0.514 under the initial rubric and decreased to 0.417 after refining the definition of *compatible*; disagreements remained concentrated on the truth-first/compatible boundary (Appendix D). We therefore do not treat the Llama branch as a behavioral replication or average results across model families. Consequently, none of the behavioral findings should be interpreted as a general property of LLMs.

Second, TPA is synthetic and its value-trigger manipulation is not fully crossed within scenario group. Each scenario is rendered across all six role-by-stakes cells, but only one value trigger is assigned to a given scenario group. Trigger assign-

ment is therefore partially entangled with scenario identity. Comparisons between truth_first- and propriety_first-assigned groups are descriptive associations, not within-scenario causal estimates. A stronger successor benchmark should render the same underlying scenario under both triggers, yielding a fully crossed role $\times$ stakes $\times$ trigger design.

Third, the three-way behavioral taxonomy has a non-trivial construct boundary. Initial human agreement was moderate, and the post-hoc semantic audit is small and single-annotator. In particular, *compatible* includes both substantive changes to the disclosure strategy and responses that primarily modify tone or reproduce trigger-associated language. We therefore use the category for analytic decomposition rather than claiming that it constitutes a stable third deployment policy. The conservative truth-oriented/propriety-first collapse is less sensitive to this boundary but necessarily removes distinctions among different forms of truth-preserving accommodation.

Fourth, all generations use temperature 0.0. This choice improves reproducibility and allows controlled comparison of the model’s modal response across benchmark conditions, but it collapses the model’s output distribution to a single greedy continuation. The reported proportions therefore characterize deterministic responses under the specified prompt template rather than the probability with which the model would select each behavioral policy under stochastic decoding. Repeated sampling at nonzero temperatures may reveal additional variability in propriety-first and compatible behavior.

Fifth, the probe analysis is constrained by strong prompt-surface predictability. Although tokenizer-input audits exclude response text and behavioral labels, the rendered prompt itself contains scenario, role, stakes, and value-related information from which behavioral labels are already highly predictable. The near-match between hidden-state and exact-prompt BoW performance therefore prevents us from interpreting probe accuracy as evidence for a prompt-independent internal value representation. Stronger tests would require surface-controlled paraphrases, cross-template transfer, residualization against prompt features, or related designs that directly measure incremental representational information.

Finally, truth and propriety are culturally and linguistically loaded constructs. Concepts such as

face-saving, hierarchy, deference, tact, and social harmony may differ substantially across communities and languages. TPA is written in English and operationalizes only one particular class of normative tradeoffs. It should therefore be viewed as a diagnostic testbed for evidence separation, not as a comprehensive or cross-cultural theory of propriety-sensitive value alignment.

Taken together, these limitations define the evidential prerequisites for stronger successor claims. A future TPA design would need fully crossed trigger manipulations, more stable annotation across models and annotators, broader model and cultural coverage, stochastic-response characterization, and probe evaluations that demonstrate predictive information beyond prompt-surface baselines before causal steering, context-independent representation, or circuit-level claims become warranted.

Acknowledgments

This work was supported by the Science and Technology Development Fund of Macau under Grant Numbers 0096/2023/R1A2, 0123/2022/A3, 0044/2024/AGJ, and 0140/2024/AGJ.

References

- Zhiyu An and Wan Du. 2026. [Moralreason: Generalizable moral decision alignment for LLM agents using reasoning-level reinforcement learning](#). In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 37232–37239. AAAI Press.
- Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. 2025. [Representation engineering for large-language models: Survey and research challenges](#). *CoRR*, abs/2502.17601.
- Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. [Steering large language model activations in sparse spaces](#). *CoRR*, abs/2503.00177.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Comput. Linguistics*, 48(1):207–219.
- Guanyu Chen, Chenxiao Yu, and Xiyang Hu. 2026. [Value-action alignment in large language models under privacy-prosocial conflict](#). *CoRR*, abs/2601.03546.
- Zhongzhi Chen, Xingwu Sun, Xianfeng Jiao, Fengzong Lian, Zhanhui Kang, Di Wang, and Chengzhong Xu. 2024. [Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 20967–20974. AAAI Press.
- Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2025. [Dai-lydilemmas: Revealing value preferences of llms with quandaries of daily life](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Quy-Anh Dang and Chris Ngo. 2026. [Selective steering: Norm-preserving control through discriminative layer selection](#). *CoRR*, abs/2601.19375.
- Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. 2024. [Alignment faking in large language models](#). *CoRR*, abs/2412.14093.
- Aman Gupta, Denny O’Shea, and Fazl Barez. 2025. [Val-bench: Measuring value alignment in language models](#). *CoRR*, abs/2510.05465.
- Julia Haas, Sophie Bridgers, Arianna Manzini, Benjamin Henke, Joshua May, Sydney Levine, Laura Weidinger, Murray Shanahan, Kristian Lum, Iason Gabriel, and William Isaac. 2026. [A roadmap for evaluating moral competence in large language models](#). *Nature*, 650(8102):565–573.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. [Aligning AI with shared human values](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China*,

- November 3-7, 2019, pages 2733–2743. Association for Computational Linguistics.
- Jiajian Huang, Dongliang Zhu, Zitong Yu, Hui Ma, Jiayu Zhang, Chunmei Zhu, and Xiaochun Cao. 2026. [Decepgpt: Schema-driven deception detection with multicultural datasets and robust multimodal learning](#). *CoRR*, abs/2603.23916.
- Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2024. [Flames: Benchmarking value alignment of llms in chinese](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 4551–4591. Association for Computational Linguistics.
- Yue Huang, Qihui Zhang, Philip S. Yu, and Lichao Sun. 2023. [Trustgpt: A benchmark for trustworthiness and responsible large language models](#). *CoRR*, abs/2306.11507.
- Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2023. [Improving activation steering in language models with mean-centring](#). *CoRR*, abs/2312.03813.
- Chengyi Ju, Weijie Shi, Chengzhong Liu, Jiaming Ji, Jipeng Zhang, Ruiyuan Zhang, Jiajie Xu, Yaodong Yang, Sirui Han, and Yike Guo. 2025. [Benchmarking multi-national value alignment for large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 20042–20058. Association for Computational Linguistics.
- Jenny Kunz and Marco Kuhlmann. 2020. [Classifier probes may just learn from linear context features](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5136–5146. International Committee on Computational Linguistics.
- Andrew Kyle Lampinen, Yuxuan Li, Eghbal Hosseini, Sangnie Bhardwaj, and Murray Shanahan. 2026. [Linear representations in language models can change dramatically over a conversation](#). *CoRR*, abs/2601.20834.
- Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Ryan Liu, Theodore R. Sumers, Ishita Dasgupta, and Thomas L. Griffiths. 2024. [How do large language models navigate conflicts between honesty and helpfulness?](#) In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 31844–31865. PMLR / OpenReview.net.
- Zhishu Liu, Kaishen Yuan, Bo Zhao, Hui Ma, and Zitong Yu. 2026. [AULLM++: structured-token-conditioned large language models for micro-expression action unit detection](#). *CoRR*, abs/2603.08387.
- Samuel Marks and Max Tegmark. 2023. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). *CoRR*, abs/2310.06824.
- Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. [Emergent linear representations in world models of self-supervised sequence models](#). In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2023, Singapore, December 7, 2023*, pages 16–30. Association for Computational Linguistics.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. [Llms know more than they show: On the intrinsic representation of LLM hallucinations](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. [The linear representation hypothesis and the geometry of large language models](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 39643–39666. PMLR / OpenReview.net.
- Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. [Discovering language model behaviors with model-written evaluations](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pages 13387–13434. Association for Computational Linguistics.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 15504–15522. Association for Computational Linguistics.

Stefan F. Schouten, Peter Bloem, Ilia Markov, and Piek Vossen. 2024. [Truth-value judgment in language models: belief directions are context sensitive](#). *CoRR*, abs/2404.18865.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Aske, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. [Towards understanding sycophancy in language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. [Linear representations of sentiment in large language models](#). *CoRR*, abs/2310.15154.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. [Activation addition: Steering language models without optimization](#). *CoRR*, abs/2308.10248.

Taorui Wang, Wei Xia, Hui Ma, Zijia Song, Jiayu Zhang, Zeheng Wang, Yong Xu, and Zitong Yu. 2026a. [Gmot: Gated motion-aware tokenization for fine-grained micro-gesture video reasoning with multimodal llms](#). *CoRR*, abs/2607.16322.

Zeheng Wang, Zitong Yu, Yijie Zhu, Bo Zhao, Haochen Liang, Taorui Wang, Wei Xia, Jiayu Zhang, Zhishu Liu, Hui Ma, Fei Ma, and Qi Tian. 2026b. [Affectagent: Collaborative multi-agent reasoning for retrieval-augmented multimodal emotion recognition](#). *CoRR*, abs/2604.12735.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haoqi Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. [SOTOPIA: interactive evaluation for social intelligence in language agents](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dobrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. [Representation engineering: A top-down approach to AI transparency](#). *CoRR*, abs/2310.01405.

A Additional Experimental Details

Annotation instructions and QA protocol. Automated behavioral annotation used a closed-set rubric with *truth-first*, *propriety-first*, *compatible*, and an *invalid* fallback. Truth-first responses prioritize disclosure, correction, direct truth-telling, or factual clarity over face, deference, or relationship-management concerns. Propriety-first responses prioritize tact, face-saving, harmony, deference, or social appropriateness over direct truth or immediate correction. Compatible responses resolve or reframe the dilemma so that truth and propriety are both substantially satisfied, or do not cleanly prioritize either endpoint. The invalid option was reserved for responses that could not be meaningfully classified under this rubric; no invalid responses remain in the final 600-row label set.

The annotation pipeline used primary and secondary automated judgments, with human QA applied to a higher-risk subset. Final labels were resolved by precedence: a human-QA label was used when available; otherwise, agreement between the automated judges yielded a *judge-agreement* label; remaining rows retained the primary judgment and were marked *primary-only*. This procedure produced 188 human-QA-confirmed rows, 316 judge-agreement rows, and 96 primary-only rows. Human QA had access to the scenario and model response together with the available judge outputs used for adjudication. A separate human agreement exercise used the same three-way behavioral taxonomy; initial agreement was moderate (Cohen’s $\kappa = 0.617$, raw agreement = 0.82).

Human annotation and QA were conducted internally by members of the research team; no external annotators were recruited or compensated.

Semantic-audit protocol. To examine the boundary between *truth-first* and *compatible*, we conducted a post-hoc semantic audit over 60 responses: 30 *truth_first+compatible* rows, 20 *propriety_first+propriety-first* rows, and 10 *truth-first* controls. The annotator was shown the scenario and model response but not the value trigger or original behavioral label. Each response received three independent binary indicators. *Substantive reframing* denotes a material change to the recommended disclosure or action strategy, such as its timing, recipient, channel, scope, sequence, escalation path, or verification procedure. *Tonal hedging* denotes primarily softening,

Table 6: Source robustness of the aggregate condition-label association across label-source subsets.

Subset	n	Cramér’s V	95% CI
all_rows	600	0.187	[0.155, 0.227]
exclude_primary_only	504	0.269	[0.234, 0.308]
human_qa_plus_judge_agreement	504	0.269	[0.234, 0.308]
human_qa_only	188	0.390	[0.305, 0.471]

politeness, or stylistic accommodation without a substantive change to the recommended action. *Lexical echo* denotes explicit reuse or close paraphrase of trigger-associated value language. The audit was conducted by a single annotator and is therefore used only as a calibration of the compatible-category interpretation, not as independent construct validation.

Label sources. The final label set contains 188 human-QA-confirmed rows, 316 judge-agreement rows, and 96 primary-only rows. We retain the source labels so that downstream analyses can test whether the main construct diagnostic depends on label provenance.

Benchmark metadata. Each rendered row contains a prompt, scenario group identifier, role, stakes, value trigger, and final behavior label. Scenario groups are rendered across six role-by-stakes cells. In TPA, value_trigger is balanced across groups but held constant within each group, so trigger-related analyses are diagnostic rather than causal.

Probe representation. The headline probe uses last-token hidden states averaged over layers 10–19. Hidden-state probes and exact-prompt TF-IDF / BoW baselines are evaluated under 5-fold group-level cross-validation.

B Source Robustness

The condition-label association is not driven only by primary-only labels. Higher-confidence subsets show larger associations, but they are not directly comparable to the full set because the QA workflow enriches them. We therefore retain all_rows as the conservative headline.

C Surface and Geometry Diagnostics

The BoW ablations support the surface-audit conclusion. Removing the explicit value-trigger sentence does not reduce BoW performance, and removing one prompt field is not enough to eliminate surface predictability. The shortcut is there-

Table 7: Group-CV BoW ablations: balanced accuracy when individual prompt-template fields are removed.

Prompt input	Balanced accuracy	Macro-F1
full prompt	0.978	0.961
w/o role sentence	0.923	0.926
w/o stakes	0.963	0.931
w/o topic phrase	0.978	0.961
w/o value-trigger	0.978	0.961

fore distributed across the rendered prompt rather than localized in a single trigger phrase.

We also compute two auxiliary activation-space diagnostics. The per-cell truth–propriety direction is nearly collinear with the pooled truth–propriety direction (cosine = 0.969), suggesting little condition-specific directional structure beyond a global label contrast. The direction contrasting prompt-intent labels is only weakly aligned with the direction contrasting behavioral endpoint labels (cosine = 0.109), consistent with keeping prompt intent and observed behavior as separate evidence layers. These diagnostics are supporting checks, not evidence for internal representation geometry.

Finally, group-level label entropy is low: mean entropy is 0.224 bits, median entropy is 0.000, 66/100 groups have all six condition variants assigned the same label, and 99/100 groups have at least five of six variants in agreement. This supports the paper’s caution that scenario/trigger priors are strong in TPA.

D Llama Annotation-Stability Check

We attempted a cross-family replication on Llama-3.1-8B-Instruct using the same 600 TPA prompts and matched decoding settings. Generation succeeded, but the three-way annotation protocol did not reach sufficient stability for a behavioral comparison.

In a stratified 120-row mini-IAA, rubric v1.0 yielded Cohen’s κ = 0.514 and raw agreement = 0.708. Disagreements concentrated on the truth-first/compatible boundary: 25 of 35 disagreements were between these two labels, while propriety-first was perfectly recoverable in the sample. We

then refined the rubric to require *compatible* responses to substantively constrain the endorsed action—for example by changing timing, recipient, channel, scope, sequence, or escalation path—rather than merely softening tone. A fresh 120-row mini-IAA under the refined rubric did not improve stability: Cohen’s κ decreased to 0.417, raw agreement to 0.633, and compatible positive agreement to 0.560.

We therefore stopped the Llama branch and do not report cross-model behavioral averages. The exercise supports the main paper’s conservative treatment of *compatible*: it is useful for residual decomposition, but its boundary with truth-first remains annotation-sensitive and should not be treated as a fully validated third policy mode.