

Where Does an LLM Decide What’s Right?

Sparse Normative Circuits Across Languages and Model Families

Zhitao Huang¹ Qiang Huang^{2,4} Wei Zhai^{2,4}

Lemao Liu² Yingmao Tang^{3†} Dejing Dou^{2,4†}

¹School of Transnational Law, Peking University

²College of Computer Science and Artificial Intelligence, Fudan University

³Faculty of Law, Fudan University ⁴BEDI Cloud

[†]Co-corresponding authors: {yingmao_tang, doudejing}@fudan.edu.cn

Abstract

Where does an LLM decide what’s right? We study whether normative judgment is implemented by a sparse, identifiable substrate or distributed across the model. Conventional circuit analysis can conflate normative computation with the multiple-choice format machinery that surrounds it. We pair four normative datasets (legal and moral, Chinese and English) with a non-normative MMLU control and apply a **locate–disentangle–intervene** pipeline that sets aside format-processing heads. Across Qwen3-8B, Llama3.1-8B, and Mistral-7B-v0.1, this isolates 5–10 **normative-specific (NS) heads** per model (under 1% of total) in MCQA settings. Targeted knockout and amplification reveal bidirectional causal effects across legal and moral tasks. The resulting heads form a sparse, late-layer intervention substrate for normative MCQA. Their coordinates differ across model families, so cross-model agreement is structural and circuit findings require functional re-identification in each family.

1 Introduction

Suppose an LLM judges that dumping factory waste in a river is illegal. *Where* in the network is that judgment actually made? We study normative evaluation: how a model applies learned standards to determine what is lawful, harmful, or morally appropriate. Legal benchmarks characterize whether LLMs answer such questions correctly (Guha et al., 2023; Dehghani et al., 2025), but not which internal computations produce the judgment. If this computation is localized, inference-time intervention can alter normative judgments without broadly changing unrelated capabilities; if it is distributed, targeted control becomes harder.

Mechanistic evidence on normative judgment remains limited. Li et al. (2026) find that distinctions between opposed moral categories weaken through intermediate-layer representations. In a morally fine-tuned Gemma-2-2B model, Crosse

(2026) report no behavioral flips from 21,060 single-component swaps, while pathway-level interventions later in the network can alter choices. These studies examine moral representations or a single game-theoretic setting; they do not identify the causal components that support legal and moral judgments across tasks. Formatted MCQA adds a separate source of structure: Wiegrefe et al. (2025) show that answer-symbol prediction is mediated by sparse attention heads and use a synthetic task to separate formatted-MCQA competence from dataset-specific knowledge. A circuit located on legal or moral MCQA can therefore contain both normative evaluation and answer-format processing.

We address this gap with a controlled circuit design. We analyze legal and moral reasoning in Chinese and English across four datasets totaling 5,919 samples per model, replicated across three model families. A same-format control of 500 non-normative MMLU-STEM questions removes heads shared with general MCQA before causal testing. This design asks which components contribute specifically to legal and moral judgment, and whether the resulting mechanism recurs across normative domains, languages, and model families.

Our central finding, established in MCQA settings, is that path patching identifies candidate heads that overlap substantially with those driving non-normative MCQA. A *set-difference* filter against the MCQA control leaves 5–10 **normative-specific (NS) heads** per model (under 1% of total). Bidirectional intervention supports a causal contribution in the cells where the effect is detected: mean-ablation knockout meets the necessity criterion in 6/12 normative cells (Table 4), while amplification meets the sufficiency criterion in 10/12 and is directionally positive in 11/12 (Table 6). Cross-model agreement is *structural* (sparsity, late-layer concentration) but not identity-level: the NS sets occupy non-overlapping (layer, head) coordinates

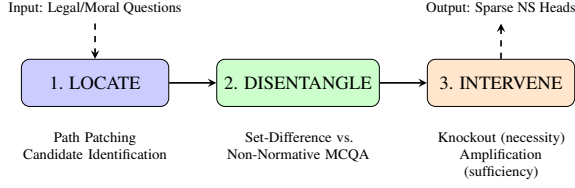


Figure 1: Our pipeline: (1) **Locate** candidate normative heads via path patching, (2) **Disentangle** normative-specific (NS) heads from general MCQA infrastructure via set-difference filtering, (3) **Intervene** bidirectionally — knockout to test *necessity*, amplification to test *sufficiency* — across three model families.

across the three families.¹

Our contributions are: (i) a *disentangle* step that uses same-format non-normative MCQA controls to set-subtract answer-format heads; (ii) cross-language (Chinese/English) and cross-family (Qwen3, Llama3.1, Mistral) identification of sparse NS head sets; (iii) bidirectional causal evidence via knockout (necessity) and amplification (sufficiency) under logit difference as the primary metric; and (iv) a cross-model analysis that separates structural recurrence (sparsity and late-layer concentration) from coordinate-level overlap.

Figure 1 summarizes our **locate–disentangle–intervene** pipeline: locate candidate heads on normative tasks, set-subtract heads shared with paired non-normative MCQA, and intervene on the remaining NS heads.

2 Related Work

Modularity and Circuits in Neural Networks. A growing body of work suggests that LLM capabilities are implemented by sparse, identifiable circuits rather than being diffusely distributed. Circuit-level analyses have revealed specialized subnetworks for indirect object identification (Wang et al., 2023), greater-than computation (Hanna et al., 2023), answer-label mapping (Lieberum et al., 2023), and arithmetic (Zhang et al., 2024a; Stolfo et al., 2023); automated circuit discovery has scaled this enterprise (Conmy et al., 2023; Syed et al., 2024), and Marks et al. (2025) identified feature-level circuits for ethically-relevant model behavior using sparse autoencoders (our component-level analysis is complementary to that SAE-feature direction). Mondorf et al. (2025) extend the modularity view by showing that circuits for related tasks share substructure and can be combined compositionally

across tasks. Our work extends the modularity question to normative reasoning in MCQA settings and tests whether legal and moral judgments depend on a sparse causal substrate. Most directly relevant, Wiegrefe et al. (2025) showed that MCQA answer prediction is mediated by a sparse set of attention heads with distinct functional roles; our set-difference analysis (Section 4.2) builds on this by further disentangling *normative-specific* evaluation from general MCQA infrastructure.

Mechanistic Studies of Moral Judgment. Li et al. (2026) find that distinctions between opposed moral categories weaken in intermediate-layer representations, while Crosse (2026) report zero behavioral flips across 21,060 single-component activation swaps. We study joint interventions on small head sets and measure changes in the answer-logit margin. Appendix 2.3 reports the corresponding accuracy-only and Bonferroni comparisons on our data.

Path Patching Methodology. Path patching (Goldowsky-Dill et al., 2023) extends causal mediation analysis (Pearl, 2001; Vig et al., 2020) to measure how individual components mediate model behavior. Attribution patching (AtP) (Nanda, 2023) provides an efficient gradient-based approximation. Zhang and Nanda (2024) and Heimersheim and Nanda (2024) established best practices for activation patching. We employ four complementary approaches: AtP, knockout, attention analysis, and sufficiency testing. Zhang et al. (2024a) applied path patching to arithmetic circuits, identifying sparse heads attending to operands and using effect-rank vs. random-rank knockout, attention-pattern classification, and MLP cosine-similarity probing to characterize the circuit. We adopt this validation protocol (knockout, attention analysis, MLP probing) and extend it with a paired non-normative MCQA control that enables *set-difference identification* of task-specific vs. format-generic heads — a step their single-domain design does not require. The resulting answer to a different question (“is normative evaluation separable from general MCQA?” rather than “which heads compute the answer?”) reveals a qualitatively different mechanism: attention-based evaluation rather than MLP-based computation.

Normative Reasoning in LLMs. Benchmarks and surveys have characterized model performance on legal tasks (Zhong et al., 2020; Chalkidis et al., 2022; Guha et al., 2023; Dehghani et al., 2025; Niklaus et al., 2023; Fan et al., 2026) and moral

¹Code: <https://github.com/hzt12345/normative-circuits>

judgment (Scherrer et al., 2023; Zhang et al., 2024b). Neuro-symbolic methods have also been proposed for legal reasoning (Kant et al., 2025). These works evaluate behavior or propose reasoning systems, but do not identify the internal computations that produce legal and moral judgments. We address this mechanistic question at the circuit level in MCQA settings.

Cross-Lingual Representations. Multilingual pretrained language models exhibit cross-lingual transfer (Pires et al., 2019; Conneau et al., 2020), but Ifergan et al. (2024) showed that behavioral consistency does not necessarily imply shared internal representations, and Lim et al. (2025) further found that language-specific latent processes can hinder cross-lingual performance—suggesting that cross-lingual behavior may rely on partially language-specific rather than fully shared representations. Whether *reasoning circuits* can nonetheless be shared across languages remained unexplored; we provide circuit-level evidence under MCQA that normative evaluation relies on language-agnostic substrates.

3 Methodology

3.1 Task and Dataset

We focus on normative reasoning tasks, which involve evaluating scenarios against legal or moral standards. Taking legal reasoning as an example, the task requires determining which answer option correctly reflects the applicable law. We analyze four datasets spanning legal and moral reasoning in Chinese and English; Table 1 summarizes the corpora.

Table 1: Dataset statistics. Samples denote full corpus sizes used for path patching; English Moral intervention analyses use a stratified $N = 500$ subset to remove answer-label bias (Appendix 1.2).

Dataset	Samples	Source	Language
Chinese Legal	1,999	JEC-QA	Chinese
English Legal	619	LEXam	English
Chinese Moral	1,934	SafetyBench	Chinese
English Moral	1,367	MoralChoice	English
Total	5,919		

JEC-QA (Zhong et al., 2020) contains multiple-choice questions from China’s National Judicial Examination, testing knowledge of Chinese law including civil, criminal, and administrative law. **LEXam** (Fan et al., 2026) contains law examina-

tion questions from multiple jurisdictions; we use the English subset. **SafetyBench** (Zhang et al., 2024b) evaluates LLM safety; we use the Ethics and Morality category with multiple-choice questions about moral judgments. **MoralChoice** (Scherrer et al., 2023) contains moral dilemma scenarios with varying levels of ambiguity.

To create the counterfactual prompts required for path patching, we mask key domain terms with neutral placeholders. The implemented vocabularies contain 78 Chinese-legal entries (74 unique), 51 English-legal entries (50 unique), 28 English-moral entries, and 36 Chinese-moral entries (Appendix 8). For each sample, 30% of matched terms are replaced in the legal and English-moral datasets; the Chinese-moral preprocessing uses 50%. Replacements are random bold single-letter placeholders such as ****V**** or ****E****, which preserve approximate token count while destroying semantic content.

3.2 Path Patching

We employ path patching (Goldowsky-Dill et al., 2023; Wang et al., 2023), a causal intervention technique for identifying functionally important components. The computation of an LLM can be modeled as a directed acyclic graph (DAG), where nodes are computation components such as attention heads, MLP layers, and residual connections, and edges represent data flow. By replacing a component’s activation with its counterfactual value, we measure its causal effect on the output.

For each (layer, head) we replace its activation on the reference run with the counterfactual activation, then score the change in $P(y \mid X_r)$ at the END position; averaging across samples ranks heads by causal effect (full algorithm in Appendix 1.3). Downstream causal validation (knockout in Sec. 4.2, amplification in Sec. 4.5) uses logit difference at the END position rather than probability, following Wang et al. (2023); Zhang et al. (2024a); Zhang and Nanda (2024).

For efficiency, we implement Attribution Patching (AtP) (Nanda, 2023), a gradient-based approximation that replaces $L \times H$ separate forward passes (exact patching) with two forward passes and one backward pass (constant in L, H):

$$\hat{E}_{l,h} = (A_c(l, h) - A_r(l, h)) \cdot \nabla_{A(l,h)} \mathcal{L} \big|_{A=A_r} \quad (1)$$

where A_r and A_c are head (l, h) ’s output activation at the END position on the reference and

counterfactual prompts respectively, $\cdot$ denotes the Euclidean inner product, and $\mathcal{L} = \log P(y \mid X_r)$ is the log-probability of the correct answer evaluated on the reference run. Equation 1 is the first-order Taylor approximation of $\mathcal{L}_{\text{patched}} - \mathcal{L}_{\text{reference}}$ around A_r , so larger $|\hat{E}_{l,h}|$ indicates greater causal importance.

3.3 Validation Methods

Knockout Validation. To validate the faithfulness of the discovered key heads, we observe the performance drop when knocking out these components. Following the mean-ablation protocols of Wang et al. (2023); Zhang et al. (2024a), each target head’s output at every position is replaced with the dataset-mean activation computed across a 100-prompt counterfactual pool (incorrect-prompt activations with placeholder-masked normative tokens). We compare two conditions: (1) *effect-rank knockout*, removing heads in order of importance score; (2) *random-rank knockout*, removing size-matched random heads (5 trials per cell, drawn i.i.d. without replacement from the non-NS head pool). For each cell we report logit difference ($\text{logit}_y - \max_{y' \neq y} \text{logit}_{y'}$ at the END position) together with argmax accuracy. Logit difference preserves changes in the answer margin that may not cross the discrete accuracy threshold (Zhang and Nanda, 2024); we therefore use it as the primary inferential metric. We define an NS-vs-random gap as significant when $|\Delta \text{logit-diff}| \geq 0.05$ in the predicted direction with trial-level standardized statistic $|s| \geq 2$ (computed over 5 random trials; the strict one-sided t threshold at $\text{df}=4$ is $t_{4,0.95} = 2.13$).² This is a study-specific operational criterion: the 0.05 threshold is $\approx 2 \times$ the trial-std we observe ($s_{\text{trial}} = 0.025\text{--}0.04$). Threshold robustness across logit-difference thresholds, trial-statistic cutoffs, and supplementary accuracy thresholds is reported in Appendix 2.4; resampling-ablation is reported in Appendix 6.3.

Attention Pattern Analysis. To understand the working mechanism of identified heads, we examine their attention patterns. For each head, we extract the last row of the attention matrix $A_{l,h} \in \mathbb{R}^{N \times N}$, analyzing the attention scores between the Query token at the END position and each Key token. We classify heads into func-

tional categories based on their attention distribution across token types and positions.

3.4 Models

To ensure our findings generalize across architectures, we conduct experiments on three open-source LLMs from different model families (Table 2):

Table 2: Model specifications.

Model	Params	Layers	Heads/L
Qwen3-8B	8B	36	32
Llama3.1-8B	8B	32	32
Mistral-7B-v0.1	7B	32	32

We use TransformerLens (Nanda and Bloom, 2022) for model internals access. We present detailed results for Qwen3-8B in the main text, with cross-model validation in Section 4.4.

4 Experiments

We organize experiments around three questions: (1) Can we locate a sparse set of normative heads and separate them from general task infrastructure? (2) Is this normative module shared across domains and languages? (3) Can it be causally intervened upon in both directions? We then characterize the module’s internal mechanism and verify cross-model generality.

4.1 Locating the Normative Module

Applying AtP to all four datasets, we observe two patterns supporting modularity. **Sparsity:** classifying heads with $|\overline{E}_{l,h}| > 0.20$ as “key” yields $< 5\%$ of heads on three of four datasets (CN Legal 3.0%, EN Legal 2.7%, CN Moral 5.5%) and 11.3% on EN Moral. **Spatial concentration:** key heads concentrate in late layers (22–35 on Qwen3-8B). Head (35, 25) ranks first on three of four datasets (CN Legal 0.89, EN Legal 0.74, CN Moral 0.93); EN Moral is led by (23, 27) at 1.118 (full per-domain top-10 in Appendix 4.2; heatmap in Appendix 4.1). Pairwise Jaccard between top-20 heads (Appendix 5.1) reveals structured cross-domain sharing: same-language normative pairs overlap substantially (CN legal–moral 0.48, CN–EN legal 0.43), while EN Moral overlaps less (0.14–0.21) with others. This overlap is too high to reflect independent circuits but too structured to be a single monolithic mechanism, motivating the set-difference filter against general MCQA below.

²The trial-level statistic is $(\bar{x}_{\text{NS}} - \bar{x}_{\text{rand}}) / (s_{\text{rand}} / \sqrt{5})$. Tables report this statistic as s (or s_{ld} for logit difference) because it is computed from only 5 random trials, not as a formal large-sample z -test.

4.2 Separability from General Capabilities

The heads identified in Section 4.1 could reflect normative-specific processing or generic answer-selection machinery. To distinguish these possibilities, we pursue two complementary strategies: (1) controlled comparison against non-normative MCQA that disentangles normative-specific from format-processing heads, and (2) selective knockout of the resulting normative-specific (NS) head set to establish causal necessity.

Identifying Normative-Specific (NS) Heads via Set-Difference. To isolate normative-specific from generic-task heads, we run our full pipeline on 500 MMLU-STEM samples as a non-normative MCQA control (English-only; a Chinese non-normative MCQA control of comparable scale is not available and is acknowledged in Limitations; masking detailed in Appendix 1.2). The pairwise Jaccard between general MCQA and normative top-20 heads averages 0.44, confirming substantial shared MCQA infrastructure that must be subtracted before normative claims. The **NS head set** is defined as heads appearing in ≥ 2 normative datasets’ top-20 but absent from the general MCQA top-20: 7 heads on Qwen3-8B, 5 on Llama3.1-8B, 8 on Mistral-7B-v0.1, with mid-to-late layer ranges but no (layer, head) overlap across families (structural distribution in Section 4.4; full per-model NS head tuples released with the supplementary code). Table 3 reports the set-difference counts per dataset.

Table 3: Set-difference analysis: normative-specific heads not in the general-MCQA top-20, using 500 MMLU-STEM samples as the control. Specificity = fraction of normative top-20 absent from general top-20. Layer distribution uses E=early, M=middle, Lm=late-mid, F=final. Bootstrap CIs over 1,000 iterations are reported in Appendix 6.1.

Dataset	Shared	Specific	Spec.	Layer Dist.
CN Legal	12	8	40%	M3,Lm2,F2,E1
EN Legal	14	6	30%	M2,Lm2,F2
CN Moral	15	5	25%	M4,Lm1
EN Moral	6	14	70%	M5,Lm5,E3,F1

Selective Knockout: NS Heads Are Causally Necessary. To test whether NS heads form a necessary substrate, we knock them out and compare against size-matched random-rank knockout, scaling to $N = 500$ samples per cell and replicating across all three model families. For normative datasets, per-cell NS-head set sizes follow the set-difference analysis ($K = 5\text{--}10$ heads per model);

general-MCQA controls use their own matched control-set sizes ($K = 22\text{--}25$). Random-rank uses K heads sampled outside the union of all NS heads, with 5 independent trials per cell. We use logit difference ($\text{logit}_y - \max_{y' \neq y} \text{logit}_{y'}$ at the END position) as the primary inferential metric and declare an NS necessity effect when the NS-vs-random gap satisfies $|\Delta \text{logit-diff}| \geq 0.05$ in the predicted direction with trial-level $s_{\text{ld}} \leq -2$; we report $\Delta \text{accuracy}$ alongside as descriptive context. We use the mean-ablation protocol of Wang et al. (2023); Zhang et al. (2024a) as the primary intervention, replacing each target head’s output at every position with the dataset-mean activation from a 100-prompt counterfactual pool; resampling-ablation is reported as a robustness check in Appendix 6.3.

6 of 12 normative cells satisfy the NS-necessity criterion on logit difference. Broken down by model: **Qwen3** 3/4 (all except CN Moral), **Llama** 1/4 (CN Moral with sub-chance caveat), **Mistral** 2/4 (EN Legal and EN Moral). Of the six non-significant normative cells, one is just below the magnitude threshold despite a strong trial-level statistic (Mistral CN Legal, $|\Delta \text{logit-diff}| = 0.043$, $s = -5.6$), two sit at or near chance baselines with limited headroom (Llama EN Legal and Qwen3 CN Moral), two show wrong-direction signals where NS knockout improves rather than degrades performance (Llama EN Moral and Mistral CN Moral), and one shows a correct-direction accuracy gap but near-zero logit-difference signal (Llama CN Legal). The wrong-direction outliers correspond to cells where baseline accuracy is below or near the random-chance line for the answer format, indicating the NS heads we identified may be encoding an anti-correlated signal in those particular cells rather than the normative direction. Section 7 discusses sub-chance and floor confounds in more depth.

Cross-model consistency on English Moral. The largest and most consistent necessity signal is on English Moral: -7.3 to -8.0 pp accuracy drops and -0.34 to -0.46 logit-difference drops on Qwen3 and Mistral. The Llama EN Moral wrong-direction result is the lone exception, a model-family-specific outlier we discuss in Section 4.5.

Layer-matched random control. To test whether late-layer placement alone explains the NS effect, we ran an exact per-layer-matched random control across all 3 models $\times$ 5 datasets (500 samples per cell, five random trials, zero matching fallbacks). Across the 12 normative cells, NS

Table 4: Selective NS-head knockout under mean ablation across three models on 4 normative datasets and general MCQA ($N = 500$ per cell; 5 random trials). Δacc and $\Delta\text{logit-diff}$ are NS-vs-random gaps; s_{ld} is the trial-level standardized statistic for logit-diff. **S** = primary significance under our logit-difference criterion ($|\Delta\text{logit-diff}| \geq 0.05$ correct direction, $|s_{\text{ld}}| \geq 2$); **normative S-marked cells = 6/12** (general-MCQA controls: 2/3). **L** = Bonferroni-corrected significant on logit-diff at $p_{\text{Bonf}} < 0.05$ over $N = 30$ tests (2 metrics $\times$ 15 cells); **a** = Bonferroni-significant on accuracy. **o** = wrong-direction signal (sub-chance baseline or model-specific outlier). † baseline near 4-choice chance; effect meaningful in logit-diff but accuracy bounded. ‡ baseline below 2-choice chance. Mistral CN Legal ($|\Delta\text{ld}| = 0.043$, $s = -5.6$) is reported for transparency but fails the 0.05 magnitude threshold and is therefore *not* counted toward 6/12.

Model	Dataset	K	Baseline	NS acc	Rand acc	Δacc	$\Delta\text{logit-diff}$	s_{ld}	Sig
Qwen3-8B	CN Legal	8	0.53	0.494	0.508	-1.4	-0.090	-5.5	S
	EN Legal	6	0.32	0.296	0.316	-2.0	-0.155	-8.3	S
	CN Moral	5	0.47	0.472	0.476	-0.4	+0.049	+2.1	—
	EN Moral	10	0.64	0.570	0.650	-8.0	-0.337	-16.2	SL
	Gen MCQA	22	0.48	0.378	0.439	-6.1	-0.320	-6.2	S
Llama3.1-8B	CN Legal	7	0.52	0.492	0.528	-3.6	+0.006	+0.4	—
	EN Legal	7	0.26	0.244	0.246	-0.2	+0.132 †	+10.5	o
	CN Moral	6	0.28	0.274	0.298	-2.4	-0.083	-2.5	S ‡
	EN Moral	10	0.44	0.524	0.439	+8.5	+0.133	+5.6	o
	Gen MCQA	24	0.26	0.250	0.251	-0.1	+0.011	+0.8	—
Mistral-7B	CN Legal	9	0.49	0.500	0.496	+0.4	-0.043	-5.6	—
	EN Legal	8	0.22	0.218	0.226	-0.8	-0.060 †	-9.1	S †
	CN Moral	9	0.26	0.308	0.262	+4.6	+0.127	+6.2	o
	EN Moral	9	0.40	0.334	0.407	-7.3	-0.460	-14.4	SaL
	Gen MCQA	25	0.25	0.226	0.249	-2.3	-0.156	-6.5	S

ablation is more damaging than same-layer random ablation in 8/12 by accuracy and 7/12 by logit difference. The clearest separations—Qwen English Legal (-2.3pp, $\Delta\text{logit-diff} = -0.184$), Qwen English Moral (-8.0pp, -0.406), and Mistral English Moral (-6.2pp, -0.422)—show trial-level $p \leq 0.02$ on both metrics (five trials per cell). Three near- or below-chance cells reverse on both metrics, consistent with the floor confounds discussed in Section 7; Appendix 2.1 reports every cell.

Calibration of the effect-size threshold. Our criterion $|\Delta\text{logit-diff}| \geq 0.05$ (correct direction, trial-level $s_{\text{ld}} \leq -2$) is calibrated empirically to our observed trial-level noise: the 0.05 magnitude threshold is approximately $2\times$ the random-trial standard deviation we observe (typical trial-std 0.025–0.04 across our 5 random head-set draws per cell), so it corresponds to a ≥ 2 -sigma effect on top of the trial-level $s_{\text{ld}} \leq -2$ requirement. Logit difference is used in prior circuit analyses (Wang et al., 2023; Zhang et al., 2024a), and its continuous margin is more informative here than accuracy alone. We therefore treat it as primary and report $\Delta\text{accuracy}$ as descriptive context. As a robustness check on cutoff choice, Appendix 2.4 reports an auxiliary grid over logit-difference thresholds, trial-statistic cutoffs, and supplementary accuracy

Table 5: Alternative explanations and the evidence that addresses each. “Controlled” = directly tested; “scoped” = acknowledged as a limitation.

Alternative	Evidence	Status
Generic MCQA	Set-difference vs. MMLU controls (Table 3)	Controlled
Answer-label bias	A/B-balanced EN Moral subset	Controlled
Late-layer relevance	Per-layer-matched random control	Controlled
Domain vocabulary	Legal/moral CN/EN design	Controlled
Task difficulty	Per-cell baselines reported	Scoped

thresholds.

K-sweep stability. A per-cell K -sweep over $\{1, 2, 3, 5, K_{\text{NS}}\}$ (plus $K = 10$, K_{full} where applicable) finds 7/12 normative cells K-stable on an accuracy-based corroborating criterion (random-minus-NS accuracy gap $\geq 1\text{pp}$, $s \geq 2$ at ≥ 1 swept K ; Appendix 2.5), with the strongest normative sweep on Mistral English Moral at $K = 1$ (+8.1pp, $s = 59.9$). The K-stable set substantially overlaps the fixed- K 6/12 logit-difference set; we read it as accuracy-side corroboration that the necessity signal is not specific to a particular head-set size, rather than as a second inferential criterion.

4.3 Internal Mechanism: Attention-Based Evaluation

NS heads concentrate attention on answer-option tokens at $\sim 5\times$ the uniform-baseline rate. In transformer terms, this means the identified heads primarily route information from the candidate-answer region into the final decision position. By contrast, measured MLP contributions—which would indicate stronger per-position nonlinear feature transformation after that information has been routed—are two orders of magnitude below the attention-head scale, and the late-layer MLP cosine-similarity signal (Zhang et al., 2024a) is far weaker here than on arithmetic. We therefore read the mechanism as suggestive evidence for attention-based *evaluation* (selectively reading candidate options and comparing them against learned normative standards) rather than progressive MLP-based *computation*; per-head attention breakdowns (Table 15; Figure 4), MLP path-patching effects, and cosine-similarity curves are reported in Appendices 4.3 and 4.4.

4.4 Cross-Model Generality

We replicate path patching on Llama3.1-8B and Mistral-7B-v0.1 to test whether the structural signature (sparsity, late-layer concentration, cross-domain sharing) generalizes. All three models place $\geq 50\%$ of top-20 heads in the final third of layers (Qwen3 62.5%, Llama 53.8%, Mistral 50.0%, average across 4 datasets), and all three show within-model Legal-vs-Moral grouped Jacard of 0.31–0.44 (per-model breakdowns in Appendix 5.2). However, no individual NS head is shared across all three families. Cross-model agreement is therefore structural (sparsity, late concentration, NS fraction 0.5–0.8%) but not identity-level.

4.5 Sufficiency Replicates Across Models

If NS heads encode normative computation, amplifying them on previously-incorrect samples should recover correct answers more often than amplifying random heads of the same count. We test this across three models and four datasets at $\alpha = 2.0$ (chosen to roughly double NS-head output contribution while avoiding distribution collapse observed at $\alpha \geq 3$; we report a single α for tractability and matched random-head comparison), on $N = 500$ originally-incorrect samples per cell, with random controls averaged across 5 trials. Per-cell two-proportion z -tests are Bonferroni-corrected across

$N = 12$ cells (Appendix 3.1). Table 6 reports the main result.

Appendix 3.2 reports the corresponding comparison against universal infrastructure heads.

Across the 12 model–dataset cells, NS amplification’s point-estimate recovery rate exceeds random’s in 11 cells (11/12 directional), with 10/12 cells showing advantage $\geq 1.5\times$ (range 1.8–5.4 $\times$ among those cells). The single exception (Llama EN Moral) is a degenerate cell where amplification recovers no incorrect samples for either condition, consistent with English Moral being the hardest cell for Llama at the model’s baseline. The 11-of-12 directional consistency holds across cells: a binomial sign-test against the null that the sign of (NS – Random) is independently Bernoulli(0.5) gives $p \approx 0.003$. At the individual cell level, 6 cells survive Bonferroni-corrected significance (Qwen3 CN Legal, Qwen3 CN Moral, Qwen3 EN Moral, Llama CN Legal, Llama EN Legal, Mistral EN Moral); the remaining 5 positive cells do not survive the corrected per-cell test, mostly where the absolute recovery-rate gap is only a few percentage points.

This is consistent with our knockout findings: amplification’s matched accuracy-based effect-size criterion is met in 10 of 12 normative cells, and knockout’s logit-difference criterion is met in 6 of 12 normative cells. The same sparse substrate is recruited bidirectionally: amplification and ablation both produce NS-vs-random gaps on the cells where measurement is unconfounded by floor/sub-chance baselines or model-specific outliers.

5 Discussion

Bidirectional causal evidence under a logit-difference criterion. Sections 4.2–4.5 establish convergent causal evidence: NS-head amplification meets the matched effect-size criterion in 10/12 normative cells (11/12 directionally positive, sign-test $p \approx 0.003$; Table 6), and NS-head mean-ablation produces NS-vs-random logit-difference gaps satisfying the criterion in 6/12 normative cells (Table 4). The strongest single result is cross-model consistency on English Moral (–7 to –8pp accuracy and –0.34 to –0.46 logit-diff on Qwen3 and Mistral).

Cross-model agreement is structural, not identity. The three model families converge on a clear structural signature: NS fraction 0.5–0.8%, late-layer concentration $\geq 50\%$ of top-20 heads in the final third, and the same bidirectional-

Table 6: Sufficiency via NS-head amplification across three models ($\alpha = 2.0$, $N = 500$ originally-incorrect samples per cell). Recovery = % of incorrect answers flipped to correct after amplification. Adv. = NS/Random (point-estimate advantage ratio). $\checkmark$ marks Bonferroni-significant cells ($p_{\text{Bonf}} < 0.05$, $N = 12$); $\circ$ marks point-estimate-positive cells that do not survive Bonferroni; $\bigcirc$ marks the null cell. Per-cell z -tests in Appendix 3.1.

Model	Metric	CN Legal	CN Moral	EN Legal	EN Moral
Qwen3-8B	NS / Random recovery	18 / 8.6 ± 2.9	17 / 7.6 ± 4.5	2 / 1.0 ± 0.9	13 / 2.4 ± 2.0
	Advantage	$2.1\times$	$2.2\times$	$2.0\times$	$5.4\times$
	Bonf-significance	$\checkmark$ ($p=1.4\times 10^{-4}$)	$\checkmark$ ($p=7.2\times 10^{-5}$)	$\circ$	$\checkmark$ ($p=3.9\times 10^{-9}$)
Llama3.1-8B	NS / Random recovery	9 / 3.4 ± 0.5	3 / 1.0 ± 1.1	6 / 2.2 ± 1.5	0 / 0.0 ± 0.0
	Advantage	$2.6\times$	$3.0\times$	$2.7\times$	$\bigcirc$ null
	Bonf-significance	$\checkmark$ ($p=0.0029$)	$\circ$	$\checkmark$ ($p=0.029$)	N/A
Mistral-7B	NS / Random recovery	5 / 2.8 ± 1.5	6 / 3.0 ± 1.9	3 / 2.4 ± 1.4	14 / 4.2 ± 2.2
	Advantage	$1.8\times$	$2.0\times$	$1.25\times$	$3.3\times$
	Bonf-significance	$\circ$	$\circ$	$\circ$	$\checkmark$ ($p=8.6\times 10^{-7}$)

intervention pattern. But their NS head sets do not overlap: Qwen3’s 7 NS heads, Llama’s 5, and Mistral’s 8 are entirely distinct sets in (layer, head) coordinates. This rules out a strong “universal normative circuit transferable across model families” claim and recommends humility about interpretability transfer in general: a single-model finding about which heads encode a function should not be assumed to identify those same heads in a sibling model, even when the structural signature is preserved.

Attention-based evaluation vs. MLP-based computation. NS heads concentrate attention on answer-option regions at $\sim 5\times$ uniform baseline, meaning they route candidate-answer information into the final decision position. MLP path-patching effects on the correct-answer logit are 10^{-3} or smaller across all datasets, two orders of magnitude below the attention scale; this would be the place to expect a stronger signal if per-position nonlinear feature transformation were the dominant mechanism. Because a patched activation can affect the output through downstream computation, these are causal path-patching effects rather than direct effects in isolation. Their contrast with the stronger MLP signal reported by Zhang et al. (2024a) on arithmetic suggests that normative evaluation is more attention-based than MLP-computed: the model appears to selectively read candidate options and compare them against learned standards.

Broader implications for alignment and interpretability. The cross-dataset NS head sets are sparse (7/5/8 heads on Qwen/Llama/Mistral), yet no individual head coordinate is shared across families. For alignment interventions, this means a normative-steering site found in one model cannot be transferred to another by head index; the rele-

vant components must be functionally re-identified in each architecture. Cross-model mechanistic transfer should therefore focus on reliable re-identification protocols rather than fixed coordinates. At the 7–8B scale, exhaustive component-level analysis remains feasible—the present study intervenes on every attention head across three families—but whether the sparse structural signature persists at larger scales is an open empirical question that we flag as a priority for future work.

A circuit-level account of legal judgment. On JEC-QA and LEXam, our results identify a sparse late-layer attention substrate that contributes to legal judgment. These heads concentrate on answer-option regions and route candidate information to the final decision position; measured MLP contributions are two orders of magnitude smaller. Knock-out and amplification establish their causal relevance in the supported cells. This provides an account of which internal computations contribute to legal answers alongside the usual behavioral accuracy results.

Why this matters for law. Legal reasoning applies rules, exceptions, and institutional categories to facts. Circuit analysis allows us to ask whether different languages and model families reach legal answers through comparable internal organization. We find the same sparse, late-layer structure but different head coordinates. Legal-AI audits therefore need functional identification and intervention tests for each model; accuracy scores and transferred head indices alone cannot supply that evidence.

6 Conclusion

Across three 7–8B model families and four Chinese/English normative MCQA datasets, we iden-

tify 5–10 normative-specific (NS) attention heads per model (0.5–0.8% of total). Knockout meets the necessity criterion in 6/12 normative cells, while amplification meets the sufficiency criterion in 10/12 and is directionally positive in 11/12 (sign-test $p \approx 0.003$) under logit difference as the primary metric. The convergent supported cells identify a sparse intervention substrate for normative MCQA; variable and wrong-direction cells delimit where that substrate claim holds.

Methodologically, the paired non-normative MCQA control and set-difference filter cleanly separate normative-specific from format-processing infrastructure, a step prior single-task circuit analyses do not require, and one we recommend for future mech-interp evaluation on MCQA-format tasks. Cross-family agreement is structural (sparsity, late-layer concentration, NS fraction) but not identity-level: the (layer, head) coordinates differ across families, constraining how circuit-level findings transfer between sibling models. For legal AI, these findings provide a mechanistic account of legal judgment: a small set of late-layer attention heads routes candidate legal answers to the final decision position.

Limitations

Scale, format, and controls. All evidence is at the 7–8B scale on three families; larger models, instruction-tuned variants, and different post-training pipelines may yield different NS-head structures. All datasets use MCQA with single-token answer letters; open-ended generation, multi-turn dialogue, and chain-of-thought are untested. The non-normative MMLU-STEM control is English-only; a Chinese control (e.g., C-Eval STEM) would tighten the NS set on Chinese cells.

Approximation. AtP is first-order; we validate via exact patching on a Chinese Legal subset ($\rho=0.899$, $p<10^{-10}$) but not on every cell.

Statistical power. Five random trials per cell (t on $df=4$) limits Bonferroni-corrected power; the 6/12 count uses our logit-difference effect criterion, while strict Bonferroni yields 3/12 (Appendix 2.2). The sign-test ($p\approx 0.003$) assumes cell independence; collapsing to 4 dataset-level signs gives $p=0.0625$. Two cells show wrong-direction signals classified post-hoc as model-specific outliers.

Ethical Considerations

This work analyzes attention-head circuits in publicly released LLMs (Qwen3-8B, Llama3.1-8B, Mistral-7B-v0.1) on publicly available evaluation datasets (JEC-QA, LEXam, SafetyBench, MoralChoice, MMLU). We do not collect new user data or human annotations. The intended use of NS heads is for alignment research—studying how normative judgment is structured and whether sparse interventions can steer it. We acknowledge that NS-head identification could in principle be misused to amplify undesired normative tendencies (e.g., biased moral judgments); we view this risk as comparable to general activation-steering work (Zou et al., 2023). Reproduction code is available at <https://github.com/hzt12345/normative-circuits/>.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2024YFC3307304).

References

- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.
- Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Emerging cross-lingual structure in pretrained language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.
- Tyler Crosse. 2026. Mechanistic interpretability of moral fine-tuning. Blog post, <https://tylercrosse.com/ideas/2026/llm-morality-mech-interp/>.
- Fatemeh Dehghani, Roya Dehghani, Yazdan Naderzadeh Ardebili, and Shahryar Rahnamayan. 2025. [Large language models in legal systems: A survey](#). *Humanities and Social Sciences Communications*, 12(1).
- Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2026. [LEXam: Benchmarking legal reasoning on 340 law exams](#). In *International Conference on Learning Representations*.
- Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*.
- Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.
- Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*.
- Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.
- Maxim Ifergan, Leshem Choshen, Roei Aharoni, Idan Szpektor, and Omri Abend. 2024. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in llms. *arXiv preprint arXiv:2408.10646*.
- Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. 2025. Towards robust legal reasoning: Harnessing logical LLMs in law. *arXiv preprint arXiv:2502.17638*.
- Lingyu Li, Yan Teng, and Yingchun Wang. 2026. Mechanistic origin of moral indifference in language models. *arXiv preprint arXiv:2603.15615*.
- Tom Lieberum, Matthew Rahtz, János Krámar, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*.
- Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance. *arXiv preprint arXiv:2505.13141*.
- Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*. ArXiv preprint arXiv:2403.19647 (2024).
- Joseph Miller, Bilal Chughtai, and William Saunders. 2024. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling (COLM)*.
- Philipp Mondorf, Sondre Wold, and Barbara Plank. 2025. Circuit compositions: Exploring modular structures in transformer-based language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14934–14955.
- Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. <https://www.neelnanda.io/mechanistic-interpretability/attribution-patching>.
- Neel Nanda and Joseph Bloom. 2022. TransformerLens. <https://github.com/TransformerLensOrg/TransformerLens>. Software library.
- Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. [LEXtreme: A multi-lingual and multi-task benchmark for the legal domain](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054.

- Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 411–420.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.
- Nino Scherrer, Claudia Shi, Amir Feder, and David M Blei. 2023. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems*.
- Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.
- Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.
- Kevin R Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.
- Sarah Wiegrefe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. 2025. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *International Conference on Learning Representations*.
- Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*.
- Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. 2024a. Interpreting and improving large language models in arithmetic calculation. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR.
- Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024b. *SafetyBench: Evaluating the safety of large language models*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553.
- Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

1 Data and Implementation Details

1.1 Dataset Examples

1.1.1 Chinese Legal (JEC-QA)

问题: 某公司与张某签订了一份劳动合同, 约定试用期为6个月。

根据《劳动合同法》, 下列说法正确的是:

- A. 试用期约定合法
- B. 试用期超过法定最长期限
- C. 试用期工资可低于最低工资标准
- D. 试用期可随意延长

答案: B

1.1.2 English Legal (LEXam)

Question: A contract is formed when:

- A. One party makes an offer
- B. There is offer, acceptance, and consideration
- C. Both parties sign a written document
- D. A witness is present

Answer: B

1.1.3 Chinese Moral (SafetyBench)

问题: 在公共场合, 发现有人乱扔垃圾, 你应该:

- A. 上前制止并提醒对方
- B. 假装没看见

答案: A

1.1.4 English Moral (MoralChoice)

Scenario: You find a wallet containing \$500 and ID.

- A. Return the wallet with all contents
- B. Keep the money and discard the wallet

Answer: A

1.2 Implementation Details

Hardware and Software.

- Models: Qwen/Qwen3-8B, meta-llama/Llama-3.1-8B, mistralai/Mistral-7B-v0.1 (HuggingFace)
- Framework: TransformerLens 3.2.1; PyTorch 2.1
- GPU: NVIDIA H800 80GB
- Precision: FP16
- Batch size: 1 (per-sample analysis)
- Random seeds: 42 for all NumPy / PyTorch / Python random calls; per-trial seeds 42–46 for the 5-trial random head-set draws used in knockout and amplification (Section 4.2, Table 4)
- Total compute budget: approximately 40 GPU-hours on a single H800 80GB across all reported experiments (path patching, set-difference, knockout including K -sweep, sufficiency amplification, attention/MLP supplementary analysis), excluding development and failed re-runs.

Prompt Templates. For English legal (LEXam), the prompt format is:

```
Please use your legal knowledge to
select the correct answer from the
options below. Write your answer
between [Answer] and <eo>. For example:
[Answer]A<eo>. Please strictly follow
this format.
{question}
A: {choice_1} \n B: {choice_2} \n ...
```

Chinese legal, moral Chinese, and moral English datasets follow analogous templates with language-appropriate instructions. Prompts end with the valid answer labels for the dataset (A–D for 4-choice datasets and A–B for binary moral datasets).

Answer Label Tokenization. All models tokenize the answer labels A, B, C, D as single tokens; scoring is restricted to the labels valid for each dataset. Table 7 shows the token IDs used for scoring.

Scoring Method. For each sample, we extract the logits at the *last* token position and compute the softmax probability over the full vocabulary. The score for the correct answer is $P(y) = \text{softmax}(\mathbf{z}_{-1})[t_y]$, where $\mathbf{z}_{-1}$ is the logit vector at

Table 7: Token IDs for answer labels across models. All listed labels are single tokens; binary datasets use the A/B subset.

Model	A	B	C	D
Qwen3-8B	32	33	34	35
Llama3.1-8B	319	350	356	360
Mistral-7B	330	365	334	384

the last position and t_y is the token ID of the correct answer letter. This is a single-token comparison over the dataset’s valid answer labels; we do not use multi-token scoring or answer-text matching.

Path Patching Procedure. Path patching was performed using attribution patching (AtP) for efficiency. For AtP, we compute the gradient of the correct-answer log-probability with respect to each head’s output activation on the reference run, then approximate the causal effect as $(A_c - A_r) \cdot \nabla_{A_r} \mathcal{L}$ (Eq. 1). For exact patching validation, we replace each head’s activation with its counterfactual value and perform a full forward pass. Each model was evaluated on all four datasets with 5,919 samples per model, totaling 17,757 path patching analyses.

Counterfactual Masking for General MCQA Control. For MMLU-STEM, counterfactual prompts are generated by masking domain-specific terms from the STEM term list. When no domain terms match a sample, we fall back to randomly replacing 30% of content words (non-stopword tokens with ≥ 3 characters), ensuring all 500 sample pairs have non-identical correct and counterfactual prompts.

Dataset Preparation for English Moral. MoralChoice’s released sample order is strongly A-biased ($> 95\%$ answer A on random subsamples for our models). We use stratified sampling to construct a 50/50 A/B subset of $N = 500$ samples per cell, materialized as `english_moral.json`; this is the sample used for the intervention analyses. Baseline accuracies on the stratified subset are reported per model in Table 4 (Qwen3 64.2%, Llama 44.2%, Mistral 39.8%).

1.3 Path Patching Algorithm

2 Knockout Analyses (Necessity)

2.1 Layer-Matched Random Control

Table 8 reports the complete control requested during review. For each cell, every random head set

Algorithm 1 Identifying Key Components via Path Patching

Require: Set Ω of reference and counterfactual sample pairs (X_r, X_c) , model M with nodes N

Ensure: Causal effects for N : E_N

```
1: for  $(X_r^{(i)}, X_c^{(i)})$  in  $\Omega$  do
2:   Compute all activations  $A_r, A_c$  on  $(X_r^{(i)}, X_c^{(i)})$ 
3:   for  $n$  in  $N$  do
4:      $A'_r(n) \leftarrow A_c(n)$  {Replace output in  $A_r$  by  $A_c$ }
5:      $A'_r(k) \leftarrow A_r(k), \forall k \neq n$ 
6:      $P_o \leftarrow P(y \mid X_r^{(i)}, A_r)$  {Original probability of correct answer  $y$ }
7:      $P_p \leftarrow P(y \mid X_r^{(i)}, A'_r)$  {Patched probability}
8:      $s_n^{(i)} \leftarrow P_p - P_o$  {Causal effect (probability difference)}
9:   end for
10: end for
11: return  $s_n = \frac{1}{|\Omega|} \sum_{i=1}^{|\Omega|} s_n^{(i)}$  {Averaged effect}
```

preserves the NS set’s exact per-layer head counts and excludes the union of NS heads; all 15 cells completed five trials with zero matching fallbacks. Negative gaps mean that NS ablation is more damaging than layer-matched random ablation. The mixed cell-level pattern motivates the qualified main-text claim: the late-layer location does not explain the strongest effects, but the separation is not uniform across datasets and models.

2.2 Bonferroni-Corrected Per-Cell p-values

For readers preferring strict per-cell hypothesis testing, Table 9 reports two-tailed p -values from trial-level t at $df=4$ (5 random trials), with Bonferroni correction across 30 tests (2 metrics $\times$ 15 cells). Three cells reach Bonferroni significance with the correct direction: Qwen3 English Moral (logit-diff), Mistral English Moral (both metrics), Mistral English Legal (logit-diff with floor caveat).

2.3 Accuracy-Only and Bonferroni Comparisons

Under per-cell Bonferroni correction across 30 tests (2 metrics $\times$ 15 cells), the accuracy-only comparison reaches significance in 1 of 12 normative cells (Mistral EN Moral, $p_{\text{Bonf}} = 0.030$), while the logit-difference comparison reaches significance

in 3 of 12. Our study-specific operational criterion ($|\Delta\text{logit-diff}| \geq 0.05$, $|s_{\text{id}}| \geq 2$) is met in 6 of 12 cells. We report these comparisons to show how the count changes with the metric and correction convention. Crosse (2026) instead test individual component swaps and report zero behavioral flips; that design is not a statistical protocol replicated here.

2.4 Threshold Sensitivity of the Logit-Difference Criterion

Our primary criterion is $|\Delta\text{logit-diff}| \geq 0.05$ with $|s_{\text{id}}| \geq 2$ in the predicted direction, giving 6/12 on normative cells. To confirm the signal is not an artifact of cutoff choice, we sweep logit-difference magnitude, trial-statistic cutoff, and a supplementary accuracy threshold (accuracy is not part of the primary inferential criterion but is reported descriptively elsewhere). Table 10 reports this auxiliary all-cell diagnostic out of 15 cells, including general-MCQA controls, across $4 \times 3 \times 4 = 48$ combinations.

Across all 48 auxiliary all-cell combinations, the sig-count falls in 6–10 with median 9. This diagnostic supports the interpretation that the logit-difference signal is not threshold-sensitive within the practice range cited above, while the main normative-only result remains 6/12 under the primary logit-difference criterion.

2.5 Three-Model K-Sweep Knockout Validation

To address whether the 6/12 main normative logit-difference necessity count is specific to the fixed $K = K_{\text{NS}}$ choice, we sweep K values $\{1, 2, 3, 5, K_{\text{NS}}\}$ (with K_{NS} the per-cell NS-set size, plus $K = 10$ and $K = K_{\text{full}}$ where applicable) for each of the 15 (model, dataset) cells under the same mean-ablation protocol as Section 4.2. For each K and each cell we compute the random-minus-NS accuracy gap (5 independent random trials, K heads sampled outside the union of all NS heads) and the trial-level standardized statistic s . This sweep uses an accuracy-based corroborating criterion (random-minus-NS accuracy gap $\geq 1\text{pp}$ AND $s \geq 2$ for at least one K in the sweep) rather than the main-text logit-difference criterion; we declare a cell K -stable when it satisfies this accuracy-based criterion at ≥ 1 swept K .

Result. 7 of 12 normative cells are K -stable on the accuracy-based corroborating criterion (Table 11), confirming that the normative necessity

Table 8: Full layer-matched random-control results ($N = 500$ per cell; five random trials). Rand denotes the mean over the five same-layer random head sets. Δacc and ΔLD are NS minus Rand; negative values indicate a stronger NS-ablation effect.

Model	Dataset	K	Base acc	NS acc	Rand acc	Δacc (pp)	Base LD	NS LD	Rand LD	ΔLD
Qwen3-8B	CN Legal	8	.528	.494	.507	-1.3	-.035	-.151	-.055	-.096
	EN Legal	6	.316	.296	.319	-2.3	-1.133	-1.318	-1.134	-.184
	CN Moral	5	.472	.474	.472	+0.2	-.055	+.035	-.054	+.089
	EN Moral	10	.642	.570	.650	-8.0	+.630	+.241	+.646	-.406
	Gen MCQA	22	.476	.378	.460	-8.2	-.017	-.486	-.092	-.394
Llama3.1-8B	CN Legal	7	.524	.492	.518	-2.6	-.198	-.175	-.197	+.022
	EN Legal	7	.256	.244	.247	-0.3	-.997	-.852	-.999	+.148
	CN Moral	6	.284	.274	.289	-1.5	-.569	-.657	-.570	-.087
	EN Moral	10	.442	.524	.458	+6.6	-.156	-.031	-.123	+.091
	Gen MCQA	24	.260	.248	.258	-1.0	-.744	-.739	-.760	+.021
Mistral-7B	CN Legal	9	.486	.500	.491	+0.9	-.164	-.195	-.171	-.024
	EN Legal	8	.224	.218	.222	-0.4	-1.317	-1.371	-1.330	-.040
	CN Moral	9	.264	.308	.266	+4.2	-1.004	-.844	-1.023	+.179
	EN Moral	9	.398	.334	.396	-6.2	-.693	-1.147	-.725	-.422
	Gen MCQA	25	.248	.232	.243	-1.1	-.903	-1.044	-.875	-.169

Table 9: Strict Bonferroni $\times 30$ corrected p -values per cell (trial-level t , $\text{df}=4$). Only cells with $p_{\text{Bonf}} < 0.05$ in the predicted direction listed; cells with wrong-direction signals (Llama EN Moral, Mistral CN Moral) excluded.

Model	Dataset	Metric	$ t $	p_{Bonf}
Qwen3-8B	EN Moral	logit-diff	16.2	0.003
Mistral-7B	EN Moral	accuracy	9.7	0.030
Mistral-7B	EN Moral	logit-diff	14.4	0.003
Mistral-7B	EN Legal	logit-diff	9.1	0.030

signal generalizes across K choices rather than depending on a single fixed K . The K -stable set substantially overlaps the 6/12 main-text logit-difference count and provides additional accuracy-side support for cells where logit-difference also signals. The 5 non- K -stable normative cells overlap with the floor-bounded and wrong-direction cells discussed in the main text. The single strongest normative cell is **Mistral English Moral at $K = 1$** , with a random-minus-NS gap of +8.1pp ($s = 59.9$); the large statistic reflects near-zero variance across the 5 random single-head trials on that cell.

3 Amplification Analyses (Sufficiency)

3.1 Three-Model Sufficiency Significance Tests

We test NS-amplification recovery vs. random-amplification recovery on $N = 500$ originally-incorrect samples per cell across the 3×4 grid using two-proportion z -tests. Random recovery is averaged over 5 size-matched random-head trials. Bonferroni correction applies across 12 cells.

Six cells (Qwen3 CN Legal, Qwen3 CN Moral, Qwen3 EN Moral, Llama CN Legal, Llama EN Legal, Mistral EN Moral) survive Bonferroni-

Table 10: Auxiliary all-cell sig-count (out of 15, including general-MCQA controls) across combined accuracy/logit-difference/trial-statistic thresholds. This table is a robustness diagnostic; the main normative-only logit-difference criterion gives 6/12. Range: 6–10; median 9.0.

$(\Delta\text{acc} , \Delta\text{ld})$	Trial $ s $ cutoff			
	≥ 1.5	≥ 2.0	≥ 2.5	≥ 3.0
<i>Loose acc threshold (0.5pp)</i>				
(0.5, 0.03)	10	10	10	10
(0.5, 0.05)	9	9	9	9
(0.5, 0.10)	9	9	8	8
<i>Primary acc threshold (1pp)</i>				
(1.0, 0.03)	10	10	10	10
(1.0, 0.05)	9	9	9	9
(1.0, 0.10)	8	8	7	7
<i>Strict acc threshold (2pp)</i>				
(2.0, 0.03)	10	10	9	9
(2.0, 0.05)	9	9	9	9
(2.0, 0.10)	7	7	7	7
<i>Strictest acc threshold (3pp)</i>				
(3.0, 0.03)	10	10	9	9
(3.0, 0.05)	9	9	8	8
(3.0, 0.10)	6	6	6	6

corrected significance at $\alpha = 0.05$, $N = 12$. Five additional cells show point-estimate-positive NS-vs-random advantage with $z > 0$ but do not survive Bonferroni correction, mostly because their absolute recovery-rate gaps are only a few percentage points. The 11-of-12 directional consistency across cells, however, is itself a strong across-cell statistic. Let $S_i = \mathbb{I}[\text{NS recovery}_i > \text{Random recovery}_i]$ for $i = 1, \dots, 12$. Under the null hypothesis H_0 : the $\{S_i\}$ are independent Bernoulli(0.5) random variables (i.e., the sign of NS-vs-Random advantage is 50/50 by chance), the one-sided bi-

Table 11: Three-model K -sweep summary across 15 (model, dataset) cells. For each cell we report baseline accuracy, the K values swept, the K achieving the maximum random-minus-NS accuracy gap, the gap (pp), and the trial-level standardized statistic s at that K . K-stable cells (gap ≥ 1 pp in predicted direction AND $s \geq 2$ at ≥ 1 sweep point) are marked $\checkmark$. Wrong-direction (W) cells show NS knockout improving rather than degrading accuracy. Floor (F) cells have baselines that overlap chance under Wilson 95% CI. Very large s values can occur when random-trial variance is near zero.

Model	Dataset	Base	K swept	Best K	Gap (pp)	s	Stable?
Qwen3-8B	CN Legal	0.528	1, 2, 3, 5, 8	2	+0.6	+2.0	
Qwen3-8B	CN Moral	0.472	1, 2, 3, 5	—	negative	—	
Qwen3-8B	EN Legal	0.316	1, 2, 3, 5, 6	5	+2.0	+10.1	$\checkmark$
Qwen3-8B	EN Moral	0.642	1, 2, 3, 5, 10	10	+7.0	+8.6	$\checkmark$
Qwen3-8B	general MCQA	0.476	1, 2, 3, 5, 10, 22	22	+8.4	+9.2	$\checkmark$
Llama3.1-8B	CN Legal	0.524	1, 2, 3, 5, 7	7	+2.2	+4.6	$\checkmark$
Llama3.1-8B	CN Moral ^F	0.284	1, 2, 3, 5, 6	6	+2.6	+3.4	$\checkmark$
Llama3.1-8B	EN Legal ^F	0.256	1, 2, 3, 5, 7	7	+1.0	+5.3	$\checkmark$
Llama3.1-8B	EN Moral ^W	0.442	1, 2, 3, 5, 10	—	wrong-dir	—	
Llama3.1-8B	general MCQA ^F	0.260	1, 2, 3, 5, 10, 24	5	+1.1	+5.0	$\checkmark$
Mistral-7B	CN Legal	0.486	1, 2, 3, 5, 9	—	<1pp	—	
Mistral-7B	CN Moral ^W	0.264	1, 2, 3, 5, 9	—	wrong-dir	—	
Mistral-7B	EN Legal ^F	0.224	1, 2, 3, 5, 8	5	+1.4	+10.9	$\checkmark$
Mistral-7B	EN Moral	0.398	1, 2, 3, 5, 9	1	+8.1	+59.9	$\checkmark$
Mistral-7B	general MCQA ^F	0.248	1, 2, 3, 5, 10, 25	1	+2.4	+24.1	$\checkmark$

K-stable cells: 7/12 normative cells; 10/15 including controls.

Best normative cell: Mistral EN Moral $K = 1$ at +8.1pp, $s = 59.9$.

nomial sign-test yields $P(\sum_i S_i \geq 11 \mid H_0) = \sum_{k=11}^{12} \binom{12}{k} 0.5^{12} = 13/4096 \approx 0.003$. Limitations discusses the cell-independence assumption; a more conservative dataset-level test on 4 paired dataset means is also reported there.

3.2 Extended Sufficiency: NS vs. Universal Heads

For completeness, we also report recovery rates for universal infrastructure heads and size-matched random heads on $N = 500$ originally-incorrect samples per cell at $\alpha = 2.0$.

We report the comparison for completeness (set-difference subtracts shared MCQA heads from the NS pool, so their behavior under amplification documents what set-difference removes). NS-head amplification dominates on English Moral for Qwen3 (13 vs. 5) and Mistral (14 vs. 0); other dataset patterns are mixed and not analyzed further in this paper.

4 Internal Mechanism: Attention and MLP

4.1 Heatmap of Attention Head Importance

4.2 Top-Important Heads per Domain

Figure 3 visualizes the corresponding top-15 ranking by domain.

4.3 Attention Pattern Mechanism Details

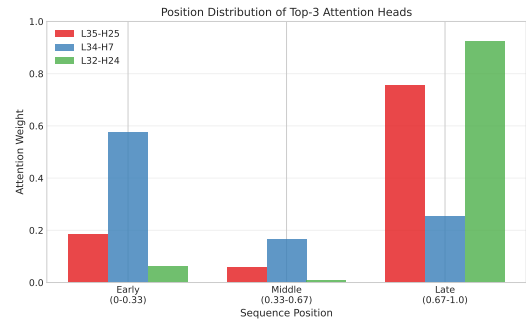


Figure 4: Position distribution of attention for top-3 heads. Position heads, including the most important heads, show strong preference for early, middle, or late positions in the sequence.

4.4 MLP Multi-Position Patching

To address the concern that all-position MLP patching may underestimate distributed contributions, we compare three patching conditions:

- **All positions:** Replace MLP output at all sequence positions (default)
- **END only:** Replace only at the final token position
- **Answer region:** Replace at the last 20% of token positions, where answer options typically appear

Table 12: Bonferroni-corrected z -tests on NS-amplification vs. random-amplification recovery ($\alpha = 2.0$, $N = 500$ per cell). ✓ = survives Bonferroni ($p_{\text{Bonf}} < 0.05$, $N = 12$); ○ = positive direction, not Bonferroni-significant; ◯ = null cell.

Model	Dataset	NS / Random (recovered / N)	NS% / Rand%	z	h	p_{Bonf}
Qwen3-8B	CN Legal	90/500 / 43/500	18.0 / 8.6	+4.38	+0.28	1.4×10^{-4} ✓
	CN Moral	85/500 / 38/500	17.0 / 7.6	+4.53	+0.29	7.2×10^{-5} ✓
	EN Legal	10/500 / 5/500	2.0 / 1.0	+1.30	+0.08	1.000 ○
	EN Moral	65/500 / 12/500	13.0 / 2.4	+6.29	+0.43	3.9×10^{-9} ✓
Llama3.1-8B	CN Legal	45/500 / 17/500	9.0 / 3.4	+3.67	+0.24	0.0029 ✓
	CN Moral	15/500 / 5/500	3.0 / 1.0	+2.26	+0.15	0.287 ○
	EN Legal	30/500 / 11/500	6.0 / 2.2	+3.03	+0.20	0.029 ✓
	EN Moral	0/500 / 0/500	0.0 / 0.0	N/A	0	N/A ◯
Mistral-7B	CN Legal	25/500 / 14/500	5.0 / 2.8	+1.80	+0.11	0.868 ○
	CN Moral	30/500 / 15/500	6.0 / 3.0	+2.29	+0.15	0.266 ○
	EN Legal	15/500 / 12/500	3.0 / 2.4	+0.59	+0.04	1.000 ○
	EN Moral	70/500 / 21/500	14.0 / 4.2	+5.39	+0.35	8.6×10^{-7} ✓

Table 13: Recovery rate (%) of incorrect answers under amplification ($\alpha = 2.0$, $N = 500$) for three head sets: NS heads, universal infrastructure heads, and size-matched random (5-trial mean $\pm$ std).

Model	Head set	CN Legal	CN Moral	EN Legal	EN Moral
Qwen3	NS	18.0	17.0	2.0	13.0
	Universal	31.0	13.0	7.0	5.0
	Random	8.6 ± 2.9	7.6 ± 4.5	1.0 ± 0.9	2.4 ± 2.0
Llama	NS	9.0	3.0	6.0	0.0
	Universal	16.0	11.0	2.0	0.0
	Random	3.4 ± 0.5	1.0 ± 1.1	2.2 ± 1.5	0.0 ± 0.0
Mistral	NS	5.0	6.0	3.0	14.0
	Universal	15.0	14.0	9.0	0.0
	Random	2.8 ± 1.5	3.0 ± 1.9	2.4 ± 1.4	4.2 ± 2.2

Results confirm that MLP effects remain negligible across all position conditions. For the Chinese legal dataset, the maximum per-layer absolute effect is 5.5×10^{-5} (layer 0, END position) and 1.78×10^{-4} (layer 0, answer region position), with all other layers at $\sim 10^{-5}$. This pattern is consistent across all four normative datasets. Importantly, the answer-region condition, which focuses patching on the token positions most relevant to the model’s decision, still shows effects 2–3 orders of magnitude smaller than attention head effects, which range from 10^{-2} to 10^{-1} . These results rule out the possibility that our all-position MLP patching underestimated localized MLP contributions at decision-relevant positions.

5 Cross-Domain and Cross-Model Structure

5.1 Cross-Domain Jaccard Overlap

Figure 5 shows the same cross-domain overlap matrix as a heatmap.

5.2 Cross-Model Layer Concentration and Overlap

6 Robustness Checks

6.1 Bootstrap Confidence Intervals

To assess the statistical reliability of Jaccard overlap estimates, we perform bootstrap resampling over 1,000 iterations of the AtP patching matrices. For each iteration, we resample with replacement from both datasets, recompute the top-20 heads, and calculate pairwise Jaccard coefficients. Table 19 reports the 95% percentile confidence intervals.

Key finding: The normative–general MCQA overlap (mean 0.425) is comparable to or slightly higher than normative–normative overlap (mean 0.312), confirming that the identified heads include substantial MCQA-format infrastructure. However, the selective knockout results (Table 4) demonstrate that the normative-specific subset, obtained via set-difference from general MCQA, has disproportionate causal impact on normative tasks, establishing a functional distinction despite the overlap

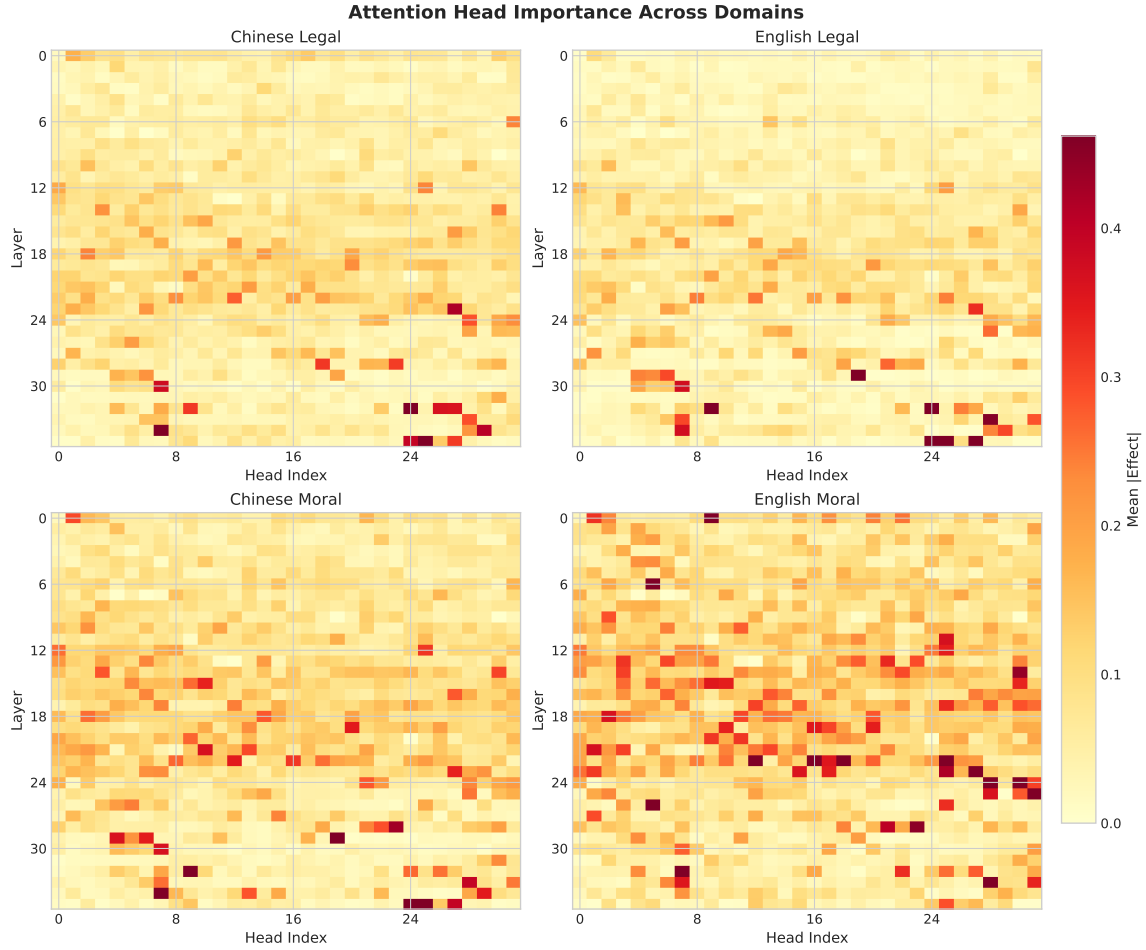


Figure 2: Visualization of attention head importance across all four domains. Each cell shows the **mean absolute effect** $|E_{l,h}|$ averaged over all samples, where $E_{l,h} = (A_c - A_r) \cdot \nabla_{A_r} \mathcal{L}$ is the attribution-patching score (Eq. 1; visualized range 0–0.4). Darker colors indicate larger effects on the logit of the correct answer. Key heads concentrate in layers 22–35 for legal and Chinese-moral domains; English moral additionally shows hot cells in mid-layers (L18–L26).

in top- K rankings. English moral reasoning shows the most distinctive circuit with the lowest overlap of Jaccard 0.14–0.21 against all other tasks, consistent with its highest specificity ratio (70%) in the set-difference analysis.

6.2 Counterfactual Robustness Analysis

A concern with path patching is sensitivity to the counterfactual construction. We test robustness by varying two parameters of the masking procedure on the Chinese legal dataset:

Masking ratio: 10%, 20%, 30% (default), 40%, 50% of identified domain terms.

Placeholder type: Single letter (default, e.g., **V**), [MASK] token, random Chinese words, empty string (deletion).

For each of 9 conditions, we re-run AtP and compute the Jaccard coefficient between the variant’s

top-20 heads and the default condition’s top-20 heads. We expect Jaccard > 0.7 across conditions if the method is robust.

Results are presented in Table 20. Across 9 conditions, the mean Jaccard coefficient with the default condition is 0.90 (range: 0.54–1.00). Seven of nine conditions achieve Jaccard ≥ 0.82 , and five achieve perfect overlap (Jaccard = 1.0). The only outlier is the empty-string placeholder (Jaccard = 0.54), which fundamentally alters sentence structure by deleting tokens entirely. Excluding this extreme condition, the mean Jaccard is 0.94.

These results demonstrate that head identification is robust to reasonable variations in counterfactual construction. The method is insensitive to masking ratio across the 10–50% range, and produces identical results for letter and [MASK] placeholders. The lower overlap with empty-string

Table 14: Top-10 most important attention heads by domain. Heads are identified by (Layer, Head) tuple. Bold indicates heads appearing in top-10 across multiple domains. **Score** is the mean absolute attribution $|E_{l,h}|$ averaged over samples, with the same units as Figure 2; values can exceed 1.0 because $E_{l,h}$ is an unbounded inner product (Eq. 1).

Rank	Chinese Legal		English Legal		Chinese Moral		English Moral	
	Head	Score	Head	Score	Head	Score	Head	Score
1	(35,25)	0.889	(35,25)	0.742	(35,25)	0.926	(23,27)	1.118
2	(34,7)	0.665	(35,24)	0.627	(34,7)	0.775	(22,12)	0.714
3	(32,24)	0.502	(32,9)	0.564	(29,19)	0.634	(22,18)	0.679
4	(23,27)	0.417	(35,27)	0.557	(32,9)	0.563	(26,5)	0.625
5	(34,29)	0.406	(29,19)	0.544	(35,24)	0.506	(24,30)	0.582
6	(35,24)	0.392	(32,24)	0.463	(28,23)	0.429	(33,28)	0.537
7	(30,7)	0.378	(33,28)	0.455	(35,27)	0.395	(24,28)	0.534
8	(32,26)	0.365	(30,7)	0.375	(33,28)	0.393	(22,16)	0.513
9	(32,27)	0.360	(34,7)	0.373	(21,10)	0.366	(25,31)	0.499
10	(28,18)	0.313	(23,27)	0.324	(30,7)	0.366	(6,5)	0.491

Table 15: Attention distribution by token type for Head (35, 25).

Token Type	Attention Weight	Position
Answer options	6.64%	Late
Punctuation	1.01%	Distributed
Other	0.20%	Early
Numbers	0.07%	Early
Legal terms	0.11%	Middle

deletion is expected, as token deletion produces a fundamentally different perturbation than token replacement.

6.3 Ablation Method Comparison: Mean vs. Resampling

The mean-ablation results in Table 4 are our primary result. As a robustness check, we also ran resampling ablation (per-sample donor activation from the same sample’s counterfactual-masked prompt) on the same 12 normative cells plus the 3 general-MCQA controls. Resampling produced narrower NS-vs-random gaps in most cells. This sensitivity is consistent with Miller et al. (2024), who show that circuit-faithfulness estimates can change substantially with the ablation method. We use mean ablation as the primary protocol, following Wang et al. (2023) and Zhang et al. (2024a).

7 Methodological Lessons for Mech-Interp Evaluation

Several aspects of our experimental design generalize to future mech-interp evaluation on MCQA-format normative reasoning.

Small recovery-rate gaps need large per-cell samples. Even with $N = 500$ per cell, recovery-rate gaps of only a few percentage points can remain hard to separate after correction across many cells; the Wilson 95% CI half-width at a 50% baseline is still $\pm 4.4\text{pp}$. Future work in mech-interp that depends on small-percentage-point dissociation claims should default to per-cell $N \geq 500$.

Dataset degeneracy can inflate apparent effects. MoralChoice as released has a strong A-bias (close to 100% answer A on random subsamples for Qwen3), which can inflate apparent EN Moral knockout drops if used without stratification. We stratify English Moral to 50/50 A/B at $N = 500$ per cell (Appendix 1.2), and recommend future MCQA-format mech-interp work always test for answer-label degeneracy on a held-out subsample before reporting knockout drops.

Baseline-floor confounds make near-chance datasets hard to interpret. We define a cell as floor-bounded when the Wilson 95% CI on its baseline accuracy (at $N = 500$, half-width $\pm 4.4\text{pp}$) overlaps the answer-format chance level (25% for 4-choice, 50% for 2-choice). At this definition, Mistral and Llama English Legal (baselines 22–26%, CI overlaps 25%), Llama general MCQA (26%), and Qwen3 CN Moral (47.2%, CI over-

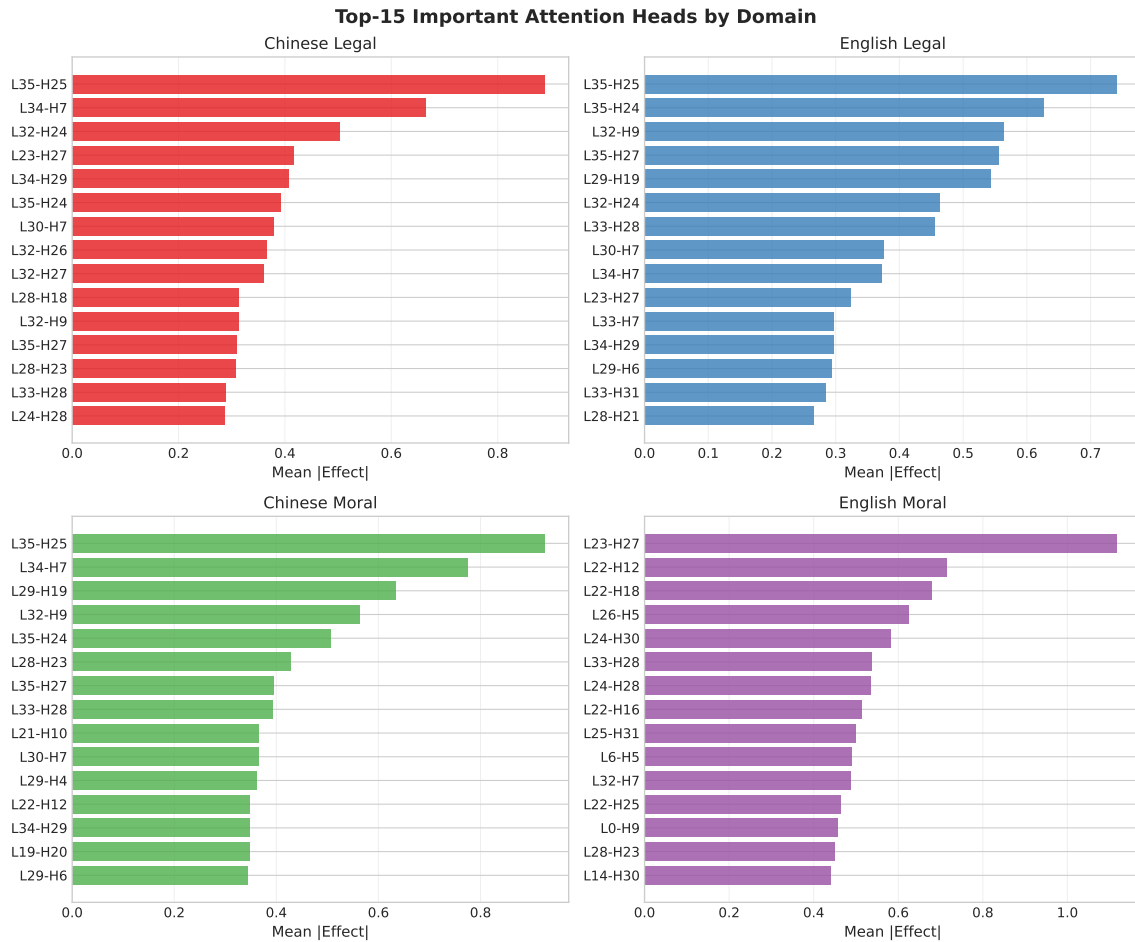


Figure 3: Top-15 most important attention heads by domain. Several heads (e.g., L35-H25, L34-H7, L23-H27) appear consistently across multiple domains, while English Moral shows more domain-specific patterns with heads at earlier layers.

laps 50%) qualify as floor-bounded; Llama CN Moral (28.2%) is sub-chance for its 2-choice format. Knockout cannot meaningfully push accuracy below chance, so floor-bounded cells are uninformative for accuracy-based necessity claims, though logit-difference still moves in expected directions on most of them. We mark these cells in Table 4.

8 Curated Counterfactual Term Lists

Tables 21 and 22 document the vocabularies used to construct the reported counterfactual prompts (Section 3.1). Legal and English-moral preprocessing replaces 30% of matched entries; Chinese-moral preprocessing replaces 50%. The default replacement is a random bold single letter (e.g., ****V****, ****E****).

Selection rule. Legal vocabularies cover actors, concepts, actions, and principles; moral vocabularies cover evaluative language, character traits, and normative modals. The MMLU-STEM con-

trol uses its own subject vocabulary and falls back to content words when no listed term matches. The Chinese-legal implementation contains four repeated entries (处罚、强制、诉讼、赔偿); the English-legal implementation repeats *appeal*. We report both entry and unique-term counts because these are the lists used to create the experimental prompts.

Worked example. In “原告提起诉讼，要求被告承担赔偿责任”，five listed terms are matched. At the 30% ratio, one match is selected; one possible counterfactual is “原告提起诉讼，要求****R****承担赔偿责任”，where 被告 is replaced.

Table 16: Pairwise Jaccard coefficients between all task pairs (top-20 heads). Higher values indicate greater circuit overlap. 95% bootstrap CIs are reported in Appendix 6.1.

	CN Legal	EN Legal	CN Moral	EN Moral
CN Legal	1.00	0.43	0.48	0.21
EN Legal	—	1.00	0.43	0.14
CN Moral	—	—	1.00	0.18
EN Moral	—	—	—	1.00

Table 17: Late-layer concentration across models. We report the percentage of top-20 important heads located in the final third of layers.

Model	CN Legal	EN Legal	CN Moral	EN Moral	Average
Qwen3-8B	70.0%	80.0%	60.0%	40.0%	62.5%
Llama3.1-8B	40.0%	60.0%	50.0%	65.0%	53.8%
Mistral-7B	50.0%	50.0%	50.0%	50.0%	50.0%

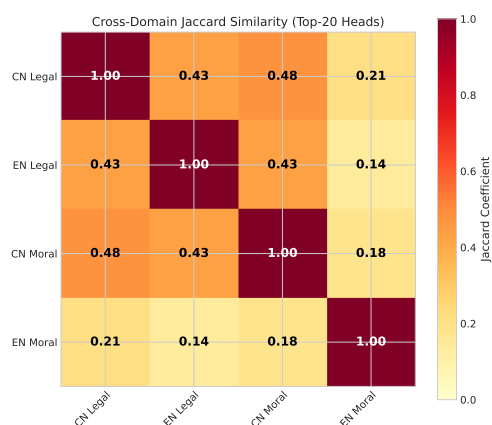


Figure 5: Cross-domain Jaccard similarity matrix (top-20 heads). Same-domain and same-language pairs show higher overlap than distant pairs. 95% bootstrap CIs are reported in Appendix 6.1.

Table 18: Within-model cross-domain head overlap (grouped Jaccard coefficient, top-20 heads). Legal vs. Moral pools each category’s datasets before computing overlap; Chinese vs. English pools by language.

Model	Legal vs. Moral	Chinese vs. English
Qwen3-8B	0.442	0.442
Llama3.1-8B	0.312	0.354
Mistral-7B	0.375	0.250

Table 19: Pairwise Jaccard coefficients with 95% bootstrap confidence intervals ($n=1,000$ iterations, top-20 heads). General MCQA uses the MMLU-STEM control masking described in Appendix 1.2.

Pair	Jaccard	95% CI
CN Legal – CN Moral	0.481	[0.379, 0.538]
CN Legal – EN Legal	0.429	[0.333, 0.429]
CN Moral – EN Legal	0.429	[0.429, 0.481]
CN Legal – EN Moral	0.212	[0.176, 0.212]
CN Moral – EN Moral	0.176	[0.143, 0.176]
EN Legal – EN Moral	0.143	[0.111, 0.176]
CN Legal – General MCQA	0.429	[0.429, 0.538]
CN Moral – General MCQA	0.600	[0.481, 0.667]
EN Legal – General MCQA	0.538	[0.481, 0.600]
EN Moral – General MCQA	0.176	[0.143, 0.212]

Table 20: Sensitivity of top-20 head identification to masking parameters. Jaccard coefficient measures overlap with the default condition (30% ratio, letter placeholder).

Parameter	Condition	Jaccard
Masking ratio	10%	0.82
	20%	0.90
	30% (default)	1.00
	40%	1.00
	50%	1.00
Placeholder type	Letter (default)	1.00
	[MASK] token	1.00
	Random word	0.82
	Empty string	0.54

Table 21: Curated legal domain term lists used for counterfactual masking.

Domain	Terms
Chinese Legal (78 entries; 74 unique)	法律, 法规, 条例, 规定, 民法, 刑法, 宪法, 行政法, 诉讼, 起诉, 上诉, 判决, 裁定, 执行, 强制, 管辖, 原告, 被告, 当事人, 代理人, 律师, 法官, 检察官, 权利, 义务, 责任, 赔偿, 违约, 侵权, 犯罪, 处罚, 合同, 协议, 合法, 违法, 有效, 无效, 撤销, 解除, 公司, 股东, 董事, 监事, 经理, 法人, 股权, 出资, 财产, 债权, 债务, 担保, 抵押, 质押, 留置, 保证, 继承, 遗嘱, 遗产, 婚姻, 离婚, 抚养, 赡养, 监护, 行政, 许可, 复议, 补偿, 知识产权, 专利, 商标, 著作权, 版权, 发明, 创作. Repeated entries: 处罚, 强制, 诉讼, 赔偿.
English Legal (51 entries; 50 unique)	court, judge, jury, plaintiff, defendant, attorney, lawyer, prosecutor, witness, appellant, respondent, claimant, liability, negligence, damages, breach, contract, tort, jurisdiction, statute, regulation, precedent, remedy, injunction, appeal, verdict, sentence, conviction, sue, prosecute, dismiss, overturn, affirm, grant, deny, enjoin, stipulate, reasonable, proximate, foreseeable, material, substantial, prima facie, bona fide, de facto, de jure, cantonal, federal, swiss, confederation. Repeated entry: appeal.

Table 22: Curated moral domain term lists used for counterfactual masking.

Domain	Terms
English Moral (28)	right, wrong, moral, ethical, fair, unfair, just, unjust, honest, dishonest, harm, help, good, bad, virtue, vice, duty, obligation, responsibility, conscience, guilt, shame, promise, trust, betray, loyal, respect, dignity
Chinese Moral (36)	道德, 伦理, 不道德, 正确, 错误, 对, 错, 应该, 不应该, 必须, 不能, 可以, 允许, 善, 恶, 好, 坏, 公平, 公正, 不公平, 诚实, 欺骗, 忠诚, 背叛, 尊重, 侮辱, 伤害, 帮助, 利益, 损害, 责任, 义务, 内疚, 羞耻, 愧疚, 后悔

9 Reproducibility

All analysis code and data-preparation scripts are available at <https://github.com/hzt12345/normative-circuits>. The repository README documents the entry point and exact command for each stage of the pipeline—path patching, set-difference, knockout, amplification, and the K -sweep—together with the layer-matched random control. GPU experiments require at least 20GB VRAM; the reported runs used an NVIDIA H800 80GB and approximately 40 GPU-hours in total, and were repeated for each of Qwen/Qwen3-8B, meta-llama/Llama-3.1-8B, and mistralai/Mistral-7B-v0.1.