# EXAMINING SOCIAL ATTRIBUTION IN LLM REASONING: A THEORY-GUIDED PROBING METHODOLOGY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large language models (LLMs) are increasingly deployed in sociotechnical systems to facilitate social interactions, in which *social attribution*, the reasoning process attributing external events to the causes and reasons of agents' social behaviors, plays a critical role. Such reasoning processes involve the judgments of social cause, responsibility, and blame/credit to agents. Although attributional models are well-studied in social psychology and social cognition, represented by Attribution Theory, it is still underexplored in the AI and computing community, and LLM social reasoning in particular. This paper aims to advance the understanding of LLM in a fundamental social aspect by providing the first systematic exploration of LLM social attribution. Our work focuses on responsibility and blame attributions, examining current LLMs' performances in making responsibility and blame judgments, as well as their internal mechanisms underlying such judgments. Specifically, guided by attribution theory, we first construct a social attribution benchmark consisting of a *Vignette* subset based on classic scenarios from attribution theory research and a *Reality* subset constructed based on real-world social narratives, yielding **7,639** responsibility/blame judgment questions. On this basis, we evaluate **32** representative LLMs and **5** basic non-LLM baselines. To further explore the internal mechanisms underlying the LLM judgment process, we develop a probing-based methodology to investigate the latent-space representations of **5** key attribution dimensions and the consistency of their influences on LLM judgments compared to those in human social attribution. Our research findings reveal that current LLMs exhibit measurable but incomplete agreement with human responsibility and blame judgments, and meanwhile, this agreement is positively correlated with model size. At a deep level, some attribution dimensions are systematically decodable from specific positions in LLM hidden states, and their influences on the final judgment are consistent with those indicated by human Attribution Theory.

## 1 INTRODUCTION

In recent years, large language models (LLMs) have shown broader impacts in the social realm and been integrated into sociotechnical applications, such as social simulation (Park et al., 2023), human-AI interactions (Hadan et al., 2025), social learning (Goindani & Neville, 2019), and normative reasoning (Yuan et al., 2024). A critical topic underlying these diverse applications is the reliance on *social attribution* (Gratch et al., 2006; Malle, 2022), the social reasoning process that centers on attributing the occurrence of external events to the causes and reasons of agents' social behaviors, and assigning responsibility, blame, or credit to the involved agents in rich social interactions. Such social inference reflects how agents make sense of and act on the social world around them, and lies at the heart of social intelligence.

Social attribution is a fundamental field of research in social psychology and social cognition, and has been well-studied. For decades, researchers have established *Attribution Theory* (Shaver, 1985; Weiner, 1995; Malle et al., 2014) as the dominant theoretical framework. In these models, the central theme involves the judgments of social cause, responsibility and blame/credit. Such judgments are not driven by outcomes alone, but systematically shaped by a set of key attributional dimensions, such as *intention*, *voluntariness*, *foreknowledge*, *controllability*, and *obligation*, that characterize agents' mental states and the nature of their involvement in social events. Recently, several works propose to leverage LLMs primarily for normative or legal reasoning (Yuan et al., 2024; Guha et al., 2023). However, social attribution is still underexplored systematically in LLM social reasoning, despite its importance in broader applications.



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.

In this work, we study LLM responsibility and blame judgments and the internal representations underlying them. We first construct a benchmark with two complementary subsets based on classic attribution scenarios with controlled dimension values (*Vignette*) and real-world social narratives (*Reality*). After expert annotation and filtering, the benchmark contains 7,639 questions. We then evaluate 32 LLMs (9 closed-source and 23 open-source) and five non-LLM models, measuring exact agreement with human labels and agreement in within-scenario agent rankings. Furthermore, given the judgment process and the intermediate cognitive variables (i.e., attribution dimensions) clearly identified by attribution theory, it remains unclear whether a similar process exists in LLM social reasoning and to what extent LLM social attribution resembles that of human judgments in the internal social and cognitive mechanisms described in attribution theory. Therefore, we develop a dimension-level probing framework to examine the existence of dimension representations in LLM hidden states and their influences on the final judgments.

Specifically, probing is a well-established technique for studying cognitive and psychological representations, providing a quantitative link between theory-derived dimensions and LLM hidden states (Tak et al., 2025; Kang et al., 2025; Zhang, 2025). In our probing-based framework, to examine the existence of dimension representations, we first train linear probes on the hidden states, leveraging the annotations of the key attribution dimensions from our constructed benchmark. We then conduct an empirical analysis across layers and dimensions to investigate the representation strength of each dimension at different latent regions. To further examine the influences of dimension representations on responsibility and blame judgments, we directionally vary the LLM hidden states along specific dimension representations to quantify the extent to which the final judgments are affected, and conduct rigorous hypothesis testing to verify the consistency of these impacts with the human judgment patterns indicated in attribution theory.

Our research contributions and findings are summarized as follows:

1. We focus on the fundamental research problem of LLM social attribution with wide applications, and present the first systematic study on LLM responsibility and blame judgments as well as the internal mechanisms underlying such judgments.
2. We construct a social attribution benchmark consisting of a controlled vignette subset and a realistic social scenario subset, and conduct a comprehensive empirical evaluation on 32 representative LLMs and 5 classical baselines.
3. We develop a latent space dimension probing methodology that employs a probe-based technique to quantify the decodability of attribution dimensions and dimension-level influences, and on this basis, investigate the existence of dimensional representations in LLM hidden states and their influence mechanisms.
4. Our empirical findings demonstrate that specific attribution dimensions are represented within the latent space of LLMs, and their influences on LLM final judgments align with those of human social attribution theory.

## 2 Related Work

**LLMs for normative judgments.** Recent research systematically evaluates LLMs' attribution-related behaviors in social contexts (Xu et al., 2024). Benchmarks typically address two categories of judgment: (1) social-norm judgments, using everyday, context-dependent rules of thumb to evaluate whether models generate or select norm-consistent explanations and behavioral recommendations (Forbes et al., 2020; Jiang et al., 2021; Yuan et al., 2024; Agarwal et al., 2024; Jin et al., 2025; Ji et al., 2025; Chakraborty et al., 2025); and (2) legal judgments, grounded in statutes, precedents, and legal reasoning, assessing performance and reliability in legal question answering, case analysis, and decision recommendation (Guha et al., 2023; Fei et al., 2024; Dai et al., 2025). These studies primarily adopt a normative or prescriptive perspective. We instead take a descriptive perspective, modeling systematic patterns in LLM responsibility and blame judgments and evaluating their alignment with human judgments, without assuming that human judgments conform to universal principles of rationality. Our benchmark further annotates five theory-derived attribution dimensions to examine the intermediate variables underlying these judgments.

**LLM Probing for Mechanism Exploration.** Probing measures the decodability of target variables from LLM hidden states using lightweight predictors (Belinkov & Glass, 2019; Belinkov, 2022).

Studies examine factual and relational knowledge (Meng et al., 2022; Hernandez et al., 2024), instruction following (Heo et al., 2025), and truth-related signals (Orgad et al., 2025), sometimes combining representation analysis with activation interventions to assess their influence on outputs. Social and cognitive studies investigate modality reliance through prompt perturbations (Qi et al., 2023), stereotypical patterns in emotion attribution (Plaza-del Arco et al., 2024), and latent value structures (Kang et al., 2025). More directly related to our methodology, Tak et al. (2025) probe theory-derived cognitive appraisal dimensions and intervene on their representations to influence emotion inference. Other studies examine the decodability and steerability of self- and personality-related constructs (Ju et al., 2025) and ideological positions (Zhang, 2025). These findings motivate theory-guided probing and interventions for LLM social reasoning; however, to the best of our knowledge, prior work has not specifically probed attribution dimensions underlying responsibility and blame judgments.

# 3 Theoretical Foundation

**Responsibility and blame judgments.** Responsibility judgment assesses the extent to which an agent is accountable for an event, considering causal involvement, controllability, and mitigating factors such as coercion or lack of foreknowledge. Blame judgment further assesses whether the agent deserves moral criticism, taking justifications and excuses into account. Responsibility thus presupposes causality but is not reducible to it; similarly, responsibility is a prerequisite for blame but does not necessarily entail blameworthiness (Shaver, 1985; Weiner, 1995).

**Attribution theories and dimensions.** Classic attribution theories characterize responsibility and blame judgments through structured assessments of attribution dimensions, which capture agents' causal involvement, mental states, and situational constraints (Weiner, 1995; Shaver, 1985; Malle et al., 2014); see Appendix A.1 for detailed accounts and illustrations. Drawing on these theories, we focus on five dimensions (Appendix A.2):

1. *Intention.* Whether the agent deliberately performed the focal action according to a goal or plan. Here, intention concerns the action itself, not the outcome it produces.
2. *Voluntariness.* Whether the agent acted freely rather than under external coercion, authority, or situational pressure, even without an explicit threat.
3. *Foreknowledge.* Whether the agent knew or, given their epistemic and cognitive conditions, should have foreseen the outcome. Foreknowledge typically increases attributed responsibility and blame (Shaver, 1985).
4. *Controllability.* Whether the agent had the practical capacity to prevent the outcome, including the necessary means, skills, or opportunity.
5. *Obligation.* Whether the agent had a duty to prevent the negative outcome given the situation and social relationship, a key consideration in attributing responsibility or blame for unintentional outcomes.

# 4 Methodology

## 4.1 Data Construction

To address the lack of a dedicated data resource for systematically studying LLM social attribution, we first construct a benchmark that pairs human responsibility/blame judgments with theory-derived attribution dimensions. The benchmark comprises two complementary subsets, *Vignette* and *Reality*, supporting both behavioral evaluation and analysis of the representations underlying these judgments. Figure 1 summarizes construction and evaluation.

The *Vignette* subset draws on classic scenarios from psychological attribution research, preserving their theoretical grounding while using structured narratives to precisely control attribution dimensions. Following the vignette paradigm (Baguley et al., 2022), we construct 11 templates spanning eight scenario types and three inter-agent relation categories: general interactions, vicarious blame, and command–obedience chains. Dimension values constrained by each role structure are fixed, while the remaining values are systematically varied. To reduce surface shortcuts, we diversify

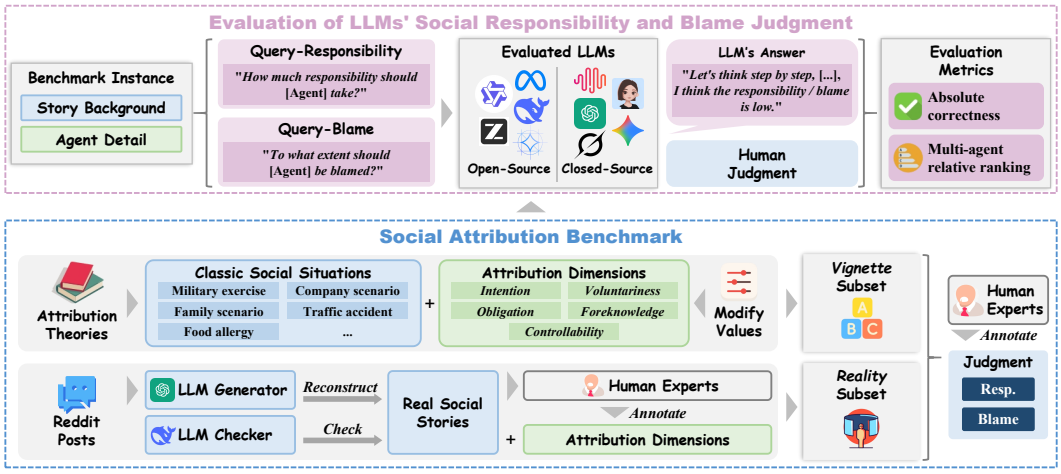



Figure 1: Benchmark construction and evaluation of LLM social attribution.

narrative details and phrasing, randomize names and genders, and avoid explicit dimension labels. Each scenario involves multiple agents, with a separate responsibility question and a separate blame question for each agent. Appendix B.1 details the construction. The *Reality* subset captures realistic and ambiguous social interactions through narratives posted by real users, complementing the controlled *Vignette* settings. Following the story-reconstruction approach of Forbes et al. (2020), we select Reddit AITA posts with clear negative outcomes and reconstruct them into concise narratives. LLM-assisted preprocessing and inspection remove irrelevant or harmful content and identifying information, replacing identities with fictional names. For each scenario, we select a single target agent and ask separate responsibility and blame questions about that agent. Appendices B.2 and B.3 specify the models, processing steps, and prompts. Both subsets are constructed in Chinese and English.

**Data Example.** During an artillery exercise, a general orders a strike through a platoon leader and a soldier. Both candidate buildings, A and B, are abandoned. The general, unaware of workers cleaning inside A, allows the platoon leader to choose either building. The platoon leader knows about the workers but orders a strike on A. The soldier also knows about them. The strike injures the workers.

**Question.** To what extent should the general (platoon leader/soldier) be held responsible (be blamed)? Choose from [high, medium, low, no].

Each question $i$ concerns one agent, scenario, language, and task $m \in \{\textit{responsibility}, \textit{blame}\}$, with judgment $a_{i,m} \in \mathcal{A} = \{\texttt{no}, \texttt{low}, \texttt{medium}, \texttt{high}\}$. The dimension set $\mathcal{K}$ comprises intention, voluntariness, foreknowledge, controllability, and obligation. For each dimension $k \in \mathcal{K}$, $y_{i,k} \in \{0, 1, \varnothing\}$ denotes a binary value or an unspecified/inapplicable label. Dimension values are set during *Vignette* construction and human-annotated for *Reality*. Two psychology experts independently annotate judgments, treating languages and tasks separately and viewing only scenarios and questions, with dimension labels hidden. Before filtering, the annotators achieve 86.8% exact agreement overall and a mean Cohen's $\kappa$ of 0.826 across the two tasks and languages, supporting the consistency of the annotations. We retain only questions for which both annotators assign the same judgment label, yielding 7,639 questions. See Appendices B.4, B.5, B.6, and B.7 for annotation and dataset details.

We designed two metrics to measure the social attribution capabilities of LLMs. Outputs are canonicalized to $\mathcal{A}$ (Appendix C.3). Accuracy is the proportion of valid predictions that exactly match human labels, measuring agreement on the absolute level of responsibility or blame assigned to an agent. Mapping $(\texttt{no}, \texttt{low}, \texttt{medium}, \texttt{high})$ to $(0, 1, 2, 3)$ through $r$, we also compute Spearman correlation between predicted and human agent rankings within each *Vignette* scenario, assigning average ranks to ties. This metric measures agreement on which agents are more responsible or blameworthy relative to others, even when the predicted judgment levels differ from human labels.

We report the mean correlation across scenarios. Invalid predictions are excluded from accuracy, and scenarios containing them are excluded from correlation (Appendix C.4).

### 4.2 Latent Space Dimension Probing

We examine whether attribution dimensions are decodable from LLM hidden states and whether interventions along their representations affect judgments in theory-predicted directions (Figure 2).

#### 4.2.1 Decodability of Attribution Dimensions

For each prompt, we parse the frozen LLM's judgment and locate its label-token prediction step $\mathrm{Idx}_i^\star$. At each probed layer $l$, we extract $z_i^{(l)} = h_{i,\mathrm{Idx}_i^\star}^{(l)} \in \mathbb{R}^d$, the state used to predict that token (Appendix D.1). For each dimension and layer, a logistic probe estimates

$$p_k^{(l)}(y = 1 \mid z) = \sigma\Big(w_k^{(l)\top} z + b_k^{(l)}\Big), \tag{1}$$

where $\sigma$ is the sigmoid. We fit probes on training folds using class-balanced cross-entropy with $L_2$ regularization; training and evaluation require available hidden states and non-missing dimension labels (Appendix D.3).

Probes are trained separately by language and subset to avoid train–test overlap between translations. *Vignette* uses eight-fold leave-one-scenario-type-out cross-validation: each fold tests one type and trains on the other seven, grouping all occurrences of that type across agent relations. This tests decoding on unseen scenario types and reduces performance inflation from shared event backgrounds and templates. *Reality* uses 10-fold random cross-validation (Appendix D.2). Within each language–subset run, we summarize *decodability* by the unweighted mean of fold balanced accuracies, $(\mathrm{Recall}_0 + \mathrm{Recall}_1)/2$. Decodability measures recovery of binary dimension labels; benchmark accuracy concerns responsibility/blame judgments. We report decodability as a percentage, with 50% indicating chance.

#### 4.2.2 Influence of Attribution Dimensions on Judgment

We investigate whether the influence of latent attribution-dimension representations on LLMs' final responsibility and blame judgments is consistent with the patterns of human social attribution described by attribution theory (Shaver, 1985; Mao & Gratch, 2012; Malle et al., 2014). To assess this consistency, we intervene along each dimension's representation and compare the resulting judgment changes with theory-derived directional expectations (Appendix A.3).

First, for each training fold and selected layer $l \in \mathcal{L}_{\mathrm{sel}}$, let $W_{-k}^{(l)}$ collect the weights of other available dimension probes. To reduce shared linear components, we residualize the target weight and normalize the residual:

$$\begin{aligned} \hat{v}_k^{(l)} &= w_k^{(l)} - W_{-k}^{(l)} \Big(W_{-k}^{(l)\top} W_{-k}^{(l)} + \gamma I\Big)^{-1} W_{-k}^{(l)\top} w_k^{(l)}, \\ v_k^{(l)} &= \frac{\hat{v}_k^{(l)}}{\|\hat{v}_k^{(l)}\|_2}. \end{aligned} \tag{2}$$

Here $\gamma > 0$ regularizes residualization. Directions are derived only from training-fold probes; degenerate residuals are skipped. The positive direction increases probe evidence for $y_{i,k} = 1$.

Second, the unedited output $\hat{a}_{i,m}$ fixes the feasible signs for all comparisons: $\mathcal{E}_i = \{+1\}$ for `no`, $\{-1\}$ for `high`, and $\{-1, +1\}$ otherwise. A single-use hook edits its label-prediction step $\mathrm{Idx}_i^\star$ at layer $l$:

$$\tilde{h}_{i,\mathrm{Idx}_i^\star}^{(l)} = h_{i,\mathrm{Idx}_i^\star}^{(l)} + \eta\alpha v_k^{(l)}, \qquad \eta \in \mathcal{E}_i, \quad \alpha > 0. \tag{3}$$

We then generate exactly one token and obtain comparison predictions by restricting the argmax to $\mathcal{A}$. Besides the unedited baseline, we use two intervention controls: a random unit direction independent of labels and probe weights, and a direction obtained from a shuffled-label probe using the same direction-construction procedure. Controls match the layer, prediction step, strength, sign, and perturbation norm, testing whether effects depend on the learned dimension direction.

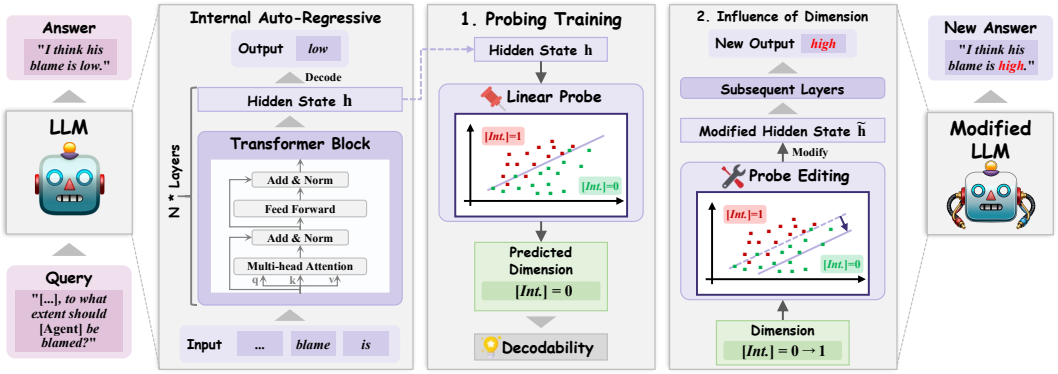



Figure 2: Attribution-dimension probing through layerwise decodability analysis and directional interventions on judgments.

In these negative-outcome scenarios, we hypothesize that strengthening an attribution-dimension representation tends to increase responsibility or blame, while weakening it tends to decrease the judgment. For each setting $(m, l, k, \alpha)$, let $\hat{a}_{i,\eta}$ be the probe-directed prediction and $\hat{a}^c_{i,\eta}$ the comparison prediction, where $c \in \{\text{base}, \text{rand}, \text{shuf}\}$ and $\hat{a}^{\text{base}}_{i,\eta} = \hat{a}_{i,m}$. Using $r$, we compute

$$\Delta^c_{i,\eta} = \eta \left[ r(\hat{a}_{i,\eta}) - r(\hat{a}^c_{i,\eta}) \right]. \tag{4}$$

Positive values indicate theory-aligned shifts relative to $c$: higher judgments under positive interventions or lower judgments under negative ones.

For $N$ held-out questions, define one binary outcome per question and comparison:

$$s^c_i = \mathbb{I}\left[ \min_{\eta \in \mathcal{E}_i} \Delta^c_{i,\eta} > 0 \right], \qquad S_c = \sum_{i=1}^{N} s^c_i. \tag{5}$$

Every feasible sign must strictly align; a tie or opposite shift on any tested sign counts as failure. For success probability $\pi_c$, we test $H_0 : \pi_c \leq 1/2$ against $H_1 : \pi_c > 1/2$ using a one-sided exact binomial test:

$$p_c = \Pr[B \geq S_c] = \sum_{j=S_c}^{N} \binom{N}{j} 2^{-N}, \qquad B \sim \text{Binom}(N, 1/2). \tag{6}$$

We report "Tests passed" if $N > 0$ and $p_c < 0.01$ for all three comparisons, and "Tests not passed" otherwise.

## 5 Experiments

### 5.1 Experimental Setup

We evaluate 32 LLMs: nine closed-source LLMs from the Claude, Doubao Seed, Gemini, GPT, Grok, Kimi, MiMo, and Step families, and 23 open-source LLMs from the Llama, Phi, GPT-oss, DeepSeek-R1, Qwen, Gemma, GLM, and InternLM families. The open-source models span 0.5–120B parameters; Appendix C.1 provides the full model list and inference settings. We evaluate each LLM with four prompt strategies: Direct (answer without reasoning), CoT (reason step by step), Self-reflection (review and revise an initial judgment), and Theory-guided (analyze the five attribution dimensions before judging). For comparison, we evaluate five non-LLM models: Random Stratified (sample labels according to their frequencies), Majority Class (predict the most frequent label), Heuristic Rule (score textual keywords), Attribution-Theory Rule (apply theory-based rules to annotated dimensions), and mBERT Classifier (predict labels with fine-tuned multilingual BERT). Performance is measured by accuracy and Spearman's $\rho$, as defined in Section 4. Prompt templates and non-LLM model details are provided in Appendices C.2 and C.5.

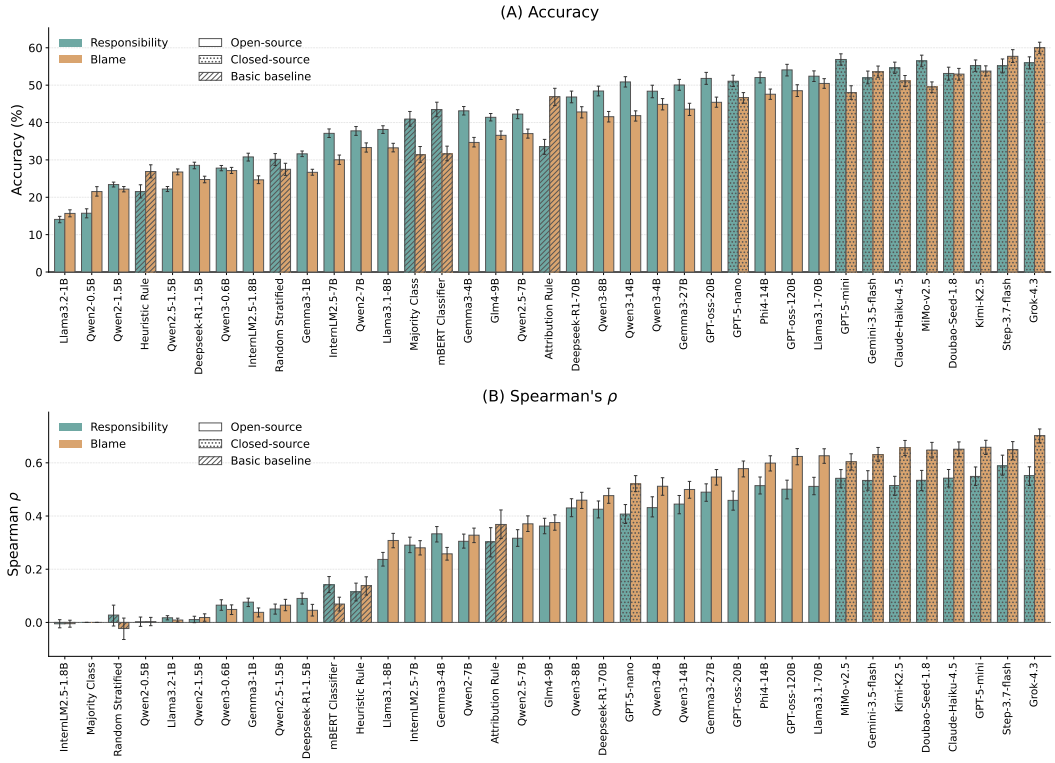



Figure 3: Accuracy and Spearman's $\rho$ across models, with 95% confidence intervals.

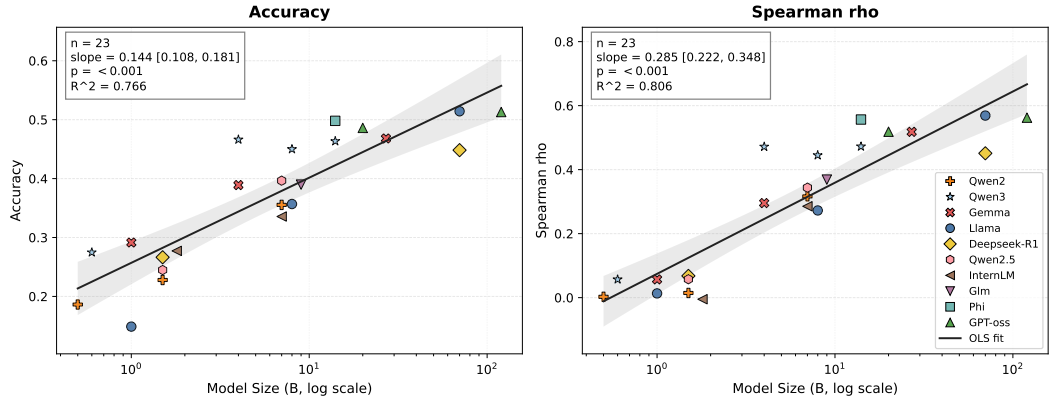



Figure 4: Model-size trends averaged over both tasks, with OLS fits and 95% confidence bands.

To understand how attribution dimensions are represented and influence LLM judgments, we examine their decodability in hidden states and whether their influences on judgments align with human attribution theory. We use DeepSeek-R1-1.5B, Llama-3.2-1B, Llama-3.1-8B, Qwen3-4B, and Qwen3-8B as backbones with CoT prompts, following the methods in Section 4.2. Reported decodability plots recompute balanced accuracy from held-out predictions pooled across languages and subsets (Appendix D.2). Hyperparameters, implementation, and testing details are provided in Appendices D.4 and D.5.

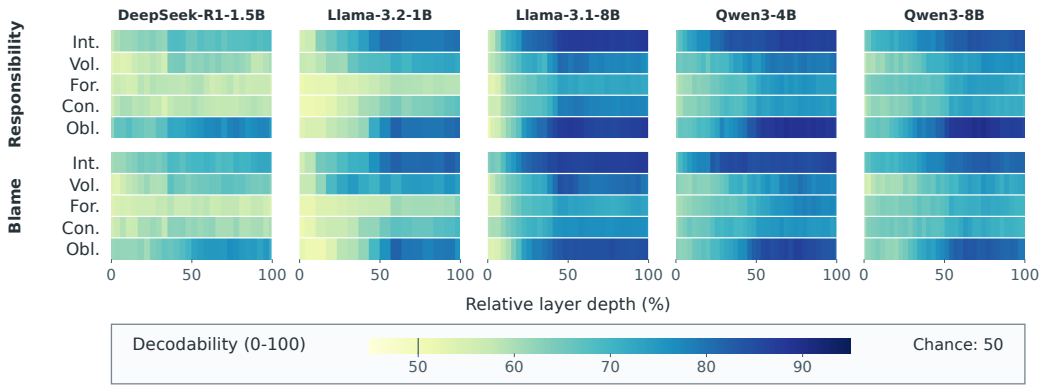



Figure 5: Layerwise dimension decodability. Panels separate models and tasks; rows denote dimensions, and darker colors indicate higher decodability (chance: 50%).

### 5.2 Performance Evaluation on LLM Social Attribution

Figure 3 compares agreement with human judgments across models, averaging each LLM's results over the four prompting strategies. Grok-4.3 achieves the highest scores on both metrics when averaged across responsibility and blame, with 58.00% accuracy and $\rho = 0.627$.

For most models, accuracy is higher on responsibility than on blame, whereas Spearman's $\rho$ shows the opposite pattern. It is because assigning a blame level requires finer judgments of moral severity and mitigating circumstances: models may preserve relative differences in blame among agents while differing from humans on the precise level assigned.

The supervised mBERT classifier achieves accuracy comparable to selected 4–8B LLMs: 43.47% on responsibility, versus 43.13% for Gemma3-4B, and 31.60% on blame, versus 33.23% for Llama3.1-8B (Appendix Table 15). Its limited accuracy may reflect difficulty learning agent-specific roles and obligations from the available supervision.

Across the 23 open-source models, both accuracy and ranking agreement increase with log-scale parameter count (Figure 4). The fitted slopes are positive for both metrics ($p < 0.001$), with $R^2 = 0.766$ for accuracy and $R^2 = 0.806$ for $\rho$. However, differences in model families and training preclude attributing this trend to parameter count alone.

### 5.3 Decodability of Attribution Dimensions in LLM Hidden States

Figure 5 shows the decodability of attribution dimensions at the label-token prediction step across five models, both judgment tasks, and normalized layer depths.

Attribution dimensions are consistently decodable from hidden states, with most middle and later layers scoring above chance. Intention and obligation show the highest decodability, approaching 90% in Llama-3.1-8B and the Qwen models. Foreknowledge is weaker, particularly in DeepSeek-R1-1.5B and Llama-3.2-1B, while voluntariness and controllability vary more across models. Responsibility and blame show similar patterns in dimension strength and layerwise variation.

Decodability generally increases from early to middle layers and remains high across later layers. Some dimensions plateau or decline slightly near the output, with peak decodability often occurring before the final layer. This pattern is consistent with attribution information becoming easier to read out as the model integrates the scenario context.

Sanity checks further support these findings (see Appendix Figures 10–14). Across all five models, the original probes outperform the shuffled-label control, which stays near chance, and generally outperform the 64-dimensional random-projection control. For Llama-3.1-8B, Qwen3-4B, and Qwen3-8B, they also exceed the scenario-text TF-IDF control across most layers.

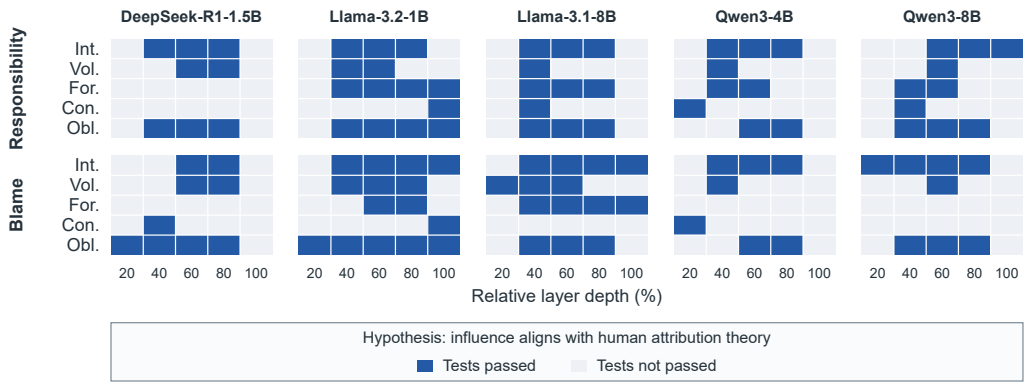



Figure 6: Layerwise influence of attribution dimensions. Blue cells pass all three question-level directional-success tests ($p < 0.01$); gray cells do not. Rows denote dimensions; 100% denotes the final layer.

### 5.4 Influence of Attribution Dimensions on Judgment

Figure 6 shows whether interventions along attribution-dimension representations shift responsibility and blame judgments in the directions predicted by attribution theory. Intention and obligation show the most consistent effects: at 60% and 80% relative depth, both dimensions satisfy the directional tests in all five models and both tasks. These effects hold against the unedited baseline and both intervention controls. This pattern also matches the high decodability of intention and obligation observed in the preceding analysis.

Theory-aligned effects are concentrated in the middle and later layers, particularly between 40% and 80% relative depth. The 60% depth has the broadest support, with 33 of 50 settings passing the tests, compared with six at 20% depth and nine at the final layer. Llama-3.2-1B retains several significant effects at the final layer, while the other models show fewer.

The remaining dimensions show more variation across models and tasks. Foreknowledge produces significant effects on both tasks in the Llama models, while the Qwen models show significant effects only on responsibility. Voluntariness passes the tests in all five models at selected intermediate layers, including both tasks at 60% and 80% depth in DeepSeek-R1-1.5B. Controllability has the fewest significant results, appearing at isolated layers.

### 5.5 Supplementary Experimental Results

The appendices report full LLM and non-LLM results (E.1, E.2), valid-output rates (E.3), scenario- and dimension-level breakdowns (E.4), decodability sanity checks (F.1), earlier-token decodability and influence (F.2, G.1), and intervention-strength sensitivity (G.2).

## 6 Conclusion

To understand social attribution as a fundamental aspect of social reasoning, we present the first systematic investigation of LLM responsibility and blame judgments. Our benchmark comprises a *Vignette* subset of classic attribution scenarios and a *Reality* subset of real-world social narratives. We evaluate 32 LLMs and develop a framework to probe the representations underlying their judgments. Closed-source models generally outperform open-source models, whose agreement with human judgments is positively associated with model size. Across five probed models, attribution dimensions are decodable from hidden states, especially intention and obligation. Interventions on these dimensions produce the most consistent shifts predicted by attribution theory, especially in middle and later layers. Effects of other dimensions vary across models, tasks, and layers. Together, these findings connect LLM social attribution behavior with internal representations of factors identified in human attribution theory. See Appendices H.1, H.2, and H.3 for limitations and responsible use.

## AI USE STATEMENT

For manuscript preparation and code development, LLM assistance was limited to language polishing and the implementation of peripheral utilities, such as plotting and visualization. **The research ideas, mathematical derivations, and implementation of the core algorithms and experimental logic were completed by the human authors without AI assistance.** Separately, LLMs assisted with narrative reconstruction, cleaning, and quality inspection for the *Reality* subset, as documented in Appendices B.2 and B.3; the final responsibility and blame labels were assigned by human experts. The authors take full responsibility for the manuscript, code, data, and reported findings.

## ETHICS STATEMENT

This work studies social attribution in fictional or reconstructed scenarios for research evaluation, and we do not anticipate significant ethical risks within this scope. To protect privacy, the *Reality* subset reconstructs publicly available narratives, removes identifying information and unrelated private details, and replaces identities with fictional names. Preprocessing and quality inspection also filter harmful or biased wording. Annotators were informed of the task and potentially negative content, participated voluntarily, and could withdraw at any time. To discourage inappropriate use, we explicitly frame the benchmark as a tool for studying model behavior; its labels and model outputs are not intended for consequential decisions about real individuals. Further details on these safeguards and responsible use are provided in Appendix H.3.

## REPRODUCIBILITY STATEMENT

We document the experimental procedures and settings in the appendices to support reproducibility. Appendix B describes dataset construction, preprocessing prompts, human annotation guidelines, and filtering. Appendix C provides model specifications, inference settings, evaluation prompt templates, answer extraction rules, metrics, and non-LLM baseline implementations. Appendix D details hidden-state extraction, data splits and random seeds, probe training, intervention implementation and hyperparameters, and statistical tests. Complete benchmark results and additional probing and intervention analyses are reported in Appendices E–G. We will publicly release the complete experimental data and code on GitHub after the paper is accepted, to facilitate independent reproduction and further research.

## REFERENCES

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of LREC-COLING*, pp. 6330–6340, 2024.

Herman Aguinis and Kyle J Bradley. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, 17(4): 351–371, 2014.

Thom Baguley, Grace Dunham, and Oonagh Steer. Statistical modelling of vignette data in psychology. *British Journal of Psychology*, 113(4):1143–1163, 2022.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

Mohna Chakraborty, Lu Wang, and David Jurgens. Structured moral reasoning in language models: A value-grounded evaluation framework. In *Proceedings of EMNLP*, pp. 30295–30323, 2025.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. LAiW: A Chinese legal large language models benchmark. In *Proceedings of COLING*, pp. 10738–10766, 2025.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, et al. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of EMNLP*, pp. 7933–7962, 2024.

Janet Finch. The vignette technique in survey research. *Sociology*, 21(1):105–114, 1987.

Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, 1973.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of EMNLP*, pp. 653–670, 2020.

Mahak Goindani and Jennifer Neville. Social reinforcement learning to combat fake news spread. In *Proceedings of UAI*, pp. 1006–1016, 2019.

Jonathan Gratch, Wenji Mao, and Stacy Marsella. Modeling social emotions and social attributions. *Cognition and Multi-agent Interaction*, pp. 219–251, 2006.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of NeurIPS*, pp. 44123–44279, 2023.

Hilda Hadan, Reza Hadi Mogavi, Leah Zhang-Kennedy, and Lennart E Nacke. Who is responsible when AI fails? Mapping causes, entities, and consequences of AI privacy and ethical incidents. *International Journal of Human–Computer Interaction*, pp. 1–45, 2025.

Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley You Ren, Andrew Miller, Udhyakumar Nallasamy, and Jaya Narain. Do LLMs "know" internally when they follow instructions? In *Proceedings of ICLR*, 2025.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *Proceedings of ICLR*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP*, pp. 2733–2743, 2019.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. Moralbench: Moral evaluation of LLMs. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71, 2025.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. Can machines learn morality? The delphi experiment. *arXiv preprint arXiv:2110.07574*, 2021.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Adauto, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, et al. Language model alignment in multilingual trolley problems. In *Proceedings of ICLR*, 2025.

Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. Probing then editing response personality of large language models. In *Proceedings of COLM*, 2025.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. Are the values of LLMs structurally aligned with humans? A causal perspective. In *Findings of ACL*, pp. 23147–23161, 2025.

Bertram F Malle. Attribution theories: How people make sense of behavior. *Theories in Social Psychology*, pp. 93–120, 2022.

Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. A theory of blame. *Psychological Inquiry*, 25(2):147–186, 2014.

Wenji Mao and Jonathan Gratch. Modeling social causality and responsibility judgment in multi-agent interactions. *Journal of Artificial Intelligence Research*, 44:223–273, 2012.

Wenji Mao, Ansheng Ge, and Xiaochen Li. From causal scenarios to social causality: An attributional approach. *IEEE Intelligent Systems*, 26(6):48–57, 2011.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Proceedings of NeurIPS*, pp. 17359–17372, 2022.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *Proceedings of ICLR*, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*, pp. 1–22, 2023.

Flor Plaza-del Arco, Amanda Curry, Alba Cercas Curry, Gavin Abercrombie, and Dirk Hovy. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of ACL*, pp. 7682–7696, 2024.

Shuhan Qi, Zhengying Cao, Jun Rao, Lei Wang, Jing Xiao, and Xuan Wang. What is the limitation of multimodal LLMs? a deeper look into multimodal LLMs through prompt probing. *Information Processing & Management*, 60(6):103510, 2023. ISSN 0306-4573.

Kelly G Shaver. *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer Science & Business Media, 1985.

Kelly G Shaver and Debra Drown. On causality, responsibility, and self-blame: a theoretical note. *Journal of Personality and Social Psychology*, 50(4):697–702, 1986.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of ACL*, pp. 13090–13120, 2025.

Bernard Weiner. An attributional theory of achievement motivation and emotion. *Psychological Review*, 92(4):548–573, 1985.

Bernard Weiner. *An attributional theory of motivation and emotion*. Springer-Verlag, 1986.

Bernard Weiner. *Judgments of responsibility: A foundation for a theory of social conduct*. Guilford Press, 1995.

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. AI for social science and social science of AI: A survey. *Information Processing & Management*, 61(3):103665, 2024.

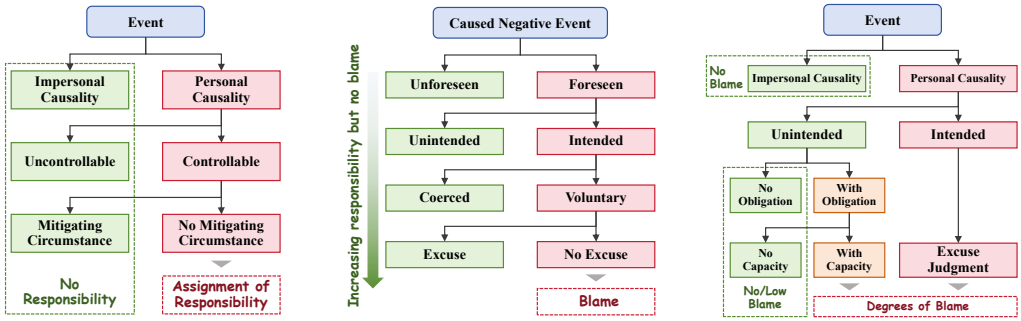
Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. Measuring social norms of large language models. In *Findings of NAACL*, pp. 650–699, 2024.

Tianyi Zhang. Probing political ideology in large language models: How latent political representations generalize across tasks. In *Findings of EMNLP*, pp. 23349–23360, 2025.

# A Theoretical Background and Operationalization

## A.1 Classical Attribution Theories

Attribution theories describe how perceivers connect an agent's actions to an outcome and assess the agent's responsibility and blameworthiness. These judgments are related but distinct: causal involvement does not by itself establish responsibility, and responsibility does not by itself establish blame. Assessments of control, mental states, coercion, and mitigating circumstances help explain why agents involved in the same negative event can receive different judgments (Shaver, 1985; Weiner, 1995). This distinction motivates our separate responsibility and blame tasks, as well as annotations of intermediate attribution dimensions rather than final judgments alone.




(a) Weiner's responsibility model. (b) Shaver's blame model. (c) Malle et al.'s path model.

Figure 7: Classic attribution theories organize responsibility and blame judgments around assessments of causality, mental states, control, and mitigating circumstances (Weiner, 1995; Shaver, 1985; Malle et al., 2014). The models provide complementary theoretical foundations for the dimensions investigated here; they are not assumed to describe the computational architecture of an LLM.

**Weiner's model of responsibility judgment.** Weiner characterizes causal explanations through separable dimensions: the *locus of causality*, distinguishing internal from external causes; *stability*, describing whether a cause persists over time; and *controllability*, concerning whether the agent could have acted otherwise (Weiner, 1985; 1986). In the responsibility model (Figure 7a), perceivers first assess whether an outcome reflects personal or situational causality. When personal causality is implicated, they evaluate controllability and consider mitigating circumstances before assigning responsibility (Weiner, 1995). The model therefore distinguishes an agent's causal contribution from the extent to which the agent can reasonably be held responsible for it.

**Shaver's model of blame assignment.** Shaver describes blame for a negative consequence as a structured assessment of responsibility-relevant variables (Shaver, 1985; Shaver & Drown, 1986). As illustrated in Figure 7b, the assessment begins by distinguishing human agency from impersonal causes. When human agency is involved, subsequent assessments concern whether the agent foresaw the outcome, intended it, and acted voluntarily rather than under coercion. Justifications and excuses can further mitigate blame. These distinctions explain why similar outcomes can support different judgments depending on the agent's knowledge, intentions, and freedom of action. They also motivate treating responsibility and blame as separate targets: mitigating information can reduce blame even when responsibility remains.

**Malle et al.'s Path Model of Blame.** The Path Model of Blame (PMoB) specifies when different information becomes relevant to blame judgments (Malle et al., 2014). After assessing causality, perceivers evaluate intentionality, which separates two subsequent paths (Figure 7c). For intentional conduct, reasons and their justificatory force are central. For unintentional conduct, preventability becomes central, including whether the agent had an obligation and the capacity to prevent the outcome. New information can revise an earlier assessment and redirect the judgment along the other path. This conditional structure is particularly relevant to our inclusion of supervisory obligations

and command–obedience relations: a social role can affect responsibility and blame even when the agent did not directly intend the negative outcome.

### A.2 Operational Definitions of Attribution Dimensions

We select five dimensions that are recurrent in attribution accounts and can be represented explicitly in controlled scenarios or annotated in reconstructed narratives. The operationalizations support comparisons across scenarios while preserving distinctions that would be obscured by a single overall judgment. They do not reproduce every construct or dependency in any one theory. In particular, causality provides part of the scenario context, whereas the five annotated dimensions describe mental states, constraints, capacities, and obligations associated with a target agent.

**Intention: deliberate action rather than intended outcome.** We operationalize intention at the level of the focal action: whether the agent deliberately performed it according to a goal or plan. This is *action-intention*, which need not imply a purpose to bring about the eventual negative outcome. Some attribution accounts use intention in the latter, outcome-directed sense (Shaver, 1985; Shaver & Drown, 1986; Malle et al., 2014). Maintaining the distinction allows the benchmark to vary deliberate action and foreknowledge separately. Their conjunction represents acting while foreseeing a consequence, without equating foresight with a desire to produce that consequence (Mao et al., 2011).

**Voluntariness: constraints on behavioral choice.** Voluntariness concerns whether the agent's choice was substantially constrained by coercion, authority, or situational pressure. An explicit threat is not required: compliance with an authority can also constrain the choice. This dimension therefore separates freely chosen actions from actions performed under orders or pressure (Shaver, 1985; Mao et al., 2011). It is distinct from practical controllability. Following the distinction used in Mao et al. (2011), information about control remains relevant when voluntariness is absent, because constrained choice alone does not specify which alternatives were practically available to the agent.

**Foreknowledge: epistemic access to the consequence.** Foreknowledge concerns whether the agent knew that the action could lead to the eventual outcome or, given the agent's epistemic and cognitive conditions, should have been able to foresee it. The assessment thus depends on information available at the time of acting, rather than knowledge acquired after the event. Foreknowledge can increase attributed responsibility and blame even when the negative outcome was not the agent's goal (Shaver, 1985).

**Controllability: practical capacity to prevent the outcome.** We restrict controllability to the agent's physical or practical ability to prevent the outcome, including relevant means, skills, and opportunities. In PMoB, preventability can also involve the capacity to foresee the outcome (Malle et al., 2014). We represent that epistemic component through foreknowledge, following the analytical separation of knowledge and control in our benchmark. This choice allows probing to distinguish knowing about a risk from having the practical capacity to avert it.

**Obligation: a duty to prevent harm.** Obligation concerns a duty of care, supervision, or prevention arising from the social situation and the agent's role or relationship. It is particularly relevant to unintentional outcomes and omissions, where an agent may bear responsibility or blame despite not performing the action that directly produced the harm (Malle et al., 2014). Unlike broader uses of obligation that also include a subordinate's duty to obey a superior (Mao et al., 2011), our operationalization concerns the duty to prevent the negative outcome. This boundary keeps supervisory responsibility distinct from obedience within a command relationship.

The binary values indicate the presence or absence of each operationalized factor; $\varnothing$ denotes an unspecified or inapplicable value, as defined in the main text. The definitions guide the construction of *Vignette* scenarios and the annotation of *Reality* narratives. The annotation materials and instructions are provided in Appendix B.4.

### A.3 Theoretical Positioning and Directional Expectations

Human attribution theory supplies a principled reference for selecting variables, constructing data, and specifying hypotheses about LLM judgments. It does not require assuming that LLMs instantiate human cognitive processes. This distinction is central to the study design: the benchmark measures agreement with human judgments, the probes measure whether theory-derived dimensions can be recovered from hidden states, and the interventions examine whether directions associated with those dimensions affect outputs in the expected way. Together, these analyses connect behavioral agreement to internal representations while allowing the model's patterns to differ from the theoretical reference.

In the negative-outcome scenarios considered here, intention, voluntariness, foreknowledge, practical control, and obligation provide reasons for assigning greater responsibility or blame when other relevant circumstances are comparable (Shaver, 1985; Weiner, 1995; Mao & Gratch, 2012; Malle et al., 2014). We use this shared directional pattern to formulate an empirical expectation: strengthening the representation associated with a factor tends to increase the judgment, whereas weakening it tends to decrease the judgment. The positive probe direction corresponds to greater evidence for the factor's presence under our operationalization.

These expectations are conditional tendencies, not universal additive rules. The classic models distinguish intentional and unintentional conduct and allow justifications, excuses, and social relationships to change the relevance of a factor. Deliberately performing an action, for example, does not by itself establish an intention to cause harm or settle blameworthiness. Our directional tests therefore assess aggregate consistency with the stated hypothesis within the benchmark and intervention setting; they do not require every individual scenario to follow the same monotonic relation. The inferential limits of this evidence are discussed in Appendix H.2.

# B Dataset Construction and Annotation

The two subsets serve complementary purposes. *Vignette* provides controlled scenarios in which attribution factors can be varied while preserving a coherent social context. *Reality* adds reconstructed everyday conflicts with less explicit and more ambiguous attribution information. Together, they support both evaluation against human judgments and probing of the dimensions underlying those judgments. This section details how the narratives, dimension labels, and judgment labels are obtained, and how agreement-based filtering changes the resulting sample.

## B.1 Vignette Construction

The vignette method uses short hypothetical situations to vary explanatory factors systematically while retaining contextual detail (Finch, 1987; Aguinis & Bradley, 2014; Baguley et al., 2022). This makes it useful for relating LLM judgments to the factors studied in human social attribution. We construct scenarios around negative outcomes using canonical situations from attribution research (Mao et al., 2011; Mao & Gratch, 2012; Malle et al., 2014). Each template specifies agents, their relations, and attribution dimensions constrained by their roles. An *inter-agent relation* describes how agents are socially connected; a *scenario type* describes the event setting. The three relation categories are:

- **General**: agents have no predefined social ties, or their ties do not entail responsibility or blame for another agent's conduct.
- **Vicarious blame**: agents occupy asymmetric social roles, such as parent–child or advisor–student, in which one agent has a supervisory obligation and may be held responsible for another agent's behavior (Malle et al., 2014).
- **Commanding chain**: agents are linked by command–obedience relations (Mao et al., 2011; Mao & Gratch, 2012).

We generate instances by varying the remaining dimension values subject to these role constraints. For example, a parent–child relation introduces a supervisory obligation for the parent without requiring the parent and child to share the same intention. The template therefore constrains which combinations are meaningful; it does not require all five dimensions to vary independently for every agent. Table 1 lists the 11 relation-specific templates, their sources, event descriptions, fixed and variable dimensions, agent counts, and instance counts. These templates cover eight scenario types: Traffic Accident, Food Allergy, Company, Drowning, Laboratory, Family, Gunman, and Military Strike. Company occurs in all three relation categories, and Laboratory occurs in General and Vicarious Blame; each remains one scenario type across relations. These eight types define the held-out groups in the probing evaluation (Appendix D.2).

To reduce the possibility that later probes exploit recurring names or sentence patterns instead of attribution-related information, we manually diversify the semantic details, syntactic structures, and narrative phrasing of the instances. Character names and genders are randomly generated for each instance. The descriptions also avoid trivially revealing the attribution variables through literal labels. These construction choices reduce repeated surface cues; the decodability controls separately examine how much information can still be recovered from surface text (Appendix D.3).

A *scenario instance* is one completed multi-agent narrative; an *agent–scenario pair* selects one of its agents; a *question* additionally specifies a judgment task and language. Construction yields 320 general, 80 vicarious-blame, and 72 commanding-chain scenario instances per language. Querying every agent produces

$$320 \times 3 + 80 \times 2 + 72 \times 3 = 1{,}336$$

agent–scenario pairs per language. Scenarios are first constructed in Chinese and translated into English item by item. Each pair is instantiated into separate responsibility and blame questions; the resulting pre-filtering question counts are given in Appendix B.4.

## B.2 Reality Preprocessing

Controlled vignettes make the attribution factors explicit by design, but cannot capture all the ambiguity of everyday social narratives. The *Reality* subset complements them with interpersonal con-

Table 1: The 11 Vignette templates span eight scenario types and three inter-agent relation categories. Construction settings precede annotation filtering. "Agents Number" gives the number of agents in each scenario, and "#Instances" gives the number of scenario instances per language. Fixed and variable dimensions are subject to agent-specific role constraints.

| Inter-agent Relation | Scenario Type | Source | Event Description | Fixed Dimensions | Variable Dimensions | Agents Number | #Instances |
|---|---|---|---|---|---|---|---|
| General Scenarios | Traffic Accident | Malle et al. (2014) | During a driving license test, the test vehicle collides with and injures a cyclist. | obligation voluntariness controllability | foreknowledge intention | 3 | 64 |
| | Food Allergy | Malle et al. (2014) | Someone orders a dish containing allergens, causing a companion to suffer an allergic reaction and require medical treatment. | | | | 64 |
| | Company Scenario | Mao and Gratch (2012) | An employee's mistake results in the loss of a critical document. | | | | 64 |
| | Drowning Incident | Malle et al. (2014) | A person drowns while surfing on the beach. | | | | 64 |
| | Laboratory Scenario | Malle et al. (2014) | A student's carelessness causes a classmate to be scalded by molten metal. | | | | 64 |
| | Total | | | | | | 320 |
| Vicarious Blame Scenarios | Company Scenario | Mao and Gratch (2012) | When an intern makes a mistake at work, determine the responsibility of the intern himself and his supervisor. | obligation voluntariness | foreknowledge intention controllability | 2 | 32 |
| | Laboratory Scenario | Malle et al. (2014) | When a graduate student makes a mistake during the experiment, determine the responsibility of the student himself and his supervisor. | | | | 32 |
| | Family Scenario | Malle et al. (2014) | When a child makes a mistake, determine the responsibility of the child himself and his parents | | | | 16 |
| | Total | | | | | | 80 |
| Commanding Chain Scenarios | Company Scenario | Mao and Gratch (2012) | When a superior orders a subordinate to carry out a certain action, an accident occurs after the subordinate executes it. | intention obligation voluntariness | foreknowledge controllability | 3 | 24 |
| | Gunman Scenario | Mao et al. (2011) | The superior ordered the gunman under his command to open fire, resulting in a pedestrian being injured. | | | | 24 |
| | Military Strike | Mao et al. (2011) | Some people were accidentally injured by shells during military strikes. | | | | 24 |
| | Total | | | | | | 72 |

flicts drawn from Reddit's AITA community,[1] where users describe conflicts and ask for judgments about the parties involved. These narratives provide agents, actions, constraints, and outcomes in naturally occurring social contexts.

We select posts with clear negative outcomes and apply story reconstruction (Forbes et al., 2020) to preserve their core events while removing irrelevant slang, dialogue, harmful or biased content, personal information, and metadata. Real identities are replaced with fictional names. The inclusion criteria require a self-contained narrative with at least two agents whose actions, roles, or knowledge are relevant to the negative outcome. For each reconstructed scenario, a single target agent is selected for separate responsibility and blame questions.

ChatGPT-5.4-mini[2] performs cleaning and preprocessing. DeepSeek-V4-Pro[3] then checks the reconstructed narratives and filters those containing residual unwanted content. The prompts are provided in Appendix B.3. English instances are translated into Chinese, yielding 864 agent–scenario pairs in each language before judgment annotation. This narrative-quality filtering precedes the separate judgment-agreement filtering in Appendix B.6.

[1] https://www.reddit.com/r/AmItheAsshole/

[2] https://developers.openai.com/api/docs/models/gpt-5.4-mini

[3] https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro

### B.3 Prompts for LLM-Assisted Data Construction

LLMs assist with reconstruction, cleaning, and quality inspection; they do not supply the final responsibility or blame labels. Those judgments are assigned by human experts. The reconstruction prompt takes an AITA post as input and returns the scenario background and agent descriptions. The inspection prompt specifies retention criteria and returns a decision with its reason. Both steps require valid JSON without additional prose.

**Prompt for story reconstruction and cleaning.** We used the following prompt to transform raw Reddit posts into concise benchmark-style social scenarios.

> You are helping construct a dataset for social attribution research. Given a Reddit AITA post, rewrite it as a concise third-person story.
>
> Requirements: (1) Keep only the core plot related to the interpersonal conflict and the negative outcome. (2) Preserve the main agents, their actions, relevant knowledge, constraints, and consequences. (3) Remove irrelevant metadata, voting information, comments, edits, usernames, hyperlinks, and platform-specific expressions. (4) Remove slang, offensive wording, harmful or biased expressions, and private personal information. (5) Replace real or specific personal names with fictional names. (6) Do not add new facts, motives, or outcomes that are not supported by the original post. (7) Write the output in a neutral and clear style.
>
> Return only a valid JSON object. Do not include markdown, comments, explanations, or any text outside the JSON object.
>
> The JSON object must contain the following fields:
>
> - `story_background`: a concise third-person description of the core event and the negative outcome.
> - `character_details`: a list of short descriptions of the main agents. Each item should describe one agent's relevant actions, knowledge, constraints, or role.

**Prompt for quality inspection and filtering.** We used the following prompt to check whether the reconstructed stories satisfied the requirements for inclusion in the *Reality* subset.

> You are checking reconstructed social scenarios for a benchmark on responsibility and blame judgment. Given a reconstructed story, determine whether it should be retained.
>
> A story should be retained only if: (1) it contains a clear negative outcome; (2) it contains at least two agents involved in the event; (3) the actions, roles, or knowledge of the agents are relevant to responsibility or blame judgment; (4) the story is understandable without the original Reddit post; (5) the story does not contain private personal information, platform metadata, irrelevant comments, or harmful and biased expressions; (6) the story does not introduce unsupported facts beyond the original narrative.
>
> Return only a valid JSON object. Do not include markdown, comments, explanations, or any text outside the JSON object.
>
> The JSON object must contain the following fields:
>
> - `decision`: either `Retain` or `Filter`.
> - `reason`: a brief reason for the decision. If the decision is `Retain`, write a short confirmation. If the decision is `Filter`, state the main problem.

### B.4 Human Annotation Protocol and Guidelines

**Annotation tasks and units.** The annotation procedure has two stages. For *Reality*, human annotators assign values to the five attribution dimensions for the target agent in each reconstructed story: intention, voluntariness, foreknowledge, controllability, and obligation. Unlike *Vignette*, these narratives have no pre-designed dimension settings. For both subsets, annotators then assign the final responsibility and blame judgments. Thus, *Vignette* dimension values come from construction, whereas its judgments are human-annotated; *Reality* has human annotations for both.

A judgment-annotation unit is one agent–scenario pair paired with one task in one language. Before judgment annotation and agreement filtering, *Vignette* contains 1,336 pairs per language and *Reality* contains 864. Instantiating both languages and tasks yields $1{,}336 \times 2 \times 2 = 5{,}344$ *Vignette* questions and $864 \times 2 \times 2 = 3{,}456$ *Reality* questions. Chinese and English versions are treated as separate samples, and responsibility and blame are annotated independently using `no`, `low`, `medium`, and

`high`. Separating the tasks allows the reference judgments to reflect the distinction between bearing responsibility for an outcome and deserving blame for it.

**Annotators and blinding.** We recruit two social-psychology experts of different genders, both fluent in English and Chinese. They are recruited independently, do not know each other before the task, and do not participate in the initial dataset design. They are not informed of the research hypotheses. For each item, annotators see the finalized scenario background, agent descriptions, target agent, question, and label options. For judgment annotation, the question identifies whether responsibility or blame is being assessed.

Annotators are not shown the construction scaffold, including template identifiers, scenario categories, designed variable settings, or expected rankings among agents. For judgment annotation, attribution-dimension labels are hidden. They also do not see model outputs, the other annotator's labels, or raw and intermediate narrative versions, including the original Reddit posts. This separation is intended to make the reference judgments depend on the same finalized narrative available to the evaluated models, rather than on the authors' intended dimension manipulations or expected model behavior.

**Training, calibration, and independent annotation.** Before formal annotation, both experts receive the same guideline, including dimension definitions, the four-level judgment scales, and examples of common boundary cases. They first read it independently, then complete a pilot set drawn from different scenario and task types. The authors collect their questions and clarify label boundaries in the guideline. Calibration establishes a shared interpretation of the scales; it is not used to direct annotators toward an expected judgment.

After calibration, the experts annotate the formal dataset independently without discussing individual samples. For responsibility and blame judgments, no consensus label is forced when they disagree. Agreement is computed on their independent labels before filtering, and only matching judgments are retained in the final benchmark. Appendices B.5 and B.6 report agreement and the resulting selection, respectively.

**Annotation guidelines.** The instructions below specify the criteria presented to annotators. Dimension labels use $\{1, 0, \varnothing\}$, with the supplied *Reality* instructions reserving $\varnothing$ for information that cannot be determined from the narrative. Final judgments use a separate four-level ordinal scale.

**Instructions for attribution-dimension annotation in the *Reality* subset.**

> You will read a reconstructed real-world social scenario. For each target agent, please annotate the following five attribution dimensions: *intention*, *voluntariness*, *foreknowledge*, *controllability*, and *obligation*.
>
> For each dimension, please choose one value from $\{1, 0, \varnothing\}$:
>
> - Use 1 if the dimension is present for the target agent.
> - Use 0 if the dimension is clearly absent for the target agent.
> - Use $\varnothing$ if the story does not provide enough information to determine the value.
>
> Please judge each dimension only based on the information explicitly stated or reasonably implied in the story. Do not infer hidden facts, motives, or background information that are not supported by the text.
>
> The five dimensions are defined as follows:
>
> - *Intention*: Judge whether the target agent deliberately performed the focal action. This refers to the intention behind the action itself, not necessarily the intention to cause the final negative outcome.
> - *Voluntariness*: Judge whether the target agent acted voluntarily, rather than under coercion, strong pressure, or direct command.
> - *Foreknowledge*: Judge whether the target agent knew, or should reasonably have been able to foresee, that the action could lead to the negative outcome.
> - *Controllability*: Judge whether the target agent had the practical ability, opportunity, or means to prevent or avoid the negative outcome.
> - *Obligation*: Judge whether the target agent had a role-based, relational, or situational duty to prevent the negative outcome.

**Instructions for responsibility annotation.**

> You will read a social scenario, the description of a target agent, and a question asking how much responsibility the target agent should bear for the negative outcome. Please choose one label from {`no`, `low`, `medium`, `high`}.
>
> When making the judgment, please consider whether the target agent contributed to the negative outcome, whether the agent could have acted otherwise, whether the agent had relevant knowledge or control, and whether the agent had a relevant role or duty in the event.
>
> The four labels are defined as follows:
>
> - `high`: Choose this label if the target agent is a central source of the negative outcome and should clearly bear strong responsibility.
> - `medium`: Choose this label if the target agent makes a substantial contribution to the negative outcome, but the degree of responsibility is limited by mitigating factors such as limited knowledge, limited control, or role constraints.
> - `low`: Choose this label if the target agent is only weakly or indirectly connected to the negative outcome, or if there are strong mitigating factors.
> - `no`: Choose this label if the target agent should not be held responsible based on the given scenario.

**Instructions for blame annotation.**

> You will read a social scenario, the description of a target agent, and a question asking how much the target agent should be blamed for the negative outcome. Please choose one label from {`no`, `low`, `medium`, `high`}.
>
> When making the judgment, please consider not only whether the target agent contributed to the outcome, but also whether the agent acted intentionally, knowingly, voluntarily, or in violation of a reasonable expectation or obligation. Please also consider whether there are excuses or mitigating circumstances.
>
> The four labels are defined as follows:
>
> - `high`: Choose this label if the target agent clearly deserves strong blame for the negative outcome.
> - `medium`: Choose this label if the target agent deserves a moderate degree of blame, but the blame is reduced by mitigating factors.
> - `low`: Choose this label if the target agent deserves only slight blame because the action, knowledge, control, or obligation is limited.
> - `no`: Choose this label if the target agent should not be blamed based on the given scenario.
>
> Please note that responsibility and blame are related but not identical. A person may be responsible for an outcome but receive lower blame if they lacked intention, foreknowledge, voluntariness, or sufficient control. Please make the responsibility and blame judgments separately according to the specific question.

### B.5 Inter-Annotator Agreement

We measure agreement on responsibility and blame judgments before agreement-based filtering. Exact agreement across all 8,800 questions is 86.8%. We also report Cohen's $\kappa$ (Cohen, 1960), linearly weighted $\kappa_{\text{lin}}$ (Cohen, 1968), and quadratically weighted $\kappa_{\text{quad}}$ (Fleiss & Cohen, 1973). Unweighted $\kappa$ adjusts exact agreement for agreement expected from the annotators' marginal label distributions. The weighted measures additionally account for the ordered labels, with quadratic weighting penalizing larger disagreements more heavily. Reporting all three distinguishes disagreement about an exact level from disagreement across more distant judgment levels.

Table 2 reports agreement by relation, subset, task, and language. The mean unweighted $\kappa$ across the four overall task–language combinations is 0.826. All reported coefficients are positive and significantly different from zero under two-sided tests of $H_0 : \kappa = 0$ ($p < 0.001$). Agreement is lower for several vicarious-blame and *Reality* conditions than for general vignettes, consistent with the greater judgment ambiguity in those settings. These results support using the expert judgments as an evaluation reference while retaining the distinction between reliable agreement and universally shared judgments. Only questions with matching judgment labels enter the final dataset.

Table 2: Inter-annotator agreement ($\times 100$). Resp. denotes responsibility. All coefficients are significantly different from zero ($p < 0.001$).

| Type | Task | $\kappa$ | | $\kappa_{\text{lin}}$ | | $\kappa_{\text{quad}}$ | |
|---|---|---|---|---|---|---|---|
| | | EN | ZH | EN | ZH | EN | ZH |
| General | Resp. | 90.95 | 91.53 | 93.89 | 94.38 | 96.50 | 96.88 |
| | Blame | 84.72 | 85.42 | 90.52 | 91.03 | 95.01 | 95.35 |
| Vicarious blame | Resp. | 77.18 | 81.69 | 79.85 | 83.63 | 83.67 | 86.50 |
| | Blame | 74.49 | 78.88 | 81.93 | 85.51 | 88.84 | 91.57 |
| Commanding chain | Resp. | 80.86 | 84.86 | 85.95 | 88.85 | 90.97 | 92.76 |
| | Blame | 83.53 | 86.02 | 90.14 | 91.63 | 95.03 | 95.78 |
| Reality | Resp. | 75.14 | 74.65 | 84.22 | 84.06 | 91.70 | 91.77 |
| | Blame | 76.42 | 75.89 | 86.11 | 85.93 | 93.12 | 93.17 |
| Overall | Resp. | 83.33 | 84.06 | 88.87 | 89.44 | 93.75 | 94.17 |
| | Blame | 81.13 | 81.80 | 88.59 | 89.11 | 94.15 | 94.53 |

### B.6 Agreement-Based Filtering and Distribution Changes

Agreement-based filtering removes questions for which the two experts assign different responsibility or blame labels. Its purpose is to provide a common reference label for exact-match evaluation without forcing agreement on ambiguous cases. Because this selection could alter the evaluated sample, we examine its effect on subset composition, relation categories, dimension values, and judgment-label frequencies. All counts in this subsection refer to task-specific questions and pool English and Chinese versions.

Table 3 gives the retained and removed counts. Table 4 compares the three *Vignette* relation categories, and Table 5 compares dimension values among questions with available labels. Table 6 reports the before/after judgment-label distributions. These comparisons complement the agreement coefficients: agreement measures annotation consistency, whereas the distribution checks show which parts of the sample remain after applying the retention rule.

Filtering removes 1,161 of 8,800 questions. The *Vignette* share increases from 60.7% to 62.5%, reflecting the larger proportion of disagreements removed from *Reality*. The three vignette relation shares change by at most 1.1 percentage points, and the dimension-value shares change by at most 2.4 percentage points in the rounded tables. The broad composition is therefore similar before and after filtering, although judgment-label frequencies do change. In particular, the retained sample represents expert-agreed cases; these marginal distribution checks do not recover the excluded judgment disagreements.

Table 3: Question counts before and after agreement-based filtering. Percentages give each subset's share within the Before or After column; both tasks and languages are pooled.

| **Subset** | **Before** | **After** | **Removed** |
|---|---|---|---|
| Vignette | 5,344 (60.7%) | 4,776 (62.5%) | 568 |
| Reality | 3,456 (39.3%) | 2,863 (37.5%) | 593 |
| Total | 8,800 | 7,639 | 1,161 |

Table 4: Distribution of the three *Vignette* relation categories before and after filtering. Percentages use the *Vignette* total at the corresponding stage.

| **Agent relation** | **Before** | **After** |
|---|---|---|
| General Scenarios | 3,840 (71.9%) | 3,488 (73.0%) |
| Vicarious Blame Scenarios | 640 (12.0%) | 541 (11.3%) |
| Commanding Chain Scenarios | 864 (16.2%) | 747 (15.7%) |

Table 5: Attribution-dimension distributions before and after filtering. Percentages use questions with non-missing labels for each dimension at the corresponding stage.

| **Dimension** | **Before** | | **After** | |
|---|---|---|---|---|
| | **Value = 1** | **Value = 0** | **Value = 1** | **Value = 0** |
| Intention | 6,097 (69.3%) | 2,703 (30.7%) | 5,405 (70.8%) | 2,234 (29.2%) |
| Voluntariness | 3,164 (67.7%) | 1,509 (32.3%) | 2,845 (69.2%) | 1,267 (30.8%) |
| Foreknowledge | 4,572 (52.0%) | 4,228 (48.0%) | 3,935 (51.5%) | 3,704 (48.5%) |
| Controllability | 5,462 (62.1%) | 3,338 (37.9%) | 4,709 (61.6%) | 2,930 (38.4%) |
| Obligation | 3,087 (50.8%) | 2,993 (49.2%) | 2,740 (53.2%) | 2,413 (46.8%) |

Table 6: Judgment-label distributions before and after agreement-based filtering, pooled across English and Chinese. Percentages use the row total.

| **Subset** | **Task** | **Stage** | **High** | **Medium** | **Low** | **No** | **Total** |
|---|---|---|---|---|---|---|---|
| Vignette | Responsibility | Before | 961 (36.0%) | 919 (34.4%) | 598 (22.4%) | 194 (7.3%) | 2,672 |
| | | After | 921 (37.8%) | 791 (32.4%) | 539 (22.1%) | 188 (7.7%) | 2,439 |
| | Blame | Before | 583 (21.8%) | 733 (27.4%) | 744 (27.8%) | 612 (22.9%) | 2,672 |
| | | After | 556 (23.8%) | 550 (23.5%) | 639 (27.3%) | 592 (25.3%) | 2,337 |
| Reality | Responsibility | Before | 815 (47.2%) | 528 (30.6%) | 189 (10.9%) | 196 (11.3%) | 1,728 |
| | | After | 730 (51.2%) | 369 (25.9%) | 146 (10.2%) | 181 (12.7%) | 1,426 |
| | Blame | Before | 684 (39.6%) | 496 (28.7%) | 232 (13.4%) | 316 (18.3%) | 1,728 |
| | | After | 634 (44.1%) | 321 (22.3%) | 183 (12.7%) | 299 (20.8%) | 1,437 |

### B.7 Final Dataset Statistics

The final dataset contains 7,639 questions: 4,776 from *Vignette* and 2,863 from *Reality*, with 3,837 English and 3,802 Chinese questions. Table 7 reports language and judgment distributions. Table 8 reports dimension-value distributions, excluding unspecified or inapplicable labels. Table 9 reports the retained *Vignette* questions by agent relation and scenario type. The lower denominators for voluntariness and obligation reflect unspecified or inapplicable labels; these are excluded from the corresponding dimension percentages rather than counted as zero. Figures 8 and 9 visualize frequent content words in the English and Chinese *Reality* narratives, respectively.

Table 7: Language and judgment distributions. Each cell reports a count and its percentage within the row. Resp. denotes responsibility.

| Subset | Task | Language | | Judgment | | | |
|---|---|---|---|---|---|---|---|
| | | EN | ZH | High | Medium | Low | No |
| Vignette | Resp. | 1,226 (50.3%) | 1,213 (49.7%) | 921 (37.8%) | 791 (32.4%) | 539 (22.1%) | 188 (7.7%) |
| Vignette | Blame | 1,171 (50.1%) | 1,166 (49.9%) | 556 (23.8%) | 550 (23.5%) | 639 (27.3%) | 592 (25.3%) |
| Reality | Resp. | 720 (50.5%) | 706 (49.5%) | 730 (51.2%) | 369 (25.9%) | 146 (10.2%) | 181 (12.7%) |
| Reality | Blame | 720 (50.1%) | 717 (49.9%) | 634 (44.1%) | 321 (22.3%) | 183 (12.7%) | 299 (20.8%) |

Table 8: Attribution-dimension distributions. Percentages are calculated within each dimension using only questions with non-missing labels.

| Dimension | Value 1 | Value 0 |
|---|---|---|
| Intention | 5,405 (70.8%) | 2,234 (29.2%) |
| Voluntariness | 2,845 (69.2%) | 1,267 (30.8%) |
| Foreknowledge | 3,935 (51.5%) | 3,704 (48.5%) |
| Controllability | 4,709 (61.6%) | 2,930 (38.4%) |
| Obligation | 2,740 (53.2%) | 2,413 (46.8%) |

Table 9: Retained *Vignette* questions by agent relation and scenario type. The 11 rows cover eight distinct scenario types, with shared types repeated under their respective relations. Percentages use all 4,776 *Vignette* questions as the denominator.

| Agent Relation | Scenario Type | Count | Percentage |
|---|---|---|---|
| General | Traffic Accident | 755 | 15.8% |
| General | Food Allergy | 744 | 15.6% |
| General | Company Scenario | 759 | 15.9% |
| General | Drowning Incident | 764 | 16.0% |
| General | Laboratory Scenario | 466 | 9.8% |
| Vicarious Blame | Company Scenario | 248 | 5.2% |
| Vicarious Blame | Laboratory Scenario | 179 | 3.7% |
| Vicarious Blame | Family Scenario | 114 | 2.4% |
| Commanding Chain | Company Scenario | 252 | 5.3% |
| Commanding Chain | Gunman Scenario | 257 | 5.4% |
| Commanding Chain | Military Strike | 238 | 5.0% |

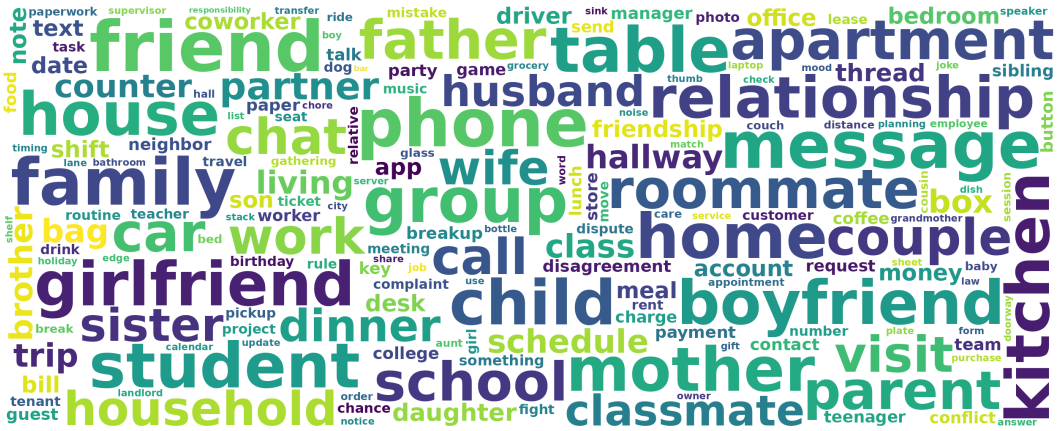



Figure 8: High-frequency content words in the English *Reality* narratives.


朋友
短信
聊天
孩子
女朋友
司机
妻子
手机
关系
客厅
室友
厨房
伴侣
丈夫
信息
母亲
男朋友
学校
沙发
旅行
走廊
聚会
同学
餐桌
工作
晚餐
父母
消息
屏幕
公寓
桌子
朋友
账单
学生
房子
女友
经理
电话
父亲
同事
家庭
回家
青少年
员工
办公室
约会
老师
儿子
笔记本电脑
卧室
钱
客人
租约
日程
教室
食物
小组
女儿
照片
活动
家务
见面
游戏
检查
公司
机会
通话
情侣
队伍
咖啡
家庭成员
标签
路线
买家
影响
时刻
女孩
兄弟姐妹
团队
主管
零食
袋子
课程
音乐
座位
顾客
表格
编辑
建议
弟弟
水槽
笔
生活
设备
角色
按钮
开口
台面
压力
细节
文件夹
亲戚
私下
设置
社交
噪音
名字
姐妹
变化
女士
婚礼
过程
饮料
班次
前门
住户
电脑
费用
礼物
前任
车辆
号码
宿舍
异地
房东
男孩
男友
个人
比赛
盒子
商店
会议
洗衣
裙子
项目
厨师
主人
房租
表格
无效
无效
双手
行程
车库
车道
车长
感觉
感染
年龄
年龄
年龄

# C BENCHMARK EVALUATION PROTOCOL

The behavioral evaluation measures agreement with human responsibility and blame judgments at two complementary levels: the label assigned to an individual agent and the relative ordering of agents within a scenario. The protocol below specifies the evaluated models, elicitation strategies, answer extraction, metrics, and comparison methods. Complete results are reported in Appendix E.

## C.1 EVALUATED MODELS AND INFERENCE SETTINGS

We evaluate 32 LLMs, comprising nine models accessed through provider APIs and 23 locally deployed open-source models spanning 0.5–120B parameters. Table 10 lists the model families, developers, versions, parameter scales, and links to their documentation. The "closed-source" grouping used in the results follows the access mode in these experiments: Kimi, MiMo, and Step are included in this group because we used their official APIs, although their weights have also been publicly released. The open-source group consists of instruction-tuned or chat models. For the behavioral benchmark, these models are deployed through Ollama[4], with temperature set to 0 and random seed set to 0. The hidden-state experiments use the separate backbone configuration described in Appendix D.

Each model is evaluated on both languages and both judgment tasks under four prompting strategies. Direct prompting elicits a judgment without an explicit reasoning trace; CoT requests a step-by-step analysis; Self-reflection asks the model to review an initial judgment; and Theory-guided prompting explicitly introduces the five attribution dimensions. Comparing these strategies examines whether additional reasoning instructions, revision, or access to the theoretical scaffold changes agreement with human judgments. The Theory-guided prompt supplies dimension names, not the human-annotated dimension values.

Table 10: List of LLMs under test and their detailed information.

| LLM Family | Developer | Open/Closed-source | Version | Size |
|---|---|---|---|---|
| Claude | Anthropic | Closed | `Claude-Haiku-4.5` | - |
| Doubao Seed | ByteDance | Closed | `Doubao-Seed-1.8` | - |
| Gemini | Google | Closed | `Gemini-3.5-flash` | - |
| ChatGPT | OpenAI | Closed | `GPT-5-mini/GPT-5-nano` | - |
| Grok | xAI | Closed | `Grok-4.3` | - |
| Kimi | Moonshot AI | Closed | `Kimi-K2.5` | - |
| MiMo | Xiaomi | Closed | `MiMo-v2.5` | - |
| Step | StepFun | Closed | `Step-3.7-flash` | - |
| Llama | Meta | Open | `Llama3.1/Llama3.2` | 1B/8B/70B |
| Phi | Microsoft | Open | `Phi4` | 14B |
| GPT-oss | OpenAI | Open | `GPT-oss` | 20B/120B |
| DeepSeek-R1 | DeepSeek | Open | `DeepSeek-R1` | 1.5B/70B |
| Qwen | Alibaba | Open | `Qwen2/Qwen2.5/Qwen3` | 0.5B/0.6B/1.5B/4B/7B/8B/14B |
| Gemma | Google | Open | `Gemma3` | 1B/4B/27B |
| GLM | Zhipu AI | Open | `GLM4` | 9B |
| InternLM | Shanghai AI Laboratory | Open | `InternLM2.5` | 1.8B/7B |

## C.2 PROMPT TEMPLATES

The English templates for the four strategies—Direct, CoT, Self-reflection, and Theory-guided—are reproduced below. For each prompt, `[SCENARIO]` is replaced with the scenario text, `[QUESTION]` is replaced with the task-specific question, and `[AGENT]` is replaced with the target character. The output label must be one of `high`, `medium`, `low`, or `no`.

**Task questions.**

> **Responsibility question:**
> In the scenario described above, to what extent should `[AGENT]` be held **responsible**? Please choose one answer from [high, medium, low, no].

[4] `https://ollama.com/`

**Blame question:**
In the scenario described above, to what extent should `[AGENT]` be **blamed**? Please choose one answer from [high, medium, low, no].

**Direct prompt.**

`[SCENARIO]`
`[QUESTION]`
Answer directly. Do not provide reasoning, analysis, explanation, or intermediate steps. Output only the final answer in `\boxed{}`. The answer inside `\boxed{}` must be exactly one of: high, medium, low, no.

**CoT prompt.**

`[SCENARIO]`
`[QUESTION]`
Please analyze the scenario step by step before answering. After your analysis, put your final answer directly in `\boxed{}`. The answer inside `\boxed{}` must be exactly one of: high, medium, low, no.

**Self-reflection prompt.**

`[SCENARIO]`
`[QUESTION]`
First give an initial answer with a brief reason, but do not put the initial answer in `\boxed{}`. Then reflect on whether your initial answer missed any important facts or should be revised. After the reflection, put only the final updated answer directly in `\boxed{}`. The answer inside `\boxed{}` must be exactly one of: high, medium, low, no.

**Theory-guided prompt.**

`[SCENARIO]`
`[QUESTION]`
Before making the final judgment, analyze the case using these five dimensions: intention, voluntariness, foreknowledge, controllability, and obligation. Then decide the final answer. Put your final answer directly in `\boxed{}`. The answer inside `\boxed{}` must be exactly one of: high, medium, low, no.

### C.3 Answer Extraction and Invalid Outputs

We use a rule-based parser to map generated answers to $\mathcal{A} = \{\texttt{no}, \texttt{low}, \texttt{medium}, \texttt{high}\}$. The parser first searches for all case-insensitive `\boxed{}` spans and scans them from last to first. It returns the first span that can be mapped to a canonical label. This order follows the requested output format and prevents an initial answer in a reasoning or self-reflection trace from taking precedence over a later final answer. If no valid boxed answer is found, the parser falls back to the last standalone occurrence of a canonical label word in the complete output. An output from which no valid label can be extracted is marked invalid.

Accuracy excludes individual questions with invalid predictions. For the within-scenario ranking metric, any scenario containing an invalid prediction is excluded. Consequently, valid-output coverage and agreement conditional on validity answer different questions: a high agreement score does not by itself establish that a model reliably produces an evaluable answer. Appendix E.3 reports valid-output rates alongside the complete benchmark results.

### C.4 Evaluation Metrics and Statistical Reporting

**Exact agreement.** Let $\mathcal{I}_m$ contain the questions with valid predictions for task $m$. Accuracy is

$$\mathrm{Acc}_m = \frac{1}{|\mathcal{I}_m|} \sum_{i \in \mathcal{I}_m} \mathbb{1}[\hat{a}_{i,m} = a_{i,m}]. \tag{7}$$

It measures agreement on the absolute responsibility or blame level assigned to the target agent. We report it as a percentage, rounded to two decimal places in the full result tables.

**Within-scenario ranking agreement.** For the multi-agent *Vignette* subset, map the labels to ordinal scores with $r(\texttt{no}) = 0$, $r(\texttt{low}) = 1$, $r(\texttt{medium}) = 2$, and $r(\texttt{high}) = 3$. For scenario $s$, let $\mathcal{I}_{m,s}$ index its agent questions and write

$$u_{m,s} = (r(a_{i,m}))_{i \in \mathcal{I}_{m,s}}, \qquad \hat{u}_{m,s} = (r(\hat{a}_{i,m}))_{i \in \mathcal{I}_{m,s}}. \tag{8}$$

The Spearman correlation is

$$\rho_{m,s} = \mathrm{Corr}(\mathrm{Rank}(u_{m,s}), \mathrm{Rank}(\hat{u}_{m,s}))\,, \tag{9}$$

with average ranks assigned to tied values. We report the mean within-scenario correlation on its original scale, rounded to three decimal places in the full tables. This metric can distinguish a model that shifts all judgment levels but preserves the ordering of agents from one that assigns responsibility or blame to the wrong agents. It is restricted to *Vignette* because each *Reality* scenario has only one target agent.

**Uncertainty and aggregation.** The reported 95% confidence intervals use 1,000 cluster-bootstrap resamples over scenario groups. Scenario-level resampling accounts for the fact that several agent questions can concern the same event. The full result tables separate the two tasks and the four prompt strategies. The main model-comparison figure averages each model's scores over the four strategies, while the model-size analysis additionally averages over responsibility and blame. Thus, the per-strategy tables retain variation that is hidden by the aggregate comparison.

**Model-size analysis.** For the 23 open-source models with identifiable parameter scales, we fit a separate ordinary least-squares regression for each metric, using the task-averaged agreement score as the response and $\log_{10}$ parameter count as the predictor. The main figure reports fitted lines and 95% confidence bands; the text reports the slope significance and $R^2$. These regressions summarize an association across the evaluated models. They do not isolate parameter count from model family, training data, architecture, or post-training differences.

### C.5 Non-LLM Baselines

The five non-LLM baselines provide complementary reference points: label-frequency information, surface textual cues, an explicit attribution-theory rule, and supervised text classification. Each predicts one of `no`, `low`, `medium`, or `high` for responsibility and blame and is evaluated with the metrics in Appendix C.4. Their information sources differ: in particular, the Attribution-Theory Rule receives annotated attribution dimensions, whereas the text-based methods must work from the narrative. The results should be interpreted with this distinction in mind.

For each task–subset combination, mBERT uses five random splits with 90% of the data for training and 10% for testing. For a fair comparison among the non-LLM baselines, the other four methods use the same test sets, and accuracy is computed from the pooled test predictions across the five runs.

**Random Stratified.** This baseline samples a label at random according to the empirical label distribution of the corresponding task and dataset subset. It does not use the scenario text, attribution dimensions, or any learned parameters. It measures how much performance can be obtained from label balance alone.

**Majority Class.** This baseline always predicts the most frequent label in the corresponding task and dataset subset. Ties are broken by the fixed label order `no`, `low`, `medium`, `high`. It provides a reference for the contribution of the dominant class to exact agreement.

**Heuristic Rule.** This baseline uses only surface cues from the scenario text and does not use the annotated attribution dimensions. For each record, it first focuses on the sentence or paragraph mentioning the target agent when possible, and otherwise uses the full scenario. It then matches English and Chinese keywords related to causal involvement, intention, foreknowledge, control, obligation,

coercion, inability, accident, justification, and lack of knowledge. Responsibility is scored mainly from causal involvement, control, obligation, foreknowledge, and intention, with penalties for coercion and mitigation. Blame gives more weight to intention and foreknowledge, and applies stronger penalties for coercion, accident, justification, and lack of knowledge. The final score is mapped to the four labels by fixed thresholds: scores at or below zero give `no`, one gives `low`, two or three give `medium`, and four or above gives `high`. This baseline tests how much of the benchmark can be handled through explicit surface matching.

**Attribution-Theory Rule.** This baseline uses the annotated attribution dimensions rather than the raw scenario text. It follows a Mao-and-Gratch-inspired rule over five binary dimensions (Mao & Gratch, 2012): intention, obligation, voluntariness, controllability, and foreknowledge. Missing or not-applicable dimension values are treated as zero. The responsibility rule uses controllability as a key prerequisite, reduces judgments for non-voluntary action, and assigns higher levels in conjunction with intention, foreknowledge, and obligation. The blame rule uses the same dimensions but is more sensitive to lack of voluntariness and lack of foreknowledge, because these factors provide stronger mitigation for blame. This comparison assesses the utility of explicitly supplied attribution factors, while also giving the rule a different input advantage from the LLMs and text classifiers.

**mBERT Classifier.** This baseline is a supervised, non-generative encoder model. We fine-tune bert-base-multilingual-cased[5] with a four-way classification head, using the scenario text as input and the gold attribution label as the target. Responsibility and blame are trained as separate classification tasks. Unlike the frequency and hand-written rule baselines, mBERT tests whether supervised learning from the available text–judgment pairs is sufficient to reproduce the annotated judgments.

Complete baseline results and their comparison with the LLMs are given in Appendix E, Table 15.

[5] `https://huggingface.co/google-bert/bert-base-multilingual-cased`

# D PROBING AND INTERVENTION DETAILS

The representation experiments test decodability and theory-aligned influence separately. The implementation below specifies the controls and token-position comparisons; Appendices F and G report their results.

## D.1 BACKBONES AND HIDDEN-STATE EXTRACTION

We use five frozen backbones with CoT prompts: DeepSeek-R1-1.5B (28 Transformer layers), Llama-3.2-1B (16), Llama-3.1-8B (32), Qwen3-4B (36), and Qwen3-8B (36). The main decodability analysis probes every Transformer layer. For cross-model comparison, layer depth is normalized by each model's total number of layers; 100% denotes the final layer. The selected intervention layers are given in Table 11.

For each baseline response, we parse the final judgment using the canonical labels in Appendix C.3 and locate its label-token prediction step $\mathrm{Idx}_i^\star$. At layer $l$, $z_i^{(l)} = h_{i,\mathrm{Idx}_i^\star}^{(l)}$ denotes the state used to predict the label token, as defined in Section 4.2.1. This position is chosen because it follows the available scenario context and the model's generated reasoning and lies at the point where that information is converted into a judgment output. Throughout these appendices, "label-token prediction step" refers to this prediction position, rather than a state obtained after feeding the generated label back into the model.

The last input token is a distinct, earlier position at the end of the prompt, before the generated CoT response. Probing this position tests whether dimension information is already accessible before the final judgment stage. Intervening there tests whether a local perturbation can propagate through subsequent autoregressive generation to affect the final answer. These complementary comparisons are reported in Appendices F.2 and G.1.

Table 11: Backbones and intervention layers. Layer numbers use the one-based convention $l \in \{1, \ldots, L\}$. The four intermediate positions are $\lfloor 0.2L \rfloor$, $\lfloor 0.4L \rfloor$, $\lfloor 0.6L \rfloor$, and $\lfloor 0.8L \rfloor$; percentages are nominal depth selections.

| Backbone | $L$ | 20% | 40% | 60% | 80% | Final |
|---|---|---|---|---|---|---|
| DeepSeek-R1-1.5B | 28 | 5 | 11 | 16 | 22 | 28 |
| Llama-3.2-1B | 16 | 3 | 6 | 9 | 12 | 16 |
| Llama-3.1-8B | 32 | 6 | 12 | 19 | 25 | 32 |
| Qwen3-4B | 36 | 7 | 14 | 21 | 28 | 36 |
| Qwen3-8B | 36 | 7 | 14 | 21 | 28 | 36 |

## D.2 DATA SPLITS AND RESULT AGGREGATION

Probes are trained and evaluated separately for each language and each subset. Keeping Chinese and English versions in separate runs prevents a translated version of a held-out scenario from appearing in that probe's training data.

**Vignette: leave one scenario type out.** The eight scenario types describe event settings, whereas the three inter-agent relation categories describe how agents are related (Appendix B.1). We use eight-fold leave-one-scenario-type-out cross-validation: each fold holds out all samples of one scenario type for testing and uses the other seven types for training. Occurrences of the same type across relation categories belong to the same fold, including all their agent questions and narrative variants. For example, a Company test fold includes Company templates from General, Vicarious Blame, and Commanding Chain; a Laboratory test fold includes both of its relation-specific templates in Table 1. Each scenario type is held out once. Withholding entire event settings limits the opportunity to memorize recurring scenario content or templates, reducing the risk of inflated decodability from such overlap. The resulting evaluation tests whether attribution dimensions can be decoded in unseen scenario types, strengthening the evidence for generalization across event settings.

**Reality: random cross-validation.** We use 10-fold random cross-validation, training on nine folds and testing on the remaining fold in each round. The same subset-specific splits are used for the original probes and decodability controls at both the label-token prediction step and the last input token.

**Eligibility and aggregation.** For dimension $k$, training and evaluation require a non-missing label $y_{i,k} \in \{0, 1\}$ and an available hidden state at the probed location. Questions with unspecified or inapplicable values are excluded for that dimension; they need not be excluded from other dimensions for which labels are available. Consequently, the eligible sample set can differ across dimensions. All probe fitting and direction construction use training-fold data; interventions are evaluated on held-out prompts.

For a fixed language–subset combination, model, task, dimension, and layer, we compute balanced accuracy on each held-out fold and take the unweighted mean over the eight *Vignette* folds or the ten *Reality* folds. This is the summary for an individual language–subset run. For the combined decodability plots, we instead collect the held-out predictions and labels from all folds of all four language–subset runs and recompute balanced accuracy on that pooled set. Each eligible question contributes the prediction from its test fold. Thus, the combined score is computed from pooled predictions, rather than by averaging the fold means or the four run-level scores; these aggregations can differ when fold sizes or class frequencies differ. Probe training is repeated with seeds 0, 1, and 2, and the resulting scores are averaged across seeds. Pooling predictions for reporting does not entail fitting a joint bilingual probe. Balanced accuracy gives the two binary dimension values equal importance, which is necessary because their label frequencies differ. It is distinct from benchmark accuracy, whose targets are the four human responsibility/blame labels.

### D.3 Probe Training and Decodability Controls

**Training objective.** For fold $f$, dimension $k$, and layer $l$, let

$$\mathcal{D}_{\mathrm{tr},k}^{(f,l)} = \left\{ i \in \mathcal{D}_{\mathrm{tr}}^{(f)} \mid y_{i,k} \neq \varnothing,\ z_i^{(l)} \text{ available} \right\}. \tag{10}$$

The logistic probe is fitted by minimizing class-balanced cross-entropy with $L_2$ regularization:

$$\min_{w,b} \sum_{i \in \mathcal{D}_{\mathrm{tr},k}^{(f,l)}} \beta_{y_{i,k}} \, \mathrm{CE}\Big(y_{i,k}, \sigma(w^\top z_i^{(l)} + b)\Big) + \lambda \|w\|_2^2, \tag{11}$$

where $\beta_y$ denotes class-balancing weights and $\mathrm{CE}(y, p) = -y \log p - (1 - y) \log(1 - p)$. The held-out set is defined by the same label and hidden-state eligibility conditions. Decodability is its balanced accuracy, $(\mathrm{Recall}_0 + \mathrm{Recall}_1)/2$, with 50% denoting chance.

High probe accuracy alone does not establish that the model represents the intended construct independently of superficial text patterns (Hewitt & Liang, 2019; Belinkov, 2022). The vignette construction therefore varies narrative phrasing and character identities and avoids explicit dimension labels; leaving entire scenario types out of training further reduces shortcuts based on recurring event settings. We complement these design choices with three controls, using the same splits, class-balanced logistic classifier, and balanced-accuracy metric.

**Shuffled-label control.** For each dimension and training fold, we randomly permute the training labels while leaving hidden states and test labels unchanged. This preserves the training-label counts but removes their original correspondence with the representations. It tests whether held-out performance depends on that correspondence rather than on label balance or the probe-fitting procedure alone.

**Random-projection control.** At each layer, we replace the hidden state with $\tilde{z}_i^{(l)} = R_l^\top z_i^{(l)}$, where $R_l \in \mathbb{R}^{d \times 64}$ and $(R_l)_{ab} \sim \mathcal{N}(0, 1/64)$. The same random matrix is applied to the training and test states at that layer. This control measures how much dimension information remains after compression to 64 features. Because it retains a projection of the original representation, it is a compression comparison, not a label-independent chance baseline.

**Lexical control.** We replace hidden states with character 3–5-gram TF-IDF features extracted from the scenario text alone. This control measures predictive information available from surface text without using the LLM's hidden representations or generated reasoning. The original-probe and control curves in Appendix F.1 therefore provide complementary evidence: shuffled labels check the label–representation association, while random projection and lexical features compare the full representation with restricted alternatives.

### D.4 Intervention Implementation and Hyperparameters

**Direction construction.** At each selected layer and training fold, we derive a direction from the target dimension's probe weight. Equation 2 removes components explained by the other available dimension-probe weights using regularized residualization, then normalizes the residual to unit length. This reduces shared linear components among correlated dimensions; it does not establish complete independence among the underlying attribution factors. Degenerate residuals are skipped. The positive direction is oriented to increase probe evidence for $y_{i,k} = 1$. The directional expectations apply to the negative-outcome scenarios considered here, as discussed in Appendix A.

**Prediction-step intervention.** For each held-out prompt, baseline generation identifies $\mathrm{Idx}_i^\star$. A single-use forward hook adds $\eta\alpha v_k^{(l)}$ at that prediction step and selected layer, following Equation 3. We generate exactly one token at the edited step and obtain the comparison prediction by restricting the argmax to the four canonical labels. The hook is local to that step: the main intervention asks whether the decoded direction can influence the final judgment at the output decision stage. The earlier-token experiment moves both probe training and the edit to the last input token and permits subsequent autoregressive generation, providing a separate test of effects initiated before this decision stage.

**Matched controls.** The unedited baseline establishes the direction of change in the judgment. The random-direction control uses a unit vector independent of dimension labels and probe weights. The shuffled-label control derives a direction from a training-fold probe fitted to shuffled labels, using the same direction-construction procedure. Both intervention controls match the target intervention's layer, prediction step, strength, sign, and perturbation norm. This matching tests whether the learned dimension direction has a stronger theory-aligned effect than a generic perturbation of the same size or a direction learned without the original label association.

Table 12: Settings shared across the probing and intervention analyses. Backbone-specific layer selections appear in Table 11.

| Setting | Value or procedure |
|---|---|
| Prompt strategy | CoT |
| Probe-training runs | Separate for each language and subset |
| Vignette cross-validation | 8 folds; one scenario type held out, seven used for training |
| Reality cross-validation | 10 random folds; one held out, nine used for training |
| Probe-training seeds | 0, 1, 2 |
| Decodability score | Balanced accuracy; chance 50% |
| Random-projection dimension | 64 |
| Lexical features | Scenario-text character 3–5-gram TF-IDF |
| Direction norm | Unit $L_2$ norm |
| Intervention signs | $+1$ for unedited `no`; $-1$ for `high`; both otherwise |
| Intervention strengths | $\alpha \in \{1, 2, 4, 8, 16\}$ |
| Main intervention strength | $\alpha = 8$, the pre-specified representative setting |
| Test unit | One binary outcome per question and comparison |
| Success criterion | Every feasible sign strictly aligns; ties count as failures |
| Directional test | One-sided exact binomial test; null threshold $\pi_c = 1/2$ |
| Passing criterion | $N > 0$ and $p_c < 0.01$ for all three comparisons |

The full strength grid characterizes how the average signed judgment change varies with perturbation size. Appendix G.2 reports this descriptive sensitivity analysis using the continuous measure $D(\alpha)$, averaged equally across the evaluated layer–dimension settings.

### D.5 Statistical Testing and Reporting Details

For each setting $(m, l, k, \alpha)$, the unedited output fixes $\mathcal{E}_i$ for all three comparisons, with one sign at the endpoints and both signs otherwise (Table 12). Equation 5 assigns $s_i^c = 1$ only when every sign in $\mathcal{E}_i$ has $\Delta_{i,\eta}^c > 0$; a tie or opposite shift on any tested sign gives $s_i^c = 0$. Each question, including endpoint cases, contributes one binary outcome per comparison and remains in $N$.

For $S_c = \sum_{i=1}^{N} s_i^c$, we apply the one-sided exact binomial test in Equation 6 to $H_0 : \pi_c \leq 1/2$ against $H_1 : \pi_c > 1/2$. The threshold $1/2$ tests whether more than half of the questions satisfy their applicable directional criterion. The binomial model assumes independent question-level indicators, without requiring independence between intervention outcomes within questions that have two feasible signs. The reported $p$-values do not account for dependence among questions derived from the same scenario. A setting is reported as "Tests passed" (blue in the heatmaps) only when $N > 0$ and all three comparisons satisfy $p_c < 0.01$. The same criterion is applied separately against the unedited baseline, random-direction output, and shuffled-label output.

"Tests not passed" (gray) denotes $N = 0$ or insufficient evidence in at least one comparison; it does not establish the absence of a dimension or its influence. The binary tests assess directional success, while $D(\alpha)$ (Appendix G.2) averages signed changes over the same feasible signs before averaging across questions.

# E FULL BENCHMARK RESULTS

The aggregate comparisons in Section 5.2 summarize agreement with human judgments, but can conceal differences between prompting strategies, tasks, and scenario groups. We therefore report the complete results and complementary breakdowns below. These analyses establish the behavioral setting for the representation experiments: successful recovery of attribution dimensions should not be confused with complete agreement between model outputs and human judgments.

## E.1 FULL LLM RESULTS ACROSS PROMPTING STRATEGIES

Tables 13 and 14 report all 32 LLMs under the four prompting strategies. Accuracy measures exact agreement with the four-level human judgments across both subsets; Spearman's $\rho$ measures within-scenario agent-ranking agreement on *Vignette*. The two metrics distinguish calibration of judgment levels from preservation of relative responsibility or blame. In the tables, N denotes the Direct prompt, SR denotes Self-reflection, and TG denotes Theory-guided prompting; Avg. averages over the four strategies.

Table 13: Accuracy of 32 representative LLMs on our constructed social attribution benchmark.

**(a) Responsibility judgment**

| Model | Prompt Strategy N | CoT | SR | TG | Avg. | 95% CI |
|---|---|---|---|---|---|---|
| Closed-source Models | | | | | | |
| Claude-Haiku-4.5 | 55.71 | 52.01 | 54.75 | 56.09 | 54.64 | [53.16, 56.16] |
| Doubao-Seed-1.8 | 52.70 | 52.91 | 52.39 | 54.51 | 53.13 | [51.34, 54.81] |
| Gemini-3.5-flash | 50.09 | 52.19 | 52.82 | 52.73 | 51.96 | [50.35, 53.77] |
| GPT-5-mini | **57.18** | 57.67 | 55.96 | 56.66 | **56.87** | [55.36, 58.38] |
| GPT-5-nano | 53.04 | 50.79 | 47.35 | 53.04 | 51.05 | [49.64, 52.66] |
| Grok-4.3 | 56.09 | 57.75 | 56.09 | 53.95 | 55.97 | [54.33, 57.58] |
| Kimi-K2.5 | 54.49 | 55.73 | 52.86 | 57.77 | 55.21 | [53.70, 56.74] |
| MiMo-v2.5 | 56.77 | 55.55 | 54.83 | **58.82** | 56.49 | [54.92, 58.03] |
| Step-3.7-flash | 53.91 | **58.47** | **56.72** | 51.52 | 55.15 | [53.27, 57.00] |
| Open-source Models | | | | | | |
| Deepseek-R1-70B | 47.24 | 49.86 | 46.42 | 43.80 | 46.83 | [45.28, 48.41] |
| Deepseek-R1-1.5B | 27.48 | 28.56 | 27.61 | 30.51 | 28.54 | [27.63, 29.38] |
| Gemma3-27B | 51.98 | 52.70 | 48.62 | 46.88 | 50.05 | [48.63, 51.54] |
| Gemma3-4B | 41.63 | 47.87 | 42.23 | 40.80 | 43.13 | [42.02, 44.28] |
| Gemma3-1B | 16.07 | 39.07 | 34.75 | 36.61 | 31.62 | [30.88, 32.38] |
| Glm4-9B | 39.38 | 47.76 | 41.37 | 37.05 | 41.39 | [40.43, 42.41] |
| GPT-oss-120B | 54.33 | 53.89 | 53.45 | 54.64 | 54.08 | [52.46, 55.57] |
| GPT-oss-20B | 52.57 | 52.52 | 51.69 | 50.40 | 51.80 | [50.22, 53.42] |
| InternLM2.5-7B | 36.12 | 41.16 | 35.89 | 35.24 | 37.10 | [35.99, 38.28] |
| InternLM2.5-1.8B | 25.59 | 33.76 | 33.45 | 30.38 | 30.80 | [29.70, 31.79] |
| Llama3.1-70B | 49.91 | 55.94 | 51.10 | 52.65 | 52.40 | [50.98, 53.82] |
| Llama3.1-8B | 36.07 | 39.90 | 38.55 | 38.06 | 38.14 | [37.07, 39.13] |
| Llama3.2-1B | 17.72 | 14.85 | 9.99 | 13.74 | 14.08 | [13.19, 14.89] |
| Phi4-14B | 49.42 | 56.40 | 52.24 | 50.09 | 52.04 | [50.58, 53.51] |
| Qwen2-7B | 36.82 | 40.60 | 38.84 | 34.83 | 37.77 | [36.57, 38.92] |
| Qwen2-1.5B | 19.66 | 15.58 | 31.57 | 26.78 | 23.40 | [22.76, 24.06] |
| Qwen2-0.5B | 16.02 | 14.26 | 16.48 | 16.14 | 15.72 | [14.50, 16.91] |
| Qwen2.5-7B | 42.90 | 49.70 | 34.75 | 41.68 | 42.26 | [41.02, 43.43] |
| Qwen2.5-1.5B | 10.97 | 31.38 | 25.92 | 20.54 | 22.21 | [21.53, 22.88] |
| Qwen3-14B | 50.43 | 53.22 | 45.30 | 54.49 | 50.86 | [49.50, 52.27] |
| Qwen3-8B | 46.70 | 50.87 | 43.10 | 53.07 | 48.43 | [47.12, 49.74] |
| Qwen3-4B | 48.23 | 48.77 | 47.58 | 48.95 | 48.38 | [46.63, 49.98] |
| Qwen3-0.6B | 27.04 | 26.34 | 26.29 | 31.57 | 27.81 | [27.09, 28.51] |

**(b) Blame judgment**

| Model | Prompt Strategy N | CoT | SR | TG | Avg. | 95% CI |
|---|---|---|---|---|---|---|
| Closed-source Models | | | | | | |
| Claude-Haiku-4.5 | 52.23 | 48.97 | 48.99 | 54.27 | 51.11 | [49.60, 52.58] |
| Doubao-Seed-1.8 | 49.92 | 52.65 | 52.17 | 57.10 | 52.96 | [51.32, 54.48] |
| Gemini-3.5-flash | 46.79 | 55.09 | 55.72 | 56.65 | 53.56 | [52.11, 55.11] |
| GPT-5-mini | 51.19 | 48.65 | 46.77 | 45.34 | 47.99 | [46.20, 49.84] |
| GPT-5-nano | 49.84 | 46.85 | 45.84 | 44.12 | 46.66 | [45.17, 48.00] |
| Grok-4.3 | **60.31** | **60.60** | **60.25** | **58.96** | **60.03** | [58.53, 61.49] |
| Kimi-K2.5 | 51.14 | 55.80 | 49.68 | 58.45 | 53.77 | [52.45, 55.17] |
| MiMo-v2.5 | 50.66 | 49.63 | 46.34 | 51.41 | 49.51 | [48.00, 50.87] |
| Step-3.7-flash | 55.81 | 60.60 | 58.45 | 56.00 | 57.72 | [56.06, 59.50] |
| Open-source Models | | | | | | |
| Deepseek-R1-70B | 40.70 | 46.79 | 41.84 | 41.97 | 42.83 | [41.23, 44.23] |
| Deepseek-R1-1.5B | 24.64 | 25.13 | 24.40 | 24.83 | 24.75 | [23.88, 25.64] |
| Gemma3-27B | 45.57 | 45.23 | 43.14 | 40.36 | 43.57 | [41.89, 45.15] |
| Gemma3-4B | 34.50 | 36.46 | 32.75 | 34.98 | 34.67 | [33.47, 36.02] |
| Gemma3-1B | 23.16 | 28.43 | 26.10 | 29.01 | 26.68 | [25.90, 27.49] |
| Glm4-9B | 32.33 | 42.53 | 32.86 | 38.58 | 36.57 | [35.46, 37.74] |
| GPT-oss-120B | 48.68 | 49.71 | 47.83 | 47.85 | 48.52 | [46.96, 50.11] |
| GPT-oss-20B | 45.28 | 45.57 | 45.07 | 45.71 | 45.41 | [44.09, 46.80] |
| InternLM2.5-7B | 30.47 | 32.43 | 28.35 | 28.86 | 30.03 | [28.79, 31.30] |
| InternLM2.5-1.8B | 25.09 | 26.34 | 21.97 | 25.25 | 24.66 | [23.65, 25.73] |
| Llama3.1-70B | 48.46 | 53.29 | 47.09 | 52.99 | 50.46 | [49.16, 51.74] |
| Llama3.1-8B | 30.15 | 36.22 | 34.18 | 32.35 | 33.23 | [32.15, 34.44] |
| Llama3.2-1B | 21.78 | 11.63 | 10.73 | 18.57 | 15.68 | [14.77, 16.64] |
| Phi4-14B | 48.20 | 52.70 | 43.35 | 46.03 | 47.57 | [46.20, 48.96] |
| Qwen2-7B | 31.90 | 36.22 | 30.45 | 34.66 | 33.31 | [32.09, 34.54] |
| Qwen2-1.5B | 22.97 | 17.33 | 25.30 | 23.13 | 22.18 | [21.44, 22.87] |
| Qwen2-0.5B | 21.52 | 21.36 | 22.36 | 20.88 | 21.53 | [20.31, 22.84] |
| Qwen2.5-7B | 34.82 | 43.96 | 30.63 | 38.74 | 37.04 | [35.82, 38.24] |
| Qwen2.5-1.5B | 24.06 | 29.09 | 27.11 | 26.79 | 26.76 | [26.01, 27.57] |
| Qwen3-14B | 40.20 | 47.38 | 36.78 | 43.00 | 41.84 | [40.39, 43.12] |
| Qwen3-8B | 41.02 | 44.57 | 34.74 | 45.92 | 41.56 | [40.15, 42.96] |
| Qwen3-4B | 44.78 | 44.49 | 44.20 | 45.95 | 44.85 | [43.38, 46.40] |
| Qwen3-0.6B | 26.29 | 26.71 | 25.01 | 30.55 | 27.14 | [26.36, 27.99] |

* The best results are highlighted in **bold** and the second-best results are underlined. Accuracy is reported as a percentage and rounded to two decimal places. The 95% confidence intervals are for Avg. and are cluster-bootstrapped over scenario groups. Strategy abbreviations: N = Direct, CoT = chain-of-thought, SR = Self-reflection, TG = Theory-guided.

**Task and metric differences.** GPT-5-mini has the highest prompt-averaged responsibility accuracy (56.87%), whereas Step-3.7-flash has the highest responsibility rank agreement ($\rho = 0.588$). Grok-4.3 leads on both blame accuracy (60.03%) and blame rank agreement ($\rho = 0.702$). Across all 32 LLMs, responsibility accuracy exceeds blame accuracy on average (43.10% versus 39.82%), but rank agreement is higher for blame (0.400 versus 0.348). Most models follow these task-specific patterns. Thus, matching the absolute judgment category and preserving the ordering of agents are empirically distinct aspects of social attribution. Models can rank agents' blame reasonably well while disagreeing with the precise severity assigned by human annotators.

Table 14: Spearman $\rho$ of 32 representative LLMs on our constructed social attribution benchmark.

**(a) Responsibility judgment**

| Model | Prompt Strategy | | | | Avg. | 95% CI |
|---|---|---|---|---|---|---|
| | N | CoT | SR | TG | | |
| *Closed-source Models* | | | | | | |
| Claude-Haiku-4.5 | 0.554 | 0.534 | 0.538 | 0.544 | 0.542 | [0.509, 0.575] |
| Doubao-Seed-1.8 | 0.535 | 0.540 | 0.530 | 0.532 | 0.534 | [0.495, 0.571] |
| Gemini-3.5-flash | 0.519 | 0.554 | 0.521 | 0.541 | 0.534 | [0.497, 0.570] |
| GPT-5-mini | 0.556 | 0.567 | 0.538 | 0.534 | 0.549 | [0.515, 0.584] |
| GPT-5-nano | 0.435 | 0.411 | 0.344 | 0.440 | 0.407 | [0.372, 0.443] |
| Grok-4.3 | 0.545 | 0.566 | 0.547 | 0.550 | 0.552 | [0.515, 0.585] |
| Kimi-K2.5 | 0.513 | 0.543 | 0.488 | 0.515 | 0.515 | [0.478, 0.549] |
| MiMo-v2.5 | 0.539 | 0.546 | 0.535 | 0.547 | 0.542 | [0.506, 0.575] |
| Step-3.7-flash | 0.550 | **0.595** | **0.655** | **0.554** | **0.588** | [0.554, 0.629] |
| *Open-source Models* | | | | | | |
| Deepseek-R1-70B | 0.464 | 0.476 | 0.430 | 0.331 | 0.426 | [0.393, 0.457] |
| Deepseek-R1-1.5B | 0.063 | 0.107 | 0.057 | 0.133 | 0.090 | [0.069, 0.110] |
| Gemma3-27B | 0.506 | 0.539 | 0.457 | 0.457 | 0.490 | [0.455, 0.521] |
| Gemma3-4B | 0.344 | 0.392 | 0.233 | 0.363 | 0.333 | [0.302, 0.360] |
| Gemma3-1B | 0.072 | 0.107 | 0.041 | 0.085 | 0.076 | [0.060, 0.091] |
| Glm4-9B | 0.426 | 0.415 | 0.193 | 0.414 | 0.362 | [0.333, 0.392] |
| GPT-oss-120B | 0.515 | 0.500 | 0.488 | 0.501 | 0.501 | [0.464, 0.535] |
| GPT-oss-20B | 0.462 | 0.455 | 0.466 | 0.453 | 0.459 | [0.422, 0.494] |
| InternLM2.5-7B | 0.310 | 0.318 | 0.228 | 0.306 | 0.290 | [0.262, 0.320] |
| InternLM2.5-1.8B | 0.044 | 0.002 | 0.030 | -0.096 | -0.005 | [-0.020, 0.010] |
| Llama3.1-70B | **0.558** | 0.521 | 0.457 | 0.509 | 0.511 | [0.480, 0.546] |
| Llama3.1-8B | 0.293 | 0.306 | 0.154 | 0.194 | 0.237 | [0.212, 0.263] |
| Llama3.2-1B | 0.000 | 0.020 | 0.035 | 0.015 | 0.017 | [0.009, 0.025] |
| Phi4-14B | 0.484 | 0.551 | 0.516 | 0.506 | 0.514 | [0.483, 0.547] |
| Qwen2-7B | 0.331 | 0.301 | 0.271 | 0.316 | 0.305 | [0.279, 0.332] |
| Qwen2-1.5B | 0.079 | 0.006 | -0.027 | -0.015 | 0.011 | [-0.001, 0.023] |
| Qwen2-0.5B | 0.037 | -0.010 | -0.012 | -0.006 | 0.002 | [-0.015, 0.020] |
| Qwen2.5-7B | 0.324 | 0.378 | 0.200 | 0.364 | 0.316 | [0.286, 0.349] |
| Qwen2.5-1.5B | -0.046 | 0.102 | 0.047 | 0.098 | 0.050 | [0.031, 0.069] |
| Qwen3-14B | 0.427 | 0.464 | 0.410 | 0.478 | 0.445 | [0.408, 0.477] |
| Qwen3-8B | 0.362 | 0.446 | 0.420 | 0.494 | 0.430 | [0.397, 0.465] |
| Qwen3-4B | 0.437 | 0.440 | 0.414 | 0.434 | 0.432 | [0.397, 0.472] |
| Qwen3-0.6B | -0.023 | 0.125 | 0.001 | 0.157 | 0.065 | [0.046, 0.085] |

**(b) Blame judgment**

| Model | Prompt Strategy | | | | Avg. | 95% CI |
|---|---|---|---|---|---|---|
| | N | CoT | SR | TG | | |
| *Closed-source Models* | | | | | | |
| Claude-Haiku-4.5 | 0.666 | 0.649 | 0.622 | 0.668 | 0.651 | [0.623, 0.679] |
| Doubao-Seed-1.8 | 0.635 | 0.655 | 0.632 | 0.670 | 0.648 | [0.618, 0.677] |
| Gemini-3.5-flash | 0.488 | 0.676 | 0.668 | 0.690 | 0.631 | [0.604, 0.658] |
| GPT-5-mini | 0.672 | 0.649 | 0.654 | 0.658 | 0.658 | [0.631, 0.685] |
| GPT-5-nano | 0.532 | 0.535 | 0.497 | 0.518 | 0.521 | [0.492, 0.552] |
| Grok-4.3 | **0.695** | **0.695** | **0.700** | **0.718** | **0.702** | [0.675, 0.727] |
| Kimi-K2.5 | 0.637 | 0.694 | 0.604 | 0.693 | 0.657 | [0.627, 0.684] |
| MiMo-v2.5 | 0.615 | 0.626 | 0.584 | 0.591 | 0.604 | [0.573, 0.634] |
| Step-3.7-flash | 0.669 | 0.668 | 0.677 | 0.584 | 0.650 | [0.611, 0.680] |
| *Open-source Models* | | | | | | |
| Deepseek-R1-70B | 0.503 | 0.549 | 0.453 | 0.402 | 0.477 | [0.448, 0.504] |
| Deepseek-R1-1.5B | 0.051 | 0.010 | 0.035 | 0.087 | 0.046 | [0.024, 0.068] |
| Gemma3-27B | 0.565 | 0.582 | 0.497 | 0.542 | 0.546 | [0.516, 0.575] |
| Gemma3-4B | 0.239 | 0.317 | 0.098 | 0.377 | 0.258 | [0.234, 0.282] |
| Gemma3-1B | 0.043 | 0.012 | 0.054 | 0.043 | 0.038 | [0.021, 0.054] |
| Glm4-9B | 0.357 | 0.472 | 0.194 | 0.479 | 0.376 | [0.347, 0.404] |
| GPT-oss-120B | 0.633 | 0.625 | 0.619 | 0.620 | 0.624 | [0.593, 0.653] |
| GPT-oss-20B | 0.554 | 0.586 | 0.568 | 0.604 | 0.578 | [0.547, 0.607] |
| InternLM2.5-7B | 0.329 | 0.321 | 0.175 | 0.297 | 0.281 | [0.254, 0.307] |
| InternLM2.5-1.8B | 0.023 | -0.012 | -0.026 | -0.004 | -0.005 | [-0.018, 0.008] |
| Llama3.1-70B | 0.652 | 0.660 | 0.535 | 0.660 | 0.627 | [0.598, 0.653] |
| Llama3.1-8B | 0.331 | 0.408 | 0.170 | 0.322 | 0.308 | [0.281, 0.335] |
| Llama3.2-1B | 0.000 | 0.011 | 0.003 | 0.022 | 0.009 | [0.002, 0.016] |
| Phi4-14B | 0.598 | 0.627 | 0.561 | 0.610 | 0.599 | [0.569, 0.626] |
| Qwen2-7B | 0.348 | 0.358 | 0.222 | 0.384 | 0.328 | [0.300, 0.355] |
| Qwen2-1.5B | 0.072 | 0.022 | -0.002 | -0.018 | 0.019 | [0.004, 0.032] |
| Qwen2-0.5B | 0.036 | -0.027 | -0.003 | 0.005 | 0.003 | [-0.012, 0.019] |
| Qwen2.5-7B | 0.347 | 0.470 | 0.236 | 0.431 | 0.371 | [0.342, 0.401] |
| Qwen2.5-1.5B | 0.044 | 0.060 | 0.089 | 0.066 | 0.065 | [0.044, 0.087] |
| Qwen3-14B | 0.464 | 0.562 | 0.428 | 0.545 | 0.500 | [0.467, 0.530] |
| Qwen3-8B | 0.348 | 0.518 | 0.401 | 0.571 | 0.460 | [0.428, 0.489] |
| Qwen3-4B | 0.512 | 0.497 | 0.490 | 0.549 | 0.512 | [0.478, 0.544] |
| Qwen3-0.6B | 0.002 | 0.136 | -0.048 | 0.105 | 0.049 | [0.030, 0.066] |

* The best results are highlighted in **bold** and the second-best results are underlined. Spearman's $\rho$ is reported on its original $[-1, 1]$ scale and rounded to three decimal places. The 95% confidence intervals are for Avg. and are cluster-bootstrapped over scenario groups. Strategy abbreviations: N = Direct, CoT = chain-of-thought, SR = Self-reflection, TG = Theory-guided.

**Prompting strategy.** Averaging over all models and both tasks, accuracy is 40.45% for Direct, 43.13% for CoT, 40.33% for Self-reflection, and 41.93% for Theory-guided prompting; the corresponding correlations are 0.373, 0.397, 0.337, and 0.387. CoT also has the highest mean score within each task. This comparison motivates examining attribution representations during CoT-based judgments: explicit reasoning improves behavioral agreement on average, while leaving substantial disagreement to investigate. Theory-guided prompting also improves the aggregate scores over Direct, but explicitly asking for the five dimensions does not guarantee the best judgment. Self-reflection does not improve agreement on average. These are aggregate comparisons, with individual model–strategy exceptions visible in the tables.

**Model families and generations.** The closed-source/API group averages 54.50% responsibility accuracy and 52.59% blame accuracy, compared with 38.64% and 34.82% for the open-source group; the corresponding correlations are 0.529/0.636 and 0.276/0.307. Family-level comparisons further show why parameter count alone is insufficient. At 8B parameters, Qwen3-8B exceeds Llama-3.1-8B on both tasks and metrics. Within the Qwen family, the 7B–8B sequence Qwen2-7B, Qwen2.5-7B, and Qwen3-8B improves in responsibility accuracy from 37.77% to 42.26% to 48.43%, and in blame accuracy from 33.31% to 37.04% to 41.56%. Rank agreement improves along the same sequence. Differences in training and model generation therefore remain relevant when interpreting the model-size association in Figure 4.

### E.2 Full Non-LLM Baseline Results

The five baselines test how much agreement can be obtained from label frequencies, surface textual cues, explicit attribution dimensions, or supervised text classification. Their implementations are given in Appendix C.5; Table 15 reports both tasks and metrics.

Table 15: Social attribution performance of basic baselines.

**(a) Responsibility judgment**

| Baseline | Acc. | 95% CI | $\rho$ | 95% CI |
|---|---|---|---|---|
| `Random Stratified` | 30.16 | [28.62, 31.68] | 0.028 | [-0.013, 0.068] |
| `Majority Class` | <u>40.93</u> | [38.87, 42.99] | 0.000 | [0.000, 0.000] |
| `Heuristic Rule` | 21.51 | [19.82, 23.16] | 0.115 | [0.078, 0.149] |
| `Attribution Rule` | 33.52 | [31.52, 35.43] | **0.303** | [0.246, 0.353] |
| `mBERT Classifier` | **43.47** | [41.43, 45.32] | <u>0.141</u> | [0.112, 0.170] |

**(b) Blame judgment**

| Baseline | Acc. | 95% CI | $\rho$ | 95% CI |
|---|---|---|---|---|
| `Random Stratified` | 27.48 | [25.89, 29.09] | -0.023 | [-0.063, 0.015] |
| `Majority Class` | 31.37 | [29.20, 33.39] | 0.000 | [0.000, 0.000] |
| `Heuristic Rule` | 26.89 | [25.07, 28.87] | <u>0.138</u> | [0.106, 0.169] |
| `Attribution Rule` | **46.91** | [44.61, 49.39] | **0.368** | [0.314, 0.421] |
| `mBERT Classifier` | <u>31.60</u> | [29.66, 33.60] | 0.069 | [0.042, 0.098] |

The best results are highlighted in **bold** and the second-best results are <u>underlined</u>.

mBERT has the highest responsibility accuracy among the baselines (43.47%), while the Attribution-Theory Rule has the highest responsibility rank agreement (0.303). On blame, the Attribution-Theory Rule leads on both accuracy (46.91%) and correlation (0.368). This contrast motivates distinguishing textual prediction from reasoning over attribution-relevant factors. However, the rule baseline receives annotated dimension values, whereas mBERT and the LLMs must infer relevant information from text; the comparison therefore does not isolate model architecture or establish that a particular reasoning mechanism is necessary. The strongest LLMs exceed these baselines, but the limited absolute agreement across methods reinforces the need to examine intermediate attribution representations.

### E.3 VALID-OUTPUT RATES

Because the evaluation excludes invalid predictions, output validity is needed to interpret the coverage of the reported agreement scores. A model that seldom produces an extractable answer should not be assessed from its agreement score alone. The parser and task-specific exclusion rules are described in Appendix C.3.

We report the proportion of valid LLM outputs for each model, task, and prompt strategy. An output is considered valid if the evaluator can extract one canonical answer label from {`no`, `low`, `medium`, `high`}. Across all evaluated LLM outputs, 98.49% are valid, and 1.51% are invalid. These invalid cases are excluded from the calculation of metrics.

Table 16: Valid output rates of evaluated LLMs by task and prompt strategy.

**(a) Responsibility judgment**

| Model | Prompt Strategy | | | | Avg. |
|---|---|---|---|---|---|
| | N | CoT | SR | TG | |
| Closed-source Models | | | | | |
| Claude-Haiku-4.5 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Doubao-Seed-1.8 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Gemini-3.5-flash | 99.90 | 100.00 | 100.00 | 100.00 | 99.97 |
| GPT-5-mini | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| GPT-5-nano | 100.00 | 99.12 | 92.94 | 98.50 | 97.64 |
| Grok-4.3 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Kimi-K2.5 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| MiMo-v2.5 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Step-3.7-flash | 100.00 | 97.87 | 99.92 | 96.54 | 98.58 |
| Open-source Models | | | | | |
| Deepseek-R1-70B | 99.90 | 97.98 | 96.07 | 89.50 | 95.86 |
| Deepseek-R1-1.5B | 95.78 | 97.26 | 97.88 | 98.68 | 97.40 |
| Gemma3-27B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Gemma3-4B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Gemma3-1B | 100.00 | 99.92 | 89.99 | 99.79 | 97.43 |
| Glm4-9B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| GPT-oss-120B | 99.97 | 100.00 | 99.97 | 100.00 | 99.99 |
| GPT-oss-20B | 100.00 | 99.84 | 99.87 | 99.97 | 99.92 |
| InternLM2.5-7B | 100.00 | 99.25 | 97.15 | 99.04 | 98.86 |
| InternLM2.5-1.8B | 100.00 | 91.38 | 89.52 | 90.74 | 92.91 |
| Llama3.1-70B | 100.00 | 99.33 | 97.57 | 99.79 | 99.17 |
| Llama3.1-8B | 100.00 | 87.83 | 86.27 | 91.85 | 91.49 |
| Llama3.2-1B | 100.00 | 91.45 | 93.79 | 86.94 | 93.05 |
| Phi4-14B | 100.00 | 100.00 | 100.00 | 99.97 | 99.99 |
| Qwen2-7B | 100.00 | 94.05 | 100.00 | 95.21 | 97.32 |
| Qwen2-1.5B | 100.00 | 66.91 | 99.09 | 96.58 | 90.65 |
| Qwen2-0.5B | 100.00 | 96.66 | 99.59 | 99.79 | 99.01 |
| Qwen2.5-7B | 100.00 | 100.00 | 99.84 | 100.00 | 99.96 |
| Qwen2.5-1.5B | 100.00 | 94.31 | 99.92 | 99.35 | 98.40 |
| Qwen3-14B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Qwen3-8B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Qwen3-4B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Qwen3-0.6B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

**(b) Blame judgment**

| Model | Prompt Strategy | | | | Avg. |
|---|---|---|---|---|---|
| | N | CoT | SR | TG | |
| Closed-source Models | | | | | |
| Claude-Haiku-4.5 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Doubao-Seed-1.8 | 99.97 | 100.00 | 100.00 | 100.00 | 99.99 |
| Gemini-3.5-flash | 98.99 | 100.00 | 100.00 | 100.00 | 99.75 |
| GPT-5-mini | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| GPT-5-nano | 100.00 | 99.84 | 97.11 | 99.02 | 98.99 |
| Grok-4.3 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Kimi-K2.5 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| MiMo-v2.5 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Step-3.7-flash | 99.97 | 97.57 | 99.91 | 97.59 | 98.76 |
| Open-source Models | | | | | |
| Deepseek-R1-70B | 99.84 | 99.39 | 97.43 | 90.99 | 96.91 |
| Deepseek-R1-1.5B | 96.24 | 96.24 | 97.56 | 97.83 | 96.97 |
| Gemma3-27B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Gemma3-4B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Gemma3-1B | 100.00 | 99.97 | 98.81 | 99.87 | 99.66 |
| Glm4-9B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| GPT-oss-120B | 99.92 | 100.00 | 99.97 | 100.00 | 99.97 |
| GPT-oss-20B | 99.97 | 99.92 | 99.92 | 99.97 | 99.95 |
| InternLM2.5-7B | 100.00 | 98.22 | 92.29 | 97.85 | 97.09 |
| InternLM2.5-1.8B | 100.00 | 87.60 | 81.61 | 89.35 | 89.64 |
| Llama3.1-70B | 100.00 | 99.50 | 97.51 | 99.97 | 99.24 |
| Llama3.1-8B | 100.00 | 97.96 | 96.18 | 93.22 | 96.84 |
| Llama3.2-1B | 100.00 | 83.26 | 84.59 | 91.69 | 89.89 |
| Phi4-14B | 100.00 | 100.00 | 99.89 | 100.00 | 99.97 |
| Qwen2-7B | 100.00 | 98.20 | 100.00 | 97.64 | 98.96 |
| Qwen2-1.5B | 100.00 | 77.66 | 99.97 | 99.68 | 94.33 |
| Qwen2-0.5B | 100.00 | 95.95 | 99.95 | 99.92 | 98.95 |
| Qwen2.5-7B | 100.00 | 100.00 | 99.79 | 100.00 | 99.95 |
| Qwen2.5-1.5B | 100.00 | 99.13 | 99.95 | 99.95 | 99.75 |
| Qwen3-14B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Qwen3-8B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Qwen3-4B | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Qwen3-0.6B | 100.00 | 99.95 | 100.00 | 100.00 | 99.99 |

* Values are valid-output rates in percentages, pooled over two subsets and over English and Chinese items. Strategy abbreviations: N = Direct, CoT = chain-of-thought, SR = Self-reflection, TG = Theory-guided.

### E.4 PERFORMANCE BREAKDOWNS

The benchmark combines controlled relations and realistic narratives, with multiple values of the attribution dimensions. The following breakdowns check whether aggregate performance conceals systematic variation across scenario groups and dimension values. They complement the latent-space experiments by identifying behavioral differences that a single overall score cannot reveal.

**Agent relations and dataset subsets.** Tables 17 and 18 break down *Vignette* results by agent relation; the accuracy table also includes *Reality*. Responsibility accuracy is highest on *Reality* for all nine API-accessed LLMs, but the ordering of these groups changes across tasks. For GPT-5-mini, responsibility accuracy is 61.89% on *Reality*, compared with 51.88% on vicarious-blame scenarios and 46.16% on command chains. For blame, its highest score among these groups is instead on vicarious-blame scenarios (66.60%), compared with 49.39% on *Reality* and 44.68% on command chains. These differences support retaining complementary groups rather than treating aggregate agreement as a uniform measure across social relations. "General" denotes interactions without the designated supervisory or command relation.

Table 17: Accuracy by inter-agent relation within *Vignette*, with *Reality* shown separately.

**(a) Responsibility judgment**

| Model | Vignette agent relation | | | |
|---|---|---|---|---|
| | General | Vicarious | Commanding | Reality |
| *Closed-source Models* | | | | |
| Claude-Haiku-4.5 | 55.74 | 42.92 | 41.44 | 58.98 |
| Doubao-Seed-1.8 | 48.83 | 44.62 | 39.15 | 63.83 |
| Gemini-3.5-flash | 46.60 | 48.84 | 43.60 | 61.46 |
| GPT-5-mini | 55.87 | 51.88 | 46.16 | 61.89 |
| GPT-5-nano | 48.16 | 40.14 | 41.11 | 59.41 |
| Grok-4.3 | 51.80 | 39.25 | 47.98 | **66.55** |
| Kimi-K2.5 | 54.65 | 48.75 | 41.37 | 60.78 |
| MiMo-v2.5 | **56.43** | 51.34 | 43.41 | 60.98 |
| Step-3.7-flash | 50.32 | 37.76 | 37.81 | 66.15 |
| *Open-source Models* | | | | |
| Deepseek-R1-70B | 48.17 | 31.36 | 38.75 | 50.28 |
| Deepseek-R1-1.5B | 28.25 | 26.08 | 32.08 | 28.47 |
| Gemma3-27B | 47.92 | 50.09 | 39.56 | 55.43 |
| Gemma3-4B | 40.20 | 47.13 | 37.13 | 47.58 |
| Gemma3-1B | 29.61 | 33.42 | 31.47 | 33.84 |
| Glm4-9B | 42.17 | 33.24 | 39.42 | 42.51 |
| GPT-oss-120B | 53.72 | 42.74 | **50.67** | 57.64 |
| GPT-oss-20B | 52.50 | 40.23 | 47.24 | 54.37 |
| InternLM2.5-7B | 38.16 | 40.68 | 34.30 | 35.80 |
| InternLM2.5-1.8B | 26.91 | 37.01 | 30.86 | 34.43 |
| Llama3.1-70B | 50.81 | 38.08 | 47.04 | 58.59 |
| Llama3.1-8B | 38.23 | 34.59 | 35.51 | 39.41 |
| Llama3.2-1B | 16.74 | 8.15 | 21.36 | 9.99 |
| Phi4-14B | 52.96 | 34.59 | 43.26 | 56.57 |
| Qwen2-7B | 39.55 | 32.44 | 39.62 | 36.10 |
| Qwen2-1.5B | 24.47 | 22.40 | 29.78 | 20.58 |
| Qwen2-0.5B | 20.16 | 3.67 | 16.11 | 12.41 |
| Qwen2.5-7B | 43.25 | 33.87 | 40.70 | 43.06 |
| Qwen2.5-1.5B | 24.11 | 18.91 | 18.73 | 21.37 |
| Qwen3-14B | 50.59 | 41.31 | 43.94 | 54.87 |
| Qwen3-8B | 46.17 | 47.13 | 37.87 | 54.28 |
| Qwen3-4B | 45.86 | 44.44 | 36.66 | 55.36 |
| Qwen3-0.6B | 25.14 | 28.67 | 28.98 | 30.68 |
| *Basic Baselines* | | | | |
| Random Stratified | 27.72 | 34.77 | 31.00 | 35.06 |
| Majority Class | 34.94 | 60.22 | 34.50 | 51.19 |
| Heuristic Rule | 21.52 | 9.68 | 32.61 | 20.62 |
| Attribution Rule | 35.33 | 20.07 | 17.25 | 43.20 |
| mBERT Classifier | 35.72 | **60.93** | 35.31 | 60.45 |

**(b) Blame judgment**

| Model | Vignette agent relation | | | |
|---|---|---|---|---|
| | General | Vicarious | Commanding | Reality |
| *Closed-source Models* | | | | |
| Claude-Haiku-4.5 | 54.03 | 53.34 | 42.55 | 48.79 |
| Doubao-Seed-1.8 | 52.56 | 52.77 | 47.07 | 55.01 |
| Gemini-3.5-flash | 52.65 | 48.57 | 57.71 | 54.47 |
| GPT-5-mini | 44.66 | **66.60** | 44.68 | 49.39 |
| GPT-5-nano | 43.44 | 58.49 | 41.89 | 49.57 |
| Grok-4.3 | 61.48 | 46.09 | **62.03** | 60.33 |
| Kimi-K2.5 | 55.36 | 59.35 | 49.00 | 52.12 |
| MiMo-v2.5 | 49.03 | 64.03 | 42.72 | 49.21 |
| Step-3.7-flash | **62.82** | 37.95 | 50.74 | 56.87 |
| *Open-source Models* | | | | |
| Deepseek-R1-70B | 47.68 | 44.94 | 37.17 | 38.19 |
| Deepseek-R1-1.5B | 25.08 | 26.53 | 24.00 | 24.23 |
| Gemma3-27B | 40.99 | 53.34 | 36.04 | 46.82 |
| Gemma3-4B | 28.56 | 36.26 | 33.38 | 41.95 |
| Gemma3-1B | 24.26 | 29.87 | 26.99 | 28.86 |
| Glm4-9B | 36.89 | 45.42 | 30.92 | 36.06 |
| GPT-oss-120B | 45.97 | 64.98 | 44.68 | 49.53 |
| GPT-oss-20B | 46.81 | 57.06 | 41.09 | 42.76 |
| InternLM2.5-7B | 33.00 | 35.31 | 21.21 | 27.85 |
| InternLM2.5-1.8B | 19.85 | 26.53 | 28.66 | 28.97 |
| Llama3.1-70B | 54.30 | 53.05 | 43.22 | 47.34 |
| Llama3.1-8B | 34.09 | 37.69 | 28.52 | 32.62 |
| Llama3.2-1B | 17.98 | 18.99 | 21.08 | 10.94 |
| Phi4-14B | 52.60 | 50.48 | 39.49 | 43.20 |
| Qwen2-7B | 35.02 | 40.74 | 26.46 | 31.72 |
| Qwen2-1.5B | 23.15 | 28.15 | 25.60 | 19.07 |
| Qwen2-0.5B | 25.87 | 21.47 | 25.80 | 15.29 |
| Qwen2.5-7B | 40.05 | 43.23 | 26.46 | 35.11 |
| Qwen2.5-1.5B | 28.53 | 26.05 | 22.87 | 25.82 |
| Qwen3-14B | 41.10 | 50.38 | 35.84 | 42.73 |
| Qwen3-8B | 39.05 | 51.24 | 36.97 | 43.96 |
| Qwen3-4B | 46.72 | 42.94 | 43.28 | 43.41 |
| Qwen3-0.6B | 26.81 | 27.86 | 27.86 | 27.21 |
| *Basic Baselines* | | | | |
| Random Stratified | 27.19 | 22.90 | 23.94 | 29.71 |
| Majority Class | 26.02 | 29.01 | 32.18 | 44.12 |
| Heuristic Rule | 31.84 | 15.65 | 25.80 | 20.74 |
| Attribution Rule | 46.38 | 29.01 | 50.00 | 50.73 |
| mBERT Classifier | 19.25 | 25.95 | 31.38 | **62.92** |

* The best results are highlighted in **bold** and the second-best results are underlined. LLM cells are averages over N, CoT, SR, and TG. Accuracy is reported as a percentage and rounded to two decimal places. Accuracy is reported separately for *Vignette* agent relations and *Reality*.

Table 18: Spearman $\rho$ by inter-agent relation within *Vignette*.

**(a) Responsibility judgment**

| Model | Vignette agent relation | | |
|---|---|---|---|
| | General | Vicarious | Commanding |
| Closed-source Models | | | |
| Claude-Haiku-4.5 | 0.661 | 0.256 | 0.258 |
| Doubao-Seed-1.8 | 0.646 | 0.252 | 0.275 |
| Gemini-3.5-flash | 0.627 | 0.324 | 0.299 |
| GPT-5-mini | 0.645 | **0.334** | 0.303 |
| GPT-5-nano | 0.496 | 0.176 | 0.211 |
| Grok-4.3 | 0.665 | 0.231 | 0.320 |
| Kimi-K2.5 | 0.637 | 0.210 | 0.232 |
| MiMo-v2.5 | 0.664 | 0.210 | 0.280 |
| Step-3.7-flash | **0.676** | 0.237 | 0.318 |
| Open-source Models | | | |
| Deepseek-R1-70B | 0.538 | 0.132 | 0.176 |
| Deepseek-R1-1.5B | 0.132 | -0.034 | 0.012 |
| Gemma3-27B | 0.612 | 0.141 | 0.242 |
| Gemma3-4B | 0.446 | 0.061 | 0.063 |
| Gemma3-1B | 0.105 | 0.000 | 0.012 |
| Glm4-9B | 0.481 | 0.011 | 0.134 |
| GPT-oss-120B | 0.609 | 0.147 | **0.322** |
| GPT-oss-20B | 0.555 | 0.143 | 0.300 |
| InternLM2.5-7B | 0.401 | 0.006 | 0.041 |
| InternLM2.5-1.8B | -0.022 | 0.055 | 0.017 |
| Llama3.1-70B | 0.617 | 0.258 | 0.258 |
| Llama3.1-8B | 0.325 | -0.021 | 0.065 |
| Llama3.2-1B | 0.019 | 0.006 | 0.020 |
| Phi4-14B | 0.639 | 0.158 | 0.262 |
| Qwen2-7B | 0.403 | 0.044 | 0.090 |
| Qwen2-1.5B | 0.011 | 0.027 | -0.006 |
| Qwen2-0.5B | 0.001 | -0.006 | 0.018 |
| Qwen2.5-7B | 0.426 | 0.019 | 0.085 |
| Qwen2.5-1.5B | 0.070 | 0.025 | -0.014 |
| Qwen3-14B | 0.564 | 0.078 | 0.229 |
| Qwen3-8B | 0.563 | 0.057 | 0.160 |
| Qwen3-4B | 0.523 | 0.197 | 0.224 |
| Qwen3-0.6B | 0.078 | -0.008 | 0.070 |
| Basic Baselines | | | |
| Random Stratified | 0.017 | 0.025 | 0.049 |
| Majority Class | 0.000 | 0.000 | 0.000 |
| Heuristic Rule | 0.176 | -0.084 | 0.046 |
| Attribution Rule | 0.473 | -0.252 | 0.158 |
| mBERT Classifier | 0.186 | 0.025 | 0.038 |

**(b) Blame judgment**

| Model | Vignette agent relation | | |
|---|---|---|---|
| | General | Vicarious | Commanding |
| Closed-source Models | | | |
| Claude-Haiku-4.5 | 0.726 | 0.450 | 0.506 |
| Doubao-Seed-1.8 | 0.715 | 0.495 | 0.495 |
| Gemini-3.5-flash | 0.686 | 0.407 | 0.579 |
| GPT-5-mini | 0.719 | **0.500** | 0.537 |
| GPT-5-nano | 0.557 | 0.423 | 0.452 |
| Grok-4.3 | **0.757** | 0.484 | **0.651** |
| Kimi-K2.5 | 0.715 | 0.498 | 0.546 |
| MiMo-v2.5 | 0.660 | 0.461 | 0.489 |
| Step-3.7-flash | 0.700 | 0.459 | 0.544 |
| Open-source Models | | | |
| Deepseek-R1-70B | 0.544 | 0.314 | 0.331 |
| Deepseek-R1-1.5B | 0.068 | -0.018 | 0.003 |
| Gemma3-27B | 0.631 | 0.332 | 0.372 |
| Gemma3-4B | 0.325 | 0.102 | 0.109 |
| Gemma3-1B | 0.051 | -0.011 | 0.021 |
| Glm4-9B | 0.437 | 0.264 | 0.215 |
| GPT-oss-120B | 0.699 | 0.443 | 0.460 |
| GPT-oss-20B | 0.641 | 0.391 | 0.470 |
| InternLM2.5-7B | 0.355 | 0.166 | 0.071 |
| InternLM2.5-1.8B | -0.006 | 0.020 | -0.017 |
| Llama3.1-70B | 0.702 | 0.436 | 0.471 |
| Llama3.1-8B | 0.344 | 0.216 | 0.233 |
| Llama3.2-1B | 0.009 | 0.002 | 0.016 |
| Phi4-14B | 0.677 | 0.402 | 0.437 |
| Qwen2-7B | 0.398 | 0.191 | 0.154 |
| Qwen2-1.5B | 0.019 | 0.032 | 0.005 |
| Qwen2-0.5B | -0.001 | -0.011 | 0.027 |
| Qwen2.5-7B | 0.454 | 0.223 | 0.151 |
| Qwen2.5-1.5B | 0.089 | 0.018 | 0.005 |
| Qwen3-14B | 0.568 | 0.325 | 0.358 |
| Qwen3-8B | 0.537 | 0.309 | 0.263 |
| Qwen3-4B | 0.550 | 0.425 | 0.425 |
| Qwen3-0.6B | 0.057 | 0.009 | 0.046 |
| Basic Baselines | | | |
| Random Stratified | -0.011 | -0.109 | -0.048 |
| Majority Class | 0.000 | 0.000 | 0.000 |
| Heuristic Rule | 0.156 | 0.027 | 0.219 |
| Attribution Rule | 0.506 | 0.182 | 0.211 |
| mBERT Classifier | 0.101 | -0.018 | 0.004 |

* The best results are highlighted in **bold** and the second-best results are underlined. LLM cells are averages over N, CoT, SR, and TG. Spearman's $\rho$ is reported on its original $[-1, 1]$ scale and rounded to three decimal places. Spearman's $\rho$ uses vignette data only.

**Attribution-dimension values.** Table 19 shows that accuracy also varies across dimension-value groups. For GPT-5-mini, responsibility accuracy is 62.82% when intention is present versus 42.01% when absent, and 65.78% versus 47.29% for foreknowledge. The pattern is not universal: Grok-4.3 has higher blame accuracy when intention is absent (61.89%) than when present (59.23%). These observational differences concern agreement within groups that can differ in scenario composition and human-label distributions. They do not measure a dimension's causal effect or show that a model necessarily assigns higher judgments when the dimension is present. This distinction motivates the controlled representation interventions in the main analysis.

Table 19: Accuracy of tested models by attribution dimension values.

| | **(a) Responsibility judgment** | | | | | | | | | | **(b) Blame judgment** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Intent. | | Volunt. | | Foreknow. | | Control. | | Obligation | | Intent. | | Volunt. | | Foreknow. | | Control. | | Obligation | |
| Model | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Closed-source Models | | | | | | | | | | | | | | | | | | | | |
| Claude-Haiku-4.5 | 44.55 | 58.68 | 53.42 | 66.64 | 46.43 | 62.27 | **48.44** | 58.47 | 56.89 | 55.00 | 45.58 | 53.46 | 49.73 | 54.69 | 43.45 | 58.55 | 46.02 | 54.62 | 56.23 | 51.69 |
| Doubao-Seed-1.8 | 41.92 | 57.62 | 54.33 | 67.43 | 43.53 | 62.06 | 40.53 | 60.93 | 56.84 | 52.10 | 49.60 | 54.40 | 57.48 | 57.03 | 45.75 | 59.83 | 47.69 | 56.26 | 55.51 | 51.88 |
| Gemini-3.5-flash | 41.10 | 56.30 | 52.86 | 63.98 | 40.59 | 62.52 | 41.40 | 58.49 | 54.46 | 53.86 | 49.98 | 55.09 | 54.00 | 58.12 | 50.96 | 56.04 | 50.15 | 55.70 | 55.45 | 54.17 |
| GPT-5-mini | 42.01 | **62.82** | **59.22** | **70.13** | 47.29 | **65.78** | 48.07 | 62.32 | 50.31 | **62.80** | 33.13 | 54.33 | 56.19 | 55.72 | 35.50 | 59.89 | 36.98 | 54.87 | 44.01 | 48.92 |
| GPT-5-nano | 44.23 | 53.79 | 56.04 | 60.17 | 44.88 | 56.79 | 44.86 | 54.89 | 53.79 | 53.19 | 35.25 | 51.53 | 52.86 | 54.33 | 34.36 | 58.39 | 37.09 | 52.65 | 45.62 | 48.85 |
| Grok-4.3 | **49.50** | 58.56 | 56.60 | 66.63 | **48.43** | 62.98 | 46.63 | 61.75 | **61.35** | 55.34 | 61.89 | **59.23** | **60.27** | **62.87** | 59.11 | 60.91 | 58.49 | **60.99** | **61.86** | **59.36** |
| Kimi-K2.5 | 42.58 | 60.27 | 55.96 | 68.38 | 44.88 | 64.82 | 46.14 | 60.83 | 57.11 | 54.99 | 44.86 | 57.57 | 57.09 | 59.63 | 45.18 | **61.96** | 46.38 | 58.39 | 53.79 | 54.91 |
| MiMo-v2.5 | 44.64 | 61.24 | 57.08 | 68.57 | 46.55 | 65.73 | 46.74 | **62.53** | 57.71 | 57.71 | 41.81 | 52.80 | 56.11 | 52.51 | 38.08 | 60.40 | 42.05 | 54.18 | 50.88 | 50.57 |
| Step-3.7-flash | 48.48 | 57.80 | 55.30 | 70.11 | 45.39 | 63.13 | 47.38 | 59.71 | 58.95 | 53.31 | **64.15** | 54.73 | 54.62 | 59.98 | 57.85 | 57.57 | **61.58** | 55.25 | 60.71 | 56.87 |
| Open-source Models | | | | | | | | | | | | | | | | | | | | |
| Deepseek-R1-70B | 40.68 | 49.29 | 46.66 | 53.16 | 43.21 | 50.20 | 41.36 | 50.22 | 50.59 | 47.27 | 39.79 | 44.12 | 40.87 | 43.83 | 38.38 | 47.06 | 38.83 | 45.33 | 43.61 | 43.34 |
| Deepseek-R1-1.5B | 26.45 | 29.38 | 32.11 | 27.56 | 28.33 | 28.74 | 27.40 | 29.25 | 26.41 | 30.60 | 17.58 | 27.81 | 27.86 | 26.39 | 20.68 | 28.63 | 18.49 | 28.67 | 24.41 | 24.59 |
| Gemma3-27B | 34.43 | 56.30 | 54.97 | 63.23 | 43.15 | 56.45 | 41.34 | 55.44 | 41.25 | 56.96 | 21.74 | 52.89 | 52.66 | 53.13 | 27.84 | 58.58 | 29.29 | 52.51 | 39.38 | 46.48 |
| Gemma3-4B | 28.05 | 49.17 | 48.01 | 53.53 | 33.18 | 52.38 | 34.78 | 48.30 | 33.57 | 49.52 | 10.78 | 44.87 | 43.93 | 49.49 | 19.92 | 48.73 | 17.61 | 45.34 | 29.69 | 35.89 |
| Gemma3-1B | 23.67 | 34.81 | 35.73 | 35.72 | 27.60 | 35.36 | 24.88 | 35.80 | 26.78 | 34.86 | 17.16 | 30.74 | 31.82 | 30.26 | 20.32 | 32.74 | 18.84 | 31.58 | 24.20 | 28.55 |
| Glm4-9B | 36.76 | 43.24 | 45.35 | 44.33 | 39.59 | 43.06 | 38.95 | 42.90 | 41.45 | 43.26 | 26.79 | 40.75 | 42.75 | 39.79 | 28.77 | 44.01 | 30.56 | 40.33 | 34.20 | 38.11 |
| GPT-oss-120B | 43.71 | 58.23 | 58.07 | 67.26 | 47.66 | 60.05 | 48.04 | 57.82 | 53.09 | 52.16 | 37.69 | 53.14 | 54.51 | 54.52 | 40.00 | 56.64 | 37.35 | 55.50 | 46.02 | 51.57 |
| GPT-oss-20B | 44.41 | 54.76 | 54.97 | 60.89 | 45.31 | 57.83 | 47.01 | 54.77 | 49.34 | 54.08 | 40.50 | 47.50 | 50.98 | 46.58 | 37.77 | 52.69 | 40.03 | 48.77 | 45.91 | 46.90 |
| InternLM2.5-7B | 32.67 | 38.88 | 40.74 | 40.61 | 34.57 | 39.45 | 36.67 | 37.37 | 34.02 | 37.54 | 23.78 | 32.69 | 36.44 | 28.35 | 22.86 | 36.87 | 27.44 | 31.64 | 29.13 | 31.02 |
| InternLM2.5-1.8B | 17.94 | 35.94 | 35.17 | 38.08 | 23.99 | 37.12 | 21.04 | 36.83 | 25.68 | 33.90 | 10.47 | 30.72 | 26.41 | 35.09 | 17.59 | 31.41 | 11.98 | 32.59 | 22.40 | 25.22 |
| Llama3.1-70B | 43.03 | 56.15 | 55.17 | 60.68 | 46.04 | 58.31 | 44.38 | 57.36 | 55.33 | 55.02 | 47.23 | 51.83 | 49.29 | 53.22 | 45.94 | 54.76 | 47.78 | 52.13 | 50.98 | 52.72 |
| Llama3.1-8B | 33.28 | 40.09 | 41.77 | 43.39 | 35.14 | 40.94 | 35.89 | 39.54 | 34.96 | 39.57 | 27.64 | 35.61 | 38.48 | 34.20 | 26.63 | 39.52 | 31.44 | 34.35 | 30.62 | 36.27 |
| Llama3.2-1B | 16.40 | 13.14 | 12.36 | 10.94 | 17.90 | 10.52 | 16.98 | 12.27 | 14.63 | 13.68 | 17.09 | 15.08 | 12.42 | 12.96 | 19.01 | 12.50 | 15.17 | 16.00 | 15.61 | 15.01 |
| Phi4-14B | 46.13 | 54.40 | 53.62 | 60.24 | 46.08 | 57.58 | 47.80 | 54.66 | 56.09 | 51.33 | 46.01 | 48.23 | 49.69 | 47.87 | 41.45 | 53.40 | 46.32 | 48.35 | 48.58 | 48.02 |
| Qwen2-7B | 32.13 | 40.03 | 40.54 | 41.30 | 35.04 | 40.30 | 35.72 | 39.03 | 34.34 | 38.22 | 23.38 | 37.54 | 37.66 | 35.11 | 25.52 | 40.73 | 27.50 | 36.94 | 31.50 | 34.65 |
| Qwen2-1.5B | 22.76 | 23.65 | 24.80 | 21.72 | 24.76 | 22.13 | 24.02 | 23.01 | 21.97 | 23.12 | 19.13 | 23.49 | 21.75 | 21.27 | 21.95 | 22.41 | 19.18 | 24.06 | 19.64 | 21.27 |
| Qwen2-0.5B | 25.05 | 11.99 | 9.22 | 8.25 | 21.48 | 10.37 | 22.68 | 11.42 | 23.42 | 13.77 | 31.11 | 17.44 | 14.54 | 14.60 | 30.93 | 12.56 | 28.20 | 17.36 | 24.77 | 18.52 |
| Qwen2.5-7B | 34.57 | 45.34 | 46.98 | 48.15 | 36.86 | 47.28 | 40.61 | 43.28 | 39.86 | 43.15 | 28.01 | 40.89 | 41.77 | 38.36 | 29.33 | 44.38 | 32.33 | 39.98 | 36.06 | 39.16 |
| Qwen2.5-1.5B | 28.62 | 19.64 | 19.08 | 18.38 | 23.89 | 20.64 | 25.61 | 20.10 | 26.78 | 20.37 | 32.15 | 24.46 | 26.33 | 23.09 | 29.18 | 24.46 | 31.27 | 23.94 | 28.44 | 25.87 |
| Qwen3-14B | 36.81 | 56.49 | 54.29 | 61.58 | 43.39 | 57.80 | 41.83 | 56.45 | 50.02 | 52.95 | 32.26 | 45.93 | 49.29 | 45.48 | 30.66 | 52.50 | 35.59 | 45.75 | 39.96 | 44.57 |
| Qwen3-8B | 35.63 | 53.56 | 51.55 | 59.99 | 38.65 | 57.53 | 39.24 | 54.13 | 44.55 | 52.10 | 27.21 | 47.68 | 49.02 | 47.54 | 28.98 | 53.56 | 27.74 | 50.20 | 39.77 | 42.69 |
| Qwen3-4B | 42.13 | 50.89 | 49.32 | 58.31 | 37.37 | 58.62 | 40.05 | 53.54 | 49.51 | 50.06 | 45.11 | 44.74 | 46.20 | 49.93 | 39.02 | 50.41 | 43.53 | 45.68 | 45.24 | 42.69 |
| Qwen3-0.6B | 28.42 | 27.56 | 29.85 | 27.55 | 27.78 | 27.83 | 26.89 | 28.37 | 29.10 | 29.03 | 27.44 | 27.01 | 24.80 | 29.81 | 27.59 | 26.71 | 25.65 | 28.07 | 27.56 | 27.13 |
| Basic Baselines | | | | | | | | | | | | | | | | | | | | |
| Random Stratified | 20.27 | 35.65 | 34.66 | 38.86 | 27.66 | 34.60 | 25.10 | 35.06 | 26.15 | 34.56 | 23.12 | 29.41 | 27.27 | 30.95 | 27.74 | 27.33 | 25.62 | 28.73 | 28.67 | 28.79 |
| Majority Class | 14.75 | 53.91 | 49.28 | 62.91 | 24.87 | 59.31 | 20.70 | 56.35 | 25.82 | 51.03 | 12.31 | 42.87 | 36.52 | 49.60 | 34.69 | 32.82 | 22.45 | 40.78 | 37.89 | 38.53 |
| Heuristic Rule | 33.39 | 16.59 | 21.14 | 17.60 | 24.06 | 18.92 | 28.69 | 16.88 | 28.85 | 15.81 | 50.66 | 15.31 | 22.10 | 11.87 | 36.21 | 16.05 | 46.69 | 12.88 | 33.61 | 25.32 |
| Attribution Rule | 32.49 | 36.56 | 5.09 | 55.71 | 42.05 | 29.21 | 25.78 | 41.35 | 53.28 | 22.97 | 58.90 | 42.19 | 25.86 | 57.90 | **60.80** | 34.21 | 49.31 | 45.87 | 61.19 | 37.16 |
| mBERT Classifier | 24.36 | 50.59 | 44.59 | 57.14 | 30.85 | 55.65 | 27.74 | 54.71 | 31.91 | 47.98 | 27.66 | 33.18 | 31.17 | 39.75 | 29.14 | 34.03 | 20.13 | 39.83 | 45.61 | 30.77 |

* The best results are highlighted in **bold** and the second-best results are underlined. LLM cells are averages over N, CoT, SR, and TG. Accuracy is reported as a percentage and rounded to two decimal places. Accuracy pools vignette and reality data; questions with unspecified or inapplicable dimension labels are excluded from the corresponding 0/1 columns.

# F Additional Decodability Analyses

The main decodability analysis asks whether theory-derived dimensions can be recovered from hidden states. Above-chance accuracy alone does not establish what information a probe uses or whether that information is specific to the final judgment stage. We therefore examine controls for alternative predictive signals and compare two token positions.

## F.1 Decodability Sanity Checks

The controls target three distinct concerns about the probing measurement. Shuffling training labels tests dependence on the original dimension–representation association; random projection tests how much decodability is retained after compressing the representation to 64 features; and scenario-text TF-IDF measures the information available from lexical surface cues. All use the same data splits, classifier family, and balanced-accuracy metric as the main probes (Appendix D.3): eight folds holding out one scenario type at a time for *Vignette*, and ten random folds for *Reality*. Together with diversified vignette wording, holding out entire scenario types reduces opportunities to exploit familiar event settings; the controls further assess dependence on the tested representation and surface features.

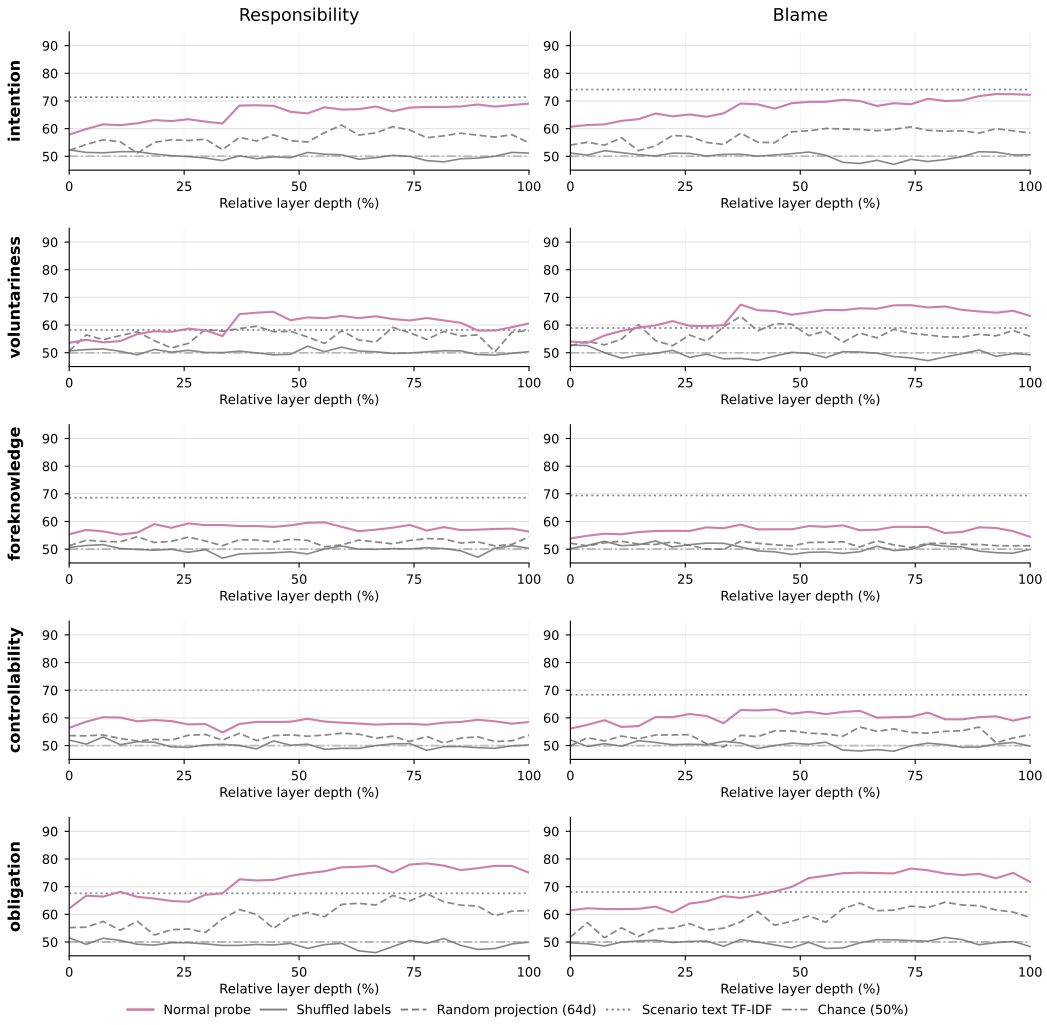



Figure 10: Sanity-check results at the label-token prediction step for DeepSeek-R1-1.5B, including shuffled-label, 64-dimensional random-projection, scenario-text lexical, and chance controls.

Figures 10–14 report the results for all five backbones. The original probes consistently outperform the shuffled-label control, which remains near chance, and generally outperform the 64-dimensional random-projection control. For Llama-3.1-8B, Qwen3-4B, and Qwen3-8B, they also exceed the scenario-text lexical control across most layers. These comparisons support the usefulness of the hidden-state measurements beyond the tested controls, without implying that all lexical or contextual correlations have been removed. In particular, the lexical comparison is not uniformly favorable across every model and layer.

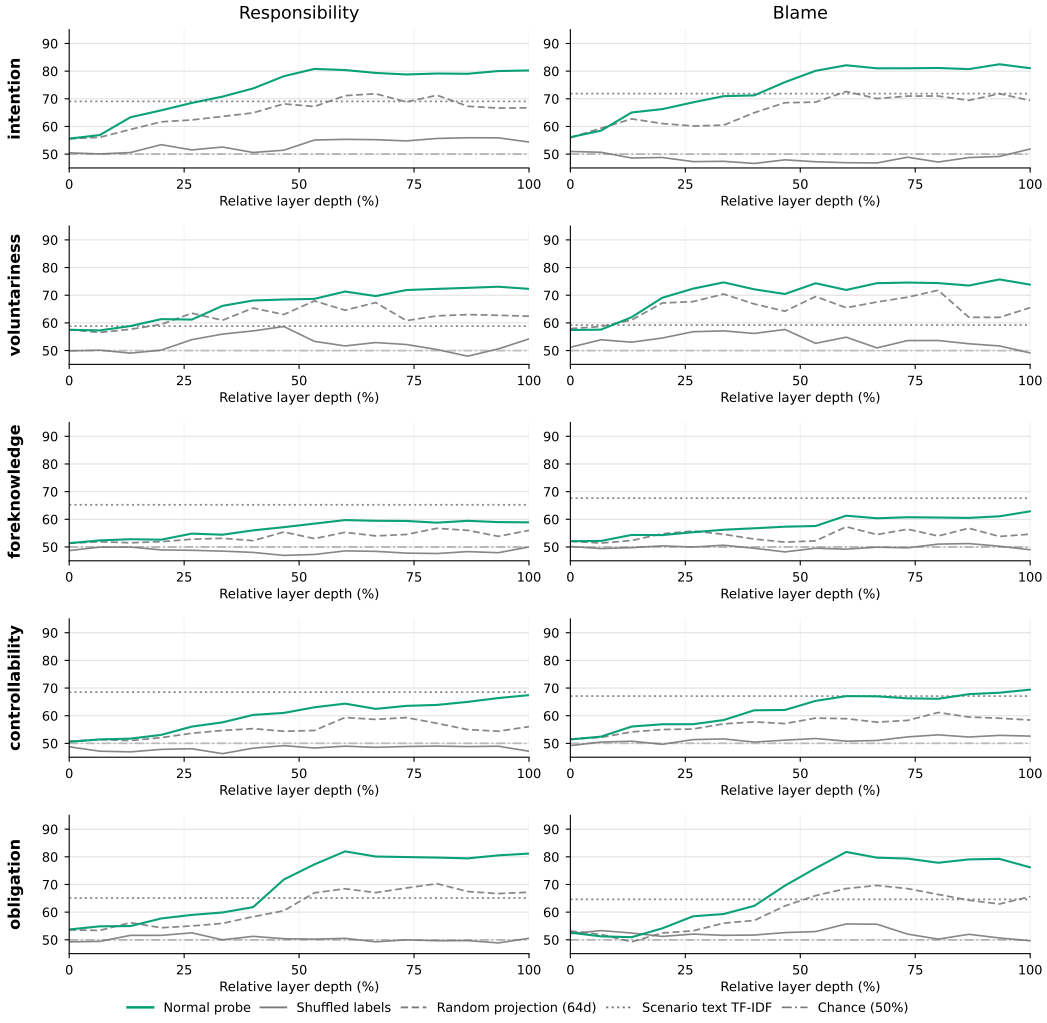



Figure 11: Sanity-check results at the label-token prediction step for Llama-3.2-1B, including shuffled-label, 64-dimensional random-projection, scenario-text lexical, and chance controls.

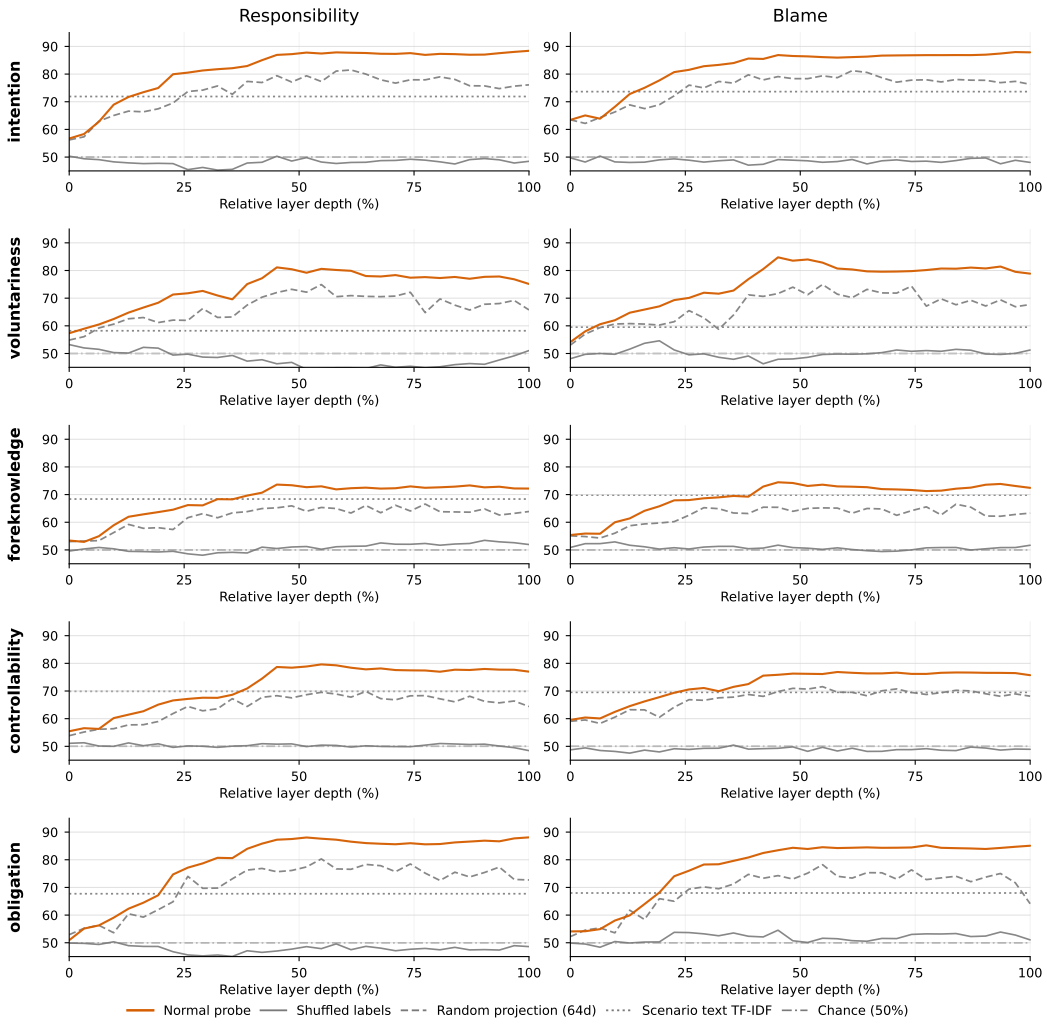



Figure 12: Sanity-check results at the label-token prediction step for Llama-3.1-8B, including shuffled-label, 64-dimensional random-projection, scenario-text lexical, and chance controls.

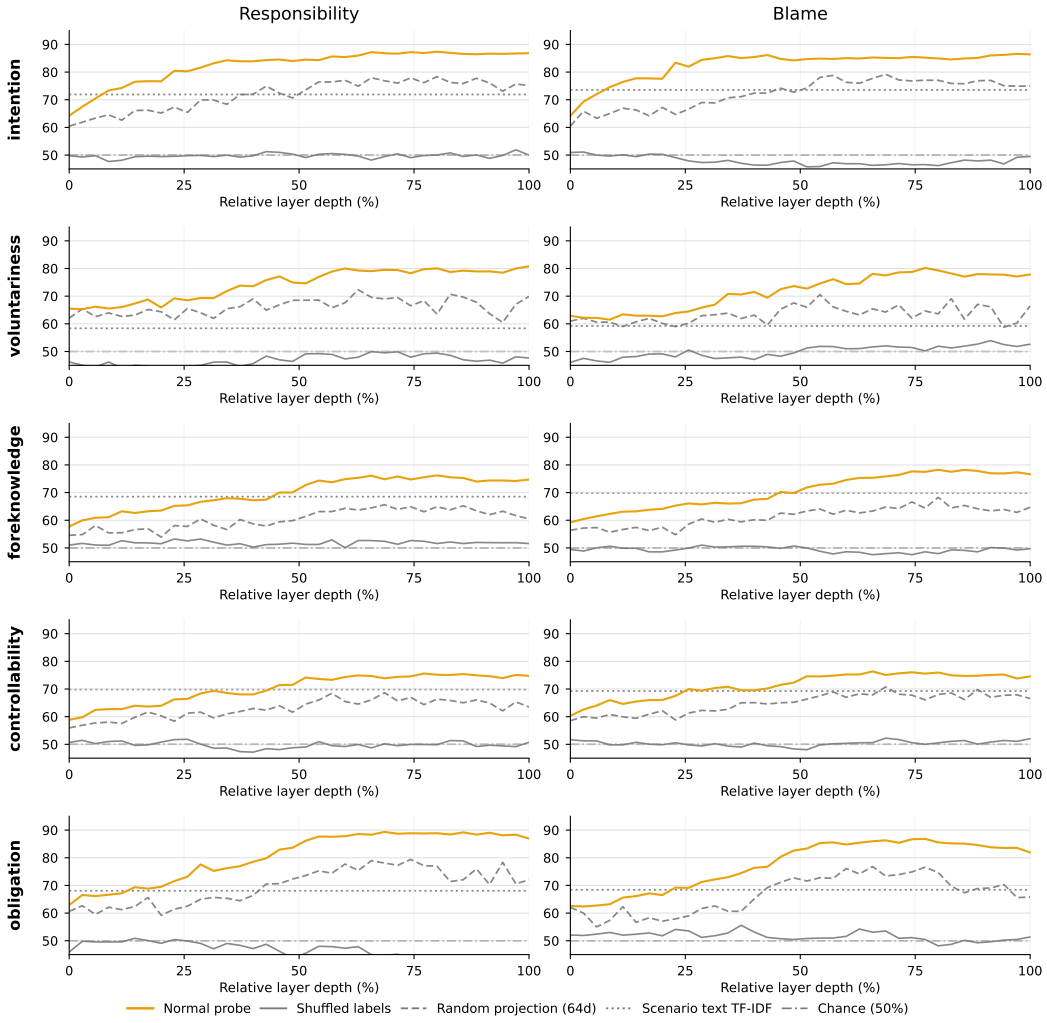



Figure 13: Sanity-check results at the label-token prediction step for Qwen3-4B, including shuffled-label, 64-dimensional random-projection, scenario-text lexical, and chance controls.

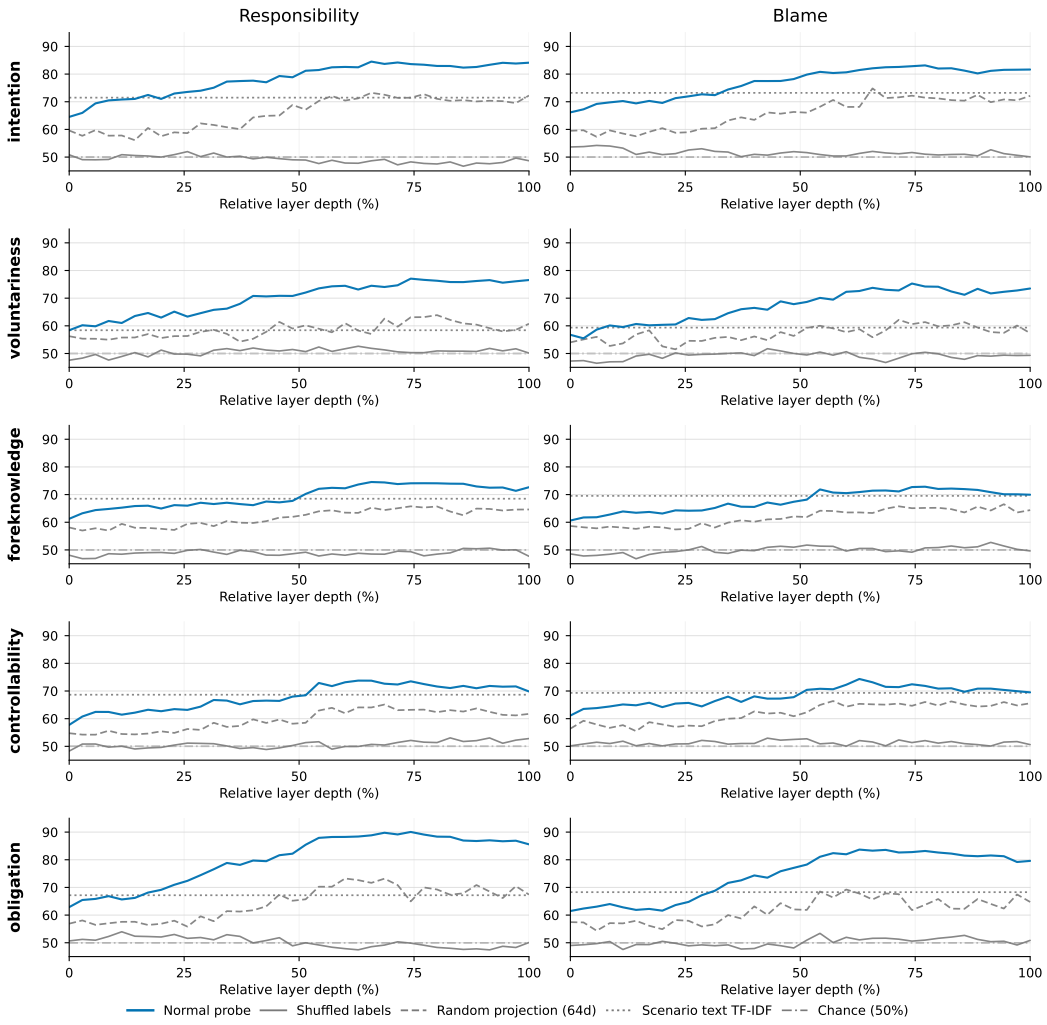



Figure 14: Sanity-check results at the label-token prediction step for Qwen3-8B, including shuffled-label, 64-dimensional random-projection, scenario-text lexical, and chance controls.

### F.2 Decodability at the Last Input Token

The main probes read the hidden state used to predict the final judgment label, where information from the prompt and preceding generated reasoning is available. A possible concern is that decodability at this position reflects only the immediate answer-selection stage. To examine whether attribution information is accessible earlier, we instead extract the hidden state at the last token of the input prompt, before the model generates its response, and train the same type of probes. Figure 15 compares this last-input-token position with the label-token prediction step.

Attribution dimensions are already strongly decodable at the earlier position. Averaged over the displayed models, dimensions, and layers, balanced accuracy is 73.8% at the last input token and 71.1% at the label-token prediction step. Four of the five models have higher average decodability at the input position, with the largest difference in DeepSeek-R1-1.5B. Qwen3-8B is the main exception: its later position improves decoding of voluntariness, foreknowledge, and controllability. Llama-3.2-1B also improves on voluntariness at the later position.

At both positions, decodability generally rises from early layers and peaks in the middle or later layers. Intention and obligation remain among the most decodable dimensions, and the layerwise curves often have similar shapes. This comparison shows that the measured attribution information is not confined to the final label-prediction step. It does not imply that generating reasoning uniformly increases decodability, or that the two positions implement the same computation. Whether an earlier representation can also influence the final judgment is examined separately in Appendix G.1.

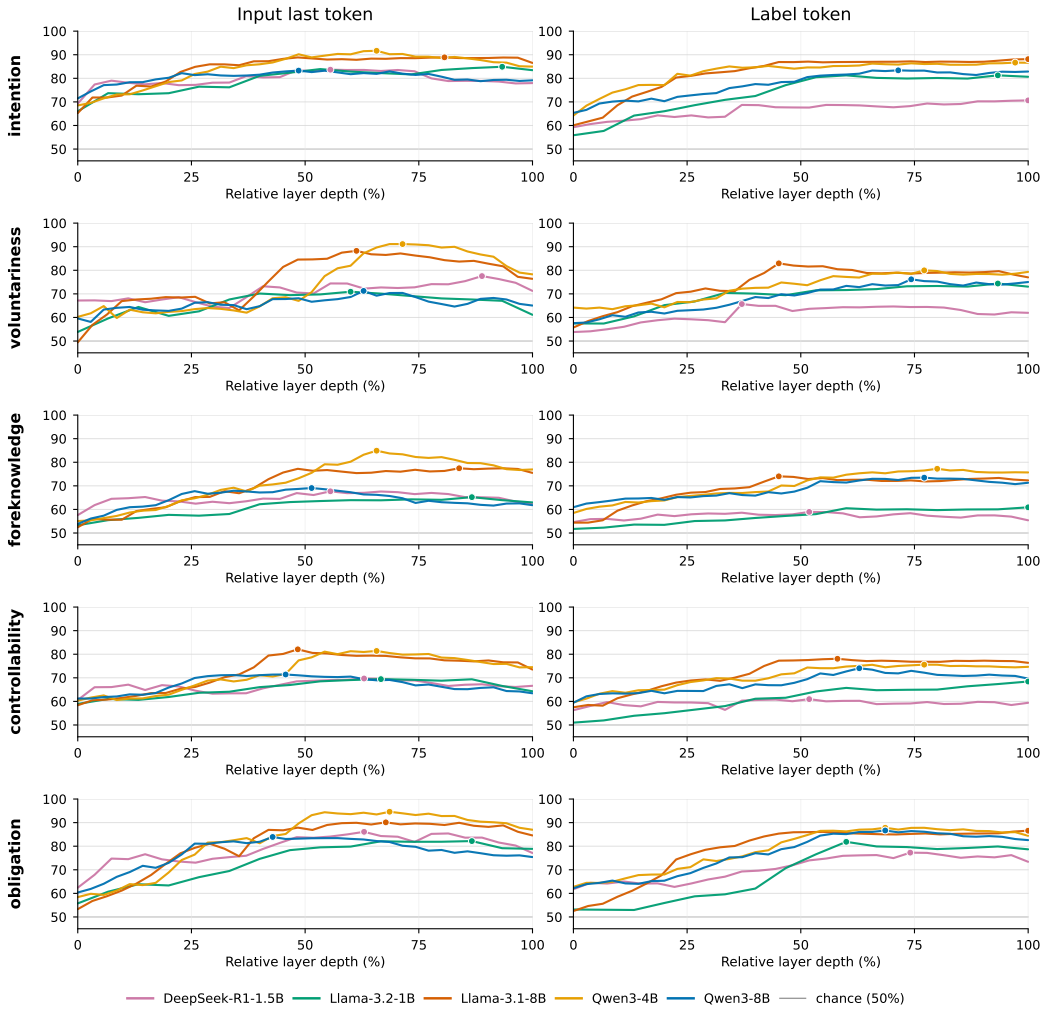



Figure 15: Layerwise decodability at the last input token and the label-token prediction step. Layer depth is normalized within each model; dots indicate peak decodability.

# G Additional Intervention Analyses

Decodability establishes that a dimension is accessible to a linear readout, but does not by itself show that changing the corresponding representation affects the model's judgment. The intervention analyses address this distinction. Complementing the main directional-test heatmap in Figure 6, we test an earlier intervention position and examine how directional changes vary with perturbation strength.

## G.1 Intervention at the Last Input Token

The main intervention edits a hidden state at the label-token prediction step. This position directly tests whether the learned attribution directions can influence the final judgment, but its proximity to the output leaves open whether the effect is restricted to local answer selection. The earlier-token comparison addresses this concern by moving both probe training and intervention to the last token of the input prompt.

For each prompt, we extract the last-input-token hidden state and train attribution-dimension probes at that position. We then perturb that state along the corresponding probe-derived direction and allow autoregressive generation to continue until the model produces its final attribution label. The selected layers, strengths, directional hypotheses, random-direction controls, and shuffled-label probe-direction controls follow the main intervention setup. Each comparison uses the same feasible signs $\mathcal{E}_i$ determined by the unedited output: a question succeeds only if every tested sign strictly aligns, as in Equation 5; ties count as failures and all questions remain in $N$. We use the same one-sided exact binomial test with threshold $1/2$, requiring $p_c < 0.01$ for all three comparisons (Appendix D.5). Unlike the main one-step label prediction, the earlier edit must propagate through subsequent generation before the final judgment is observed.

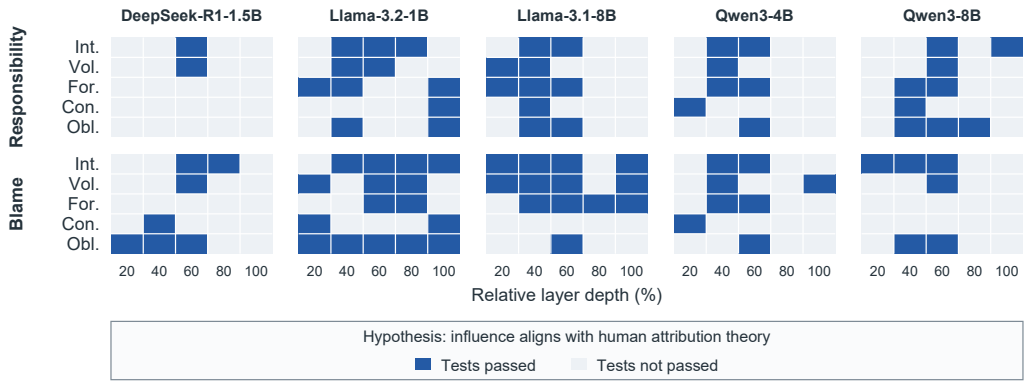



Figure 16: Layerwise influence of attribution dimensions when interventions are applied at the last input token, with $\alpha = 8$. Blue cells pass all three question-level directional-success tests ($p < 0.01$); gray cells do not. Rows denote dimensions; 100% denotes the final layer. Int., Vol., For., Con., and Obl. denote intention, voluntariness, foreknowledge, controllability, and obligation.

Figure 16 reports the earlier-position results at $\alpha = 8$, using the same visual encoding as Figure 6. Theory-aligned effects occur in several models and intermediate layers, particularly in the Llama family, but are unevenly distributed across models, dimensions, and layers. Together with Appendix F.2, this separates two questions: attribution information can already be read out before response generation, whereas the directional tests assess whether an edit at that position affects the eventual judgment.

The surviving effects show that influence along these probe-derived directions is not confined to an edit immediately before the answer label. However, the model- and layer-dependent pattern limits claims about a broadly stable mechanism throughout generation. An earlier local perturbation may be attenuated or transformed during subsequent computation; the present comparison measures its eventual effect rather than identifying the specific process responsible for that attenuation.

### G.2 Sensitivity to Intervention Strength

The binary tests assess question-level directional success, but do not quantify the magnitude of judgment changes. We complement them with a descriptive sensitivity analysis, varying $\alpha \in \{1, 2, 4, 8, 16\}$. For a fixed model and task, let $\mathcal{I}_{l,k}$ denote the evaluated held-out questions at layer $l$ and dimension $k$. Using the same feasible signs $\mathcal{E}_i$ determined by the unedited output and the baseline-relative differences in Equation 4, we first calculate

$$D_{l,k}(\alpha) = \frac{1}{|\mathcal{I}_{l,k}|} \sum_{i \in \mathcal{I}_{l,k}} \frac{1}{|\mathcal{E}_i|} \sum_{\eta \in \mathcal{E}_i} \Delta^{\text{base}}_{i,\eta}. \tag{12}$$

An increase under a positive intervention and a decrease under a negative intervention contribute positively; unchanged judgments contribute zero. Each curve in Figure 17 then gives the equal-weight mean across the evaluated layers $\mathcal{L}$ (Table 11) and the five dimensions $\mathcal{K}$:

$$D(\alpha) = \frac{1}{|\mathcal{L}|\,|\mathcal{K}|} \sum_{l \in \mathcal{L}} \sum_{k \in \mathcal{K}} D_{l,k}(\alpha). \tag{13}$$

This continuous measure first averages the feasible signed changes within each question, then averages across questions and layer–dimension settings. The binary test instead assigns one success only when every feasible sign strictly satisfies the criterion. Thus, $D(\alpha) > 0$ does not imply a directional-success rate above $1/2$ or that the directional tests pass. A value near zero indicates little net directional change, including unchanged judgments or changes that cancel.

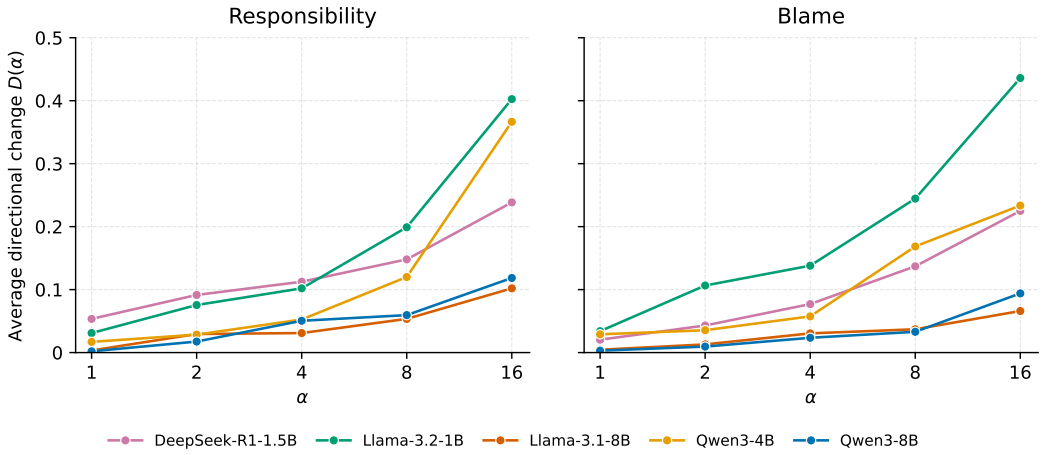



Figure 17: Average signed judgment change $D(\alpha)$ at different intervention strengths, shown separately for responsibility and blame. Each model's curve averages $D_{l,k}(\alpha)$ equally across evaluated layers and dimensions. Positive values indicate a net theory-aligned change, not a directional-success rate or a significance result.

Figure 17 shows that the average directional change generally increases as $\alpha$ grows from 1 to 16 for both tasks. Over this full range, Llama-3.2-1B and Qwen3-4B show the largest increases, while the 8B models and DeepSeek-R1-1.5B show smaller positive increases. The ordering of model sensitivities can differ at individual strengths. These curves describe how average change magnitude varies with intervention strength; they do not establish that the directional-success tests pass at every strength. A baseline-relative magnitude alone also does not establish specificity to an attribution dimension, which is assessed by the matched-control tests.

Cross-model differences should be interpreted within this intervention setup. The same numerical $\alpha$ specifies the norm of a unit-direction perturbation, but need not represent the same relative change in hidden-state scale across backbones. The observed sensitivity therefore does not by itself establish that smaller models have intrinsically simpler attribution mechanisms or that larger models are generally more robust.

# H Discussion and Limitations

## H.1 Dataset Coverage and Human-Judgment Reference

The two subsets address complementary requirements. Controlled *Vignette* scenarios make attribution dimensions and agent relationships explicit enough to vary systematically, supporting analysis that would be difficult with unconstrained narratives alone. The *Reality* subset introduces more ambiguous situations grounded in everyday social conflicts. Together, they broaden the evaluation beyond a single form of evidence, but neither subset exhausts social attribution. Vignettes remain constructed scenarios, and reconstructed narratives omit details of the original interactions. The benchmark also focuses on negative outcomes and responsibility/blame judgments; the findings do not directly establish performance on positive credit attribution or other forms of social explanation.

The *Reality* subset draws from a particular online community whose posting practices and selection of interpersonal conflicts shape the source material. Its cultural and platform coverage is therefore limited. Providing translated versions does not broaden the underlying cultural coverage.

The reference labels reflect the independently elicited judgments of two psychology experts under the supplied guidelines. They provide a consistent basis for comparison, rather than a population-wide consensus or a uniquely correct moral verdict. Keeping only questions with matching judgments avoids forcing consensus on disputed cases, but may preferentially retain clearer or less contentious items. Similar aggregate distributions before and after filtering would not rule out all such selection effects. Model accuracy should therefore be read as agreement with this defined human reference, within the dataset's coverage.

## H.2 Interpretation of Probing and Intervention Evidence

The framework distinguishes three evidential questions: whether models agree with human judgments, whether attribution dimensions are linearly decodable, and whether edits along probe-derived directions influence judgments. This separation makes it possible to study intermediate variables in social reasoning rather than treating final-label agreement as the sole indicator of capability. Decodability alone, however, establishes that information is accessible to a particular classifier; it does not establish that the model uses that information in the same way as a human reasoner (Hewitt & Liang, 2019; Belinkov, 2022).

Several design choices strengthen the interpretation of the probing results. Vignette phrasing, names, and genders are varied, and dimension labels are not explicitly supplied in the scenario text. Its eight-fold leave-one-scenario-type-out evaluation holds each type out across all agent relations, testing generalization to unseen event settings and reducing reliance on recurring scenario content. Shuffled-label, random-projection, and lexical controls test complementary explanations for probe performance. Matched random-direction and shuffled-label interventions further test whether the observed judgment shifts depend on the learned dimension direction. These safeguards provide evidence beyond an unconstrained probe fit, while leaving open representational explanations that the controls do not directly test.

The dimensions can be correlated within narratives, and residualizing one probe weight against other probe weights removes only shared linear components captured by those probes. It does not establish complete independence among the underlying constructs or isolate every feature affected by an edit. Interventions at the label-token prediction step also demonstrate influence at a particular stage of output production; they do not recover the entire process by which the model arrived at its judgment. The earlier-token comparisons help characterize this position dependence, but do not establish equivalence to the sequential cognitive architectures of attribution theory. Dimension interactions and nonlinear representations remain incompletely characterized by the present approach.

The internal analyses cover five backbones and specific prompts, layers, token positions, and perturbation strengths. Their patterns should not be generalized to all models or inference settings. In particular, differences among the backbones confound parameter scale, model family, and training. Explanations based on more distributed representations in larger models or information distributed across a reasoning trajectory remain hypotheses rather than established mechanisms. Likewise, failure to pass the directional tests indicates insufficient evidence under the tested setting; it does not prove that the model ignores the corresponding dimension.

### H.3 Data Ethics and Responsible Use

The study concerns judgments about fictional or reconstructed agents for research evaluation. The *Vignette* scenarios use fictional characters. For *Reality*, publicly available posts serve as sources for reconstruction rather than being reproduced as the benchmark text. The preprocessing described in Appendix B.2 removes identifying information and unrelated private details, and substitutes fictional names. These measures aim to reduce privacy risks while retaining the event structure needed for attribution analysis.

Annotators were informed of the task purpose, the annotation content, and the intended use of their work. Participation was voluntary, and annotators could withdraw without providing a reason. The task did not require disclosure of personal experiences or sensitive personal information. Because scenarios include accidents, injury, conflict, and negligence, annotators were informed that they could encounter negative content; construction sought to avoid unnecessarily graphic, humiliating, or inflammatory descriptions.

The benchmark and dimension annotations can support comparative evaluation, analysis of systematic judgment differences, and investigation of representations underlying social attribution. These research uses do not establish suitability for adjudicating real disputes. Responsibility and blame involve social, moral, and sometimes legal considerations that the benchmark cannot fully represent. The labels and model outputs should not serve as the basis for decisions about real individuals in legal proceedings, disciplinary action, hiring, insurance compensation, or other consequential assessments. Agreement with the benchmark is therefore evidence about model behavior under a specified research protocol, not authorization to replace human judgment or accountable decision procedures.