Probing Implicit Ethical Frameworks in Large Language Models

Anonymous ACL submission

Abstract

As AI systems make moral judgments—from content moderation to medical triage—their ethical consistency becomes critical, yet their internal processing remains opaque. We probe hidden representations to reveal whether normative frameworks are encoded coherently and if internal conflicts predict behavioral inconsistency. Mid-to-late layer probes show framework-specific rather than unified encodings: ethical systems exhibit asymmetric cross-framework transfer and severe miscalibration out-of-distribution. When deontological and utilitarian probes make opposing predictions on the same scenario, models under stochastic sampling produce significantly higher choice entropy—internal conflicts propagate to behavioral variability. This reveals behavioral alignment testing alone is insufficient: models maintain misaligned internal representations that surface under distribution shift, motivating probe-based auditing as a necessary complement.

1 Introduction

Understanding how LLMs encode moral reasoning is essential for alignment. While behavioral studies show LLMs can produce ethically-aligned outputs (Hendrycks et al., 2021), the underlying mechanisms remain unknown. Do models represent deontology and utilitarianism as distinct concepts, or are they entangled? Where does ethical reasoning emerge architecturally?

Figure 1 illustrates our hypothesis: ethical frameworks exist as overlapping representations in high-dimensional space rather than discrete modules. The deontology-utilitarianism-commonsense overlap motivates our investigation into how these patterns manifest across architectures.

We employ representation probing—training linear classifiers on hidden layer activations to reveal encoded information. Extracting activation patterns from ETHICS dataset scenarios (Hendrycks et al., 2021), we train classifiers to identify framework-specific signals. If classifiers reliably distinguish frameworks from internal representations, models encode those distinctions regardless of output behavior. This reveals both where ethical information concentrates across layers and how frameworks relate geometrically.

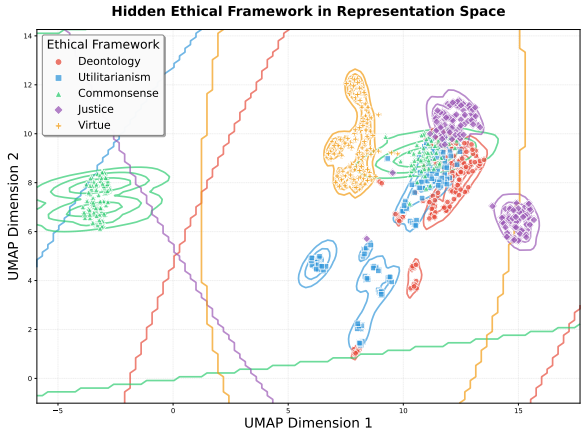


Figure 1: UMAP projection of ethical framework representations at layer 20 (of 32) in Mistral-7B-Instruct. Deontology (red), utilitarianism (blue), and commonsense morality (green) substantially overlap centrally, indicating shared representational substrate for action-oriented frameworks. Virtue ethics (orange) forms a dense, independent cluster reflecting its character-focused nature. Notably, commonsense morality exhibits two distinct subclusters, which preliminary analysis suggests correspond to scenarios emphasizing personal moral intuitions versus social norm violations—a distinction the model appears to encode separately within the broader commonsense framework. This intermediate-layer entanglement suggests behavioral fine-tuning may produce unintended cross-framework effects, as adjusting representations for one framework inevitably affects overlapping ones.

Three key findings emerge. First, ethical representations concentrate in the final third of model layers (65%+ depth) across architectures and scales, where frameworks share overlapping features that

prevent independent modification. Second, probes maintain high confidence even during catastrophic cross-framework failures, demonstrating that representational entanglement prevents models from detecting when they apply inappropriate ethical frameworks—a critical miscalibration that behavioral testing alone cannot reveal. Third, when deontological and utilitarian probes make confident opposing predictions on the same scenario, models exhibit significantly higher behavioral entropy under stochastic sampling, demonstrating that late-layer representational conflicts propagate to inconsistent ethical judgments despite identical inputs.

These findings challenge current alignment approaches. Framework entanglement reduces controllability: adjusting representations for one framework affects overlapping ones. The concentration and conflict patterns identify specific intervention points—mid-layers for framework separation, late layers for consistency. The link between internal conflict and behavioral inconsistency indicates that probes can detect instability invisible to behavioral testing alone. By mapping the geometric structure of moral reasoning, we provide empirical grounding for mechanistic alignment strategies.

2 Related Work

2.1 Probing Internal Representations

Probing techniques reveal information encoding in neural networks through simple classifiers trained on frozen representations (Alain and Bengio, 2016; Hewitt and Manning, 2019). Belinkov et al. (2017) demonstrated that layers specialize in different linguistic information, from word structure in early layers to word meaning in later ones. Diagnostic classifiers successfully extract syntactic structure (Tenney et al., 2019), factual recall (Ettinger, 2020), and abstract reasoning (Ravichander et al., 2020), revealing that LLMs encode nontrivial patterns without direct supervision.

Recent advances extend beyond supervised probing. Burns et al. (2023) introduced Contrast-Consistent Search (CCS) to extract latent knowledge without labels, while Stoehr et al. (2024) developed Contrast-Consistent Ranking (CCR) to identify semantic dimensions through unsupervised ranking. Both methods reveal models encode truth-related and semantic concepts not always reflected in outputs—highlighting potential for discovering latent ethical structures influencing behavior.

Probing offers distinct advantages for analyzing ethical representations. While steering methods actively modify behavior through fine-tuning or activation interventions (Tlaie, 2024; Choi et al., 2024), they risk overwriting representations or conflating training effects with native inclinations. Probing is purely diagnostic: it preserves model integrity, isolates representational patterns from training artifacts, and identifies optimal intervention points—making it essential for both understanding and modifying ethical reasoning in pretrained LLMs.

2.2 Moral Reasoning in Language Models

The ETHICS benchmark (Hendrycks et al., 2021) established systematic evaluation of moral judgment across five normative theories, extended through behavioral audits (Sachdeva and van Nuenen, 2025), normative consistency checks (Srivastava et al., 2022), and psychometric profiling (Li et al., 2024). Jiang et al. (2024) found moral decision patterns vary significantly with model scale and training approach. Scherrer et al. (2024) analyzed 40+ LLMs, revealing consistent utilitarian biases in proprietary models and deontological tendencies in open-source alternatives, while Yuan and Singh (2024) showed larger LLMs increasingly default to deontological responses even in scenarios with clear negative consequences.

Value alignment studies examine behavioral patterns. Neuman and Shah (2025) found six proprietary and open-weight LLMs exhibit widespread rationalist, consequentialist reasoning prioritizing harm minimization and fairness. Ganguli et al. (2023) showed constitutional AI can steer moral behavior but leaves internal representations opaque. Santurkar et al. (2023) demonstrated RLHF-tuned models reflect specific demographic moral preferences, while Hartmann et al. (2023) found political biases correlate with training data sources. Moral Foundations Theory assessments revealed consistent liberal biases across model types (Tlaie, 2024).

However, behavioral studies cannot determine whether observed patterns emerge from genuine moral representations or surface-level pattern matching. Chun and Elkins (2024) revealed authoritarian biases through systematic prompt variations, while Duan et al. (2024) proposed DeNEVIL to uncover latent ethical biases using adversarial prompts, showing underlying ethical misalignment when probed effectively.

Recent work links behavior to underlying representations. Park et al. (2024) used activation

steering to modify moral judgments, suggesting dedicated moral circuits exist. Rimsky et al. (2023) demonstrated ethical behavior can be manipulated through targeted activation interventions. Huang et al. (2024) explored dialog-based ethical persuasion, showing user-facing prompt design can shift LLM moral decisions. Nevertheless, ethical steering faces limitations: Marraffini et al. (2024) found most models rejected instrumental harm in favor of impartial beneficence—behavior diverging from both classical ethical theory and human preferences, highlighting tensions between normative fidelity and practical alignment.

Most closely related, Tlaie (2024) examined ethical alignment using classical dilemmas, Moral Foundations Questionnaire assessments, and activation steering (SARA). They found proprietary models tend toward utilitarian reasoning while open-source models exhibit more deontological patterns, with activation steering demonstrating ethical orientations could be causally manipulated toward different frameworks. Tanmay et al. (2023) applied the Defining Issues Test to measure moral reasoning stages, finding GPT-4 achieved post-conventional moral reasoning comparable to graduate students while earlier models performed at or below adult human levels, with inconsistent reasoning across dilemmas pointing to important gaps.

These studies provide valuable insights into LLM moral behavior; we complement this work by probing internal representations to reveal how ethical frameworks are encoded and interact, without behavioral confounds.

3 Methods

3.1 Ethical Framework Datasets

We use the ETHICS benchmark (Hendrycks et al., 2021), which provides moral scenarios across five philosophical frameworks: *deontological* ethics evaluates rule adherence regardless of consequences; *utilitarian* ethics judges impact on aggregate well-being; *virtue* ethics assesses character traits; *justice* ethics emphasizes fairness; and *commonsense* morality reflects social intuitions. All scenarios are converted into standardized binary-choice prompts (prompt engineering templates in Appendix A).

3.2 Probing Methodology

We train linear probes on frozen model representations to decode framework-specific moral judgments. For hidden representation $h_i^{(l)} \in R^d$ at layer l and input i, a logistic regression probe learns:

$$P(y = 1|h_i^{(l)}) = \sigma(w^T h_i^{(l)} + b) \qquad (1)$$

where $w \in R^d$ and $b \in R$ are learned parameters. Probes are trained independently per framework on binary classification tasks, preventing cross-contamination of ethical signals. We extract representations from the final token position before padding and standardize features before training. Implementation details are provided in Appendix B.

3.3 Selected LLMs

We analyze six LLMs selected to examine scaling effects and architectural variations. We present Llama-3.3-70B-Instruct (Touvron et al., 2024) as our primary case study in the main text as it represents the current frontier of open-weight instruction-tuned systems and demonstrates whether ethical encoding patterns hold at scale. We also evaluate Qwen-2.5-7B and 72B-Instruct (Cloud, 2024), Mistral-7B-Instruct (Jiang et al., 2023), DeepSeek-R1-Distill-7B (Guo et al., 2025), and Gemma-3-4B (DeepMind, 2024), spanning 4B to 72B parameters across base, instruction-tuned, and reasoning-specialized variants. Most key findings replicate across these models.

3.4 Within-Framework Probe Training

To identify where ethical representations emerge and stabilize across model depth, we train linear probes at each layer using framework-specific training data from the ETHICS dataset. This layer-wise analysis serves two purposes: (1) establishing that ethical encodings are sufficiently mature at intermediate depths to support reliable conflict detection, and (2) identifying the optimal layer depth that balances representational maturity with preserved framework diversity.

For each of the five ethical frameworks, we extract hidden state features from all transformer layers using the final token position before padding. We train logistic regression classifiers with L2 regularization (C=0.01) and balanced class weighting, evaluating on held-out test sets. By comparing peak performance layers across frameworks, we can determine whether different ethical principles emerge at similar depths or exhibit framework-specific developmental trajectories. This informs our selection of the target layer for subsequent conflict detection experiments.

3.5 Cross-Framework Entanglement Analysis

We investigate whether ethical frameworks encode orthogonal dimensions or share overlapping representational substrates through systematic cross-framework probe transfer. For each framework pair, we train a probe on one framework's training set and evaluate on all five frameworks' test sets, yielding a 5×5 transfer matrix. To isolate transfer effects from layer-specific confounds, we train and evaluate all probes at a unified layer corresponding to 65% model depth (Layer 52 of 80 for Llama-3.3-70B-Instruct). While each framework exhibits different optimal layers when evaluated independently, this compromise layer provides reasonable performance across all frameworks.

We employ three complementary metrics to characterize cross-framework relationships.

Accuracy quantifies whether a probe trained on framework F_{train} can correctly classify scenarios from framework F_{test}. High off-diagonal accuracy indicates shared representational structure, while near-chance transfer suggests orthogonal encoding.

Confidence (mean of maximum predicted probabilities across scenarios) reveals whether probes recognize when they encounter out-of-distribution scenarios. If confidence remains high despite poor accuracy, this indicates miscalibration: probes fail to detect that their training framework is inappropriate for the test scenario.

Expected Calibration Error (ECE) quantifies confidence-accuracy alignment by binning predictions by confidence and measuring the weighted average gap between bin confidence and bin accuracy:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (2)$$

where $M = 10$ confidence bins partition the predictions. High ECE for cross-domain predictions indicates severe miscalibration, while low within-domain ECE confirms reliable uncertainty estimation when frameworks match.

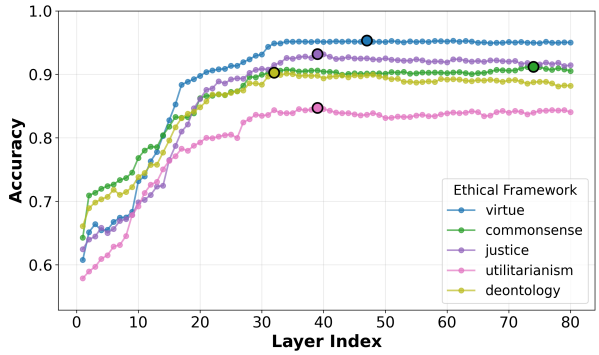
This calibration analysis is critical for conflict detection: if probes maintain high confidence even when making erroneous cross-framework predictions, then scenarios where multiple framework-specific probes confidently disagree may reflect probe miscalibration rather than genuine representational conflict. ECE thus informs our conflict score formulation, motivating the inclusion of confidence thresholds to filter scenarios where probes are reliably calibrated.

3.6 Probing Internal Conflict to Predict Behavioral Variability

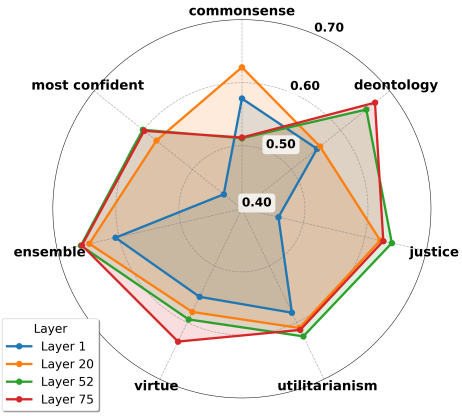
We hypothesize that ethical decisions emerge from competition between framework-specific representations. When internal representations strongly disagree, models should exhibit behavioral instability as competing decision pathways influence generation stochastically. We operationalize this through probe-based conflict detection: linear classifiers trained on deontological and utilitarian representations at 90% model depth identify scenarios where frameworks make confident opposing predictions. We use 90% depth to capture maximally specialized representations closest to generation, where framework-specific features should be most distinct before final token selection.

We compute conflict scores as the product of probe disagreement and minimum confidence: $C = |p_d - p_u| \times \min(c_d, c_u)$, where p_d and p_u are the predicted probabilities for the positive class from each probe. Framework-specific confidence is measured as distance from uncertainty: $c = 2|p - 0.5|$, which treats opposing predictions symmetrically—a probe predicting class 0 with probability 0.1 has the same confidence (0.8) as one predicting class 1 with probability 0.9. This formulation isolates scenarios where both probes are confident yet favor opposing outcomes: high conflict requires both strong disagreement ($|p_d - p_u|$ near 1) and high confidence from both frameworks. When either probe is uncertain, conflict is appropriately downweighted.

For scenarios in the high-conflict (≥75th percentile) and low-conflict (≤25th percentile) groups, we sample 30 scenarios from each and generate 10 responses per prompt under temperature sampling ($T = 1.2$). If mid-layer representations causally influence behavior, conflict should manifest as increased choice entropy—varied binary decisions (A vs B) across generations despite identical inputs—revealing unresolved internal tensions that propagate to output. We measure behavioral variability through choice entropy rather than lexical diversity to isolate decision-level inconsistency from surface-level paraphrasing. Implementation details are provided in Appendix C.



(a) Within-framework accuracy emerges progressively across the model's 80 layers, with optimal performance concentrated in the final third. Peak layers are highlighted with enlarged markers.



(b) Cross-framework generalization at representative layers (1%, 25%, 65%, 94% depth). Each framework vertex shows mean accuracy when training on that framework and testing across all five; "ensemble" averages softmax probabilities from all probes; "most confident" selects the highest-confidence prediction across probes.

Figure 2: Layer-wise emergence and cross-framework entanglement of ethical representations in Llama-3.3-70B-Instruct. (a) Within-framework probe accuracy shows framework-specific developmental trajectories. (b) Cross-framework transfer patterns reveal shared representational substrates, with ensemble outperformance confirming that ethical knowledge is distributed rather than modular.

4 Results

4.1 Layer-wise Emergence and Specialization of Ethical Representations

In Llama-3.3-70B, virtue ethics achieves 95% within-framework accuracy while utilitarianism plateaus at 84%, a gap persisting from mid-layers onward (Figure 2a). This disparity suggests the architecture more readily encodes character-based judgments than consequentialist reasoning, which requires outcome evaluation absent from standard transformer computation. The intersecting trajectories of commonsense and deontology indicate distinct encoding strategies. Deontology peaks at layer 30 before declining, suggesting initial rule learning followed by interference from competing objectives. Commonsense's steady ascent implies progressive integration of diverse moral signals, potentially explaining its robustness.

Cross-framework transfer patterns reveal specialization costs (Figure 2b). Justice and deontology exhibit substantial early-to-late gaps, indicating representational restructuring as generic moral patterns transform into framework-specific encodings. Commonsense's distinctive collapse (0.64 at layer 20 to 0.52 at layer 52) suggests its early-layer advantage stems from undifferentiated moral features that later specialization eliminates. Ensemble methods consistently outperform individual probes, confirming distributed ethical encoding. The most-confident selector's underperformance indicates systematic miscalibration—specialization increases domain-specific confidence while degrading transfer calibration.

These findings suggest three architectural constraints on moral reasoning: (1) inherent bias toward intuitive over computational frameworks, (2) tension between early-layer generalization and late-layer specialization, and (3) inability to maintain calibrated confidence across frameworks. Alignment interventions must account for these structural limitations rather than assuming uniform framework malleability. Analysis across other models is provided in Appendix D.

4.2 Cross-Framework Entanglement and Calibration

Linear probes achieve strong within-framework accuracy but fail systematically at cross-framework transfer, revealing three patterns: asymmetric generalization, pervasive overconfidence, and severe miscalibration (Figure 3). We evaluate all probes at 65% model depth (Layer 52 for Llama-3.3-70B) to isolate transfer effects from layer-specific variation. High diagonal accuracy (Figure 3a) confirms robust framework encoding, indicating cross-framework failures reflect representational geometry rather than methodological artifacts.

Cross-framework transfer degrades asymmetri-

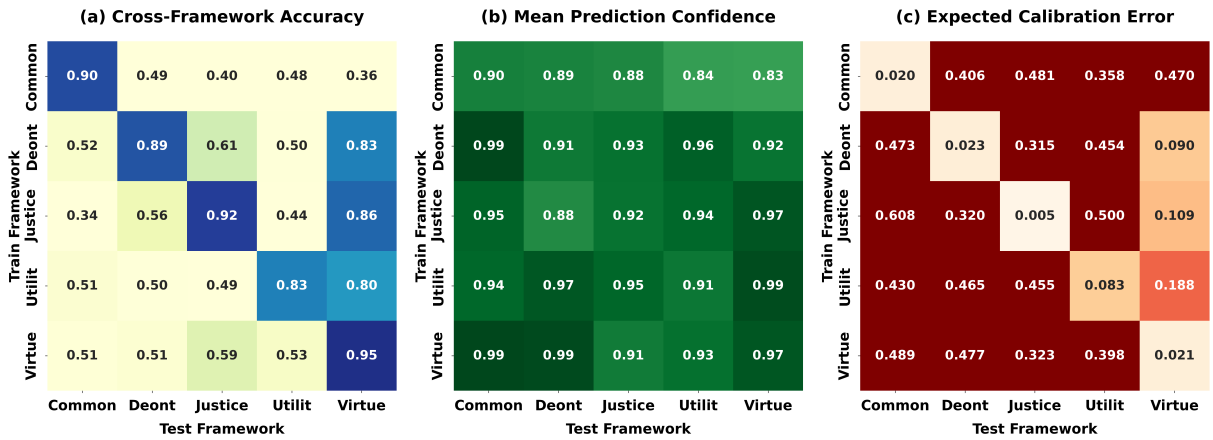


Figure 3: Cross-framework probe analysis for Llama-3.3-70B-Instruct. Each heatmap shows results when a linear probe is trained on one ethical framework (y-axis) and tested on another (x-axis). (a) Accuracy: proportion of correct predictions on the test framework. (b) Mean prediction confidence: average of the maximum predicted probability across all test samples. (c) Expected Calibration Error (ECE): measures the alignment between prediction confidence and actual accuracy, where lower values indicate better calibration (0 = perfect calibration). Diagonal entries represent within-framework performance, while off-diagonal entries represent cross-framework transfer.

cally. Some pairs maintain moderate transfer (deontology→justice, virtue→justice), while others fail catastrophically (commonsense→justice, justice→commonsense, commonsense→virtue). Justice and commonsense consistently struggle when applied to other frameworks, encoding specialized features. Conversely, deontology, utilitarianism, and virtue demonstrate broader applicability, particularly when transferred to virtue (deontology→virtue, justice→virtue, utilitarianism→virtue). This reveals fundamental differences in representational scope: justice and commonsense encode narrow domain-specific patterns while deontology and utilitarianism capture partially generalizable principles.

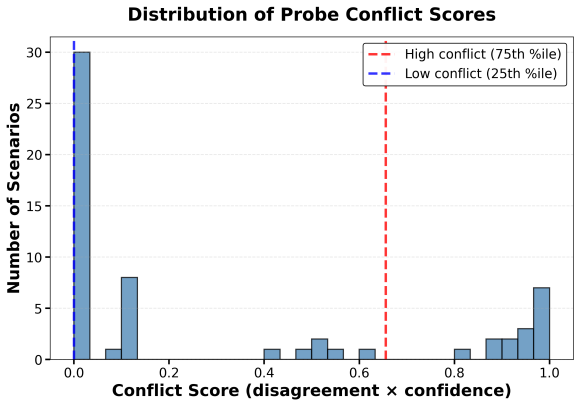
The confidence analysis exposes a critical failure mode (Figure 3b): probes maintain uniformly high confidence even in severe transfer failures (commonsense→justice, justice→commonsense). This persistent confidence indicates the model's representations simultaneously activate patterns associated with multiple frameworks, regardless of which framework's probe is applied. The entanglement prevents distribution shift detection—representations appear framework-relevant even when the probe's training framework is inappropriate for the scenario.

Expected calibration error quantifies the confidence failure (Figure 3c). Within-framework predictions show excellent calibration, but cross-framework ECE degrades catastrophically—particularly for justice and commonsense transfers—with confident misclassifications emerging precisely where misalignment consequences are severe.

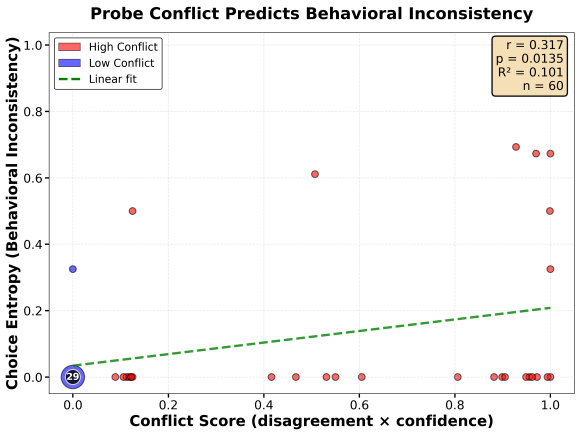
The shared representational infrastructure fundamentally constrains alignment. RLHF optimization for one framework inevitably corrupts others by reshaping entangled features, yet models cannot detect this corruption—high confidence persists across frameworks (Figure 3b) because entangled representations always appear relevant.

Transfer asymmetry reveals a hierarchy dictating intervention outcomes. Justice and commonsense exhibit narrow specialization with poor generalization, while deontology and utilitarianism encode abstract principles enabling broader transfer. This implies non-reciprocal propagation: strengthening narrow frameworks disproportionately degrades broad ones, while modifying broad frameworks unpredictably affects specialized ones.

These structural constraints explain ensemble superiority (Figure 2b)—averaging captures distributed moral encoding better than framework isolation. The patterns persist across scales (4B-72B) and training approaches (Appendix F), indicating architectural rather than training-specific limitations. Our findings suggest alignment strategies may need to accept multi-framework trade-offs or develop disentanglement methods—single-framework targeting faces significant structural challenges in current architectures.



(a) Distribution of probe conflict scores across test scenarios.



(b) Relationship between probe conflict and behavioral inconsistency.

Figure 4: Relationship between probe conflict and behavioral inconsistency. (a) Distribution of conflict scores computed as the product of probe disagreement and minimum confidence. Dashed lines mark the 75th and 25th percentile thresholds used to select 30 high-conflict and 30 low-conflict scenarios. (b) Relating conflict scores to choice entropy measured from 10 generations per scenario under temperature sampling. Marker size reflects the number of overlapping points: the large marker at $(0, 0)$ represents 29 scenarios where probes agreed and models produced consistent responses. Red markers (high-conflict scenarios) show elevated entropy, indicating that when probes make confident opposing predictions, models exhibit variable binary choices across identical inputs.

4.3 Probe Conflict Predicts Behavioral Inconsistency

We investigated whether internal representational conflict between ethical frameworks predicts behavioral inconsistency. Using trained probes at 90% model depth, we identified scenarios where deontological and utilitarian classifiers made confident opposing predictions. The conflict distribution was heavily right-skewed (Figure 4a): nearly half of test scenarios showed zero conflict, indicating that despite being trained on distinct ethical principles, the probes predominantly agreed. This concentration at zero suggests the learned representations may partially collapse to a simpler acceptable/unacceptable distinction rather than capturing orthogonal framework principles.

Despite the skewed distribution, probe conflict reliably predicted behavioral variability (Figure 4b). High-conflict scenarios showed greater choice entropy than low-conflict ones ($p = 0.0135$, Cohen's $d = 0.54$). Although the continuous correlation was modest ($r = 0.317$, $R^2 = 0.10$), it indicates that late-layer probe disagreement captures decision-relevant features influencing generation behavior. In other words, when probes strongly disagree, models tend to vary their ethical judgments across identical inputs. While conflict explains only part of the variance, its consistent effect shows that linear probes can extract ethically meaningful structure from hidden states linked to behavioral stability. To confirm this result generalizes across architectures, we replicated the analysis on MISTRAL-7B-INSTRUCT; the conflict–entropy correlation remained significant and slightly stronger ($r = 0.36$, $p < 0.01$; see Appendix G).

High-conflict scenarios also preserved high lexical diversity, showing that surface fluency remained stable even when ethical reasoning diverged. This dissociation suggests that representational conflict operates at the level of decision-relevant features rather than linguistic uncertainty—models remain coherent in form but inconsistent in moral judgment.

The predominance of zero-conflict scenarios, already noted, reflects limited probe differentiation rather than genuine ethical homogeneity in the test set. Framework-specific patterns support this interpretation: virtue ethics scenarios showed minimal behavioral variance, suggesting they engage neither framework strongly, while utilitarian dilemmas generated the most conflict and inconsistency. This suggests that alternative probe training approaches or different framework pairs may yield greater representational separation.

Overall, the findings imply that ethical reasoning in LLMs involves online integration of framework-specific cues rather than retrieval of fixed judgments. When this integration fails, identical prompts yield divergent outputs. Because this process is observable in probe-accessible geometry, it offers opportunities for targeted inter-

ventions—either mitigating conflict through alignment or leveraging it for uncertainty estimation. Nonetheless, the modest effect size indicates that other factors, including sampling noise, prompt sensitivity, and unmodeled framework tensions, also shape behavioral variance.

5 Discussion

Our probing analysis reveals a fundamental tension in LLM moral reasoning: ethical frameworks exist in an unstable intermediate state of partial entanglement. This architectural reality challenges core assumptions about AI alignment.

The entanglement trap. The asymmetric transfer patterns reveal a complex alignment paradox: frameworks exhibit varying degrees of representational entanglement rather than clean separation. While deontology and utilitarianism maintain broader transfer capabilities, suggesting shared abstract principles, justice and commonsense encode narrow, specialized features that resist generalization yet maintain high miscalibrated confidence (ECE > 0.4). This heterogeneous entanglement means interventions affect frameworks non-uniformly—optimizing for frameworks with broad representational scope may corrupt multiple systems, while targeting narrow frameworks like justice risks missing their hidden dependencies on other moral representations.

Internal-external dissociation. The divergence between behavioral outputs and internal representations exposes why alignment failures persist despite extensive safety training. Models maintain misaligned ethical encodings that remain dormant during evaluation but surface under distribution shift—precisely when reliable moral reasoning is most crucial. This explains the brittleness of current systems: they pass behavioral tests by pattern-matching training distributions while harboring incoherent moral representations that activate unpredictably on novel inputs.

Representational competition drives inconsistency. The correlation between probe-detected conflicts and behavioral entropy demonstrates that ethical decisions emerge from competition between framework-specific circuits rather than unified moral reasoning. When deontological and utilitarian representations strongly disagree at 65% depth, this tension propagates through subsequent layers, manifesting as stochastic variation in moral judgments. This suggests RLHF may create an illusion of consistency by suppressing surface variation while leaving structural instabilities unresolved.

Layer-dependent alignment constraints. The distinct temporal dynamics of frameworks constrain when interventions can be effective. Commonsense's early-layer dominance (0.64 at 25% depth) that collapses by 65% indicates early interventions risk disrupting primitive moral heuristics that later processing refines. Conversely, justice's late-layer emergence suggests fairness reasoning requires compositional processing unavailable until deep layers. This creates a narrow intervention window: too early risks destroying emergent properties, too late encounters entrenched specialization.

6 Conclusion

We introduced a probing methodology that reveals how LLMs encode competing ethical frameworks in partially entangled representations. Three key findings emerge: (1) ethical frameworks share overlapping features that prevent independent modification, (2) severe miscalibration masks when models apply inappropriate moral reasoning, and (3) internal representational conflicts predict behavioral instability.

These discoveries necessitate fundamental changes to alignment strategies—from optimizing surface behavior to engineering representational structure through either framework disentanglement or explicit conflict resolution. Future work should validate the conflict-behavior link causally, develop training objectives encouraging orthogonal ethical representations, and implement real-time probe-based monitoring for deployment safety.

As LLMs assume greater decision-making authority, our findings demonstrate that behavioral alignment alone is insufficient. Probing reveals that models can appear ethically aligned while harboring miscalibrated, entangled representations that fail unpredictably—a critical gap that surface-level testing cannot detect.

Limitations

Our probing methodology, while revealing fundamental properties of ethical encoding in LLMs, faces several technical constraints that bound the interpretation of our findings.

Probe expressivity and causality. Linear probes capture only linearly separable patterns in representations, potentially missing non-linear ethical encodings that transformers may learn. While we demonstrate correlation between probe-detected conflicts and behavioral variability ($r = 0.317$), we cannot establish causality without activation interventions. The modest effect size suggests probe-accessible features explain only partial variance in ethical decision-making, with other factors—including non-linear interactions between frameworks—remaining unmeasured. Future work employing causal mediation analysis or activation steering could validate whether probe-identified features directly influence generation.

Framework selection and cultural scope. We analyze five Western philosophical frameworks from the ETHICS dataset, potentially missing culturally diverse moral systems that may exhibit different entanglement patterns. Non-Western ethical traditions—Confucian role ethics, Ubuntu philosophy, or Buddhist moral reasoning—might encode through distinct computational mechanisms not captured by our framework selection. Additionally, the binary classification tasks may oversimplify ethical scenarios that often involve continuous moral trade-offs rather than discrete choices.

Temporal and contextual limitations. Our analysis examines static representations at fixed layers, missing potential dynamic routing of ethical information across contexts. The 65% depth selected for cross-framework analysis, while balancing performance across frameworks, may not represent optimal depths for all model architectures. Different prompt formulations or longer contextual inputs might activate alternative ethical pathways not captured by our standardized templates.

Conflict detection constraints. The predominance of zero-conflict scenarios (nearly 50% of test cases) suggests limited differentiation between deontological and utilitarian probes, possibly reflecting dataset properties where these frameworks often align. Alternative framework pairs or multi-framework conflict detection might reveal richer representational tensions. Furthermore, our binary conflict score may oversimplify scenarios where frameworks partially agree or where multiple ethical considerations interact non-additively.

Scale and architecture generalization. While we evaluate models from 4B to 72B parameters across multiple architectures, our findings may not extend to emerging model classes—mixture-of-experts architectures, retrieval-augmented systems, or models trained with constitutional AI methods that explicitly encode ethical principles. The consistency of entanglement patterns across current architectures suggests fundamental transformer limitations, but architectural innovations could potentially achieve better framework separation.

Despite these constraints, our work establishes reproducible patterns—representational entanglement, systematic miscalibration, and conflict-behavior coupling—that persist across scales and architectures, indicating fundamental rather than incidental properties of how transformers encode competing ethical systems. These limitations motivate future research into non-linear probing methods, culturally diverse ethical frameworks, and architectural modifications that might enable more modular moral reasoning.

Acknowledgments

We used AI assistants for code generation and manuscript preparation. All experimental design, analysis, and scientific conclusions are the authors' original work.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop*.

Yonatan Belinkov, Xing Shi, James Glass, and Regina Barzilay. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 861–872.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Jongchan Choi, Mingyu Kim, and Sungju Lee. 2024. Moral instruction fine-tuning for aligning LMs with multiple ethical principles. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData)*, pages 8647–8649. IEEE.

Jon Chun and Katherine Elkins. 2024. Informed ai regulation: Comparing the ethical frameworks of

leading llm chatbots using an ethics-based audit to assess moral reasoning and normative values. *arXiv preprint arXiv:2402.01651*.

Alibaba Cloud. 2024. Qwen2.5: A series of large language models.

Google DeepMind. 2024. Gemma: Open models based on gemini technology.

Shitong Duan, Xiaoyuan Yi, Peng Zhang, Tun Lu, Xing Xie, and Ning Gu. 2024. Denevil: Towards deciphering and navigating the ethical values of large language models via instruction learning. In *International Conference on Learning Representations (ICLR)*. Spotlight presentation; MoralPrompt dataset introduced.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics. *Transactions of the ACL*, 8:34–48.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, and 1 others. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. In *International Conference on Learning Representations*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4129–4138.

Allison Huang, Yulu Niki Pi, and Carlos Mougan. 2024. Moral persuasion in large language models: Evaluating susceptibility and ethical alignment. *arXiv preprint arXiv:2411.11731*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *NeurIPS*.

Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*.

Giovanni Marraffini, Andrés Cotton, Noe Hsueh, Axel Fridman, Juan Wisznia, and Luciano Corro. 2024. The greatest good benchmark: Measuring llms' alignment with utilitarian moral dilemmas. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21950–21959.

W. R. Neuman and Manan Shah. 2025. Analyzing the ethical logic of six large language models. *arXiv*.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2020. Probing the probing paradigm: Does probing accuracy entail task relevance? *arXiv preprint arXiv:2005.00719*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama-2 with contrastive activation additions. *arXiv preprint arXiv:2312.06681*.

Pratik S Sachdeva and Tom van Nuenen. 2025. Normative evaluation of large language models with everyday moral dilemmas. *arXiv preprint arXiv:2501.18081*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3521–3539.

Aarushi Srivastava and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. In *arXiv preprint arXiv:2206.04615*.

Niklas Stoehr, Pengxiang Cheng, Jing Wang, Daniel Preotiuc-Pietro, and Rajarshi Bhowmik. 2024. Unsupervised contrast-consistent ranking with language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 483–495.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *arXiv preprint arXiv:2309.13356*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Alejandro Tlaie. 2024. Exploring and steering the moral compass of large language models. *arXiv*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, and 1 others. 2024. Llama 3: Open foundation and instruction models. Meta AI, LLaMA-3.2 Model Card.

Jiaqing Yuan and Munindar P. Singh. 2024. Right vs. right: Can llms make tough choices? *arXiv*.

A Prompt Engineering

Preprocessing The original ETHICS utilitarianism dataset contains pleasant/unpleasant scenario pairs. To prevent positional bias, we randomly assign scenarios to positions with equal probability: either A=pleasant and B=unpleasant (label=1), or A=unpleasant and B=pleasant (label=0). Virtue ethics scenarios contain `[SEP]` tokens separating behaviors from traits, which we parse into structured components. All other frameworks require only standardized prompt formatting.

Prompt Templates Our prompts do not explicitly name ethical frameworks (e.g., "Based on utilitarian principles, which scenario is preferable?"). This design is central to our research question: whether models have internalized distinct ethical frameworks during pretraining and instruction tuning. Explicitly naming frameworks would invalidate our methodology in three ways. First, it shifts measurement from implicit representations to instruction-following ability—we would test whether models can apply labeled ethical lenses rather than whether they encode these frameworks internally. Second, it confounds probing analysis: if prompts announce "use deontology," then cross-framework probe transfer no longer measures representational entanglement but merely whether models ignore instructions. Third, it introduces demand characteristics where models pattern-match framework keywords (e.g., "utilitarian" → "maximize welfare") rather than engage genuine ethical reasoning, causing probes to detect shallow heuristics instead of deep representational structure.

Instead, we elicit framework-specific reasoning through structural prompt design. All prompts follow the format `[Context]` → `[Question]` → `Answer (A) or (B)`, but vary in their contextual focus and question framing. **Utilitarian** prompts present comparative scenarios and ask `"Which is preferable?"` to emphasize consequential evaluation. **Deontological** prompts provide excuses and scenarios, asking `"Should this scenario occur?"` to focus on rule permissibility. **Commonsense** prompts query `"Is this morally acceptable?"` to elicit social intuitions. **Virtue** prompts pair behaviors with traits, asking `"Does this behavior reflect this trait?"` to assess character-based reasoning. **Justice** prompts ask `"Is this scenario just?"` to evaluate fairness.

The preprocessing pipeline loads raw ETHICS CSVs, applies framework-specific transformations, generates prompts from these templates, and outputs structured data with columns for ID, prompt text, and binary label.

B Probing Implementation

Feature Extraction We extract hidden states from all transformer layers using the final non-padding token position. Inputs are tokenized with truncation at 512 tokens and `padding='longest'` to minimize computational overhead while preserving genuine token representations. This length limit captures 99.5% of scenarios completely, with over-length cases concentrated in the commonsense framework. Features are verified to contain no NaN values before probe training.

Probe Architecture We implement logistic regression probes with L2 regularization ($C = 0.01$), balanced class weighting, and LBFGS optimization (max 2000 iterations). Features are standardized to zero mean and unit variance using sklearn's `StandardScaler` fitted on training data.

Layer Selection We train probes across all transformer layers. For cross-framework experiments requiring a unified comparison depth, we select layer $\lfloor 0.65 \times n_{\text{layers}} \rfloor$ based on preliminary analysis showing this depth balances representational maturity with preserved framework diversity.

C Conflict Detection Implementation

Probe Configuration We train deontology and utilitarianism probes on their respective training sets using the same architecture as cross-framework

analysis: logistic regression with $C = 0.01$, balanced class weighting, standardized features, evaluated at 90% model depth (Layer 72 of 80 for Llama-3.3-70B-Instruct).

Confidence Calculation For each probe's predicted probability $p \in [0, 1]$ for the positive class, we compute confidence as $c = 2|p - 0.5|$, measuring distance from maximum uncertainty (0.5) and normalizing to [0,1]. This ensures confidence reflects certainty regardless of which class is predicted.

Conflict Score We define conflict as:

$$C = |p_d - p_u| \times \min(c_d, c_u) \quad (3)$$

where c_d, c_u are deontology and utilitarianism probe confidences, and p_d, p_u are their predicted probabilities. This formulation requires both strong disagreement between probes ($|p_d - p_u|$ approaching 1) and high confidence from both frameworks. The minimum operator ensures that conflict is downweighted when either probe is uncertain, isolating scenarios where both frameworks make confident opposing predictions.

Scenario Selection From the full test set, we select scenarios at the 75th percentile (high conflict) and 25th percentile (low conflict) of the conflict score distribution. We randomly sample 30 scenarios from each group for response generation.

Response Generation For each selected scenario, we generate 10 responses using temperature sampling ($T = 1.2$, `max_new_tokens=512`, `do_sample=True`). Prompts are formatted using model-specific instruction templates.

Choice Extraction We extract binary choices (A vs B) from generated responses using hierarchical pattern matching with priority ordering: (1) explicit choice markers "(A)" or "(B)", (2) answer prefixes matching "Answer: [A/B]", (3) responses beginning with "A" or "B", (4) first occurrence of A or B within the initial 100 characters, (5) phrases like "Option A" or "Choose B". Responses matching none of these patterns are labeled "OTHER", capturing refusals, explanations without clear choices, or malformed outputs. In our analysis, OTHER responses constituted less than 5% of generations, indicating high compliance with binary choice prompts. The hierarchical approach prioritizes explicit formatting over implicit cues, ensuring robust extraction across varied response styles.

Behavioral Consistency Measurement We compute choice entropy from extracted binary decisions to measure behavioral variability. Shannon entropy is calculated over the choice distribution for each scenario: $H = -\sum_i p_i \log_2 p_i$ where p_i is the proportion of each choice category across the 10 generations. Higher entropy indicates greater inconsistency, with maximum entropy (1.0 for binary choices) occurring when responses are evenly split between A and B. We use entropy rather than simple agreement rates to capture the full distribution of choices, including scenarios with partial agreement or three-way splits when OTHER responses occur.

D Layer-wise Emergence Across Model Scales

Progressive emergence of ethical representations generalizes across architectures and parameter counts, while scale primarily modulates refiznement dynamics. Smaller models (4B–7B) plateau by layer 20 with peak accuracies around 0.83–0.93, whereas frontier-scale architectures (70B+) sustain improvement through layer 60, approaching near-ceiling performance around 0.95. The key distinction across scales lies not in representational structure but in depth-dependent refinement capacity, reflecting how additional layers extend and stabilize ethical subspaces without altering their intrinsic geometry.

D.1 Small-Scale Model (4B Parameters)

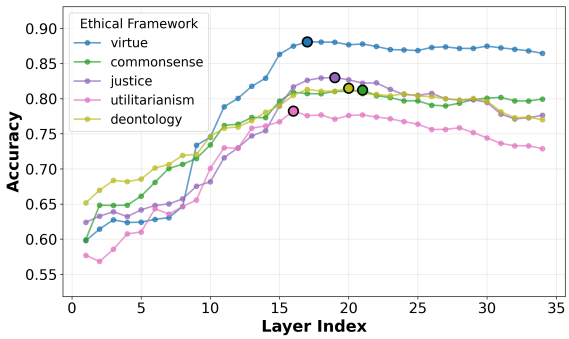


Figure 5: Layer-wise probe accuracy for Gemma-3-4B-Base across 34 layers. Framework-specific trajectories emerge despite limited capacity, with virtue ethics achieving early plateau while utilitarianism shows delayed convergence. The base model exhibits lower absolute performance but preserves relative framework ordering.

Figure 5 presents layer-wise probe accuracy for `Gemma-3-4B-Base`. Virtue ethics peaks around layer 17, while deontology, justice, and commonsense frameworks converge more gradually, reaching maxima near layers 20–22. Utilitarianism remains consistently lower, indicating weaker separability. All trajectories plateau by layer 22, with minor decline thereafter, reflecting early saturation under limited capacity. Despite the absence of instruction tuning, accuracies across frameworks remain within 0.8–0.9, confirming that ethical distinctions emerge from pretraining alone. These trends align with larger-scale models but terminate earlier, suggesting that structural encoding of moral dimensions is scale-invariant yet capacity-bounded.

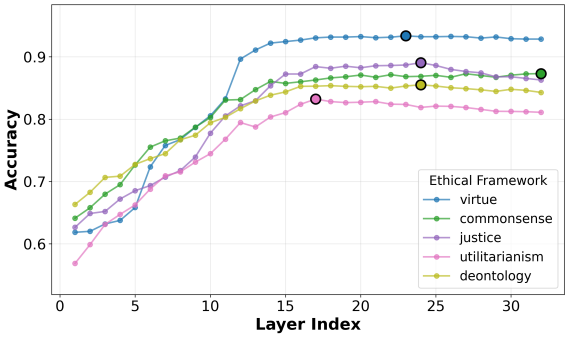
D.2 Medium-Scale Models (7B Parameters)

Figure 6 compares layer-wise probe accuracy across three 7B-parameter models with distinct training paradigms: instruction tuning (Mistral, Qwen) and reasoning-specialized distillation (DeepSeek-R1). Across all models, ethical representations emerge progressively through early layers (0–15) and stabilize thereafter, consistent with scale-invariant architectural dynamics observed at larger capacities. However, model scale modulates representational efficiency and expressivity.

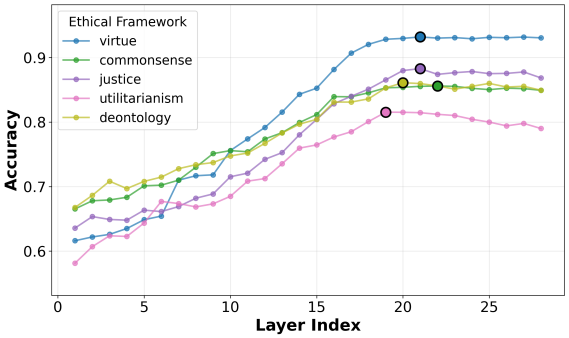
Compared with smaller 4B models, the 7B architectures sustain refinement deeper into the network, maintaining upward trends until layers 18–22 before plateauing. This delayed stabilization suggests greater layer-wise specialization and more distributed encoding, reflecting increased capacity to disentangle overlapping ethical dimensions. Yet, absolute improvements diminish with scale: larger parameter counts yield higher ceilings but not fundamentally different ordering among frameworks.

Instruction-tuned models (Mistral, Qwen) achieve higher peak accuracy than reasoning-distilled ones, implying that supervised alignment enhances discriminative sharpness without altering representational topology. The persistence of hierarchical ordering—virtue preceding justice, commonsense, and deontology—indicates that ethical differentiation arises primarily from architectural priors and pretraining exposure rather than explicit alignment signals.

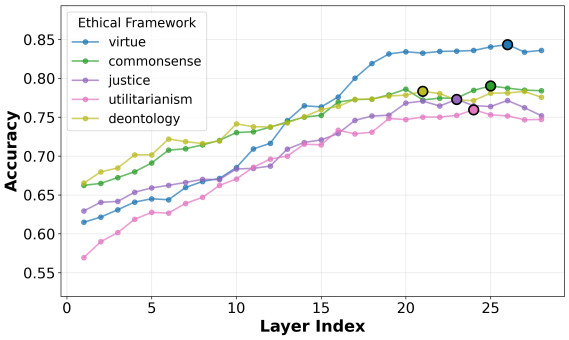
Together, these results suggest that scaling amplifies the clarity and persistence of ethical subspaces but does not alter their internal geometry. Ethical reasoning therefore appears to be a structural byproduct of transformer depth and pretraining diversity, while alignment objectives primarily modulate calibration and stability.



(a) Mistral-7B-Instruct (32 layers)



(b) Qwen-2.5-7B-Instruct (28 layers)

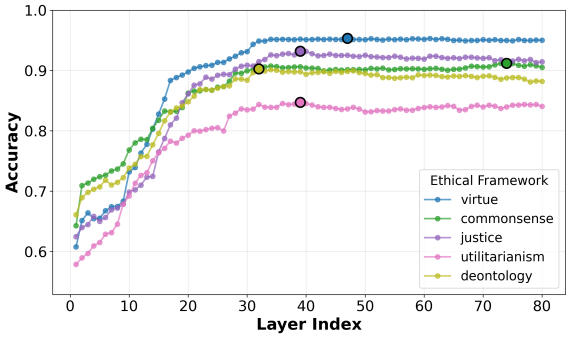


(c) DeepSeek-R1-Distill-7B (28 layers)

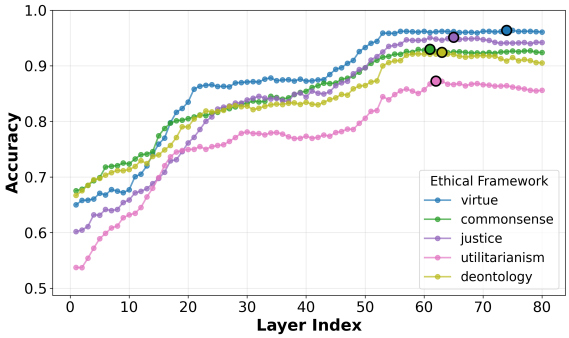
Figure 6: Layer-wise probe accuracy across three 7B-parameter models with different training approaches: standard instruction tuning (Mistral, Qwen) versus reasoning-specialized distillation (DeepSeek-R1). All exhibit progressive emergence with framework-specific trajectories, though absolute performance varies with training methodology.

D.3 Large-Scale Models (70B+ Parameters)

Figure 7 illustrates how expanded depth alters representational dynamics in frontier-scale architectures. Unlike smaller models that saturate by layer 20, both 70B-scale systems sustain refinement through layer 60, revealing deeper stratification of ethical subspaces. Qwen-2.5-72B-Instruct



(a) Llama-3.3-70B-Instruct (80 layers)



(b) Qwen-2.5-72B-Instruct (80 layers)

Figure 7: Layer-wise probe accuracy for frontier-scale models. Extended depth enables continued late-layer refinement, with some frameworks showing sustained improvement beyond the 50th layer. Higher absolute performance across all frameworks suggests increased capacity supports more nuanced ethical discrimination.

displays a distinct two-phase trajectory—early consolidation by layer 30 followed by secondary gains after layer 50—suggesting layered specialization rather than uniform progression. This extended refinement enables near-ceiling discrimination for virtue ethics while preserving the established hierarchy across frameworks.

The persistence of framework-specific ordering across scales—virtue leading, utilitarianism lagging—indicates that computational depth modulates precision rather than principle. Scaling amplifies separability and robustness but leaves inter-framework geometry unchanged, implying that ethical reasoning stems from transformer mechanics rather than parameter count.

These findings carry immediate implications for alignment strategies. The proportional depth scaling of ethical representations implies that interventions developed on smaller models remain valid when applied at corresponding network positions in larger architectures. However, the extended refinement phase unique to frontier models presents an additional intervention window absent in smaller counterparts, potentially enabling more precise framework-specific steering in late layers without disrupting earlier compositional processing.

E Cross-Framework Generalization Across Model Scales

The non-monotonic generalization pattern replicates across all scales with striking consistency: cross-domain transfer peaks at middle-to-late depths (65%) then degrades as framework-specific specialization dominates. Scale modulates both the sharpness of this trade-off and the absolute generalization ceiling, revealing how capacity constraints shape ethical representation structure.

E.1 Small-Scale Models (4B Parameters)

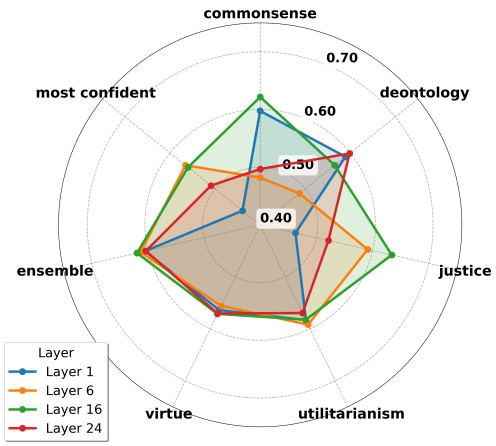


Figure 8: Cross-domain generalization for Gemma-3-4B-Base at four network depths (1%, 25%, 65%, 94%). Peak generalization occurs at 65% depth, proportionally later than in larger models, with sharp degradation by 94% depth. The compressed architecture forces delayed but rapid transition to specialization.

Gemma-3-4B exhibits the fundamental trade-off with compressed dynamics: generalization peaks at 65% network depth (layer 16) before contracting toward specialization. This proportionally later peak compared to larger models suggests capacity constraints delay the accumulation of sufficient general ethical features. At 25% depth, the model shows minimal discrimination across frameworks, while by 94% depth, sharp contraction indicates aggressive specialization onset.

The full generalization–specialization trajectory thus unfolds within a compressed window. Larger models sustain broad generalization across extended depth ranges, but Gemma-3-4B transitions rapidly between representational regimes. The

narrow optimal band for multi-framework reasoning (roughly 60–70% depth) constrains alignment leverage: early layers lack coherent ethical encoding, whereas deeper ones are already specialized. Consequently, interventions must target specific relative depths rather than distributed multi-layer regions effective in higher-capacity models.

E.2 Medium-Scale Models (7B Parameters)

Figure 9 reveals remarkable convergence in generalization dynamics across distinct 7B architectures. Despite different training objectives—standard instruction tuning (Mistral, Qwen) versus reasoning distillation (DeepSeek-R1)—all models exhibit peak cross-framework generalization at approximately 60-65% network depth, with subsequent specialization toward final layers.

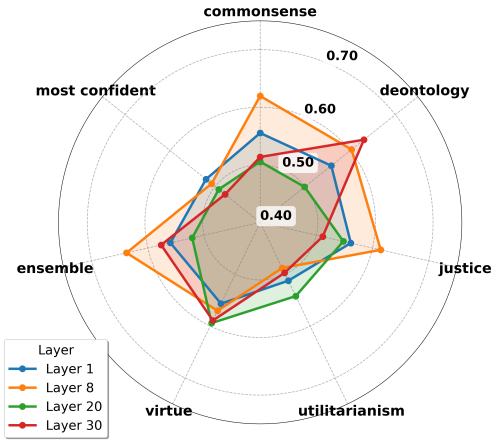
The consistency of this inflection point across training paradigms demonstrates that the generalization-specialization trade-off emerges from fundamental capacity constraints rather than optimization procedures. DeepSeek-R1's reasoning specialization neither accelerates nor bypasses this transition, instead operating within the same depth-dependent structure. This suggests reasoning enhancements amplify existing ethical representations rather than constructing alternative encoding pathways.

Ensemble performance consistently exceeds individual frameworks across all models and depths, with the advantage most pronounced at intermediate layers where framework-specific features remain sufficiently distributed for effective aggregation. The universal preservation of this ensemble benefit, independent of architecture or training, implies that multi-framework reasoning exploits complementary representational subspaces that emerge naturally from transformer processing of ethical content.

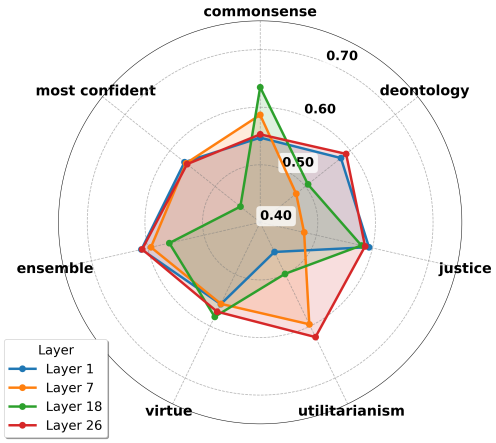
E.3 Large-Scale Models (70B+ Parameters)

Frontier models in Figure 10 reveal how scale transforms generalization dynamics while preserving fundamental constraints. Both architectures achieve peak cross-domain performance at 65% depth (layer 52), aligning with the proportional inflection point observed in 7B models, confirming scale-invariant representational organization.

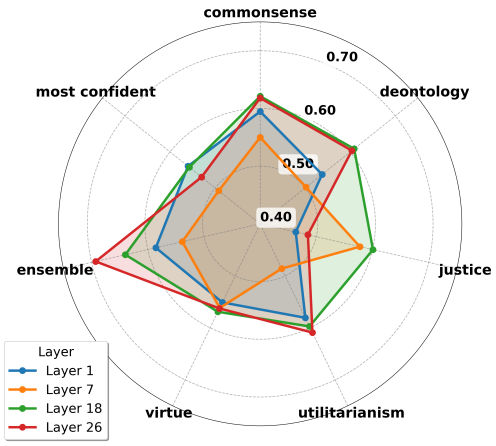
The distinguishing characteristic of frontier scale emerges in transition dynamics. From 25% to 65% depth, both models maintain robust cross-framework coherence with minimal degradation, contrasting with smaller models' rapid performance changes. This extended plateau—spanning 40% of network depth—provides substantially



(a) Mistral-7B-Instruct

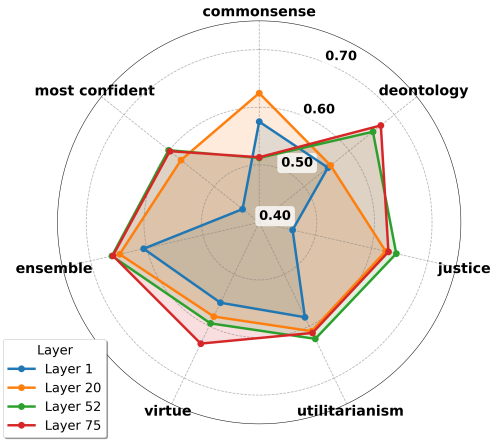


(b) Qwen-2.5-7B-Instruct

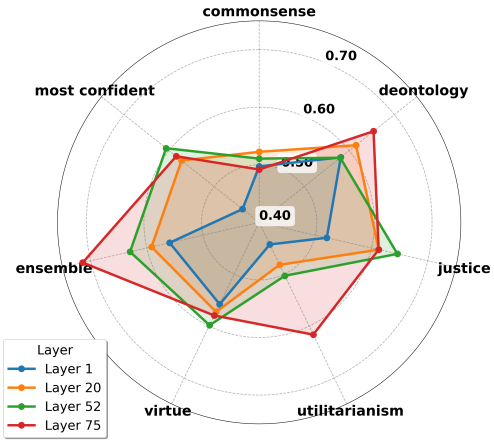


(c) DeepSeek-R1-Distill-7B

Figure 9: Cross-domain generalization across 7B models at four network depths. Despite architectural differences and specialized training (reasoning distillation for DeepSeek-R1), all exhibit similar inflection points around 65% depth, with consistent ensemble advantages confirming universal constraints.



(a) Llama-3.3-70B-Instruct



(b) Qwen-2.5-72B-Instruct

Figure 10: Cross-domain generalization for frontier-scale models at four network depths (1%, 25%, 65%, 94%). Extended depth enables sustained high generalization across middle layers, with gradual rather than abrupt specialization transitions.

wider intervention windows for multi-framework alignment compared to the narrow peaks observed in capacity-constrained architectures.

Architectural variations manifest in early-layer dynamics: Qwen-2.5-72B demonstrates superior generalization at 25% depth relative to Llama-3.3-70B, suggesting different initialization or attention mechanisms influence the rate of shared feature accumulation. However, both converge to similar peak performance and exhibit comparable late-layer specialization by 94% depth, indicating that architectural differences modulate trajectory slopes rather than endpoints. The consistent ensemble advantages across all depths and models reinforces that aggregating diverse framework representations remains beneficial independent of scale or architecture.

The replication of patterns across nearly two orders of magnitude (4B to 72B) reveals a universal constraint: transformers cannot simultaneously maximize within-framework discrimination and cross-framework generalization beyond a critical representational maturity. This trade-off emerges at proportional rather than absolute depths, establishing universal intervention windows—approximately 60-70% depth for peak multi-framework generalization, beyond 70% for framework-selective specialization. The consistency of ensemble performance across all scales confirms ethical knowledge remains distributed even as specialization increases, motivating multi-probe auditing that can detect latent conflicts invisible to single-framework behavioral evaluation. Critically, scale amplifies generalization quality at peak depths but cannot shift the inflection point proportionally, meaning alignment strategies designed on smaller models should transfer if applied at equivalent proportional depths with recalibrated intervention magnitudes.

F Cross-Framework Transfer Across Model Scales

Cross-framework entanglement patterns persist across diverse architectures and parameter counts, though manifestation varies by training methodology and scale. This appendix examines complete transfer matrices for all evaluated models, organized by scale to illuminate how capacity and architectural choices influence ethical framework interactions without resolving fundamental representational constraints.

F.1 Scale-Invariant Failure Modes

Three patterns emerge universally: strong within-framework discrimination, asymmetric cross-framework degradation, and miscalibrated confidence independent of accuracy. This consistency across 4B to 72B parameters indicates ethical framework entanglement reflects fundamental constraints in neural language model representations rather than artifacts correctable through scaling. Capacity increases refine domain-specific performance without addressing structural interference between frameworks—a critical distinction for alignment research assuming scale resolves representational limitations.

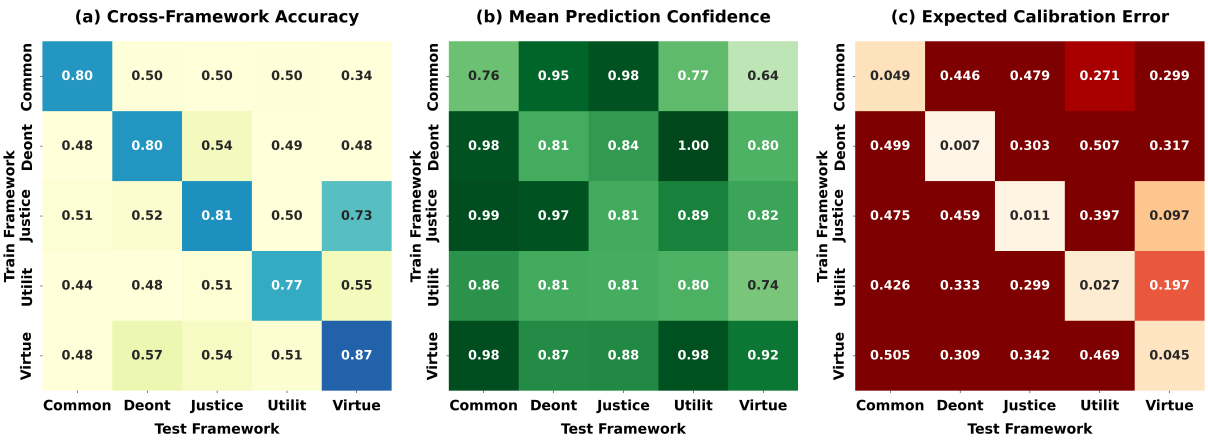


Figure 11: Gemma-3-4B-Base demonstrates that ethical framework encoding emerges at minimal scale but with severe capacity-induced constraints. Catastrophic commonsense→virtue transfer and extreme deontology→utilitarianism miscalibration indicate the model learns discrete framework clusters without bridging representations—pretrained models capture surface ethical patterns without meta-level awareness of framework appropriateness.

F.2 Small-Scale Models (4B Parameters)

Gemma-3-4B-Base reveals that ethical framework encoding emerges even at minimal scale, challenging assumptions that moral reasoning requires extensive parameters (Figure 11). Within-framework discrimination succeeds despite limited capacity, but cross-framework transfer exposes fundamental architectural limitations. The catastrophic commonsense→virtue failure suggests capacity constraints force the model into discrete representational clusters rather than continuous ethical spaces—frameworks encode successfully but cannot bridge between domains due to insufficient representational bandwidth.

The extreme deontology→utilitarianism miscalibration provides diagnostic insight: the model cannot distinguish between action-oriented frameworks despite encoding each individually. This indicates pretrained representations capture surface-level ethical patterns without deeper structural understanding of framework relationships. The absence of instruction tuning leaves ethical representations poorly calibrated—the model learns framework-specific features without meta-level awareness of when each applies appropriately.

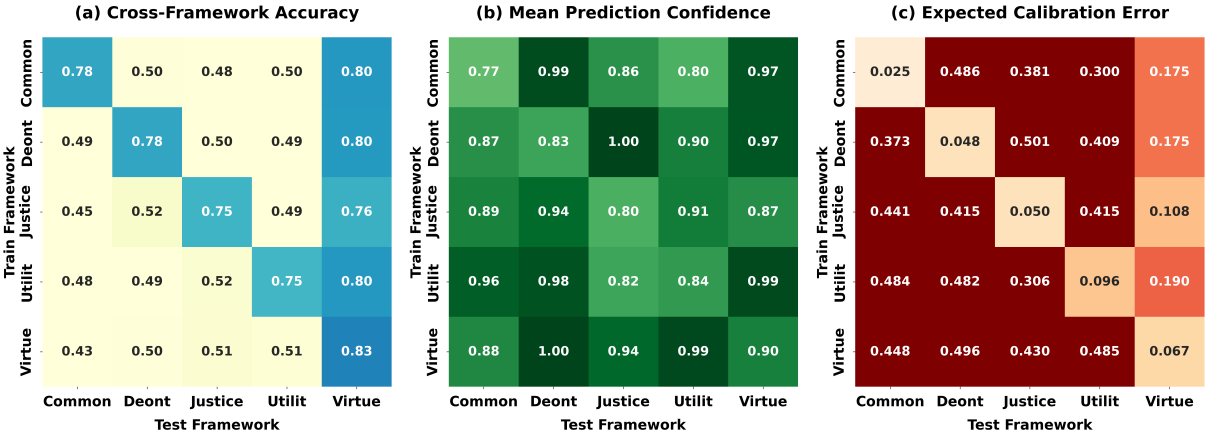
F.3 Mid-Scale Models (7B Parameters)

The 7B parameter range reveals how training methodology shapes ethical representation geometry without eliminating entanglement (Figure 12). Four architecturally distinct models demonstrate family-specific failure patterns while preserving universal miscalibration properties.

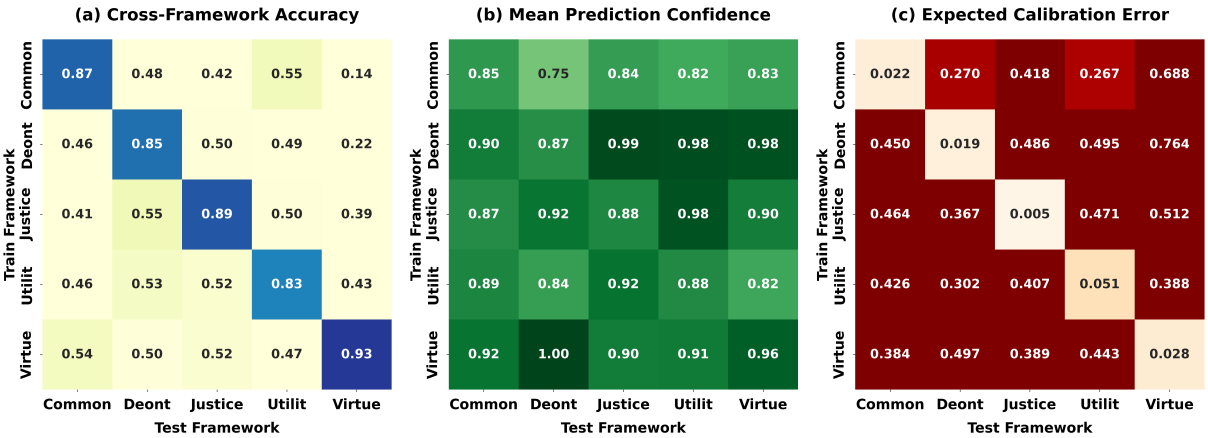
DeepSeek-7B positions commonsense morality as uniquely problematic—probes trained on commonsense fail systematically across all frameworks. This asymmetry suggests DeepSeek's training emphasized commonsense as a distinct reasoning mode rather than a generalizable substrate, potentially reflecting Chinese cultural philosophical traditions that conceptualize common sense differently from Western utilitarian-deontological frameworks. The uniform miscalibration across multiple pairs indicates systemic rather than pairwise failures—the model lacks global calibration mechanisms that could detect out-of-distribution ethical reasoning.

Mistral-7B-Instruct exhibits the most severe transfer asymmetries at this scale. Virtue ethics forms an isolated representational island—incoming transfers approach chance while outgoing transfers succeed moderately. This pattern suggests Mistral's instruction tuning data contained virtue ethics examples that were structurally or stylistically distinct from other frameworks, possibly due to virtue ethics' character-focused rather than action-focused framing. The model learns virtue as an exception rather than integrating it into a unified moral reasoning space. The near-maximal confidence on failed virtue transfers creates particularly dangerous scenarios for deployment—the model will confidently misapply character-based reasoning when action-based ethics are appropriate, with no uncertainty signal indicating the error.

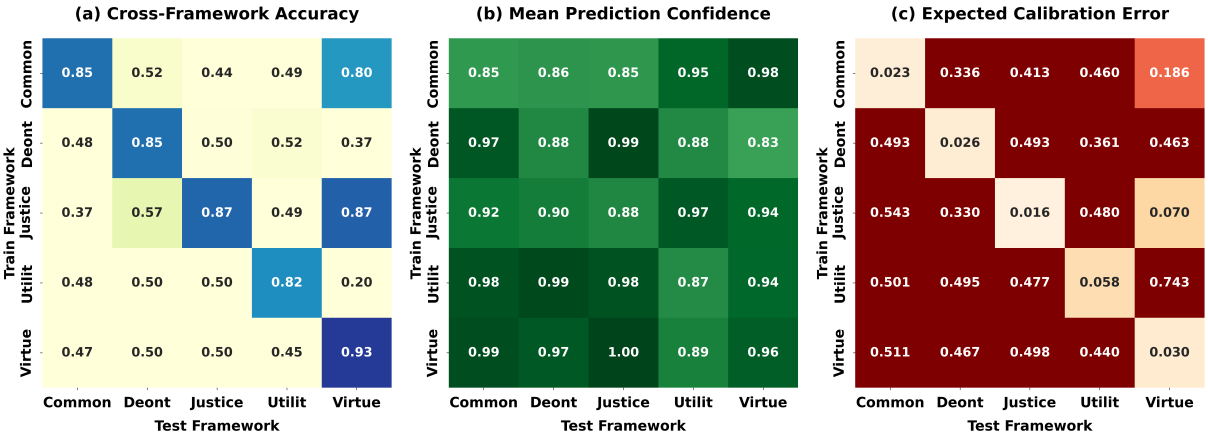
Qwen-2.5-7B–Instruct demonstrates family-specific architectural biases that persist across scale.



(a) DeepSeek-R1-Distill-7B



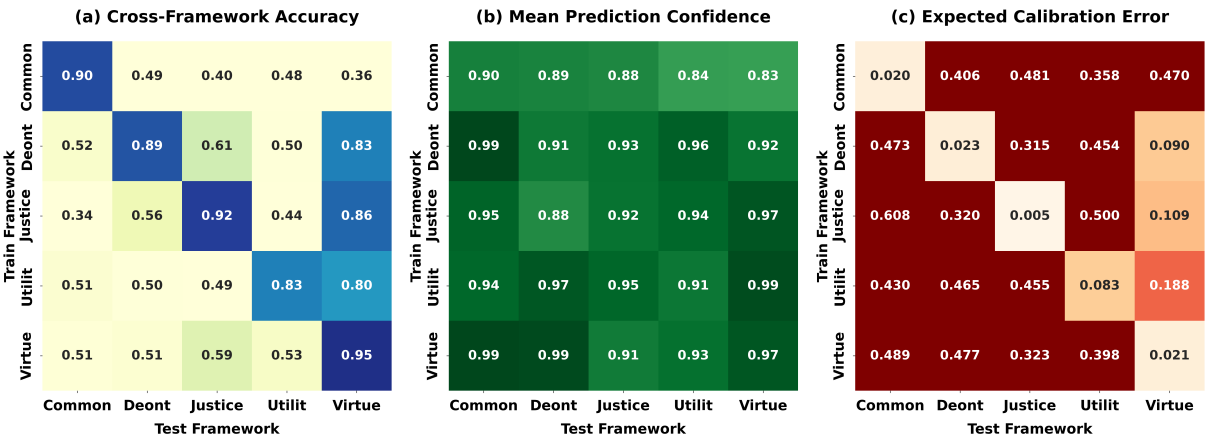
(b) Mistral-7B-Instruct



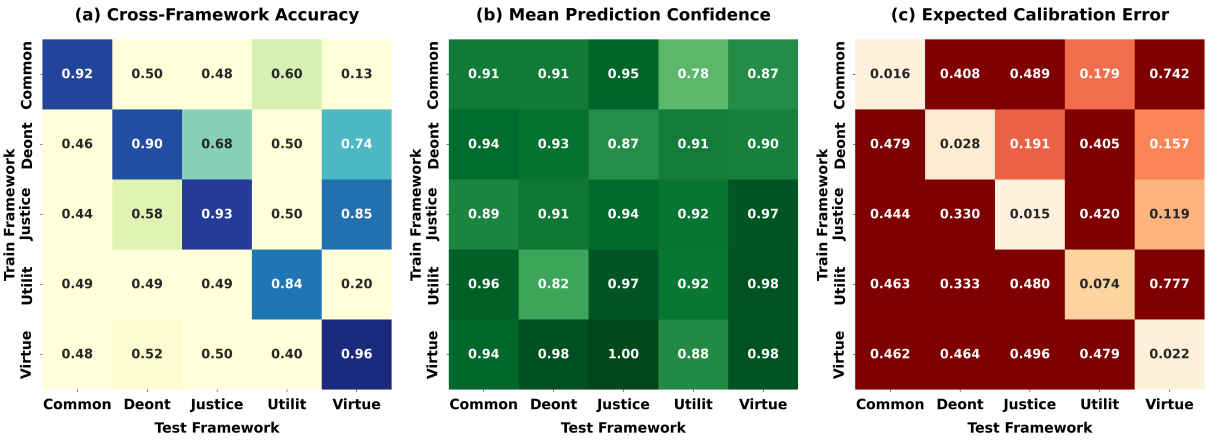
(c) Qwen-2.5-7B-Instruct

Figure 12: Mid-scale models (7B) exhibit family-specific failure patterns. DeepSeek-R1-Distill-7B positions commonsense as uniquely problematic across all transfers, potentially reflecting cultural-philosophical training biases. Mistral-7B-Instruct shows extreme virtue ethics isolation with catastrophic incoming transfer but moderate outgoing transfer. Qwen 2.5 demonstrates scale-invariant architectural biases—virtue isolation pattern identical to Qwen-2.5-72B-Instruct reveals training methodology creates systematic structure independent of capacity.

The virtue ethics isolation pattern nearly identical to Qwen-2.5-72B (10× larger) reveals that Qwen's training methodology creates systematic representational structure independent of capacity. This consistency suggests deliberate architectural or data curation choices rather than emergent properties—the same framework relationships appear regardless of parameter count. The scale-invariant failure modes



(a) Llama-3.3-70B-Instruct



(b) Qwen-2.5-72B-Instruct

Figure 13: Frontier-scale models (70B-72B) demonstrate capacity limits on resolving entanglement. Llama-70B shows accuracy-calibration dissociation—improved within-framework performance but persistent severe out-of-domain miscalibration. Qwen-72B exhibits the most catastrophic virtue isolation despite massive scale, preserving patterns from Qwen-7B—indicating systematic training biases that parameter count cannot overcome.

indicate that adding parameters within the Qwen architecture reinforces rather than corrects underlying biases—capacity increases refine within-framework performance while preserving problematic cross-framework structures.

F.4 Large-Scale Models (70B-72B Parameters)

Llama-3.3-70B-Instruct achieves the strongest overall performance but preserves fundamental entanglement patterns (Figure 13). Increased capacity enables more nuanced within-framework discrimination and moderately improved cross-framework transfer, yet miscalibration on failed transfers remains catastrophic. Justice and commonsense as problematic source frameworks persist from smaller models, suggesting these patterns reflect dataset properties or architectural constraints that scale does not address.

The critical insight: accuracy and calibration dissociate at scale. Llama achieves higher domain-specific accuracy while maintaining severe out-of-domain miscalibration, indicating standard training objectives optimize for confident correct responses without teaching distribution shift detection. This dissociation constrains alignment approaches assuming larger models naturally develop better uncertainty quantification—capacity improves performance but not epistemic awareness of framework boundaries.

Qwen-2.5-72B-Instruct demonstrates that extreme scale cannot overcome architectural or training methodology biases. Despite 72B parameters, the model exhibits the most catastrophic virtue ethics isolation across all evaluations—near-chance transfer accuracy with extreme miscalibration. The preservation of this pattern from Qwen-2.5-7B indicates systematic rather than capacity-limited

failures. If 72B parameters cannot resolve representational biases introduced during training, targeted interventions in data curation or architectural design become necessary rather than optional. Simply adding parameters amplifies existing structures—both beneficial patterns and problematic entanglements—without fundamentally restructuring ethical representation geometry.

F.5 Cross-Scale Insights and Alignment Implications

Three architectural insights emerge from cross-scale analysis. First, model families display consistent failure modes independent of scale—Qwen's virtue isolation, Mistral's strong asymmetries, and Llama's balanced yet entangled structure—indicating that training methodology shapes which frameworks interfere, while architecture governs interference magnitude. Second, instruction tuning amplifies within-framework confidence without improving cross-framework calibration, suggesting that RLHF favors decisiveness over calibrated uncertainty. Third, scaling enhances accuracy within frameworks but fails to mitigate inter-framework interference or enable detection of out-of-distribution ethical contexts.

These patterns impose fundamental limits on alignment. Expanding capacity alleviates low accuracy but not representational entanglement. Effective alignment will likely require architectural separation of ethical subspaces, objectives promoting distribution-shift awareness, or ensemble approaches combining specialized models rather than enforcing unified moral reasoning within a shared latent space.

The persistence of high-confidence miscalibration across scales and architectures is the most consequential result: models cannot detect when they apply inappropriate ethical frameworks. This failure mode appears invariant to capacity, architecture, and training, reflecting a deeper limitation in how language models represent normative uncertainty. Alignment efforts must confront this epistemic blindness directly rather than assuming scale or fine-tuning will induce it spontaneously.

G Probe Conflict and Behavioral Inconsistency in Mistral-7B

To assess whether the link between internal representational conflict and behavioral inconsistency generalizes across architectures, we replicated the conflict–entropy analysis on Mistral-7B-Instruct. Results are shown in Figure 14.

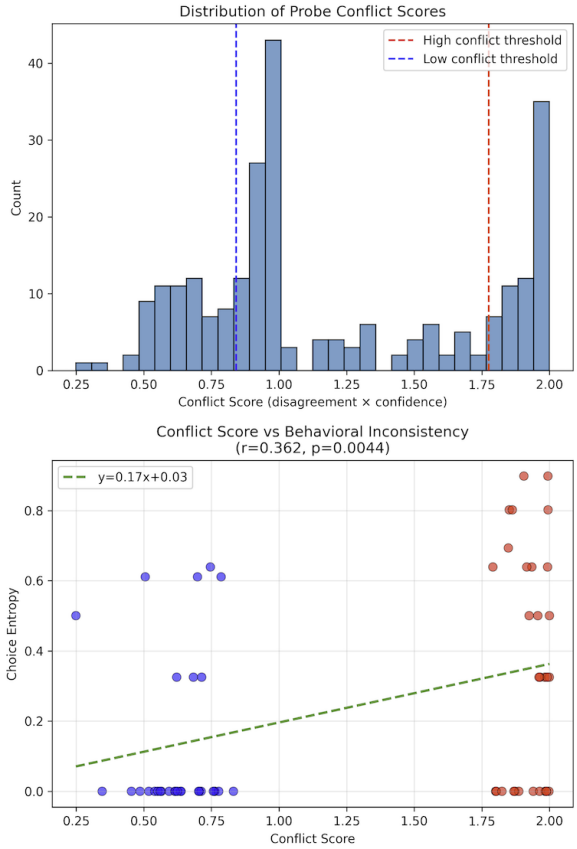


Figure 14: Probe conflict predicts behavioral inconsistency in Mistral-7B-Instruct. (Top) Distribution of conflict scores computed as the product of probe disagreement and minimum confidence. The bimodal pattern contrasts with the right-skewed distribution in Llama-3.3-70B, indicating stronger and more frequent representational disagreement between deontological and utilitarian encodings. (Bottom) Relationship between conflict scores and choice entropy under temperature sampling ($T = 1.2$). Scenarios with higher internal disagreement exhibit greater behavioral variability, reflecting that representational conflict propagates to output instability.

Unlike the heavily right-skewed distribution observed in LLAMA-3.3-70B, Mistral exhibits a bimodal conflict-score profile, with both low- and high-conflict regions densely populated. This indicates that Mistral's hidden representations preserve greater variability between deontological and utilitarian encodings rather than collapsing toward a single acceptability dimension. The model therefore sustains partially disentangled moral subspaces that come into tension more frequently during decision formation.

Consistent with this geometric interpretation, scenarios characterized by stronger probe disagree-

ment produce measurably higher behavioral entropy under temperature sampling. The relationship between conflict and inconsistency remains significant ($r = 0.36$, $p < 0.01$), suggesting that Mistral's internal ethical competition manifests more transparently in its outputs. In contrast to larger models whose alignment objectives suppress such tension, Mistral's mid-scale architecture appears to integrate competing frameworks dynamically rather than hierarchically.

This pattern carries two implications. First, behavioral variability in Mistral reflects genuine representational plurality rather than noise—conflict between ethical subsystems remains behaviorally expressive instead of being overwritten by final-layer convergence. Second, the stronger linear association between conflict and entropy implies that smaller or less heavily aligned models may offer clearer diagnostic windows into moral reasoning mechanisms. Probe-based auditing thus benefits from examining models at multiple scales, where internal competition is not yet masked by reinforcement-driven homogenization.