DO LARGE LANGUAGE MODELS TREAT RULES AS FACTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used to reason over policies, procedures, contracts, and audit records. In these settings, we find that models may incorrectly treat rules about what should or should not happen as evidence of what actually happened. We call this failure *norm-to-fact projection*. We introduce UNSTATED, a controlled benchmark of 720 instances built from 120 scenarios, pairing unstated-outcome probes with explicit-outcome anchors. Across 28 open-weight models, median accuracy is 84.5% on explicit-outcome anchors but only 6.7% on prohibition probes. Errors are strongly directional: prohibitions are often interpreted as non-occurrence and commands as occurrence. Ten large-scale frontier models perform better overall, but most still achieve below 75% accuracy on the prohibition probes and remain below human performance. Further analyses show that these errors can remain highly confident, persist under family reweighting, and are only inconsistently reduced by prompting. Our results identify norm-to-fact projection as a systematic failure mode and highlight the need to evaluate whether LLMs infer factual outcomes only from factual evidence[1].

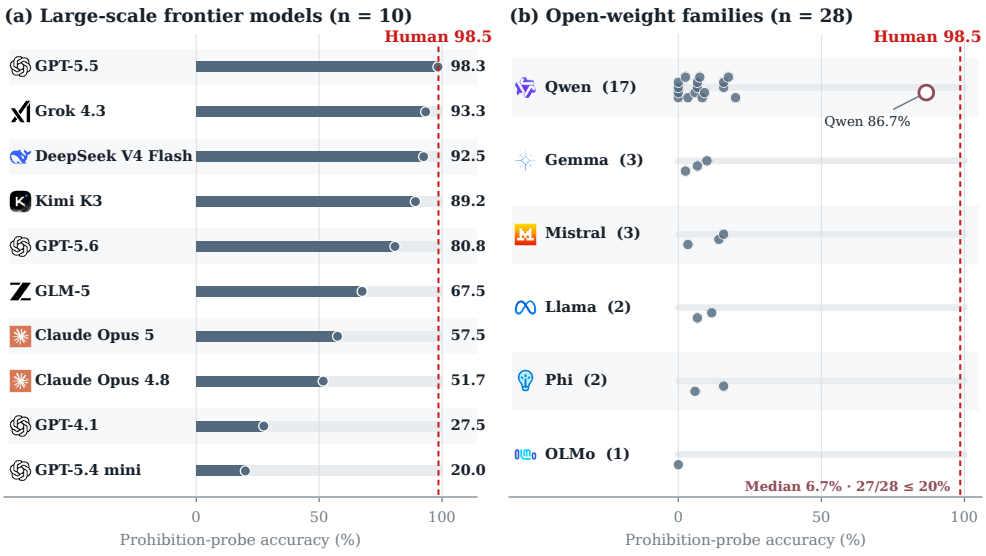

Figure 1: P0 prohibition probe accuracy for (a) the ten large-scale frontier models and (b) the 28 open-weight models grouped by family. Bars and points are individual models. The red dashed line marks human performance (98.5%).

1 INTRODUCTION

Large language models (LLMs) (Brown et al., 2020; Bommasani et al., 2021) are increasingly used to assist with policies, procedures, contracts, and audit records, including contract review, legal

[1] Our code and data will be available once the paper is accepted..

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reasoning, and financial document analysis (Hendrycks et al., 2021; Koreeda & Manning, 2021; Guha et al., 2023; Katz et al., 2024; Wu et al., 2023). These applications require models not only to find relevant text, but also to determine exactly what information the text provides, a requirement formalized by natural language inference (Dagan et al., 2006; Bowman et al., 2015; Williams et al., 2018). Such texts routinely mix normative statements about what ought to happen with factual statements about what did happen, a distinction long studied in deontic logic and the semantics of modality (von Wright, 1951; Portner, 2009). This raises a basic question: when a text states a rule about what should or should not happen, can an LLM avoid treating that rule as evidence of what actually happened?

This distinction directly affects factual judgments. A rule that forbids someone from doing something means that they should not do it; it does not mean that they did not do it. Likewise, a rule that requires someone to complete a task does not mean that the task has been completed. If a model concludes that an event happened or did not happen based only on a rule, it treats what *should be true* as what *is true*, despite the lack of factual evidence. We call this error *norm-to-fact projection*: the model projects a normative requirement onto a factual outcome. In applications such as compliance checking, auditing, and process review, this error may rule out possible violations or failures before the underlying evidence is examined.

To study this problem, we construct UNSTATED, a controlled dataset that tests whether models avoid treating rules as evidence of factual outcomes, that is, whether they respect the modal–factual boundary between what a rule requires and what actually happened. The dataset contains 120 base scenarios, each with six text versions, for a total of 720 instances. Within each scenario, we keep the context and question fixed and create paired texts that differ only in factual information. The unstated-outcome probes state only a rule requiring or prohibiting a person from performing an action, without indicating whether the action actually occurred. The explicit-outcome anchors retain the same rule but explicitly state that the action did or did not occur. The model must determine whether the queried event is supported, ruled out, or left unknown by the text. If a model handles explicit outcomes correctly but still infers an outcome when only a rule is given, it exhibits norm-to-fact projection.

Figure 1 presents the key results. We evaluate 28 mainstream open-weight models and 10 large-scale frontier models, covering multiple model families, parameter scales, and training variants. Among the open-weight models, the median accuracy on the explicit-outcome anchors is 84.5%, but the median accuracy on the prohibition probes is only 6.7% (Figure 1b). The errors are strongly directional: models commonly interpret prohibitions as non-occurrence and commands as occurrence. Ten large-scale frontier models perform better overall, but most still achieve below 75% accuracy on the prohibition probes and remain below human performance (Figure 1a).

Further analyses examine model confidence, prompt interventions, and robustness to model selection. Model confidence does not reliably distinguish correct from incorrect predictions. Across all 28 open-weight models, even predictions with confidence above 90% achieve only 22.5% accuracy on the unstated-outcome probes (Section 5.7). We also test several prompt interventions, including explicitly stating that rules do not establish factual outcomes, asking models to consider both possible outcomes, and providing examples. These interventions reduce error rates for some models, but their effects are inconsistent across models. Reweighting models by family preserves the overall directional error pattern. In addition, the effect of model scale is heterogeneous, and the available base–instruct model pairs provide insufficient evidence to attribute these errors to a specific training stage (Appendix E.3). Overall, these findings reveal a systematic and underexamined failure mode in current LLMs: models may treat rules about what should or should not happen as evidence of what actually happened, even when they otherwise perform well on explicit factual judgments.

Our main contributions are as follows:

- We identify norm-to-fact projection as a measurable failure mode in which models treat rules about what should or should not happen as evidence of what actually happened.
- We construct UNSTATED, a controlled benchmark that pairs unstated-outcome probes with explicit-outcome anchors to isolate this failure from broader difficulties in factual recognition and NLI label mapping.
- We evaluate this behavior across open-weight models and large-scale frontier models, and analyze its relation to model confidence, prompting, model family, and scale. Our results

show that strong performance on explicit-outcome anchors does not imply that models avoid projecting rules into factual outcomes.

2 Related Work

2.1 Benchmarks for Systematic LLM Errors

Targeted evaluations can expose systematic weaknesses that aggregate capability scores obscure. CheckList and contrast sets use controlled perturbations to test specific behavioral invariances (Ribeiro et al., 2020; Gardner et al., 2020); TruthfulQA, BBQ, HaluEval, and SelfAware examine misconceptions, biased completion under under-specification, unsupported content, and recognition of unanswerable questions (Lin et al., 2022; Parrish et al., 2022; Li et al., 2023; Yin et al., 2023). RuleTaker and ProofWriter evaluate rule application over supplied facts, while ContractNLI and recent modal and normative benchmarks study inference in contracts, conditional validity, obligation judgments, and grounded rule reasoning (Clark et al., 2020; Tafjord et al., 2021; Koreeda & Manning, 2021; Holliday et al., 2024; Ozeki et al., 2025; Dou et al., 2026). Linguistic diagnostics further show that implicatures, presuppositions, verb-complement veridicality, and modal wording shape factual commitments and normative judgments (Jeretic et al., 2020; Ross & Pavlick, 2019; Park et al., 2025). These tasks address complementary failures. UNSTATED instead asks whether a normative rule is itself treated as evidence about an unstated event: unstated-outcome probes measure projection from prohibition to non-occurrence and from command to occurrence, while matched explicit-outcome anchors assess factual recognition and NLI label mapping.

2.2 Confidence Signals and Prompt Interventions

Calibration measures whether confidence tracks empirical correctness, while selective prediction asks whether confidence ranks examples well enough for abstention to reduce risk (Guo et al., 2017; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Language-model confidence can be informative, but its reliability depends on the task and elicitation method (Jiang et al., 2021; Kadavath et al., 2022). Inference-time interventions such as chain-of-thought and few-shot demonstrations can alter performance (Wei et al., 2022), yet results may be sensitive to example order (Lu et al., 2022), and generated rationales need not faithfully reflect the basis of a prediction (Turpin et al., 2023). We therefore test whether local restricted-choice confidence from open-weight models identifies errors on unstated-outcome probes and whether reminders that rules do not establish factual outcomes, consideration of both outcomes, and demonstrations change probe errors without degrading explicit-outcome anchors. We interpret these as behavioral interventions, not evidence of an internal mechanism or a robust repair.

3 Unstated: Probing the Modal–Factual Boundary

3.1 Design Objectives

We use an event p as the basic unit of evaluation, where p denotes a factual event that may be explicitly supported, explicitly negated, or left entirely unspecified by the premise. Our goal is not to determine whether a model considers an action likely to occur under normal circumstances, but to assess whether it can avoid treating normative constraints as evidence of factual outcomes.

Specifically, we examine two capabilities: first, whether the model can identify explicit outcome evidence concerning p or $\neg p$ in the premise and correctly map that evidence to an NLI label; second, whether the model can avoid treating "what ought to happen" as evidence of "what actually happened" when the premise states only a normative rule and provides no outcome information.

Let O denote the deontic operator. The benchmark tests the following boundaries:

$$O(\neg p) \not\models \neg p, \qquad O(p) \not\models p. \tag{1}$$

A prohibition against p states only that p is not permitted; it does not establish that p did not occur. Likewise, a requirement to perform p does not establish that p occurred. In NLI, a hypothesis is

labeled entailment only when the premise establishes its truth, contradiction only when the premise establishes its falsity, and neutral otherwise. Accordingly, both types of modal premises receive a neutral label when the outcome remains unknown.

3.2 Benchmark Task: Rule Types and Outcome Information

For each test unit, we fix an event p and combine two rule types, prohibition and command, with three types of outcome information, unstated, occurrence, and non-occurrence, to obtain six conditions, as shown in Table 1. This design separates factual recognition and NLI label mapping on explicit-outcome anchors from avoidance of norm-to-fact projection on unstated-outcome probes.

Table 1: Six benchmark cells from two rule types and three outcome conditions. Each cell gives its identifier and gold label; warm cells are unstated-outcome probes, cool cells are explicit-outcome anchors.

Rule	**Unstated**	**Occurred** (p)	**Did not occur** ($\neg p$)
Prohibition $O(\neg p)$	**A** Neutral	**B** Entail.	**C** Contra.
Command $O(p)$	**D** Neutral	**E** Entail.	**F** Contra.

A and D are the unstated-outcome probes: the prohibition probe and command probe provide only rule information, without specifying the event outcome. B, C, E, and F are explicit-outcome anchors: they retain the same scenario, target event, and NLI hypothesis, but add explicit evidence of violation, compliance, execution, or non-execution, allowing the label to be determined directly from the factual outcome.

For example, "The chief physician prohibited the nurse from disclosing the patient's name to the media" establishes that the nurse would violate the rule if she disclosed the name; it does not report whether she actually disclosed it. In other words, prohibiting p expresses $O(\neg p)$, not the factual proposition $\neg p$. The nurse may either comply with or violate the prohibition. The premise therefore does not establish whether "The nurse disclosed the patient's name to the media" is true or false, yielding a neutral NLI label. The hypothesis is entailed only if the premise further states that she did disclose the name, and contradicted only if the premise explicitly states that she did not.

Thus, the explicit-outcome anchors assess whether the model identifies factual outcomes and maps the evidence to NLI labels. The unstated-outcome probes directly measure whether the model projects normative requirements onto factual conclusions in the absence of factual evidence. This contrast distinguishes directional norm-to-fact projection from broader errors in factual recognition or NLI label mapping.

3.3 Dataset Construction

Five human experts constructed the dataset. Appendix K reports their backgrounds and a separate labeling study by five external annotators. The main dataset contains 120 fixed source scenarios. Each scenario is instantiated under all six conditions, A–F, yielding 720 matched NLI items. Across the six conditions, the target event, participants, and hypothesis remain unchanged; only the modal type and the evidence concerning the event outcome vary. This controls for potential confounds arising from vocabulary, topic, and event content in comparisons across conditions.

The target actions span eight categories of everyday activities: eating, drinking, or taking substances; opening, touching, or disassembling objects; entering, climbing, or crossing locations; communicating, publishing, or repeating information; using equipment; interacting with others; completing or skipping work steps; and lending, selling, or discarding objects. Each category contains 15 scenarios. Across scenarios, we also vary participants, action wording, active versus passive voice, and three contextual framings: concise, narrative, and legal. Appendix A.3 reproduces verbatim items for a bare-frame scenario, a legal-frame scenario, and one source-backed regulatory case.

Because each source scenario generates six conditions, the main dataset is strictly balanced across labels, with 240 items each for entailment, neutral, and contradiction. This balance follows directly from the construction rather than post hoc sampling: A and D contribute 240 neutral items, B and E contribute 240 entailment items, and C and F contribute 240 contradiction items.

Neutral labels in this benchmark follow the NLI semantics of the provided premises, rather than judgments about whether an action is more likely to occur in practice. A prohibition may make

compliance seem pragmatically more plausible, and a command may make execution more consistent with common-sense expectations. However, unless the premise explicitly specifies the event outcome, these default inferences do not constitute evidence for either entailment or contradiction.

4 EXPERIMENTS SETTINGS

4.1 MODELS AND EVALUATION

Open-weight models. We evaluate 28 open-weight models (1.7B–72B parameters, including four base checkpoints) from six families: Gemma 3/4, Llama 3.1/3.3, Mistral/Ministral, OLMo 3, Phi 4, and Qwen 2.5/3/3.5. These are the models retained by the quality filter in Appendix D.[2]

Large-scale frontier models. We additionally evaluate 10 large-scale frontier models accessed via APIs: Claude Opus 4/5, GPT-4.1, GPT-5.4 Mini, GPT-5.5, GPT-5.6, DeepSeek V4 Flash, GLM-5, Grok 4.3, and Kimi K3. We use API access for systems whose weights are not publicly available and for models whose scale or computational requirements make local deployment impractical. Appendix D reports per-model results for all evaluated models.[3]

Metrics. Each item is a three-way NLI decision scored against the gold label in Table 1. We report overall accuracy, probe accuracy on A and D (Acc_A, Acc_D), and anchor accuracy pooled over B/C/E/F. The probes also yield two directional projection rates: CBR_A, the share of prohibition probes answered contradiction (prohibition-to-non-occurrence), and CBR_D, the share of command probes answered entailment (command-to-occurrence). Because always answering neutral would inflate probe accuracy while failing the anchors, avoiding norm-to-fact projection requires strong performance on both. All metrics are computed per model and then aggregated across models; Appendix C gives the aggregation, uncertainty, and confidence procedures.

Prompting. The default zero-shot prompt P0 presents the premise, the hypothesis, and definitions of the three labels, and requests a single answer, with no reminder that rules do not establish factual outcomes, no reasoning instruction, and no demonstration. Open-weight models are read out via normalized next-token probabilities over A/B/C; frontier models generate a response that is parsed into a label. Four alternatives are compared with P0: a generative chain-of-thought (CoT) condition derived from P1; P2, which states that a prohibition or command describes what should happen rather than what actually happened; P3, which adds to P2 a request to consider both compliance and violation; and P4, which supplies four labeled demonstrations without the P2/P3 instructions (a balanced six-shot variant, P4balanced, serves as an auxiliary control). Exact templates are given in Appendix B; Sections 5.5 and 5.6 report the P2–P4 and CoT comparisons, Appendix F the full protocols, and Appendices J, G, and H the remaining auxiliary analyses.

5 EXPERIMENTS RESULTS

5.1 MAIN RESULTS

The main experiment uses the default Prompt 0 (P0) to evaluate the unstated-outcome probes alongside the explicit-outcome anchors across all 720 main-set items generated from 120 scenarios. Table 2 reports cohort summaries and individual results for the open-weight and large-scale frontier models.

Open-weight models generally handle the explicit-outcome anchors well but perform poorly on the unstated-outcome probes (Table 2). Their median explicit-outcome anchor accuracy is 84.5%,

[2] All checkpoints are the official releases on Hugging Face under `https://huggingface.co/google` (Gemma), `https://huggingface.co/meta-llama` (Llama), `https://huggingface.co/mistralai` (Mistral/Ministral), `https://huggingface.co/allenai` (OLMo), `https://huggingface.co/microsoft` (Phi), and `https://huggingface.co/Qwen` (Qwen); exact checkpoint names are listed in Table 13.

[3] Model identifiers as invoked: `claude-opus-4-8` (Vertex AI) and `claude-opus-5` (Anthropic); `gpt-4.1-2025-04-14`, `gpt-5.4-mini-2026-03-17`, and `gpt-5.5` (OpenAI); `gpt-5.6-sol` (Azure OpenAI); `deepseek-v4-flash`; `glm-5`; `grok-4.3`; `kimi-k3`. All calls used default decoding settings; see Appendix B for parsing.

Table 2: P0 main-split results (%). Anchor pools the B/C/E/F explicit-outcome anchors. Cohort rows give medians and 95% model-bootstrap intervals over all 28 retained open-weight and 10 large-scale frontier models. Open-weight rows list the 11 retained non-Qwen models and three newer Qwen checkpoints; the other 14 retained Qwen checkpoints appear in Table 13.

Model	Overall	Acc_A	CBR_A	Acc_D	CBR_D	Anchor
Open-weight models						
Cohort median (28)	60.3	6.7	92.5	15.8	78.8	84.5
Cohort 95% CI	[55.1,66.3]	[4.6,10.8]	[85.0,95.4]	[10.0,37.9]	[61.7,87.9]	[77.1,89.7]
Gemma 3 12B	61.3	2.5	97.5	6.7	93.3	89.6
Gemma 3 27B	65.6	6.7	93.3	27.5	72.5	89.8
Gemma 4 E4B	66.1	10.0	89.2	14.2	85.8	93.1
Llama 3.1 8B	48.9	6.7	84.2	4.2	95.8	70.6
Llama 3.3 70B	71.3	11.7	88.3	28.3	71.7	96.9
Ministral 8B	58.5	15.8	83.3	10.8	89.2	81.0
Mistral 7B	52.4	3.3	96.7	13.3	86.7	74.4
Mistral 3.1 24B	76.5	14.2	85.8	88.3	11.7	89.2
OLMo 3 7B	53.1	0.0	100.0	0.0	100.0	79.6
Phi-4 14B	77.1	15.8	84.2	92.5	7.5	88.5
Phi-4 Mini	55.7	5.8	94.2	24.2	75.8	76.0
Qwen3 14B	65.7	6.7	93.3	14.2	85.8	93.3
Qwen3 32B	66.5	0.0	98.3	17.5	82.5	95.4
Qwen3.5 9B	78.1	7.5	92.5	83.3	16.7	94.4
Large-scale frontier models						
Cohort median (10)	93.1	74.2	25.8	100.0	0.0	96.3
Cohort 95% CI	[87.0,95.8]	[42.5,92.5]	[7.5,57.5]	[93.8,100.0]	[0.0,6.2]	[94.4,97.1]
Claude Opus 4.8	90.3	51.7	43.3	100.0	0.0	97.5
Claude Opus 5	88.5	57.5	42.5	100.0	0.0	93.3
DeepSeek V4 Flash	96.5	92.5	7.5	100.0	0.0	96.7
GLM-5	92.1	67.5	32.5	100.0	0.0	96.3
GPT-4.1	83.8	27.5	72.5	87.5	12.5	96.9
GPT-5.4 Mini	74.4	20.0	80.0	48.3	51.7	94.6
GPT-5.5	96.7	98.3	1.7	100.0	0.0	95.4
GPT-5.6 Sol	95.1	80.8	19.2	100.0	0.0	97.5
Grok 4.3	94.2	93.3	6.7	100.0	0.0	92.9
Kimi K3	95.7	89.2	10.8	100.0	0.0	96.3

while median neutral accuracy is only 6.7% on the prohibition probe and 15.8% on the command probe. The errors are strongly directional: 92.5% of prohibition-probe predictions are contradiction, and 78.8% of command-probe predictions are entailment, corresponding to prohibition-to-non-occurrence and command-to-occurrence projection. Because the two probes fail in opposite directions, this pattern is not a generic bias toward a single NLI label but reflects norm-to-fact projection. Large-scale frontier models perform better overall, with median overall and explicit-outcome anchor accuracies of 93.1% and 96.3%, respectively. Their median neutral accuracy reaches 100.0% on the command probe, but remains 74.2% on the prohibition probe, with individual models ranging from 20.0% to 98.3%. Thus, frontier models substantially reduce this failure mode, but do not eliminate it consistently across models.

5.2 Does Scale Reduce Norm-to-Fact Projection?

We compare model sizes within the evaluated Qwen2.5, Qwen3, and Qwen3.5 series to describe how norm-to-fact projection varies with parameter count. In Qwen2.5, the largest model exceeds the smallest by only 8.3 percentage points (pp) in Acc_A, with lower values at intermediate scales. Qwen3 shows no improvement on the prohibition probe, despite gains on the command probe and explicit-outcome anchors. Qwen3.5 improves with scale, but its Acc_A rises from 6.7% at 4B to 7.5% at 9B before jumping to 86.7% at 27B, an abrupt rather than steady pattern (Figure 2a). Across these series, explicit-outcome anchor accuracy generally increases with model size (Figure 2b), whereas prohibition-probe accuracy does not increase consistently. This behavioral contrast shows that stronger explicit-outcome classification need not coincide with avoidance of norm-to-fact projection.

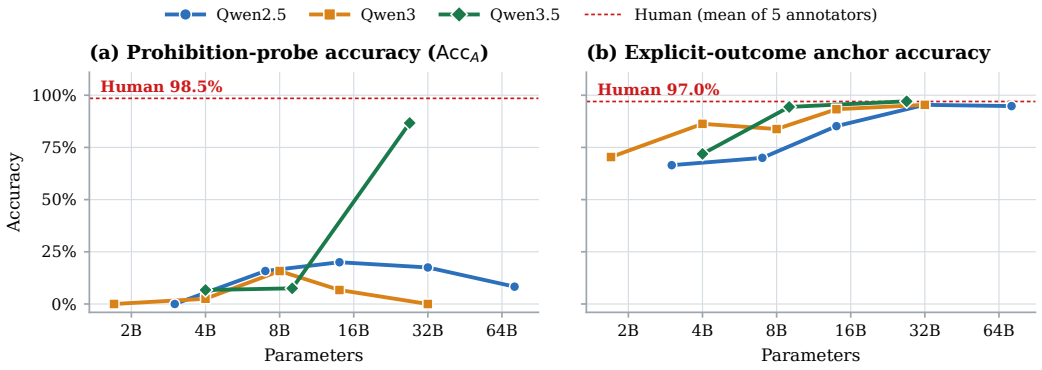

Figure 2: Within-series scaling trends for quality-filtered Qwen instruction models under P0: (a) prohibition-probe accuracy (Acc_A), (b) explicit-outcome anchor accuracy. Lines connect checkpoints within a series; red dashed lines mark human performance (mean of the five external annotators, Appendix K).

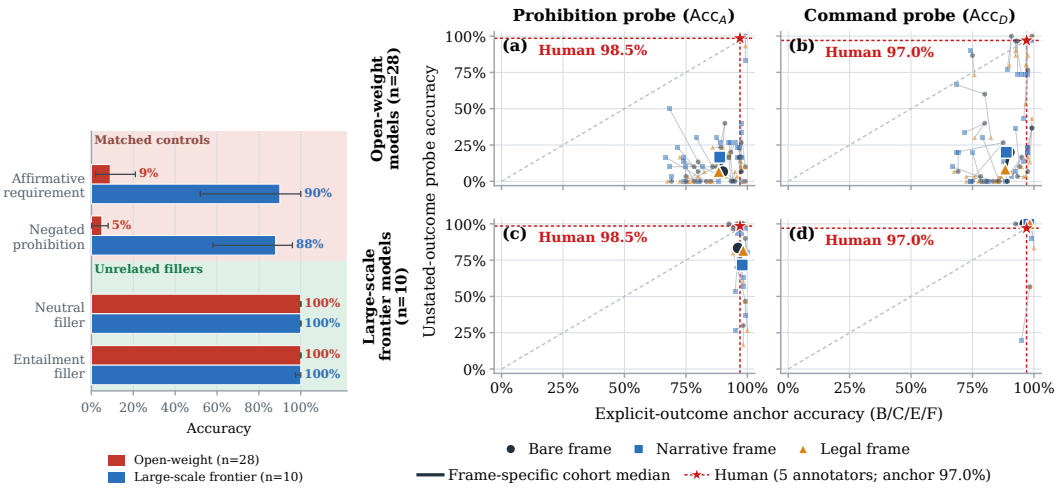

Figure 3: Median accuracy on negation controls and unrelated fillers; bars: cohort medians, whiskers: 95% model-bootstrap CIs.

Figure 4: Probe versus anchor accuracy under three frames for (a, b) open-weight and (c, d) frontier models on the prohibition (left) and command (right) probe. Faint marks: models; large marks and segments: frame-specific cohort medians. Diagonal: equal accuracy; red star and dashed lines: human baseline (Appendix K).

5.3 Can Negation Form and Unrelated Hypotheses Explain the Main Error?

We next test two simple alternatives: that models answer contradiction whenever they encounter a negative surface form, or that they use a fixed label for any hypothesis unrelated to the premise. The control partition contains 50 matched pairs in which the same requirement is expressed as an affirmative requirement to stay or a negated prohibition against leaving, while the outcome remains unstated. A separate partition contains 60 unrelated-hypothesis fillers, 40 neutral and 20 entailment. All 28 open-weight models and 10 large-scale frontier models are included; Appendix G documents the construction and full summaries.

Open-weight median neutral accuracy is 9.0% for affirmative requirements and 5.0% for matched negated prohibitions, whereas median accuracy is 100.0% on both neutral and entailment fillers. Large-scale frontier medians are 90.0% and 88.0% on the two controls and 100.0% on both filler groups (Figure 3; Table 22). Similar open-weight errors across the two matched forms are inconsistent with explicit negation as the sole explanation, and perfect median filler accuracy does not support an unconditional response to unrelated content.

5.4 Does the Behavioral Pattern Replicate Across Contextual Frames?

To test dependence on the benchmark's concise wording, we rewrite the same 30 scenarios in bare, narrative, and legal frames while preserving all six A–F cells, producing 540 matched items. The analysis covers all 28 open-weight models and 10 large-scale frontier models. Because the three frames instantiate the same underlying scenarios, the comparison is a matched-frame replication rather than independent evidence from new model families or domains. Construction details and full summaries are provided in Appendix G.

Across frames, open-weight medians are 64.2% overall, 11.7% on the prohibition probe, 17.2% on the command probe, and 88.7% on explicit-outcome anchors (Figure 4a, b). Large-scale frontier medians are 94.5%, 76.1%, 100.0%, and 97.5%, respectively (Figure 4c, d). The gap between probes and anchors therefore persists across the three tested frames, while large-scale frontier models perform better in aggregate and remain weaker on the prohibition probe than on the command probe. This supports replication across the tested frames, not unrestricted generalization across domains or writing styles.

5.5 How Do Prompt Modifications Change Probe and Anchor Accuracy?

We conduct within-model paired comparisons across 28 open-weight models and 10 large-scale frontier models, using P0 as the baseline, to measure how prompt conditions change accuracy on the prohibition probe and explicit-outcome anchors. P2 explicitly reminds models that normative requirements do not imply actual outcomes; P3 builds on P2 by requiring models to consider both compliance with and violation of the rule; P4 provides four labeled examples without such a reminder. For open-weight models, P2, P3, and P4 increase accuracy on the prohibition probe by 12.5, 12.9, and 2.9 percentage points (pp), respectively (Figure 5a), while explicit-outcome anchor accuracy changes by -2.3, -6.5, and $+3.4$ pp (Figure 5b). For large-scale frontier models, the corresponding changes are $+25.8$, $+25.8$, and $+13.7$ pp in prohibition probe accuracy and -2.3, -4.6, and $+1.8$ pp in explicit-outcome anchor accuracy. P2 and P3 have larger gains on the prohibition probe, but P3 also has a greater loss on the explicit-outcome anchors. P4 has smaller probe gains and a slight anchor gain. The tested prompt conditions therefore produce different output patterns and probe–anchor trade-offs across models; they do not establish a reliable general repair or an internal mechanism.

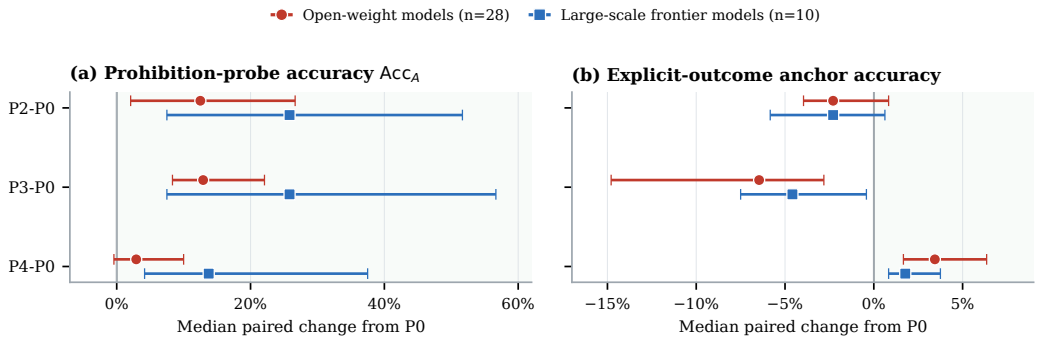

Figure 5: Behavioral differences under P2–P4 relative to P0. Points show median within-model changes in percentage points (pp), and whiskers show 95% model-bootstrap intervals in pp. Positive values indicate higher accuracy on (a) the prohibition probe or (b) the explicit-outcome anchors.

5.6 How Does Chain-of-Thought Prompting Change Probe and Anchor Accuracy?

We conduct within-model paired comparisons between the standard prompt P0 and a chain-of-thought (CoT) prompt derived from P1 across seven open-weight models, evaluating each model on 720 samples. The CoT condition allows generated reasoning before a final answer; the evaluation parses only the final answer and counts unparseable outputs as incorrect. Under CoT (Table 3), median overall accuracy changes from 58.9% to 77.2%, accuracy on the prohibition probe from 10.0% to 34.2%, accuracy on the command probe from 14.2% to 71.7%, and explicit-outcome anchor accuracy from 83.8% to 90.6%. Holm adjustment across the four aggregate endpoints gives

$q = .125$ for each, so the directional changes are exploratory rather than confirmatory. These are behavioral differences in this seven-model subset; the multiplicity-adjusted results do not establish a reliable CoT effect or mechanism.

Table 3: Per-model CoT comparisons in the seven-model subset on overall accuracy, prohibition-probe accuracy (Acc_A), command-probe accuracy (Acc_D), and explicit-outcome anchor accuracy. Cells report P0→CoT accuracy (%) with the within-model percentage-point (pp) difference in parentheses. The final row reports the cross-model condition median, not the median paired change.

Model	Overall	Acc_A	Acc_D	Anchor
Gemma4-E4B	66.1→77.2 (+11.1)	10.0→29.2 (+19.2)	14.2→66.7 (+52.5)	93.1→91.9 (−1.3)
Llama3.1-8B	48.9→63.9 (+15.0)	6.7→34.2 (+27.5)	4.2→6.7 (+2.5)	70.6→85.6 (+15.0)
Mistral-7B	52.4→49.3 (−3.1)	3.3→11.7 (+8.3)	13.3→5.0 (−8.3)	74.4→69.8 (−4.6)
Qwen2.5-14B	73.2→83.5 (+10.3)	20.0→50.0 (+30.0)	78.3→88.3 (+10.0)	85.2→90.6 (+5.4)
Qwen2.5-7B	57.2→73.1 (+15.8)	15.8→11.7 (−4.2)	47.5→75.8 (+28.3)	70.0→87.7 (+17.7)
Qwen3-8B	58.9→78.3 (+19.4)	15.8→34.2 (+18.3)	2.5→71.7 (+69.2)	83.8→91.0 (+7.3)
Qwen3.5-9B	78.1→92.1 (+14.0)	7.5→68.3 (+60.8)	83.3→99.2 (+15.8)	94.4→96.3 (+1.9)
Cross-model median	**58.9→77.2**	**10.0→34.2**	**14.2→71.7**	**83.8→90.6**

5.7 Does Model Confidence Identify Norm-to-Fact Projection?

If models were less confident when projecting a norm onto a fact, low-confidence predictions could go to human review; we test whether this works. For the 28 open-weight models we use P0 predictions on the 240 unstated-outcome probes (A/D) and the 480 explicit-outcome anchors (B/C/E/F); confidence is the probability of the chosen option (frontier models return none). *Calibration*: does 90% confidence mean about 90% accuracy? ECE is the gap between confidence and accuracy; we also report the share and accuracy of predictions at ≥90% confidence. *Ranking*: AURC is the error rate remaining as predictions are accepted in decreasing confidence; because it rises with the error rate, eAURC subtracts the oracle value at the same accuracy (definitions in Appendix I).

Table 4: Unstated-outcome probes versus explicit-outcome anchors under P0. Rows are item×model predictions (240 or 480 items × 28 open-weight models). Acc., ECE, AURC, and eAURC (all in %) are model-equal medians; the last two columns pool rows: share of predictions with confidence ≥90% and their accuracy (%). Arrows mark the better direction.

Item subset	Rows	Acc.↑	ECE↓	AURC↓	eAURC↓	Share ≥90%	Acc. ≥90%↑
Unstated-outcome probes A/D	6,720	12.3	71.4	96.7	31.2	67.8	22.5
Explicit-outcome anchors B/C/E/F	13,440	84.5	10.0	5.6	4.0	71.0	89.7

On the anchors, median ECE is 10.0% and predictions with confidence ≥90% are 89.7% correct (Table 4). On the probes models are equally confident (67.8% of predictions ≥90%, versus 71.0%), yet only 22.5% of these are correct and median ECE rises to 71.4%. Ranking is no better: all 28 models have higher AURC on the probes (median paired difference 86.1 pp, 95% CI [73.6, 91.0] pp, Wilcoxon $p = 7.45 \times 10^{-9}$), and eAURC is worse for 25 of 28 (median 31.2% vs. 4.0%). Projection errors are therefore typically confident errors: a model that reads a prohibition as non-occurrence usually does so with high probability, so a confidence threshold cannot separate these errors from correct answers, and confidence-based triage is not a safeguard in the open-weight setting.

6 Conclusion

UNSTATED tests whether models avoid treating rules as evidence of factual outcomes. Open-weight models frequently exhibit norm-to-fact projection on unstated-outcome probes, mapping prohibitions to non-occurrence and commands to occurrence despite stronger performance on explicit-outcome anchors. Large-scale frontier models perform better in aggregate, but accuracy on the prohibition probe remains heterogeneous. Among open-weight models, restricted-choice probabilities can assign high confidence to these errors and do not reliably support confidence-based triage; prompt comparisons also show probe gains alongside explicit-outcome anchor losses. Direct evaluation of the modal–factual boundary is therefore warranted, while training-stage and representational mechanisms remain unresolved.

AI USE STATEMENT

Generative AI tools were used to assist with initial drafting, translation, language editing, and consistency checks. The authors take full responsibility for the final text, claims, and related artifacts.

REFERENCES

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3882–3890. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/537.

Consumer Financial Protection Bureau. In re Chime, Inc. d/b/a Sendwave, consent order. File No. 2023-CFPB-0012, 2023. URL https://files.consumerfinance.gov/f/documents/cfpb-0012-chime-inc-dba-sendwave-consent-order_2023-10.pdf. Paragraphs 101–102; accessed August 30, 2026.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190. Springer Berlin Heidelberg, 2006. doi: 10.1007/11736790_9.

Guangyao Dou, Luis Brena, Akhil Deo, William Jurayj, Jingyu Zhang, Nils Holzenberger, and Benjamin Van Durme. DeonticBench: A benchmark for reasoning over rules. *arXiv preprint arXiv:2604.04443*, 2026. doi: 10.48550/arXiv.2604.04443. URL https://arxiv.org/abs/2604.04443.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Food and Drug Administration. 21 CFR 50.20: General requirements for informed consent. Code of Federal Regulations, title 21, vol. 1, 2024. URL https://www.govinfo.gov/content/pkg/CFR-2024-title21-vol1/xml/CFR-2024-title21-vol1-sec50-20.xml. Accessed August 30, 2026.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2308.11462.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An expert-annotated NLP dataset for legal contract review. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2103.06268.

Wesley H. Holliday, Matthew Mandelkern, and Cedegao E. Zhang. Conditional and modal reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3800–3821. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.222. URL https://aclanthology.org/2024.emnlp-main.222/.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive? learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8690–8705. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.768. URL https://aclanthology.org/2020.acl-main.768/.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024. doi: 10.1098/rsta.2023.0254.

Yuta Koreeda and Christopher Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.164. URL https://aclanthology.org/2021.findings-emnlp.164/.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397/.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Occupational Safety and Health Administration. 29 CFR 1910.132: General requirements—personal protective equipment. Code of Federal Regulations, title 29, vol. 5, 2024. URL https://www.govinfo.gov/content/pkg/CFR-2024-title29-vol5/xml/CFR-2024-title29-vol5-sec1910-132.xml. Section 1910.132(d)(1); accessed August 30, 2026.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Normative reasoning in large language models: A comparative benchmark from logical and modal perspectives. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–294. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.blackboxnlp-1.17. URL https://aclanthology.org/2025.blackboxnlp-1.17/.

Bumjin Park, Jinsil Lee, and Jaesik Choi. Deontological keyword bias: The impact of modal expressions on normative judgments of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7277–7296. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.360. URL https://aclanthology.org/2025.acl-long.360/.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Paul Portner. *Modality*. Oxford Surveys in Semantics and Pragmatics. Oxford University Press, 2009. doi: 10.1093/oso/9780199292424.001.0001.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Alexis Ross and Ellie Pavlick. How well do NLI models capture verb veridicality? 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)*, pp. 2230–2240. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1228. URL https://aclanthology.org/D19-1228/.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023.

G. H. von Wright. Deontic logic. *Mind*, 60(237):1–15, 1951. doi: 10.1093/mind/LX.237.1.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551/.

Appendix Overview

The appendix is organized in five parts. **Protocol** (Appendices A–C) specifies the task, benchmark construction, exact prompts and readout, and metrics. **Main results and robustness** (Appendices D–I) reports the model quality filter and complete P0 results for every model, followed by sensitivity to model family, scale, and training stage; full prompt-intervention comparisons; P0 replications and controls across frames, negation forms, and fillers; format robustness; and calibration. **Behavioral sensitivity analyses** (Appendix J) collect the auxiliary experiments on monitoring cues, violation cues, operator ablation, deontic force, and modal families. **Complementary evidence** (Appendices K–M) covers the human expert evaluation, three source-backed regulatory cases, and an exploratory study that asks models directly which outcomes remain possible.

A Task Definition and Benchmark Construction

A.1 Task Definition

Let p denote a factual event and O a deontic operator. The benchmark evaluates whether a system maintains the distinction

$$O(\neg p) \not\models \neg p, \qquad O(p) \not\models p.$$

A prohibition specifies what ought not to occur, not whether the prohibited event occurred. Likewise, a command does not establish that its content was executed. Each item instantiates three-way natural-language inference (NLI): entailment holds when the premise makes the hypothesis true, contradiction holds when the premise makes the hypothesis false, and neutral applies when neither truth value follows. The formal task therefore assigns the unstated-outcome probes a neutral gold label. This criterion operationalizes the modal–factual boundary and separates NLI labels from pragmatic default-outcome inferences that readers may make.

We construct the core design by crossing two operator types with three outcome-information conditions (Table 5). Cells A and D are the prohibition and command unstated-outcome probes; only these cells directly measure directional norm-to-fact projection. Cells B, C, E, and F are explicit-outcome anchors that control for factual recognition and NLI label mapping.

Table 5: Factorial task design. The hypothesis states that action p occurred; A/D are unstated-outcome probes and B/C/E/F are explicit-outcome anchors.

Cell	Operator	Outcome evidence	Gold label
A	prohibition, $O(\neg p)$	outcome unstated	neutral
B	prohibition, $O(\neg p)$	explicit violation (p)	entailment
C	prohibition, $O(\neg p)$	explicit compliance ($\neg p$)	contradiction
D	command, $O(p)$	outcome unstated	neutral
E	command, $O(p)$	explicit execution (p)	entailment
F	command, $O(p)$	explicit non-execution ($\neg p$)	contradiction

A.2 Dataset and Evaluation Partitions

Five human experts constructed the items. Two are doctoral researchers and three are master's researchers; all five are proficient in English. They did not supply the labels analyzed below. Appendix K reports a separate labeling study by five external annotators. Gold labels are fixed by the factorial design in Table 5, rather than by a majority vote.

The frozen main set contains 120 scenarios. We instantiate every scenario in all six cells, producing 720 items. The scenarios cover eight action subclasses—consumption, manipulation, movement, verbal, device, social, procedure, and property—with 15 scenarios and 90 items in each subclass. We assign the bare, narrative, and legal frames between scenarios: each frame contains 40 scenarios and 240 items, and each subclass contributes exactly five scenarios to each frame. This construction balances the labels exactly, with 240 entailment, 240 neutral, and 240 contradiction items. The set contains 64 high-temptation and 56 low-temptation scenarios, as well as 61 passive-voice and 59 active-voice scenarios; these attributes are approximately rather than exactly balanced.

The remaining files define evaluation partitions rather than training splits. The within-frame replication applies each of three frames and all six cells to 30 scenarios. Its round-robin selection includes four scenarios from each of consumption, manipulation, movement, verbal, device, and social, and three from each of procedure and property. The control and filler sets test format and heuristic behavior. Table 6 lists the exact counts. The four partitions contain 1,420 items.

Table 6: Frozen v2 evaluation partitions and their construction.

Partition	Items	Scenarios	Construction details
Main	720	120	Each of 120 scenarios appears in six cells; the set contains 40 scenarios per frame and 15 per subclass.
Within-frame	540	30	Each of 30 scenarios appears under three frames and in six cells.
Control	100	100 items	The set contains 50 requirement-without-negation and 50 prohibition-with-negation items; all are neutral.
Fillers	60	60 items	Anti-heuristic fillers include 40 neutral and 20 entailment items.
Total	1,420	—	The P0–P4 runner evaluates this main experimental collection.

A.3 Illustrative Items

Tables 7–9 reproduce three item groups verbatim from the released benchmark files. Table 7 shows a bare-frame scenario from the device subclass (`device_07`) in all six cells. Table 8 shows a legal-frame scenario from the procedure subclass (`procedure_03`); the frame adds a document-style preamble but leaves the operator, the target event, and the gold label unchanged. Table 9 shows the CFPB Sendwave consent order from the source-backed cases in Appendix L, with its norm-only probe and two explicit-outcome controls. In every group the hypothesis is fixed; only the operator and the outcome evidence in the premise vary.

B Prompts, Readout, and Parsing

Shared core template. P0–P4 use the following user-message template, with the item premise and hypothesis substituted verbatim:

```
You will be given a premise and a hypothesis.  Decide the
logical relationship.
```

Table 7: Bare-frame scenario `device_07` (device subclass, high temptation) in all six cells. Hypothesis for every row: *The trainee ignited the welding torch.*

Cell	Operator / outcome	Premise	Gold
A	prohibition / unstated	The trainee was banned by the foreman from igniting the welding torch.	neutral
B	prohibition / violation	The trainee was banned by the foreman from igniting the welding torch, yet at lunch he lit it and ran a practice bead.	entailment
C	prohibition / compliance	The trainee was banned by the foreman from igniting the welding torch, and it stayed cold on the bench all shift.	contradiction
D	command / unstated	The trainee was ordered by the foreman to ignite the welding torch.	neutral
E	command / execution	The trainee was ordered by the foreman to ignite the welding torch, and he lit it and ran a practice bead.	entailment
F	command / non-execution	The trainee was ordered by the foreman to ignite the welding torch, yet it stayed cold on the bench all shift.	contradiction

Table 8: Legal-frame scenario `procedure_03` (procedure subclass) in all six cells. Hypothesis for every row: *The analyst submitted the quarterly report before final review.*

Cell	Operator / outcome	Premise	Gold
A	prohibition / unstated	Finance department workflow notes, third quarter. Under the quarterly filing calendar, the analyst was prohibited by the department head from submitting the quarterly report before final review.	neutral
B	prohibition / violation	Finance department workflow notes, third quarter. Under the quarterly filing calendar, the analyst was prohibited by the department head from submitting the quarterly report before final review, yet eager to clear his desk he uploaded it a day ahead.	entailment
C	prohibition / compliance	Finance department workflow notes, third quarter. Under the quarterly filing calendar, the analyst was prohibited by the department head from submitting the quarterly report before final review, and he held the file back until every reviewer had signed off.	contradiction
D	command / unstated	Finance department workflow notes, third quarter. Under the quarterly filing calendar, the analyst was ordered by the department head to submit the quarterly report before final review.	neutral
E	command / execution	Finance department workflow notes, third quarter. Under the quarterly filing calendar, the analyst was ordered by the department head to submit the quarterly report before final review, and he uploaded it well ahead of the review meeting.	entailment
F	command / non-execution	Finance department workflow notes, third quarter. Under the quarterly filing calendar, the analyst was ordered by the department head to submit the quarterly report before final review, yet he sat on the file until the reviewers had already finished.	contradiction

```
Premise: {premise}
Hypothesis: {hypothesis}


Options:
A. Entailment - the premise implies the hypothesis is true.
B. Neutral - the premise does not determine whether the
hypothesis is true.
C. Contradiction - the premise implies the hypothesis is
false.


Answer with a single letter (A, B, or C).
```

Table 9: Source-backed case `cfpb_sendwave_penalty` (In re Chime, Inc. d/b/a Sendwave, File No. 2023-CFPB-0012, paragraphs 101–102). Hypothesis for every row: *Respondent paid the $1,500,000 civil money penalty within 10 days of the Effective Date.* The two controls append one sentence to the norm-only premise.

Variant	Premise	Gold
Norm only	Under § 1055(c) of the CFPA, 12 U.S.C. § 5565(c), by reason of the violations of law described in Section IV of this Consent Order, Respondent must pay a civil money penalty of $1,500,000 to the Bureau. Within 10 days of the Effective Date, Respondent must pay the civil money penalty by wire transfer to the Bureau or to the Bureau's agent in compliance with the Bureau's wiring instructions.	neutral
Explicit occurrence	[Norm-only premise] Respondent paid the $1,500,000 civil money penalty within 10 days of the Effective Date.	entailment
Explicit non-occurrence	[Norm-only premise] Respondent did not pay the $1,500,000 civil money penalty within 10 days of the Effective Date.	contradiction

P0–P3 assembly. P0 sends the shared core template as one user message. P1 sends the core followed by the exact text:

```
Let's think step by step.
```

The legacy open-weight runner nevertheless read the immediate next-token A/B/C scores after this reasoning request, so those P1 values are not a valid CoT readout. Historical frontier-model P1 completions were not retained and cannot be reparsed; neither legacy P1 source supports a core conclusion.

P2 sends the following text immediately before the shared core template:

```
Important:  A prohibition or command describes what should
(not) happen, not what actually happened.  Whether the rule
was followed is a separate factual question.
```

P3 sends that same P2 prefix, then the shared core template, then the exact text:

```
Before answering, consider both the possibility that the
rule was followed and the possibility that it was broken.
```

P4 message sequence. P4 does not include the P2 prefix or P3 suffix. It sends four user–assistant demonstration pairs followed by the target item as a ninth message. Every demonstration user message instantiates the shared core template above with the exact premise and hypothesis in Table 10.

Table 10: Exact P4 role-message sequence. "Core" denotes the complete shared template above.

Message	Role	Content
1	user	Core with premise: "The referee warned the goalkeeper not to leave the penalty box." Hypothesis: "The goalkeeper left the penalty box."
2	assistant	`B`
3	user	Core with premise: "The pharmacist told the customer to take the tablets with water, and he swallowed them with a full glass." Hypothesis: "The customer took the tablets with water."
4	assistant	`A`
5	user	Core with premise: "The gardener was banned from pruning the hedge, and the branches stayed exactly as they were." Hypothesis: "The gardener pruned the hedge."
6	assistant	`C`
7	user	Core with premise: "The captain instructed the diver to check the oxygen gauge." Hypothesis: "The diver checked the oxygen gauge."
8	assistant	`B`
9	user	Core with the current item's premise and hypothesis.

The P4 answer-choice order is therefore B, A, C, B: neutral appears twice, while entailment and contradiction appear once each.

Corrected generative CoT. P1gen is a separate auxiliary condition, not a replacement value for legacy P1. It appends the following exact instruction to the shared core:

```
Let's think step by step.  Reason briefly, then end your
reply with exactly one line of the form:
Final answer:  <A, B, or C>
```

P1gen generates up to 256 new tokens without sampling. We strictly reparse saved completions and require an explicit final-answer/answer/option marker or an independent A/B/C answer line; unparsed outputs count as incorrect.

Balanced few-shot control. P4balanced is also separate from P4. It uses the same construction as P4 but replaces the four demonstrations with the six examples in Table 11, yielding two examples for each NLI answer choice.

Table 11: Exact P4balanced demonstrations in message order. Each row is a core-template user message followed by the listed assistant answer.

Demo	Premise	Hypothesis	Answer
1	The referee warned the goalkeeper not to leave the penalty box.	The goalkeeper left the penalty box.	B
2	The curator forbade the visitor to photograph the manuscript, yet the visitor took several photographs.	The visitor photographed the manuscript.	A
3	The gardener was banned from pruning the hedge, and the branches stayed exactly as they were.	The gardener pruned the hedge.	C
4	The captain instructed the diver to check the oxygen gauge.	The diver checked the oxygen gauge.	B
5	The pharmacist told the customer to take the tablets with water, and he swallowed them with a full glass.	The customer took the tablets with water.	A
6	The supervisor instructed the technician to restart the server, but the technician left it running without a restart.	The technician restarted the server.	C

Readout and parsing. For open-weight causal language models in P0/P2/P3/P4, we use restricted-choice readout. The runner obtains next-token log probabilities for single-token realizations of A, B, and C, with and without a leading space where available; selects the largest score; and normalizes the three scores with a softmax. These probabilities are relative to the three permitted answers rather than to the full vocabulary. We apply chat templates without an added substantive system message and disable explicit model "thinking" modes where supported.

For large-scale frontier models accessed via APIs, the parser first seeks an explicit "final answer," "answer," or "option" marker and otherwise accepts only a final standalone answer line, optionally accompanied by the matching label word. It does not use broad fallbacks that could match an article or an incidental letter in an explanation. When bounded retries still fail to parse the output, the row is marked `unparsed`. None of the audited frontier-model responses supplied complete, comparable A/B/C token log probabilities. We therefore store parsed frontier-model labels as `onehot_no_logprobs`; they support discrete-label analyses only.

C Metrics and Statistical Procedures

For an item set S, accuracy is the percentage of items assigned the gold label. We report overall accuracy (`acc`); prohibition- and command-probe accuracy (`acc_A`, `acc_D`); and pooled B/C/E/F explicit-outcome anchor accuracy (`anchor`). We retain the legacy symbols CBR_A and CBR_D for compatibility with existing analysis artifacts. Only the unstated-outcome A/D probes define the directional norm-to-fact projection rates: CBR_A is the prohibition-to-non-occurrence projection rate and CBR_D is the command-to-occurrence projection rate,

$$\mathrm{CBR}_A = \Pr(\widehat{y} = \text{contradiction} \mid A), \qquad \mathrm{CBR}_D = \Pr(\widehat{y} = \text{entailment} \mid D).$$

We also measure TgBR, the prohibition-probe entailment rate; COF_B, the B-anchor contradiction-output rate despite explicit violation; and COF_F, the F-anchor entailment-output rate despite explicit non-execution. The latter two diagnose failures of factual recognition or label mapping, not directional projection. Because A has three exhaustive labels, the underlying proportions satisfy

$$\texttt{acc_A} + \mathrm{CBR}_A + \texttt{TgBR} = 1.$$

Unless stated otherwise, we compute each metric for each model and then summarize across models, so models with more outputs receive no additional weight. Confidence intervals use 10,000 bootstrap replicates with fixed seed 20260829. Model-level group intervals resample models. Parent-family intervals use two levels: the outer loop resamples parent families and the inner loop resamples models within sampled families. The primary P2–P4 table reports paired model-bootstrap intervals without multiplicity-adjusted hypothesis tests. Exploratory P1gen, P4balanced, and auxiliary comparison families use model-wise differences, Wilcoxon signed-rank tests, and Holm correction within each reported family. Scale endpoint intervals use paired scenario bootstrap. Spearman tests assess monotone association only; a non-significant result does not establish statistical independence.

D Complete P0 Results for All Models

The primary group retains the 28 open-weight models that meet two criteria on P0 main: a most-common-label share no greater than 80%, which excludes near-constant predictors, and explicit-outcome anchor accuracy of at least 60%. All ten large-scale frontier models enter the frontier-model summary; we analyze the two groups separately. Table 12 reports the primary group summaries; full model-level values follow in Tables 13 and 14.

Table 12: P0 performance by model group: model-level medians, 95% model-bootstrap intervals, and ranges, all in percent.

Model group	Measure	n	Median [95% CI] (%)	Range (%)
Open-weight models, retained	Overall accuracy	28	60.3 [55.1, 66.3]	46.4–95.7
Open-weight models, retained	Prohibition-probe accuracy	28	6.7 [4.6, 10.8]	0.0–86.7
Open-weight models, retained	CBR_A	28	92.5 [85.0, 95.4]	13.3–100.0
Open-weight models, retained	Command-probe accuracy	28	15.8 [10.0, 37.9]	0.0–99.2
Open-weight models, retained	CBR_D	28	78.8 [61.7, 87.9]	0.8–100.0
Open-weight models, retained	Explicit-outcome anchor accuracy	28	84.5 [77.1, 89.7]	66.5–97.1
Large-scale frontier models	Overall accuracy	10	93.1 [87.0, 95.8]	74.4–96.7
Large-scale frontier models	Prohibition-probe accuracy	10	74.2 [42.5, 92.5]	20.0–98.3
Large-scale frontier models	CBR_A	10	25.8 [7.5, 57.5]	1.7–80.0
Large-scale frontier models	Command-probe accuracy	10	100.0 [93.8, 100.0]	48.3–100.0
Large-scale frontier models	CBR_D	10	0.0 [0.0, 6.2]	0.0–51.7
Large-scale frontier models	Explicit-outcome anchor accuracy	10	96.3 [94.4, 97.1]	92.9–97.5

D.1 All Retained Open-Weight Models

Table 13: P0 results (%) for the 28 retained open-weight models.

Model	Accuracy	acc_A	CBR_A	acc_D	CBR_D	Anchor accuracy
gemma3-12b	61.3	2.5	97.5	6.7	93.3	89.6
gemma3-27b	65.6	6.7	93.3	27.5	72.5	89.8
gemma4-e4b	66.1	10.0	89.2	14.2	85.8	93.1
llama31-8b	48.9	6.7	84.2	4.2	95.8	70.6
llama33-70b	71.3	11.7	88.3	28.3	71.7	96.9
ministral-8b	58.5	15.8	83.3	10.8	89.2	81.0
mistral-7b	52.4	3.3	96.7	13.3	86.7	74.4
mistral31-24b	76.5	14.2	85.8	88.3	11.7	89.2
olmo3-7b	53.1	0.0	100.0	0.0	100.0	79.6
phi4-14b	77.1	15.8	84.2	92.5	7.5	88.5
phi4-mini	55.7	5.8	94.2	24.2	75.8	76.0
qwen25-14b	73.2	20.0	80.0	78.3	21.7	85.2
qwen25-14b-base	54.0	5.8	92.5	19.2	80.8	74.8
qwen25-32b	80.0	17.5	82.5	80.8	19.2	95.4

Continued on next page

Table 13 (continued)

Model	Accuracy	acc_A	CBR_A	acc_D	CBR_D	Anchor accuracy
qwen25-3b	46.4	0.0	100.0	12.5	70.8	66.5
qwen25-72b	73.8	8.3	91.7	55.0	45.0	94.8
qwen25-7b	57.2	15.8	84.2	47.5	52.5	70.0
qwen3-1.7b	46.9	0.0	100.0	0.0	76.7	70.4
qwen3-14b	65.7	6.7	93.3	14.2	85.8	93.3
qwen3-32b	66.5	0.0	98.3	17.5	82.5	95.4
qwen3-4b	59.4	2.5	97.5	9.2	90.8	86.3
qwen3-4b-base	54.6	3.3	96.7	0.0	100.0	81.0
qwen3-8b	58.9	15.8	84.2	2.5	97.5	83.8
qwen3-8b-base	50.7	9.2	69.2	2.5	97.5	73.1
qwen35-27b	95.7	86.7	13.3	99.2	0.8	97.1
qwen35-4b	62.2	6.7	93.3	79.2	20.8	71.9
qwen35-4b-base	52.1	0.0	98.3	0.0	98.3	78.1
qwen35-9b	78.1	7.5	92.5	83.3	16.7	94.4

D.2 All Large-Scale Frontier Models

Table 14: P0 main-split results (%) for all ten large-scale frontier models. "Unparsed" gives the P0-main count; no unparsed frontier-model output occurs in the unstated-outcome A/D probes.

Model	Accuracy	acc_A	CBR_A	acc_D	CBR_D	Anchor accuracy	Unparsed
claude-opus-5	88.5	57.5	42.5	100.0	0.0	93.3	2
gpt-5.6-sol	95.1	80.8	19.2	100.0	0.0	97.5	0
deepseek-v4-flash	96.5	92.5	7.5	100.0	0.0	96.7	0
glm-5	92.1	67.5	32.5	100.0	0.0	96.3	3
gpt-5.4-mini-2026-03-17	74.4	20.0	80.0	48.3	51.7	94.6	0
grok-4.3	94.2	93.3	6.7	100.0	0.0	92.9	0
kimi-k3	95.7	89.2	10.8	100.0	0.0	96.3	0
gpt-4.1-2025-04-14	83.8	27.5	72.5	87.5	12.5	96.9	0
openai.gpt-5.5	96.7	98.3	1.7	100.0	0.0	95.4	0
claude-opus-4-8	90.3	51.7	43.3	100.0	0.0	97.5	0

E Sensitivity to Model Family, Scale, and Training Stage

E.1 Family- and Series-Equal Reweighting

Table 15 reports model-equal means and sensitivity estimates that aggregate first within each parent family or series, and Table 16 lists the per-family means that enter the family-equal aggregate. Both reweightings preserve the directional pattern, but their intervals describe this model selection rather than an i.i.d. population of models.

Table 15: Family- and series-equal sensitivity estimates (%) for 28 retained open-weight models.

Measure	Model-equal mean	Family-equal [95% clustered CI]	Series-equal [95% clustered CI]
acc_A	10.7	8.3 [4.2, 12.4]	9.9 [6.1, 15.5]
CBR_A	88.0	90.6 [86.2, 95.2]	88.9 [83.4, 92.9]
acc_D	32.5	27.3 [10.7, 48.6]	29.6 [14.7, 47.5]
CBR_D	66.0	72.3 [51.0, 89.1]	69.8 [52.1, 84.8]
Explicit-outcome anchor accuracy	83.6	83.5 [78.9, 88.2]	83.9 [79.4, 88.2]

Table 16: P0 means (%) within each parent family.

Parent family	n	acc_A	CBR_A	acc_D	CBR_D	Anchor accuracy
Gemma	3	6.4	93.3	16.1	83.9	90.8
Llama	2	9.2	86.3	16.3	83.8	83.8
Mistral	3	11.1	88.6	37.5	62.5	81.5
OLMo	1	0.0	100.0	0.0	100.0	79.6
Phi	2	10.8	89.2	58.3	41.7	82.3
Qwen	17	12.1	86.3	35.3	62.2	83.0

Across the observed model selections, the Spearman correlations between anchor accuracy and acc_A are $\rho = .356$ ($p = .063$, 28 retained open-weight models) and $-.232$ ($p = .519$, ten large-scale frontier models). These tests show no stable monotone relationship in the current model selections; they do not establish independence.

E.2 Scale within Series

The confirmatory scale analysis requires at least three sizes and evaluates Qwen2.5, Qwen3, and quality-filtered Qwen3.5 (Table 17). Qwen2.5 and Qwen3 are non-monotone on all three reported measures. Quality-filtered Qwen3.5 is monotone in this three-point subset. Restoring small checkpoints that fail the quality gate yields a prohibition-probe endpoint difference of approximately zero (-0.8 pp) with an interval that crosses zero. Gemma 3 and Phi-4 provide only two-point supplements and do not establish a general scaling law.

Table 17: Quality-filtered within-series scale trends. Endpoint differences, in percentage points (pp), compare the largest and smallest retained sizes with paired scenarios.

Series	Measure	n	Spearman ρ (p)	Monotone	Endpoint Δ [95% CI] (pp)
Qwen2.5	acc_A	5	.300 (.624)	No	8.3 [4.2, 13.3]
Qwen2.5	acc_D	5	.700 (.188)	No	42.5 [32.5, 52.5]
Qwen2.5	Anchor accuracy	5	.900 (.037)	No	28.3 [25.4, 31.2]
Qwen3	acc_A	5	.103 (.870)	No	0.0 [0.0, 0.0]
Qwen3	acc_D	5	.900 (.037)	No	17.5 [10.8, 24.2]
Qwen3	Anchor accuracy	5	.900 (.037)	No	25.0 [22.7, 27.3]
Qwen3.5	acc_A	3	1.000 (.333)	Yes	80.0 [72.5, 86.7]
Qwen3.5	acc_D	3	1.000 (.333)	Yes	20.0 [13.3, 27.5]
Qwen3.5	Anchor accuracy	3	1.000 (.333)	Yes	25.2 [21.9, 28.3]

For the three-point Qwen3.5 series, the table reports the exact two-sided permutation value ($2/3! = .333$) rather than an unreliable asymptotic Spearman value. Some series improve, especially on the command probe and explicit-outcome anchors, but larger scale does not guarantee monotone improvement on the prohibition probe.

E.3 Base–Instruct Comparison

We compare six same-size Qwen base–instruct pairs to describe checkpoint-level behavioral differences, not to identify a causal training stage. The attribution gate is stricter than the primary quality filter: both checkpoints must be non-degenerate, reach explicit-outcome anchor accuracy of at least 80%, and have a most-common-label share of at most 80%. In five of the six pairs, at least one checkpoint falls below the anchor threshold, so differences on the unstated-outcome probes cannot be cleanly separated from insufficient task performance (Figure 6a); only the Qwen3-4B pair passes. Relative to Qwen3-4B-Base, Qwen3-4B shows a 0.8 pp decrease in Acc_A, a 9.2 pp increase in Acc_D, and a 5.2 pp increase in explicit-outcome anchor accuracy (Figure 6b). Table 18 lists all six pairs and is descriptive except for that comparison. With one passing pair and opposite changes on the prohibition and command probes, these results do not attribute norm-to-fact projection to instruction tuning, nor do they isolate pretraining, supervised fine-tuning, preference optimization, or reinforcement learning.

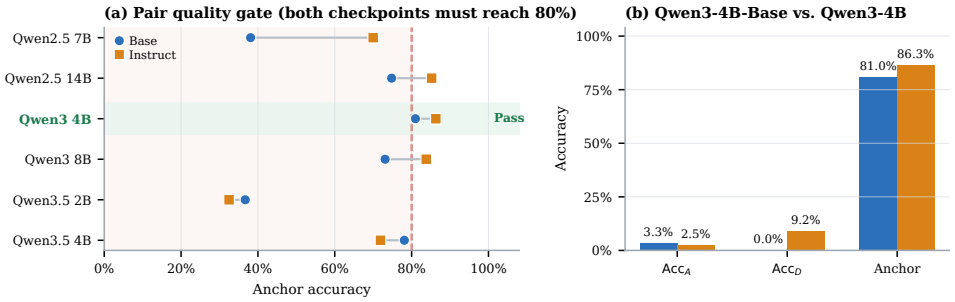

Figure 6: Base–instruct comparison under P0. Panel (a) shows explicit-outcome anchor accuracy for six same-size Qwen pairs; both checkpoints must reach 80% to pass the quality screening. Panel (b) reports accuracy on the unstated-outcome probes and explicit-outcome anchors for the only pair that passes.

Table 18: P0 behavioral differences from base to instruct models. Anchor accuracies are percentages; deltas are instruct minus base in percentage points (pp).

Pair	Quality gate	Base anchor accuracy (%)	Instruct anchor accuracy (%)	$\Delta\mathrm{acc}_A$	$\Delta\mathrm{acc}_D$	Δ anchor accuracy
qwen25-7b-base→qwen25-7b	Descriptive	38.1	70.0	−84.2	40.0	31.9
qwen25-14b-base→qwen25-14b	Descriptive	74.8	85.2	14.2	59.2	10.4
qwen3-4b-base→qwen3-4b	Pass	81.0	86.3	−0.8	9.2	5.2
qwen3-8b-base→qwen3-8b	Descriptive	73.1	83.8	6.7	0.0	10.6
qwen35-2b-base→qwen35-2b	Descriptive	36.7	32.5	−17.5	38.3	−4.2
qwen35-4b-base→qwen35-4b	Descriptive	78.1	71.9	6.7	79.2	−6.2

F Prompt Interventions: Full Comparisons

F.1 P2–P4 versus P0

Table 19 provides the complete P2–P4 comparison with P0; the exact prompt construction appears in Appendix B.

Table 19: Median within-model changes from P0, in percentage points (pp), with 95% model-bootstrap intervals.

Model group	Contrast	n	$\Delta\mathrm{acc}_A$ [95% CI] (pp)	Δ anchor accuracy [95% CI] (pp)
Open-weight models, retained	P2−P0	28	12.5 [2.1, 26.7]	−2.3 [−4.0, 0.8]
Open-weight models, retained	P3−P0	28	12.9 [8.3, 22.1]	−6.5 [−14.8, −2.8]
Open-weight models, retained	P4−P0	28	2.9 [−0.4, 10.0]	3.4 [1.7, 6.4]
Large-scale frontier models	P2−P0	10	25.8 [7.5, 51.7]	−2.3 [−5.8, 0.6]
Large-scale frontier models	P3−P0	10	25.8 [7.5, 56.7]	−4.6 [−7.5, −0.4]
Large-scale frontier models	P4−P0	10	13.7 [4.2, 37.5]	1.8 [0.8, 3.7]

Five of the six prohibition-probe intervals exclude zero; the exception is P4 for open-weight models ([−0.4, 10.0] pp). In both groups the instruction-based prompts (P2, P3) yield larger probe gains than the demonstration-only prompt (P4). The anchor columns show the trade-off: P3 lowers anchor accuracy in both groups with intervals excluding zero, P2 does so with intervals that include zero, and P4 alone raises anchor accuracy with intervals excluding zero. Wide probe intervals (up to 50 pp for frontier models) indicate that the gains are dominated by a few responsive models. The prompts therefore shift the probe–anchor balance in a prompt-specific way; none reliably reduces norm-to-fact projection across all models without cost on the anchors.

F.2 Corrected CoT and Balanced Few-Shot Prompts

Corrected P1gen covers seven open-weight models and 720 main items per model; minimum strict parse coverage is 92.1%, and unparsed completions count as errors. Table 20 gives the complete exploratory comparison; generated rationales need not faithfully expose internal reasoning.

Table 20: Corrected generative CoT results for the seven-model subset. P0 and P1gen columns are percentages; changes and their intervals are percentage points (pp).

Measure	P0 (%)	P1gen (%)	Median Δ (pp)	95% CI (pp)	p	Holm q
Overall accuracy	58.9	77.2	14.0	[10.3, 15.8]	.031	.125
acc_A	10.0	34.2	19.2	[8.3, 30.0]	.031	.125
acc_D	14.2	71.7	15.8	[2.5, 52.5]	.047	.125
Anchor accuracy	83.8	90.6	5.4	[−1.3, 15.0]	.109	.125

All four medians move in the same direction, and the largest gain is on the command probe (14.2% to 71.7%), whereas prohibition-probe accuracy remains at 34.2% even after reasoning. The anchor interval includes zero, and every Holm-adjusted q equals .125, so no endpoint is confirmatory in this seven-model subset. Generative reasoning is thus associated with fewer projected outcomes, most clearly for commands, but it does not remove the effect on prohibitions and does not establish a reliable repair.

P4balanced covers eight open-weight models and 720 main items per model. Table 21 provides both paired comparisons; Appendix B gives the exact demonstrations.

Relative to P0, the balanced six-shot prompt raises overall, command-probe, and anchor accuracy with intervals excluding zero, whereas the prohibition-probe interval includes zero, and only the

Table 21: Balanced six-shot paired comparisons for eight open-weight models. Changes and intervals are in percentage points (pp).

Contrast	Measure	Median Δ (pp)	95% CI (pp)	p	Holm q
P4balanced−P0	Overall accuracy	8.7	[3.6, 16.4]	.008	.062
	acc_A	9.2	[−2.5, 32.5]	.055	.273
	acc_D	15.8	[0.8, 42.5]	.008	.062
	Anchor accuracy	3.5	[0.4, 10.8]	.023	.141
P4balanced−P4	Overall accuracy	1.7	[0.1, 2.4]	.180	.719
	acc_A	3.3	[0.0, 10.0]	.438	.875
	acc_D	0.0	[−5.0, 1.7]	.688	.875
	Anchor accuracy	1.6	[−0.4, 4.8]	.195	.719

overall and command-probe contrasts approach the adjusted threshold ($q = .062$). Relative to P4, all four medians are within 3.3 pp and every adjusted q exceeds .7, so balancing the demonstrations does not measurably change the outcome beyond what the original four-shot prompt already achieves. Label balance in the demonstrations is therefore not the missing ingredient: few-shot prompting yields modest, mostly command-side gains regardless of the label mix.

G Replications and Controls under P0: Frames, Negation, and Fillers

The P0 within-frame, control, and filler outputs are complete for all 28 retained open-weight and 10 large-scale frontier models: 540, 100, and 60 rows per model, respectively. Table 22 summarizes all three partitions, and Table 23 breaks the within-frame partition down by frame. Every partition has zero duplicate keys and zero unknown UIDs. Unparsed completions remain in the denominator and count as incorrect; parsed coverage is 100.0% for open-weight models, while the lowest frontier-model parsed coverage is 98.9% for within-frame and 99.0% for controls.

The within-frame partition selects 30 subclass-balanced scenarios and expands every scenario into all six A–F cells under all three frames. The bare frame uses the original premise. The narrative frame prepends the scenario's validated narrative opener. The legal frame prepends its validated legal opener, lowercasing the original premise only when the opener ends with a comma. Thus, the 180 items per frame share the same scenarios and cell structure rather than constituting independent samples.

Table 22: Validated P0 partition summaries (%). Values are model-level medians [95% model-bootstrap interval]. Control rows report neutral accuracy; filler rows report gold-label accuracy.

Partition	Cohort	n	Statistic (%)
Within-frame	Open-weight	28	Overall 64.2 [57.8,70.0]; Acc_A 11.7 [6.1,18.3]; Acc_D 17.2 [7.8,43.3]; Anchor 88.7 [78.6,93.9]
Within-frame	Large-scale frontier	10	Overall 94.5 [89.4,97.4]; Acc_A 76.1 [51.7,97.8]; Acc_D 100.0 [95.0,100.0]; Anchor 97.5 [95.8,98.9]
Controls	Open-weight	28	Requirement without negation 9.0 [2.0,21.0]; negated prohibition 5.0 [0.0,8.0]
Controls	Large-scale frontier	10	Requirement without negation 90.0 [52.0,100.0]; negated prohibition 88.0 [58.0,96.0]
Fillers	Open-weight	28	Neutral 100.0 [100.0,100.0]; entailment 100.0 [100.0,100.0]
Fillers	Large-scale frontier	10	Neutral 100.0 [100.0,100.0]; entailment 100.0 [97.5,100.0]

Table 23: Frame-specific model-level accuracy medians (%) for the matched within-frame partition.

Cohort	Frame	Overall	Acc_A	Acc_D	Anchor
Open-weight	Bare	64.7	6.7	20.0	90.0
	Narrative	65.6	16.7	20.0	88.8
	Legal	62.8	6.7	8.3	88.3
Large-scale frontier	Bare	94.4	83.3	100.0	96.3
	Narrative	94.4	71.7	100.0	97.9
	Legal	94.7	81.7	100.0	98.3

H FORMAT ROBUSTNESS: LENGTH, OPTION ORDER, AND LABEL SURFACE

H.1 LENGTH-MATCHED NEUTRAL CONTINUATIONS

Unstated-outcome probes are shorter than explicit-outcome anchors, so a model might respond to premise length rather than to the absence of an outcome. We compare baseline probes with versions that append a neutral setting description without mentioning the actor, action, or result. The length-matched auxiliary dataset contains 240 rows per model: one A and one D version for each of 120 scenarios. For each scenario, the builder measures the additional word counts in the explicit-outcome B and C premises relative to A, averages them, and chooses the closest of 12 fixed setting-only continuations describing weather, light, ambient sound, or other setting details. The formal paired analysis covers 16 retained open-weight models and filters this dataset to the 40 scenario IDs assigned to the bare frame in the main set, yielding 40 matched observations per probe and condition.

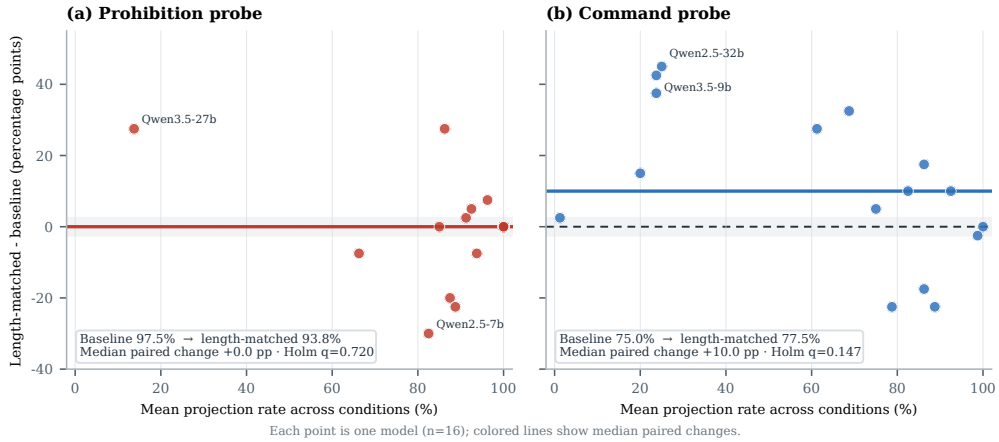

Figure 7: Within-model differences after adding length-matched neutral continuations for (a) the prohibition probe and (b) the command probe. Each point is one model; the horizontal coordinate averages its baseline and length-matched projection rates, and the vertical coordinate is the within-model change in percentage points (pp).

The median prohibition-to-non-occurrence projection rate changes from 97.5% at baseline to 93.8% after length matching; the median within-model paired change is 0.0 pp (95% CI $[-7.5, 2.5]$ pp, $p = .720$, Holm $q = .720$; Figure 7a). The median command-to-occurrence projection rate changes from 75.0% to 77.5%; the median paired change is 10.0 pp (95% CI [0.0, 27.5] pp, $p = .073$, Holm $q = .147$; Figure 7b). Neither comparison survives Holm correction, and the command-probe change, if anything, points toward more rather than less projection. In this subset, simple length mismatch therefore does not explain the effect, although this control does not exhaust possible effects of syntactic complexity or information density.

H.2 ANSWER ORDER AND LABEL SURFACE

A second format control asks whether the effect depends on where the neutral option appears or on how the labels are written. The option-robustness experiment covers seven models and contains 12,960 rows per model. It crosses all six displayed semantic-label orders—E–N–C, E–C–N, N–E–C, N–C–E, C–E–N, and C–N–E—with three surfaces. The *letters* surface uses A/B/C tied to displayed slots; the *truth terms* surface uses True/Unknown/False tied to entailment/neutral/contradiction; and the *class names* surface uses Entailment/Neutral/Contradiction. Each of the 18 interfaces evaluates all 720 main-set items, and each matrix cell in Figure 8a–c uses its 120 prohibition-probe items per model.

Across the 18 interfaces, the model-level median prohibition-to-non-occurrence projection rate (CBR_A) ranges from 78.3% to 97.5% (Figure 8a–c). Predicted answers occupy displayed positions 1, 2, and 3 with shares of 35.2%, 32.9%, and 31.9% (Figure 8d), close to the uniform 33.3%. A fixed answer-position preference therefore does not by itself explain the effect: the projection rate

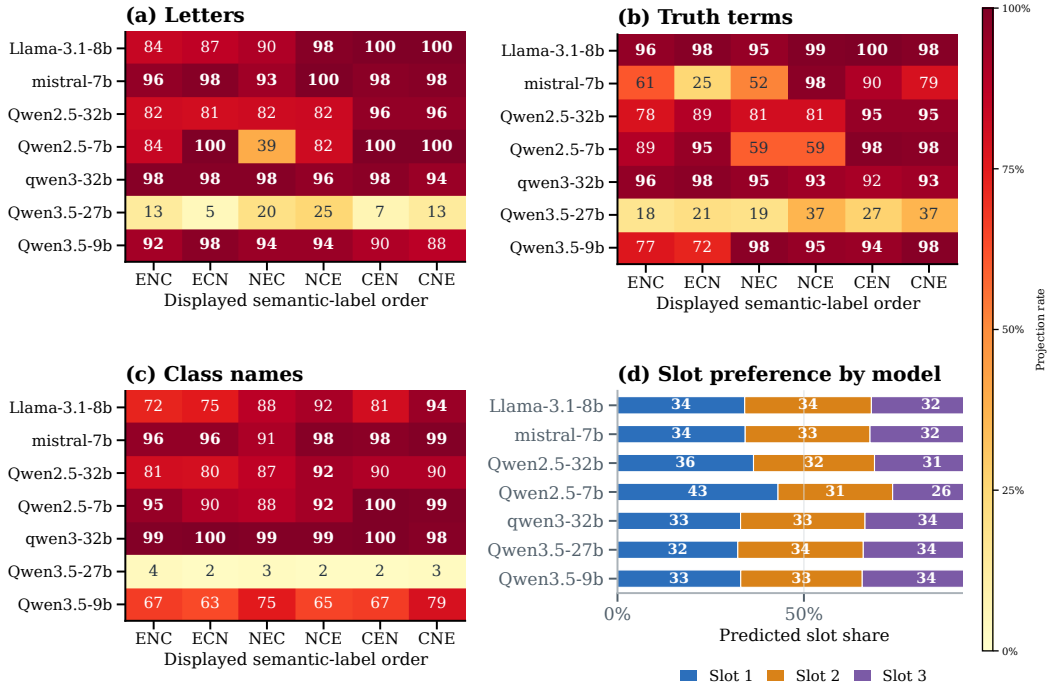

Figure 8: Answer-order and label-surface robustness for seven open-weight models. Panels (a)–(c) report the prohibition-to-non-occurrence projection rate (CBR_A) for the six semantic-label orders under letter, truth-term, and class-name interfaces; E, N, and C denote entailment, neutral, and contradiction in displayed order. Panel (d) shows each model's distribution over the three displayed answer slots. All panels use prohibition-probe predictions only.

stays high under every order and surface for most models. Individual models are nevertheless sensitive to the interface—Qwen3.5-27b, for example, projects far less under the letter and class-name surfaces than under truth terms—so this evidence is descriptive robustness across interfaces rather than strict invariance.

I Calibration and Selective Prediction

This appendix documents the data, metric definitions, and complete results behind the confidence analysis in Section 5.7. The analysis uses only P0 main rows from the 28 retained open-weight models with genuine restricted-choice probabilities. It contains 20,160 rows: 6,720 A/D probes and 13,440 B/C/E/F anchors. No frontier-model row enters this analysis. Legacy five-decimal serialization caused small probability-sum errors. We explicitly renormalized 3,587 rows within tolerance 1.1×10^{-5}; the largest observed error was approximately 1.00×10^{-5}.

For confidence $c_i = \max_k p_{ik}$, we define ten-bin equal-width expected calibration error as

$$\mathrm{ECE} = \sum_b \frac{|I_b|}{n} \left| \mathrm{acc}(I_b) - \frac{1}{|I_b|} \sum_{i \in I_b} c_i \right|.$$

The multiclass Brier score is $n^{-1} \sum_i \sum_k (p_{ik} - \mathbf{1}[y_i = k])^2$, where $\mathbf{1}$ is an indicator, and NLL is $-n^{-1} \sum_i \log p_{i,y_i}$. For risk–coverage, we sort by confidence, retain complete confidence-tie groups, and use right-step integration for AURC. For error rate e, we report $\mathrm{eAURC} = \mathrm{AURC} - [e + (1-e)\log(1-e)]$, which subtracts the continuous accuracy-matched oracle baseline. This baseline is an asymptotic reference rather than the exact finite-sample right-step optimum.

Table 24 reports model-equal medians for all metrics, and Table 25 reports the pooled confidence bins from which the high-confidence columns of Table 4 are derived. Pooled ECE (computed over

all rows rather than per model) is 66.8% for the unstated-outcome probes and 9.3% for the explicit-outcome anchors, close to the model-equal medians of 71.4% and 10.0%. The paired probe-minus-anchor AURC difference is 86.1 pp (95% model-bootstrap CI [73.6, 91.0] pp, Wilcoxon $p = 7.45 \times 10^{-9}$), and eAURC is worse on the probes for 25 of 28 models. These results do not extend to frontier-model calibration or unrestricted full-vocabulary confidence.

Table 24: Model-equal median probability metrics; n denotes the pooled row count. Accuracy, ECE, AURC, and eAURC are percentages; Brier and NLL retain their native scales.

Item subset	n	Accuracy (%)	ECE (%)	Brier	NLL	AURC (%)	eAURC (%)
Unstated-outcome probes A/D	6,720	12.3	71.4	1.433	11.617	96.7	31.2
Explicit-outcome anchors B/C/E/F	13,440	84.5	10.0	.267	.787	5.6	4.0

Table 25: Pooled equal-width calibration bins. Confidence and accuracy are percentages; empty bins are omitted.

Item subset	Confidence bin (%)	n	Mean confidence (%)	Accuracy (%)
Unstated-outcome probes A/D	30–40	36	37.7	30.6
	40–50	290	45.9	14.5
	50–60	535	55.2	17.6
	60–70	443	64.9	16.5
	70–80	375	75.2	24.0
	80–90	482	85.5	23.9
	90–100	4,559	99.1	22.5
Explicit-outcome anchors B/C/E/F	30–40	214	37.3	48.6
	40–50	790	45.4	58.9
	50–60	781	54.9	68.8
	60–70	675	65.0	69.3
	70–80	643	74.8	75.4
	80–90	796	85.4	77.5
	90–100	9,541	99.3	89.7

J Sensitivity to Pragmatic Cues and Normative Wording

All auxiliary analyses use P0 and specified subsets of open-weight models. They test sensitivity to controlled textual manipulations, not a training cause or internal mechanism. We state coverage separately for every experiment.

J.1 Monitoring and Enforcement Cues

Supervision refers to whether an authority is present to observe behavior and whether the rule is checked or enforced. We use this manipulation to measure whether factual outputs differ across enforcement cues rather than remaining fixed under every prohibition. The experiment covers 22 open-weight models and 120 scenarios, each evaluated as a prohibition-only baseline, under continuous monitoring, with no observer present, and with a rule described as never checked or enforced. None of the four versions states the outcome, so each is an unstated-outcome prohibition probe with a neutral gold label. Section J.4 gives the exact rewrites and the full paired statistics in Tables 27 and 28.

The median contradiction-output rate is 95.4% at baseline and 95.0% under continuous monitoring (Figure 9a), but is 62.1% when no observer is present (Figure 9b) and 7.1% when the rule is never checked or enforced (Figure 9c). The high-monitoring condition does not differ reliably from baseline, whereas the other two conditions have lower prohibition-to-non-occurrence projection rates.

J.2 Violation Cues

A prohibition does not establish that the prohibited act failed to occur. We measure whether outputs differ when the unstated-outcome prohibition probe adds one of four cues that makes violation more plausible: inability to comply, habitual rule breaking, a prior record of the same act, or a strong motive with no obstacle. The experiment evaluates 16 open-weight models on 120 scenarios in five versions, for 600 items per model. Each cue is compared with the same model's baseline; exact wording and complete paired results appear in Section J.4.

The median contradiction-output rate is 96.7% at baseline, compared with 82.9% under inability to comply, 83.3% for a habitual offender, 78.3% with a prior record, and 53.7% under a strong motive

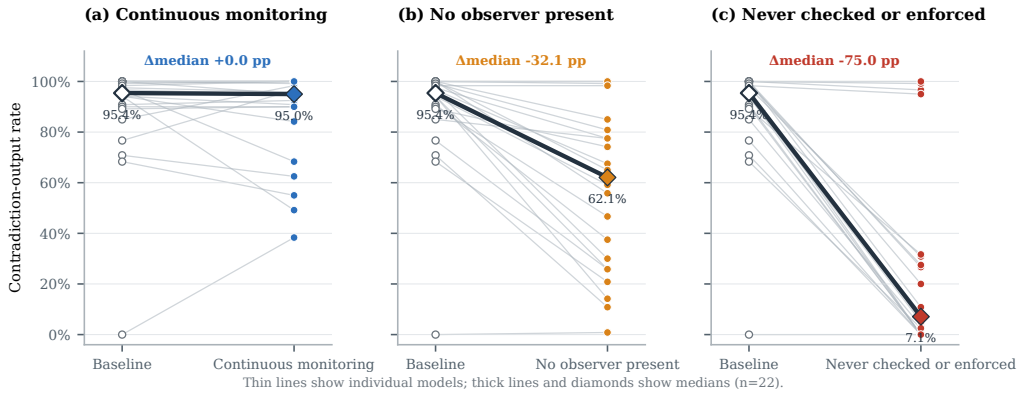

Figure 9: Within-model differences across monitoring and enforcement cues: (a) continuous monitoring, (b) no observer present, and (c) rule never checked or enforced. Thin lines connect each model's baseline contradiction-output rate to the manipulated condition; thick lines and diamonds show cross-model medians.

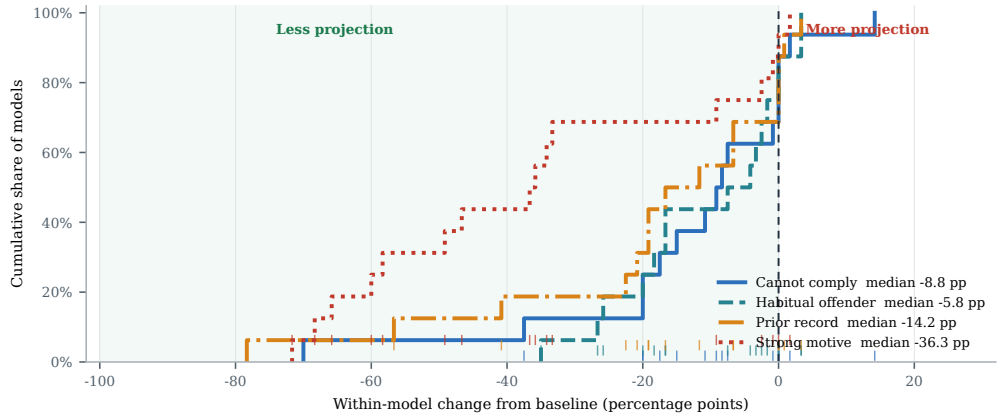

Figure 10: Empirical cumulative distributions of within-model changes in percentage points (pp) from the prohibition-only baseline. Negative values indicate less prohibition-to-non-occurrence projection. The four cues are separate manipulations, not levels on a continuous intensity scale.

(Figure 10). All four cue conditions have fewer non-occurrence outputs than baseline, with the largest difference under strong motive. Thus, factual outputs vary with the tested information about ability, history, and motivation; the primary pattern is not a fixed reaction to the prohibition wording alone.

J.3 Prohibitive Wording versus Behavioral Propensity

We further separate the normative prohibition from cues about whether the behavior is likely. Sixteen open-weight models evaluate four neutral-gold versions of the same 120 scenarios: a prohibition-only baseline, a low-occurrence cue without a prohibition, a prohibition combined with a high-violation cue, and a neutral possibility control containing neither a prohibition nor a directional propensity cue. Table 26 reports contradiction-output rates for all conditions; for rows without a prohibition, this is an output rate rather than CBR_A.

The low-occurrence-cue condition has a 100.0% median contradiction-output rate, whereas adding a high-violation cue to a prohibition corresponds to a 55.4% median rate. The neutral possibility control has a 0.0% median rate. These contrasts show that the observed output pattern varies with both normative wording and behavioral-propensity cues; no single prohibition token accounts for the pattern. Section J.4 gives the exact templates and the paired comparisons in Table 28.

Table 26: Separating prohibitive wording from behavioral-propensity cues (16 open-weight models). All outcomes are unstated and all gold labels are neutral.

Condition	Prohibition	Cue	**Median contradiction-output rate (%)**
Prohibition only	Yes	None beyond the prohibition	96.7
Low-occurrence cue only	No	No prior inclination to perform the act	100.0
Prohibition plus high-violation cue	Yes	Strong temptation and no intervention	55.4
Neutral possibility control	No	Opportunity to act, without a directional cue	0.0

J.4 Cue Experiments: Exact Rewrites and Paired Statistics

The supervision set covers 22 of 28 retained open-weight models and contains 480 rows for each complete model. Every version retains the original prohibition and an unstated outcome. For person authorities, the appended sentences are "*[Authority] stayed right there the whole time*," "*[Authority] left and was away the whole time*," and "*[Authority] never checks and has never enforced that rule*." For institutional authorities, the corresponding sentences are "*A supervisor was present and monitoring throughout*," "*No one was present to observe at any point*," and "*The rule is never checked and has never been enforced*." The baseline receives no appended sentence.

The high-violation set covers 16 models and contains 600 rows per model. It appends one of four outcome-silent cues to the original prohibition: "*The instruction was far too complex for anyone there to follow*" (cannot comply), "*Every rule set there had been broken before, without exception*" (habitual offender), "*There were records of the same act on eleven previous occasions*" (prior record), or "*The temptation was overwhelming and no obstacle stood in the way*" (strong motive).

The operator ablation also covers 16 models and contains 480 rows per model. Its internal condition names map to the paper labels as follows: `deontic_lowprob` is the original prohibition ("prohibition only"); `prob_only` states "*[Subject] had never shown any inclination to [act]*" ("low-occurrence cue only"); `deontic_highprob` appends "*and the temptation was overwhelming with no one nearby to intervene*" to the prohibition; and `neutral_base` states only that the subject was in the room where the act was possible. All conditions have neutral gold labels. Table 27 reports the median proportion of contradiction outputs. This quantity is the prohibition-to-non-occurrence projection rate when a prohibition is present. In operator-free ablations, it is only an output rate and not a norm-to-fact projection measure.

Table 27: Auxiliary condition results (%). Intervals bootstrap models.

Experiment	Condition	Models evaluated	Median contradiction-output rate (%)	95% CI (%)
Supervision	baseline	22	95.4	[90.4, 99.2]
	high watch	22	95.0	[90.0, 98.8]
	low watch	22	62.1	[31.7, 77.5]
	no enforcement	22	7.1	[1.3, 27.5]
High violation	baseline	16	96.7	[89.2, 100.0]
	cannot comply	16	82.9	[66.7, 92.5]
	habitual offender	16	83.3	[67.5, 95.8]
	prior record	16	78.3	[48.3, 90.8]
	strong motive	16	53.7	[22.5, 72.5]
Operator ablation	deontic, high probability	16	55.4	[20.0, 69.2]
	deontic, low probability	16	96.7	[89.2, 100.0]
	neutral base	16	0.0	[0.0, 0.0]
	probability cue only	16	100.0	[99.2, 100.0]

Nominally identical baselines in independently run auxiliary sets have median prediction disagreement of 0.0% under both P0 and P2; the largest model-by-dataset disagreements are 0.8% and 3.3%, respectively. We do not assign a mechanism interpretation to effects near this scale.

J.5 Negative-Normative Wording Tiers

We compare contradiction outputs as negative-normative wording changes from requests and advice to explicit prohibition. Twenty-two open-weight models evaluate 120 scenarios under 12 expres-

Table 28: Within-model auxiliary contrasts in percentage points (pp). Holm correction applies within each reported comparison family.

Within-model contrast	n	Median Δ [95% CI] (pp)	p	Holm q
Supervision: high watch − baseline	22	0.0 [−3.3, 0.4]	.394	.394
Supervision: low watch − baseline	22	−32.1 [−47.5, −15.0]	1.82×10^{-4}	3.63×10^{-4}
Supervision: no enforcement − baseline	22	−75.0 [−87.5, −65.8]	8.82×10^{-5}	2.65×10^{-4}
Habitual offender − baseline	16	−5.8 [−18.3, −1.7]	.006	.011
Prior record − baseline	16	−14.2 [−20.8, 0.0]	.003	.009
Cannot comply − baseline	16	−8.8 [−17.5, 0.0]	.011	.011
Strong motive − baseline	16	−36.3 [−58.3, −9.2]	9.87×10^{-4}	.004
Probability only − deontic low-probability	16	1.7 [0.0, 9.2]	.167	.167
Deontic high- − low-probability	16	−39.6 [−56.7, −19.2]	.001	.003
Neutral base − deontic low-probability	16	−95.4 [−99.2, −83.3]	6.32×10^{-4}	.002

sions grouped a priori into five tiers, yielding 1,440 items per model. Because the tiers jointly change lexical choice and syntax, they are treated as discrete wording groups rather than points on a precisely measured continuous scale. Section J.6 lists the exact expressions, and Table 29 reports the intervals and paired comparisons.

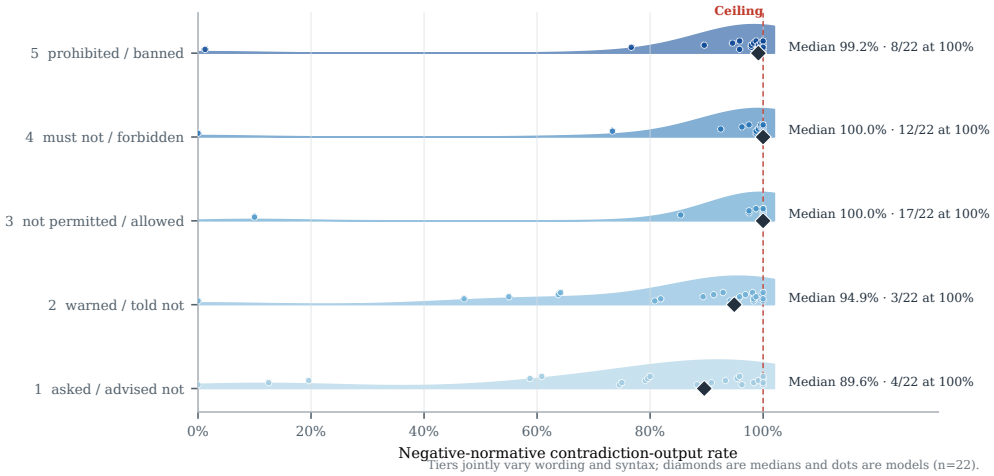

Figure 11: Model-level distributions of the negative-normative contradiction-output rate across five discrete wording tiers. Diamonds show cross-model medians, dots show individual models, and the annotations report the number of models at the 100% ceiling.

The median negative-normative contradiction-output rate is 89.6%, 94.9%, 100.0%, 100.0%, and 99.2% from tiers 1 through 5 (Figure 11). Every tier is reliably above tier 1 (paired differences of 2.9–9.6 pp, Holm $q \leq .010$; Table 29), but the rate reaches the ceiling by tier 3, and the strongest tier does not exceed tier 4 (median difference -0.2 pp, 95% CI [-0.8, 0.0] pp, Holm $q = .102$). Outputs therefore differ across wording tiers, but the observed pattern is not restricted to words such as *prohibited* or *forbidden*, nor is it a smooth five-step dose response: requests and advice alone already yield contradiction outputs for about nine in ten items. Because requests and advice are not strict prohibitions, only the strict-prohibition tier supports the narrower interpretation as prohibition-to-non-occurrence projection.

J.6 Deontic Force: Exact Expressions and Tier Results

The deontic-force experiment covers 22 models and contains 1,440 rows per model. The 12 builder-defined expressions and their a priori tiers are: tier 1, "*[Authority] asked [subject] not to [act]*" and "*[Authority] advised [subject] not to [act]*"; tier 2, "*[Authority] warned [subject] never to [act],*" "*[Authority] told [subject] not to [act],*" "*[Subject] was expected by [authority] not to [act],*" and "*According to [authority], [subject] should not [act]*"; tier 3, "*[Subject] was not permitted by [authority] to [act]*" and "*[Subject] was not allowed by [authority] to [act]*"; tier 4, "*Under [authority]'s instructions, [subject] must not [act]*" and "*[Subject] was forbidden by [authority] to [act]*"; and tier 5, "*[Subject] was prohibited by [authority] from [acting]*" and "*[Subject] was banned by [authority] from [acting].*" The bracketed fields are instantiated from each scenario, with

grammatical object and gerund forms where required. Table 29 reports the tier medians, model-bootstrap intervals, and paired contrasts against tier 1.

Table 29: Deontic-force tiers for 22 models. Rates and intervals are percentages; paired differences are percentage points (pp).

Tier	Median negative-normative contradiction-output rate (%)	95% CI (%)	Selected paired contrast vs. tier 1
1	89.6	[76.9, 96.0]	—
2	94.9	[85.1, 98.2]	$\Delta = 2.9$ pp, Holm $q = .010$
3	100.0	[100.0, 100.0]	$\Delta = 9.6$ pp, Holm $q = 9.82 \times 10^{-4}$
4	100.0	[99.0, 100.0]	$\Delta = 7.5$ pp, Holm $q = .001$
5	99.2	[97.9, 100.0]	$\Delta = 7.5$ pp, Holm $q = .001$

J.7 Modal Families

The modal-family experiment covers 22 models and contains 1,080 rows per model. It reports the rates of choosing the "did not occur" (contradiction) and "did occur" (entailment) readings (Table 30). These rates measure response tendencies rather than accuracy across heterogeneous modal statements.

Table 30: Median directional response rates (%) by modal family, with 95% model-bootstrap intervals.

Modal family	Did-not-occur rate	95% CI	Did-occur rate	95% CI
Ability	0.0	[0.0, 0.0]	84.2	[68.3, 97.1]
Attempt	0.0	[0.0, 0.0]	50.4	[10.0, 87.9]
Desire	0.0	[0.0, 0.0]	11.3	[0.8, 46.7]
Intention	0.0	[0.0, 0.0]	28.3	[2.5, 67.5]
Likely high	0.0	[0.0, 0.0]	24.6	[0.0, 97.5]
Likely low	98.8	[82.1, 100.0]	0.0	[0.0, 0.0]
Plan	0.0	[0.0, 0.0]	39.2	[8.3, 81.2]
Prohibition	95.8	[90.0, 98.3]	0.0	[0.0, 0.0]
Seldom explicit	12.5	[2.9, 34.2]	0.0	[0.0, 0.0]

These heterogeneous directions indicate that default-outcome completion extends beyond deontic vocabulary but does not produce a uniform response to every modal expression.

K Human Expert Evaluation

This evaluation examines whether external annotators avoid inferring factual outcomes directly from rules when those outcomes are unstated. Five annotators, distinct from the five experts who constructed the dataset, independently completed all 720 English items, covering 120 source scenarios under the six A–F conditions. The constructors are two doctoral researchers and three master's researchers, and all five are proficient in English. The annotators did not construct the items. Gold labels follow Table 5 rather than a majority vote, so the accuracies below measure agreement with that design. The prohibition probe A and command probe D each contain 120 items, while the explicit-outcome anchors B/C/E/F contain 480 items in total. Table 31 reports each annotator's performance relative to the gold labels, and Table 32 reports agreement among the annotators.

Table 31: Individual external-annotator results and means (%). Accuracy is computed against the dataset gold labels. CBR_A is the proportion of prohibition probes labeled as contradiction, and CBR_D is the proportion of command probes labeled as entailment.

Annotator	Overall accuracy	Acc_A	CBR_A	Acc_D	CBR_D	Anchor accuracy
1	98.3	100.0	0.0	98.3	1.7	97.9
2	97.8	99.2	0.8	97.5	1.7	97.5
3	96.7	98.3	1.7	96.7	2.5	96.3
4	98.5	98.3	0.8	97.5	1.7	98.8
5	95.0	96.7	2.5	95.0	2.5	94.6
Mean	97.3	98.5	1.2	97.0	2.0	97.0

Across the five external annotators, mean overall accuracy is 97.3%, with mean accuracies of 98.5% on the prohibition probe, 97.0% on the command probe, and 97.0% on the explicit-outcome anchors.

Table 32: Unanimous agreement among the five external annotators, defined as the proportion of items for which all five assign the same label.

Item subset	Items	Unanimous agreement (%)
All items	720	85.8
Explicit-outcome anchors B/C/E/F	480	84.8
Unstated-outcome probes A/D	240	87.9

The corresponding directional projection rates are low, with $\mathrm{CBR}_A = 1.2\%$ and $\mathrm{CBR}_D = 2.0\%$. Of the 720 items, 618 receive the same label from all five annotators, yielding 85.8% unanimous agreement. This splits into 407 of 480 explicit-outcome anchors (84.8%) and 211 of 240 unstated-outcome probes (87.9%), with slightly higher agreement on the probes, where mean accuracy is also higher. Overall, these annotators rarely treat rules as evidence of factual outcomes, supporting the neutral gold labels for the unstated-outcome probes. Because the sample contains five external annotators, these results should not be interpreted as agreement in the general population.

L Source-Backed Illustrative Cases

We construct three source-backed cases from (i) the OSHA workplace PPE hazard-assessment requirement, 29 CFR 1910.132(d)(1) (Occupational Safety and Health Administration, 2024); (ii) the FDA informed-consent requirement, 21 CFR 50.20 (Food and Drug Administration, 2024); and (iii) the CFPB Sendwave consent order, File No. 2023-CFPB-0012, paragraphs 101–102 (Consumer Financial Protection Bureau, 2023). Each case includes a norm-only probe (neutral), an explicit-occurrence control (entailment), and an explicit-non-occurrence control (contradiction), producing nine rows per model. The frozen source manifest records exact URLs, source locators, document dates, and SHA-256 hashes.

All 28 retained open-weight models completed P0 with no missing UID, duplicate key, or probability-sum failure. Automated source-fidelity and semantic checks passed. Table 33 reports model-equal accuracy for each case variant, and Table 34 breaks the norm-only probe down by case. These results are illustrative only; they do not measure legal, regulatory, audit, medical, or financial domain accuracy.

Table 33: Model-equal accuracy (%) across the three illustrative cases. Intervals bootstrap models.

Case variant	Models evaluated	Mean	Median	95% CI	Range
Unstated-outcome norm only	28	29.8	33.3	[33.3, 33.3]	0.0–66.7
Explicit occurrence	28	95.2	100.0	[100.0, 100.0]	33.3–100.0
Explicit non-occurrence	28	95.2	100.0	[100.0, 100.0]	33.3–100.0
Both explicit-outcome controls	28	95.2	100.0	[100.0, 100.0]	66.7–100.0

Table 34: Unstated-outcome norm-only accuracy (%) for each illustrative case.

Illustrative case	Mean	Median	95% CI	Range
CFPB Sendwave penalty	71.4	100.0	[100.0, 100.0]	0.0–100.0
FDA informed consent	0.0	0.0	[0.0, 0.0]	0.0–0.0
OSHA hazard assessment	17.9	0.0	[0.0, 0.0]	0.0–100.0

The pattern from the synthetic benchmark carries over to these real regulatory texts. Models read both explicit-outcome controls almost perfectly (median 100.0%), but on the norm-only probe the model-equal mean is 29.8% and the median 33.3%: two of every three predictions treat the regulatory requirement as if it settled what happened. The projection is case-dependent rather than uniform. All 28 models fail the FDA informed-consent probe, most fail the OSHA hazard-assessment probe (mean 17.9%), whereas most pass the CFPB consent-order probe (mean 71.4%), whose source text already describes an enforcement action and therefore reads less like an open rule.

M Exploratory Mechanism Studies

M.1 Asking Directly Which Outcomes Remain Possible

This exploratory study changes the question rather than the premise. Instead of asking for an NLI label, it asks the model which factual outcomes are still possible given the text and offers three

options: the action occurred, the action did not occur, or either could still be the case. The correct answer for every unstated-outcome probe is "either", and for every explicit-outcome anchor it is the stated outcome. We evaluate all six option orders. Two wordings are compared: "guided" adds an explicit warning not to assume that the rule was followed; "neutral" asks only which outcomes remain possible. In each wording, each of three models answers the 720 main items under six option orders, producing 4,320 rows per model, with no duplicate, unknown-UID, or prompt-mode mismatch. For each item, we average the option probabilities across the six orders and take the largest average as the prediction used for A, D, and anchor accuracy. "Order consistency", computed on anchor items only, is the proportion of items for which all six order-specific predictions agree with this aggregate prediction. Table 35 reports the results for both wordings.

Table 35: Direct outcome question (%). "Pass" indicates that the model and wording meet the experiment's task-control criteria; it does not provide evidence about internal representations.

Mode	Model	A-probe accuracy	D-probe accuracy	Explicit-outcome anchor accuracy	Order consistency	Task control
Guided	`qwen35-9b`	81.7	100.0	83.1	91.0	Fail
Guided	`qwen3-8b`	91.7	100.0	39.8	73.6	Fail
Guided	`qwen35-27b`	98.3	100.0	86.7	96.0	Fail
Neutral	`qwen35-9b`	50.8	95.0	91.0	94.5	Fail
Neutral	`qwen3-8b`	85.0	98.3	40.2	75.7	Fail
Neutral	`qwen35-27b`	93.3	100.0	93.1	97.2	Pass

The guided wording fails the explicit-outcome anchor controls for all three models: once warned not to assume compliance, the models become less willing to commit even to outcomes that the text states outright. Under the neutral wording, `qwen3-8b` scores high on the A/D probes but reaches only 40.2% on the anchors, so it is simply answering "either" regardless of the text. `qwen35-9b` has B-anchor accuracy of 84.2%, just below the prespecified 85% cell threshold, and gives the correct "either" answer on 46.8% of the prohibition-probe items it had projected under P0. Only neutral `qwen35-27b` passes all task controls. Because the preregistered aggregate criterion requires successful controls across all three models, both wordings have `all_acceptance_criteria_pass=false`. Asking the question differently can change the answer for some models, but these results do not show that models already "know" under P0 that both outcomes remain open and lose this only when choosing an NLI label.