SCoUT: A Framework for Structured Stereotype Analysis in Language Models

Jinxuan Wu[1], **Bin Li**[1], **Xiangyang Xue**[1,2*]

[1]College of Computer Science and Artificial Intelligence, Fudan University, China
[2]Institute of Big Data, Fudan University, China
24210240054@m.fudan.edu.cn, {libin, xyxue}@fudan.edu.cn

Abstract

Existing stereotype auditing methods for Large Language Models (LLMs) typically rely on isolated rating schemes or task-specific probes, lacking theoretical grounding and failing to reveal internal organization beyond surface-level output patterns. In this paper, we introduce **SCoUT** (**S**tereotype **C**ontent-**o**riented **U**tility structure via **T**hurstonian modeling), a closed-loop framework that structurally models, explicitly probes, and functionally steers stereotype dimensions (warmth and competence) in LLMs. SCoUT first reconstructs a global *stereotype utility structure* aligned with Stereotype Content Model theory via Thurstonian comparative judgments. Across multiple open-source LLMs, this modeling achieves high pairwise-preference prediction accuracy ($\geq$ 0.90 on larger-scale models) and exhibits strong cross-model consistency. Probing internal attention mechanisms localizes this structure to specific heads (Spearman's ρ up to 0.83 for warmth and 0.90 for competence) and surfaces a salient asymmetry between warmth and competence. Further, targeted inference-time activation modifications on these dimension-sensitive heads consistently steer model outputs along the intended axes. By bridging behavioral measurement with internal representation and controllable steering, SCoUT offers an end-to-end framework that uncovers and interprets the latent structure of stereotypes, advancing stereotype auditing from surface detection to structural analysis.

Introduction

The rapid deployment of Large Language Models (LLMs) has raised serious concerns about their propensity to reproduce and amplify harmful social stereotypes (Schramowski et al. 2022; Bolukbasi et al. 2016). Existing stereotype auditing methods typically fall into two broad categories (Morehouse, Swaroop, and Pan 2025): Association-level methods (e.g., WEAT, StereoSet, (Caliskan, Bryson, and Narayanan 2017; Nadeem, Bethke, and Reddy 2021)) measure semantic associations between group labels and attributes. Although straightforward and context-agnostic, these approaches provide only superficial associative measurements, failing to reveal how stereotypes are systematically organized and represented internally within models. Decision or generation-based methods (e.g., CrowS-Pairs, BiasInBios (Nangia et al. 2020; De-Arteaga et al. 2019)) probe stereotypes through artificial scenarios or generation tasks. While capturing more realistic use cases, their results are critically sensitive to prompt phrasing and scenario design, making findings subjective, fragmented, and difficult to systematically compare or generalize. Crucially, both families of methods share the same fundamental limitation: they focus solely on measuring *external outputs*, failing to answer whether stereotypes reflect deeper *internal structure* or merely surface-level output patterns. This restricts current auditing to descriptive detection rather than providing meaningful explanations or actionable insights for stereotype control. Furthermore, because each scenario or association is evaluated independently, existing methods fail to uncover the global structural relationships that potentially underpin stereotypical outputs, forcing researchers to design numerous disparate contexts that provide only incomplete, fragmented coverage.

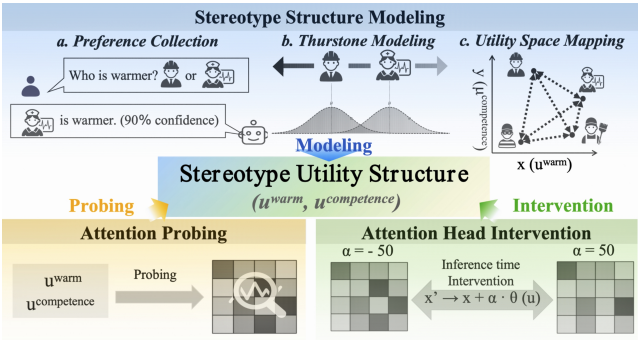


Figure 1: SCoUT: A closed-loop framework for modeling, probing, and intervening on stereotype structures in LLMs. Pairwise preference data is modeled with Thurstone analysis to infer a structured utility space (top), which is then used to probe and identify relevant attention heads (bottom left). Targeted interventions on these heads (bottom right) demonstrate that the internal stereotype structure is both interpretable and functionally controllable.

Psychological theories suggest that human stereotypes are not merely isolated associations, but rather organized along meaningful latent dimensions, which consistently guide judgments across contexts (Fiske et al. 2002). If LLMs similarly internalize stereotypes as coherent, latent structures,

*Corresponding author.

then directly modeling these underlying utilities would offer deeper insight and greater control. Recent advances like Stereomap (Jeoung, Ge, and Diesner 2023) apply the Stereotype Content Model (SCM) to rate groups on warmth and competence, improving interpretability over ad-hoc benchmarks. However, these absolute Likert ratings remain isolated, prone to scale anchoring and score compression, and cannot recover how groups relate to one another in a unified space. In contrast, Thurstone's law of comparative judgment (Thurstone 1927) models perceptions as latent utilities inferred from pairwise comparisons, yielding a normalized structure that supports consistent, generalizable comparisons across groups.

Based on these insights, we introduce **SCoUT** (Stereotype Content-oriented Utility structure via Thurstonian modeling), a closed-loop, theory-grounded framework designed specifically to diagnose, interpret, and validate stereotype structures in LLMs. SCoUT first reconstructs a globally consistent *stereotype utility structure* using pairwise comparative judgments aligned with SCM. Unlike prior absolute rating approaches, our comparative design can robustly recover relational positioning between groups. Then, by probing model internals, we explicitly identify and localize these stereotype dimensions within specific attention heads, demonstrating that stereotypes are deeply and systematically encoded inside models. Finally, targeted functional interventions provide validation for the internal representational structure, showing it can be directly manipulated to steer model outputs along warmth and competence dimensions. This closed-loop integration of modeling, probing, and intervention fundamentally shifts stereotype auditing from isolated, descriptive analyses toward interpretable, actionable governance of LLM stereotypes. In summary, our main contributions are:

- Theory-grounded structural auditing: Unlike fragmented and prompt-sensitive evaluations, SCoUT constructs a globally consistent stereotype utility structure grounded in established psychometric (Thurstonian modeling) and social psychological (SCM) theories, offering robust and relationally meaningful insights.
- Internal localization and interpretability: Through comprehensive probing, we demonstrate that stereotype dimensions are explicitly encoded in identifiable attention heads (Spearman's ρ up to 0.90), offering unprecedented transparency and interpretability.
- Functional validation of internal structures: We demonstrate that the identified internal stereotype structures are not merely correlational but functionally actionable. Our inference-time activation modifications serve as a mechanistic proof-of-concept, confirming that these structures can influence model outputs.

Overall, this framework reframes stereotype auditing from simply asking *"Does the model output biased content?"* to exploring *"How are stereotypes internally organized, and how can we systematically diagnose and control them?"*, laying a solid foundation for more transparent and accountable governance of LLM stereotypes.

Preliminaries

Stereotype Content Model (SCM). SCM is a well-established and validated social psychological theory proposing that stereotypes about social groups universally organize along two dimensions: *warmth* (reflecting intent and trustworthiness) and *competence* (reflecting capability and status) (Fiske et al. 2002). This model classifies social groups into four distinct quadrants—Admiration (high warmth, high competence), Envy (low warmth, high competence), Pity (high warmth, low competence), and Contempt (low warmth, low competence)—each associated with specific emotions and predictable social attitudes. SCM provides a theoretically grounded framework suitable for systematic stereotype analysis and interpretation.

Thurstonian Comparative Judgment. Thurstone's law of comparative judgment (Thurstone 1927) is a foundational psychometric framework that models attitudes or perceptions as latent utilities on a continuous scale. Rather than relying on absolute ratings, it infers these utilities from pairwise comparisons, assuming each item has a hidden value and that observed preferences reflect noisy differences between them. Typically, the model assumes normally distributed perceptual noise, allowing utilities to be recovered through probabilistic estimation. Compared to direct scoring, this approach mitigates scale anchoring and compression, and yields a globally consistent structure that supports robust, interpretable comparisons across arbitrary groups. Recent applications (Mazeika et al. 2025) show its effectiveness in revealing stable internal value systems in LLMs.

Probing and Intervention in LLMs. Probing techniques test whether specific attributes are encoded within neural network activations by fitting simple classifiers or regressors (often linear) to internal representations (Belinkov 2022). Interventions extend probing by directly manipulating activations during inference, verifying whether such manipulations functionally affect model outputs (Li et al. 2023; Kim, Evans, and Schein 2025). Together, these interpretability methods provide principled tools for localizing and controlling abstract concepts like stereotypes, making internal biases transparent and actionable.

Methodology

SCoUT aims to recover and interpret the internal organization of stereotypes in large language models (LLMs), moving beyond surface-level output analysis. As shown in Figure 1, our pipeline consists of three main stages: (1) constructing a latent stereotype utility space via Thurstonian modeling; (2) probing attention heads to localize the encoding of these dimensions; and (3) functional intervention to manipulate stereotype-related output.

Stage 1: Modeling Stereotype Utility Structure

As illustrated in Figure 2, SCoUT first *models* the global, theory-driven stereotype utility structure in three steps: We begin by collecting pairwise preferences and fitting a Thurstonian comparative judgment model; the resulting utilities are then mapped into an interpretable SCM-aligned space.

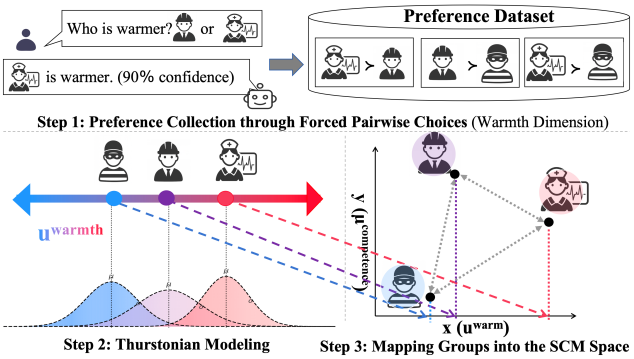


Figure 2: Overview of Stereotype Utility Structure construction via Thurstonian Modeling.

Contrastive Prompting and Preference Collection. To ground the utility structure in observable model behavior, we collect a dataset of forced-choice preferences. For each stereotype dimension $d \in \{\text{Warmth}, \text{Competence}\}$, the model is prompted with two social groups and asked which is higher on d. Aggregating across both orderings and repeated trials ($K = 10$) yields an empirical preference probability:

$$\hat{P}(g_i \succ g_j \mid d) = \frac{1}{K} \sum_{k=1}^{K} y_k, \tag{1}$$

where $y_k = 1$ if g_i is preferred in trial k. The collected dataset provides the raw behavioral signal against which we fit a latent utility model.

Thurstonian Utility Modeling. Following previous work (Mazeika et al. 2025), we assume each group g_i has a latent utility $u_i^{(d)}$ on dimension d. According to Thurstone's Law of Comparative Judgment:

$$P(g_i \succ g_j \mid d) = \Phi\left(\frac{u_i^{(d)} - u_j^{(d)}}{\sqrt{(\sigma_i^{(d)})^2 + (\sigma_j^{(d)})^2}}\right), \tag{2}$$

where $\Phi(\cdot)$ is the standard normal CDF. Given all empirical preferences $\{\hat{P}(g_i \succ g_j \mid d)\}$, we estimate $\{u_i^{(d)}\}$ and $\{\sigma_i^{(d)}\}$ by maximizing the likelihood:

$$\mathcal{L} = \prod_{i<j} p_{i,j}^{\hat{p}_{i,j}} [1 - p_{i,j}]^{1-\hat{p}_{i,j}}, \tag{3}$$

where $p_{i,j} = P(g_i \succ g_j \mid d)$ denote the predicted probability that group g_i is preferred over g_j on dimension d, and $\hat{p}_{i,j} = \hat{P}(g_i \succ g_j \mid d)$ the empirical preference estimated from model outputs. We then z-normalize per dimension to ensure cross-model comparability for subsequent analyses.

To assess generalization, we partition the pairwise comparisons into training and test sets. The recovered utilities are then evaluated by computing prediction accuracy on the held-out test pairs as ($\mathbf{1}$ is the indicator function):

$$\text{Acc} = \frac{1}{N_{\text{pairs}}} \sum_{i<j} \mathbf{1}\left[\text{sign}(u_i^{(d)} - u_j^{(d)}) = \text{sign}(\hat{p}_{i,j} - 0.5)\right]. \tag{4}$$

We further test the stability of the structure by comparing the induced group orderings across models using Spearman's rank correlation. These metrics ensure that the latent utilities are not only a good fit to the collected preferences, but also robust and model-agnostic.

Stereotype Utility Space and Evaluation. Finally, we embed each group as $z_i = (u_i^{(\text{Warmth})}, u_i^{(\text{Competence})})$ in a two-dimensional space $\mathcal{U} \subset \mathbb{R}^2$. This mapping enables structural analysis beyond pairwise comparisons. To examine whether the recovered structure aligns with social-psychological theory, we follow (Fiske et al. 2002) by clustering the embeddings with k-means ($k = 4$):

$$C = \arg\min_{C} \sum_{c=1}^{4} \sum_{z_i \in C_c} \|z_i - m_c\|^2, \tag{5}$$

where m_c is the centroid of cluster C_c. SCM theory further posits that different combinations of high or low warmth and competence elicit distinct emotions towards these groups. For example, groups perceived as high in both dimensions tend to evoke admiration, whereas low warmth and high competence evoke envy, high warmth and low competence evoke pity, and low on both evoke contempt. For each cluster C_c, we prompt the model and average the Likert scores for emotions:

$$s_c(e) = \tfrac{1}{|C_c|} \sum_{g_i \in C_c} \text{LikertScore}(g_i, e), \tag{6}$$

with $e \in \{\text{contempt, admiration, pity, envy}\}$. $s_c(e)$ tests whether the clusters correspond to the four SCM quadrants, providing qualitative and quantitative evidence that the learned utility space captures human-theorized social categories.

Stage 2: Probing Internal Representations

After modeling a global stereotype utility space, we then examine whether these utilities are explicitly encoded in the model's *internal activations*.

Attention Head Representation Extraction. For each social group g and dimension $d \in \{\text{Warmth}, \text{Competence}\}$, we construct a neutral prompt containing the group label and a brief dimension description. After tokenization and encoding, we extract the mean-pooled output vector $\mathbf{h}_{g,l,h}$ from each attention head h at layer l:

$$\mathbf{h}_{g,l,h} = \text{MeanPool}\big(\text{HeadOutput}(g, l, h)\big).$$

This representation captures how each head encodes the input group in context.

Linear Probing on Group-Level Utilities. To test whether a given head encodes stereotype utilities, we train a ridge regression probe $f_{l,h}^{(d)}$ to predict the modeled utility $u_g^{(d)}$:

$$\hat{u}_g^{(d)} = f_{l,h}^{(d)}(\mathbf{h}_{g,l,h}) = \mathbf{w}_{l,h}^{(d)\top} \mathbf{h}_{g,l,h} + b. \tag{7}$$

Performance is measured by Spearman's rank correlation:

$$r_{l,h}^{(d)} = \text{Spearman}\Big(\{\hat{u}_g^{(d)}\}, \{u_g^{(d)}\}\Big), \tag{8}$$

where $r_{l,h}^{(d)}$ reflects how well head (l, h) encodes the utility scores across groups. Heads with top-k highest $r_{l,h}^{(d)}$ are designated as dimension-sensitive for subsequent analyses.

Word-Level Probing. To evaluate whether the learned utilities extend beyond specific group labels to more abstract semantic concepts, we adopt the lexicon from previous work (Fraser, Kiritchenko, and Nejadgholi 2022), which provides human-rated warmth and competence scores for common descriptors. For each word w, we insert it into a neutral template (e.g., *"The person is known as* $\{w\}$ *by others."*) and feed the sentence into the model. We then extract the activation $\mathbf{h}_{w,l,h}$ at the token position of w for each head h and layer l. Using the group-level trained probes $f_{l,h}^{(d)}$, we compute the predicted score from each head and then average over the top-k dimension-sensitive heads ($k = 20$):

$$\hat{u}_w^{(d)} = \frac{1}{k} \sum_{(l,h) \in H_d^{\text{top}}} f_{l,h}^{(d)}\big(\mathbf{h}_{w,l,h}\big), \tag{9}$$

where H_d^{top} is the set of top k dimension-sensitive heads. The separation between predicted scores for the top-25 and bottom-25 words on each dimension is then used to evaluate whether these attention heads encode more abstract lexical semantics aligned with the warmth and competence utilities. To quantify the separation between high- and low-rated sets, we compute a mean-score gap for each dimension d:

$$gap_{\text{set}}^{(d)} = avg_{\text{set}}^{(d)}(W_d^+) - avg_{\text{set}}^{(d)}(W_d^-), \tag{10}$$

where

$$avg_{\text{set}}^{(d)}(W_d) = \frac{1}{|W_d|} \sum_{w \in W_d} \hat{u}_w^{(d)} \tag{11}$$

and set $\in \{\text{group}, \text{descriptor}\}$ indicates whether the gap is computed over social groups or lexical descriptors.

Stage 3: Intervention on Sensitive Heads

After confirming that stereotype utilities are encoded in specific heads, we further assess their *functional influence* on generation to show that these internal representations are actionable at inference.

Inference-Time Activation Steering. Following (Li et al. 2023; Kim, Evans, and Schein 2025), we intervene on stereotype-sensitive heads by shifting their activation along the learned probe direction. For a selected head (l, h) at timestep t, the activation is updated as:

$$\mathbf{x}_{l,h}^{(\alpha,t)} = \mathbf{x}_{l,h}^{(t)} + \alpha \cdot \mathbf{w}_{l,h}^{(d)}, \tag{12}$$

where α is a scalar controlling the magnitude and direction of intervention. By varying α, we can systematically modulate the model's output toward higher or lower values on the stereotype dimension d.

Evaluating Functional Steering Effects. We evaluate intervention effectiveness through two complementary setups. First, we conduct qualitative case studies: for each dimension, we provide a scenario prompt and let the model generate short narratives under different α values. By examining generated sentences, we observe whether descriptive framing shifts consistently with the intended steering direction. Second, we adopt a quantitative evaluation using a pretrained classifier by (Wan and Chang 2024) to score generated sentences along communality axis, which is semantically aligned with warmth. For each generated sentence s, we obtain a classifier score $c(s)$; the intervention effect is then measured as the change in mean scores before and after steering:

$$\Delta c(\alpha) = \tfrac{1}{N} \sum_{s \in S_\alpha} c(s) \; - \; \tfrac{1}{N} \sum_{s \in S_0} c(s), \tag{13}$$

where S_0 is the set of sentences with no intervention and S_α is the set under steering strength α. A consistent positive or negative $\Delta c(\alpha)$ indicates that manipulating selected heads is associated with directional changes in model framing along dimension d, as measured by our proxy.

Experiments

Experimental Setup

Following Stereomap (Jeoung, Ge, and Diesner 2023), we adopt a structured set of 98 social groups derived from (Cuddy, Fiske, and Glick 2007; Fiske et al. 2002), spanning race, gender, occupation and ideological affiliations. The groups include both commonly studied stereotypes (e.g., "doctors," "criminals") and intersectional identities (e.g., "poor White"), enabling comprehensive analysis across multiple social dimensions. We fix the random seed to 42 and use a decoding temperature of 1.0 in our experiments. We primarily report results and visualizations based on `LLaMA3.3-70B` (Touvron et al. 2023). Analyses of cross-model accuracy and consistency additionally involve `DeepSeek` (Bi et al. 2024), `LLaMA3.1` (Touvron et al. 2023) and `Qwen2.5` (Yang et al. 2024).

Evaluation of the Constructed Structure

Our experiments aim to validate the feasibility, robustness, and interpretability of our stereotype utility structure. Specifically, we address four key questions: (1) Does comparative modeling yield high-accuracy across models and prompts? (2) Is the recovered utility structure robust and consistent across different model architectures? (3) How does our method compare to direct scoring approaches such as Stereomap? (4) Does the utility space align with classic social psychology theory in a meaningful way?

Modeling Accuracy Across Model Size and Reasoning Prompts To validate the first stage (preference modeling), we measure how well Thurstonian modeling predicts held-out pairwise preferences across model scales and prompt settings. As shown in Table 1, larger models consistently outperform smaller ones, and reasoning prompts deliver substantial gains in prediction accuracy, especially for weaker models (up to $+15.6\%$ absolute for warmth). Notably, apart from the DeepSeek variants, most models achieve over 90% pairwise prediction accuracy on at least one dimension, indicating that the proposed comparative method can faithfully reflect pairwise preference distribution. We attribute these trends to two factors: (1) larger models possess richer, more

Model	Competence	Warmth
DeepSeek-7B	74.7 ± 5.3	68.8 ± 6.3
+ Reasoning Prompt	87.8 ± 3.8 (+13.1)	84.4 ± 4.6 (+15.6)
DeepSeek-67B	86.9 ± 4.2	74.7 ± 5.7
+ Reasoning Prompt	95.8 ± 2.5 (+8.9)	89.5 ± 3.8 (+14.8)
LLaMA3.1-8B	90.3 ± 3.8	89.0 ± 3.8
+ Reasoning Prompt	90.7 ± 3.6 (+0.4)	91.1 ± 3.6 (+2.1)
LLaMA3.1-70B	$\mathbf{95.4 \pm 2.5}$	$\mathbf{95.8 \pm 2.3}$
+ Reasoning Prompt	97.0 ± 2.3 (+1.6)	$\mathbf{95.4 \pm 2.7}$ (-0.4)
Qwen2.5-7B	90.3 ± 3.8	79.3 ± 5.7
+ Reasoning Prompt	89.5 ± 3.6 (-0.8)	84.8 ± 4.6 (+5.5)
Qwen2.5-72B	94.9 ± 2.7	93.7 ± 3.2
+ Reasoning Prompt	$\mathbf{97.9 \pm 1.9}$ (+3.0)	94.5 ± 3.0 (+0.8)

Table 1: Prediction accuracy (%) on Warmth and Competence dimensions. We report the 95% confidence interval. Increment vs base shown in brackets.

stable representations of social groups due to their scale and pretraining, which is also consistent with previous research findings (Mazeika et al. 2025) and (2) reasoning-augmented prompts encourage explicit step-by-step comparison rather than shallow pattern matching, reducing noise in pairwise judgments. Together, these findings support the scalability and robustness of our method as a reliable backbone for downstream structural analysis.

Consistency Across Models Having established accuracy, we then assess whether the recovered utility structure is stable across different LLM architectures by computing Spearman rank correlations between model-induced rankings. Figure 3 demonstrates high cross-model consistency, particularly among larger-scale (around 70B) models and under reasoning prompts ($\bar{\rho}_{\text{warm}} = 0.923$, $\bar{\rho}_{\text{comp}} = 0.969$; Kendall's $W = 0.948/0.979$ and ICC(2,1)$= 0.937/0.963$). This suggests that the recovered utility space captures a model-agnostic latent stereotype structure.

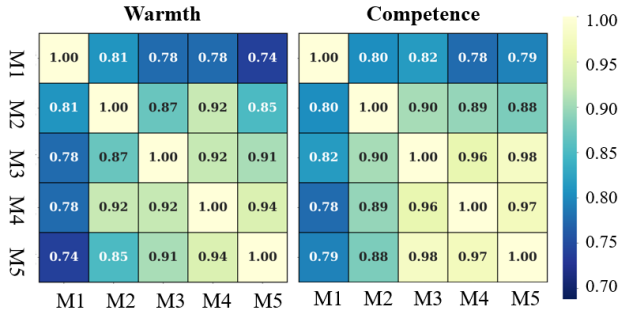
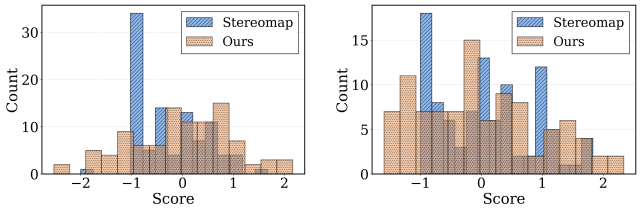


Figure 3: Spearman correlations of group rankings between models with reasoning (left: warmth, right: competence). Models correspond to M1–M5: Qwen-7B, LLaMA2-8B, DeepSeek-67B, LLaMA2-70B, and Qwen-72B.



(a) Warmth dimension (b) Competence dimension

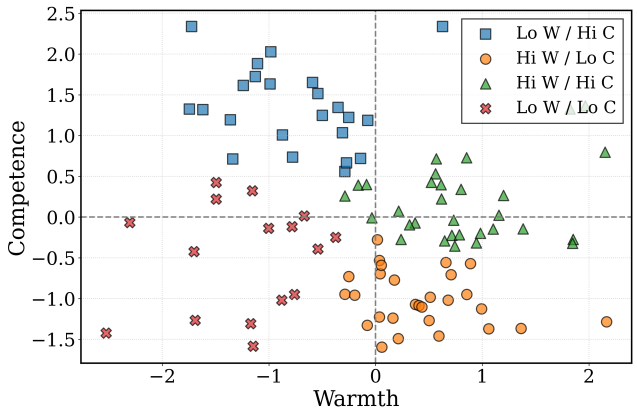
Figure 4: Distribution of group scores from direct scoring (Stereomap-blue) vs. Thurstonian modeling (Ours-orange).

Comparison with Direct Scoring (Stereomap) To contextualize our approach,we contrast the distribution of recovered scores against a direct rating baseline (Stereomap) (Jeoung, Ge, and Diesner 2023).

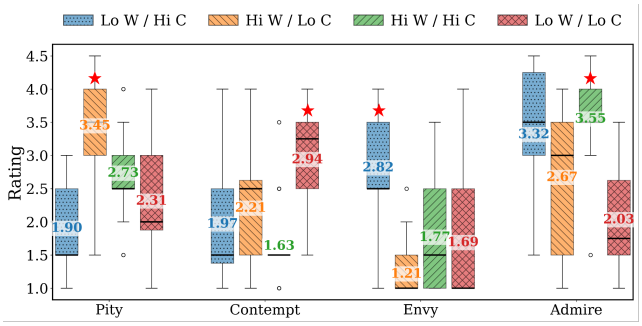
As shown in Figure 4, direct scoring (Stereomap) exhibits strong *anchoring* and *compression* effects on both dimensions. For example, over one third of all groups receive identical warmth scores (Figure 4a), resulting in a narrow range ($[-2.0, 1.67]$) and a smaller standard deviation (Std $= 0.708$). In contrast, our Thurstonian modeling yields standardized distributions (mean 0, Std $= 1.0$ by construction) with wider ranges (e.g., $[-2.53, 2.17]$ for warmth) and finer granularity across groups. A similar pattern holds on the competence dimension (Figure 4b). These quantitative differences show that the comparative method not only mitigates the anchoring bias inherent in Likert-style scores but also preserves fine-grained distinctions on both dimensions. The normalization in our model allows scores to be directly comparable across dimensions and models, while the broader variance demonstrates its ability to capture subtle inter-group differences that direct scoring compresses.

Interpretability and Social Psychology Alignment Finally, we evaluate whether the recovered utility space aligns with social-psychological theory by clustering groups in the 2D space and comparing with canonical SCM quadrants. Following previous works (Fiske et al. 2002; Jeoung, Ge, and Diesner 2023), we apply k-means clustering ($k = 4$) to embeddings $(u_{\text{warmth}}, u_{\text{competence}})$ obtained from Thurstonian modeling. As shown in Figure 5a, the layout naturally separates into four regions: Cluster 3 (top right) includes high-warmth/high-competence groups such as *doctors* and *teachers*; Cluster 2 (top left) groups low-warmth/high-competence entities like *CEOs* and *lawyers*; Cluster 1 (bottom right) contains high-warmth/low-competence groups such as *housekeepers* and *southerners*; and Cluster 4 (bottom left) includes low-warmth/low-competence groups such as *criminals* and *goths*.

To validate whether these clusters reflect SCM's affective predictions, we quantitatively test the alignment between clusters and SCM emotions. For each group, we elicit Likert ratings on pity, contempt, envy, and admiration. As summarized in Figure 5b, each cluster exhibits a clear peak on one emotion: Cluster 1 on pity (3.26±0.68), Cluster 4 on contempt (2.62±0.95), Cluster 2 on envy (2.52±1.02), and Cluster 3 on admiration (3.58±0.65). This alignment between



(a) Visualization of embeddings in stereotype utility space, colored by k-means cluster ($k = 4$).



(b) Likert ratings across clusters (numeric labels indicate the mean score within each cluster; star marks the cluster with the highest mean for the corresponding emotion).

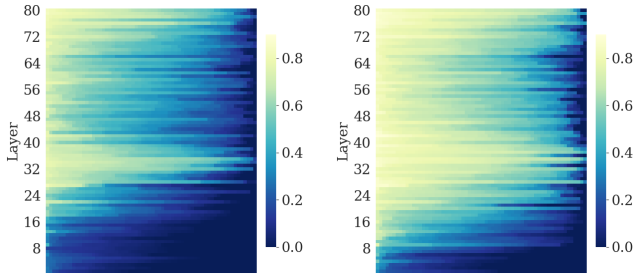
Figure 5: Alignment with social-psychological theory.

cluster-specific emotion and theoretical quadrants provides converging evidence that the recovered utility space captures the affective structure predicted by SCM.

Summary These experiments demonstrate that SCoUT not only models pairwise stereotype preferences with high fidelity, but also reconstructs a stable and generalizable utility space across prompts and models. The learned structure yields finer-grained and psychologically interpretable distinctions than direct scoring methods, serving as a robust foundation for locating how such social dimensions are internally represented.

Probing Attention Heads

While the previous section establishes that our modeling produces accurate and interpretable stereotype utility structure, a key question remains: *Is this structure genuinely encoded within model's internal representations, or merely reflected in its surface behavior?* To address this, we conduct attention probing experiments to test whether, where, and how LLM attention heads internally encode the learned stereotype utility dimensions, thereby connecting external behavioral findings with mechanistic interpretability.



(a) Warmth (b) Competence

Figure 6: Probing correlations across layers and heads.

Head-Wise Probing: Predicting Utility from Attention For each attention head in every layer, we fit a linear probe to predict group utility scores on both dimensions and measure performance by Spearman's correlation between predicted and true utilities. As shown in Figure 6, strong correlations are concentrated in upper–middle layers, peaking at $r = 0.83$ (warmth) and $r = 0.90$ (competence). The top 1% of heads average $\bar{r} = 0.728/0.794$, indicating that stereotype-utility information is sparsely localized rather than uniformly distributed.

Generalization to Human-Annotated Stereotype Words To further test whether heads encode semantic information about dimensions, we evaluate them on annotated descriptors from (Fraser, Kiritchenko, and Nejadgholi 2022). Figure 7 compares the predicted scores for top and bottom-10 groups and top and bottom-25 descriptors from an SCM-aligned lexicon using the top-20 sensitive heads. Across both dimensions, the predicted scores for high-rated words are consistently higher than for low-rated words, confirming that the attention heads generalize beyond training groups to lexical-level cues. Interestingly, the separation is asymmetric: at group level, competence shows a larger gap ($gap_{group}^{(competence)} = 0.125$ vs. $gap_{group}^{(warm)} = 0.087$), whereas at the semantic descriptor level, warmth shows a larger gap ($gap_{descriptor}^{(warm)} = 0.081$ vs. $gap_{descriptor}^{(competence)} = 0.038$). This suggests that competence-sensitive heads are more robustly tied to group identity, while warmth-sensitive heads are more readily related to semantic information. Findings from social psychology (Cuddy, Fiske, and Glick 2008) may partly explain this asymmetry: Human competence judgments often rely on status-related *group knowledge* (e.g., profession or social class), whereas warmth judgments depend on *contextual cues* such as competition or cooperation. In large-scale human-authored corpora used for LLM training, this distinction is implicitly reflected: status information about groups frequently co-occurs with competence-related descriptions, allowing the model to learn group-level competence associations. In contrast, contextual cues for warmth cannot be inferred solely from group status, so corpora tend to include more direct affective descriptors attached to groups, enabling the model to encode warmth more transparently through lexical indicators.

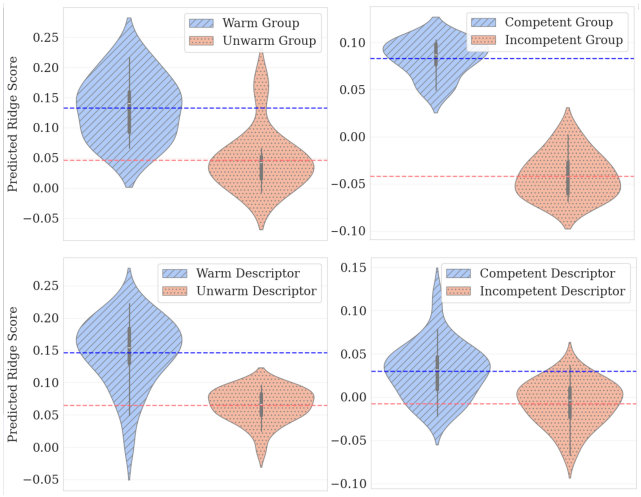


Figure 7: Predicted head scores for human-annotated SCM dictionary words. High-warmth/competence words (blue) receive higher probe scores than low-scoring words (orange), indicating the probe's generalization to out-of-distribution stereotype cues.

Summary These experiments confirm that SCoUT's modeled utilities are explicitly encoded within the model's attention mechanisms and generalize beyond the training groups to human-interpretable words. Moreover, the observed asymmetry between warmth and competence enriches our understanding of how different stereotype dimensions are represented internally, providing evidence that SCoUT not only bridges behavioral measurement and interpretability, but also surfaces psychologically meaningful distinctions in how LLMs encode social knowledge.

Evaluating Functional Steering Effects

Having established that attention heads encode interpretable stereotype utility dimensions, we move from internal diagnosis to behavioral control by investigating whether these internal representations are *functionally actionable*.

Qualitative Case Study: Steering Social Behavior Generation We first conducted qualitative case studies using hand-crafted prompts designed to elicit warmth- or competence-related behaviors. For example, under a warmth prompt (*"Imagine a CEO meeting a stranger on the street..."*), steering with $\alpha = -50$ led to guarded and cautious responses (e.g., "...*expression mixed with a hint of caution...*"), while $\alpha = +50$ produced overt friendliness and proactive offers of help (e.g., "...*being a natural communicator...*"). Similarly, under a competence prompt (*"Describe how an intern presents a research talk..."*), steering from $\alpha = -50$ to $\alpha = +50$ shifted the output from hesitant and self-doubting descriptions to confident, organized presentation behaviors. These clear directional changes provide intuitive, human-readable evidence that modifying activations along the probe direction directly modulates stereotype-relevant content.

Race	**Gender**	$c(-30)$	$c(0)$	$c(30)$	$\Delta c(-30)$	$\Delta c(30)$
White	Male	36.43	61.25	90.83	-24.82	+29.58
	Female	42.36	62.29	91.88	-19.93	+29.58
Black	Male	43.33	63.33	76.77	-20.00	+13.44
	Female	39.38	58.45	83.81	-19.07	+25.36
Hispanic	Male	57.02	49.38	86.43	+7.65	+37.05
	Female	30.89	56.31	81.76	-25.42	+25.45
Asian	Male	34.20	61.04	91.77	-26.84	+30.73
	Female	41.88	52.29	88.42	-10.41	+36.13
Average		**40.44**	**57.54**	**86.71**	**-17.10**	**+29.17**

Table 2: Professor review task results. Per communal scores $c(\alpha)$ and relative changes $\Delta c(\alpha)$ compared to $\alpha = 0$.

Quantitative Evaluation: Controlling Communality We conduct a professor-review writing task and report the communal scores $c(\alpha)$ for different demographic subgroups under intervention strengths $\alpha = -30, 0, 30$ in Table 2. On average across all demographic subgroups, steering in the negative direction (α = -30) reduces communal framing by 17.10 points, while steering in the positive direction (α = 30) increases it by 29.17 points. A paired analysis confirms that both directional shifts are statistically reliable (Cohen's d = 1.5 and 3.8), indicating a robust and predictable relationship between manipulating these heads and the resulting changes in communal framing. Together with the qualitative case studies, this quantitative analysis provides strong evidence that the stereotype-sensitive heads identified by our framework are not merely correlational artifacts. Instead, they represent an actionable link between internal model representations and downstream social behaviors, validating the utility of SCoUT for interpreting the functional structure of stereotypes in LLMs.

Summary These experiments support the validity of the SCoUT framework, confirming that the identified stereotype-sensitive heads exert functional influence over generation. By demonstrating that targeted internal modifications lead to predictable behavioral shifts, we close the loop from modeling and probing to a functional validation of the entire diagnostic process.

Conclusion and Future Work

In this work, we present SCoUT, a unified, theory-grounded framework for diagnosing and interpreting stereotype structures in LLMs. Moving beyond surface-level auditing, SCoUT reconstructs a globally consistent utility space and shows that these dimensions are casually encoded in identifiable attention heads. Our key contribution is reframing stereotype auditing from detection to structural diagnosis. This representational perspective enables more transparent analysis and provides a foundation for targeted fairness interventions. While our current steering primarily serves as a mechanistic verification, future work can integrate these diagnostic insights with permanent mitigation approaches such as fine-tuning or model editing.

Acknowledgments

This work was supported by NSFC Project (No. 62176061) and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. We would like to thank the reviewers for their thoughtful feedback and constructive suggestions, which greatly helped improve the quality of this work. We also appreciate all those who offered encouragement and guidance during the most uncertain times of this research. This work would not have been possible without their support.

References

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

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; Gao, H.; Gao, K.; Gao, W.; Ge, R.; Guan, K.; Guo, D.; Guo, J.; Hao, G.; Hao, Z.; He, Y.; Hu, W.; Huang, P.; Li, E.; Li, G.; Li, J.; Li, Y.; Li, Y. K.; Liang, W.; Lin, F.; Liu, A. X.; Liu, B.; Liu, W.; Liu, X.; Liu, X.; Liu, Y.; Lu, H.; Lu, S.; Luo, F.; Ma, S.; Nie, X.; Pei, T.; Piao, Y.; Qiu, J.; Qu, H.; Ren, T.; Ren, Z.; Ruan, C.; Sha, Z.; Shao, Z.; Song, J.; Su, X.; Sun, J.; Sun, Y.; Tang, M.; Wang, B.; Wang, P.; Wang, S.; Wang, Y.; Wang, Y.; Wu, T.; Wu, Y.; Xie, X.; Xie, Z.; Xie, Z.; Xiong, Y.; Xu, H.; Xu, R. X.; Xu, Y.; Yang, D.; You, Y.; Yu, S.; Yu, X.; Zhang, B.; Zhang, H.; Zhang, L.; Zhang, L.; Zhang, M.; Zhang, M.; Zhang, W.; Zhang, Y.; Zhao, C.; Zhao, Y.; Zhou, S.; Zhou, S.; Zhu, Q.; and Zou, Y. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint*, arXiv:2401.02954. Submitted on 5 Jan 2024.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv preprint arXiv:1607.06520*.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Cuddy, A. J.; Fiske, S. T.; and Glick, P. 2007. The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, 92(4): 631.

Cuddy, A. J.; Fiske, S. T.; and Glick, P. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology*, 40: 61–149.

De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 120–128. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Fiske, S. T.; Cuddy, A. J.; Glick, P.; and Xu, J. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6): 878–902.

Fraser, K. C.; Kiritchenko, S.; and Nejadgholi, I. 2022. Computational modeling of stereotype content in text. *Frontiers in artificial intelligence*, 5: 826207.

Jeoung, S.; Ge, Y.; and Diesner, J. 2023. Stereomap: Quantifying the awareness of human-like stereotypes in large language models. *arXiv preprint arXiv:2310.13673*.

Kim, J.; Evans, J.; and Schein, A. 2025. Linear Representations of Political Perspective Emerge in Large Language Models. *arXiv preprint arXiv:2503.02080*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.

Mazeika, M.; Yin, X.; Tamirisa, R.; Lim, J.; Lee, B. W.; Ren, R.; Phan, L.; Mu, N.; Khoja, A.; Zhang, O.; et al. 2025. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. *arXiv preprint arXiv:2502.08640*.

Morehouse, K.; Swaroop, S.; and Pan, W. 2025. Rethinking LLM Bias Probing Using Lessons from the Social Sciences. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP*, 5356–5371.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on EMNLP*, 1953–1967.

Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C. A.; and Kersting, K. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3): 258–268.

Thurstone, L. L. 1927. A law of comparative judgment. *Psychological Review*, 34(4): 273–286.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wan, Y.; and Chang, K.-W. 2024. White Men Lead, Black Women Help? Benchmarking Language Agency Social Biases in LLMs. *arXiv preprint arXiv:2404.10508*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.